

ESTADÍSTICA INDUSTRIAL  
(Temas de estadística para Ingenieros)

Rosa Rodríguez Huertas, Antonio Gámez Mellado,  
Luis Marín Trechera y Santiago Fandiño Patiño

Escuela Superior de Ingeniería.  
Universidad de Cádiz

Diciembre de 2005



# Índice General

<b>I</b>	<b>ESTADÍSTICA BÁSICA</b>	<b>11</b>
<b>1</b>	<b>Estadística Descriptiva</b>	<b>13</b>
1.1	Introducción a la Estadística . . . . .	13
1.1.1	Concepto . . . . .	13
1.1.2	Aplicaciones de la estadística . . . . .	15
1.1.3	Notas Históricas . . . . .	17
1.1.4	Paquetes estadísticos . . . . .	19
1.2	Estadística Descriptiva Unidimensional . . . . .	20
1.2.1	Conceptos básicos . . . . .	20
1.2.2	Tipos de muestreo . . . . .	21
1.2.3	Presentación de los datos: tablas y representaciones gráficas . . . . .	23
1.2.4	Representaciones gráficas . . . . .	26
1.2.5	Medidas de posición . . . . .	30
1.3	Estadística Descriptiva bidimensional . . . . .	35
1.3.1	Introducción: distribución conjunta y tablas de doble entrada . . . . .	35
1.3.2	Representaciones gráficas . . . . .	38
1.3.3	Distribuciones marginales. Distribuciones condicionadas	39
1.4	Regresión y Correlación . . . . .	41
1.4.1	Independencia de variables estadísticas. Dependencia funcional y dependencia estadística . . . . .	41
1.4.2	Medias, varianzas y covarianzas . . . . .	42
1.4.3	Ajustes. Método de mínimos cuadrados . . . . .	44
1.4.4	Regresión lineal mínimo cuadrática . . . . .	45
1.4.5	Coefficiente de determinación. Coeficiente de correlación lineal . . . . .	48
1.5	EJERCICIOS PROPUESTOS . . . . .	50
<b>2</b>	<b>Cálculo de probabilidades</b>	<b>57</b>
2.1	Introducción a la teoría de la probabilidad . . . . .	57

2.2	Definiciones de Probabilidad . . . . .	58
2.2.1	Fenómenos aleatorios . . . . .	58
2.2.2	Relaciones y Operaciones con sucesos . . . . .	59
2.2.3	Propiedades de las operaciones entre sucesos . . . . .	61
2.2.4	Definición frecuentista de probabilidad . . . . .	61
2.2.5	Definición clásica de probabilidad . . . . .	62
2.2.6	Propiedades de la probabilidad. . . . .	63
2.2.7	Definición axiomática de probabilidad. . . . .	65
2.3	Recursos para el cálculo de probabilidades . . . . .	68
2.3.1	Regla de multiplicación . . . . .	68
2.3.2	Diagramas de árbol . . . . .	69
2.3.3	Combinatoria . . . . .	69
2.3.4	De lo particular a lo general . . . . .	74
2.3.5	La probabilidad geométrica . . . . .	75
2.4	Probabilidad condicionada. . . . .	77
2.4.1	Independencia de un par de sucesos . . . . .	77
2.4.2	Independencia de más de dos sucesos . . . . .	79
2.5	Teorema de la Probabilidad Total. . . . .	80
2.6	Teorema de Bayes. . . . .	83
2.7	EJERCICIOS PROPUESTOS . . . . .	84
2.7.1	Repaso de combinatoria . . . . .	84
2.7.2	Probabilidad . . . . .	85
<b>3</b>	<b>Distribuciones Estadísticas</b>	<b>91</b>
3.1	Introducción al concepto de variable aleatoria . . . . .	91
3.2	Variables aleatorias discretas . . . . .	92
3.2.1	Función de probabilidad . . . . .	92
3.2.2	Función de distribución de una variable aleatoria discreta . . . . .	93
3.2.3	Media y varianza de una variable aleatoria discreta . . . . .	95
3.2.4	La distribución uniforme discreta . . . . .	95
3.2.5	La distribución de Bernoulli . . . . .	96
3.2.6	La distribución binomial . . . . .	97
3.2.7	La distribución geométrica . . . . .	99
3.2.8	La distribución de Poisson . . . . .	100
3.2.9	La distribución hipergeométrica . . . . .	103
3.3	Variables aleatorias continuas . . . . .	106
3.3.1	Función de densidad de probabilidad . . . . .	106
3.3.2	Función de distribución de una variable aleatoria continua . . . . .	107
3.3.3	Media y varianza de una variable aleatoria continua . . . . .	108
3.3.4	La distribución uniforme . . . . .	108

3.3.5	La distribución exponencial . . . . .	110
3.3.6	La distribución normal . . . . .	111
3.3.7	Distribuciones asociadas a la distribución normal . . .	115
3.3.8	La distribución de Weibull . . . . .	118
3.3.9	La distribución triangular . . . . .	119
3.3.10	La distribución gamma . . . . .	120
3.3.11	La distribución beta . . . . .	122
3.4	EJERCICIOS PROPUESTOS . . . . .	124
<b>4</b>	<b>Simulación y Teorema Central del Límite</b>	<b>131</b>
4.1	Introducción a la Simulación . . . . .	131
4.2	Un ejemplo muy sencillo . . . . .	132
4.3	Método Montecarlo . . . . .	135
4.4	Notas históricas sobre el Método Montecarlo . . . . .	136
4.5	Generación de números aleatorios . . . . .	137
4.5.1	Propiedades de un buen generador de números aleatorios	138
4.6	Método de la transformación inversa . . . . .	138
4.6.1	Método de la transformación inversa aplicado a la dis- tribución exponencial . . . . .	139
4.7	Simulación de una cola con una línea y un servidor . . . . .	139
4.8	Integración Montecarlo . . . . .	145
4.9	El Teorema Central del Límite . . . . .	152
4.9.1	Convergencia en distribución (o en ley) . . . . .	153
4.9.2	Aplicaciones del Teorema Central del Límite . . . . .	154
4.10	Simulación del Teorema Central del Límite . . . . .	156
4.11	EJERCICIOS PROPUESTOS . . . . .	157
<b>5</b>	<b>Inferencia Estadística.</b>	<b>159</b>
5.1	Introducción a la Inferencia Estadística . . . . .	159
5.2	Tipos de estimación . . . . .	160
5.3	Estadísticos y Estimadores . . . . .	161
5.4	Propiedades de los estimadores . . . . .	162
5.5	Estimadores insesgados de la media y la varianza . . . . .	162
5.6	Distribución de los estadísticos muestrales . . . . .	163
5.7	Intervalos de confianza para la media . . . . .	163
5.7.1	Con varianza conocida . . . . .	164
5.7.2	Con varianza desconocida . . . . .	166
5.8	Intervalos de confianza para la varianza . . . . .	167
5.9	Contrastes de hipótesis . . . . .	169
5.9.1	Introducción: . . . . .	169
5.9.2	Conceptos generales . . . . .	171
5.10	Prueba de hipótesis para la media . . . . .	175

5.10.1	Poblaciones normales . . . . .	175
5.11	Distribuciones no normales . . . . .	178
5.12	Prueba de hipótesis para una proporción . . . . .	179
5.13	Prueba de hipótesis para la varianza . . . . .	180
5.14	Prueba de bondad de ajuste . . . . .	181
5.15	Contrastes de hipótesis para dos poblaciones . . . . .	184
5.15.1	Test para comparar la igualdad entre las varianzas de dos poblaciones . . . . .	184
5.15.2	Comparación entre las medias de dos poblaciones normales . . . . .	186
5.15.3	Test para la diferencia entre dos proporciones . . . . .	190
5.15.4	Test de independencia de dos variables cualitativas . . . . .	190
5.16	EJERCICIOS PROPUESTOS . . . . .	193
<b>II CONTROL DE CALIDAD</b>		<b>197</b>
<b>6</b>	<b>Control de Calidad. Control por atributos.</b>	<b>199</b>
6.1	Introducción. . . . .	199
6.2	Control por atributos. Capacidad . . . . .	200
6.3	Gráficos de control. . . . .	202
6.3.1	Fracción de defectos ó número de defectos. Interpretación	202
6.3.2	Un ejemplo de gráficos de Control con Statgraphics . . . . .	202
6.4	EJERCICIOS PROPUESTOS . . . . .	204
<b>7</b>	<b>Control por variables</b>	<b>207</b>
7.1	Control por variables según una distribución Normal. . . . .	207
7.2	Control de la variabilidad del sistema . . . . .	210
7.2.1	Límites de control para la varianza. Comparación con un estandar . . . . .	210
7.2.2	Límites de control para el recorrido . . . . .	211
7.2.3	Resumen: . . . . .	211
7.3	Limites de tolerancia . . . . .	212
7.3.1	Capacidad de un proceso . . . . .	213
7.4	EJERCICIOS PROPUESTOS . . . . .	214
<b>8</b>	<b>Control de Recepción</b>	<b>223</b>
8.1	Introducción. . . . .	223
8.2	El control simple por atributos . . . . .	223
8.3	Diseño de un plan de muestreo . . . . .	225
8.4	Planes de muestreos tabulados . . . . .	226
8.5	Muestreo doble y múltiple . . . . .	228

8.6 Muestreo Secuencial . . . . .	229
8.7 EJERCICIOS PROPUESTOS . . . . .	231

**III FIABILIDAD 233**

**9 Fiabilidad y Fallos 235**

9.1 Introducción. Fallos y clases de fallos . . . . .	235
9.2 Distribución de los fallos y función de fiabilidad . . . . .	236
9.3 Vida media y tasa de fallo . . . . .	237
9.4 La vida media en función de la fiabilidad . . . . .	238
9.5 EJERCICIOS PROPUESTOS . . . . .	240

**10 Distribuciones de tiempos de fallos 243**

10.1 Introducción. La curva de bañera. . . . .	243
10.2 La curva de bañera . . . . .	243
10.3 La distribución exponencial . . . . .	244
10.4 La distribución normal . . . . .	245
10.5 La distribución log-normal . . . . .	248
10.6 La distribución de Weibull . . . . .	248
10.6.1 Uso del papel probabilístico de Weibull . . . . .	251
10.7 La distribución gamma . . . . .	253
10.8 El Test Chi cuadrado de bondad de Ajuste . . . . .	254
10.9 EJERCICIOS PROPUESTOS . . . . .	257

**11 Modelos para sistemas. Redundancia 263**

11.1 Introducción. Modelo matemático. . . . .	263
11.2 Redundancia . . . . .	264
11.3 Sistemas en serie. . . . .	264
11.4 Sistemas en paralelo. . . . .	266
11.5 Redundancia activa parcial . . . . .	268
11.6 Combinaciones serie-paralelo . . . . .	270
11.7 Fiabilidad de sistemas Complejos . . . . .	271
11.8 Redundancia secuencial . . . . .	273
11.9 Redundancia secuencial con bloques exponenciales. . . . .	274
11.10 EJERCICIOS PROPUESTOS . . . . .	276

**12 Inferencia con pruebas de vida 283**

12.1 Pruebas o ensayos de vida . . . . .	283
12.1.1 Tipos de pruebas de vida . . . . .	284
12.2 Estimadores de máxima verosimilitud. . . . .	285
12.3 Estimación de la vida media útil . . . . .	286
12.3.1 Ensayos terminados a r fallos con reposición. . . . .	286

12.3.2	Ensayos terminados a $r$ fallos sin reposición . . . . .	288
12.3.3	Ensayos terminados a tiempo $T$ con reposición . . . . .	289
12.3.4	Ensayos terminados a tiempo $T$ sin reposición . . . . .	290
12.3.5	Contraste con hipótesis alternativa. Planes de muestreo. Curva característica. . . . .	290
12.4	Nota sobre planes de muestreo tabulados para pruebas de vida.	294
12.5	Pruebas de vida para el periodo de desgaste . . . . .	294
12.5.1	Ensayo hasta el fallo de todas las unidades. Estimación de la vida media en el periodo de desgaste. Estimación de la varianza. . . . .	294
12.5.2	Intervalos de confianza y contraste de hipótesis para la vida media y la varianza . . . . .	294
12.6	Parámetros de la distribución de Weibull. . . . .	296
12.7	EJERCICIOS PROPUESTOS . . . . .	298
<b>IV</b>	<b>ANÁLISIS DE LA VARIANZA</b>	<b>301</b>
<b>13</b>	<b>Análisis de la varianza con un factor</b>	<b>303</b>
13.1	Generalidades sobre el diseño de experimentos . . . . .	303
13.2	Análisis de varianza con un factor. . . . .	304
13.2.1	Introducción . . . . .	304
13.2.2	Diseño completamente aleatorizado . . . . .	305
13.3	El modelo del análisis de la varianza . . . . .	306
13.3.1	Suma de cuadrados . . . . .	307
13.3.2	Grados de libertad de las sumas de cuadrados y medias cuadráticos . . . . .	308
13.3.3	Construcción del test de hipótesis . . . . .	309
13.4	Comparación de dos muestras . . . . .	311
13.5	Validación del modelo . . . . .	311
13.5.1	Coefficiente de Determinación . . . . .	312
13.6	Comparaciones parciales entre las medias . . . . .	313
13.6.1	Intervalos de confianza individuales para la media de cada grupo . . . . .	313
13.6.2	Comparaciones multiples . . . . .	313
13.7	EJERCICIOS PROPUESTOS . . . . .	317
<b>14</b>	<b>Análisis de la varianza con varios factores</b>	<b>321</b>
14.1	Introducción . . . . .	321
14.2	Modelo de Análisis de Varianza con dos factores. . . . .	322
14.2.1	Descripción del modelo con interacción . . . . .	322
14.2.2	Descripción del test de hipótesis . . . . .	324



14.2.3 Descripción del modelo sin interacción . . . . .	325
14.3 Diseño por bloques completos al azar . . . . .	330
14.4 Principios básicos para el diseño de experimentos . . . . .	334
14.5 Otros diseños de experimentos . . . . .	336
14.6 EJERCICIOS PROPUESTOS . . . . .	339

**V ANÁLISIS MULTIVARIANTE 343**

**15 Análisis multivariante. Regresión 345**

15.1 Generalidades . . . . .	345
15.2 Regresión Múltiple . . . . .	346
15.2.1 Introducción . . . . .	346
15.2.2 Modelo de Regresión lineal . . . . .	347
15.3 Estimación mínimo cuadrática . . . . .	348
15.4 Test de hipótesis para la regresión lineal multiple . . . . .	350
15.5 Intervalos de confianza para los coeficientes . . . . .	352
15.6 Predicción. Intervalos de confianza . . . . .	353
15.7 Modelos de regresión polinomiales . . . . .	354
15.8 Regresión paso a paso. . . . .	355
15.9 Estudio de un caso práctico . . . . .	358
15.10 EJERCICIOS PROPUESTOS . . . . .	360

**16 Diversas técnicas de Análisis Multivariante . 363**

16.1 El Análisis Discriminante . . . . .	363
16.2 El Análisis Cluster o de conglomerados . . . . .	365
16.3 El Análisis de componentes Principales . . . . .	368
16.4 El Análisis Factorial . . . . .	372
16.5 EJERCICIOS PROPUESTOS . . . . .	376

**VI SERIES TEMPORALES 379**

**17 Series temporales . Modelos Clásicos. 381**

17.1 Introducción y ejemplos de series temporales . . . . .	381
17.2 Software para el análisis de series temporales . . . . .	385
17.3 Parámetros para el análisis de series temporales . . . . .	386
17.4 Variable de ruido blanco . . . . .	389
17.5 Estudio de la tendencia . . . . .	390
17.5.1 Tendencia polinómica. . . . .	393
17.5.2 Modelos multiplicativos . . . . .	393
17.6 Métodos de suavizado . . . . .	394

17.6.1	Método de las medias móviles . . . . .	394
17.6.2	Método de Alisado exponencial simple . . . . .	398
17.6.3	Método de alisado exponencial doble o de Brown . . . . .	401
17.6.4	Método de Holt . . . . .	403
17.7	Análisis de la estacionalidad . . . . .	404
17.7.1	El método de la razón a la media movil . . . . .	405
17.7.2	El método de Holt-Winters . . . . .	408
17.8	EJERCICIOS PROPUESTOS . . . . .	412
<b>18</b>	<b>Series temporales. Modelos Arima.</b>	<b>417</b>
18.1	Procesos estocásticos . . . . .	417
18.2	Estacionaridad Funciones de autocorrelación . . . . .	418
18.3	La función de autocorrelación parcial. . . . .	420
18.4	Procesos lineales . . . . .	420
18.5	Estudio teórico del modelo AR(1) . . . . .	421
18.6	Análisis del modelo AR(1) . . . . .	423
18.6.1	Identificación del modelo AR(1) . . . . .	423
18.6.2	Estimación del modelo AR(1) . . . . .	424
18.6.3	Predicción en el modelo AR(1) . . . . .	426
18.6.4	Validación . . . . .	427
18.7	Identificación de los modelos ARMA(p,q) . . . . .	432
18.8	Procesos no estacionarios . . . . .	432
18.8.1	Modelos ARIMA(p,d,q). Eliminación de la tendencia . . . . .	432
18.8.2	Eliminación de la estacionalidad . . . . .	433
18.9	EJERCICIOS PROPUESTOS . . . . .	439

Unidad Temática I

**ESTADISTICA BÁSICA**



# Tema 1

## Estadística Descriptiva

*“El pensamiento estadístico será un día tan necesario para el ciudadano eficiente como la capacidad de leer y escribir”.*

*(H.G. Wells)*

*”Llegará un día en que la Estadística ocupe en la enseñanza un puesto. ligeramente posterior al de la Aritmética”*

*(L. H. C Tippett, 1947). (Discipulo de Fisher y de Pearson)*

### 1.1 Introducción a la Estadística

#### 1.1.1 Concepto

Desde un punto de vista muy primitivo usamos la estadística continuamente en nuestra vida. A veces oímos frases como las siguientes: “No voy a comprar todavía un ordenador porque espero que baje de precio”, “No voy a salir a las diez de la noche porque es casi seguro que aún no habrá salido ninguno de mis amigos”, “El precio de las casas subirá para el año que viene más del 10%”... Expresamos opiniones sobre muchos temas: Los trenes llegan a menudo con retraso, las mujeres conducen peor que los hombres, ciertos profesores suspenden mucho. También hay opiniones sobre algunos temas que se discuten en las conversaciones y a veces en los medios de comunicación: ¿Los catalanes son peseteros?, ¿Los andaluces somos vagos? Las respuestas a estas preguntas dependen de la experiencia personal de cada individuo, que la interpreta subjetivamente según su estado de ánimo, su situación social, su educación o su ideología. Como es natural, es difícil ponerse de acuerdo. Algo similar sucede en el campo de las ciencias experimentales. Si una cierta propiedad se presenta en un número finito de experimentos el científico pretenderá declarar esta propiedad experimental en forma de ley

general. Pero los experimentadores pueden cometer errores en la realización de experimentos, el material puede sufrir variaciones. Si el experimentador desea comprobar una hipótesis en la que confía, inconscientemente tenderá a dar más importancia a los datos que la corroboran que a los que la rebaten. Por lo tanto deberá seleccionar los datos e interpretarlos de una forma que no sea subjetiva. En todos los casos comentados hay una idea general: se dispone de una información particular que deseamos generalizar.

En situación similar se encuentra el economista, que disponiendo de datos anteriores, desea hacer previsiones sobre las subidas de interés o la variación del índice de precios de consumo, las compañías de seguros que necesitan actualizar los precios de las pólizas, los empresarios que desean organizar la producción de sus fábricas, etc...

La Estadística nos va a ayudar a seleccionar las conclusiones generales más adecuadas a partir de datos parciales y representativos. Distinguiremos los tres campos básicos de la Estadística: La estadística descriptiva, el cálculo de probabilidades y la estadística inferencial.

La **estadística descriptiva** trata del estudio de los datos particulares (la **muestra**). La **estadística inferencial** se ocupa de lo referente a la selección de las conclusiones generales. Pero como estas conclusiones dependen de la muestra considerada, tendremos que considerar la probabilidad de error que se origina por la selección de una muestra inadecuada por no ser suficientemente representativa. Cuestiones de este tipo son las que se resuelven por medio del **cálculo de probabilidades**.

En los razonamientos estadísticos se emplea con frecuencia el **método inductivo**. Existen dos formas principales de pensamiento lógico: deductivo e inductivo. El pensamiento deductivo se debe principalmente a los griegos. Consiste en proponer axiomas, hechos admitidos, y deducir de ellos otras propiedades. El razonamiento inductivo, que es el más usado en las aplicaciones estadísticas, nos conduce a inferir conclusiones generales a partir de hechos experimentales.

En el razonamiento deductivo, usado frecuentemente en Matemáticas, los teoremas se deducen de los axiomas siguiendo las leyes de la Lógica. En este sentido son absolutamente ciertos. En cambio, en el razonamiento inductivo las conclusiones tienen un cierto grado de incertidumbre.

La base del razonamiento inductivo es admitir que los fenómenos de la naturaleza son demasiado complejos para permitir una información completa, así que no podemos recolectar toda la información y debemos contentarnos con la información parcial suministrada por una **muestra**. Las cuestiones principales que uno puede hacerse sobre las muestras son las siguientes:

¿Cómo se describe una muestra de forma útil y clara?

¿Cómo sacar conclusiones a partir de una muestra que sea generalizable al colectivo total?

¿Hasta qué punto son de fiar estas conclusiones?

¿Cómo se deben tomar las muestras para que realicen las funciones anteriores de la forma más eficaz posible?

La materia que responde a la primera pregunta es la Estadística Descriptiva. Este es el tipo de Estadística más divulgado por los medios de comunicación: tablas donde se resumen los datos, gráficas más o menos sugerentes y quizá, algunos valores promedios. También es interesante en esta fase dar algún parámetro que nos indique si los datos son entre sí más o menos parecidos. A estos parámetros se les suele llamar medidas de dispersión. La gente piensa frecuentemente que este mero reflejo de una realidad observada es el único papel de la Estadística. Sin embargo éste es sólo el primer escalón. Estaríamos en la fase que hemos llamado experimental en el método inductivo y aún quedaría la fase consistente en establecer las conclusiones (Estadística Inferencial) y determinar el grado de fiabilidad de estas conclusiones (Cálculo de Probabilidades).

La respuesta a la última pregunta que hemos formulado sobre las muestras se estudia en una rama de la Estadística que se llama Teoría de Muestras. Es fácil darse cuenta que algunas muestras no serán reflejo de la realidad global que pretendemos investigar. Si deseamos estimar el sueldo medio de los trabajadores gaditanos no sería adecuado tener en cuenta solamente el sueldo de las personas que viven en los barrios residenciales, o si queremos tener una idea de la salud de los españoles, no sería lógico investigar exclusivamente en un único hospital. La muestra debe ser representativa de la población que se pretende investigar. Como no todas las muestras van a ser igualmente representativas tendremos que investigar cómo pueden variar todas las posibles muestras entre sí. Otra rama de la Estadística, el Diseño de Experimentos, nos indica cómo deben diseñarse las muestras para extraer la mayor cantidad posible de información minimizando el esfuerzo requerido en la extracción de la muestra.

*En resumen: La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa o cualitativa referente a individuos, grupos, series de hechos, etc. analizar los datos obtenidos y deducir, a partir de este análisis y mediante técnicas propias, conclusiones generales o previsiones para el futuro con un cierto grado de incertidumbre.*

### 1.1.2 Aplicaciones de la estadística

Una de las utilidades de la estadística es dar servicio a los estados. La Estadística influye en las decisiones de los gobiernos y las administraciones estatales. Con sus técnicas, los estados pueden conseguir un conocimiento claro de la población con la que cuenta, permitiéndole, por ejemplo, establecer su política fiscal. En relación con este hecho el estado organiza numerosas

encuestas para el conocimiento de la población como, por ejemplo, el Censo de población. En España, como en otros países, se realizan numerosas encuestas a nivel nacional: la Encuesta de población activa (EPA), la Encuesta de Presupuestos Familiares, las encuestas del CIS...

La realización por parte de los estados del recuento de la población y la investigación sobre algunas características de ésta es muy antigua (los faraones egipcios lograron recopilar, hacia el año 3050 antes de Cristo, muchos datos sobre la población y la riqueza de su país). En concreto, y como puede apreciarse fácilmente, la palabra estadística deriva de la palabra estado.

La Estadística colabora en el conocimiento de las necesidades de la sociedad, permitiendo planificar los servicios sociales de esta, como los hospitales, las subvenciones, las necesidades asistenciales, etc. Para estudiar estas necesidades sociales normalmente se recurre a encuestas, ya que investigar a toda la población sería lento y caro. Las técnicas estadísticas de Muestreo, permiten obtener muestras válidas para que las conclusiones sean extrapolables a toda la población.

Las técnicas de Investigación de Mercados permiten planificar la producción y saber si un producto nuevo, o un nuevo centro comercial va a ser viable económicamente, estudiando el número de personas que prefieren ese tipo de productos o estarían gustosos de comprar en este tipo de establecimiento. También se puede conocer la audiencia en Televisión y Radio, para ver el impacto esperado de las campañas publicitarias.

En Medicina se emplea la estadística para seleccionar, entre un conjunto de tratamientos para una enfermedad, el mejor posible.

Con el estudio de las Series Temporales se puede tener una mejor comprensión del comportamiento aleatorio de los fenómenos meteorológicos que pueden ayudar en la previsión del tiempo, hacer pronósticos sobre el comportamiento en bolsa de ciertas acciones, de las fluctuaciones de las ventas, etc.

En cuanto a la Industria, el Control de Calidad permite seguir la calidad de un producto en todas las fases de la cadena de producción dentro de la fábrica, tomando decisiones correctivas si procede. Como las correcciones se realizan en la fase de aceptación de la materia prima y en el proceso de producción, antes de que el producto esté terminado por completo, se elimina el gasto superfluo de seguir produciendo artículos defectuosos. Esto permite conseguir un producto, no solo de mejor calidad, sino incluso más barato. También es muy útil en la Industria los estudios estadísticos sobre la duración sin fallos de los productos una vez que están siendo usados por el consumidor, para lo cual se emplean las técnicas estadísticas de Fiabilidad, que permiten, entre otras cosas, establecer los periodos ofertados de garantía para el producto, evaluando el costo total esperado para la fábrica por este concepto. Asimismo, en Agricultura se aplican técnicas estadísticas para



estimar los rendimientos obtenidos en una cosecha, o seleccionar qué producto será más rentable económicamente en el mercado.

La Estadística es en la actualidad una herramienta auxiliar para todas las ramas del saber; inclusive en Lingüística se aplican técnicas estadísticas, para atribuir un escrito a un cierto autor o para establecer las características propias de un idioma. Su utilidad se entiende mejor si tenemos en cuenta que los quehaceres y decisiones diarias embargan cierto grado de incertidumbre y que la Estadística ayuda a tomar las decisiones más adecuadas en cada situación reduciendo esta incertidumbre.

### 1.1.3 Notas Históricas

Los comienzos de la Estadística pueden ser hallados en el antiguo Egipto, cuyos faraones lograron recopilar, hacia el año 3050 antes de Cristo, una gran cantidad de datos relativos a la población y la riqueza del país. Así que podemos decir que la Estadística es más antigua que las pirámides de Egipto. Se cree que este registro de la riqueza y de la población se hizo, precisamente, con el objetivo de preparar la construcción de las pirámides.

El libro bíblico de *Números* da referencias de dos censos de la población de Israel y el de *Crónicas* describe el bienestar material de las diversas tribus judías. En China existían registros numéricos similares con anterioridad al año 2000 A.C.

Los griegos clásicos realizaban censos cuya información se utilizaba para recabar información tributaria y estimar los elementos de la población susceptibles de ser militarizados en caso de guerra.

Los romanos fueron, entre los pueblos antiguos, quienes mejor supieron emplear los recursos de la Estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos anotaban nacimientos, defunciones y matrimonios y realizaban registros periódicos del número de cabezas de ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio, como queda registrado en los evangelios.

Aunque Carlomagno, en Francia, y Guillermo el Conquistador, en Inglaterra, intentaron recobrar estas costumbres romanas, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo da Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes aportaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió el comercio internacional existía ya un método capaz de aplicarse a los datos económicos.

En la primera mitad de siglo XVI en Francia se exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de

peste que apareció a fines de la década de 1500, el gobierno inglés comenzó a publicar estadísticas semanales del número de muertes y sus causas. Esa costumbre continuó muchos años, y en 1632 estos *Bills of Mortality* (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt usó estos documentos, que abarcaban treinta años, y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y sobre las proporciones de nacimientos de varones y mujeres que cabía esperar. El trabajo de Graunt, condensado en su obra *Natural and Political Observations...Made upon the Bills of Mortality* (Observaciones Políticas y Naturales ... Hechas a partir de las Cuentas de Mortalidad), fue un esfuerzo innovador en el análisis estadístico. Los eruditos del siglo XVII cultivaron la Estadística Demográfica para responder a la cuestión de saber si la población aumentaba, decrecía o permanecía estática.

El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann. Este investigador se propuso acabar con creencia popular de que en los años terminados en siete moría más gente que en los restantes. Para ello investigó los archivos parroquiales de la ciudad. Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la duración de la vida humana. Sus cálculos sirvieron de base para las tablas de mortalidad que hoy utilizan todas las compañías de seguros.

A mediados del siglo XVII los juegos de azar eran frecuentes en los salones europeos. El caballero De Méré, jugador empedernido, consultó al famoso matemático y filósofo Blaise Pascal (1623-1662) para que le revelara las leyes que controlan el juego de los dados, el cual, interesado en el tema, sostuvo una correspondencia epistolar con Pierre de Fermat (1601-1665) dando origen a la teoría de la probabilidad, que llegaría a constituir la base primordial de la Estadística.

Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos.

Jacques Quételet (1796-1874) interpretó la teoría de la probabilidad para su uso en las ciencias sociales. Quételet fue el primero en realizar la aplicación práctica de todo el método Estadístico entonces conocido, a las diversas ramas de la ciencia.

En el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados

desarrollada por Laplace, Gauss y Legendre. En 1840 Sir Francis Galton partió de una distribución discreta y la fue refinando hasta llegar a una continua similar a la normal. Incluso invento una máquina que permite ilustrar la distribución normal.

El calculo de probabilidades se comenzó a usar en demografía y en la matemática actuarial. La mecánica estadística, que introdujeron Maxwell (1831-1879) y Boltzman dio una justificación a la distribución normal en la teoría de los gases. A finales del siglo XIX, Quételet aplicó por primera vez análisis estadísticos en biología humana y Sir Francis Galton, primo de Darwin, estudio la variación genética humana usando métodos de regresión y correlación. De aquí partió, ya en el siglo XX, el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica inglesa.

Durante los siglos XVIII y XIX la estadística se desarrollo muchísimo, a pesar del estancamiento de la teoría de probabilidades debido a que no se disponía de una definición general de probabilidad. Esta definición no se lograría hasta que Andrey Nickolaevich Kolmogorov (1903-1987) definiera axiomáticamente la probabilidad, hecho que marca el nacimiento de la Estadística Matemática moderna y convierte a la Estadística en una ciencia independiente, ya que hasta este momento su estudio estaba ligado muy íntimamente con sus aplicaciones a las otras ciencias.

En la primera mitad de este siglo Gossett, con el seudónimo de Student, estudió el problema del tratamiento de pequeñas muestras desarrollando el test de Student y Fisher desarrolló el Análisis de Varianza, de gran interés en el diseño de experimentos.

La segunda mitad del siglo XX destacan los trabajos de Wilcoxon, que estudió los pesticidas desarrollando un test no paramétrico para comparar dos muestras, de Kruskal Wallis, que aportó un test no paramétrico para comparar más de dos muestras, de Spearman y Kendall que desarrollaron sendos coeficientes de correlación no paramétricos, de Tukey, que desarrollo procedimientos de comparación múltiple...

La llegada de los ordenadores revolucionó el desarrollo de la estadística, propiciando la aparición de nuevas técnicas. Benzecri, en Francia, y Tukey, en Estados Unidos, fueron pioneros en repensar la estadística en función de los ordenadores, adaptando, mejorando y creando nuevos instrumentos, técnicas analíticas y gráficas para estudiar una gran cantidad de datos.

#### 1.1.4 Paquetes estadísticos

La sociedad genera una gran cantidad de información que necesita dar a conocer, resumir, interpretar y emplear para tomar decisiones. Con el avance de la Informática y la vinculación de ésta con la Estadística se ha conseguido

manejar de manera rápida, fiable y relativamente sencilla estos volúmenes de información, y obtener conclusiones a partir de esta información. En la actualidad, y con la ayuda de la informática, la estadística ha dejado de ser patrimonio exclusivo del estado y de científicos brillantes, pasando a impregnar la sociedad en las vertientes económica, social, industrial, sanitaria, etc. En esta extensión de la estadística, la informática ha jugado un papel fundamental, propiciando el uso de paquetes estadísticos. En el mercado existen paquetes estadísticos muy completos. Destacamos algunos de los más utilizados: STATGRAPHICS, SPSS, SAS, STATISTICA, EXCEL SOLVER y R (éste último gratuito).

## 1.2 Estadística Descriptiva Unidimensional

### 1.2.1 Conceptos básicos

Damos, en primer lugar, algunas definiciones básicas de interés general y que nos ayudarán a clasificar los tipos de datos que se nos presenten.

**Población:** Conjunto sobre el cual se va a realizar la investigación. Está compuesta por elementos. Puede ser de tamaño finito o infinito.

**Muestra:** Subconjunto de la población del que se dispone de información necesaria para realizar el estudio.

**Caracteres:** Cualidades o propiedades de los elementos de una población que son objeto del estudio. Atendiendo a que sean o no medibles, los caracteres se pueden clasificar en **cuantitativos** (o variables) y **cualitativos** (o atributos). Las variables cuantitativas pueden ser a su vez discretas o continuas.

$$\text{caracteres} \begin{cases} \text{cuantitativos} \\ \text{cualitativos} \end{cases} \begin{cases} \text{variables discretas} \\ \text{variables continuas} \end{cases}$$

**Ejemplo 1 :** *Supongamos que se desea investigar ciertas características de los alumnos de un instituto. Se han seleccionado al azar 50 de ellos para realizar una encuesta.*

*Se ha registrado para cada uno de los alumnos seleccionado: su talla, el tipo de estudios realizados por su padre, el número de personas que conviven en su domicilio y su peso.*

Los datos están registrados en la tabla siguiente:

TALLA	ESTUDIOS	HABITANTES	PESO	TALLA	ESTUDIOS	HABITANTES	PESO
1.63	bachiller	2	54.88	1.75	diplomado	2	78.951
1.44	fp	3	33.5	1.68	diplomado	3	67.15
1.9	bachiller	3	89.38	1.72	bachiller	4	71.94
1.58	bachiller	3	53.28	1.65	bachiller	4	64.1
1.38	bachiller	5	36.62	1.8	bachiller	6	76.68
1.8	bachiller	6	79.47	1.94	bachiller	4	96.18
1.64	superior	3	72.47	1.99	primario	4	103.029
1.93	bachiller	5	85.26	1.36	primario	3	44.105
1.95	bachiller	2	102	1.69	primario	5	70.16
1.59	bachiller	2	62.28	1.69	bachiller	4	66.79
1.78	primario	4	81.17	1.51	bachiller	3	57.27
1.96	primario	4	100.61	1.98	bachiller	5	94.71
1.89	bachiller	5	94.43	1.84	bachiller	3	81.25
1.52	bachiller	4	57.96	2.02	bachiller	6	101.3
1.74	diplomado	4	67.86	1.76	primario	4	73.68
1.68	diplomado	5	63.2	1.96	fp	3	90.7
2	diplomado	6	105.01	1.78	fp	3	84.59
1.67	diplomado	3	66.57	1.54	superior	4	53.97
1.46	primario	5	43.46	1.8	diplomado	5	82.13
1.98	primario	4	96.4	1.53	diplomado	4	52.8
1.47	primario	2	42.38	1.74	diplomado	5	77.22
1.74	primario	3	80.41	1.7	primario	3	74.7
1.9	diplomado	4	92.7	1.66	primario	4	69.34
1.65	fp	5	62.81	1.83	primario	5	92.24
1.34	fp	8	39.707	1.52	bachiller	4	61.05

En este caso la población está formada por todos los alumnos del instituto, la muestra por los 50 alumnos seleccionados. Los caracteres seleccionados son: talla, estudios del padre, número de personas que viven en su domicilio y el peso. La talla y el peso son caracteres cuantitativos continuos, los habitantes de la casa es un carácter cuantitativo discreto y los estudios del padre es un carácter cualitativo o atributo.

### 1.2.2 Tipos de muestreo

A la hora de decidir sobre la forma de recoger la información de la muestra se utilizan distintos criterios, originando distintos tipos de muestreos.

### Muestreos aleatorios

Se seleccionan los elementos de la muestra por un procedimiento de azar (un sorteo). El investigador no decide que elementos van a tomar parte de la muestra, aunque debe conocer la probabilidad de selección de cada elemento. Estos tipos de muestreo permiten aplicar las técnicas de inferencia estadística. Entre ellos se usan los siguientes:

**Muestreo aleatorio simple con y sin reemplazamiento:** Todos los elementos de la Población tienen la misma probabilidad de ser incluido en la muestra y la selección de cada uno de los elementos es independiente de la selección de otro. Si cuando se extrae un elemento de la Población para formar parte de la muestra, ya no puede extraerse de nuevo (no se reemplaza en la Población) el muestreo se llama *Muestreo aleatorio simple sin reemplazamiento*. Si por el contrario se devuelve a la Población y puede formar de nuevo parte de la muestra, el muestreo se dice *Muestreo aleatorio simple con reemplazamiento*.

**Muestreo estratificado:** Este muestreo requiere que la Población esté dividida en grupos más o menos homogéneos con respecto a la característica que se investiga. A cada uno de estos grupos se le llama *clase o estrato*. Dentro de cada uno de estos estratos se selecciona la muestra con un muestreo aleatorio simple. La muestra que resulta se llama una *muestra estratificada*.

**Muestreo por conglomerados o Agrupado:** Consiste en dividir la población en grupos parecidos entre sí y seleccionar aleatoriamente un conjunto de estos grupos. Para que sea eficiente los grupos han de ser bastante parecidos entre sí, ya que todos ellos han de ser modelos en miniatura de la población. La diferencia de un grupo con un estrato consiste en que los estratos han de ser diferentes entre sí, aunque homogéneos interiormente. Sin embargo, los grupos son parecidos entre sí, pero interiormente reflejan la variabilidad de la población de la que proceden.

**Muestreo Sistemático:** Se supone que los elementos de la población están ordenados con arreglo a algún criterio. Se seleccionan sucesivamente los elementos de  $k$  en  $k$ , comenzando por un elemento seleccionado aleatoriamente.

**Muestreo Doble Múltiple y Secuencial:** Este tipo de muestreo se usa principalmente en Control de Calidad.

El muestreo Doble es un procedimiento mediante el cual se selecciona en primer lugar una muestra pequeña. Si la información obtenida con esta

muestra nos parece suficiente, hemos terminado. Si esto no fuera así se procede a tomar una segunda muestra, normalmente más grande con la que completaremos la información. En el muestreo múltiple este procedimiento se repite sucesivamente un número finito de veces. Una modificación de este tipo de muestreo múltiple consiste en decidir para cada elemento que se incorpora a la muestra si tomamos un siguiente elemento o ya la muestra extraída es suficiente para nuestro propósito. El número de elementos de la muestras no es conocido a priori, ya que dependerá de la propia muestra ya extraída y de la regla de decisión empleada para cerrar la muestra o seguir muestreando.

### **Muestreos no Aleatorios**

Este tipo de muestreo no permite, rigurosamente hablando, aplicar técnicas de inferencia estadísticas, ya que la formulación de estas técnicas se realiza bajo la hipótesis de la aleatoriedad de las muestras.

**Muestreo Dirigido o Adaptado:** Se seleccionan para formar parte de la muestra elementos, que según la opinión de los encuestadores, sean representativos. Se suele emplear en las primeras fases del estudio para construir una muestra piloto.

**Muestreo por cuotas:** Cada encuestador debe entrevistar a un cierto número de personas de unas características definidas. Por ejemplo: 15 hombres solteros con edades comprendidas entre 25 y 30 años, 22 mujeres casadas de edades comprendidas entre 30 y 50 años, 20 personas con hijos en edad escolar, etc..

**Muestro deliberado:** Se selecciona la muestra en un sector de la Población por comodidad de acceso. Por ejemplo cuando se dispone fácilmente de una lista de personas, como la guía de teléfono, las matrículas de los automóviles, etc..

### **1.2.3 Presentación de los datos: tablas y representaciones gráficas**

Una primera manera en que pueden presentarse los datos es mediante una relación exhaustiva de todas las ocurrencias de la variable. Esto es lo que se conoce como una *tabla de tipo I*. Por ejemplo, si estamos estudiando el número de hermanos que tienen los alumnos de un colegio, se nos podrían presentar los siguientes datos:

1, 2, 1, 0, 0, 1, 3, 1, 0, 2, 1, 2, 0

Esta manera de presentar los datos solo es factible cuando se tiene un número muy pequeño de observaciones. Si el número de observaciones es grande, lo que se hace es agrupar los datos, indicando a continuación el número de ocurrencias de cada uno. Esto es lo que se llama una *tabla de tipo II*. En el ejemplo anterior, la tabla tipo II correspondiente sería:

$x_i$	$n_i$
0	4
1	5
2	3
3	1

Donde  $n_i$  representa el número de veces que se presenta cada una de las observaciones.

Para la variable *estudios* de los datos de la página 21 la tabla de tipo II es la siguiente:

$x_i$	$n_i$
<i>bachiller</i>	20
<i>diplomado</i>	10
<i>fp</i>	5
<i>primario</i>	13
<i>superior</i>	2

Podemos considerar que una tabla tipo I es una tabla tipo II en la que todos los  $n_i$  valen 1. En el caso de que sean muchos los posibles valores que pueda tomar la variable, agrupamos los datos en intervalos. Obtendremos entonces una *tabla de tipo III*. Por ejemplo, al estudiar la talla de los alumnos del instituto de la tabla 3.4 podemos agrupar a los que tengan una talla parecida. En la siguiente tabla se han tomado 8 clases. Cada una de ellas tiene una amplitud de 10 cm. Es decir, que hemos agrupado los datos de la variable talla en intervalos de idéntico tamaño.

Intervalo	marca de clase	frecuencia absoluta
[1.3, 1.4]	1.35	3
(1.4, 1.5]	1.45	3
(1.5, 1.6]	1.55	7
(1.6, 1.7]	1.65	11
(1.7, 1.8]	1.75	11
(1.8, 1.9]	1.85	5
(1.9, 2.0]	1.95	9
(2.0, 2.1]	2.05	1



A cada uno de los intervalos se le denomina *intervalo de clase*, y al punto medio de cada uno lo llamaremos *marca de clase*. La longitud de los intervalos de clase no tiene que ser siempre la misma, aunque es preferible que así sea. Para realizar determinados cálculos (por ejemplo la media), nos será útil pasar de una tabla tipo III a una tabla tipo II, considerando que todas las ocurrencias corresponden a la marca de clase.

Daremos a continuación algunas definiciones de interés. A cada una de las  $n_i$  se le llama *frecuencia absoluta* de la observación  $x_i$ . Si tenemos en total  $n$  observaciones y se presentan  $k$  casos distintos  $x_1, x_2 \dots x_k$ , entonces se cumple que:

$$\sum_{i=1}^k n_i = n$$

donde  $n_i$  es la frecuencia absoluta de cada dato de diferente valor contenido en la muestra.

Es decir que la suma de las frecuencias absolutas de todas las observaciones es, como es natural, el número total de observaciones realizadas (número de elementos de la muestra). Puede comprobarse que la suma de las frecuencias de la tabla anterior es 50, que era el número de alumnos entrevistados.

Llamaremos *frecuencia relativa* de la observación  $x_i$  a

$$f_i = \frac{n_i}{n} \quad \forall i = 1, 2, \dots, k.$$

La frecuencia relativa es por tanto un número comprendido entre 0 y 1. Se cumple que

$$0 \leq f_i \leq 1 \quad \forall i = 1, 2, \dots, k.$$

Del mismo modo pueden definirse  $N_j$ , *frecuencia absoluta acumulada* correspondiente a la observación  $x_j$ , como la suma de las frecuencias absolutas correspondientes a observaciones menores o iguales a la observación  $j$ . Es decir :

$$N_j = \sum_{i \leq j} n_i \quad j = 1, 2, \dots, k$$

Y la *frecuencia relativa acumulada* como

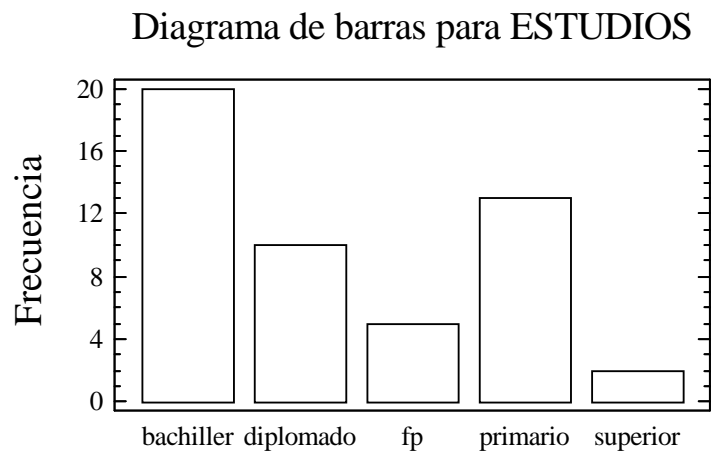
$$F_j = \sum_{i \leq j} f_i = \frac{N_j}{n} \quad j = 1, 2, \dots, k$$

**Ejemplo 2** En la siguiente tabla se dan las distintas frecuencias correspondientes a la tabla de tipo II para las tallas de los alumnos del instituto:

Intervalo	Marca de clase	Frecuencia absoluta ( $n_i$ )	Frecuencia absoluta acumulada ( $N_i$ )	Frecuencia relativa ( $f_i$ )	Frecuencia relativa acumulada ( $F_i$ )
[1.3, 1.4)	1.35	3	3	0.06	0.06
[1.4, 1.5)	1.45	3	6	0.06	0.12
[1.5, 1.6)	1.55	7	13	0.14	0.26
[1.6, 1.7)	1.65	11	24	0.22	0.48
[1.7, 1.8)	1.75	11	35	0.22	0.70
[1.8, 1.9)	1.85	5	40	0.10	0.80
[1.9, 2.0)	1.95	9	49	0.18	0.98
[2.0, 2.1]	2.05	1	50	0.02	1.00

### 1.2.4 Representaciones gráficas

**Diagrama de barras:** Se construye un gráfico poniendo en el eje horizontal los valores observados y elevando sobre cada valor una barra de altura proporcional a su frecuencia. El diagrama de barras correspondiente a la tabla tipo II para la variable estudios es:

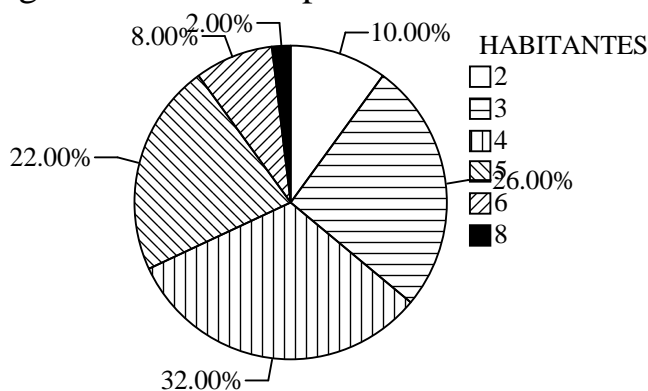


Representar la frecuencia absoluta o relativa como altura de la barra no influye en la forma de la gráfica, ya que sólo se realiza un cambio de escala.

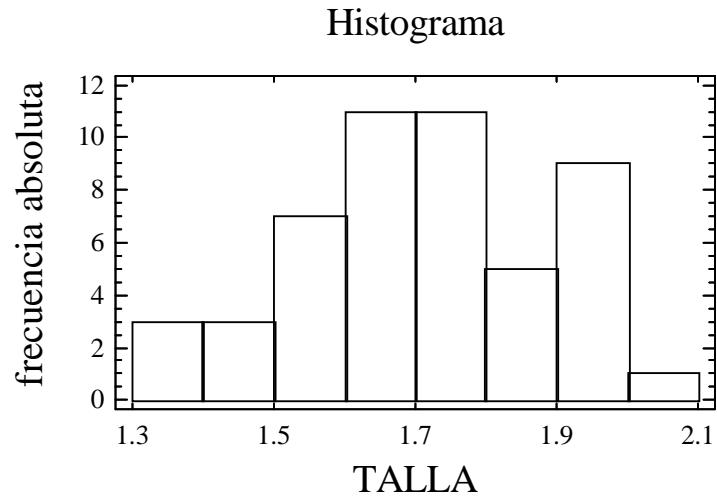
Este tipo de diagramas, y los de sectores que describiremos inmediatamente, son adecuados para variables cualitativas o cuantitativa discreta si el número de datos diferentes es pequeño, como ocurre en el caso de la variable habitantes

**Diagramas sectoriales:** Se dibuja un círculo y se divide en sectores circulares, de modo que cada uno represente la frecuencia de aparición en la muestra de un valor observado. Cada sector debe tener un área proporcional a su frecuencia, que suele venir indicada en la tabla en tanto por ciento. Si optamos por indicar la frecuencia relativa, el gráfico presenta el mismo aspecto, pero la frecuencia relativa viene indicada en tanto por uno.

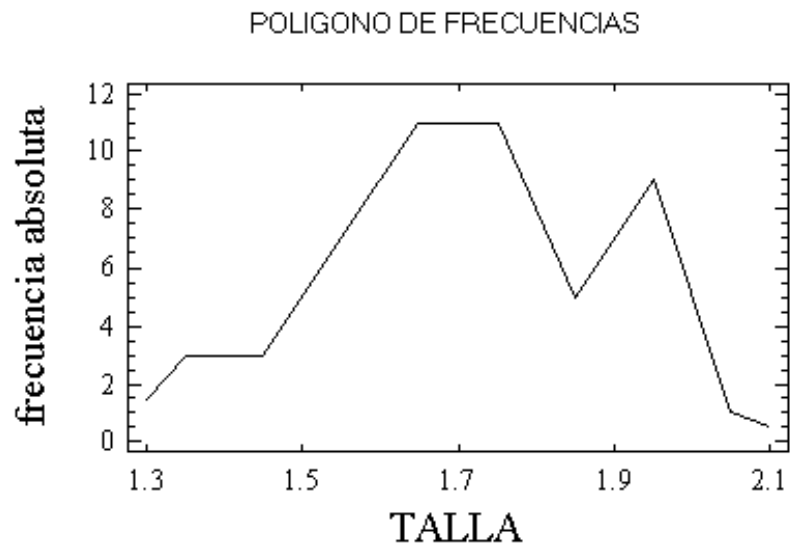
Diagrama de sectores para HABITANTES



**Histograma:** Se utiliza en el caso de las tablas de tipo III. Se construye el gráfico representando en el eje horizontal los intervalos de clase y elevando sobre cada uno de ellos un rectángulo cuya área ha de ser proporcional a su frecuencia. El histograma correspondiente a la tabla tipo III mostrada anteriormente sería:



**Polígono de frecuencias:** Se construye una curva uniendo los puntos medios de los lados superiores de cada rectángulo del histograma. El polígono de frecuencias correspondiente al anterior histograma de la variable Talla es:



**Diagrama de tallo y hojas:** A continuación mostramos un diagrama de tallo y hojas para la variable Talla.

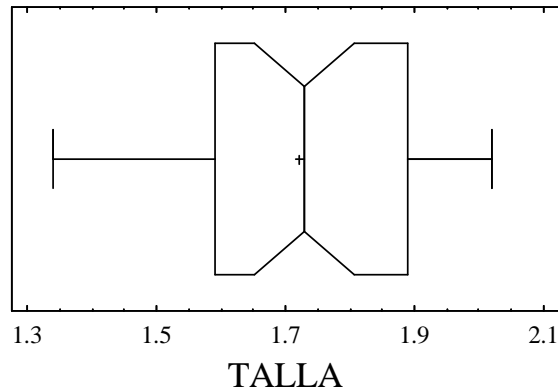
3	1.3   468
6	1.4   467
13	1.5   1223489
23	1.6   3455678899
(9)	1.7   024445688
18	1.8   000349
12	1.9   0034566889
2	2.0   02

El recorrido de la variable se ha dividido en 8 partes (los tallos), que vienen representados por los valores 1.3, 1.4, 1.5, etc. Los valores que le siguen, tras la línea vertical, son las hojas que corresponden a cada tallo. Así en el primer tallo tenemos las hojas 4, 6, 8. Esta rama corresponde a los datos más pequeños de la variable talla 1.34, 1.36, 1.38. La frecuencia acumulada de cada rama esta especificada a su izquierda. Así la frecuencia de la primera rama es 3, la de la segunda también es 3, pero la acumulada es 6. En este caso la acumulación de las frecuencias se hace por ambos lados de la tabla hasta llegar al tallo que contiene a la mediana. Este tallo contiene 9 elementos como está indicado entre paréntesis.

Esta representación tiene la ventaja de que superpone una tabla de frecuencias y una representación gráfica dada por la forma que toman los números, y que es similar al histograma de frecuencias. Además no hay pérdida de información, ya que se puede reconstruir todos los datos de la variable primitiva contenida en la muestra a partir de esta representación

**Gráfica de caja y bigotes (Box and Whisker):** En esta gráfica los datos se dividen en cuatro intervalos de igual frecuencia. La parte ancha, llamada *Caja*, contiene el 50% central de los datos de la variable. Comienza en el primer cuartil y termina en el tercer cuartil. La muesca de la caja marca la mediana (la definición de mediana y de cuartil se verá más adelante en el apartado de medidas de posición). En el gráfico de Box-Whisker correspondiente a la variable Talla, que aparece a continuación, se ha marcado además un punto, que corresponde a la media aritmética de los valores muestrales

## Gráfica de Box-Whisker



Las dos líneas horizontales se llaman *Bigotes* y se extienden a derecha e izquierda de la Caja. El bigote de la izquierda comienza por el dato más pequeño que dista del primer cuartil menos que 1.5 veces el rango intercuartílico (distancia entre el primer y tercer cuartil). En este caso corresponde al valor 1.34. El bigote de la derecha acaba en el mayor valor de la variable talla que diste del tercer cuartil menos que 1.5 veces el rango intercuartílico. Corresponde en este caso al valor mayor de la variable talla que es 2.02. A veces hay valores de la variable que sobresalen de los bigotes. Estos valores se clasifican como valores atípicos (Outliers).

Las tablas y las gráficas pretenden ordenar y clarificar la información contenida en la muestra. En los casos tratados, excepto en el caso del diagrama de tallo y hojas, siempre se hace perdiendo parte de información. En el siguiente apartado se darán algunas definiciones que pretenden reducir la información contenida en la muestra de una forma aún más drástica: a sólo unos cuantos valores, los parámetros estadísticos de la muestra. Entre ellos destacamos las medidas de posición y las de dispersión.

### 1.2.5 Medidas de posición

Suponemos los datos ordenados de menor a mayor. Las medidas de posición caracterizan ciertos datos por la posición que ocupan en esta serie. Entre las medidas de posición tenemos la siguiente:

**Mediana:** Definimos la mediana como aquel valor que hace que el 50% de las observaciones sean menores o iguales a él y otro 50% mayor o igual que él. Si el número total de observaciones es  $n$ , y ordenamos los datos de menor a mayor, la mediana será la que ocupe el lugar  $\frac{n+1}{2}$  si  $n$  es impar, o estará entre los valores  $\frac{n}{2}$  y  $\frac{n}{2} + 1$  si  $n$  es par. En este caso la mediana se obtiene como la semisuma de estos dos valores centrales.

Si partimos de una tabla de tipo III, y no conocemos los valores primitivos de la variable, calculamos en primer lugar en qué intervalo se encuentra la mediana. Dicho intervalo, al que denominaremos intervalo mediano y denotaremos por  $(L_i, L_{i+1}]$ , será aquel en el que la frecuencia absoluta acumulada sea igual o supere a  $\frac{n}{2}$ .

Entonces se aplicará la fórmula:

$$Me = L_i + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i$$

siendo  $a_i$  la amplitud del intervalo mediano. Esta fórmula puede obtenerse aplicando el procedimiento de interpolación en la frecuencia usando como datos los extremos del intervalo mediano y las frecuencias acumuladas que les correspondan.

### Medidas de posición no central

La mediana se conoce como una medida de posición central, ya que divide las observaciones en dos partes de igual frecuencia. Definimos ahora otras medidas de posición que dividen la muestra en partes de distinta frecuencia. Reciben el nombre genérico de **cuantiles**. Destacaremos los cuartiles, deciles y percentiles.

El *cuartil*  $n$ ,  $Q_n$  ( $n = 1, 2, 3$ ) es aquel valor que hace que las  $n$  cuartas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales. El segundo cuartil coincide con la mediana. Las fórmulas que permiten seleccionar un cuartil para una distribución tipo III es similar a la de la mediana. Por ejemplo para el primer cuartil

$$Q_1 = L_i + \frac{\frac{n}{4} - N_{i-1}}{n_i} a_i$$

pero ahora el intervalo  $i$  que hay que seleccionar es el que contenga un punto que deje delante el 25% de los datos.

El *decil*  $n$ ,  $D_n$  ( $n = 1, 2, \dots, 9$ ) es aquel valor que hace que las  $n$  décimas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales.

El *percentil*  $n$ ,  $P_n$  ( $n = 1, 2 \dots 99$ ) es aquel valor que hace que las  $n$  centésimas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales.

### Medidas de posición central

Las medidas de posición central pretenden ser representantes o ejemplos ilustrativos del tamaño de los datos contenidos en la muestra. La mediana es la única medida de posición central propiamente dicha. No obstante la media y la moda, toman con frecuentemente valores parecidos a la mediana y se suelen conocer también como medidas de posición central.

**Media** La media se define como el cociente entre la suma de todos los valores y el número total de elementos de la muestra.

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

No obstante, si los datos están repetidos, hay  $n$  elementos en la muestra pero sólo hay  $k$  elementos diferentes cada uno de los cuales aparece con una frecuencia  $n_i$ , se puede obtener también la media por medio de las expresiones siguientes:

$$\bar{X} = \sum_{i=1}^k \frac{n_i x_i}{n}, \quad \bar{X} = \sum_{i=1}^k \frac{n_i}{n} x_i = \sum_{i=1}^k f_i x_i \quad (1.1)$$

Con el objeto de destacar la igualdad entre estas expresiones de 1.1, ilustramos esta igualdad en el siguiente ejemplo.

**Ejemplo 3** Si la muestra está formada por los datos  $\{6, 8, 8, 8, 9, 9, 9, 9\}$  la tabla de frecuencias tipo II es:

$x_i$	$n_i$	$f_i$
6	1	$\frac{1}{8} = 0.125$
8	3	$\frac{3}{8} = 0.375$
9	4	$\frac{4}{8} = 0.5$

Hallar la media usando las expresiones de 1.1.

La media puede obtenerse de las formas siguientes:

$$\begin{aligned} \bar{X} &= \frac{6+8+8+8+9+9+9+9}{8} = \frac{1 \times 6 + 3 \times 8 + 4 \times 9}{8} = \frac{1}{8} \times 6 + \frac{3}{8} \times 8 + \frac{4}{8} \times 9 \\ &= 0.125 \times 6 + 0.375 \times 8 + 0.5 \times 9 = 8.25 \end{aligned}$$

Los cálculos realizados corresponden sucesivamente con las distintas expresiones dadas previamente para la media en 1.1.



**Moda** Es el valor que presenta una mayor frecuencia. Si es único, se dice que la distribución es unimodal, si no lo es, se dice que es multimodal.

En el caso de tablas tipo III, siendo  $a_i = L_{i+1} - L_i$ ,  $h_i = \frac{n_i}{a_i}$  es la altura que debe tener el histograma en el intervalo correspondiente. Llamaremos *intervalo modal*,  $(L_i, L_{i+1}]$ , al intervalo que presenta mayor altura en el histograma.

Aunque hay distintas expresiones para seleccionar un valor concreto para la moda dentro del intervalo modal, lo más sencillo es seleccionar la marca de clase del intervalo modal. La siguiente expresión para la moda:

$$Mo = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_i$$

$$\Delta_1 = h_i - h_{i-1}, \quad \Delta_2 = h_i - h_{i+1},$$

tiene la ventaja de que la moda queda, dentro del intervalo modal, pero más cerca del intervalo adyacente de más altura.

**Medidas de dispersión** Pretenden dar una idea sobre si los datos son muy parecidos entre sí o por el contrario están dispersos, es decir son bastante distintos unos de otros. Para aclarar este concepto consideremos las dos muestras siguientes que suponemos que son las calificaciones obtenidas por dos alumnos en las cuatro preguntas de un examen:

$$\text{Notas del alumno A} = \{5, 4, 6, 5\}$$

$$\text{Notas del alumno B} = \{1, 9, 10, 0\}$$

Si usáramos la media para obtener la calificación del examen ambos alumnos recibirían la misma calificación, un cinco. Pero percibimos que los exámenes de estos alumnos tienen características bien diferentes. Intentamos describir esta diferencia con ciertos parámetros que llamamos medidas de dispersión. Entre los más usados destacamos los siguientes:

**Recorrido o Rango** Es la diferencia entre el valor máximo y el mínimo de la variable.

$$\text{Recorrido alumno A} = 6 - 4 = 2$$

$$\text{Recorrido alumno B} = 10 - 0 = 10$$

**Rango intercuartílico** Es la distancia entre el primer y tercer cuartil.

Por ejemplo, para obtener los cuartiles de las *Notas del alumno A* = {5, 4, 6, 5}, ordenamos estas {4 - 5 - 5 - 6}. Dividiendo en cuatro partes la frecuencia obtenemos 4 clases con frecuencia 1. Como la separación de

las clases no está en ningún elemento, tomando la media de los valores más cercanos obtenemos:

Primer cuartil = 4.5

Segundo cuartil = 5.5

Así que el rango intercuartílico es  $5.5 - 4.5 = 1$

A veces se define el Rango semintercuartílico, que es la mitad del Rango intercuartílico. En este caso su valor sería 0.5.

**Desviación media** Se define de la siguiente forma:

$$\text{des}(X) = \frac{1}{n} \sum_{i=1}^k n_i |x_i - \bar{x}|$$

Mide el promedio de las distancias entre los datos y su media.

La desviación media para el primer alumno es:

$$\begin{aligned} \text{des}(\text{Notas de A}) &= \frac{1}{4} (2 \times |5 - 5| + 1 \times |4 - 5| + 1 \times |6 - 5|) = 0.5 \\ \text{des}(\text{Notas de B}) &= \\ \frac{1}{4} (1 \times |1 - 5| + 1 \times |9 - 5| + 1 \times |10 - 5| + 1 \times |0 - 5|) &= 4.5 \end{aligned}$$

Los segundos datos son más dispersos que los primeros.

**Varianza y Cuasivarianza** Aunque la desviación media nos da una definición que representa la dispersión de una forma muy intuitiva, a menudo se usa la varianza como medida de dispersión debido, entre otros motivos, a que el valor absoluto presenta ciertos inconvenientes en el cálculo, por no ser una función derivable. La varianza se define como

$$\text{var}(X) = S^2 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2$$

Puede demostrarse que esta expresión es equivalente a  $\frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2$ , que

presenta algunas ventajas de cálculo.

La varianza de las notas del alumno A es

$$\text{Var}(\text{Notas de A}) = \frac{1}{4} (2 \times (5 - 5)^2 + 1 \times (4 - 5)^2 + 1 \times (6 - 5)^2) = 0.5$$

La cuasivarianza se define como:

$$\text{cuasivar}(X) = s^2 = \frac{1}{n-1} \sum_{i=1}^k n_i (x_i - \bar{x})^2$$

La cuasivarianza de las notas del alumno A es

$$s^2 = \frac{1}{4-1} \left( 2 \times (5-5)^2 + 1 \times (4-5)^2 + 1 \times (6-5)^2 \right) = 0.66667$$

**Desviación típica y cuasidesviación** Se define la desviación típica como la raíz cuadrada de la varianza.  $S = \sqrt{\text{var}(X)}$ . Tiene la ventaja sobre la varianza de venir expresada en la misma unidad que los datos. La desviación típica de la variable notas del alumno A es

$$S = \sqrt{\text{var}(\text{Notas de A})} = \sqrt{0.5} = 0.70711$$

Definimos la cuasi desviación como la raíz cuadrada de la cuasivarianza. La cuasidesviación de la variable notas del alumno A es

$$s = \sqrt{\text{cuasivar}(\text{Notas de A})} = \sqrt{0.66667} = 0.8165$$

**Coefficiente de Variación de Pearson** Se denomina Coeficiente de Variación de Pearson al cociente:

$$CV = \frac{S}{\bar{x}}$$

que es una medida relativa de variabilidad y que permite comparar la dispersión de dos conjuntos de datos de diferentes escalas. Este parámetro es invariante frente al cambio de escala.

## 1.3 Estadística Descriptiva bidimensional

### 1.3.1 Introducción: distribución conjunta y tablas de doble entrada

En muchas ocasiones deseamos estudiar más de un carácter de una población determinada y nos interesa comprobar si existe relación entre dichos caracteres. Por ejemplo, podemos realizar un estudio de la relación entre la edad de los niños y su altura, para hacer una tabla de *alturas segun edad* de utilidad para los profesionales de la medicina. También podríamos preguntarnos si existe relación entre el número de habitantes de un país y su consumo energético, o entre el peso y la altura de sus habitantes, o entre los niveles de colesterol, glucosa, transaminasas y bilirrubina en la sangre. Centraremos nuestro estudio en el caso de dos variables.

Al igual que en el caso de una sola variable, los datos pueden venir presentados de diversas maneras. En el caso de tener pocos datos, pueden presentarse mediante una relación exhaustiva de todas las ocurrencias de las dos variables. Por ejemplo, si estamos estudiando el número de hijos e hijas que

tienen los empleados de una empresa, se nos podrían presentar los siguientes datos:

$$(1, 0), (2, 1), (0, 1), (1, 1), (0, 0), (0, 2), (3, 1), (1, 0), (2, 1), (2, 0)$$

Esta manera de presentar los datos solo es factible cuando se tiene un número muy pequeño de observaciones. Si el número de observaciones es grande, lo que se hace es agrupar los datos, indicando a continuación el número de ocurrencias de cada uno. Esto normalmente se realiza mediante una tabla de doble entrada, indicando en la intersección de cada fila y columna el número de ocurrencias.

En el ejemplo anterior, la tabla de doble entrada correspondiente sería:

	0	1	2
0	1	1	1
1	2	1	0
2	1	2	0
3	0	1	0

También es posible dar los datos con una tabla lineal. La que damos a continuación se refiere también al caso de los hijos e hijas de los empleados de la empresa. Los valores se han tomado de la tabla de doble entrada.

$X = hijos$	0	0	0	1	1	2	2	3
$Y = hijas$	0	1	2	0	1	0	1	1
$frecuencias$	1	1	1	2	1	1	2	1

Si llamamos  $X$  a la primera variable, que toma los valores  $x_1, x_2, \dots, x_r$  y llamamos  $Y$  a la segunda variable, pudiendo tomar los valores  $y_1, y_2, \dots, y_s$ , la tabla de doble entrada sería de la siguiente forma:

	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_s$	
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1j}$	$\dots$	$n_{1s}$	$n_{1\bullet}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2j}$	$\dots$	$n_{2s}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$n_{i2}$	$\dots$	$n_{ij}$	$\dots$	$n_{is}$	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_r$	$n_{r1}$	$n_{r2}$	$\dots$	$n_{rj}$	$\dots$	$n_{rs}$	$n_{r\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet s}$	$N$

donde  $n_{ij}$  representa el número de veces que se presenta la observación  $(x_i, y_j)$ . La última fila se obtiene sumando los elementos de la columna

correspondiente y la última columna sumando los elementos de su misma fila:

$$n_{i\bullet} = \sum_j n_{ij} \quad n_{\bullet j} = \sum_i n_{ij}$$

La suma de los  $n_{i\bullet}$  coincide con la suma de los  $n_{\bullet j}$  y vale  $N$ , el número total de pares de elementos de la muestra.

En el caso de que sean muchos los posibles valores que pueda tomar la variable, agrupamos los datos en intervalos. Obtendremos entonces una tabla de doble entrada equivalente a las tablas de tipo III para la variable unidimensional.

**Ejemplo 4** La tabla de doble entrada siguiente corresponde a la distribución conjunta de las variables Talla y Peso de los alumnos del instituto. Los datos se han clasificado en cuatro intervalos para el peso y otros cuatro para la talla.

	30-50	50-70	70-90	90-110	Total
1.30-1.50	6	0	0	0	6
1.50-1.70	0	15	3	0	18
1.70-1.90	0	1	12	3	16
1.90-2.10	0	0	1	9	10
Total	6	16	16	12	50

También podemos construir tablas de frecuencias relativas sin más que dividir todos los elementos de la tabla por el número total de datos  $N$ . Así pues, denominamos frecuencia relativa de la pareja  $(x_i, y_j)$  a

$$fr(x_i, y_j) = f_{ij} = \frac{n_{ij}}{N}$$

Es evidente comprobar que la suma de todas las frecuencias relativas es 1. Análogamente se pueden definir las cantidades  $f_{i\bullet}$  y  $f_{\bullet j}$ .

$$f_{i\bullet} = \sum_j f_{ij} = \frac{n_{i\bullet}}{N} \quad f_{\bullet j} = \sum_i f_{ij} = \frac{n_{\bullet j}}{N}$$

La tabla de frecuencias relativas correspondiente a las variables peso y talla es la siguiente

	30 – 50	50 – 70	70 – 90	90 – 110	Total
1.30 – 1.50	0.12	0	0	0	$f_{1\bullet} = 0.12$
1.50 – 1.70	0	0.30	0.06	0	$f_{2\bullet} = 0.36$
1.70 – 1.90	0	0.02	0.24	0.06	$f_{3\bullet} = 0.32$
1.90 – 2.10	0	0	0.02	0.18	$f_{4\bullet} = 0.2$
Total	$f_{\bullet 1} = 0.12$	$f_{\bullet 2} = 0.32$	$f_{\bullet 3} = 0.32$	$f_{\bullet 4} = 0.24$	1

### 1.3.2 Representaciones gráficas

La representación gráfica más usual para variables aleatorias bidimensionales es la llamada **Nube de Puntos**. En el plano delimitado por dos ejes que sirvan para representar las variables  $X$  e  $Y$  se dibuja un punto  $(x, y)$  por cada vez que las variables tomen este par de valores. Si coinciden varias observaciones en un mismo punto puede optarse por dibujar un pequeño círculo de radio proporcional a su frecuencia o indicar en la gráfica esta frecuencia al lado del punto.

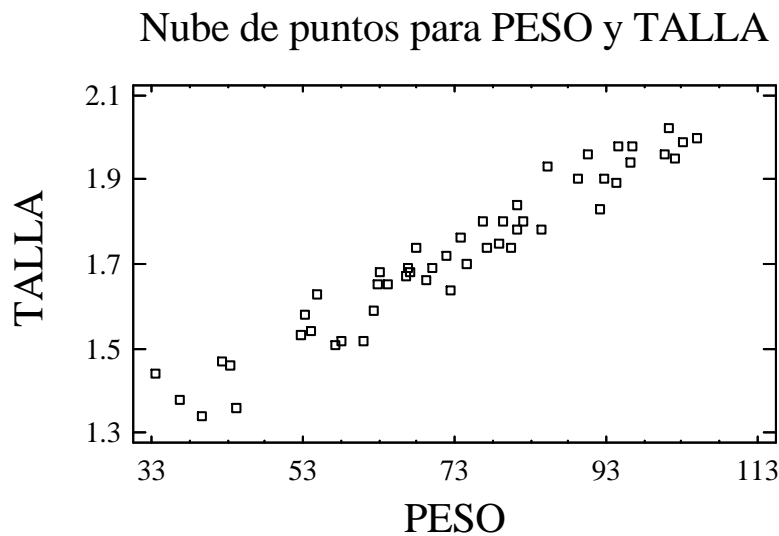
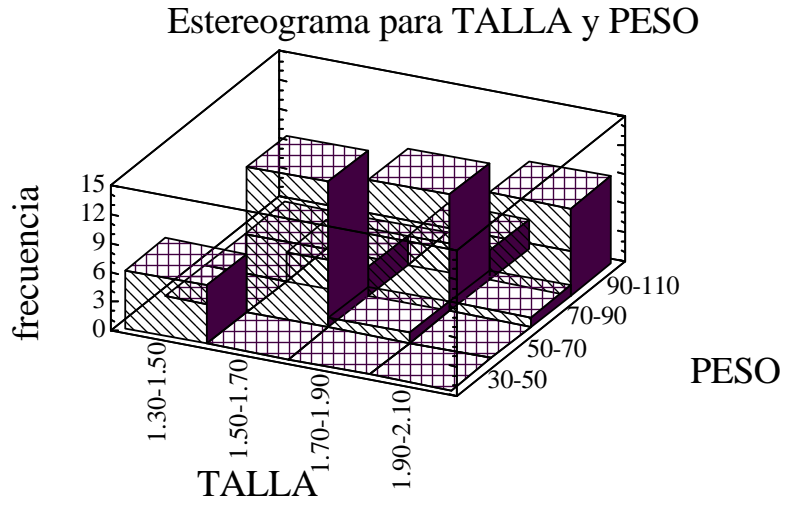


Figura 1.1:

Otra manera de representar los datos es mediante un **diagrama de barras tridimensional**. Sobre cada punto del plano se levanta una barra de altura proporcional a su frecuencia. Queda por tanto un gráfico tridimensional.

En el caso de que los datos vengan agrupados en intervalos se dibuja un histograma tridimensional, también llamado **estereograma**. Sobre cada uno de los rectángulos determinados por un intervalo de  $X$  y otro de  $Y$  se levanta un paralelepípedo rectángulo. En este caso, su volumen ha de ser proporcional a la frecuencia con que aparecen puntos contenidos en dicho rectángulo. A continuación aparece un estereograma para las variables Talla y Peso correspondiente a la tabla de la página 37.



**1.3.3 Distribuciones marginales. Distribuciones condicionadas**

En las distribuciones de frecuencia bidimensionales cabe estudiar por separado cada una de las variables unidimensionales que la componen, haciendo caso omiso de la otra. Estas distribuciones reciben el nombre de *distribuciones marginales*. Son obviamente dos: la distribución marginal de  $X$  y la distribución marginal de  $Y$ . Las frecuencias absolutas asociadas a los distintos valores  $x_i$  de la variable  $X$  son las  $n_{i\bullet}$  y las de los  $y_j$  son las  $n_{\bullet j}$ .

Así pues, la distribución marginal de  $X$  se obtiene tomando, en la tabla de doble entrada, la primera y última columnas

$x_1$	$n_{1\bullet}$
$x_2$	$n_{2\bullet}$
$\vdots$	$\vdots$
$x_i$	$n_{i\bullet}$
$\vdots$	$\vdots$
$x_r$	$n_{r\bullet}$
	$N$

y la marginal de  $Y$  tomando la primera y última fila.

$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_s$		
$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet s}$		$N$

Las frecuencias relativas de las distribuciones marginales se obtendrán dividiendo las frecuencias absolutas entre el número total de observaciones  $N$ . Es decir, la frecuencia relativa de  $x_i$  será:

$$fr(x_i) = \frac{n_{i\bullet}}{N} = f_{i\bullet}$$

y la frecuencia relativa de  $y_j$  será:

$$fr(y_j) = \frac{n_{\bullet j}}{N} = f_{\bullet j}$$

Las tablas de frecuencia relativa para la marginal correspondiente a la variable talla sería

1.30 – 1.50	$f_{1\bullet} = 0.12$
1.50 – 1.70	$f_{2\bullet} = 0.36$
1.70 – 1.90	$f_{3\bullet} = 0.32$
1.90 – 2.10	$f_{4\bullet} = 0.2$
Total	1

y la de la variable peso:

	Total
30 – 50	$f_{\bullet 1} = 0.12$
50 – 70	$f_{\bullet 2} = 0.32$
70 – 90	$f_{\bullet 3} = 0.32$
90 – 110	$f_{\bullet 4} = 0.24$
	1

En otras ocasiones nos interesará analizar los datos obtenidos por una de las variables cuando se presenta exactamente un determinado valor de la otra variable. Esta idea da lugar a las llamadas *distribuciones condicionadas de frecuencias*.

Podemos estudiar la distribución de  $X$  condicionada a que la variable  $Y$  tome el valor  $y_j$ . A esta variable la denotaremos por  $X/y_j$ , obteniéndose a partir de la primera columna y la correspondiente al valor  $y_j$ .

$x_1$	$n_{1j}$
$x_2$	$n_{2j}$
$\vdots$	$\vdots$
$x_i$	$n_{ij}$
$\vdots$	$\vdots$
$x_r$	$n_{rj}$
Total	$n_{\bullet j}$



También podemos estudiar la distribución de  $Y$  condicionada a que la variable  $X$  tome el valor  $x_i$ . A esta variable la denotaremos por  $Y/x_i$ , obteniéndose a partir de la primera fila y la correspondiente al valor  $x_i$ .

$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_s$	<i>Total</i>
$n_{i1}$	$n_{i2}$	$\dots$	$n_{ij}$	$\dots$	$n_{is}$	$n_{i\bullet}$

Las frecuencias relativas de las distribuciones condicionadas se obtendrán dividiendo las frecuencias absolutas entre el número total de observaciones que cumplen la condición requerida, que en los casos anteriores son, respectivamente,  $n_{\bullet j}$  y  $n_{i\bullet}$ .

Es decir, la frecuencia relativa de  $x_i/y_j$  será:

$$fr(x_i/y_j) = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}$$

y la frecuencia relativa de  $y_j/x_i$  será:

$$fr(y_j/x_i) = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$$

La distribución del peso condicionada a que la talla esté en el intervalo 1.70-1.90 viene dada en la tabla siguiente, donde se muestran la frecuencias absoluta de cada uno de los pesos de los individuos de cuya talla está en el intervalo de 1.70 a 1.90.

<i>Peso</i>	30 – 50	50 – 70	70 – 90	90 – 110	<i>Total</i>
<i>Frecuencia absoluta</i>	0	1	12	3	16

Para obtener la tabla de frecuencias relativa hay que dividir estas frecuencias por el total de individuos de la talla considerada, que en este caso son 16.

La tabla de frecuencias relativa para la distribución condicionada resulta:

<i>Peso</i>	30 – 50	50 – 70	70 – 90	90 – 110	<i>Total</i>
<i>Frec. relativa</i>	0	$\frac{1}{16} = .0625$	$\frac{12}{16} = 0.75$	$\frac{3}{16} = 0.1875$	$\frac{16}{16} = 1$

## 1.4 Regresión y Correlación

### 1.4.1 Independencia de variables estadísticas. Dependencia funcional y dependencia estadística

Dos variables estadísticas se dicen dependientes cuando el conocimiento de que se ha presentado una determinada ocurrencia en una de ellas condiciona

en algún sentido el valor que pueda tomar la otra. Así si observamos la nube de puntos de las variable peso y talla apreciamos que los valores bajos de talla dan valores bajos para el peso, y en cambio, a las tallas mayores corresponden pesos mayores. A la vista de los valores de los tallas sabemos que están comprendidas entre 1.34 y 2.02. Sin embargo conociendo el peso de una persona podemos dar una información más precisa sobre su talla. Así, mirando la nube de puntos observamos que las personas cuyo peso es aproximadamente 73, tienen una altura entre 1.60 y 1.80. En cambio las personas que pesan más o menos 99 tienen una talla comprendida entre 1.90 y 2. Por tanto, conociendo el peso de una persona obtenemos alguna información suplementaria sobre su peso.

Damos, a partir de esta idea, la siguiente definición: dos variables  $X$  e  $Y$  (que constituyen una variable bidimensional) son *independientes* si las distribuciones de frecuencias relativas de la variable  $X$  condicionada a cualquier valor  $y_j$  de la variable  $Y$  son todas idénticas, sin depender del valor de  $y_j$ . La distribución de  $X$  no depende del valor que tome la variable  $Y$ .

Es decir:

$$fr(x_i/y_1) = fr(x_i/y_2) = \dots = fr(x_i/y_s)$$

De aquí se deduce que

$$fr(x_i/y_j) = f_i. \quad \forall i = 1, 2, \dots, r$$

Es decir, las distribuciones condicionadas coinciden exactamente con la distribución de frecuencias relativas marginal de  $X$ .

Debemos resaltar la diferencia entre dependencia funcional y dependencia estadística. Cuando la variable  $Y$  depende funcionalmente de la variable  $X$  eso significa que conociendo el valor que toma la variable  $X$  tenemos perfectamente determinado el valor que tomará la variable  $Y$ . Sin embargo, cuando se produce dependencia estadística eso significa que al conocer el valor que toma la variable  $X$  obtenemos “alguna” información sobre la distribución de frecuencias de los valores de la variable  $Y$ , pero no obtenemos un valor concreto para esta variable.

### 1.4.2 Medias, varianzas y covarianzas

Tanto las distribuciones marginales como condicionadas que hemos visto son distribuciones unidimensionales, y por tanto podemos calcular para ellas todas las medidas descriptivas expuestas en el apartado correspondiente a variables unidimensionales, sin más que tener en cuenta las frecuencias relativas de cada caso. En particular, la media de la distribución marginal de  $X$  será:

$$\bar{X} = E[X] = \sum_{i=1}^r f_{i\bullet} x_i$$

Del mismo modo, la media de la distribución marginal de  $Y$  será:

$$\bar{Y} = E[Y] = \sum_{j=1}^s f_{\bullet j} y_j$$

Análogamente pueden calcularse las respectivas varianzas.

Para las distribuciones condicionadas tendremos los siguientes valores:

$$E[X/y_j] = \sum_{i=1}^r \frac{f_{ij}}{f_{\bullet j}} x_i$$

$$E[Y/x_i] = \sum_{j=1}^s \frac{f_{ij}}{f_{i\bullet}} y_j$$

Es inmediato deducir que si  $X$  e  $Y$  son independientes se verifica que

$$E[X/y_j] = E[X] \quad E[Y/x_i] = E[Y]$$

Las medidas vistas hasta ahora corresponden a distribuciones unidimensionales. También existen parámetros conjuntos para ambas variables, característicos de la distribución bidimensional y que, como veremos más adelante, van a estar ligados a la dependencia de las variables. Una de estas medidas recibe el nombre de *covarianza* de las variables  $X$  e  $Y$ :

$$\text{cov}(X, Y) = S_{XY} = \sum_{i=1}^r \sum_{j=1}^s f_{ij} (x_i - \bar{X})(y_j - \bar{Y})$$

Esta fórmula puede simplificarse hasta quedar:

$$\text{cov}(X, Y) = \sum_{i=1}^r \sum_{j=1}^s f_{ij} x_i y_j - \bar{X} \bar{Y}$$

Si las dos variables son independientes, se verifica que

$$\text{cov}(X, Y) = 0$$

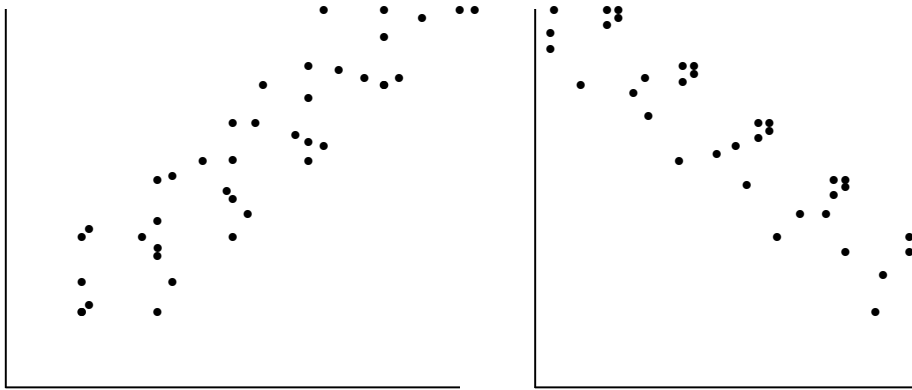
En muchas situaciones prácticas es frecuente encontrar que existe una cierta dependencia de tipo estadístico entre dos variables. Así, si estudiamos el peso de un coche y su gasto de combustible observaremos que guardan una cierta relación. Una relación de dependencia es de tipo funcional cuando podemos encontrar una función matemática de modo que para cada valor de  $X$  podamos encontrar el valor correspondiente de  $Y$ . En las dependencias de tipo estadístico, sin embargo, no es posible establecer tal función, y lo normal

es que a un valor determinado de  $X$  le puedan corresponder distintos valores de  $Y$ .

Si se representa la nube de puntos correspondiente a los datos observados es posible establecer la relación de dependencia entre las variables. En los casos de dependencia funcional se podría encontrar una función cuya gráfica pasara por todos los puntos dibujados. En el caso de la dependencia estadística se podría encontrar una función de modo que la distancia entre la nube de puntos y su gráfica sean pequeños.

En la figura 1.2 se consideran ejemplos de nubes de puntos entre los que existe dependencia estadística de tipo lineal entre variables. En la figura 1.3 hay un primer ejemplo en el que no existe dicha dependencia estadística y otro ejemplo en el que existiendo dependencia estadística no es de tipo lineal.

Figura 1.2: Dependencia estadística de tipo lineal entre variables

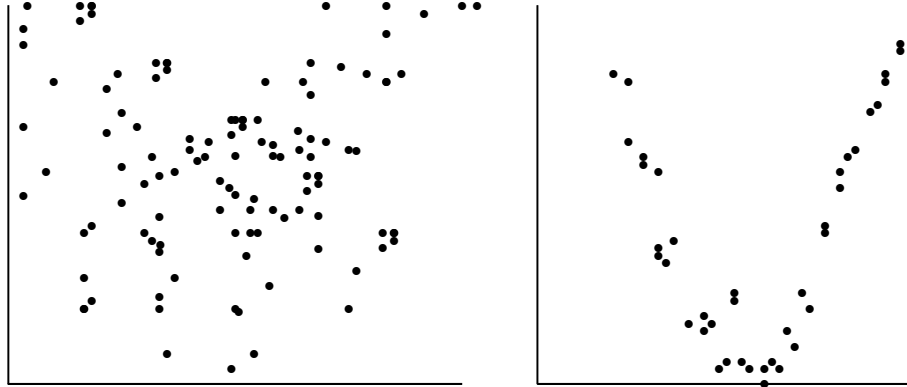


### 1.4.3 Ajustes. Método de mínimos cuadrados

Consideramos  $N$  observaciones que son pares de valores del tipo  $(x_i, y_i)$ . Si tomamos como variable independiente a  $X$  y como variable dependiente a  $Y$ , debemos de hallar una expresión de  $Y$  como función de  $X$ . En este caso diremos que estamos realizando una regresión de  $Y$  sobre  $X$ .

La función de regresión,  $y = f(x)$ , puede adoptar distintas formas: línea recta, parábola, polinomio de grado  $n$ , función exponencial, etcétera. El modo usual de proceder es prefijar el tipo de función que se va a considerar. Todas las posibles funciones de ese tipo tendrán una formulación general que dependerá de unos parámetros. La determinación de dichos parámetros se

Figura 1.3: No existe dependencia estadística entre variables o bien no es lineal



hará de modo que los valores observados estén “próximos” a los puntos de la función de regresión.

Consideremos como variable independiente a  $X$  y como variable dependiente a  $Y$ . Si hemos observado un punto  $(x_i, y_i)$ , llamamos  $y_i^*$  al valor previsto por la función de regresión para  $x_i$ , es decir  $y_i^* = f(x_i)$ . Denominaremos *error o residuo* y lo denotaremos por  $e_i$  a la diferencia entre el valor observado y el valor previsto, es decir:

$$e_i = y_i - y_i^*$$

Lógicamente se desea que los residuos o errores sean lo más pequeños posible. El método que más se utiliza para obtener los parámetros es el de *ajuste por mínimos cuadrados*, que consiste en obtener el valor de los parámetros que hagan mínima la suma de los cuadrados de los residuos. Es decir, habría que minimizar la expresión

$$\sum_{i=1}^N e_i^2$$

#### 1.4.4 Regresión lineal mínimo cuadrática

Estudiaremos el caso más sencillo y de mayor importancia, que es aquel en que la función de regresión es una línea recta. La expresión general de una línea recta será del tipo

$$y = a + bx$$

con lo que tendremos que los valores previstos o predichos por la regresión serían

$$y_i^* = a + bx_i$$

de donde deducimos que

$$e_i = y_i - y_i^* = y_i - a - bx_i$$

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - y_i^*)^2 = \sum_{i=1}^N (y_i - a - bx_i)^2$$

Para determinar los valores de los parámetros  $a$  y  $b$  que minimizan dicha suma igualamos a cero las derivadas respecto a  $a$  y  $b$ .

$$-2 \sum_{i=1}^N (y_i - a - bx_i) = 0$$

$$-2 \sum_{i=1}^N (y_i - a - bx_i)x_i = 0$$

Operando obtenemos que estas ecuaciones se convierten en:

$$\sum_{i=1}^N y_i = b \sum_{i=1}^N x_i + Na$$

$$\sum_{i=1}^N y_i x_i = b \sum_{i=1}^N x_i^2 + a \sum_{i=1}^N x_i$$

que recibe el nombre de *sistema de ecuaciones normales*.

Dividiendo ambas ecuaciones por  $N$  obtenemos:

$$\bar{Y} = b\bar{X} + a$$

$$\frac{\sum_{i=1}^N y_i x_i}{N} = b \frac{\sum_{i=1}^N x_i^2}{N} + a\bar{X}$$

Para calcular los valores de  $a$  y  $b$ , que son la únicas incógnitas de este sistema hay que resolverlo. La primera ecuación del sistema

$$\bar{Y} = a + b\bar{X} \tag{1.2}$$

nos indica que la recta de regresión de  $Y$  sobre  $X$  pasa por el punto  $(\bar{X}, \bar{Y})$ , que es el centro de gravedad de la nube de puntos.

Despejando en esta ecuación el valor de  $a$  y sustituyendo en la segunda ecuación del sistema obtenemos:

$$b = \frac{S_{XY}}{S_X^2} \tag{1.3}$$

que nos indica que el parámetro  $b$  de la recta de regresión puede calcularse como el cociente entre la covarianza y la varianza de la variable independiente. Este parámetro, llamado *coeficiente de regresión de Y sobre X*, representa la pendiente de la recta. Por tanto una expresión de la recta de regresión es

$$y - \bar{Y} = \frac{S_{XY}}{S_X^2}(x - \bar{X}) \quad (1.4)$$

que se obtiene usando la ecuación *punto-pendiente* de una recta.

Usando las expresiones 1.2 y 1.3, o también operando en la ecuación 1.4 obtenemos que

$$a = \bar{Y} - \frac{S_{XY}}{S_X^2} \bar{X}$$

**Ejemplo 5** *Calculamos ahora la covarianza y la recta de regresión correspondiente a los datos de la siguiente tabla que se refieren a los hijos e hijas de los empleados de la empresa. Los datos vienen dados en la siguiente tabla:*

	0	1	2	Marginal de X
0	1	1	1	3
1	2	1	0	3
2	1	2	0	3
3	0	1	0	1
Marginal de Y	4	5	1	total = 10

Comenzamos hallando la media de  $X$  e  $Y$ , ya que son necesarias para evaluar la covarianza e igualmente los coeficientes de la recta de regresión.

$$\bar{X} = \frac{3 \times 0 + 3 \times 1 + 3 \times 2 + 1 \times 3}{10} = 1.2$$

$$\bar{Y} = \frac{4 \times 0 + 5 \times 1 + 1 \times 2}{10} = 0.7$$

$$\begin{aligned} S_{XY} &= \sum_{i=1}^r \sum_{j=1}^s f_{ij}(x_i - \bar{X})(y_j - \bar{Y}) = \\ &= \frac{1}{10}(0 - 1.2)(0 - 0.7) + \frac{1}{10}(0 - 1.2)(1 - 0.7) + \frac{1}{10}(0 - 1.2)(2 - 0.7) + \\ &\quad + \frac{2}{10}(1 - 1.2)(0 - 0.7) + \frac{1}{10}(1 - 1.2)(1 - 0.7) + \\ &\quad + \frac{1}{10}(2 - 1.2)(0 - 0.7) + \frac{2}{10}(2 - 1.2)(1 - 0.7) + \\ &\quad + \frac{1}{10}(3 - 1.2)(1 - 0.7) = -0.04 \end{aligned}$$

Para hallar la recta de regresión calculamos también la varianza de  $Y$ .

Usando la expresión alternativa de la varianza:

$$\frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2 = \frac{1}{10} (3 \times 0^2 + 3 \times 1^2 + 3 \times 2^2 + 1 \times 3^2) - 1.2^2 = 0.96$$

La recta de regresión es

$$(y - 0.7) = \frac{-0.04}{0.96}(x - 1.2)$$

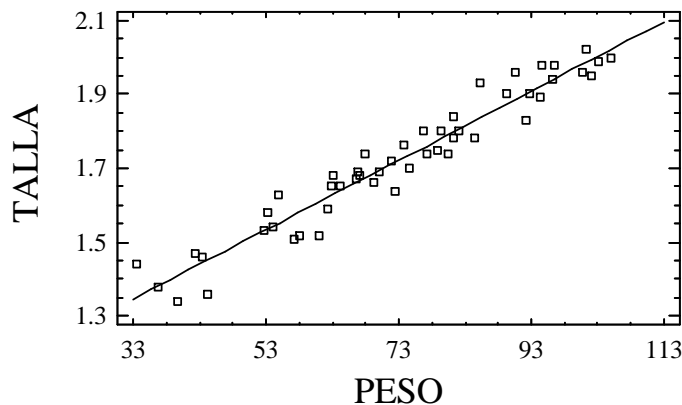
$$y = -4.1667 \times 10^{-2}x + 0.75$$

La recta de regresión correspondiente a la nube de puntos representada en la figura 1.1 es

$$Talla = 0.0094 \times peso + 1.038$$

La representación gráfica de esta última recta de regresión puede verse en la siguiente figura:

Recta de Regresión de la TALLA sobre el PESO



#### 1.4.5 Coeficiente de determinación. Coeficiente de correlación lineal

Si consideramos el caso de una regresión de  $Y$  sobre  $X$ , para medir el grado de dependencia estadística entre  $X$  e  $Y$  puede utilizarse el llamado *coeficiente de determinación*, que denotaremos por  $R^2$ , y que se calcula como

$$R^2 = \frac{S_{y^*}^2}{S_y^2}$$



El numerador es la varianza de los valores calculados para cada  $x_i$  mediante la función de regresión y el denominador es la varianza de los valores observados para  $Y$ . Se demuestra que  $R^2$  sólo puede tomar valores en el intervalo  $[0, 1]$ .

Si  $R^2$  vale 1 nos indica que existe una dependencia exactamente funcional, todos los puntos observados están sobre la gráfica de la función de regresión obtenida. En cambio, si  $R^2$  vale 0, entonces el modelo de regresión seleccionado no explica nada sobre la variación de  $Y$ . Si  $R^2$  está próximo a 1 se acepta que el modelo lineal, es decir la recta de regresión explica la relación de dependencia.

Todo lo anterior es válido para el caso de una función de regresión cualquiera, sin importar la forma que adopte. En el caso particular de regresión de tipo lineal podemos calcular el *coeficiente de correlación lineal* mediante la expresión:

$$r = \frac{S_{XY}}{S_X S_Y}$$

es decir, como el cociente entre la covarianza y el producto de las desviaciones típicas de  $X$  e  $Y$ . Se verifica que  $r^2 = R^2$ , pero mientras que  $R^2$  puede calcularse para cualquier tipo de regresión,  $r$  sólo tiene sentido en el caso de regresión lineal.

**Ejemplo 6** *El valor obtenido para estos parámetros en el caso de los hijos y las hijas de los empleados es:*

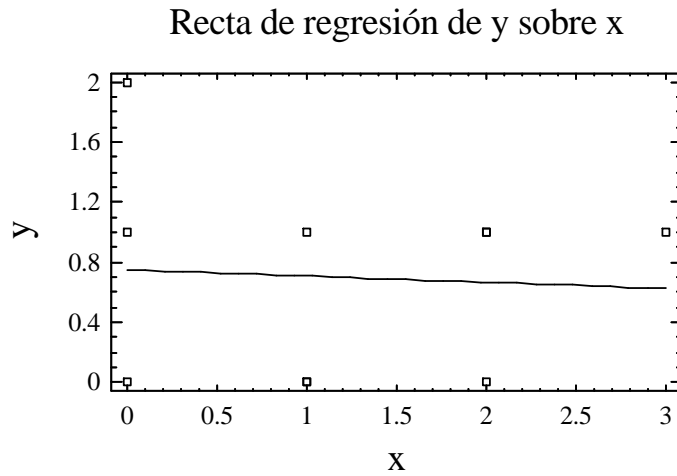
$$r = \frac{S_{XY}}{S_X S_Y} = \frac{-0.04}{\sqrt{0.96}\sqrt{0.41}} = -0.06375$$

donde la varianza de  $Y$  se ha obtenido de la expresión

$$\frac{1}{n} \sum_{j=1}^k n_j y_j^2 - \bar{y}^2 = \frac{1}{10}(4 \times 0^2 + 5 \times 1^2 + 1 \times 2^2) - 0.7^2 = 0.41$$

$$\text{Podemos obtener } R^2 = (-0.06375)^2 = 0.004064.$$

Como estos valores son próximos a 0, concluimos que el ajuste es muy pobre. Es decir que los datos no están cerca de la recta de regresión. En efecto, esto puede apreciarse en la siguiente gráfica que nos da la representación gráfica de los puntos y de la recta de regresión.



Si calculamos ambos parámetros en el caso de regresión lineal de la variable talla sobre la variable peso de los alumnos del instituto obtendremos:

$$r = \text{coeficiente de Correlación} = 0.967641$$

$$R^2 = \text{coeficiente de Determinación} = 0.93633$$

Como estos valores son cercanos al valor 1, nos indican un buen ajuste de los puntos a la recta de regresión.

El signo de  $r$  coincide con el de  $S_{XY}$ . Si  $r > 0$  la recta tiene pendiente positiva, es decir cuando una variable crece la otra también. Si  $r < 0$  cuando una variable crece la otra decrece.

Si las variables son independientes, la covarianza es nula, y por tanto  $r = 0$ . El recíproco no tiene por qué ser cierto.

La teoría de regresión nos permite hacer predicciones del valor que tomará la variable dependiente conociendo el valor que toma la variable independiente, sustituyendo el valor de esta última en la función de regresión. Hay que tener en cuenta, sin embargo, que las predicciones tienen mayor validez si se consideran valores de la variable cercanos a su media. Conforme los valores van estado más alejados de la media más arriesgada será la predicción.

## 1.5 EJERCICIOS PROPUESTOS

**Ejercicio 1** Una empresa de alimentación se dedica a enviar pizzas a domicilio. El número de pizzas enviadas en cada uno de los 30 días del mes de Abril son:

47 63 66 58 32 61 57 44 44 56  
 38 35 76 58 48 59 67 33 69 53  
 51 28 25 36 49 78 48 42 72 52

1. Construir una tabla de frecuencia relativa y de frecuencia acumulada usando una tabla tipo III e intervalos de clase con una amplitud de 5 pizzas.
2. Calcular la media en los siguientes casos: a) Usando todos los datos, b) Usando solamente los valores de la tabla de frecuencias construida.
3. Calcular la, mediana, moda y la desviación típica usando los valores de la tabla
4. Determinar los tres cuartiles a partir de la tabla de frecuencias
5. Construir el histogramas de frecuencias y el polígono de frecuencias correspondiente.

**Ejercicio 2** Los alumnos de un cierto grupo han obtenido las siguientes calificaciones en un test:

7, 5, 5, 0, 3, 9, 5, 3, 7, 9, 7, 5, 5, 3, 3, 5, 3, 5, 5, 3.

1. Hallar las tablas de frecuencias y de frecuencias acumuladas.
2. Diagrama de barras y poligono de frecuencia simple y el histograma de frecuencia acumulada.
3. Calcular la media, la mediana y la moda
4. Calcular el recorrido, la desviación media, la varianza, la desviación típica, la cuasivarianza o varianza muestral, la cuasidesviación y el coeficiente de variación.
5. Calcular el primer y tercer cuartil

**Ejercicio 3** La siguiente tabla muestra los tiempos en segundos empleados en establecer una conexión a Internet en 75 ocasiones. Los datos se han

agrupados mediante la siguiente tabla de frecuencias.

<i>Tiempos en seg.</i>	$n_i$
1,30 – 1.35	3
1.35 – 1.40	6
1,40 – 1.45	6
1.45 – 1.50	14
1,50 – 1.55	12
1.55 – 1.60	13
1,60 – 1.65	15
1.65 – 1.70	2
1,70 – 1.75	2
1.75 – 1.80	2
<i>Total</i>	$N = 75$

1. Construir el histograma de frecuencias. Indicar el intervalo modal.
2. Calcular la media y la mediana.
3. Calcular la varianza y el recorrido intercuartílico.

**Ejercicio 4** La distribución correspondiente al peso en kilos de 100 mujeres de 20 años viene resumida en la siguiente tabla de frecuencias:

<i>Intervalos de la variable</i>	<i>fr. absolutas)</i>
$[60, 65)$	20
$[65, 67)$	20
$[67, 69)$	25
$[69, 74)$	35

1. Forma la tabla de frecuencias de la distribución detallando, aparte de las frecuencias absolutas ( $n_i$ ) el intervalo de clase,  $[x_i, x_{i+1})$ , la marca de clase ( $a_i$ ), la amplitud del intervalo,  $(x_{i+1} - x_i)$ , la altura ( $h_i$ ) del histograma de frecuencia para cada clase, así como las frecuencias relativas ( $f_i$ ), absolutas acumuladas ( $N_i$ ) y relativas acumuladas ( $F_i$ ).
2. Representa los datos en un histograma
3. Estima cuántas mujeres pesan menos de 72 kilos
4. Determina la moda y la mediana.
5. Halla el rango intercuartílico, explicando su significado.
6. A partir de qué valor se encuentra el 25% de las mujeres con más peso?

**Ejercicio 5** Las calificaciones obtenidas por 40 alumnos/as de Bachillerato en las asignaturas de Matemáticas y las horas de estudio semanales que dedican a esta materia figuran en la siguiente tabla estadística bidimensional. En ella, la variable  $X$  hace referencia a la calificación lograda en Matemáticas e  $Y$ , al número de horas de estudio a la semana.

$Y_j$	$X_i$	3	4	5	6	7	8	10	
2		4							4
5			7	11					18
6					5	3			8
7					5	2			7
9							1		1
10								2	2
Total		4	7	11	10	5	1	2	40

1. Utilizar una tabla simple y otra de doble entrada para hallar la covarianza y el coeficiente de correlación lineal de Pearson
2. Analiza el grado de dependencia entre las calificaciones y las horas dedicadas al estudio.
3. En caso de que exista correlación ¿qué nota cabe esperar en Matemáticas un alumno que dedica 8 horas semanales al estudio?
4. ¿Cuántas horas se estima que dedica al estudio un alumno que haya obtenido un 5 en Matemáticas?

**Ejercicio 6** Ajustar una recta a los puntos dados en la siguiente tabla por

medio de sus coordenadas  $x$  e  $y$ . Estudiar la calidad del ajuste

$x$	$y$
1.444	0.565
1.898	0.132
1.171	2.584
7.006	-1.84
8.337	-4.12
9.323	-7.2
19.13	-119
11.63	-16.1
12.13	-17.3
17.55	-82.6
28.11	-393
32.29	-603
29.44	-449
34.88	-764
33.72	-686
35.35	-791
40.31	-1187

**Ejercicio 7** Dada la siguiente tabla de valores correspondiente a una variable estadística bidimensional  $(X, Y)$

$X$	1	1	2	2	4	4	5	5
$Y$	1	5	2	4	2	4	1	5

calcular la recta de regresión de  $Y$  con respecto a  $X$  y el coeficiente de correlación. ¿Son  $X$  e  $Y$  incorreladas? ¿Son  $X$  e  $Y$  independientes?

**Ejercicio 8** La tabla siguiente muestra los mejores tiempos mundiales en Juegos Olímpicos hasta 1976 en carrera masculina para distintas distancias. La variable  $y$  registra el tiempo en segundos y la variable  $x$  la distancia recorrida en metros.

$y$	9.9	19.8	44.26	103.5	214.9	806.4	1658.4	7795
$x$	100	200	400	800	1500	5000	10000	42196

1. Calcular la recta de regresión de  $y$  sobre  $x$
2. Calcular la varianza residual y el coeficiente de correlación. Indicar si el ajuste lineal es adecuada, usando este último coeficiente.

**Ejercicio 9** Las rectas de regresión de una distribución son: la de  $Y$  sobre  $X$ ,  $y = 0.5x + 2$  y la de  $X$  sobre  $Y$ ,  $x = 1.8y + 5$ . hallar:

1. El coeficiente de correlación
2. El centro de gravedad de los puntos

**Ejercicio 10** Una empresa inmobiliaria ofrece apartamentos en régimen de alquiler, cuyos precios mensuales y número de ellos para cada intervalo de precio son:

Precio de alquiler	Nº de apartamentos
De 700 a 1000	21
1000 a 1100	27
1100 a 1300	34
1300 a 1500	14
1500 a 1800	8

1. Completar la tabla de frecuencias y realizar la representación gráfica más adecuada.
2. Obtener los coeficientes de centralización y dispersión.
3. Si una persona quiere gastar en alquiler entre 1250 y 1350 euros al mes, aproximadamente, ¿a qué porcentaje del total de apartamentos tiene opción?

**Ejercicio 11** Dada la siguiente tabla de doble entrada, que recoge las frecuencias observadas para la variable bidimensional  $(X, Y)$  ( $X$  en horizontal,  $Y$  en vertical)

$Y \setminus X$	5	10	15	20
10	30	2	0	0
20	0	23	1	0
30	0	2	8	0
40	0	0	2	4

Se pide:

1. Distribución marginal de las variables  $X$  e  $Y$ .
2. Varianza de  $Y$  y covarianza de  $(X, Y)$ .
3. Coeficiente de determinación.

**Ejercicio 12** *Al realizar un estudio para comprobar la relación entre el tiempo en días tardado en diferentes empresas en el desarrollo de sus aplicaciones informáticas y en su posterior implantación se obtuvieron los siguientes resultados:*

<i>Desarrollo (X)</i>	75	80	93	65	87	71
<i>Implantación (Y)</i>	82	78	86	73	91	80

1. *Hallar la recta de regresión de Y respecto de X.*
2. *Realizar un ajuste del tipo  $Y = ab^X$*
3. *¿Qué ajuste te parece más conveniente?*



## Tema 2

# Cálculo de probabilidades

### 2.1 Introducción a la teoría de la probabilidad

Vamos a analizar en primer lugar el uso del término “probable” en el lenguaje cotidiano. Supongamos que llaman por teléfono y digo: “es muy probable que sea mi hermano”. ¿Cuál es el sentido de esta frase? Muestra una falta de información sobre la identidad de la persona que me llama. Por otra parte indica que, aunque no sé si es mi hermano el que llama, mi opinión es bastante más favorable a esta posibilidad que a cualquier otra, manifestando por tanto un cierto grado de preferencia entre las opciones posibles. Esta preferencia debe estar fundamentada en algún tipo de información previa que puede ser, por ejemplo, que mi hermano llama con mayor frecuencia que cualquier otra persona, o en otros datos que aunque no sean definitivos para tomar una decisión segura, sí son suficientes para decantarse a favor de una de las opciones posibles. Lo mismo ocurre si, al tirar un dado, comentamos que es menos probable sacar un cinco que no sacarlo. Mostramos un desconocimiento del resultado, pero al mismo tiempo que nuestra opinión es más favorable a una de las opciones.

En esta situación estamos frente a la mayoría de los hechos que ocurren a nuestro alrededor. No en vano se dice que *no hay nada seguro salvo la muerte*. Cuando un científico realiza un experimento o analiza un fenómeno natural es frecuente que los resultados no puedan predecirse con certeza, aunque los experimentos u observaciones se hayan realizado en idénticas condiciones. Un ejemplo muy claro son los fenómenos meteorológicos. Es difícil hacer previsiones sobre la magnitud, intensidad y extensión de las lluvias, el porcentaje de humedad, la dirección y velocidad de los vientos, el valor de la temperatura máxima o mínima, etc. Lo mismo podemos decir sobre las consecuencias de estos fenómenos: volumen de agua en las presas, magnitud de los daños provocados por inundaciones, sequías, accidentes de tráfico, etc. A

la imposibilidad de predecir con certeza los resultados de un fenómeno se le llama azar o aleatoriedad. Los fenómenos con esta característica se llaman fenómenos aleatorios (al principio esta palabra estuvo relacionada exclusivamente con los juegos de azar. En latín *alea* significa suerte).

No obstante, el hecho de no tener certeza de los resultados que se obtendrán en cada prueba particular, no significa que no dispongamos de alguna información sobre estos. Así, aunque no podamos saber si una pareja determinada va a tener hijo o hija como primer retoño, si se sabe que el porcentaje de nacidos varones es aproximadamente del 51%. Una compañía de seguros no sabe a qué edad va a morir un cliente determinado, pero tiene información sobre la proporción de muertes por edades, sexo, etc. Esto le ayudará a decidir sobre el importe que debe exigir por las primas a los asegurados. El resultado de nuestro conocimiento sobre la mayor o menor frecuencia con que aparecen cada uno de los resultados posibles es lo que se pretende cuantificar con el concepto de Probabilidad.

## 2.2 Definiciones de Probabilidad

### 2.2.1 Fenómenos aleatorios

Un *experimento o prueba* es una acción que se realiza con el propósito de hacer algún tipo de observación y obtener una serie de datos a partir de su resultado. *Un experimento es aleatorio* cuando su resultado no es completamente predecible: Aunque el experimento se repita de la misma forma y bajo idénticas condiciones puede dar lugar a diferentes resultados. El concepto contrario de fenómeno aleatorio es el de *fenómeno determinista*, que se caracteriza por obtener los mismos resultados bajo idénticas condiciones. Un ejemplo típico de fenómeno determinista es la velocidad con que llega al suelo un móvil en caída libre, que según la Física Clásica, se obtiene con la fórmula  $v = \sqrt{2gh}$ . De esta forma el resultado sería único. Sin embargo no puede hablarse de fenómenos puramente deterministas. Supongamos que se desea medir la velocidad anterior experimentalmente. En ese caso, la medición viene influenciada por los instrumentos de medida, por la habilidad de los experimentadores, por la resistencia del aire, por las condiciones ambientales, por el lugar geográfico, etc. Todas estas condiciones influyen en los resultados de una forma más o menos desconocida, lo que contribuiría a obtener distintos resultados para el experimento consistente en medir la velocidad de caída de un móvil. El fenómeno sería, desde esta nueva óptica, un fenómeno aleatorio. Por contra, un ejemplo típico de experimento aleatorio es el lanzamiento de un dado. Como el movimiento de un dado se rige asimismo por las leyes de la Física, se comprende que si pudiéramos controlar perfectamente la posición inicial, así como las características de la fuerza de lanzamiento,

deberíamos ser capaces de predecir el resultado, y por tanto se convertiría en un fenómeno determinista. Lo que varía en ambas ocasiones es el enfoque con que tratamos el problema. Lo que es cierto es que en cada caso, uno de estos enfoques, aleatorio o determinista, resulta ser el más adecuado para tratar el problema concreto que estemos intentando resolver.

Centrandonos en los fenómenos aleatorios damos las definiciones siguientes:

Se denomina *espacio muestral*, y lo denotamos por  $\Omega$ , al conjunto de posibles resultados de un experimento aleatorio

Llamaremos *suceso* a cualquier proposición formulada en relación con el experimento, y que en función del resultado del mismo pueda afirmarse categóricamente si ha ocurrido o no. Por ejemplo, si realizamos el experimento aleatorio “lanzar un dado”, el espacio muestral será  $\Omega = \{1,2,3,4,5,6\}$ . Ejemplos de sucesos asociados a este experimento pueden ser los siguientes: “sacar un 6”, “sacar un número par”, “sacar menos de tres”, ... Es directo comprobar que cada suceso puede identificarse con un subconjunto de  $\Omega$ , el de los elementos que satisfagan la proposición. En el ejemplo anterior, los subconjuntos correspondientes a estos sucesos serían, respectivamente,  $\{6\}$ ,  $\{2,4,6\}$ ,  $\{1,2\}$ . *Los sucesos son subconjuntos del espacio muestral*. Los sucesos consistentes en sólo un elemento del espacio muestral suelen llamarse *sucesos elementales*. Se puede considerar el espacio muestral completo  $\Omega$  como suceso, ya que es un subconjunto de sí mismo. Llamamos a este suceso *suceso seguro*.

### 2.2.2 Relaciones y Operaciones con sucesos

A menudo conviene describir un suceso en relación con otros. Como los sucesos son subconjuntos pueden realizarse con ellos las operaciones y relaciones definidas entre conjuntos.

Se dice que el suceso  $A$  *implica el suceso*  $B$  ( $A \subseteq B$ ) si siempre que ocurre  $A$  en el resultado del experimento ocurre también  $B$ . Por ejemplo si  $A$  consiste en sacar un tres al tirar un dado,  $A = \{3\}$ , y  $B$  consiste en sacar una puntuación que sea múltiplo de 3,  $B = \{3, 6\}$ , siempre que se cumpla  $A$ , haya salido un tres al tirar el dado, también saldrá una puntuación múltiplo de 3, y por tanto se cumplirá también el suceso  $B$ .

Se dice que dos sucesos  $A$  y  $B$  *son iguales*,  $A = B$ , si el suceso  $A$  implica el suceso  $B$  y el suceso  $B$  implica el suceso  $A$ .

*La unión de dos sucesos*  $A$  y  $B$  es un suceso  $C = A \cup B$  que se verifica si se verifica al menos uno de ellos ( $A$  o  $B$  o ambos)

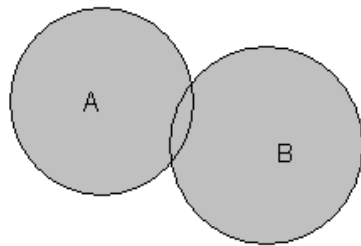
*La intersección de dos sucesos*  $A$  y  $B$  es un suceso  $C = A \cap B$  que se verifica si se verifican ambos sucesos ( $A$  y  $B$ ). *Dos sucesos son incompatibles* si no pueden verificarse simultáneamente. En este caso su intersección es un

suceso que no contiene ningún elemento del espacio muestral, ya que no se verifica nunca. Llamamos *suceso imposible* a este suceso y lo representamos con el signo  $\emptyset$ .

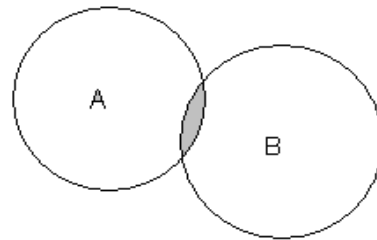
El suceso  $A'$  se llama *contrario*<sup>1</sup> de  $A$  si ocurre *siempre* que no ocurra  $A$ . En el caso de tirar un dado, el suceso contrario de  $\{3, 6\}$  es  $\{1, 2, 4, 5\}$ . Se cumple que  $A \cup A' = \Omega$  y que  $A \cap A' = \emptyset$ .

Se define *diferencia entre dos sucesos*  $A - B = A \cap B'$ , al suceso que se verifica si se cumple  $A$  pero no  $B$ .

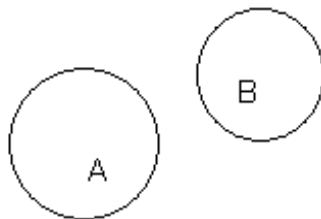
Es bastante útil representar los sucesos, al igual que los conjuntos, por medio de diagramas de Venn. En las gráficas siguientes están representados los diagramas correspondientes a las operaciones y relaciones que acabamos de definir.



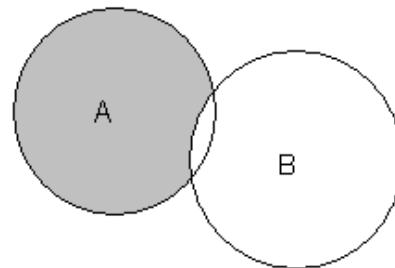
Unión de dos Sucesos



Intersección de dos Sucesos



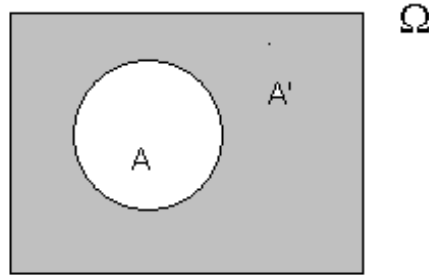
Sucesos incompatibles



Diferencia entre Sucesos

---

<sup>1</sup>También suele emplearse para el suceso contrario la notación  $A^c$  y  $\bar{A}$



Sucesos Contrarios

### 2.2.3 Propiedades de las operaciones entre sucesos

Es inmediato comprobar las siguientes propiedades, que son idénticas a las del Algebra de Boole de Conjuntos.

1. Conmutativa:  $A \cup B = B \cup A$ ;  $A \cap B = B \cap A$
2. Asociativa:  $(A \cup B) \cup C = A \cup (B \cup C)$ ;  $(A \cap B) \cap C = A \cap (B \cap C)$
3. Distributiva  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ ;  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
4. Doble complementación ó Involutiva:  $(A')' = A$
5. Idempotente:  $A \cup A = A$ ;  $A \cap A = A$
6. Elemento neutro:  $A \cup \emptyset = A$ ;  $A \cap \Omega = A$
7. De absorción:  $A \cup (A \cap B) = A$ ;  $A \cap (A \cup B) = A$
8. Leyes de De Morgan:  $(A \cup B)' = A' \cap B'$ ;  $(A \cap B)' = A' \cup B'$

### 2.2.4 Definición frecuentista de probabilidad

Supongamos que repetimos  $n$  veces un experimento aleatorio, y que el suceso  $A$  ha ocurrido en  $n_A$  ocasiones. Entonces, la frecuencia relativa de ocurrencia del suceso  $A$ , tal como se ha definido en el tema anterior será:

$$fr(A) = \frac{n_A}{n}$$

Son evidentes las siguientes propiedades de la frecuencia:

1. La frecuencia relativa de un suceso está comprendida entre 0 y 1, puesto que se cumple que  $0 \leq n_A \leq n$ .

2. La frecuencia de la unión de dos sucesos incompatibles es la suma de las frecuencias de cada uno de estos sucesos.

Si consideramos dos sucesos incompatibles  $A$  y  $B$ , es decir que  $A \cap B = \emptyset$ , se cumple que:

$$\text{card}(A \cup B) = \text{car}(A) + \text{card}(B) = n_A + n_B$$

Por tanto

$$fr(A \cup B) = \frac{n_A + n_B}{n} = \frac{n_A}{n} + \frac{n_B}{n} = fr(A) + fr(B)$$

La idea de probabilidad no es más que una generalización de dicha medida de frecuencia relativa, que podríamos considerar como el límite al que tiende  $fr(A)$  conforme se vaya aumentando el valor de  $n$ . Es un hecho experimental que el valor de la frecuencia relativa se va estabilizando según aumentamos el número de pruebas. La existencia de dicho límite, ya era claramente intuitiva desde la antigüedad en los juegos de azar, sin embargo, este límite se basa en resultados experimentales o empíricos, y por tanto no da un procedimiento de cálculo para la probabilidad, sino sólo un valor aproximado de ésta. Por ejemplo, esta definición nos sirve para asignar un valor a la probabilidad de que una chincheta caiga boca arriba. No obstante, necesitamos una definición que posea un mayor rigor matemático.

### 2.2.5 Definición clásica de probabilidad

La teoría de la probabilidad, como hemos comentado en la sección anterior, tiene su origen en los juegos de azar y de ahí procede también la definición clásica de probabilidad, que está inspirada en la idea de frecuencia relativa de un suceso. Si consideramos las ocurrencias de cara o cruz en el lanzamiento de una moneda, podríamos hacer la experiencia de tirar la moneda una gran cantidad de veces y decidir que la frecuencia relativa obtenida es el valor aproximado de la probabilidad de cada cara, o también, podemos razonar que puesto que sólo hay dos resultados posibles y parece que la moneda es más o menos simétrica, la frecuencia de cara o cruz debe ser más o menos la misma y por tanto su valor debe ser aproximadamente  $\frac{1}{2}$ . Este razonamiento es el que se hace en la definición clásica de probabilidad. Que este razonamiento sea correcto sólo puede comprobarse con la experiencia, haciendo una gran cantidad de lanzamientos de una moneda bien construida. Aunque pueda parecer extraño, así se ha hecho históricamente.

Una probabilidad tiene que ser definida de forma que a cada suceso le corresponda un número, y que se cumplan las propiedades de la frecuencia relativa. Estos requisitos se cumplen en la siguiente definición que se conoce como Regla de Laplace.

Sea  $\Omega$  un espacio muestral con un número finito de elementos. Si puede admitirse que todos los elementos del espacio muestral tienen la misma probabilidad entonces se define como probabilidad de un suceso  $S$ :

$$P(S) = \frac{\text{número de elementos del conjunto } S}{\text{número de elementos del espacio muestral } \Omega} = \frac{\text{Card}(S)}{\text{Card}(\Omega)}$$

Para aclarar el significado de esta fórmula podemos aplicarla en un par de ejemplos sencillos:

**Ejemplo 7** *Calcular la probabilidad de sacar un número múltiplo de tres en el lanzamiento de un dado*

El suceso cuya probabilidad queremos calcular es  $S = \{3, 6\}$ , y el espacio muestral está formado por todos los resultados posibles  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Si admitimos que todos estos sucesos tienen la misma probabilidad, la regla de Laplace nos da un valor para esta probabilidad:

$$P(S) = \frac{\text{número de elementos del conjunto } S}{\text{número de elementos del espacio muestral } \Omega} = \frac{2}{6} = \frac{1}{3}$$

**Ejemplo 8** *Calcular la probabilidad de sacar un número menor que 4 en una baraja española (de 40 cartas)*

En este caso el espacio muestral está formado por todos los resultados posibles, las cuarenta cartas de la baraja. El suceso contiene 12 cartas, 3 por cada uno de los cuatro palos, luego  $P(S) = \frac{12}{40} = \frac{3}{10}$

Es obvio que la probabilidad definida cumple las siguientes propiedades de la frecuencia relativa:

1.  $0 \leq P(S)$
2.  $P(\Omega) = 1$
3. Si  $A_1, A_2$  son sucesos, cumpliendo que  $A_1 \cap A_2 = \emptyset$ , entonces se cumple que

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

### 2.2.6 Propiedades de la probabilidad.

De las anteriores propiedades pueden deducirse las que siguen

1.  $P(\emptyset) = 0$ .
2. Si  $A$  y  $B$  son dos sucesos cumpliendo  $A \subset B$ , entonces  $P(B - A) = P(B) - P(A)$ .

3. Si  $A$  y  $B$  son dos sucesos cumpliendo  $A \subset B$ , entonces  $P(A) \leq P(B)$ .
4.  $P(A) \leq 1 \quad \forall A$
5.  $P(A') = 1 - P(A)$ .
6. Si  $A_1, A_2, \dots, A_n$  son sucesos, cumpliendo que  $A_i \cap A_j = \emptyset \quad \forall i \neq j$ , entonces se cumple que  $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$ .
7.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .
8.  $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$ .

**Demostraciones de las propiedades enunciadas:**

1.  $\emptyset = \emptyset \cup \emptyset$  aplicando el axioma  $P(\emptyset) = P(\emptyset) + P(\emptyset) \implies P(\emptyset) = 0$ .
2.  $B = (B - A) \cup A \implies P(B) = P(B - A) + P(A)$ .
3. Se deduce de la propiedad 2 y del axioma 1 ya que se cumple que  $P(B - A) \geq 0$ .
4.  $A \subset \Omega$ , por lo tanto, aplicando por la propiedad 3 se tiene que  $P(A) \leq P(\Omega) = 1$ .
5. Como  $\Omega = A \cup A'$  y  $A \cap A' = \emptyset$  puede aplicarse el axioma 3 que nos dará  $P(A) + P(A') = P(\Omega) = 1$
6. Por inducción usando el axioma 3.
7. Se parte de la igualdad  $A \cup B = A \cup (B - (A \cap B))$  y se aplica la propiedad 2.
8. Se demuestra por inducción sobre  $n$  partiendo de la propiedad 7.

La definición clásica de probabilidad tiene algunos problemas. En primer lugar sólo es válida para espacios muestrales finitos y además no queda claro cuando deberíamos entender que “los sucesos elementales son equiprobables”, ya que lo que se está definiendo forma parte de la definición. Por ejemplo, en el caso del nacimiento de un niño, varón o mujer, aunque también hay únicamente dos resultados del experimento como en el caso de la moneda, no hay motivos evidentes para aplicar un principio de simetría, por lo que no podemos usar la regla de Laplace y sería más adecuado usar una definición frecuentista, usando el valor de la frecuencia relativa. En la actualidad se atribuye una probabilidad de 0.51 para el nacimiento de varón y de 0.49 para mujer.



Tampoco la fórmula de Laplace es apropiada para el siguiente caso: Supongamos que queremos calcular la probabilidad de sacar dos caras tirando dos monedas. Podíamos decir que el espacio muestral está formado por tres elementos consistente en los tres resultados posibles: sacar dos caras: CC, sacar dos cruces: XX o sacar una cara y una cruz: CX. Como sólo uno de éstos, CC, forma parte del suceso cuya probabilidad queremos conocer, el número de elementos de  $S$  es 3. Si suponemos que los tres elementos del espacio muestral son igualmente probables la probabilidad del suceso CC sería  $\frac{1}{3}$ . Este resultado, al que se llega usando la fórmula de Laplace, aunque formalmente correcto, no está de acuerdo con la experiencia, que no da la misma frecuencia relativa a los tres sucesos. El motivo es que la composición CX, puede obtenerse de dos formas variando la moneda en la que aparece la cara y la cruz. La regla de Laplace nos daría un resultado de acuerdo con la experiencia si suponemos que el espacio muestral está formado por los resultados  $\{CC, CX, XC, XX\}$ , con lo que obtendríamos para el suceso sacar dos caras una probabilidad  $\frac{1}{4}$ , y que coincidiría aproximadamente con el valor para la frecuencia relativa del suceso, que obtendríamos experimentalmente realizando una gran cantidad de tiradas con las dos monedas.

A pesar de las limitaciones de la definición clásica, en la práctica se usa frecuentemente, teniendo cuidado de adaptar al caso real las consideraciones de equiprobabilidad que exige la formulación clásica. No obstante, muchas cuestiones y problemas no pueden adaptarse al modelo de definición de probabilidad de Laplace, ni al modelo de definición experimental como frecuencia relativa. Por ejemplo, ninguna de estas definiciones son muy adaptables para resolver preguntas como las siguientes: ¿cuál es la probabilidad de que estalle una tercera guerra mundial?, ¿cuál es la probabilidad de que un automóvil de un nuevo modelo dure más de 8 años?, ¿cuál es la probabilidad de que se encuentre una nueva galaxia en el presente siglo?....

### 2.2.7 Definición axiomática de probabilidad.

La definición axiomática de probabilidad, debida a Kolmogorov, es más general y se adapta a un conjunto más amplio de situaciones. Para poder dar dicha definición necesitamos disponer de un espacio muestral  $\Omega$  sobre el que esté definida un  $\sigma$ -álgebra de sucesos. es decir imponemos una cierta estructura al conjunto de sucesos

Un  $\sigma$ -álgebra de sucesos es el conjunto de todos los sucesos de interés asociados a un experimento aleatorio. Denotamos por  $\mathcal{A}$  al  $\sigma$ -álgebra de sucesos. Se verifica que

$$\mathcal{A} \subset \mathcal{P}(\Omega)$$

donde  $\mathcal{P}(\Omega)$  representa el conjunto de las partes de  $\Omega$ , esto es, el conjunto de todos los subconjuntos de  $\Omega$ .

Al conjunto  $\mathcal{A}$  le impondremos que cumpla las siguientes propiedades:

- a)  $\Omega \in \mathcal{A}$
- b) las operaciones unión, intersección, y complementación de elementos de  $\mathcal{A}$  realizadas un número finito (o infinito numerable) de veces, son internas en  $\mathcal{A}$ .

No siempre tiene que ocurrir que  $\mathcal{A} = \mathcal{P}(\Omega)$ , es decir que no es necesario que todos los subconjuntos de  $\Omega$  sean sucesos. Por ejemplo, si lanzando un dado y estamos apostando a par o impar solamente, entonces

$$\mathcal{A} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\}$$

es el conjunto de interés. Puede comprobarse que este conjunto cumple las propiedades de un  $\sigma$ -álgebra de sucesos. Es conveniente observar que  $\emptyset \in \mathcal{A}$ , ya que es el complementario de  $\Omega$ . Es decir, el conjunto vacío siempre será un suceso.

La definición axiomática de probabilidad, dada por Kolmogorov, sería la siguiente:

Sea  $\Omega$  un espacio muestral y  $\mathcal{A}$  un  $\sigma$ -álgebra de sucesos definida sobre  $\Omega$ . Una probabilidad es una aplicación  $P$  cumpliendo:

1.  $P : \mathcal{A} \longrightarrow [0, 1]$
2.  $P(\Omega) = 1$
3. Si  $A_1, A_2, \dots, A_n$  son sucesos, cumpliendo que  $A_i \cap A_j = \emptyset \quad \forall i \neq j$ , entonces se cumple que

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

Si la unión de sucesos es infinita numerable también se cumple:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \quad i \in \mathbb{N}$$

Es de notar que esta definición contiene a la de Laplace como un caso particular: En efecto si  $\Omega$  es un espacio muestral finito formado por  $n$  sucesos elementales  $s_1, s_2, \dots, s_n$ , que suponemos equiprobables, se tiene que la probabilidad de cualquier suceso sería  $\frac{1}{n}$ , ya que usando los axiomas 2 y 3 obtenemos

$$P(\Omega) = P\{s_1, s_2, \dots, s_n\} = P\{s_1\} + P\{s_2\} + \dots + P\{s_n\} = 1$$

y como los sucesos tienen por hipótesis la misma probabilidad, se deduce que la probabilidad de cualquiera de estos sucesos elementales es  $\frac{1}{n}$ . La probabilidad de un suceso compuesto se obtendría sumando la probabilidad de cada uno de los sucesos elementales que lo formen. Por ejemplo:

$$\begin{aligned} P\{s_i, s_j, s_k\} &= P\{s_i\} + P\{s_j\} + P\{s_k\} = \frac{1}{n} + \frac{1}{n} + \frac{1}{n} = \frac{3}{n} = \\ &= \frac{\text{número de elementos del suceso}}{\text{número de elementos del espacio muestral}} \end{aligned}$$

que sigue la regla de Laplace.

Es de notar que la definición axiomática no nos suministra un procedimiento de cálculo de la probabilidad de un experimento concreto, limitándose a indicar qué propiedades ha de cumplir una función  $P$  para que sea considerada una probabilidad en sentido matemático. También aquí se cumplen las propiedades enumeradas en la página 63, puesto que los axiomas son también las propiedades de la frecuencia relativa. Estas propiedades nos van a permitir hallar la probabilidad de cualquier suceso a partir de la probabilidad de unos cuantos sucesos. ¿Cómo definimos la probabilidad de estos cuantos sucesos? Con tal de que se cumplan los axiomas de la probabilidad, tenemos libertad en este sentido. La conveniencia de una forma u otra de dar valores concretos a la probabilidad de estos pocos sucesos viene determinado por el caso particular que estemos tratando y los valores numéricos de la probabilidad se deben asignar buscando un buen acuerdo entre el modelo teórico y el real o experimental. Para asignar valores a las probabilidades de estos cuantos sucesos se emplean los siguientes enfoques: *frecuentista, clásico y subjetivo*.

El enfoque frecuentista está basado en la interpretación de la probabilidad como límite de la frecuencia relativa de un suceso cuando el número de pruebas tiende a infinito. El enfoque clásico consiste en la aplicación de regla de Laplace, siempre que sea adaptable. El enfoque subjetivo asigna valores a la probabilidad de algunos sucesos en función de la opinión de determinadas personas, que se supone que son expertos en el tema. Esta última forma de asignar valores para la probabilidad se usa, por ejemplo, en las apuestas o en las quinielas.

Una vez asignados valores de probabilidad a algunos sucesos, a menudo los sucesos elementales, por medio de cualquiera de los enfoques indicados, debemos calcular las probabilidades de los otros sucesos de interés. Ahora podemos calcular indirectamente el valor para las probabilidades de sucesos más complejos, para los que quizá no es fácil asignar directamente valores de probabilidad. Se puede recurrir al uso de propiedades, como por ejemplo las indicadas en la página 63, que como hemos indicado también se cumplen

si usamos la definición de Kolmogorov, u otras propiedades, como las que damos en el resto del capítulo.

## 2.3 Recursos para el cálculo de probabilidades

Si queremos calcular la probabilidad de un suceso usando la fórmula de Laplace:

$$\frac{\text{número de elementos del suceso}}{\text{número de elementos del espacio muestral}}$$

debemos poder contar el número de resultados posibles, *y equiprobables*, del experimento y también el número de elementos del suceso. Estos valores, a veces, no son fáciles de calcular si no se dispone de algunas reglas de conteo. A continuación damos algunas reglas que pueden ser útiles para este propósito.

### 2.3.1 Regla de multiplicación

Supongamos que el experimento aleatorio consiste en tirar una moneda y después un dado. La variable aleatoria que consideramos está formada por el resultado de la moneda, cara o cruz, y la puntuación del dado. Los posibles resultados del experimento podemos indicarla por medio de una tabla de doble entrada.

	1	2	3	4	5	6
C	C1	C2	C3	C4	C5	C6
X	X1	X2	X3	X4	X5	X6

Está claro que el número de elementos del espacio muestral se obtiene multiplicando el número de elementos obtenidos en la primera prueba (2 posiciones de la moneda) por el número de resultados obtenidos en la segunda prueba (6 posibles puntuaciones del dado). Así que el espacio muestral asociado a este experimento compuesto estaría formado por  $2 \times 6 = 12$  elementos.

Si el experimento aleatorio consiste en tirar una moneda, tirar un dado y después sacar una carta de una baraja, el espacio muestral cuyos elementos constarían de tres resultados (por ejemplo *Cara, tres y sota de espadas*) estaría formado por  $2 \times 6 \times 40 = 480$  elementos.

**Ejemplo 9** *Calcular la probabilidad, realizando el experimento anterior, de sacar una puntuación par en el dado y una carta de oros .*

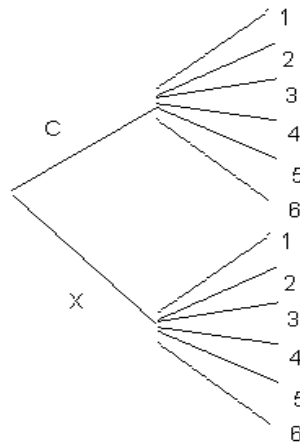
Ya hemos calculado el número de elementos del espacio muestral: 480 resultados posibles. Por la misma regla podemos calcular el número de elementos del suceso: *2 figuras de la moneda  $\times$  3 puntuaciones pares para el dado  $\times$  10 cartas de oros. Por lo tanto la probabilidad pedida sería:*

$$\text{Probabilidad} = \frac{\text{número de elementos del suceso}}{\text{número de elementos del espacio muestral}} = \frac{2 \times 3 \times 10}{480} = \frac{1}{8}.$$

### 2.3.2 Diagramas de árbol

Una representación en forma de diagrama de árbol es también útil para poner de manifiesto la regla de multiplicación:

Para el ejemplo de la moneda y el dado tal diagrama tomaría la forma de la figura



Recorriendo cada rama del árbol desde la raíz, obtenemos los distintos resultados del experimento.

### 2.3.3 Combinatoria

A continuación incluimos algunas nociones de combinatoria que pueden resultar útiles para acometer algunos problemas de probabilidad. La combinatoria estudia las posibles agrupaciones, cumpliendo ciertas condiciones, que pueden formarse con un número finito de elementos. A veces, nos sirve de ayuda para evaluar la probabilidad siguiendo la regla de Laplace ya que, en ocasiones, permite evaluar el número total de elementos equiprobables pertenecientes al espacio muestral y cuántos de éstos cumplen la propiedad característica del suceso cuya probabilidad queremos calcular.

#### Variaciones

Las variaciones de  $n$  elementos tomados de  $m$  en  $m$  son los distintos grupos con  $m$  elementos que se pueden formar eligiendo  $m$  elementos de los  $n$  totales, considerando que:

a) Una variación es distinta de otra si se distingue en algún elemento, o si se distingue en el orden de los elementos entre sí.

b) Cada elemento sólo puede aparecer como máximo una vez dentro de cada variación.

Por lo tanto, las variaciones son todas las agrupaciones ordenadas de  $m$  objetos seleccionados entre los  $n$  elementos de un conjunto dado.

Por ejemplo, si se toma una baraja con cuarenta cartas, cada una de las distintas formas en que se pueden repartir 2 cartas, teniendo en cuenta cuál ha sido la primera y cuál la segunda, es una variación de las 40 cartas tomadas en grupos de dos. El número de pares de cartas, puede obtenerse por medio del siguiente razonamiento: La primera carta puede darse de 40 maneras diferentes, y la segunda de 39, ya que queda una carta menos. Por lo tanto el número de pares de cartas diferentes serán  $40 \times 39$ . En general para obtener el número de variaciones diferentes de  $n$  elementos tomados de  $m$  en  $m$  que se denota por  $V_{n,m}$ , se usa la fórmula siguiente:

$$V_{n,m} = n \times (n - 1) \times (n - 2) \times \dots \times [n - (m - 1)]$$

Por ejemplo:  $V_{5,3} = 5 \times 4 \times 3 = 60$ .

### Permutaciones

Las permutaciones de  $n$  elementos son las variaciones de estos  $n$  elementos tomados de  $n$  en  $n$ . En este caso una permutación sólo puede distinguirse de otra en el orden en que están ordenados sus elementos, ya que todas ellas contendrán los  $n$  elementos del conjunto completo. Por tanto, las permutaciones son las distintas formas en que se pueden ordenar los  $n$  elementos de un conjunto. Supongamos ahora que tenemos cuatro cartas y que las señalamos con las letras a,b,c,d. ¿De cuántas formas pueden ordenarse? La primera carta puede colocarse de 4 formas, la segunda sólo de tres, porque una de ellas ya se ha tomado en la primera extracción, la tercera de dos formas. La última sólo puede ser la que nos quede. Por lo tanto hay  $4 \times 3 \times 2 \times 1 = 24$  ordenaciones diferentes para estas cuatro cartas. A continuación detallamos cada una de estas 24 permutaciones. Comenzamos escribiendo las variaciones con un único elemento:

a, b, c, d

Ahora con dos elementos:

ab, ac, ad,  
ba, bc, bd  
ca, cb, cd  
da, db, dc

Con tres elementos:

abc, abd	acb, acd	adb, a dc
bac, bad	bca, bcd	bda, bdc
cab, cad	cba, cbd	cda, cdb
dab, dac	dba, dbc	dca, dc b

Para formar las permutaciones añadimos ahora el elemento que falta en cada una de las anteriores

abcd, abdc	acbd, acdb	adbc, adcb
bacd, badc	bcad, bcda	bdac, bdca
cabd, cadb	cbad, cbda	cdab, cdba
dabc, dacb	dbac, dbca	dcab, dcba

El número de permutaciones de  $n$  elementos se denota por  $P_n$  y se calcula con la expresión

$$P_n = V_{n,n} = n \times (n-1) \times (n-2) \times \dots \times [n - (n-1)] = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1 = n!$$

Esta última expresión,  $n!$ , se conoce con el nombre de *factorial de  $n$* , y como se ve representa el producto de todos los enteros positivos de 1 a  $n$ . Por conveniencia de cálculo se define  $0! = 1$ .

Se puede comprobar que se cumple

$$V_{n,m} = n \times (n-1) \times (n-2) \times \dots \times (n - (m-1)) = \frac{P_n}{P_{n-m}} = \frac{n!}{(n-m)!}$$

### Combinaciones

Si consideramos como iguales las variaciones que tengan exactamente los mismos elementos y como distintas aquellas que se distinguen en, al menos, un elemento, los grupos distintos que resulten forman las combinaciones. En concreto:

Las combinaciones de  $n$  elementos tomados de  $m$  en  $m$ ,  $C_{n,m}$ , son los distintos grupos con  $m$  elementos que se pueden formar eligiendo  $m$  elementos entre los  $n$  totales. Una combinación es distinta de otra si:

- Se distingue en algún elemento.
- Cada elemento sólo puede aparecer como máximo una vez dentro de cada combinación.

Siguiendo con el ejemplo del reparto de dos cartas de cuarenta. En el caso de las variaciones hemos tenido en cuenta el orden en que nos han dado las cartas, pero en algunos juegos no importa el orden en que se reciben éstas

cartas. Si suponemos formadas las posibles variaciones de 40 cartas tomadas de dos en dos, observamos que la variación AB, y la variación BA forman ahora una única combinación, ya que no vamos a tener en cuenta el orden de reparto.

En el ejemplo detallado en el párrafo anterior, en el que se consideraban las variaciones de las 4 cartas tomados de 2 en 2, podemos observar que cada pareja de dos letras aparece repetida en dos variaciones. Por lo tanto

$$C_{4,2} = \frac{V_{4,2}}{2} = \frac{V_{4,2}}{P_2} = \frac{4 \times 3}{2 \times 1} = 6$$

Estas seis combinaciones posibles son:

ab, ac, ad,  
bc, bd  
cd

En general, el número de combinaciones de  $n$  elementos tomados de  $m$  en  $m$  se escribe  $C_{n,m}$ , o  $\binom{n}{m}$ , y su valor está dado por la fórmula:

$$C_{n,m} = \frac{V_{n,m}}{P_m} = \frac{V_{n,m} P_{n-m}}{P_m P_{n-m}} = \frac{n!}{(n-m)!m!} = \binom{n}{m}$$

Los valores  $C_{n,m} = \binom{n}{m}$  suelen conocerse con el nombre de *números combinatorios*.

### Variaciones con repetición

Las variaciones con repetición de  $n$  elementos tomados de  $m$  en  $m$  son los distintos grupos con  $m$  elementos que se pueden formar eligiendo  $m$  elementos de los  $n$  totales. Se debe cumplir que:

- a) Cada elemento puede repetirse hasta  $m$  veces dentro de cada variación con repetición
- b) Una variación con repetición se distingue de otra por que contenga elementos distintos, por el orden en que están colocados estos elementos o por el número de veces que esté repetido cada elemento.

En este caso  $m$  puede ser mayor que  $n$ .

Por ejemplo, si consideramos las distintas variaciones con repetición que pueden formarse a partir de los diez dígitos, tomados en grupos de tres, obtenemos los números desde 000, hasta el 999, es decir  $V'_{10,3} = 10^3 = 1000$  variaciones con repetición.

En general representamos el número de variaciones con repetición como  $V'_{n,m} = n^m$



### Permutaciones con repetición

Si queremos calcular las distintas ordenaciones de un conjunto de  $n$  elementos cuando algunos de estos elementos son indistinguibles entre sí, como ocurre con las 8 letras de la palabra PAPANATA en la que aparece, 4 veces la letra A, dos veces la P, y una vez cada una de las letras N y T, aplicamos la siguiente expresión:

$$P'_{8; 4,2,1,1} = \frac{8!}{4!2!1!1!}$$

Para escribir esta expresión se parte del número total de permutaciones que se obtendrían si, de forma artificial, suponemos distintos los elementos iguales de entre los  $n$  elementos de partida. En el caso del ejemplo, podemos distinguir las letras iguales con subíndices:  $P_1A_1P_2A_2N_1A_3T_1A_4$ . Se divide el número total de permutaciones por el número de veces que cada una de ellas aparecería repetida si igualáramos las letras, suprimiendo los subíndices.

En general, el número de variaciones con repetición de  $n$  elementos entre los que hay  $m$  grupos de elementos diferentes, cada uno de ellos con  $n_1, n_2, \dots, n_m$  elementos iguales siendo  $n = n_1 + n_2 + \dots + n_m$ , se obtienen con la expresión:

$$P'_{n; n_1, n_2, \dots, n_m} = \frac{n!}{n_1!n_2!\dots n_m!}$$

### Combinaciones con repetición

Si tenemos un conjunto con  $n$  elementos distintos y admitimos que cada combinación puede tener elementos repetidos, obtenemos las combinaciones con repetición. En concreto:

Las combinaciones con repetición de  $n$  elementos tomados de  $m$  en  $m$ ,  $C'_{n,m}$ , son los distintos grupos con  $m$  elementos que se pueden formar eligiendo  $m$  elementos de los  $n$  totales.

- a) Una combinación es distinta de otra si se distingue en algún elemento.
- b) Cada elemento puede repetirse hasta un total de  $m$  veces dentro de cada combinación.

En este caso  $m$  puede ser mayor que  $n$ .

La expresión de cálculo del número de combinaciones con repetición es:

$$C'_{n,m} = C_{n+m-1,m} = \binom{n+m-1}{m} = \frac{(n+m-1)!}{m!(n-1)!}$$

Las combinaciones con repetición de las dos letras {a, b} en grupos con cuatro elementos son:

$$C'_{2,4} = \frac{(2+4-1)!}{4!(2-1)!} = 5$$

Escribimos estas 5 combinaciones con repetición:

a a a a    b b b b    a a a b    a a b b    a b b b

Como puede verse en el ejemplo no se tiene en cuenta los cambios de orden para formar diferentes combinaciones. Por ejemplo la combinación con repetición “aaab” es la misma que la combinación “aaba”.

### 2.3.4 De lo particular a lo general

Cuando tenemos un problema bastante complejo, un procedimiento bastante usado es comenzar resolviendo casos particulares de este problema, que debido a su simplicidad, sean más fáciles de resolver. Esperamos que su resolución nos arroje alguna luz que nos permita vislumbrar la solución del caso general. No es raro que este método sea útil para calcular probabilidades de sucesos. Lo ilustramos con el siguiente ejemplo.

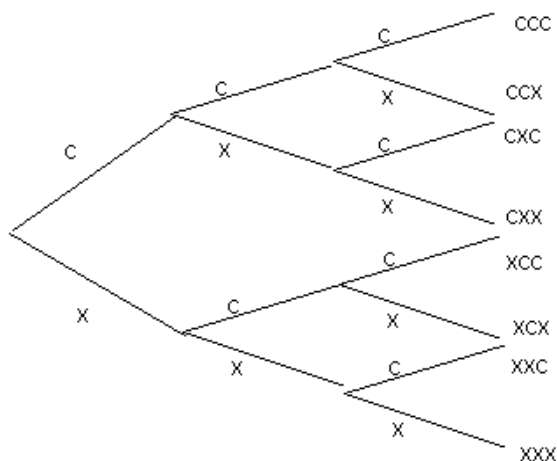
**Ejemplo 10** *Se realiza una tirada de  $n$  monedas ( $n \geq 3$ ). Calcular la probabilidad de obtener exactamente tres caras.*

Aquí se nos pide una expresión general en función de  $n$ .

Comenzamos por el caso particular más sencillo:  $n = 3$ . Usamos la fórmula de Laplace:

$$\text{Probabilidad} = \frac{\text{número de elementos del suceso}}{\text{número de elementos del espacio muestral}}$$

Calculamos el denominador. Para ello necesitamos contar todos los resultados posibles en las tres monedas. Usamos un diagrama de árbol



Usando la regla de multiplicación, el número de elementos del espacio muestral es  $2 \times 2 \times 2 = 2^3 = 8$ . Estos 8 resultados están especificados en la columna derecha del diagrama de árbol.

¿Cuántos sucesos de entre ellos cumplen que tienen tres caras? Solamente el resultado CCC. Por tanto la probabilidad de obtener exactamente tres caras tirando tres monedas es  $\frac{1}{8}$ .

Nos lo ponemos ahora un poco más difícil. Consideraremos que tiramos cuatro monedas.

El número de elementos del espacio muestral lo calculamos, análogamente al caso anterior. Éste número es ahora  $2 \times 2 \times 2 \times 2 = 2^4 = 16$ .

No es demasiado difícil describir los casos favorables al suceso: CCCX, CCXC, CXCC, XCCC. Observamos que son todas las permutaciones con repetición de 3 caras y una cruz. Usando la fórmula de las permutaciones con repetición de 4 elementos donde se repiten 3 de ellos comprobamos que efectivamente es así:

$$P'_{4; 3,1} = \frac{4!}{3!1!} = 4$$

Por tanto la probabilidad de obtener tres caras en una tirada de cuatro monedas es  $\frac{4}{16} = \frac{1}{4}$ .

Resolvemos ahora el problema primitivo considerando  $n$  monedas. El número de elementos del espacio muestral sería  $2^n$ , y el número de elementos con exactamente tres caras es  $P'_{n; 3, n-3}$

La probabilidad de obtener tres caras en una tirada de  $n \geq 3$  monedas es:

$$\frac{P'_{n; 3, n-3}}{2^n} = \frac{\frac{n!}{3!(n-3)!}}{2^n} = \frac{1}{6}n(n-1)(n-2)2^{-n}$$

### 2.3.5 La probabilidad geométrica

En algunas ocasiones la fórmula de Laplace no es adecuada. Consideremos el caso siguiente

**Ejemplo 11** *En un segmento de longitud 1 marcamos dos puntos interiores. Calcular la probabilidad de que los tres segmentos resultantes puedan formar un triángulo.*

No parece que en este caso tenga sentido contar el número de elementos del espacio muestral, que estaría formado por todas las ternas de segmentos cuyas longitudes sumen uno. Evidentemente este conjunto es de cardinal infinito, así que la regla de Laplace no es aplicable en esta ocasión. Para calcular esta probabilidad se suele identificar el número de elementos del espacio muestral

con el área del conjunto que lo represente. Es lo que haremos aquí para dar una solución a este problema:

Sean  $x, y, z$  las longitudes de los tres segmentos. Como hemos dicho, el espacio muestral está formado por las ternas de longitudes que cumplen  $x + y + z = 1$ . Las ternas que forman el suceso cuya probabilidad queremos calcular son los que, además, cumplan las tres condiciones siguientes:

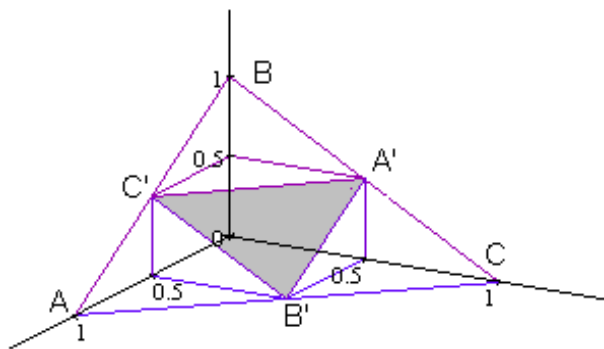
$$\begin{aligned} x &< y + z \\ y &< x + z \\ z &< x + y \end{aligned} \quad (2.1)$$

ya que cada lado de un triángulo ha de ser menor que la suma de los otros dos.

Teniendo en cuenta que  $x + y + z = 1$  y las expresiones 2.1, deducimos

$$\begin{aligned} x < y + z = 1 - x &\implies x < 0.5 \\ y < x + z = 1 - y &\implies y < 0.5 \\ z < x + y = 1 - z &\implies z < 0.5 \end{aligned}$$

En la siguiente figura se ha realizado una representación gráfica tridimensional de ambos conjuntos. Los puntos del triángulo  $ABC$  forman el espacio muestral (las longitudes de los segmentos son magnitudes positivas, por eso están representadas en el octante con las tres coordenadas positivas del plano  $x + y + z = 1$ ). Los puntos interiores del triángulo  $A'B'C'$  son las ternas  $x, y, z$  que cumplen el suceso. Verifican las condiciones necesarias:  $x < 0.5, y < 0.5, z < 0.5, x + y + z < 1$ .



Calculamos la probabilidad como cociente entre las dos áreas

Probabilidad de que los segmentos de longitudes  $x, y, z$  formen un triángulo =

$$= \frac{\text{Área del triángulo } A'B'C'}{\text{Área del triángulo } ABC} = \frac{1}{4}$$

Esta enfoque de la probabilidad, basado en las medidas de la representación gráfica de los sucesos, se conoce con el nombre de *probabilidad geométrica*.

## 2.4 Probabilidad condicionada.

Con este concepto tratamos de registrar el cambio que experimenta la probabilidad de un suceso si aumenta la información de que disponemos sobre el resultado del experimento. Supongamos que tenemos el billete 347 de una lotería de 1000 números. Nuestra probabilidad de ganar el premio es, siguiendo la regla de Laplace  $\frac{1}{1000}$ . Pero si aparece un amigo nuestro y nos dice. “No me acuerdo en que número ha caído el premio, pero me acuerdo que acababa en 7”. ¿Cuál es ahora nuestra probabilidad de ganar el premio? Ya sólo hay 100 casos posibles, el de los números de tres cifras que terminen en 7, y nosotros tenemos uno de ellos, nuestra probabilidad es ahora  $\frac{1}{100}$ .

Supongamos que queremos calcular la probabilidad del suceso  $A$  sabiendo que se ha presentado el suceso  $B$ . Esto es lo que se denomina *probabilidad de  $A$  condicionada a  $B$*  y se denota como  $P(A/B)$ . La siguiente definición nos relaciona la probabilidad condicionada, también llamada “a posteriori” en función de las probabilidades “a priori”, definidas al principio, cuando se carecía de información adicional.

$$\text{Por definición } P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Así, en el ejemplo anterior, si  $A$  es el suceso que consiste en “que salga el número 347” y  $B$  el suceso que consiste en “que salga un número terminado en 7”

$$\begin{aligned} P(A/B) &= P(\text{“que salga el número 347”/si “ha salido un número terminado} \\ &\quad \text{en 7”}) = \\ &= \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{1000}}{\frac{100}{1000}} = \frac{1}{100} \end{aligned}$$

Del mismo modo, tendríamos que  $P(B/A) = \frac{P(A \cap B)}{P(A)}$  de donde deducimos la fórmula para la probabilidad de la intersección de dos sucesos:

$$P(A \cap B) = P(A/B)P(B) = P(B/A)P(A) \quad (2.2)$$

Se cumple que  $P(A'/B) = 1 - P(A/B)$ .

### 2.4.1 Independencia de un par de sucesos

*Dos sucesos se dicen independientes* cuando  $P(A/B) = P(A)$ . Como  $P(A \cap B) = P(A/B)P(B)$  si los sucesos  $A$  y  $B$  son independientes se cumple que

$$P(A \cap B) = P(A)P(B).$$

Recíprocamente, si  $P(A \cap B) = P(A)P(B)$  despejando  $P(A)$  obtenemos:

$$P(A) = \frac{P(A \cap B)}{P(B)} = P(A/B)$$

y por tanto  $A$  y  $B$  son independientes.

**Ejemplo 12** Lanzamos un dado, hallar la probabilidad de sacar un número mayor o igual que 5 sabiendo que se ha sacado un número par.

$A = \{\text{sacar un número mayor o igual que 5}\} = \{5, 6\}$ ,  $B = \{\text{sacar par}\} = \{2, 4, 6\}$ ,  $A \cap B = \{6\}$

$$1. P(A) = 2/6; \quad P(B) = 3/6; \quad P(A \cap B) = 1/6.$$

Por tanto

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = 1/3$$

Los sucesos  $A$  y  $B$  son independientes porque  $P(A) = P(A/B) = 1/3$ .

Otra forma de calcular  $P(A/B)$  es usar la regla de Laplace considerando el suceso  $B = \{2, 4, 6\}$  como universo o espacio muestral (en el diagrama de Venn se considera solamente lo que está dentro de  $B$ , marcado en gris, haciendo caso omiso del resto). El espacio muestral sólo tiene ahora 3 elementos y sólo uno de ellos, 6, verifica el suceso  $A$ .

$$P(A/B) = \frac{\text{número de veces que se cumple el suceso } A \text{ dentro de } B}{\text{número de elementos de } B} = \frac{1}{3}$$

A veces se confunden los conceptos de sucesos independientes y los incompatibles. Aprovechamos este ejemplo para resaltar la diferencia entre estos conceptos. Como hemos dicho los sucesos  $A$  y  $B$  son independientes porque  $P(A) = P(A/B) = \frac{1}{3}$ . Sin embargo no son incompatibles. Dos sucesos son incompatibles cuando su intersección es el suceso imposible y por tanto no pueden verificarse simultáneamente. En este caso  $A \cap B = \{6\} \neq \phi$ . Cuando se obtenga un 6 al arrojar el dado se cumplirán simultáneamente los sucesos  $A$  y  $B$ .

**Ejemplo 13** Lanzamos un dado, hallar la probabilidad de sacar un número mayor o igual que 4 sabiendo que se ha sacado un número par.

Sean  $A = \{\text{sacar un cuatro, un cinco o un seis}\}$ ,  $B = \{\text{sacar par}\}$

$$A \cap B = \{4, 6\}$$

$P(A) = 3/6$ ;  $P(B) = 3/6$ ;  $P(A \cap B) = 2/6$ ;  $P(A/B) = 2/3$ . Los sucesos  $A$  y  $B$  no son independientes.

**Ejemplo 14** En una empresa los trabajadores están distribuidos en dos plantas. En cada una de ellas los trabajadores entran en tres turnos, porque la maquinaria no puede quedar inactiva completamente. En la tabla siguiente aparece el número de trabajadores que tiene la empresa por planta y turno.

	Primer turno 0 a 8 horas	Segundo turno 8 a 16 horas	Tercer turno 16 a 24 horas	Total planta
planta I	250	150	200	600
planta II	100	100	200	400
Total	350	250	400	1000

Se selecciona un empleado al azar y se consideran los sucesos:

$A$ , que se cumple si se elige un empleado de la primera planta.

$B$ , que se cumple si se elige un trabajador del segundo turno (8 a 16).

$C$ , que se cumple si se selecciona un trabajador del primer turno (de 0 a 8).

Estudiamos la independencia de los sucesos  $A$  y  $B$  y también la independencia entre los sucesos  $A$  y  $C$ :

Una tabla como la precedente, en la que constan la frecuencia de las intersecciones de los sucesos considerados se conoce con el nombre de *tabla de contingencia* y puede ser de gran ayuda en la resolución de algunos problemas de probabilidad.

$$P(A) = \frac{600}{1000} = 0.6; \quad P(A/B) = \frac{150}{250} = 0.6.$$

Como ambos valores son iguales, los sucesos  $A$  y  $B$  son independientes.

Estudiamos ahora la independencia de los sucesos  $A$  y  $C$ :

$$P(A) = \frac{600}{1000} = 0.6; \quad P(A/C) = \frac{250}{350} = 0.71429.$$

Como los valores obtenidos no son iguales, los sucesos  $A$  y  $C$  no son independientes.

### 2.4.2 Independencia de más de dos sucesos

La probabilidad de la intersección de más de dos sucesos viene dada por la expresión

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1) \times P(A_2/A_1) \times P(A_3/A_1 \cap A_2) \times \dots \times P(A_k/A_1 \cap A_2 \cap \dots \cap A_{k-1})$$

Se dice que los sucesos  $A_1, A_2, \dots, A_k$  son *independientes* si para todo subconjunto  $\{A_{i_1}, A_{i_2}, \dots, A_{i_l}\}$  de  $\{A_1, A_2, \dots, A_k\}$ , es decir que  $\{A_{i_1}, A_{i_2}, \dots, A_{i_l}\} \subseteq \{A_1, A_2, \dots, A_k\}$ , se verifica:

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_l}) = P(A_{i_1}) \times P(A_{i_2}) \times \dots \times P(A_{i_l})$$

## 2.5 Teorema de la Probabilidad Total.

Veremos en primer lugar un ejemplo que nos ayudara a entender el campo de aplicación del teorema de la probabilidad total:

**Ejemplo 15** *Los trabajadores de una fábrica pueden ser directivos, obreros u oficinistas. Si se elige un trabajador al azar se sabe que:*

- 1) *La probabilidad de que sea un directivo =  $P(A_1) = 0.10$ .*
  - 2) *La probabilidad de que sea un obrero =  $P(A_2) = 0.75$ .*
  - 3) *La probabilidad de que sea oficinista =  $P(A_3) = 0.15$ .*
  - 4) *La probabilidad de que sea mujer, si es un directivo =  $P(B/A_1) = 0.20$ .*
  - 5) *La probabilidad de que sea mujer, si es un obrero =  $P(B/A_2) = 0.45$ .*
  - 6) *La probabilidad de que sea mujer, si es un oficinista =  $P(B/A_3) = 0.50$ .*
- ¿Cuál es la probabilidad de que un trabajador elegido al azar sea mujer =  $P(B)$ ?*

Esta última probabilidad es un ejemplo de la llamada probabilidad total. El espacio muestral es el conjunto total de trabajadores. Las probabilidades de ser mujer dentro de cada categoría de empleados de la fábrica son probabilidades condicionadas (la información suplementaria en cada caso es que se conoce la categoría profesional a la que pertenece el trabajador seleccionado).

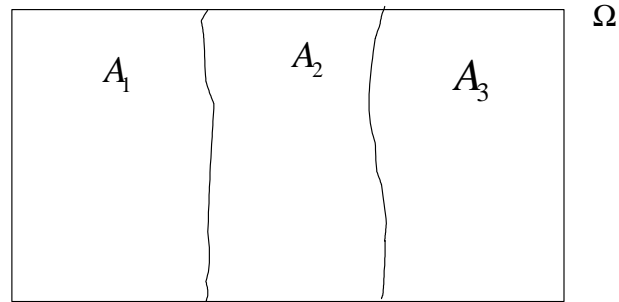
Un conjunto de sucesos  $A_1, A_2, \dots, A_n$  se dicen que forman una *partición* de  $\Omega$  ó también que forman un *sistema exhaustivo y excluyente*, si se cumple:

1. 
$$\bigcup_{i=1}^n A_i = A_1 \cup A_2 \cup \dots \cup A_n = \Omega$$
2. 
$$A_i \cap A_j = \emptyset \quad \forall i \neq j$$

En el ejemplo estos sucesos podrían ser:  $A_1$ , que se cumple si la persona seleccionada pertenece a la categoría de directivos,  $A_2$ , que se cumple si la persona seleccionada pertenece a la de obreros, y  $A_3$ , que se cumple si la persona seleccionada pertenece a la categoría de oficinistas. Estos conjuntos no tienen elementos en común y su unión cubre el conjunto de todos los empleados de la fábrica, cumpliéndose por tanto las propiedades de una partición (ó de un sistema exhaustivo y excluyente).

El diagrama de Venn correspondiente a una partición del espacio muestral en tres sucesos puede presentar el siguiente aspecto.



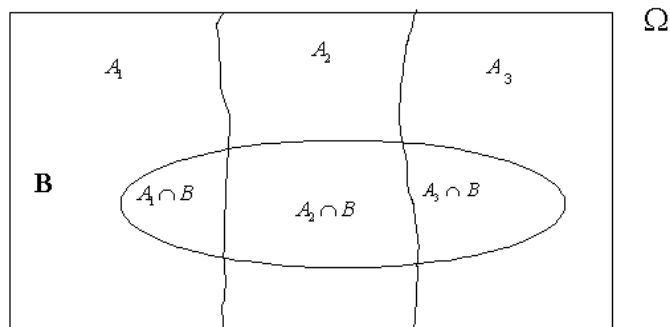


El teorema de la probabilidad total tiene el siguiente enunciado:

Sea  $A_1, A_2, \dots, A_n$  una partición del espacio muestral. Sea  $B$  un suceso. Entonces se cumple:

$$P(B) = \sum_{i=1}^n P(A_i)P(B/A_i) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_n)P(B/A_n).$$

La siguiente gráfica representa un diagrama de Venn apropiado para ilustrar el enunciado de este teorema:



El suceso  $B$  es el de *probabilidad total* (sin añadir ninguna condición). En el ejemplo, el suceso que se cumple si se selecciona una mujer de esta fábrica es de probabilidad total, en contraposición con  $P(B/A_i)$  que serían probabilidades parciales (condicionadas por algún otro suceso). Por ejemplo  $P(B/A_1) = 0.20$  indica la probabilidad de que la persona seleccionada sea una mujer si sabemos que pertenece a la categoría de directivos.

**Demostración del teorema de probabilidad total:**

$$P(B) = P(\Omega \cap B) = P((A_1 \cup A_2 \cup \dots \cup A_n) \cap B) =$$

aplicando la propiedad distributiva :

$$= P((A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_n \cap B)) =$$

De  $A_i \cap A_j = \emptyset$  se deduce que  $(A_i \cap B) \cap P(A_j \cap B) = \emptyset$ . Por el segundo axioma de la probabilidad tenemos:

$$= P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B) =$$

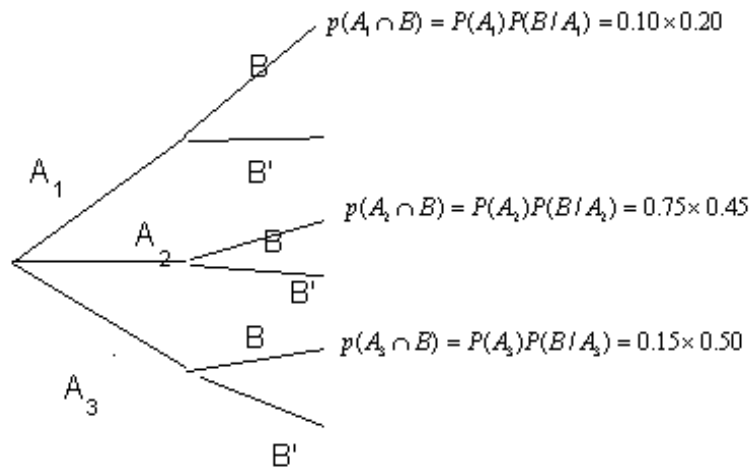
Usando ahora la fórmula 2.2 de la página 77 para la probabilidad de la intersección obtenemos:

$$= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_n)P(B/A_n) = \sum_{i=1}^n P(A_i)P(B/A_i).$$

En el problema del ejemplo obtenemos que la probabilidad de seleccionar una mujer entre el total de los empleado de la fábrica es:

$$\begin{aligned} P(B) &= P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3) = \\ &= 0.10 \times 0.20 + 0.75 \times 0.45 + 0.15 \times 0.50 = 0.4325 \end{aligned}$$

En la siguiente figura está esquematizado el diagrama de árbol del problema junto con las probabilidades que le corresponden a las ramas en las que se cumple el suceso  $B$ .



## 2.6 Teorema de Bayes.

Este teorema se usa para calcular una probabilidad condicionada de una forma indirecta. Volviendo al ejemplo enunciado en el párrafo anterior, supongamos que una vez seleccionado el trabajador de la fábrica se nos informa que es una mujer. ¿Cuál es la probabilidad de que esta mujer sea directiva? La probabilidad de que la persona seleccionada fuera un directivo, sin ninguna información adicional es  $P(A_1) = 0.10$ , que llamamos probabilidad “a priori”. Pero lo que queremos hallar es la probabilidad de que sea directivo sabiendo que se ha seleccionado una mujer. La probabilidad pedida es  $P(A_1/B)$  que llamamos probabilidad “a posteriori”, o sea que esta probabilidad viene influida por la información conocida sobre el resultado del experimento.

El teorema de Bayes se enuncia de la forma siguiente:

Sea  $A_1, A_2, \dots, A_n$  una partición de  $\Omega$ . Sea  $B$  un suceso con  $P(B) \neq 0$ . Entonces se verifica:

$$P(A_k/B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B/A_k)}{P(B)} = \frac{P(A_k)P(B/A_k)}{\sum_{i=1}^n P(A_i)P(B/A_i)}$$

**Ejemplo 16** En el ejemplo anterior, calcular la probabilidad de que un empleado sea directivo si se sabe que es mujer

La probabilidad de que un empleado sea directivo sabiendo que es mujer,  $P(A_1/B)$ , se calcularía aplicando el teorema de Bayes. En el denominador aparece la expresión de la probabilidad total,  $P(B)$ , que ya hemos calculado anteriormente. En el numerador aparecería el primer sumando de dicha probabilidad. Así tenemos que

$$P(A_1/B) = \frac{P(A_1)P(B/A_1)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)}$$

$$P(A_1/B) = \frac{0.10 \times 0.20}{0.10 \times 0.20 + 0.75 \times 0.45 + 0.15 \times 0.50} = \frac{0.02}{0.4325} = 0.04624$$

es la probabilidad de seleccionar un directivo entre las mujeres trabajadoras de esta fábrica.

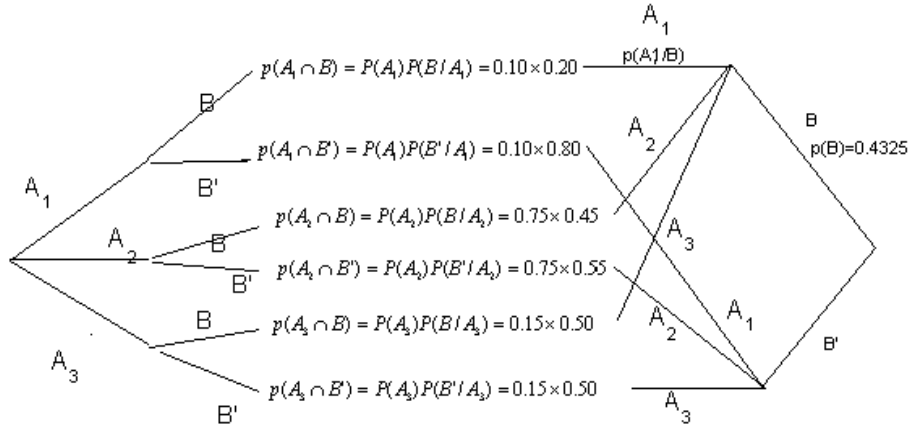
Se puede ilustrar el teorema de Bayes por medio de un diagrama de árbol. En la figura siguiente aparece representado el que corresponde al ejemplo. Las probabilidades de la intersección de los sucesos  $A_1$  y  $B$  deben ser la misma si se obtienen usando las ramas del árbol de la derecha o de la izquierda. El caso del ejemplo está ilustrado en las ramas superiores de ambos árboles:

$$\text{Rama de la izquierda: } p(A_1 \cap B) = P(A_1)P(B/A_1) = 0.10 \times 0.20$$

Rama de la derecha:  $p(A_1 \cap B) = P(B)P(A_1/B) = 0.4325 \times P(A_1/B)$

Igualando y despejando obtenemos

$$P(A_1/B) = \frac{0.10 \times 0.20}{0.4325} = 0.04624$$



Si quisiéramos hallar  $P(A_3/B')$ , probabilidad de que el empleado seleccionado sea un oficinista si sabemos que es un hombre, usaríamos las ramas inferiores de ambos árboles:

Rama de la izquierda:  $p(A_3 \cap B') = P(A_3)P(B'/A_3) = 0.15 \times 0.50$

Rama de la derecha:  $p(A_3 \cap B') = P(B')P(A_3/B') = 0.5675 \times P(A_1/B)$

Igualando y despejando obtenemos

$$P(A_3/B') = \frac{0.15 \times 0.50}{0.5675} = 0.1322$$

## 2.7 EJERCICIOS PROPUESTOS

### 2.7.1 Repaso de combinatoria

**Ejercicio 13** ¿ De cuantas formas pueden sacarse tres cartas de una baraja de 40 cartas?

1. Teniendo en cuenta el orden
2. Sin tener en cuenta el orden

**Ejercicio 14** Si sacamos tres cartas de una baraja de 40 cartas ¿ De cuantas formas pueden sacarse una pareja?

1. *Teniendo en cuenta el orden*
2. *Sin tener en cuenta el orden*

**Ejercicio 15** *Si sacamos tres cartas de una baraja de 40 cartas ¿ De cuantas formas pueden sacarse tres cartas del mismo número?*

1. *Teniendo en cuenta el orden*
2. *Sin tener en cuenta el orden*

**Ejercicio 16** *Si sacamos tres cartas de una baraja de 40 cartas ¿ De cuantas formas pueden sacarse tres cartas de distinto número?*

1. *Teniendo en cuenta el orden*
2. *Sin tener en cuenta el orden*

**Ejercicio 17** *¿Cuántos números de 3 cifras pueden formarse con los dígitos impares, 1, 3, 5, 7, 9? ¿Y con los pares 0, 2, 4, 6, 8?*

**Ejercicio 18** *¿De cuántas maneras pueden elegir 20 operarios de una fábrica una comisión formada por 3 de ellos, que los represente ante la Empresa?*

**Ejercicio 19** *¿Cuántas palabras diferentes pueden formarse con las letras de la palabra ESTADISTICA si las consonantes han de ocupar los lugares impares y las vocales los pares.*

### 2.7.2 Probabilidad

**Ejercicio 20** *Si sacamos tres cartas de una baraja de 40 cartas ¿ Cual es la probabilidad de sacar una pareja?*

**Ejercicio 21** *Si sacamos tres cartas de una baraja de 40 cartas ¿ Cual es la probabilidad de sacar un trio?*

**Ejercicio 22** *Si sacamos tres cartas de una baraja de 40 cartas ¿ Cual es la probabilidad de sacar tres cartas de distinto número?*

**Ejercicio 23** *¿ Son independientes los sucesos sacar una carta de oro y sacar un cuatro?*

**Ejercicio 24** *En una fábrica hay 8000 obreros (80% hombres), 1500 administrativos (1000 mujeres y 500 hombres ) y 500 personas que realizan labores de dirección (10% mujeres)? . Si elegimos una persona al azar,*

1. ¿ Cual es la probabilidad de que sea administrativo?
2. ¿ Cual es la probabilidad de que sea una mujer?
3. ¿ Cual es la probabilidad de que una mujer sea administrativo? ¿ Son independientes los sucesos ser administrativo y ser mujer?
4. ¿ Son independientes los sucesos ser obrero y ser mujer?
5. Cual es la probabilidad de que una mujer sea directiva? ¿ Cual es la probabilidad de que hombre sea directivo?

**Ejercicio 25** Tenemos tres urnas  $A, B, C$ .  $A$  tiene dos bolas blancas y una negra,  $B$  tiene dos bolas blancas y dos negras,  $C$  tiene 3 bolas blancas y una negra. Realizamos el experimento consistente en tirar un dado y sacar luego una bola de una urna. Si el resultado del dado es un número par elegimos la urna  $A$  y sacamos de ella una bola. Si el resultado es un 1 elegimos la urna  $B$  para sacar la bola. En los restantes casos sacamos la bola de la urna  $C$ .

1. ¿ Cual es la probabilidad de que la bola extraída sea blanca?
2. Si se sabe que la bola resulto ser blanca, ¿ Cual es la probabilidad de que proceda de la primera urna?

**Ejercicio 26** En un sistema hay instalada una alarma. La probabilidad de que se produzca un peligro es 0.1. Si se produce, la probabilidad de que la alarma funcione es 0.95. La probabilidad de que la alarma funcione sin haber peligro es 0.03. Hallar:

1. La probabilidad de que funcione la alarma
2. Probabilidad de que habiendo funcionado la alarma no haya habido peligro.
3. Probabilidad de que haya un peligro y, para colmo, la alarma no funcione.
4. Probabilidad de que no habiendo funcionado la alarma haya peligro.

**Ejercicio 27** Tengo mis películas clasificada en tres estantes,  $A$ ,  $B$ , y  $C$ . El estante  $A$  contiene 10 películas, de las cuales aún no he visto 4 de ellas; El  $B$  contiene 8 películas y no he visto 3 de ellas y el  $C$  contiene 6 películas de las cuales sólo me falta por ver una de ellas. Si selecciona al azar un estante y cojo una película.

1. ¿Cuál es la probabilidad de que no la haya visto?

2. Si estoy viendo una película por primera vez, pero no recuerdo el estante del que procede, ¿cuál es la probabilidad de que provenga del estante A?

**Ejercicio 28** Sean  $A$  y  $B$  dos sucesos tales que  $P(A) = \frac{1}{2}$ ,  $P(B) = \frac{1}{3}$  y  $P(A \cap B) = \frac{1}{7}$ . Calcular:  $P(A/B)$ ,  $P(A \cup B)$ ,  $P(A' \cap B)$  y  $P(A' \cap B')$

**Ejercicio 29** En una encuesta sobre hábitos de lectura se han sacado las siguientes conclusiones: El 55% de los ciudadanos no lee el periódico A. El 65% de los ciudadanos no lee el periódico B. El 10% de los ciudadanos no lee ni el periódico A ni el B. ¿Es esto posible?

**Ejercicio 30** En una ciudad el 56% de sus habitantes adultos son hombres, de los cuales el 34% afirma ser fumador, mientras que de las mujeres el 40% se confiesa fumadora. Eligiendo una persona al azar,

1. Calcula la probabilidad de que sea fumadora.
2. Probabilidad de que si la persona seleccionada al azar es fumadora sea varón.
3. Probabilidad que sea varón y no fumador.

**Ejercicio 31** En una universidad los estudiantes se matriculan en primer curso de ingeniería, ciencias, letras u otras especialidades. De entre ellos, acaban la carrera el 50% de ingeniería, el 70% de ciencias y el 80% de letras. El número de estos estudiantes que concluye la carrera en las otras especialidades representa el 10% del total de estudiantes matriculados en primero. Se sabe que el 20% se matriculan en ingeniería, el 15% en ciencias y el 50% en letras. Se supone que no se matriculan en más de una especialidad. Calcular:

1. Probabilidad de que un estudiante que se matricule en primero vaya a acabar la carrera y sea de ingeniería.
2. Qué porcentaje de alumnos matriculados en otras especialidades acaba la carrera?
3. Globalmente, ¿qué porcentaje de alumnos de primero se espera que vaya a acabar la carrera?
4. Probabilidad de que un estudiante que haya acabado la carrera sea de ingeniería.

**Ejercicio 32** Se ha comprobado que el 40% de las personas que toman ciertos productos farmacéuticos sufren efectos secundarios. En un grupo de 15 personas que toman estos productos, ¿Cuál es la probabilidad de que exactamente 4 de ellas sufran efectos secundarios? ¿Cuál es la probabilidad de que sufran efectos secundarios más de 12 personas?

**Ejercicio 33** Una variable aleatoria  $X$ , que puede tomar valores dentro del espacio muestral  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , tiene una función de probabilidad dada por la expresión  $P(X = i) = \frac{i}{21}$ , donde  $i \in \Omega$ .

1. Calcular el valor medio de esta variable
2. Calcular  $P(1 \leq X \leq 3)$
3. Si extraemos una muestra de tres elementos (con reemplazamiento), ¿Cuál es la probabilidad de que, al menos uno de ellos, sea mayor que 3?

**Ejercicio 34** Se sabe que la probabilidad que tiene una persona de padecer una cierta enfermedad es 0.10. Para detectar si una persona padece esa enfermedad se le realiza una prueba médica. Esta prueba no es absolutamente fiable, ya que si una persona está enferma no detecta la enfermedad en el 5% de los casos y si está sana la considera como enferma el 7% de los veces. Calcular

1. La probabilidad de que la prueba detecte la enfermedad en una persona.
  2. La probabilidad de que una persona esté sana si la prueba médica le ha detectado la enfermedad.
  3. La probabilidad de que una persona esté enferma si la prueba no se la ha detectado
1. Sean  $S$  el suceso correspondiente a que una persona esté sana y  $E$  el suceso que se verifica cuando está enferma.

**Ejercicio 35** Una compañía de seguros de automóviles clasifica sus pólizas según el riesgo de accidente de cada uno de los vehículos asegurados. El 20% de las pólizas son de alto riesgo ( $R_A$ ), el 30% de riesgo medio ( $R_M$ ) y el 50% de riesgo bajo ( $R_B$ ). Se sabe que un vehículo con una póliza de alto riesgo tiene una probabilidad 0.3 de tener accidente ( $A$ ) en el próximo año, una de riesgo medio una probabilidad de 0.1, y una de bajo riesgo una probabilidad de 0.001. Calcular

1. Calcular la probabilidad de que un vehículo asegurado, seleccionando al azar, sea de alto riesgo y tenga un accidente.
2. Calcular la probabilidad de que un vehículo asegurado, seleccionando al azar, tenga un accidente.
3. Si un vehículo determinado ha tenido un accidente, ¿Cual es la probabilidad de que tenga una póliza de alto riesgo?



**Ejercicio 36** *Durante el mes de Enero (20 días laborales), la probabilidad de que una persona pida un día de baja para asistir a una boda es 0.05. Si en una empresa hay 10 empleados. ¿Cual es la probabilidad de que alguno de ellos pida baja por dicho motivo durante ese mes?*



## Tema 3

# Distribuciones Estadísticas

### 3.1 Introducción al concepto de variable aleatoria

En el tema anterior hemos definido el concepto de probabilidad y sus propiedades. Se han asignado probabilidades a los sucesos, así que la probabilidad es una función que está definida sobre subconjuntos del espacio muestral. No obstante, estamos más acostumbrados a manejar funciones que estén definidas sobre variables numéricas. Por este motivo, el estudio de la probabilidad de los sucesos asociados a los experimentos aleatorios suele hacerse asignando un número a cada resultado del experimento. A veces el propio resultado es numérico y no es necesario realizar esta asociación, como en el caso de las puntuaciones obtenidas en un dado. En otros casos, como ocurre cuando se lanza una moneda a cara o cruz, el resultado no es numérico. Podemos decidir asignarle el valor 1 al resultado consistente en sacar cara y 0 al resultado que se obtiene cuando se saca cruz. Una *variable aleatoria* es una función que asigna un número a cada resultado posible de un experimento aleatorio. Suele designarse el nombre de la variable con una letra mayúscula, por ejemplo  $X$ . La correspondiente letra minúscula  $x$  indica un valor posible, aunque desconocido, de esta variable.

Conviene tener en cuenta que a un mismo experimento aleatorio pueden asociarse diversas variables aleatorias, dependiendo de la observación que realicemos. Así cuando se arrojan dos dados la variable aleatoria puede ser la suma de las puntuaciones, su producto, la mayor puntuación, etc. Lógicamente, en cada uno de estos casos también variará el espacio muestral asociado  $\Omega$ .

Formalmente, una *variable aleatoria unidimensional*  $X$  es una aplicación del conjunto muestral asociado al experimento en el conjunto  $\mathbb{R}$  de los números reales,  $X : \Omega \longrightarrow \mathbb{R}$ . Se debe cumplir la siguiente condición: Para cualquier valor real  $r$ , la imagen inversa por la aplicación  $X$  de cada conjunto de

números reales  $(-\infty, r)$ ,  $A_r = \{\omega \in \Omega / \omega = X^{-1}(x), x \leq r\}$ , ha de ser un suceso perteneciente al  $\sigma$ -álgebra de sucesos  $\mathcal{A}$ .

## 3.2 Variables aleatorias discretas

### 3.2.1 Función de probabilidad

Cuando la variable aleatoria sólo toma una cantidad finita o infinita numerable de valores se denota como *variable aleatoria discreta*. En este caso se puede definir la probabilidad especificando su valor para cada uno de los sucesos elementales. Por ejemplo, en el caso de un dado  $\Omega = \{1, 2, 3, 4, 5, 6\}$  y los valores de la variable aleatoria  $x$  pertenecen a  $\{1, 2, 3, 4, 5, 6\}$ . Podemos definir la probabilidad correspondiente a cada valor de esta variable aleatoria como  $P(x) = \frac{1}{6}, \forall x$ . En este caso todos los valores posibles de la variable aleatoria tienen la misma probabilidad.

Una *función de probabilidad definida sobre una variable aleatoria discreta* es una función que asigna a cada valor posible de esta variable aleatoria una probabilidad.

Si la variable aleatoria puede tomar los valores  $x_1, x_2, \dots, x_n, \dots$ , la tendremos perfectamente definida si conocemos los valores

$$p_1 = P(x_1), \quad p_2 = P(x_2), \quad \dots p_n = P(x_n), \dots$$

Para que tal función esté bien definida debe respetar las propiedades de la probabilidad:

a) Todos los valores de  $P(x)$  deben pertenecer al intervalo cerrado  $[0, 1]$

$$0 \leq P(x) \leq 1$$

b) La probabilidad de la unión de conjuntos disjuntos (finita o infinita numerable) ha de ser igual a la suma de la probabilidad de cada uno de los sucesos. Por tanto la probabilidad de la suma de los sucesos elementales ha de ser 1, ya que su unión es el espacio muestral completo. Si el espacio muestral es finito  $\Omega = \{x_1, x_2, \dots, x_n\}$ :

$$P(x_1) + P(x_2) + \dots + P(x_n) = 1$$

En el caso de que la variable tome un número infinito numerable de valores, la primera propiedad se expresaría como

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

que es una serie de números reales.

Una vez conocida la probabilidad de los sucesos elementales puede hallarse la probabilidad de cualquier otro suceso, realizando la suma (finita o infinita numerable) de las probabilidades de los elementos que contenga este suceso. Por ejemplo la probabilidad del suceso “sacar par en un dado” es:

$$P\{2, 4, 6\} = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Uno de los objetivos del Cálculo de Probabilidades es construir funciones de probabilidad que puedan servir de modelos para el comportamiento de los fenómenos aleatorios que se presentan en la naturaleza. Más adelante describimos algunas funciones de probabilidad de variable aleatoria discretas que pueden usarse en la práctica, como la distribución Uniforme Discreta, la de Bernoulli, la Binomial, la Geométrica y la de Poisson, y que se usan para modelar fenómenos aleatorios reales. Estas distribuciones pueden, en este sentido ser consideradas como *leyes de la naturaleza*, pues, como ya hemos comentado, el azar no es sinónimo de desinformación completa sino que existen leyes que rigen el azar.

### 3.2.2 Función de distribución de una variable aleatoria discreta

La función de distribución de una variable aleatoria es una abstracción del concepto de frecuencia relativa acumulada. La función de distribución,  $F(x)$ , se define:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(x_i)$$

**Ejemplo 17** *Consideremos el experimento aleatorio consistente en lanzar dos dados. La variable aleatoria que vamos a asociar a dicho experimento es la diferencia (en valor absoluto) de las puntuaciones de ambos dados. Calcular la función de probabilidad asociada a esta variable aleatoria.*

En este caso el espacio muestral  $\Omega = \{0, 1, 2, 3, 4, 5\}$ . Asignamos ahora probabilidades a los elementos de este espacio muestral. Aunque desde el punto de vista formal, esta asignación puede hacerse de diferentes formas (con tal que se cumplan las propiedades de una función de probabilidad), damos una definición que está acorde con los resultados experimentales. En la siguiente tabla se indican los resultados posibles del experimento según las puntuaciones obtenidas con los dados. Cada uno de estos 36 casos son equiprobables.

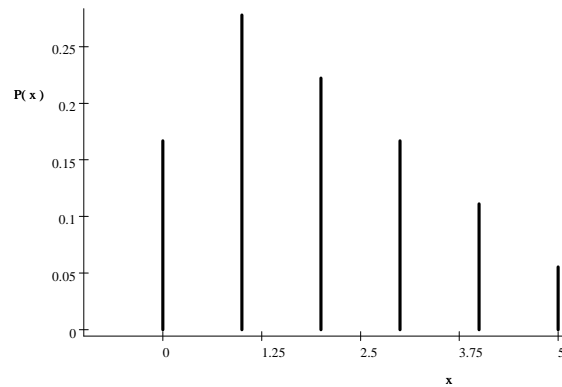
	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

La función de probabilidad viene dada en la siguiente tabla:

$x$	0	1	2	3	4	5
$P(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

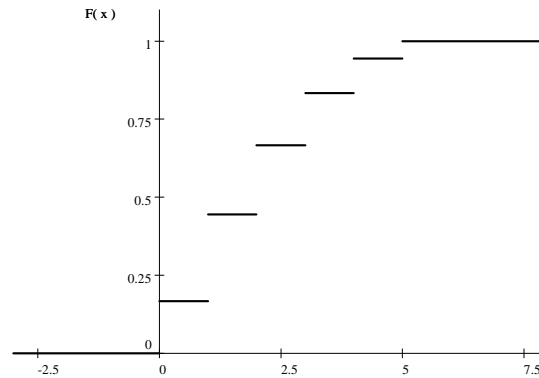
La representación gráfica de esta función de probabilidad usando un diagrama de barras es:

$x$	$P(x)$
0	0.16667
1	0.27778
2	0.22222
3	0.16667
4	0.11111
5	0.05555



Su función de distribución, que está definida para todo  $x \in \mathbb{R}$ , viene detallada a continuación.

$x < 0$	$F(x) = 0$
$0 \leq x < 1$	$F(x) = 0.16667$
$1 \leq x < 2$	$F(x) = 0.44444$
$2 \leq x < 3$	$F(x) = 0.66667$
$3 \leq x < 4$	$F(x) = 0.83333$
$4 \leq x < 5$	$F(x) = 0.94444$
$5 \leq x$	$F(x) = 1$



**Propiedades de la función de distribución:**

- 1)  $0 \leq F(x) \leq 1$
- b)  $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$
- c)  $F(x)$  es no decreciente.
- d)  $F(x)$  es continua a la derecha.

### 3.2.3 Media y varianza de una variable aleatoria discreta

El valor esperado, esperanza o media de una variable aleatoria discreta se define como

$$E(X) = \sum_{i=1}^{\infty} p_i x_i = \mu$$

donde  $p_i = P(x_i)$ . La varianza se define como

$$Var(X) = \sum_{i=1}^{\infty} p_i (x_i - \mu)^2 = \sigma^2$$

La desviación típica,  $\sigma$ , es la raíz cuadrada positiva de la varianza. Obsérvese como en estas expresiones se ha sustituido la frecuencia relativa, que se usaba en el caso de los parámetros muestrales correspondientes, por la probabilidad.

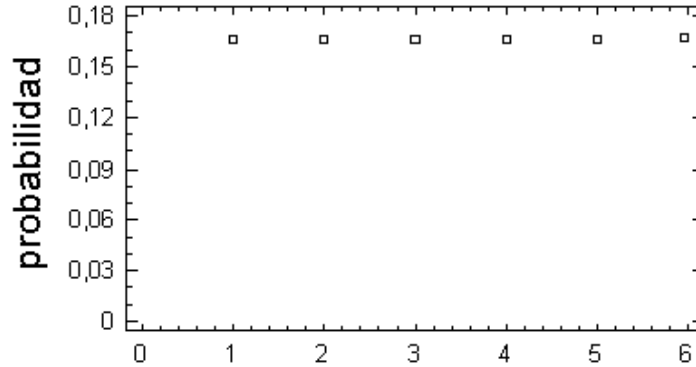
### 3.2.4 La distribución uniforme discreta

Si cuando se realiza un experimento aleatorio con sólo un número finito de resultados observamos que cada uno de estos resultados ocurre más o menos con la misma frecuencia, como ocurre cuando lanzamos un dado, podemos adaptarle un modelo de distribución uniforme discreta. En este caso  $\Omega = \{x_1, x_2, \dots, x_n\}$  y la función de probabilidad de la variable aleatoria uniforme discreta es la siguiente:

$$P(x_i) = \frac{1}{n}, \text{ para todo } i.$$

Para el caso del dado, la representación gráfica de su función de probabilidad puede ser:

funcion de probabilidad de uniforme discreta (1, 6)



### 3.2.5 La distribución de Bernoulli

Un *Proceso de Bernoulli* es una sucesión de  $N$  pruebas que satisfacen las siguientes condiciones:

- Los resultados de las  $N$  pruebas son sucesos independientes entre sí.
- Cada prueba sólo tiene dos resultados. Llamaremos a estos posibles resultados éxito y fracaso.
- La probabilidades de estos dos resultados permanecen constantes durante las  $N$  pruebas.

A cada una de estas pruebas la llamamos *experimento o prueba de Bernoulli*.

Un ejemplo de este tipo de pruebas podría ser el lanzamiento de una moneda, llamando éxito a sacar cara y fracaso a sacar cruz.

Supongamos que definimos una variable aleatoria  $X_j$ , asociada al resultado de la prueba  $j$ , que tome el valor *uno* si es un éxito y el *cero* si es un fracaso.

Entonces en virtud de la independencia de las pruebas se tiene que

$$P(x_1, x_2, \dots, x_N) = P(x_1)P(x_2)\dots P(x_N)$$



Como la probabilidad de éxito o fracaso permanece constante a lo largo de cada prueba se tiene que la función de probabilidad de una prueba de Bernoulli es,

$$P(1) = p \quad P(0) = q = 1 - p$$

donde naturalmente  $0 \leq p \leq 1$ .

Ésta es la función de probabilidad de la distribución de Bernoulli

La media de la variable aleatoria de Bernoulli se calcula como:

$$E(X) = 0 \cdot q + 1 \cdot p = p$$

y la varianza

$$\begin{aligned} \text{Var}(X) &= (0 - p)^2 q + (1 - p)^2 p = p^2(1 - p) + (1 - p)^2 p = \\ &= p - p^2 = p(1 - p) = pq \end{aligned}$$

**Ejemplo 18** *Si un proceso de fabricación produce un 2% de elementos defectuosos. y llamamos éxito el resultado de obtener un producto correcto, la probabilidad asociada a cada salida del producto sigue la distribución de Bernoulli:*

$$P(1) = 0.980 = p \quad P(0) = 0.02 = 1 - p$$

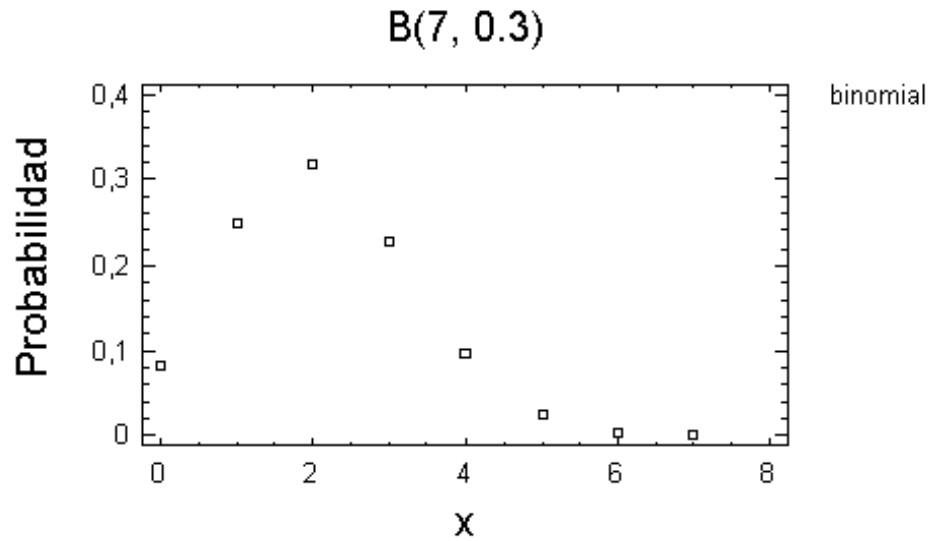
### 3.2.6 La distribución binomial

Es la distribución de la variable aleatoria  $X$  que representa el número de éxitos obtenidos cuando se realizan un total de  $n$  pruebas de Bernoulli. En este caso  $\Omega = \{0, 1, 2, 3, \dots, n\}$

La expresión para esta función de probabilidad es

$$P(x) = \binom{n}{x} p^x q^{n-x}, x \in \Omega$$

La representación gráfica de la función de probabilidad de una binomial  $n = 7, p = 0.3, B(7, 0.3)$  es la siguiente



Los primeros valores de la gráfica se corresponden con los del siguiente ejemplo.

**Ejemplo 19** *La probabilidad de curación de un cierto tipo de cáncer es de 0.3. En un grupo de siete de estos enfermos, ¿cuál es la probabilidad de que sanen 0, 1, 2, 3 de ellos?*

$$P(0) = \binom{7}{0} (0.3)^0 (0.7)^7 = 0.08235$$

$$P(1) = \binom{7}{1} (0.3) (0.7)^6 = 0.24706$$

$$P(2) = \binom{7}{2} (0.3)^2 (0.7)^5 = 0.31765$$

$$P(3) = \binom{7}{3} (0.3)^3 (0.7)^4 = 0.22689$$

Para calcular la media y la varianza de una distribución Binomial se considera que la variable aleatoria correspondiente a una  $B(n, p)$  es la suma de  $n$  variables independientes de Bernoulli y que la media de la suma de varias variables aleatorias es la suma de las respectivas medias. Idéntica propiedad tiene la varianza de una suma si estas variables son independientes

$$E(X) = p + p + \dots + p = np$$

y la varianza puede expresarse por:

$$Var(X) = pq + pq + \dots pq = npq$$

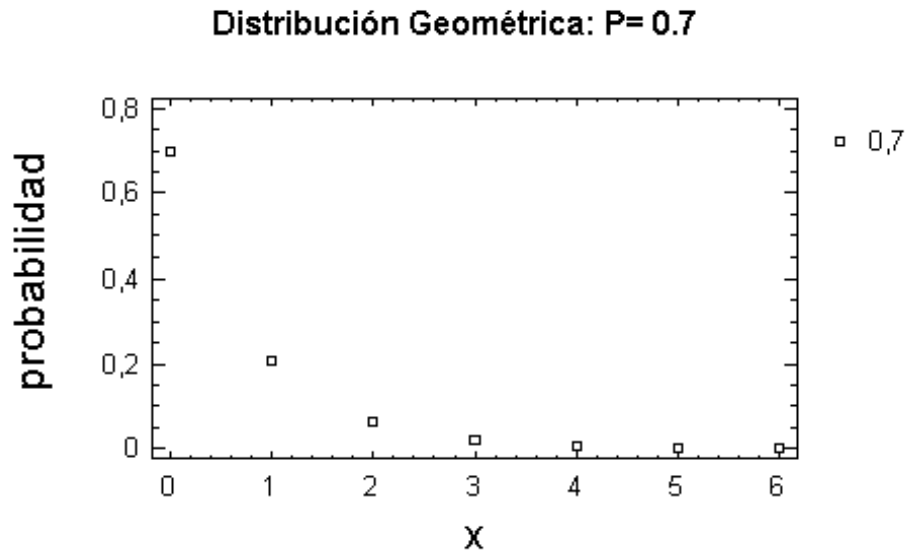
### 3.2.7 La distribución geométrica

También está relacionada con los experimentos de Bernoulli y es la distribución de la variable aleatoria  $X$  consistente en el número de pruebas que se realizan hasta obtener el primer éxito. En este caso el espacio muestral es infinito, el conjunto de todos los números naturales, ya que no hay seguridad de que el éxito se va a producir antes de un determinado valor. La función de probabilidad se calcula con la expresión:

$$p(x) = q^{x-1}p \quad \text{si } x = 1, 2, 3, \dots$$

Se entiende que la probabilidad de fracaso es  $q = 1 - p$ , siendo  $p$  la probabilidad de éxito.

Si  $p = 0.7$ , la representación gráfica es:



Las expresiones para la media y la varianza de esta distribución son:

$$E(X) = \frac{1}{p}$$

y

$$Var(X) = \frac{q}{p^2}$$

A veces se define esta distribución como la de la variable aleatoria que cuenta el número fracasos (en lugar del número total de pruebas) que ocurren antes del primer éxito, es decir que expresa la probabilidad de que se den  $y$  fracasos antes del primer éxito al repetir un experimento de Bernoulli. La variable aleatoria  $Y$ , es el número de fracasos.

$$P(y) = q^y p \quad y = 0, 1, 2, \dots$$

$$E(Y) = \frac{1-p}{p} = \frac{1}{p} - 1 \quad Var(Y) = \frac{q}{p^2}$$

*Nota:* Podemos observar que la variable  $Y$  es siempre una unidad menor que la  $X$ . Por este motivo las medias respectivas también se diferencian en una unidad. En cambio la varianza se mantiene, porque este estadístico es invariante frente a traslaciones.

**Ejemplo 20** *En un minucioso control de calidad se ha encontrado que sólo el 40% de los elementos de un proceso son aceptados. Se desea calcular la probabilidad de que haya que inspeccionar cuatro elementos para aceptar únicamente el cuarto de ellos.*

Consideramos  $x$  el número total de pruebas. Se realizan 4 pruebas. La probabilidad de tener que realizar 4 pruebas para obtener el primer éxito es:

$$P(x) = q^{x-1} p; \quad P(4) = 0.6^3 \times 0.4 = 0.0864$$

Si por el contrario, decidimos emplear como variable aleatoria el número de fracasos, son 3 en este caso, ya que hay que rechazar 3 elementos antes de aceptar el cuarto, la probabilidad se obtiene

$$P(y) = q^y p; \quad P(3) = 0.6^3 \times 0.4 = 0.0864$$

En ambos casos se obtiene la misma probabilidad, la probabilidad de que el primer elemento aceptable sea el cuarto. El resultado del experimento es el mismo, lo que cambia es la variable aleatoria.

### 3.2.8 La distribución de Poisson

Se llama también *ley de los sucesos raros*.

Comenzamos con un ejemplo que nos va a servir para aclarar la relación existente entre la distribución binomial y la de Poisson.

**Ejemplo 21** *Calcular la función de probabilidad del número de accidentes en una carretera peligrosa en una semana (1 semana = 10080 minutos), sabiendo que el número medio de accidentes por semana es 4.*

Para poder usar la distribución binomial como una aproximación se supone:

a) La probabilidad de accidentes en cada minuto  $p$  es independiente de lo que haya ocurrido en los minutos anteriores.

b) En cada minuto puede ocurrir como máximo un accidente.

Si  $p$  es la probabilidad de accidente por minuto, la probabilidad de  $x$  accidentes en el cruce en los 10080 minutos de una semana es

$$P(x) = \binom{10080}{x} p^x q^{10080-x}$$

Como la media de la binomial es  $np = 10080 \times p = 4$  y  $p = \frac{4}{10080} = 3.9683 \times 10^{-4}$

Por ejemplo si  $x = 5$

$$P(5) = \binom{10080}{5} (3.9683 \times 10^{-4})^5 (1 - 3.9683 \times 10^{-4})^{10080-5} = 0.15633$$

Para obtener la distribución de Poisson a partir de la Binomial se supone que  $n$  es muy grande y  $p$  muy pequeño, como ocurre en el ejemplo, y que se conoce el número medio de accidentes  $np = \lambda$ , que permanece constante en el intervalo de tiempo considerado.

La función de probabilidad de una distribución de Poisson se calcula como el límite de la Binomial cuando con  $n \rightarrow \infty$ ,  $np = \lambda = \text{constante}$ . Por tanto  $p = \lambda/n$  con lo cual  $p \rightarrow 0$ .

$$\begin{aligned} P(x) &= \lim_{n \rightarrow \infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} = \\ &= \lim_{n \rightarrow \infty} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \left(1 - \frac{\lambda}{n}\right)^{-x} = \\ &= \lim_{n \rightarrow \infty} \frac{\lambda^x}{x!} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = \frac{\lambda^x}{x!} e^{-\lambda} \end{aligned}$$

ya que todos los otros términos tienden a 1. La media y la varianza de esta distribución toman el mismo valor.

$$E(X) = \lambda \quad \text{Var}(X) = \lambda$$

En el ejemplo,  $\lambda = 4$  representa el promedio de accidentes en una semana.

Para calcular el grado de aproximación de ambos enfoques del problema (el primero usando  $n = 10080$ , el segundo con  $n \rightarrow \infty$ ) calculamos el mismo valor anterior por medio de la expresión de la distribución de Poisson. Para  $x = 5$

$$P(X = 5) = \frac{4^5}{5!} e^{-4} = 0.15629.$$

Como puede verse este valor es bastante parecido al que se ha obtenido usando la aproximación binomial. Por este motivo se usa en ocasiones la distribución de Poisson como una aproximación de la binomial en los casos en que  $n$  es grande y  $p$  pequeño.

Si se consideraran el número de veces que ocurre el suceso considerado en  $t$  periodos de tiempo ( $t$  semanas en el ejemplo) el promedio sería  $\lambda t$  (en el ejemplo  $\lambda t$  sería el número medio de accidentes en  $t$  semanas) y la distribución quedaría

$$P(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$$

Otras aplicaciones en las que suele usarse la distribución de Poisson son:

El número de llamadas a una central telefónica en un periodo de tiempo.

El número de imperfecciones de un tejido por unidad de superficie.

El número de bacterias en un líquido por unidad de volumen.

El número de clientes que llega a una estación de servicios en un periodo de tiempo.

El número de veces que falla una máquina por día.

El número de accidentes que se producen al día en una fábrica con un gran número de empleados.

**Ejemplo 22** *En una gran empresa el número de accidentes de trabajo sigue un promedio de tres por semana. Calcular:*

1. *La probabilidad de que no haya accidentes en una semana*
2. *La probabilidad de que haya exactamente 3 accidentes en una semana*
3. *La probabilidad de que no se superen los cuatro accidentes en una semana*
4. *La probabilidad de que haya más de 5 accidentes*

Si usamos la distribución de Poisson, obtenemos:

1.  $P(x = 0) = P(0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-3} \frac{3^0}{0!} = 4.97871 \times 10^{-2}$
2.  $P(x = 3) = P(3) = e^{-\lambda} \frac{\lambda^3}{3!} = e^{-3} \frac{3^3}{3!} = 0.224042$
3.  $P(x \leq 4) = F(4) = \sum_{x=0}^4 e^{-\lambda} \frac{\lambda^x}{x!} = \sum_{x=0}^4 e^{-3} \frac{3^x}{x!} = 0.815263$
4.  $P(x > 5) = 1 - F(5) = 1 - \sum_{x=0}^5 e^{-\lambda} \frac{\lambda^x}{x!} = 1 - \sum_{x=0}^5 e^{-3} \frac{3^x}{x!} = 8.39179 \times 10^{-2}$

### 3.2.9 La distribución hipergeométrica

Hemos advertido que la distribución binomial se aplica cuando las pruebas de Bernoulli son independientes. La distribución binomial suele aplicarse si la población en que se realizan las pruebas de Bernoulli es muy grande de modo que la probabilidad de éxito y de fracaso puede suponerse constante cualquiera que haya sido el resultado de la prueba anterior. Otra forma de conseguir probabilidad constante es que las pruebas se realicen con reemplazamiento, porque así volvemos siempre a las condiciones del primer experimento. Si no se dan estas circunstancias se suele emplear la distribución hipergeométrica. Esto es lo que ocurre en el siguiente ejemplo:

**Ejemplo 23** *En una urna hay 8 bolas blancas y tres negras. Se sacan cinco bolas. Calcular la probabilidad de que tres de ellas sean blancas.*

Puede observarse que la probabilidad es la misma cualquiera que sea el orden en que se extraigan las bolas.

$$P(BBBNN) = \frac{8}{11} \times \frac{7}{10} \times \frac{6}{9} \times \frac{3}{8} \times \frac{2}{7} = 3.6364 \times 10^{-2}$$

$$P(NBNBB) = \frac{3}{11} \times \frac{8}{10} \times \frac{2}{9} \times \frac{7}{8} \times \frac{6}{7} = 3.6364 \times 10^{-2}$$

Las ordenaciones posibles son las permutaciones con repetición de 5 elementos repitiendo tres y dos

$$\binom{5}{3} = \frac{5!}{3!2!}$$

así que la probabilidad pedida sería:

$$\binom{5}{3} \times \frac{8}{11} \times \frac{7}{10} \times \frac{6}{9} \times \frac{3}{8} \times \frac{2}{7} = 0.36364 = \binom{5}{3} \frac{V_{8,3} V_{3,2}}{V_{11,5}} = \frac{5!}{3!2!} \frac{V_{8,3} V_{3,2}}{V_{11,5}} = \frac{V_{8,3} V_{3,2}}{\frac{V_{11,5}}{5!}}$$

Esta última expresión puede ponerse como

$$\frac{\binom{8}{3} \binom{3}{2}}{\binom{11}{5}}$$

En general si se tiene un conjunto de  $N$  elementos  $K$  de una clase y el resto,  $N - K$ , de otra y se extraen  $n$  de ellos, la probabilidad de extraer  $x$  elementos de la primera clase es

$$P(X = x) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} \quad (3.1)$$

que es la expresión para la función de probabilidad de la distribución hipergeométrica.

**Ejemplo 24** De un grupo de 20 empleados, 15 hombres y 5 mujeres, se desean seleccionar 6 personas para realizar un trabajo.

- ¿Cuál es la probabilidad de que haya dos mujeres en el grupo?
- ¿Cuál es la probabilidad de que no haya ninguna mujer?
- Da una expresión de la función de probabilidad y de la función de distribución asociada a este experimento aleatorio si se toma como variable aleatoria el número de mujeres.
- Representa gráficamente esta función de probabilidad

Si llamamos  $X$  a la variable aleatoria que representa el número de mujeres seleccionadas, esta variable sigue una distribución hipergeométrica donde  $N = 20$ ,  $K = 5$ ,  $N - K = 15$ ,  $n = 6$ .

- a) La probabilidad de que haya 2 mujeres en el grupo es:

$$P(X = 2) = \frac{\binom{5}{2} \binom{15}{4}}{\binom{20}{6}} = 0.35217$$

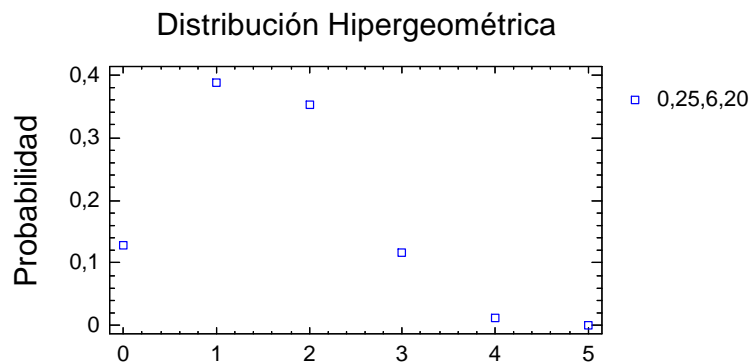
- b) La probabilidad de que no haya ninguna mujer en el grupo es:

$$P(X = 0) = \frac{\binom{5}{0} \binom{15}{6}}{\binom{20}{6}} = 0.12913$$

- c) Si  $x$  es la variable aleatoria que da el número de mujeres seleccionadas en un grupo de 6 personas, la función de probabilidad es:

$$P(x) = \frac{\binom{5}{x} \binom{15}{6-x}}{\binom{20}{6}}, \quad x = 0, 1, 2, 3, 4, 5$$

- d) La representación gráfica de esta función de probabilidad es:





Es de notar que el rango de definición de la variable aleatoria  $X$  que da la función de distribución de la hipergeométrica (expresión 3.1) ha de cumplir:

$$\max \{0, n - (N - K)\} \leq x \leq \min \{K, n\}$$

Así en el ejemplo anterior el número de mujeres que pueden ser seleccionadas cumpliría:

$$\begin{aligned} \max \{0, 6 - (20 - 5)\} &\leq x \leq \min \{5, 6\} \\ 0 &\leq x \leq 5 \end{aligned}$$

No se pueden seleccionar 6 mujeres porque sólo hay 5 empleadas en total. Si la variable aleatoria fuera el número de hombres,  $Y$ , tomaríamos  $N = 20$ ,  $K = 15$ ,  $N - K = 5$ ,  $n = 6$ .

Las restricciones serían:

$$\begin{aligned} \max \{0, n - (N - K)\} &\leq y \leq \min \{K, n\} \\ \max \{0, 6 - 5\} &\leq y \leq \min \{15, 6\} \\ 1 &\leq y \leq 6 \end{aligned}$$

Hay que seleccionar al menos un hombre, pues las 5 mujeres no pueden completar el grupo de 6 personas seleccionadas.

Si de una urna con un total 20 bolas de las cuales 10 son blancas y 10 son negras, deseamos seleccionar un total de 15 bolas, la variable aleatoria que cuenta el número de bolas blancas se rige con la distribución hipergeométrica de parámetros  $N = 20$ ,  $K = 10$ ,  $N - K = 10$ ,  $n = 15$

$$P(X = x) = \frac{\binom{10}{x} \binom{20-10}{15-x}}{\binom{20}{15}} \quad (3.2)$$

El número de bolas blancas entre las 15 ha de estar en el siguiente rango

$$\begin{aligned} \max \{0, n - (N - K)\} &\leq x \leq \min \{K, n\} \\ \max \{0, 15 - 10\} &\leq x \leq \min \{10, 15\} \\ 5 &\leq x \leq 10 \end{aligned}$$

Si se toma  $p = \frac{K}{N}$ , y  $q = 1 - p = 1 - \frac{K}{N}$ , puede comprobarse que la media y la varianza de la distribución hipergeométrica son:

$$E(X) = np, \quad Var(X) = npq \frac{N-n}{N-1} = npq \frac{1-\frac{n}{N}}{1-\frac{1}{N}}$$

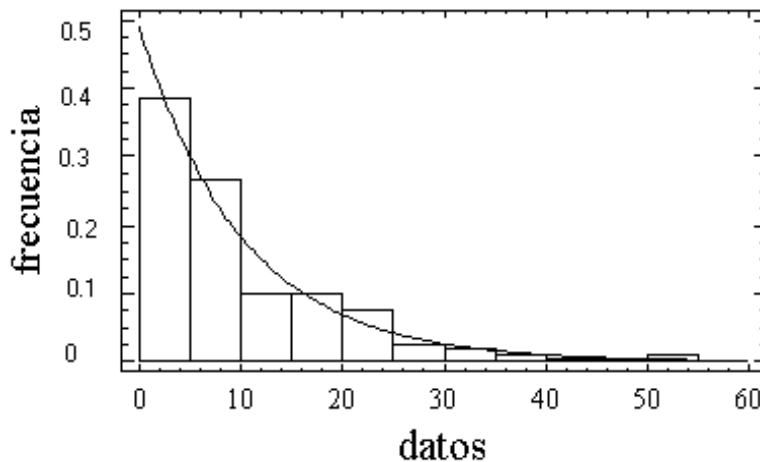
Si  $N$  es grande y la relación  $\frac{n}{N}$  pequeña, estas expresiones y las que corresponden a la media y varianza de la distribución binomial tomarían valores muy parecidos. De hecho, la distribución hipergeométrica,  $H(N, n, p = \frac{K}{N})$ , se puede aproximar con la distribución binomial,  $B(n, p)$ , si  $N > 40$  y  $\frac{n}{N} \leq 0.10$ .

### 3.3 Variables aleatorias continuas

#### 3.3.1 Función de densidad de probabilidad

Para modelar la distribución de probabilidad de una variable aleatoria continua recordamos que este tipo de variables se caracterizan por poder tomar cualquier valor dentro de un intervalo. Por ejemplo, sería una variable aleatoria continua la medida de altura de los varones españoles de 20 años, ya que, potencialmente, puede tomar cualquier valor dentro de un intervalo. Recordamos que para este tipo de variable las tablas de frecuencia se hacen subdividiendo el intervalo de variación de los datos en distintos subintervalos. La representación gráfica más adecuada para este tipo de variables es el histograma de frecuencias. En este histograma, la frecuencia relativa se representa por medio del área del rectángulo elevado sobre cada subintervalo. Si queremos respetar la idea de interpretar la probabilidad como el límite de la frecuencia relativa cuando el número de pruebas tiende a infinito parece natural definir modelos de probabilidad por medio de funciones que se asemejen a estos histogramas de frecuencia. En el siguiente histograma se ha ajustado a los datos una función del tipo exponencial :  $f(x) = \lambda e^{-\lambda x}$

**Histograma y función de densidad ajustada**



Las propiedades que debe tener esta función de ajuste para que pueda ser considerada una función de densidad de probabilidad de una variable aleatoria continua que puede tomar una cantidad infinita de valores asociados con intervalos de la recta real son las siguientes:

- 1)  $f(x) \geq 0$ , para toda  $x$

$$2) \int_{-\infty}^{\infty} f(x)dx = 1$$

La probabilidad de que la variable aleatoria pertenezca a un subintervalo dado se representa, como se hacía en el histograma, por el área que se eleva sobre este subintervalo y que está situada bajo la función de densidad:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

En consecuencia la probabilidad de un punto aislado debe definirse como cero, ya que

$$P(a) = P(a \leq x \leq a) = \int_a^a f(x)dx = 0$$

### 3.3.2 Función de distribución de una variable aleatoria continua

El concepto de función de distribución se corresponde con el de frecuencia relativa acumulada.

La función de distribución se define, tanto para una variable aleatoria discreta como continua por

$$F(x) = P(X \leq x).$$

Si la variable aleatoria es discreta

$$F(x) = P(X \leq x) = \sum_{x_j < x} P(x_j).$$

Frecuentemente la variable discreta toma valores consecutivos y enteros. En este caso la función de distribución se expresaría como

$$F(x) = P(X \leq x) = \sum_{x_j < x} P(x_j) = \sum_{i=0}^x P(i)$$

En el caso de que la variable aleatoria sea continua

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x)dx$$

es una función continua y  $F'(x) = f(x)$  en los puntos de continuidad de  $f(x)$ . La función derivada de la función de distribución de una variable aleatoria continua es su función de densidad de probabilidad. Se deduce que

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

Las propiedades características de la función de distribución de una variable aleatoria discreta dadas en la página 95 se cumplen también para el caso de que la variable aleatoria sea continua.

### 3.3.3 Media y varianza de una variable aleatoria continua

La *media o esperanza matemática* de una variable aleatoria  $X$  se denotará por  $E(X)$  o por  $\mu$  y se calculará de la siguiente manera en el caso continuo:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \mu$$

La *varianza* de una variable aleatoria  $X$  continua la denotaremos por  $Var(X)$  o bien  $\sigma^2$  y será una medida de la dispersión de los datos. Se calculará de la siguiente manera:

$$Var(X) = E((X - \mu)^2) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2$$

La *desviación típica*,  $\sigma$ , es la raíz cuadrada positiva de la varianza.

En las siguientes secciones describimos algunas funciones de densidad que se usan para modelar distribuciones de probabilidad de algunos fenómenos aleatorios que aparecen con frecuencia en la realidad, tal como la uniforme, exponencial, normal, etc.

### 3.3.4 La distribución uniforme

La variable aleatoria  $X$  sigue una distribución uniforme,  $U(a, b)$  si puede tomar valores en un intervalo  $[a, b]$  y su función de densidad de probabilidad es

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in (a, b) \\ 0 & \text{si } x \notin (a, b) \end{cases}$$

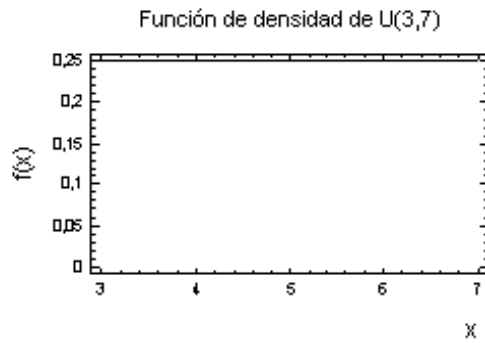
La función de distribución de la uniforme será

$$F(x) = \begin{cases} 0 & \text{si } x \leq a \\ \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a} & \text{si } x \in (a, b) \\ 1 & \text{si } x \geq b \end{cases}$$

Así la función de densidad de una  $U(3, 7)$  es:

$$f(x) = \begin{cases} \frac{1}{7-3} = 0.25 & \text{si } x \in (3, 7) \\ 0 & \text{si } x \notin (3, 7) \end{cases}$$

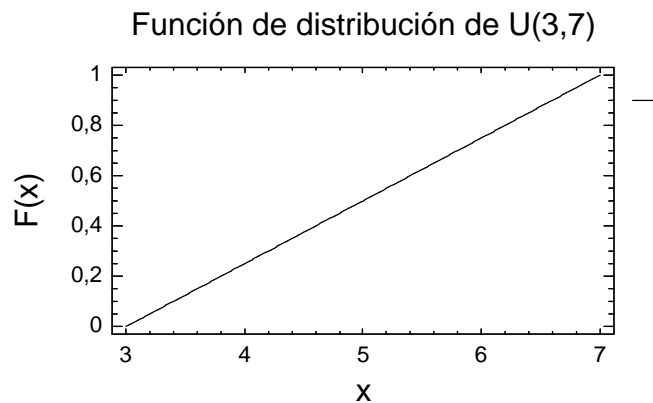
que aparece representada en la siguiente gráfica en el intervalo  $(3, 7)$



Su función de distribución es:

$$F(x) = \begin{cases} 0 & \text{si } x \leq 3 \\ \frac{x-3}{4} & \text{si } x \in (3, 7) \\ 1 & \text{si } x \geq 7 \end{cases}$$

que aparece representada en el mismo intervalo en la gráfica que sigue.



Los valores de la media y la varianza de la distribución uniforme son:

$$E(X) = \frac{b+a}{2} \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

La mayoría de los programas de ordenador tienen una función de generación de números que siguen una distribución uniforme en el intervalo  $[0, 1]$ , llamados números pseudoaleatorios, que sirven de base para generar números

aleatorios procedentes de cualquier otro tipo de distribución. Por ejemplo, para obtener números  $v$ , uniformemente distribuidos en el intervalo  $[a, b]$ , se transforman los valores,  $u$ , obtenidos de una distribución uniforme en el intervalo  $[0, 1]$  mediante la transformación  $v = a + (b - a) \times u$ .

### 3.3.5 La distribución exponencial

La variable aleatoria  $X$  sigue una distribución exponencial (a veces llamada exponencial negativa) de parámetro  $\lambda$ , si su función de densidad de probabilidad es

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

El cálculo de la media es

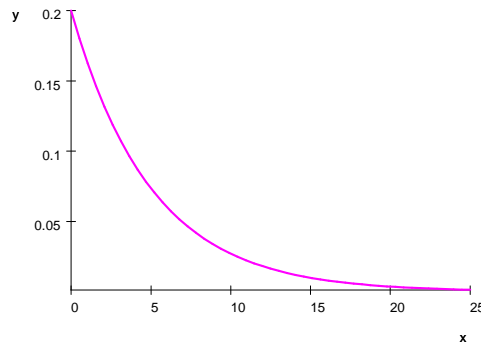
$$\mu = E(X) = \int_0^{\infty} x f(x) dx = \int_0^{\infty} x \lambda \exp(-\lambda x) dx = \frac{1}{\lambda}$$

y de la varianza

$$Var(X) = \int_0^{\infty} \left(t - \frac{1}{\lambda}\right)^2 \lambda \exp(-\lambda t) dx = \int_0^{\infty} (t)^2 \lambda \exp(-\lambda t) dx - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Ambas expresiones pueden encontrarse realizando por partes la integrales definidas correspondientes.

En la siguiente figura se muestra la representación gráfica de la función de densidad de probabilidad de una exponencial con parámetro  $\lambda = \frac{1}{5}$ .



$$\lambda = \frac{1}{5}$$

La distribución exponencial rige el comportamiento de la variable aleatoria “tiempo transcurrido entre dos acontecimientos”, cuando el número de

acontecimientos por unidad de tiempo se distribuye según una Poisson de parámetro  $\lambda$ . Se usa por tanto para modelar intervalos de tiempo transcurridos entre las llegadas consecutivas de individuos a una cola y el tiempo de duración sin fallo de ciertos dispositivos electrónicos. Está relacionada con la distribución de Poisson en los términos indicados por el siguiente teorema

**Teorema 1** *Los intervalos entre las ocurrencias de dos sucesos siguen una distribución exponencial de parámetro  $\lambda$  si y sólo si el número de veces que ocurren estos sucesos en un intervalo de tiempo  $t$  sigue una distribución de Poisson de parámetro  $\lambda t$ .*

**Ejemplo 25** *Supongamos que el tiempo de duración de una válvula está distribuido exponencialmente con un promedio de  $1/3$  de fallos por mil horas. ¿Cuál es la probabilidad de que una válvula nueva dure al menos 1000 horas? Si la válvula ya ha durado 3000 horas, ¿cuál es la probabilidad de que dure al menos 1000 horas más?*

$\lambda = \frac{1}{3}$  fallos cada mil horas, es decir que las válvulas duran por término medio 3000 horas  $= \frac{1}{\lambda}$ , así que  $\lambda = \frac{1}{3000}$  horas<sup>-1</sup>.

La probabilidad de que una válvula nueva dure al menos 1000 horas se calcula:

$$P(X \geq 1000) = \int_{1000}^{\infty} \frac{1}{3000} e^{-\frac{1}{3000}x} dx = \quad (3.3)$$

$$\frac{1}{3000} \left. e^{-\frac{1}{3000}x} \right|_{1000}^{\infty} = -e^{-\frac{1}{3000}x} \Big|_{1000}^{\infty} = e^{-\frac{1}{3}} = 0.71653$$

Tomando la unidad de medida de tiempo en “miles de horas” en lugar de en horas, puede tomarse para  $\lambda$  el valor  $\frac{1}{3}$ .

La probabilidad de que una válvula que ya ha durado 3000 horas dure al menos 1000 horas más se calcula como una probabilidad condicionada:

$$P(X \geq 4000 / X \geq 3000) = \frac{P(X \geq 4000)}{P(X \geq 3000)}$$

$$P(X \geq 3000) = -e^{-\frac{1}{3000}x} \Big|_{3000}^{\infty} = e^{-1}$$

$$P(X \geq 4000) = -e^{-\frac{1}{3000}x} \Big|_{4000}^{\infty} = e^{-\frac{4}{3}}$$

$$P(X \geq 4000 / X \geq 3000) = \frac{e^{-\frac{4}{3}}}{e^{-1}} = e^{-\frac{1}{3}}$$

### 3.3.6 La distribución normal

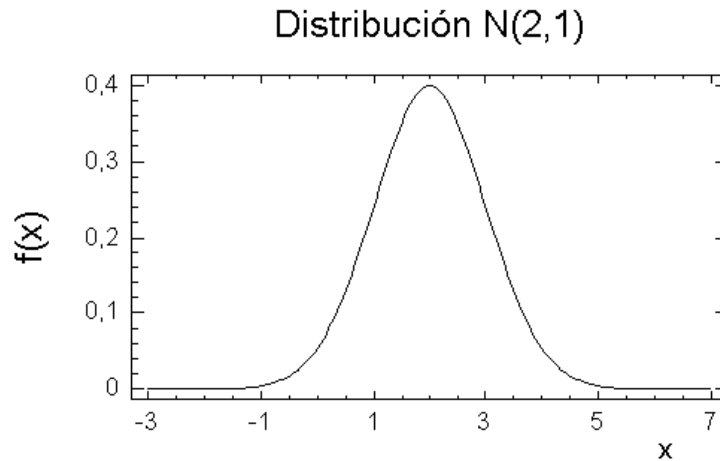
Decimos que una variable aleatoria  $X$  se distribuye según una Normal de media  $\mu$  y de desviación típica  $\sigma > 0$ , y se representará por  $X \in N(\mu, \sigma)$ , si

su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathfrak{R}$$

También es conocida como distribución de Gauss. Siguen esta distribución una gran cantidad de fenómenos aleatorios naturales. Modela errores de medidas, alturas, pesos, etc. También modela la mayoría de las medidas biológicas y las medias de muestras generadas por cualquier distribución de probabilidad, siempre que el número de elementos de las muestras sea grande.

La siguiente gráfica es una representación de la función de densidad de una distribución normal de media 2 y desviación típica 1.



#### Propiedades:

- La distribución es simétrica respecto de la recta  $x = \mu$
- La función de densidad alcanza su máximo (moda) en  $x = \mu$ . En este caso la moda coincide con la media y con la mediana.
- Si  $X \in N(\mu, \sigma)$ , entonces se cumple que

$$Y = aX + b \in N(a\mu + b, |a|\sigma)$$

- Una combinación lineal de variables normales,  $X = \sum_{i=1}^n a_i X_i$ , donde cada  $X_i \in N(\mu_i, \sigma_i)$  sigue una distribución  $N(\sum_{i=1}^n a_i \mu_i, \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2})$



- Si  $X \in N(\mu, \sigma)$ , entonces

$$Z = \frac{X - \mu}{\sigma} \in N(0, 1)$$

Esta propiedad es consecuencia de la anterior, y nos permite obtener los valores de una distribución normal cualquiera conociendo los de la  $N(0,1)$ , que es la que está tabulada.

Para calcular los valores de la función de distribución de la distribución normal  $N(\mu, \sigma)$  es preciso evaluar

$$F(x) = \int_{-\infty}^x f(x)dx = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

No es posible calcular esta expresión usando funciones elementales. Frecuentemente se recurre a realizar la transformación  $Z = \frac{X - \mu}{\sigma}$ , que se conoce como tipificación o estandarización de la variable  $X$  y que transforma cualquier normal en una  $N(0, 1)$ , que se encuentra tabulada. Aunque cada vez es más frecuente usar calculadoras o programas de ordenador para calcular los valores de la probabilidad de la normal o de cualquier otra de las distribuciones de probabilidad más usuales, aún se emplean las tablas debido a su facilidad de manejo y transporte.

Mostramos a continuación ejemplos de cálculo de valores para la probabilidades de sucesos que se generan a partir de una distribución normal. Las probabilidades se obtendrán a partir de los valores de la tabla de la página 128:

#### a) Caso de la distribución $N(0,1)$

**a1) Función de distribución en un valor positivo o nulo.** La probabilidad pedida puede deducirse directamente de la tabla.

La tabla contiene valores de la función de distribución de una variable aleatoria  $N(0,1)$  desde 0 a 4.49 en intervalos de una centésima. Por ejemplo si queremos hallar:  $F(2.34) = P(z \leq 2.34)$  usamos la tabla de la página 3.4. Como  $2.34 = 2.3 + 0.04$ , localizamos el valor que aparece en la fila que comienza en 2.3 y en la columna encabezada con 0.04. Este valor resulta ser 0.99036, por lo que concluimos que:

$$F(2.34) = P(z \leq 2.34) = P(z < 2.34) = 0.99036.$$

El último valor de la tabla es  $F(4.49) = P(z \leq 4.49) = 1.00000$ , que es un valor obtenido por redondeo, ya que su valor ha de ser algo menor que 1, pero la tabla no permite más precisión. Si el valor de  $F(z)$  que queremos hallar corresponde a un número mayor que 4.49, por ejemplo  $P(z \leq 5.00)$ ,

tomaremos, con más razón, también 1 como valor aproximado para esta probabilidad.

Si el número contiene más cifras por ejemplo :  $P(z \leq 2.347)$  pueden usarse las siguientes aproximaciones:

1. Realizar una interpolación entre los valores más cercanos:

$$P(z \leq 2.34) < P(z \leq 2.347) < P(z \leq 2.35):$$

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}, \quad \frac{y - P(z \leq 2.34)}{x - 2.34} = \frac{P(z \leq 2.35) - P(z \leq 2.34)}{2.35 - 2.34};$$

$$\frac{y - 0.99036}{2.347 - 2.34} = \frac{0.99061 - 0.99036}{0.01}$$

La solución es  $y = 0.99053$ . Este procedimiento da un valor más cercano al verdadero, que los dos siguientes, pero requiere un cálculo más largo. El valor que se obtiene usando un programa informático es

$$P(z \leq 2.347) = 0.99054$$

2. Usar como aproximación la semisuma de ambos valores:

$$P(z < 2.347) = \frac{P(z < 2.34) + P(z < 2.35)}{2} = \frac{0.99036 + 0.99062}{2} = 0.99049$$

3. Aproximar simplemente con el valor más cercano que venga en la tabla.  
En este caso

$$P(z < 2.347) \simeq P(z < 2.35) = 0.99061$$

**a2) Función de distribución en un valor negativo.** En este caso, lo transformamos de la siguiente forma usando la simetría de la  $N(0, 1)$  con respecto al eje vertical

$$F(-2.34) = P(z \leq -2.34) = P(z > 2.34) = 1 - P(z \leq 2.34) = 1 - 0.99036 = 0.00964$$

### Caso general de la distribución $N(\mu, \sigma)$

En este caso se procede a realizar la transformación de tipificación, que reduce una  $N(\mu, \sigma)$  en una  $N(0, 1)$

Por ejemplo, si es  $X$  una variable aleatoria  $N(3, 2)$ , para calcular  $P(x \leq 5)$  se sigue el siguiente proceso:

$$P(x \leq 5) = P\left(\frac{x - 3}{2} \leq \frac{5 - 3}{2}\right) = P(z \leq 1) = 0.84134$$

La distribución Normal es la más usada en Estadística. No sólo porque en la Naturaleza aparecen frecuentemente fenómenos que pueden estudiarse con esta distribución, sino también porque otras muchas distribuciones pueden aproximarse, bajo determinadas condiciones, por medio de la distribución Normal.

**Ejemplo 26** Una máquina produce ejes de acero con una longitud media de 1.005 m y una desviación típica de 0.01 m = 1 cm. Sólo son válidos los ejes que midan  $1 \pm 0.02$  m. Suponiendo que la longitud de los ejes producidos se distribuye de acuerdo con una distribución normal, ¿qué porcentaje de ejes de acero se espera que haya que desechar?

Hay que calcular  $P(0.98 < x < 1.02)$  con una  $N(1.005, 0.01)$

$$\begin{aligned} P(0.98 < x < 1.02) &= F(1.02) - F(0.98) \\ F(1.02) &= P(x \leq 1.02) = P\left(\frac{x-1.005}{0.01} \leq \frac{1.02-1.005}{0.01}\right) = P(z \leq 1.5) = \\ &0.93319 \\ F(0.98) &= P(x \leq 0.98) = P\left(\frac{x-1.005}{0.01} \leq \frac{0.98-1.005}{0.01}\right) = P(z \leq -2.5) = \\ &1 - P(z \leq 2.5) = \\ &= 1 - 0.99379 = 0.00621. \end{aligned}$$

Por lo tanto

$$P(0.98 < x < 1.02) = F(1.02) - F(0.98) = 0.93319 - 0.00621 = 0.92698$$

Se espera que el 92.7 % serán válidos y el 7.3 % serán desechables.

### 3.3.7 Distribuciones asociadas a la distribución normal

Son de uso muy frecuente en contraste de hipótesis. Sus valores se encuentran en tablas.

#### Distribución $\chi^2$ de Pearson

Sean  $Z_1, Z_2, \dots, Z_n$  variables aleatorias  $N(0, 1)$  e independientes. Entonces la suma

$$Z_1^2 + Z_2^2 + \dots + Z_n^2$$

se distribuye según una chi-cuadrado con  $n$  grados de libertad. Se representa por  $\chi_n^2$ .

Su función de densidad es de la forma:

$$f(x; n) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \text{ para } x > 0$$

La media de una  $\chi_n^2$  es  $n$  y su varianza es  $2n$ .

Al sumar dos  $\chi^2$  independientes de grados de libertad  $n_1$  y  $n_2$ , el resultado es otra  $\chi^2$  de  $n_1 + n_2$  grados de libertad.

La distribución Chi-cuadrado se usa en Inferencia Estadística para estudiar las variaciones de la varianza muestral con respecto a la varianza poblacional, la adaptación de datos experimentales a ciertas distribuciones teóricas de probabilidad, así como en el estudio de la independencia entre variables cualitativas, mediante el empleo de tablas de contingencia. Esta distribución también viene tabulada, especialmente para los valores más usuales. Se debe prestar especial atención al tipo de tabla empleada, ya que pueden dar distintos valores para la probabilidad: En unas se da la función de distribución (área a la izquierda del valor buscado) y en otras el valor complementario (cola derecha).

Las siguientes funciones, que también se encuentran tabuladas son de gran interés en Inferencia Estadística.

### Distribución F de Fisher-Snedecor

Se define como el cociente de dos  $\chi^2$  independientes divididas por sus grados de libertad.

$$F_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$

es una  $F$  con grados de libertad  $n$  y  $m$ .

La media y la varianza de esta distribución son respectivamente:

$$E(X) = \frac{m}{m-2} \quad \text{si } m > 2$$

$$Var(X) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)} \quad \text{si } m > 4.$$

### Distribución t de Student

La distribución  $t$  se define por

$$t_n = \frac{X}{\sqrt{\frac{\chi_n^2}{n}}}$$

donde  $X$  es una v.a.  $N(0,1)$  independiente del denominador.

La variable  $t$  es simétrica, con mayor dispersión que la normal estándar y tiende a ésta al aumentar el valor de  $n$  (prácticamente coinciden si  $n > 100$ ).

Para valores de  $n > 30$  se puede considerar que la normal estándar da una buena aproximación de la  $t$  de Student.

La media de la  $t$  de Student es 0, y su varianza, depende de sus grados de libertad:

$$\frac{n}{n-2}$$

Obsérvese que la media coincide con la media de una  $N(0, 1)$  y que la varianza tiende a 1 si  $n$  tiende a infinito.

### La distribución log-normal

La distribución log-normal es la distribución de una variable  $t$  cuyo logaritmo neperiano  $x = \ln t$  tiene una distribución normal. Si la media de la normal es  $\mu'$  y la desviación típica  $\sigma'$ , haciendo el cambio de variable  $x = \ln t$  en la función de distribución de la normal se obtiene que la función de densidad de la log-normal resulta ser:

$$f(t) = \frac{1}{t\sigma'\sqrt{2\pi}} e^{-\frac{(\ln t - \mu')^2}{2\sigma'^2}} \quad (3.4)$$

La media de la distribución log-normal, es decir de la variable  $t$ , es:

$$\mu = \exp\left(\mu' + \frac{\sigma'^2}{2}\right) \quad (3.5)$$

y su varianza

$$\sigma^2 = (\exp \sigma'^2 - 1) \exp(2\mu' + \sigma'^2) \quad (3.6)$$

**Ejemplo 27** Si la distribución normal asociada tiene de media  $\mu' = 4$  y la desviación típica es  $\sigma' = 2$ , entonces la media de la log-normal es:

$$\mu = \exp\left(\mu' + \frac{\sigma'^2}{2}\right) = \exp\left(4 + \frac{2^2}{2}\right) = e^6 = 403.43$$

y la varianza

$$\sigma^2 = (\exp(2^2) - 1) \exp(8 + 4) = 8.7234 \times 10^6$$

Por tanto la desviación típica de la log-normal será la raíz cuadrada de la varianza

$$\sigma = \sqrt{8.7234 \times 10^6} = 2953.5$$

Para evaluar la función de distribución de la log-normal se emplean las tablas de la normal usando el cambio de variables

$$\frac{\ln t - \mu'}{\sigma'} \quad (3.7)$$

Así si se desea calcular la probabilidad de que la variable log-normal  $t$  de media  $\mu$  y desviación típica  $\sigma$  tome valores menores o iguales que  $a$ , se procede de la siguiente forma:

$$P(t \leq a) = P(\ln t \leq \ln a) = P\left(\frac{\ln t - \mu'}{\sigma'} \leq \frac{\ln a - \mu'}{\sigma'}\right)$$

Si se conocen la media y la desviación típica,  $\mu$  y  $\sigma$ , de la log-normal, esta última expresión se evalúa teniendo en cuenta que la variable

$$z = \frac{\ln t - \mu'}{\sigma'} \in N(0, 1)$$

despejando  $\mu'$  y  $\sigma'$  de las expresiones 3.5 y 3.6.

### 3.3.8 La distribución de Weibull

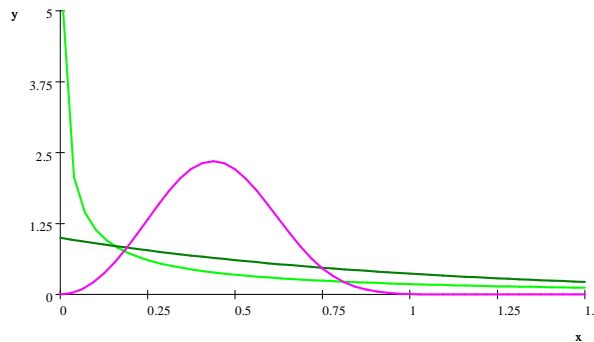
El investigador sueco Weibull propuso esta función para el estudio de fatiga de los metales (1939). Posteriormente se ha aplicado (H.J. Kao de la universidad de Cornell, 1950) al estudio del tiempo de vida de tubos electrónicos. Esta distribución se emplea mucho en fiabilidad (estudios relacionados con la duración sin fallos de los productos industriales) y para modelar la duración de la vida de los seres vivos, incluidos los humanos.

Su función de densidad es

$$f(t) = \frac{\beta}{\eta^\beta} (t - \gamma)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta}, \alpha > 0, \beta > 0 \quad (3.8)$$

Los parámetros de esta distribución  $\gamma$ ,  $\eta$ ,  $\beta$  se conocen, respectivamente, como parámetros de posición, de escala y de forma.  $\gamma$  suele ser el origen de los tiempos y frecuentemente toma el valor cero. Se observa que para  $\beta = 1$  y  $\gamma = 0$  obtenemos una distribución exponencial.

La versatilidad de la distribución de Weibull queda patente en la siguiente gráfica, donde se ha representado esta función para distintos valores de los parámetros.



Funciones de densidad de Weibull  
 $(\beta, \eta, \gamma) = (0.5, 1, 0), (1, 1, 0), (3, 0.5, 0)$

La función de distribución de la Weibul es

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad (3.9)$$

La media resulta

$$\mu = \gamma + \eta \Gamma\left(1 + \frac{1}{\beta}\right) \quad (3.10)$$

donde  $\Gamma$  indica la función *gamma completa* que sólo puede ser obtenida numéricamente (con calculadora o algún programa) o recurriendo a tablas, y que se define como:

$$\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$$

donde  $a$  debe ser real y mayor o igual que 1.

Si  $a = 1$ ,  $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$ . En el caso de que  $a$  se un número entero  $n$ ,  $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx = (n-1)!$ , como puede comprobarse realizando la integración sucesivamente por partes.

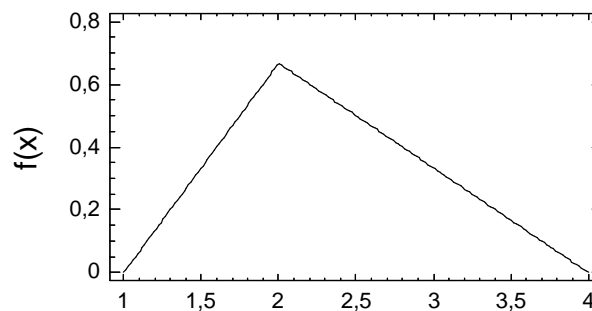
### 3.3.9 La distribución triangular

Una variable  $X$  sigue una distribución triangular de parámetros  $a, b, c$  siendo  $a < b < c$  si su función de densidad es

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{si } a \leq x \leq b \\ \frac{2(c-x)}{(c-b)(c-a)} & \text{si } b \leq x \leq c \\ 0 & \text{en el resto} \end{cases}$$

La moda, máximo de la función de densidad, se alcanza en  $x = b$ . La siguiente figura muestra la gráfica de la función de densidad de probabilidad de una distribución triangular de parámetros 1, 2, 4.

Función de densidad de una distribución triangular

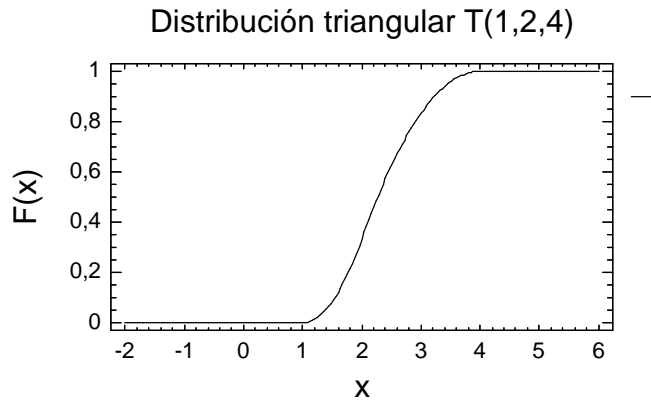


La media de la distribución es  $E(X) = \frac{a+b+c}{3}$  y la varianza  $Var(X) = \frac{1}{18}(a^2 + b^2 + c^2 - ab - ac - bc)$

La función de distribución resulta ser

$$F(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{(x-a)^2}{(b-a)(c-a)} & \text{si } a \leq x \leq b \\ 1 - \frac{(c-x)^2}{(c-b)(c-a)} & \text{si } b \leq x \leq c \\ 1 & \text{si } x \geq c \end{cases}$$

La siguiente gráfica muestra la representación gráfica de la función de distribución de la triangular de parámetros 1, 2, 4.



La distribución triangular se ha usado, junto con la distribución beta, como variable que modela el tiempo de duración de las actividades de un proyecto. Normalmente se emplea cuando se dispone de poca información, ya que es suficiente para modelarla conocer el valor mínimo, el máximo y el más probable. Tiene la ventaja de su simplicidad de cálculo. Además, puede adaptarse a distribuciones asimétricas.

### 3.3.10 La distribución gamma

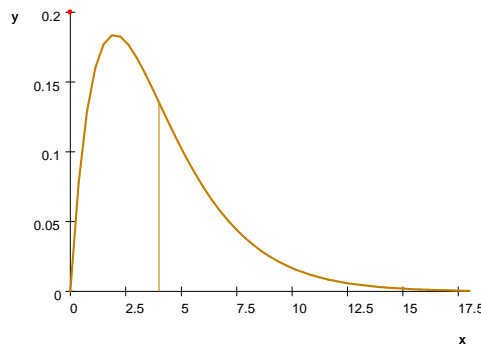
La distribución gamma tiene por función de densidad

$$f(x; a, \lambda) = \frac{\lambda}{\Gamma(a)} (\lambda x)^{a-1} e^{-\lambda x} \text{ para } x > 0, a > 0, \lambda > 0 \quad (3.11)$$



siendo  $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$ . Los parámetros  $a$  y  $\lambda$  se llaman respectivamente parámetros de forma y de escala. La *media* de esta distribución es  $\frac{a}{\lambda}$  y la *varianza* es  $\frac{a}{\lambda^2}$ . Si  $a = 1$  se obtiene la distribución exponencial. Esta distribución se aplica principalmente en problemas de Fiabilidad, siendo la variable aleatoria que modela el tiempo entre dos fallos no consecutivos.

La siguiente gráfica es la función de densidad de una gamma con  $a = 2$  y  $\lambda = \frac{1}{2}$



$$a = 2, \lambda = \frac{1}{2}$$

En el caso particular de ser  $\lambda = \frac{1}{2}$  y  $a = \frac{n}{2}$ , se obtiene la distribución  $\chi^2$  con  $n$  grados de libertad.

Si el parámetro de forma  $a$  es un entero positivo se conoce con el nombre de distribución de Erlang. En este caso la función de densidad es:

$$f(x; n, \lambda) = \frac{\lambda}{(n-1)!} (\lambda x)^{n-1} e^{-\lambda x} \text{ para } x > 0$$

La distribución de Erlang se utiliza para modelar el tiempo que transcurre desde que sucede un acontecimiento hasta que suceden los  $n$  acontecimientos siguientes, siempre que el número de veces que suceden estos acontecimientos siga una distribución de Poisson. Por ejemplo el tiempo que pasa desde que llega un cliente a una cola hasta que llegan los  $n$  clientes siguientes. Recordemos que el intervalo de tiempo entre dos llegadas consecutivas seguía una distribución exponencial. En efecto si tomamos  $n = 1$  obtenemos:

$$f(x; 1, \lambda) = \frac{\lambda}{(1-1)!} (\lambda x)^{1-1} e^{-\lambda x} = \frac{\lambda}{0!} (\lambda x)^0 e^{-\lambda x} = \lambda e^{-\lambda x}.$$

### 3.3.11 La distribución beta

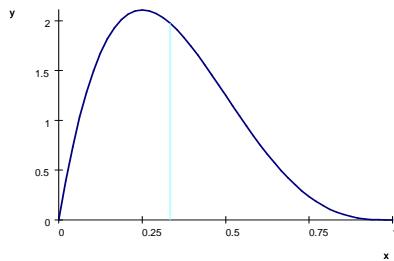
La función de densidad de la distribución beta estándar de parámetros  $\alpha$  y  $\beta$  es:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad 0 \leq x \leq 1$$

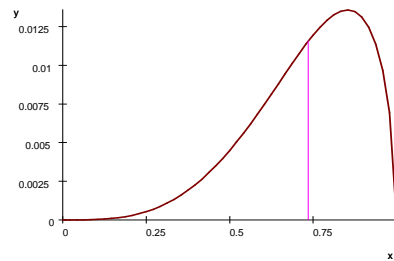
siendo  $B(\alpha, \beta)$  la función beta, que está relacionada con la función gamma como muestra la siguiente expresión:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

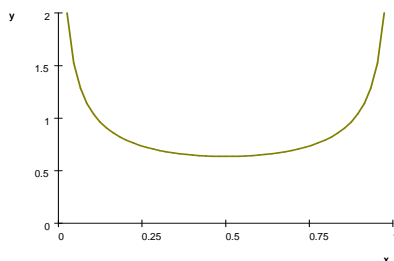
Dependiendo de los valores de  $\alpha$  y  $\beta$  la forma de la distribución puede presentar muchas variaciones. Por esto es un modelo de distribución de gran versatilidad. Podemos observar este hecho en las siguientes gráficas en las que se ha representado la función de densidad de la función beta correspondientes a algunos valores concretos de los parámetros:



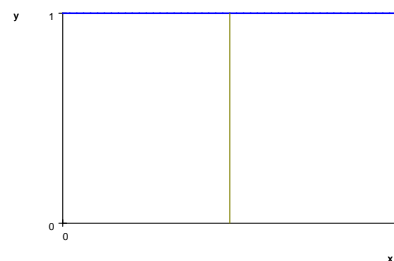
$$\alpha = 2, \beta = 4, \mu = \frac{1}{3}$$



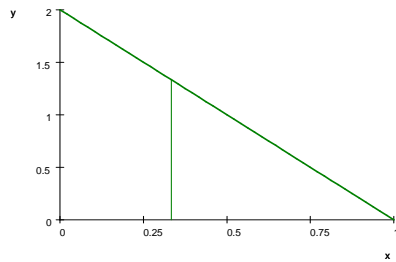
$$\alpha = 3 + \sqrt{2}, \beta = 3 - \sqrt{2}, \mu = \frac{3+\sqrt{2}}{6}$$



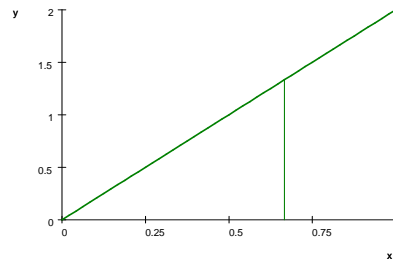
$$\alpha = \beta = \frac{1}{2}, \mu = \frac{1}{2}$$



$$\alpha = \beta = 1, \mu = \frac{1}{2}$$



$$\alpha = 1, \beta = 2, \mu = \frac{1}{3}$$



$$\alpha = 2, \beta = 1, \mu = \frac{2}{3}$$

La función de densidad de una distribución beta definida en el intervalo  $a \leq x \leq b$  viene dada por

$$f(x; \alpha, \beta, a, b) = \frac{(x-a)^{\alpha-1} (b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1} B(\alpha, \beta)} \quad \text{si } a \leq x \leq b$$

la media de la distribución es

$$\mu = a + (b-a) \frac{\alpha}{\alpha+\beta}$$

la varianza

$$\sigma^2 = \frac{(b-a)^2 \alpha \beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

y la moda

$$m = \frac{a(\beta-1)+b(\alpha-1)}{\alpha+\beta-2}$$

La relación entre la media y la moda es

$$\mu = \frac{(a+b)+m(\alpha+\beta-2)}{\alpha+\beta}$$

Este modelo es aún más versátil, ya que contiene más parámetros. Tiene diferentes aplicaciones entre las que destaca su uso en el método PERT para modelar la variable aleatoria “tiempo de ejecución de las actividades que forman un proyecto”. En este caso se toma  $\alpha = 3 + \sqrt{2}$  y  $\beta = 3 - \sqrt{2}$ .

La distribución beta se ha usado en diferentes situaciones como por ejemplo para modelar la proporción de área cubierta por una determinada planta, la proporción de impurezas en un compuesto químico, la fracción de tiempo en que una máquina está en reparación, etc.

### 3.4 EJERCICIOS PROPUESTOS

**Ejercicio 37** Dada la función:

$$f(x) = \begin{cases} cx^2 & \text{si } 0 < x < 3 \\ 0 & \text{en el resto} \end{cases}$$

1. Calcular  $c$  para que sea una función de probabilidad .
2.  $P(1 < x < 2)$
3. Calcular la función de distribución correspondiente

**Ejercicio 38** Sea  $f(x) = \frac{1}{2}e^{-\frac{x}{2}}$  para  $x > 0$ , la función de densidad de la variable que controla la duración un tipo de transistores (en cientos de horas).

1. Comprobar que es una función densidad
2. Hallar la función de distribución
3. Hallar la probabilidad de que dure entre 100 y 300 horas, usando la función de densidad y la de distribución
4. Halla la probabilidad de que uno de estos transistores dure más de 800 horas.

**Ejercicio 39** El encargado de una gasolinera, recogiendo los datos de ventas durante bastantes semanas, ha llegado a la conclusión de que la demanda semanal de gasolina (en Kl) sigue proxímadamente la distribución dada por la función de densidad

$$f(x) = \begin{cases} x & \text{si } 0 \leq x \leq 1 \\ \frac{1}{2} & \text{si } 1 < x \leq 2 \\ 0 & \text{en el resto} \end{cases}$$

- 1) Calcular la probabilidad de que en una semana se demanden entre 0.5 y 1 Kl.
- 2) Entre 0.9 y 1.1 Kl.
- 3) Más de 1500 l.
- 4) La demanda semanal esperada

**Ejercicio 40** Sea  $X$  una variable aleatoria  $N(3, 2)$  calcular:  $P(x \leq 5)$ ,  $P(x > 3)$ ,  $P(0.4 < x < 3.2)$

**Ejercicio 41** *La demanda por segundo de agua de una estación de bombeo tiene una media de  $100 \text{ m}^3/\text{seg.}$  y sigue una distribución exponencial*

1) *Calcular la probabilidad de que el agua demandada en un cierto segundo sea superior a  $200 \text{ m}^3$*

2) *Se quiere que la demanda sea atendida al menos en el 99% de los casos. ¿Cuánta agua ha de estar disponible?*

**Ejercicio 42** *Un lote de piezas contiene un 20% de defectuosas. Un cliente decide comprar el lote si tomando 100 piezas de éste elegidas al azar, como máximo 12 son defectuosas. Calcular la probabilidad de aceptar el lote.*

**Ejercicio 43** *Se supone que el número de automóviles que pasan por un cruce de carretera en 5 minutos sigue una distribución de Poisson de media 20. Calcular:*

1. *Probabilidad de que pasen menos de 2 automóviles durante 5 minutos de observación.*

2. *Probabilidad de que pasen menos de 20.*

**Ejercicio 44** *Se sabe que el número de matrimonios que se registran cada mes en una ciudad española sigue una distribución normal de media 124 y desviación típica 8. Calcular:*

1. *La probabilidad de que un cierto mes el número de matrimonios registrados esté comprendido entre 112 y 130.*

2. *La probabilidad de que la media mensual de matrimonios, obtenida computando los matrimonios registrados durante cada uno de los 12 meses del año 2005, esté entre 122 y 126.*

**Ejercicio 45** *La variable aleatoria  $X$  sigue una distribución de probabilidad cuya función de densidad es la siguiente*

$$f(x) = \begin{cases} \frac{x}{8} & \text{si } 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

*Se pide*

1. *Calcular el valor medio de  $X$*

2. *La mediana de  $X$*

3. *La función de distribución de  $X$*

4.  $P(2 \leq X \leq 3)$
5. Si extraemos una muestra de tres elementos, ¿Cuál es la probabilidad de que, al menos uno de ellos, sea mayor que 2?

**Ejercicio 46** En medicina es importante la dosis recomendada para un medicamento. Un laboratorio comercializa unos comprimidos cuyo peso sigue una distribución normal con un peso medio de 3 gramos con una desviación típica de 0.05 gramos.

1. Calcular la probabilidad de que un comprimido pese más de 3.025 gramos.
2. Un comprimido se considera defectuoso si su peso se aparta de la media en más de 0.08 gramos. Calcular la probabilidad de que un comprimido sea defectuoso.
3. Estos comprimidos se venden en cajas de 10 unidades. Si una caja contiene más de dos unidades defectuosas se retiran del mercado. ¿Qué porcentaje de cajas se retirarán del mercado?

**Ejercicio 47** El intervalo de tiempo promedio entre la llegada de dos clientes consecutivos a la caja de un supermercado es de 12 seg en un día de promoción. Con motivo de esta promoción, se cuentan los clientes que se van incorporando a la cola y se hace entrega de un pequeño obsequio a los que les correspondería un número múltiplo de 50.

1. ¿Cuál es la distribución que rige el intervalo de tiempo entre la entrega de dos obsequios consecutivos?
2. ¿Cuál es su media y su desviación típica?

**Ejercicio 48** Una caja contiene 15 tornillos de los cuales 5 son defectuosos. Calcular la distribución de probabilidad que corresponde a la variable aleatoria correspondiente al número de tornillos defectuosos obtenidos al sacar 4 tornillos de la citada caja.

**Ejercicio 49** La probabilidad de que cuando llames por teléfono a cierta oficina de información de RENFE esté comunicando es 0.40.

1. Calcular la probabilidad de poder comunicar al primer intento
2. Calcular la probabilidad de no poder hacerlo hasta el segundo intento
3. Calcular la probabilidad de no poder hacerlo hasta el tercer intento
4. Calcular la probabilidad de poder comunicar antes del quinto intento

5. Calcular la probabilidad de no poder comunicar hasta después del quinto intento.

**Ejercicio 50** El tiempo en horas hasta que se produce un fallo de un tipo de componentes electrónicos sigue una distribución de Weibull de parámetros  $\gamma = 50$ ,  $\beta = \frac{1}{3}$ ,  $\eta = 100$ . Calcular

1. La probabilidad de que uno de estos elementos falle antes de 300 horas.
2. El tiempo medio hasta el fallo de este tipo de dispositivos.

**Ejercicio 51** Se tira un dado hasta obtener tres veces el cinco.

1. Calcular la probabilidad de que esto ocurra a la séptima tirada.
2. Calcular la probabilidad de que esto ocurra antes de la séptima tirada.

**Ejercicio 52** Se ha estimado que el número de enfermos atendidos en un consultorio médico cada 10 minutos se distribuye según una ley de Poisson con una media de 3.8. Calcular la probabilidad de que en un intervalo de 10 minutos sean atendidos:

1. Ningún enfermo,
2. Un enfermo
3. Al menos dos enfermos

**Ejercicio 53** Supongamos que el ingreso mensual de un camarero es una variable aleatoria cuya función de densidad está determinada por:

$$f(x) = ke^{-\frac{x}{800}}, \quad x > 0$$

1. ¿Cuánto debe valer  $k$  para que sea una función de densidad?
2. Obtener la función de distribución.
3. Calcular la probabilidad de que el ingreso mensual exceda el ingreso promedio.
4. Determinar los ingresos medianos y el recorrido interdecil.









## Tema 4

# Simulación y Teorema Central del Límite

### 4.1 Introducción a la Simulación

La Simulación es una técnica para el análisis y estudio de sistemas complejos. Esta técnica se emplea cuando, o bien no se conocen soluciones analíticas del problema planteado, o conociendo algún modelo analítico su aplicación al estudio del problema impone demasiadas simplificaciones a la realidad, por lo que la solución obtenida se va a apartar sustancialmente de la verdadera.

La simulación pretende imitar el comportamiento del sistema real, evolucionando como éste. Lo más frecuente es estudiar la evolución del sistema en el tiempo. Para ello se formula un modelo de simulación que tiene en cuenta los elementos que vamos a considerar del modelo real y las relaciones entre éstos. Una vez determinados los objetos y las relaciones que vamos a tomar en consideración, se formula la evolución del sistema por medio de un algoritmo. Establecido el estado inicial del sistema, el algoritmo permite generar muestras simuladas de su comportamiento. Son estas muestras las que se usan para estudiar el problema tratado y dar una solución aproximada de éste. Por lo general estos algoritmos se desarrollan con programas de ordenador. Ejecutando el programa las veces deseadas se pueden obtener tantas muestras del comportamiento del sistema como queramos. Con estas muestras podremos obtener estimaciones de los parámetros en estudio muy próximas a los valores verdaderos, siempre que el modelo refleje adecuadamente el sistema real que se trata de estudiar.

Entre los muchos problemas a los que se han aplicado técnicas de Simulación citamos los siguientes.

- 1) Simulación del tráfico de vehículos en cruces de vías con mucho tráfico con el objeto de estudiar si la colocación de nuevas señales de tráfico o de de-

terminadas modificaciones en el flujo de vehículos mejorarían o empeorarían la circulación.

2) Simulación de la conducta de un modelo de inventarios. Es decir se pretende determinar la ganancia que se obtendría si los pedidos de las diferentes mercancías de un comercio se realizaran en determinada cantidad y se usaran ciertos criterios para determinar los momentos más convenientes para efectuar estos pedidos. El objetivo es realizar esta operación de la forma más conveniente para el comerciante.

3) Simulación de los movimientos sísmicos con el objeto de actuar de la mejor forma posible para paliar los efectos de estos fenómenos.

4) Simulación de las condiciones de vuelo de los aviones con el objetivo de entrenar a los futuros pilotos.

5) Simulación de las urgencias clínicas que suelen producirse en una ciudad con el objetivo de gestionar los recursos de los servicios de urgencia de manera óptima.

#### **Ventajas y desventajas de la simulación:**

Ventajas:

a) Modelos más fáciles de aplicar, por lo que se pueden acometer problemas más complejos sin imponer demasiadas simplificaciones, acercándonos más al problema real.

b) Una vez que el modelo se ha construido sirve para estudiar distintas estrategias y para determinar todos los parámetros del sistema. En un modelo analítico la teoría y el desarrollo pueden ser distintos para cada parámetro a determinar.

c) Facilidad de experimentación con el consiguiente ahorro económico. Además, las pruebas están libres de las posibles situaciones de peligro que son inherentes a algunas situaciones reales.

Desventajas:

a) Son generalmente más lentos que los cálculos analíticos.

b) Suelen ser métodos que dan soluciones aproximadas.

De todas formas no se debe establecer una competencia entre modelos analíticos y simulados. Por lo general han de complementarse mutuamente.

Desarrollamos a continuación un sencillo ejemplo que nos va servir para mostrar de una forma simple en qué consiste esta técnica de Simulación.

## **4.2 Un ejemplo muy sencillo**

**Ejemplo 28** *Consideramos el caso de una cadena de tiendas que se dedica a vender pescado por cajas. Por experiencia se sabe que la demanda es de 3 a 8 cajas diarias. Cada una de estas cajas se compra por 25 euros y se vende en 40 euros, pero las cajas que no se vendan al final del día, hay que venderlas*

en unas drásticas rebajas, a 10 euros. cada una. Si la demanda supera a la oferta suponemos que hay una pérdida de 15 euros por cada unidad que no se puede ofrecer al cliente (en concepto de pérdida de prestigio, fuga de clientes a otras tiendas, etc.). Se sabe que la demanda se puede clasificar en alta media y baja, con probabilidades 0.3, 0.45 y 0.25 respectivamente. La distribución de la demanda por categorías aparece en la tabla:

Demanda	Alta (0.3)	Media (0.45)	Baja (0.25)
3	0.05	0.10	0.15
4	0.10	0.20	0.25
5	0.25	0.30	0.35
6	0.30	0.25	0.15
7	0.20	0.10	0.05
8	0.10	0.05	0.05

Por ser un producto perecedero, el comerciante ha decidido adquirir diariamente 5 cajas. Se desea simular el comportamiento de la demanda durante 10 días calculando la ganancia media por día y determinar el número óptimo de cajas que se deben adquirir diariamente para maximizar los beneficios. ¿Cómo se puede resolver este problema por simulación?

Con el objeto de ilustrar el procedimiento vamos a hacer una simulación manual, es decir sin emplear ordenador. Para ello generamos números aleatorios. Los ordenadores tienen una función para generar estos números, pero como de momento no vamos a emplear ordenador puede emplearse una tabla de números aleatorios o una lista de premios de la lotería. También podemos recurrir a realizar un sorteo con un juego de Bingo. Necesitamos una secuencia de 20 números, diez para generar el tipo de demanda de cada uno de los diez días y otros diez para generar la cantidad demandada. Vamos a utilizar los siguientes, que se han obtenido con una tabla de números aleatorios comprendidos entre 00 y 99.

69 56 30 32 66 79 55 24 80 35 10 98 92 92 88 82 13 04 86 31

Para respetar los valores de la probabilidad indicada en la tabla anterior realizamos la siguiente asignación, haciendo corresponder a cada probabilidad una cantidad de números proporcional a ésta.

Demanda	Alta (00 a 29)	Media (30 a 74)	Baja (75 a 99)
3	0.05 (00 a 04)	0.10 (00 a 09)	0.15 (00 a 14)
4	0.10 (05 a 14)	0.20 (10 a 29)	0.25 (15 a 40)
5	0.25 (15 a 39)	0.30 (30 a 59)	0.35 (41 a 74)
6	0.30 (40 a 69)	0.25 (60 a 84)	0.15 (75 a 90)
7	0.20 (70 a 89)	0.10 (85 a 89)	0.05 (90 a 94)
8	0.10 (90 a 99)	0.05 (90 a 99)	0.05 (95 a 99)

Generamos la demanda para el primer día: usando el primer número aleatorio (69) que está entre 30 y 74, con lo que obtenemos para el día 1 una demanda media. Ahora tendremos que determinar la cantidad demandada. Para generar el número de cajas demandada en este día empleamos el segundo número (56). Mirando la columna que corresponde a la demanda media vemos que está entre 30 y 59, así que seleccionamos una demanda de 5 cajas para el primer día. La ganancia obtenida en este caso será  $40 \times 5 - 25 \times 5 = 75$  euros, ya que en este día la demanda es igual que la oferta. De forma similar se obtiene la ganancia de los días siguientes, según está indicado en la siguiente tabla.

DIA:	Compra principio del día	Sorteo tipo demanda	Sorteo demanda	Ganancia día
1	5	69 media	56 (5)	$40 \times 5 - 25 \times 5 = 75$
2	5	30 media	32 (5)	$40 \times 5 - 25 \times 5 = 75$
3	5	66 media	79 (6)	$40 \times 5 - 25 \times 5 - 15 \times 1 = 60$
4	5	55 media	24 (4)	$40 \times 4 - 25 \times 5 + 10 \times 1 = 45$
5	5	80 baja	35 (4)	$40 \times 4 - 25 \times 5 + 10 \times 1 = 45$
6	5	10 alta	98 (8)	$40 \times 5 - 25 \times 5 - 15 \times 3 = 30$
7	5	92 Baja	92 (7)	$40 \times 5 - 25 \times 5 - 15 \times 2 = 45$
8	5	88 Baja	82 (6)	$40 \times 5 - 25 \times 5 - 15 \times 1 = 60$
9	5	13 Alta	04 (3)	$40 \times 3 - 25 \times 5 + 10 \times 2 = 15$
10	5	86 baja	31 (4)	$40 \times 4 - 25 \times 5 + 10 \times 1 = 45$

Sumando la ganancia obtenida en estos diez días y dividiendo por el número de estos se obtiene la ganancia diaria media:

$$\text{Media} = 490/10 = 49 \text{ euros por día}$$

De momento hemos realizado la simulación con un pedido de 5 cajas durante 10 días. Si queremos responder a la pregunta de cuál es la cantidad de cajas por pedido que produce a la larga una ganancia máxima, podemos actuar de forma similar a como hemos hecho para el pedido de 5 cajas con todas las cantidades razonables de pedido (de 3 a 8 cajas son las demandas posibles). Es conveniente no obstante hacer simulaciones más largas, para que el valor medio de la ganancia sea más estable. Como ejemplo podemos hacer la simulación durante un año (365 días). En este caso la simulación manual, que hemos realizado anteriormente sería demasiado laboriosa. Por eso las simulaciones se realizan frecuentemente en ordenador.

El algoritmo que hay que implementar puede resumirse de la siguiente forma:

Para cada pedido (3 a 8)

Para cada día (1 a 365) se realizan los siguientes pasos:

**paso 1**

Determinar el tipo de demanda (alta media, baja)

Se genera un número aleatorio entre 0 y 1. Si este número es menor que 0.30 la demanda es alta, si está entre 0.30 y 0.75 la demanda es media. Demanda baja en otro caso.

**paso 2**

Se genera otro número aleatorio.

Generar la demanda del día seleccionando el valor correspondiente según los valores indicados en la tabla anterior.

**paso 3**

Se calcula el beneficio que corresponde a este día.

Se calcula la media de los beneficios obtenidos en los 365 días.

Repetiendo esto para todas las estrategias (pedidos de 3 a 8 cajas) se puede estimar cuál es la mejor elección.

Con un programa realizado en FORTRAN hemos estimado la ganancia media diaria en función del número de cajas pedidas, llegando a los resultados siguientes:

Cajas del pedido	3	4	5	6	7	8
Beneficio medio	10.119	36.04	53.71	58.56	51.70	39.78

Estos resultados nos permiten decidir que un pedido de 6 cajas diarias es el que reportaría mayor beneficio diario medio.

### 4.3 Método Montecarlo

Aunque las técnicas de Simulación pueden ser deterministas, es decir que se pueden simular fenómenos que no sean aleatorios, lo más frecuente, como ocurre en el ejemplo anterior, es que el fenómeno que se pretende simular tenga algún componente aleatorio. En este caso decimos que se usa el método Montecarlo. La esencia del método Montecarlo es la experimentación con números aleatorios. El procedimiento usado consiste en diseñar juegos de azar con estos números, esperando obtener de su observación conclusiones útiles para la resolución del problema que se esté estudiando. Aunque se han publicado algunos trabajos relacionados con el método de Montecarlo que no han precisado el uso de ordenadores, lo cierto es que la utilidad del método de Montecarlo se ha visto enormemente incrementada con el uso de las modernas computadoras.

Resulta difícil creer que basándose en el puro azar puedan obtenerse conclusiones que merezcan la pena y, de hecho, algunos investigadores desconfían todavía de las estimaciones que se consiguen con este método, a pesar de sus múltiples éxitos en el campo de la Investigación Operativa, de la Física y de otras ramas de las Ciencias, como la Biología, la Química, e incluso la Medicina.

Los métodos de Montecarlo suelen clasificarse en dos tipos: probabilistas y deterministas.

En el Montecarlo probabilista se simulan, con números aleatorios, fenómenos que son aleatorios en la realidad. Los números se eligen de tal forma que reproduzcan la distribución de probabilidad de la población estudiada y, de su observación, se deducen características de ésta. Por ejemplo, la Física Nuclear suministra las funciones que rigen el movimiento de los neutrones. Reproduciendo estas leyes con números aleatorios se puede simular un reactor nuclear y “experimentar” con él, evitando los problemas de dinero, tiempo y seguridad que implicaría la experimentación con un reactor nuclear verdadero.

En el Montecarlo determinista se resuelven problemas que no son aleatorios en la realidad, asociándolos con algún experimento aleatorio diseñado expresamente con este propósito. Un ejemplo de este tipo es el cálculo numérico de integrales definidas.

#### 4.4 Notas históricas sobre el Método Montecarlo

El nombre y el comienzo del desarrollo sistemático del método Montecarlo datan aproximadamente de 1944, época en la que se realizaron las investigaciones relacionadas con las primeras bombas atómicas. En tales investigaciones, llevadas a cabo principalmente en el laboratorio americano de Los Álamos, los procesos de absorción de neutrones se simulaban mediante un conjunto de ruletas adecuadamente graduadas, que originaron el nombre de “Montecarlo” con el que Von Neuman y sus colaboradores designaron a esta técnica.

Sin embargo, ya desde el siglo XVIII es posible encontrar algunos vestigios de las ideas que subyacen en el método Montecarlo. En 1777 el conde de Buffon hizo un estudio del juego siguiente, de moda por aquella época: una aguja de longitud  $L$  se arroja sobre un plano en el que hay dibujadas varias rectas paralelas con una distancia  $d$  ( $d > L$ ) entre ellas. Se gana si la aguja cae sobre alguna de las rectas paralelas. El conde de Buffon determinó la probabilidad ( $P$ ) de ganar experimentalmente (a base de tirar la aguja una gran cantidad de veces), y analíticamente, calculando para  $P$  la expresión:

$$P = 2L/\pi d$$



Años mas tarde, en 1886, Laplace sugirió que este procedimiento podría ser útil para calcular experimentalmente el valor del número  $\pi$ . Este momento es considerado, en ocasiones, como el punto de partida de las aplicaciones “serias” del método Montecarlo.

Otros trabajos pioneros sobre Montecarlo fueron los de Thompson (Lord Kelvin) en 1901, sobre la evaluación de algunas integrales de uso en la teoría de los gases. Gosset -con el seudónimo de Student- aplicó el método Montecarlo para obtener la distribución del coeficiente de correlación (1908). En 1930 Fermi empleó el método Montecarlo para sus trabajos sobre difusión y transporte de los neutrones, que resultaron esenciales para el desarrollo de las bombas y centrales nucleares.

Como ya se ha apuntado, durante la segunda guerra mundial se trabajó en estos temas. Aparte de Von Neuman, ya citado, cabe resaltar las aportaciones de Fermi, Ulam y Metrópolis. Durante esa época, la aparición de las primeras computadoras digitales dio un fuerte impulso al desarrollo del método Montecarlo. Paradójicamente, estos trabajos propiciaron a la vez un cierto descrédito del método, pues se aplicó a casi cualquier cosa, sin tener en cuenta para nada los problemas de eficiencia que le son inherentes.

En los últimos años, debido al avance experimentado en el campo de los ordenadores, a la aparición de diversas técnicas para reducir la varianza de las estimaciones obtenidas, y al muestreo de Metrópolis, el método de Montecarlo parece haber entrado en un nuevo periodo de florecimiento.

## 4.5 Generación de números aleatorios

Ya que casi siempre la simulación es aleatoria normalmente necesitamos un generador de estos números. Los ordenadores suelen tener un comando para generarlos. Nos referimos con el nombre de *números aleatorios* a muestras procedentes de una distribución uniforme en el intervalo  $[0,1]$ .

Los métodos de generación de números aleatorios pueden clasificarse en las categorías siguientes:

a) Métodos manuales: Loterías, Ruletas. Suelen ser lentos y no reproducibles. Durante bastante tiempo se creyó que era el único procedimiento para producir verdaderos números aleatorios.

b) Métodos analógicos. En este caso los números se obtienen de algún experimento físico que pueda recibirse en el ordenador. Se pueden generar rápidamente, pero no son reproducibles.

c) Tablas de números aleatorios: Es el procedimiento que hemos empleado en el ejemplo anterior. Es un procedimiento lento y presenta el inconveniente de que la tabla puede ser insuficiente para una simulación larga. La primera fue preparada por Tippett (1927). Un método que se ha usado es preparar

la tabla y almacenarla en la memoria del ordenador. En 1955 se publicó la Tabla de la Rand Corporation con un millón de dígitos. Para realizar estas tablas se usaron métodos analógicos: los datos se extrajeron del “ruido” de un generador de pulsos electrónicos.

d) Algoritmos para ordenador. Estos métodos están basados en la generación de números usando un programa de ordenador. El algoritmo usado es determinístico así que, estrictamente hablando, los números generados no serían aleatorios, aunque se comportan como si lo fueran, ya que cumplen los test de independencia y de aleatoriedad, así que se pueden usar en lugar de éstos. Los números obtenidos de esta manera se conocen con el nombre de números pseudoaleatorios.

#### 4.5.1 Propiedades de un buen generador de números aleatorios

Un generador de números aleatorios debe tener las propiedades siguientes:

a) Debe generar números aleatorios (uniformemente distribuidos e independientes).

b) Debe generarlos rápidamente.

c) No debe requerir mucho lugar de almacenamiento en el ordenador.

d) No entrar en ciclos, o al menos que los ciclos sean de periodo suficientemente largo.

f) La secuencia de números ha de ser reproducible. Es decir que se pueda repetir, si se considera conveniente, una secuencia de números que se haya producido anteriormente. De esta forma se podría repetir exactamente cualquier prueba ya realizada. En los programas de ordenador esto se consigue usando la misma semilla (número que inicializa el algoritmo).

## 4.6 Método de la transformación inversa

Los números generados por la función RANDOMIZE (o análogas) de los ordenadores siguen una distribución uniforme en  $(0,1)$ , es decir que cualquier número en este intervalo tiene la misma probabilidad de ser generado. No obstante, a veces queremos generar números cuya distribución de probabilidad no sea uniforme. Para ello hay una gran variedad de métodos. Describimos únicamente el método llamado de la transformación inversa.

Si queremos generar valores de una variable aleatoria, con función densidad  $f(x) > 0$ , se usa el hecho conocido de que si denotamos su función de Distribución por  $F(x)$ ,  $F(\xi) = \eta$  se distribuye uniformemente en el intervalo  $(0,1)$ .

Por lo tanto se pueden generar números aleatorios de una variable aleatoria  $X$  cualquiera, actuando del modo siguiente:

#### 4.7. SIMULACIÓN DE UNA COLA CON UNA LÍNEA Y UN SERVIDOR 139

- a) Se generan valores de con una distribución uniforme en el intervalo(0,1).
- b) Se calcula el nuevo número aleatorio  $\xi$ , mediante la relación  $\xi = F^{-1}(\eta)$ , donde  $\eta$  es el número generado en el paso anterior.

Este método presenta dos dificultades de tipo práctico: Calcular y resolver la ecuación  $F(\xi) = \eta$ , lo que, en muchos casos, no es tarea fácil.

##### 4.6.1 Método de la transformación inversa aplicado a la distribución exponencial

En este caso el método anteriormente descrito es de fácil aplicación. La función de densidad es  $f(x) = \lambda.exp(-\lambda x)$  :

Si  $\eta \in U[0, 1]$  es el número generado por el ordenador

$$\eta = F(x) = \int_0^x \lambda.exp(-\lambda x)dx = 1 - exp(-\lambda x)$$

Despejando  $x$  se obtiene:

$$x = -\frac{1}{\lambda}L(1 - \eta).$$

Como  $1 - \eta$  sigue la misma distribución que  $\eta$  , se puede calcular :

$$x = -\frac{1}{\lambda}L\eta$$

Si  $\eta \in U[0, 1]$ , entonces  $x$  se distribuye como una exponencial de parámetro  $\lambda$ .

#### 4.7 Simulación de una cola con una línea y un servidor

Los sistemas de colas suelen modelarse con la distribución exponencial. En este caso puede hacerse uso de la expresión anterior, lo que nos va a permitir contrastar los resultados analíticos con los obtenidos por medio de una simulación de la cola. Por este motivo incluiremos ahora un ejemplo de simulación de una cola de este tipo (modelo exponencial).

Las colas o líneas de espera son situaciones bastante corrientes: Clientes esperando servicio en un banco, alumnos que esperan matricularse, productos en una línea de producción esperando ser procesados... Los sistemas que se caracterizan por elementos que tienen que esperar para recibir un servicio se llaman fenómenos de Espera. Comenzamos estableciendo la nomenclatura empleada para referirse a las características de los fenómenos de espera.

**Terminología**

*Cola* : Elementos esperando recibir servicio

*Servidor*: Elemento que presta el servicio requerido por los elementos de la cola.

*Sistema*: Incluye cola, servidor y el elemento que está siendo servido. Pueden ser limitado o ilimitado.

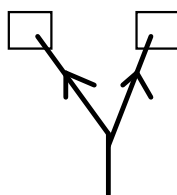
*Cadena*: Número de líneas del sistema. Los sistemas de colas son mono o multicadenas.

*Número de fases*: Es el número de servicios diferentes que hay que realizar antes de completar el servicio total

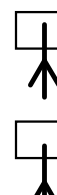
**Ejemplos:**



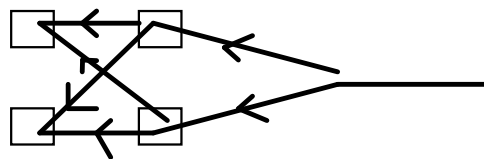
(a)



(b)



(c)



(d)

(a) Una cadena, una sola fase: Una taquilla de un cine

(b) Multi-cadena, una sola fase: Cajeros en un banco.

(c) Una cadena, multi-fase: Líneas de montaje con distintos elementos que hay que fabricar.

(d) Multi-cadena, multi-fase : Semáforos.

**Otras características que han de considerarse en las colas**

*Distribución* de las llegadas, *Distribución* de los tiempos de servicios.

*Disciplina*: Primero que llega primero que se le sirve, o que haya alguna prioridad, etc.

*Abandono*: Elemento que se va de la cola.

**Simulación de una cola simple con intervalo entre llegadas y tiempo de servicio exponencial**

A veces es posible estudiar los fenómenos de espera analíticamente, pero esto ocurre solamente en los casos más simples, así que es frecuente acometer el análisis y estudio de los sistemas de colas por Simulación. El siguiente ejemplo es lo suficientemente sencillo como para que pueda realizarse analíticamente, lo que nos permitirá contrastar los resultados obtenidos en la simulación con los que se obtienen teóricamente:

**Ejemplo 29** *Simular una cola de una sola línea y un solo servidor siendo la razón de llegada (número de personas que llegan por unidad de tiempo)  $\lambda=15$  personas por hora y la razón de servicio (número de elementos servidos en cada unidad de tiempo)  $\mu = 18$  personas por hora.*

Comenzamos generando números aleatorios. Para transformarlos en elementos de una distribución exponencial de los intervalos entre llegadas usaremos la transformación:

$$\xi_1 = -\frac{1}{15} \ln \eta_1,$$

expresando el intervalo entre llegadas en horas. Si se expresa en minutos sería:

$$\xi_1 = -\frac{1}{15} \ln \eta_1 \times 60,$$

y para los tiempos empleados en prestar servicio, también en minutos quedaría

$$\xi_2 = -\frac{1}{18} \ln \eta_2 \times 60.$$

Si la sucesión de números aleatorios que generamos para intervalos entre llegadas es:

$$1.83156 \times 10^{-2}, 4.97871 \times 10^{-2}, 0.22313, 4.08677 \times 10^{-3}$$

y para tiempos de servicios

$$6.09675 \times 10^{-3}, 4.51658 \times 10^{-3}, 0.011109, 2.47875 \times 10^{-3}, 4.97871 \times 10^{-2}$$

entonces el primer intervalo entre llegadas que se obtiene es:

$$\xi_1 = -\frac{1}{15} \ln \eta_1 \times 60 = -\frac{1}{15} \ln(1.83156 \times 10^{-2}) \times 60 = 16.0 \text{ minutos.}$$

Continuando las operaciones con los siguientes números aleatorios obtendríamos los valores de la siguiente tabla que representa los intervalos entre llegadas consecutivas y el tiempo de estancia en el servidor de los primeros elementos que van llegando al sistema

tiempo entre llegadas	Comienza la simulación	16	12	6	22
tiempo de servicio	17	18	15	20	10

Para simular el sistema podemos seguir el cambio de las variables definidas en la tabla siguiente. Cada renglón puede representar el estado del sistema. La primera llegada se produce con el reloj a 0 (comienzo de la simulación).

suceso	tipo de suceso	Tiempo de reloj	libre=1 ocupado=0	ncola = long de la cola	tiproll = hora próx llegada	tiproser = hora próx partida
0	Inicio	0	1	0	0	9999
1	llegada	0	0	0	16	17
2	llegada	16	0	1	28	17
3	partida	17	0	0	28	35
4	llegada	28	0	1	34	35
5	llegada	34	0	2	56	35
6	partida	35	0	1	56	50
⋮	⋮	⋮	⋮	⋮	⋮	⋮

El algoritmo correspondiente puede seguir el diagrama de flujo de la figura 4.1 (las variables tienen el significado descrito en la primera línea de la tabla de la página 144).

4.7. SIMULACIÓN DE UNA COLA CON UNA LÍNEA Y UN SERVIDOR<sup>143</sup>

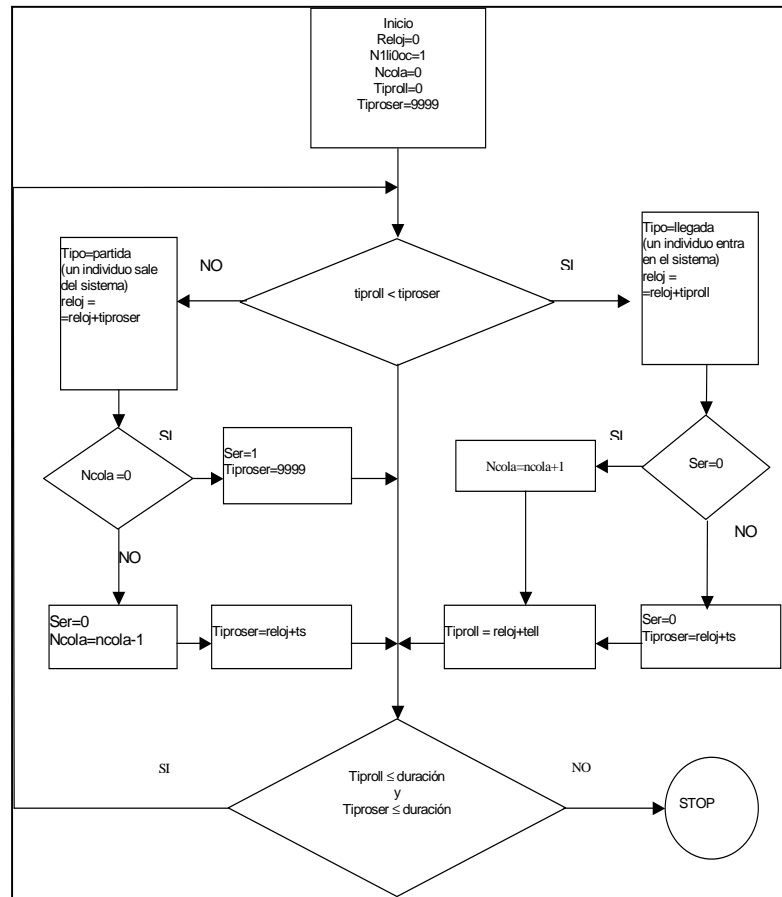


Figura 4.1: Diagrama de flujo del programa de simulación de una cola M/M/1.

tipo de suceso	Tiempo de reloj	long de la cola	libre=1 ocupado=0	hora próx llegada	hora próx partida
TIPO	RELOJ	NCOLA	N1LI0OC	TIPR0LL	TIPR0SER
LLEGADA	7.75	7	0	7.77	7.76
PARTIDA	7.76	6	0	7.77	7.81
LLEGADA	7.77	7	0	7.84	7.81
PARTIDA	7.81	6	0	7.84	7.90
LLEGADA	7.84	7	0	7.90	7.90
LLEGADA	7.90	8	0	7.90	7.90
LLEGADA	7.90	9	0	7.91	7.90
PARTIDA	7.90	8	0	7.91	7.91
LLEGADA	7.91	9	0	8.00	7.91
PARTIDA	7.91	8	0	8.00	7.93
PARTIDA	7.93	7	0	8.00	7.95
PARTIDA	7.95	6	0	8.00	8.04

Esta última tabla es la salida, obtenida con un programa de ordenador, siguiendo el esquema del diagrama 4.1. Presenta la evolución de la cola en los últimos momentos de una simulación de 8 horas.

Los parámetros obtenidos con el citado programa resultaron:

RESUMEN DE LA SIMULACION (8 horas):

ACTIVIDAD DEL SERVIDOR (fracción): 0.73

LONGITUD MEDIA DE LA COLA: 2.10

NUMERO MEDIO DE PERSONAS EN EL SISTEMA: 2.84

TIEMPO MEDIO DE ESPERA EN LA COLA: 0.14

PROBABILIDAD DE TENER QUE ESPERAR: 0.78

---

RESUMEN DE LA SIMULACION (de 10000 horas):

ACTIVIDAD DEL SERVIDOR (fracción): 0.84

LONGITUD MEDIA DE LA COLA: 4.23

NUMERO MEDIO DE PERSONAS EN EL SISTEMA: 5.07

TIEMPO MEDIO DE ESPERA EN LA COLA: 0.28

PROBABILIDAD DE TENER QUE ESPERAR: 0.84

La tabla siguiente presenta un resumen de la simulación a 8 horas y a 10000 horas. Se comparan con los resultados analíticos, observándose la similitud de estos valores con los obtenidos en la simulación más larga. A veces las simulaciones cortas, como en este ejemplo, no son suficientes para obtener una precisión aceptable, debido, no sólo a la escasez de datos, sino también a que no se ha llegado a obtener aún el régimen estacionario.



Tiempo simulado	Actividad del Servidor	Longitud media de la cola	Número medio de personas en el sistema	Tiempo medio de espera en la cola
8 horas	0.73	2.10	2.84	0.14
10000 horas	0.84	4.23	5.07	0.28
Resultados analíticos	$\frac{\lambda}{\mu} = 0.833$	$\frac{\lambda^2}{\mu(\mu-\lambda)} = 4.16$ .	$\frac{\lambda}{\mu-\lambda} = 5$	$\frac{\lambda}{\mu(\mu-\lambda)} = 0.277$

## 4.8 Integración Montecarlo

Consideremos el problema de calcular una integral unidimensional:

$$I = \int_a^b g(x) dx$$

donde supondremos que el integrando,  $g(x)$ , es una función acotada:

$$0 \leq g(x) \leq c, \quad x \in [a, b]$$

Sea  $\Omega$  el rectángulo de la figura ??:

$$\Omega = \{(x, y) \in \mathbb{R}^2, x \in [a, b], y \in [0, c]\} = [a, b] \times [0, c]$$

y sea  $(X, Y)$  una variable aleatoria uniformemente distribuida sobre  $\Omega$  con función de densidad:

$$f_{XY} = \begin{cases} \frac{1}{c(b-a)} & \text{si } (x, y) \in \Omega \\ 0 & \text{en otro caso} \end{cases}$$

¿Cuál es la probabilidad  $P$  de que el vector aleatorio  $(x, y)$  caiga en el área situada por debajo de la curva  $g(x)$  ?

Denotemos por  $S = \{(x, y) / y < g(x)\}$ . Se observa que el área bajo  $g(x)$  es igual al área de  $S$ , que a su vez coincide con el valor de la integral

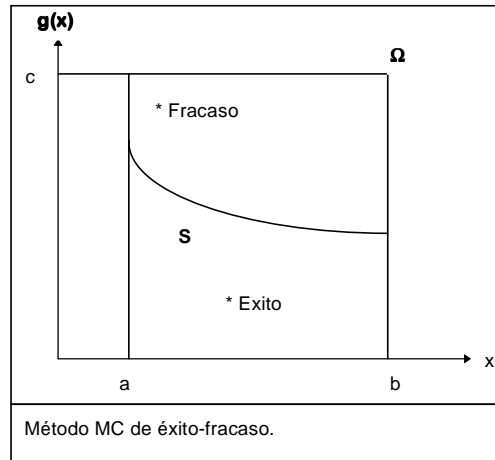
$$I = \int_a^b g(x) dx.$$

Con ayuda de la figura, se puede deducir que:

$$P = \frac{\text{area } S}{\text{area } \Omega} = \frac{\int_a^b g(x) dx}{c(b-a)} = \frac{I}{c(b-a)}$$

Por tanto

$$I = c(b-a)P$$



Para estimar el valor de  $P$  generamos  $N$  puntos aleatorios independientes:

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

dentro del rectángulo  $\Omega$ .

La probabilidad  $P$  puede ser estimada por:

$$\hat{P} = \frac{N_h}{N}$$

siendo  $N_h$  el número de estos puntos que se verifica  $g(x_i) > y_i$  (es decir, que caen dentro del área que quiere calcularse). Por lo tanto un valor aproximado de  $I$  puede obtenerse de la forma siguiente

$$I = c(b - a)\hat{P}$$

Si el extremo del vector aleatorio “cae” en  $S$  se interpreta como un éxito, y si no pertenece a  $S$  como un fracaso, de ahí el nombre del método.

Este método que presentamos aquí por su simplicidad, no suele ser muy eficiente, existiendo diversos procedimientos alternativos que permiten reducir los errores en las estimaciones de la integral. De todas formas conviene tener en cuenta que el método Montecarlo no es el más indicado para el cálculo de integrales unidimensionales, siendo sin embargo de los más eficientes para el cálculo de integrales multidimensionales si la dimension del integrando es elevada.

Describimos algunos programas de simulación que han sido realizados por alumnos de la Universidad de Cádiz, y que pueden servir para sugerir la realización de otros similares y para mostrar que estas simulaciones son susceptibles de ser realizadas dentro del ámbito académico. Todos ellos pueden cargarse en

<http://www.rosaweb.org/descargas/simul/simu.htm>

### Ejemplo 1

#### PROPÓSITO:

Aplicar el método de integración Montecarlo, éxito-fracaso, de forma práctica.

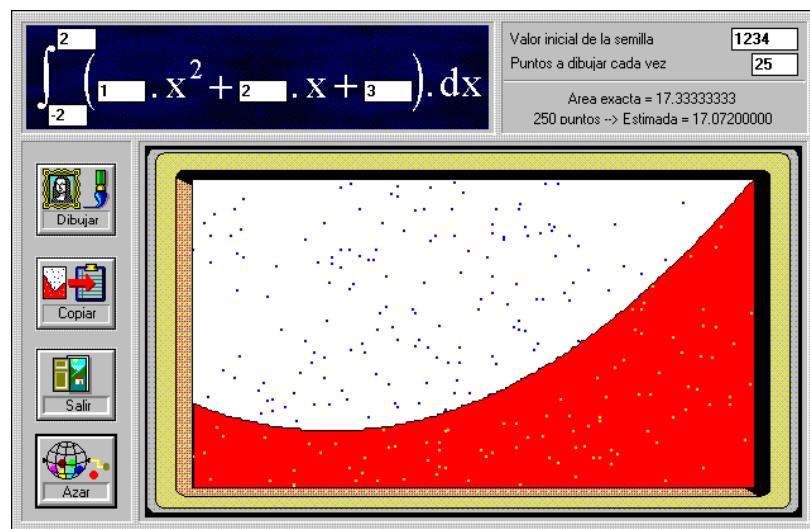
Se limita a funciones polinómicas de 2º grado.

#### UN EJEMPLO DE APLICACIÓN

En la salida siguiente del programa se puede observar que el programa obtiene como valor aproximado de

$$\int_{-2}^2 (x^2 + 2x + 3) dx$$

usando 250 puntos 17.072. Obsérvese que el valor exacto es 17.333...



#### OPERATORIA

Se sigue el orden siguiente :

- (1) Introducir los límites de integración a y b.
- (2) Indicar los coeficientes de la función polinómica de 2º grado.
- (3) Introducir el valor inicial de la semilla, para la generación de números aleatorios.

- (4) Establecer el número de puntos a generar en cada ejecución.

Pulsando el botón DIBUJAR se obtiene la gráfica de la función en el intervalo indicado y desde los valores máximo y mínimo que en él alcanza.

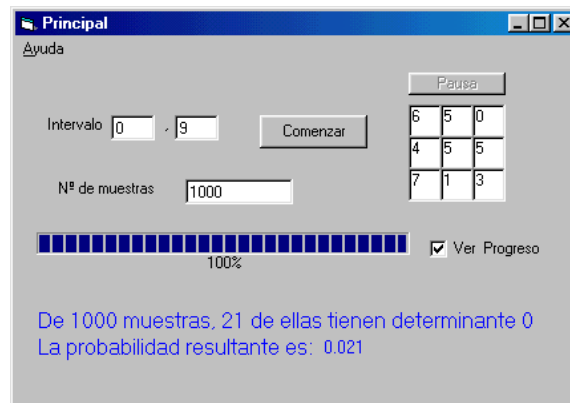
Cada vez que pulse AZAR, se representa el número de puntos indicado en (4), calculando el valor estimado del área. Puede COPIAR el gráfico en

el portapapeles Windows, para incorporar en documentos (razón por la que se habilita el acceso a Paintbrush desde el menú de opciones).

### Ejemplo 2

PROPÓSITO: Estimar la probabilidad de que el determinante de una matriz de orden 3 por 3, cuyos elementos son números naturales pertenecientes a un cierto intervalo sea nulo.

El programa muestra los determinantes generados según se muestra en la siguiente figura.

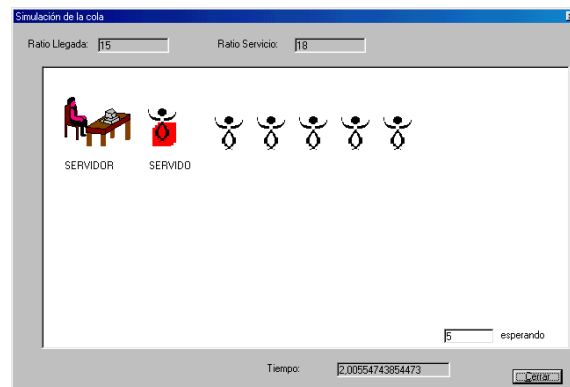


### Ejemplo 3

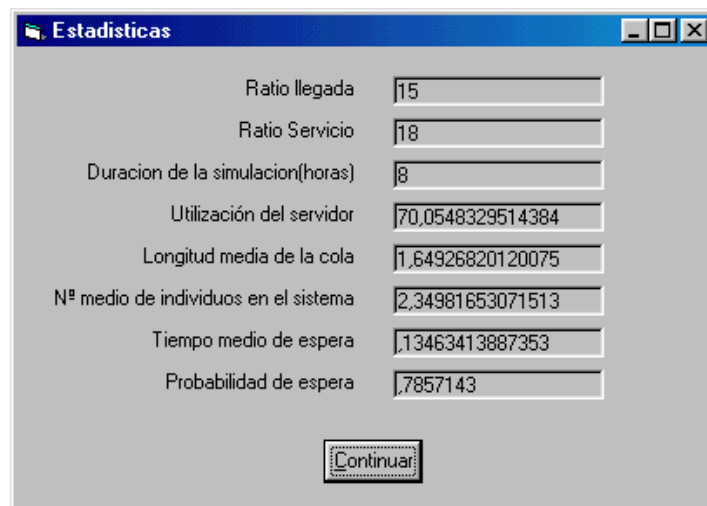
Programa realizado por los alumnos José Miguel Pérez y Manuel Jesús Pecci, de tercer curso de Ingeniería Técnica en Informática de Gestión. Curso 1999.

PROPÓSITO: Calcular los parámetros de una cola con una sola línea de espera y un solo servidor (distribuciones exponenciales).

El programa presenta una evolución visual de la cola, como se ve en la siguiente figura.



A continuación se presenta el resumen de los parámetros del sistema.



The screenshot shows a window titled "Estadísticas" with the following data:

Parámetro	Valor
Ratio llegada	15
Ratio Servicio	18
Duración de la simulación(horas)	8
Utilización del servidor	70,0548329514384
Longitud media de la cola	1,64926820120075
Nº medio de individuos en el sistema	2,34981653071513
Tiempo medio de espera	,13463413887353
Probabilidad de espera	,7857143

Below the table is a button labeled "Continuar".

#### Ejemplo 4

PROPÓSITO: Estudio detallado de diversos sistemas de colas, con intervalos entre llegadas y tiempos de servicios generados por distintas distribuciones.

FUNCIONALIDAD DEL PROGRAMA:

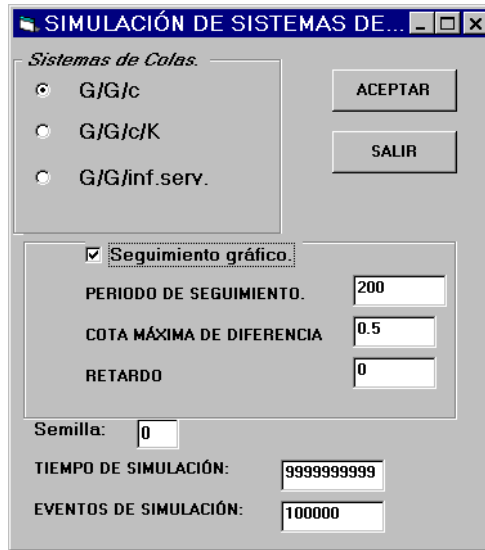
1) Tipos de problemas planteados:

Sistemas con  $c$ - canales de servicio

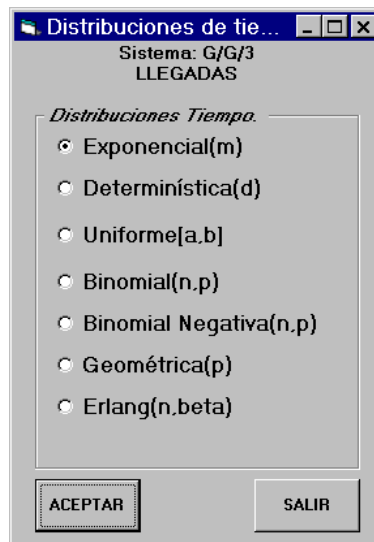
Sistemas con  $c$  canales de servicio y capacidad restringida

Sistemas con infinitos canales de servicios.

La selección se realiza en el siguiente menú:



2) Distribuciones de llegada y de servicio empleadas. Se puede seleccionar cualquiera de las siguientes:



3) Gráfico de la evolución de las medidas de efectividad para verificar la estacionariedad del sistema:

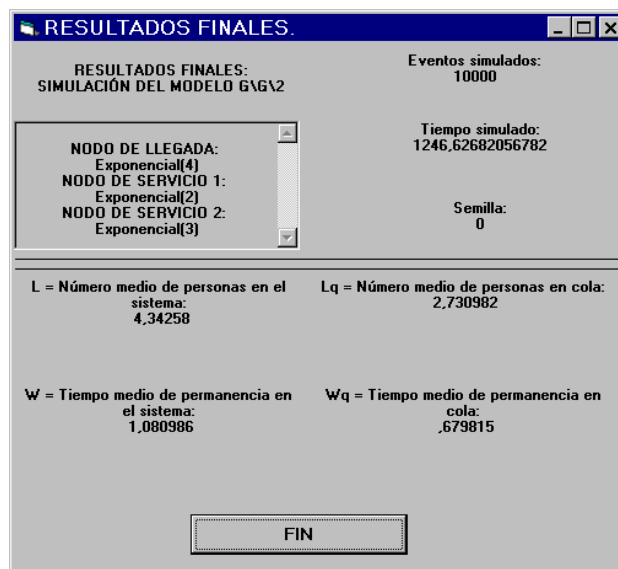
Si se ha elegido el seguimiento gráfico cada vez que se rellene la gráfica, habrá que decidir si se ha alcanzado el estado estacionario del sistema. Si

no se ha alcanzado pulsar “SEGUIR”. Si se hubiera alcanzado se pulsa “ESTADO ESTACIONARIO” y a continuación “PÚLSAME”, que iniciará la simulación del sistema, hasta obtener los parámetros de la cola, respetando el estado en que se encuentre en el momento de pulsar.

El gráfico representa el valor absoluto de las diferencias de valores de L (número de elementos en el sistema) en los dos últimos periodos. Las barras de cuadros representan diferencias mayores que las permitidas. El criterio empleado para caracterizar el estado estacionario es que las diferencias máximas entre los cuatro valores de L que se presentan sean menores que un número prefijado de antemano.

4) Cálculo de las medidas de efectividad:

Como resultado final aparecen los parámetros del sistema con una presentación como la siguiente:



#### APLICACIÓN A CASOS PRÁCTICOS

Caso 1: En un aeropuerto con tres pistas de aterrizaje se ha contrastado que los tiempos entre las llegadas de los aviones se distribuyen según una ley exponencial ( $\lambda = 0,04$ ) y que los tiempos de maniobra en el aterrizaje se distribuyen como una uniforme (14 min., 20 min) en la pista 1, exponencial ( $\lambda = 0.02$  min) en la pista 2 y una Erlang (15, 6.5) en la pista 3. Si se produce una llegada y las pistas están ocupadas, el avión se mantendrá en vuelo en espera de aterrizar.

Mostramos a continuación los resultados obtenidos por el programa para este problema usando simulaciones cada vez más largas. Puede observarse

que los parámetros de la cola son bastante estables.

No. de eventos	100000	200000	500000	800000
Tiempo simulado	1239722	2486862	6245916	9982817
Media de aviones en el sistema	1.4152	1.4127	1.4040	1.4044
Media de aviones en espera	0.0931	0.0917	0.0888	0.0889
Tiempo medio en el sistema	35.0896	35.1320	35.0767	35.0514
Tiempo medio de espera	2.3084	2.2800	2.2184	2.2195

Caso 2. En este caso se ha realizado una simulación con llegadas y tiempo de servicios exponenciales, para poder contrastar las salidas del programa con los resultados obtenidos analíticamente. También aquí se constata la precisión de los resultados obtenidos.

	Nodo de llegada exp. (1)			
G/G/3/7	Nodos de servicio (1,2,3) exp. (0.1666)			
Eventos	200000	400000	800000	Teórico
L	6.0711	6.0693	6.0615	6.0631
L <sub>q</sub>	3.0999	3.0982	3.0910	3.0920
W	12.3350	12.3175	12.2391	12.2444
W <sub>q</sub>	6.2981	6.2877	6.2412	6.2442

## 4.9 El Teorema Central del Límite

Las técnicas de Simulación tienen una gran cantidad de aplicaciones. En realidad, pueden aplicarse casi a cualquier cosa. Por supuesto la enseñanza de la estadística también queda dentro de su campo de aplicación. Mostramos como ejemplo el uso de la simulación en la visualización del teorema central del límite. Este teorema es de difícil demostración y, además, no es raro que los alumnos tengan dificultades para comprender las ideas contenidas en su enunciado. La simulación por ordenador de varios ejemplos puede ayudar a aclarar el sentido del enunciado de este teorema y a comprobar de una forma experimental su veracidad. A continuación presentamos el enunciado del citado teorema y alguna de sus aplicaciones dentro de la Estadística, dejando para el final las líneas maestras con las que se puede diseñar un programa que simule ejemplos para ilustrarlo.

Este teorema ocupa un papel *central* en Estadística. Lo expresamos en primer lugar de una manera informal: *Si una variable aleatoria  $X$  puede expresarse como suma de  $X_1, X_2, \dots, X_n$  variables aleatorias independientes,  $X = X_1 + X_2 + \dots + X_n$ , entonces, si  $n$  es suficientemente grande y se cumplen ciertas condiciones,  $X$  sigue aproximadamente una distribución normal. La aproximación mejora conforme aumenta  $n$ .*



### Consecuencias

Una de las primeras consecuencias del Teorema es que nos permite dar alguna explicación al hecho de que sea tan frecuente la aparición de variables normales cuando se realizan estudios reales. Si admitimos que estas variables, las normales, no son más que el resultado de la suma de muchas causas, pueden interpretarse como la suma de una gran cantidad de variables aleatorias distintas, por lo que, en virtud del Teorema, se distribuirán de una manera muy similar a la normal. A veces no puede admitirse que la variable en estudio sea suma de otras, sino proporcional a muchas otras. En este caso la variable en estudio podría expresarse como producto de varias variables.  $X = X_1 X_2 \dots X_n$  y por tanto la variable  $\log X = \log X_1 + \log X_2 + \dots + \log X_n$  será la que debe considerarse normal, aunque la variable  $X$  no lo fuera.

Daremos ahora enunciados más formales del Teorema Central del límite. Para ello es necesario aclarar previamente el sentido que hay que darle a la frase “la variable  $X$  sigue *aproximadamente* una distribución normal”, lo que se hará en la siguiente definición de convergencia.

#### 4.9.1 Convergencia en distribución (o en ley)

Diremos que la sucesión de variables aleatorias  $X_n$  converge en ley hasta  $X$  si la sucesión de las funciones de distribución de las  $X_n$ ,  $F_n(X)$ , converge hacia  $F(X)$  en cada punto  $X$  de continuidad de la función  $F(X)$

$$\lim_{n \rightarrow \infty} F_n(X) = F(X)$$

#### Forma de Lyapunov del Teorema central del límite

Dadas  $X_1 \dots X_n$ , variables aleatorias independientes con media  $\mu_i$ , varianza  $\sigma_i^2 < \infty$  y algún  $\sigma_i > 0$ , y con funciones de distribución cualesquiera, no necesariamente la misma, formamos la variable suma  $X = X_1 + \dots + X_n$ . Si se cumple la condición de Lyapunov para algún  $\delta > 0$  (existe  $\delta > 0$  tal que  $\lim_{n \rightarrow \infty} \frac{1}{(\sqrt{\text{var}(X)})^{2+\delta}} \sum_{k=1}^n E|X_k - \mu_k|^{2+\delta} = 0$ ) entonces, cuando  $n$  crece la variable

$$\frac{X - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

converge en ley hacia la función de distribución de la  $N(0, 1)$ . El resultado anterior implica que si  $n$  es grande podemos aproximar las probabilidades de  $X$  usando una  $N(\sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2})$

**Forma de Lindeberg-Levy (caso particular de la forma de Lyapunov)**

Si  $X_1 \dots X_n$  son variables aleatorias independientes y con la misma distribución, por tanto todas las variables tendrán la misma media  $\mu$  y varianza  $\sigma^2 \neq 0$ , y formamos la variable suma  $X = X_1 + \dots + X_n$ , entonces cuando  $n$  crece la variable

$$\frac{X - n\mu}{\sigma\sqrt{n}}$$

converge en ley hacia una la función de distribución  $N(0,1)$ . El resultado anterior implica que si  $n$  es grande (se suele considerar suficiente  $n \geq 30$ ) podemos aproximar las probabilidades de  $X$  usando una  $N(n\mu, \sigma\sqrt{n})$

**4.9.2 Aplicaciones del Teorema Central del Límite****Distribución de la media muestral**

Dada una variable aleatoria  $X$ , de media  $\mu$  y varianza  $\sigma^2$ , seleccionando distintas muestras de  $n$  elementos y hallando sus respectivas medias, obtendremos, por regla general, un resultado diferente para cada una de estas medias muestrales, debido a la influencia que tiene el azar en la selección de cada muestra. Por tanto la media de las muestras de tamaño  $n$  es otra variable aleatoria. Denotemos esta media muestral por  $\bar{X}$ . ¿Qué distribución sigue esta variable aleatoria? No hay una respuesta genérica ya que depende de las distribuciones de partida. Pero si el número de elementos de la muestra es amplio, el Teorema Central del Límite nos da una solución aproximada para la distribución del estadístico media muestral: La distribución de las medias muestrales se van aproximando a la distribución Normal al ir creciendo el valor de  $n$ . Para ver que es un caso particular del teorema central del límite basta observar que  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}$  es la suma de  $n$  variables aleatorias independientes con la misma distribución de probabilidad. La media de cada una de estas variables aleatorias es  $\frac{\mu}{n}$  y su varianza es  $\frac{\sigma^2}{n^2}$ . Usando el teorema en la forma de Lindeberg-Levy se deduce que  $\bar{X}$  se distribuye aproximadamente según una distribución  $N(n\frac{\mu}{n}, \frac{\sigma}{n}\sqrt{n}) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ .

Por tanto si la variable aleatoria  $X$  tiene media  $\mu$  y varianza finita  $\sigma^2$ , la distribución de las medias de las muestras de tamaño  $n$  tiende hacia una  $N(\mu, \sigma/\sqrt{n})$  y por tanto la variable aleatoria  $\frac{X - \mu}{\sigma/\sqrt{n}}$  converge en ley hacia la  $N(0,1)$ .

### Relaciones entre las distribuciones Binomial y Poisson con la Normal

Como otra aplicación del Teorema Central del Límite podemos citar la aproximación de la distribución Binomial por medio de la distribución Normal. En efecto, una variable aleatoria binomial no es más que la suma de variables aleatorias independientes de tipo Bernoulli. Por tanto, podemos aplicar el Teorema central del Límite, con lo que obtendremos que si  $X \in B(n, p)$ , para valores grandes de  $n$ ,  $X$  se distribuye aproximadamente como una  $N(np, \sqrt{npq})$ . Se considera que esta aproximación es adecuada si  $np > 5$ ,  $nq = n(1 - p) > 5$  y  $0.5 < p < 0.95$ ,  $0.5 < q < 0.95$ .

**Ejemplo 30** *La probabilidad de curación de un cierto tipo de cáncer es de  $\frac{1}{3}$ . En un grupo de 100 de estos enfermos, ¿cuál es la probabilidad de que sanen al menos 30 de ellos?*

Si llamamos  $i$  a la variable aleatoria que representa el número de enfermos que sanen, y usamos directamente la distribución binomial, debemos calcular

$$P(i \geq 30) = \sum_{i=30}^{100} \binom{100}{i} \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{100-i} = 1 - \sum_{i=0}^{29} \binom{100}{i} \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{100-i} = 0.79073$$

Debido a que en ocasiones es complicado calcular este tipo de probabilidades, en especial si  $i$  toma valores altos, es frecuente usar su valor aproximado dado por el teorema central del límite. Es decir, aproximamos la distribución de la variable binomial  $i$  por medio de una normal que llamaremos  $X$ .

$$N(np, \sqrt{npq}) \approx N\left(100 \times \frac{1}{3}, \sqrt{100 \times \frac{1}{3} \times \frac{2}{3}}\right) = N(33.333, 4.714)$$

Dado que la variable binomial es discreta y la Normal es continua hay que hacer una asignación de probabilidades de una en otra. Esta asignación se realiza de la siguiente forma:

$$P(i = a) \approx P(a - 0.5 < X < a + 0.5)$$

Esta modificación, llamada *corrección por continuidad*, hay que hacerla siempre que una variable discreta se aproxime por medio de una variable continua.

Así que

$$P(i \geq 30) \approx P(X \geq 29.5)$$

La probabilidad  $P(X \geq 29.5)$ , usando la distribución  $N(33.333, 4.714)$  es 0.79192 que como se ve tiene un valor bastante parecido al obtenido usando la distribución binomial.

Si tomáramos directamente  $X = 30$  para la variable normal,  $P(X \geq 30) = 0.76023$ , valor que, aunque próximo al valor obtenido con la binomial, es bastante peor que el obtenido tomando  $X = 29.5$ .

Pasamos ahora a describir las circunstancias en las que la distribución de Poisson puede aproximarse por medio de la Normal. Se puede demostrar que la suma de  $n$  variables de Poisson de parámetro  $\lambda_1$  se distribuye como una Poisson de parámetro  $n\lambda_1$ . Por tanto podemos interpretar la distribución de Poisson de parámetro  $\lambda$  como la suma de  $n$  variables de Poisson de parámetro  $\frac{\lambda}{n}$ , con lo cual si  $n$  aumenta (lo que requiere que  $\lambda$  sea grande) se aproximará a una normal. En estas circunstancias se tiene que: Si  $X$  se distribuye según una Poisson  $P(\lambda)$ , entonces  $X$  es aproximadamente  $N(\lambda, \sqrt{\lambda})$ . Esta aproximación se considera buena si  $\lambda > 10$ .

#### 4.10 Simulación del Teorema Central del Límite

Como ya hemos anotado anteriormente, un caso particular del Teorema Central del límite, nos permite asegurar que si  $X$  es una variable aleatoria con media  $\mu$  y desviación típica  $\sigma$ , entonces el estadístico media muestral,  $\bar{X}$ , de las muestras de tamaño  $n \geq 30$  sigue una distribución  $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ .

Para visualizar este teorema, usando muestras simuladas, podemos seguir los siguientes pasos:

**Paso 1:** Generar  $N$  muestras de  $n$  elementos cada una que sigan la función de distribución de la variable  $X$ .

**Paso 2:** Calcular las medias de cada una de estas  $N$  muestras.

**Paso 3:** Considerando la muestra  $M$ , de  $N$  elementos, formada con las  $N$  medias obtenidas en el Paso 2, calcular:

3.a) La media de  $M$  (la media de las medias muestrales de  $X$ )

3.b) La cuasidesviación típica de  $M$ , ya que la cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

3.c) Realizar una tabla de frecuencias y un histograma de frecuencias para la variable  $M$ .

**Paso 4 (Comprobación):**

4.a) Se comparará el valor obtenido en 3.a) con  $\mu$

4.b) Se comparará el valor obtenido en 3.b) con  $\frac{\sigma}{\sqrt{n}}$ .

4.c) Se comparará la tabla de frecuencias con las probabilidades correspondientes a los intervalos de la tabla si se usa la distribución  $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$  indicada en la tesis del teorema. Igualmente se comparará el histograma de

frecuencias obtenido en 3.c) con la representación gráfica de la distribución teórica  $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ .

Se podrá apreciar un buen acuerdo entre la simulación y el enunciado del teorema si se tienen las siguientes precauciones: El valor de  $N$  (número de muestras generadas) debe ser lo bastante grande para permitir dibujar un histograma de frecuencias razonable. El número de elementos,  $n$ , de cada muestra debe ser al menos 30.

También se puede usar la simulación para comprobar que si el número de elementos de las muestras no es suficientemente grande, por ejemplo tomando  $n = 2$  o  $n = 3$  elementos en las muestras, la distribución de las medias no tiene por qué ser normal, aunque si lo sería si las muestras procedieran de variables normales. En la red hay numerosos applets que suministran ejemplos de este tipo.

Otra utilidad de la simulación es obtener datos para realizar ejemplos ilustrativos de muchos temas estadísticos. Así, es posible obtener muestras de distribuciones de cualquier tipo para los ejemplos de clase.

También se pueden abordar diversos temas de investigación. Por ejemplo: estudiar la distribución del estadístico *diferencia entre la media y la mediana*, estudiar la distribución del valor *máximo de una muestra de determinado tamaño*, del *coeficiente de variación*, etc, tanto si la variable de partida es normal como si no lo es. Para realizar el citado estudio se puede recurrir a generar muestras de estas distribuciones, obtener en cada una de ellas el valor del estadístico de interés y estudiar las propiedades de la distribución muestral obtenida considerando, por ejemplo, la función de distribución empírica.

## 4.11 EJERCICIOS PROPUESTOS

**Ejercicio 54** *Generar 100 valores con el método de generación de números aleatorios de algún paquete estadístico y comprobar la calidad del algoritmo usando esta muestra:*

1. *Por medio del test Chi-cuadrado para comprobar el ajuste de la muestra a la distribución  $U[0,1]$ .*
2. *Usando las funciones de autocorrelación muestral para comprobar la independencia de los valores muestrales.*

**Ejercicio 55** *Construir un generador de números aleatorios usando el método congruencial, basado en la relación*

$$X_{i+1} \equiv (aX_i + c) \pmod{m}$$

$$c = 7, a = 5, m = 2^{16}.$$

1. ¿Cuántos números distintos genera este algoritmo?
2. Usando este algoritmo anterior genera valores procedentes de una distribución  $U[3,6]$ .

**Ejercicio 56** Simula con el ordenador la tirada de tres dados.

**Ejercicio 57** Implementa un algoritmo que genere valores controlados por una distribución exponencial.

**Ejercicio 58** Implementa un algoritmo que genere valores controlados por una distribución binomial.

**Ejercicio 59** Teniendo en cuenta que la suma de variables exponenciales es una variable de Erlang, implementar un algoritmo para generar valores que se rijan por una distribución de Erlang.

**Ejercicio 60** Una variable que se distribuye uniformemente entre 7500 y 10500. hallar la probabilidad de que la media de una muestra de 1000 elementos sea mayor que 9050.

**Ejercicio 61** En la placa correspondiente a las características de un ascensor se lee: “nº máximo de personas: 6, peso máximo admitido: 450 kg.”. La población de usuarios tiene un peso que se distribuye según una Ley Normal de media 60 Kg y desviación típica 20 Kg.

1. Hallar la probabilidad de que al montarse 6 personas en el ascensor se supere el peso máximo señalado.
2. La indicación de la placa tiene un margen de seguridad. El peligro es real si se superan los 550 kg. ¿A partir de qué número de personas habrá una probabilidad mayor del 10% de que haya peligro real? ¿Cual es la probabilidad de peligro real si se suben 6 personas?

## Tema 5

# Inferencia Estadística.

### 5.1 Introducción a la Inferencia Estadística

La palabra inferir significa extraer consecuencias, o deducir un conocimiento a partir de otro. La Inferencia Estadística es la parte de la estadística que se encarga de deducir características de la población a partir de los resultados obtenidos en muestras de esta población. Las decisiones se basan en la información contenida en muestras extraídas de ella.

En muchas circunstancias hay que tomar decisiones basándose sólo en la información contenida en una muestra: Un gerente de Control de Calidad debe determinar si un proceso funciona correctamente. Para ello, cada cierto tiempo, analiza la calidad de una pequeña cantidad de productos fabricados por este proceso. Con esta información debe decidir si continúa fabricando nuevas piezas, o si debe realizar algún ajuste o reparación de la maquinaria de la fábrica antes de continuar el proceso de fabricación. Un gerente de Marketing debe determinar si una nueva estrategia de mercado aumentará las ventas. Para ello se basará fundamentalmente en encuestas realizadas a unos cuantos clientes potenciales, etc. Para adoptar estas decisiones se toma toda la información posible de la muestra seleccionada y se estudia, en términos de probabilidad, el grado de fiabilidad de las decisiones adoptadas.

Podemos distinguir de modo general dos grandes métodos dentro de la Inferencia Estadística:

**Métodos Paramétricos.-** Se supone que los datos provienen de una familia de distribuciones conocida (Normal, Poisson, . . .) y que lo único que se desconoce es el valor concreto de alguno de los parámetros que la definen ( $\mu$  y  $\sigma$  para la Normal,  $\lambda$  para la Poisson, . . .).

Se pueden hacer inferencias acerca de los parámetros poblacionales de dos maneras. Dando valores aproximados para los parámetros (Es-

timación) o tomando decisiones con respecto a ellos (Contrastes de Hipótesis).

**Métodos No Paramétricos.-** No suponen conocida la distribución, y solamente suponen hipótesis muy generales respecto a las mismas. Estos métodos se aplican en los tests de bondad de ajuste, que prueban la adecuación de los datos a ciertos modelos de distribuciones teóricas, los test de independencia, etc.

Evidentemente, las conclusiones que obtengamos y que generalizaremos para toda la población dependerán de los valores concretos que se hayan observado en la muestra. Muchas personas manifiestan su desconfianza y su recelo sobre las conclusiones obtenidas con métodos estadísticos, debido, entre otras causas, a que estas conclusiones dependen de la muestra extraída, y que las muestras presentan fluctuaciones aleatorias. Sin embargo, en la vida cotidiana, nuestras opiniones y nuestros comportamientos se basan en generalizaciones que hacemos a partir de muestras. Así, es muy frecuente que manifestemos que los productos de una determinada marca son mejores que los de la competencia. Dicha afirmación no la hacemos, evidentemente, tras un análisis exhaustivo de todos los productos de una y otra marca, sino basándonos en nuestra propia experiencia personal, que es claramente muy limitada. Es decir, generalizamos a partir de observaciones realizadas en muestras pequeñas.

## 5.2 Tipos de estimación

Cuando se desean estimar los parámetros de la población a partir de los de la muestra se consideran dos formas de realizar dicha estimación.

**Estimación puntual.-** En la estimación puntual damos un solo punto como valor estimado del parámetro. Por ejemplo, si queremos estimar la altura media,  $\mu$ , de los varones españoles de 20 años, obtendremos una muestra aleatoria de cierto tamaño de esta población, hallaremos la altura media de las personas seleccionadas en esta muestra y diremos que este valor, el de la media muestral, es una estimación puntual de la altura media de la población de varones de 20 años.

**Estimación por intervalos.-** En realidad, cuando realizamos una estimación puntual, nos damos cuenta que es muy difícil que ésta estimación sea realmente el verdadero valor del parámetro desconocido. Tendremos más oportunidades de acertar si indicamos que el parámetro desconocido pertenece a un cierto intervalo. En el ejemplo de la altura media de los varones de 20 años, si la media muestral resultara 1.75 m.,



podríamos decidir manifestar que la media verdadera pertenece al intervalo  $(1.75 - 0.05, 1.75 + 0.05)$ . El intervalo en el que se afirma que se encuentra el parámetro poblacional se denomina *intervalo de confianza*. Tampoco en este caso podemos estar seguros de que el valor real pertenezca a dicho intervalo. Por este motivo suele decirse que el valor real del parámetro pertenece a dicho intervalo con un cierto “grado de confianza”. La cuantificación de la confianza que se tiene en que el parámetro desconocido esté verdaderamente en el intervalo dado se denomina *grado de confianza* y es una medida relacionada con la función de distribución de probabilidad del parámetro en estudio.

### 5.3 Estadísticos y Estimadores

Un *estadístico* es una función de los elementos de la muestra. Si tenemos una población en la que estamos observando una característica que se distribuye según una variable aleatoria  $X$ , y consideramos una muestra aleatoria simple de tamaño  $n$

$$x_1, x_2, \dots, x_n$$

podemos calcular el siguiente estadístico  $\bar{X}$ :

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Evidentemente, el valor del estadístico dependerá de los valores que hayan tomado los elementos de la muestra. Si repetimos el experimento de tomar una muestra y calculamos de nuevo el valor del mismo estadístico, obtendremos, por lo general, otro valor distinto. Tenemos por tanto que el estadístico es una variable aleatoria. La distribución que seguirá dicha variable aleatoria dependerá de la distribución de la variable  $X$ . En determinados casos podremos calcular la distribución del estadístico.

Un *estimador* de un parámetro poblacional es un estadístico que se utiliza para obtener un valor aproximado de ese determinado parámetro de la población. Por ejemplo, la media muestral es el estadístico que suele usarse más frecuentemente para estimar la media poblacional. Entonces, la media muestral es un estimador de la media poblacional. La mediana y la moda son también estimadores de la media poblacional. Para indicar que  $T$  es un estimador del parámetro poblacional  $\theta$  se indicará

$$T = \hat{\theta}$$

El valor que toma este estimador en la muestra concreta que estamos considerando es una *estimación* del parámetro desconocido.

## 5.4 Propiedades de los estimadores

¿Cómo se eligen estimadores para la media y la varianza de la población? Más en general, ¿cómo se eligen los estimadores de los parámetros poblacionales? Es normal que exijamos que los estimadores, al menos en promedio, se parezcan al parámetro que quiere estimarse. También es conveniente que no fluctúen demasiado con las distintas muestras y que mejoren si aumentamos el tamaño de ésta. Estas condiciones son las que están formuladas en las siguientes definiciones.

### Centrado o insesgado

Una de las propiedades que con más frecuencia se le exige a los estimadores es que sean *insesgados*. Decimos que  $T$  es un estimador centrado o insesgado del parámetro  $\theta$  si para cualquier tamaño muestral se cumple que

$$E(T) = \theta$$

### Eficiencia de dos estimadores

Si tenemos dos estimadores  $T_1$  y  $T_2$  de un parámetro  $\theta$ , decimos que  $T_1$  es más eficiente que  $T_2$  si para cualquier tamaño muestral se verifica que

$$\text{Var}(T_1) \leq \text{Var}(T_2)$$

Entre dos estimadores posibles sería preferible el más eficiente.

### Consistencia de un estimador

Diremos que un estimador es consistente si cumple:

$$\lim_{n \rightarrow \infty} E(T) = \theta \quad \text{y} \quad \lim_{n \rightarrow \infty} \text{Var}(T) = 0$$

donde  $n$  es el tamaño de la muestra.

## 5.5 Estimadores insesgados de la media y la varianza

Para una variable aleatoria  $X$ , con media  $\mu$  y varianza  $\sigma^2$ , un estimador insesgado de la media poblacional es

$$\hat{\mu} = \bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

y un estimador insesgado de la varianza es

$$\widehat{\sigma^2} = S_c^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Es decir, la media muestral es un estimador insesgado de la media poblacional y la cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

## 5.6 Distribución de los estadísticos muestrales

Cuando se realiza una estimación por medio de intervalos de confianza se da un *grado de confianza*. Este valor se basa en la proporción de muestras en las que el parámetro que se desea estimar quedaría dentro del intervalo de confianza dado. Para calcular esta proporción es necesario conocer la distribución del estimador en el muestreo. Con el propósito de conocer el grado de confianza asociado a las estimaciones por intervalo de la media y de la varianza poblacionales son útiles los siguientes teoremas.

**Teorema 2** *En el caso de que la variable  $X$  sea normal, podemos afirmar que*

$$\bar{X} \in N(\mu, \sigma/\sqrt{n})$$

y que

$$\frac{(n-1)S_c^2}{\sigma^2} \in \chi_{n-1}^2$$

*esto es, una chi-cuadrado con  $n-1$  grados de libertad.*

*Nota:  $S_c$  es la cuasidesviación muestral.*

Si la población no es normal y  $n$  es grande podemos afirmar, por el Teorema Central del Límite que

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

A efectos prácticos se emplea esta aproximación si  $n \geq 30$ .

**Teorema 3** *Si  $X$  es una variable aleatoria  $N(\mu, \sigma)$*

$$T = \frac{\bar{X} - \mu}{S_c/\sqrt{n}}$$

*se distribuye según una  $t$  de Student con  $n-1$  grados de libertad.*

Si la  $X$  no sigue una distribución normal, este resultado puede aplicarse a efectos prácticos si el número de elementos de la muestra es mayor o igual que 60.

## 5.7 Intervalos de confianza para la media

En este epígrafe consideraremos únicamente poblaciones normales

### 5.7.1 Con varianza conocida

Supongamos que la población se distribuye según una variable

$$X \in N(\mu, \sigma)$$

y que la desviación típica  $\sigma$  es conocida.

Si tomamos muestras de tamaño  $n$  y calculamos el estadístico  $\bar{X}$ , se tiene que según el teorema 17.3

$$\bar{X} \in N(\mu, \sigma/\sqrt{n})$$

Un *intervalo de confianza* para la media poblacional, centrado en la media muestral,  $(\bar{X} - b, \bar{X} + b)$ , con una *grado de confianza*  $100(1 - \alpha)\%$ , es un intervalo que cumple la siguiente condición:

$$P(\bar{X} - b < \mu < \bar{X} + b) = 1 - \alpha$$

lo que quiere decir que la media poblacional estaría contenida en el intervalo obtenido a partir de la muestra, en el  $100(1 - \alpha)\%$  de las muestras. El valor de  $\alpha$  se conoce con el nombre de *nivel de significación*.

Para obtener dicho intervalo se tienen en cuenta las siguientes consideraciones:

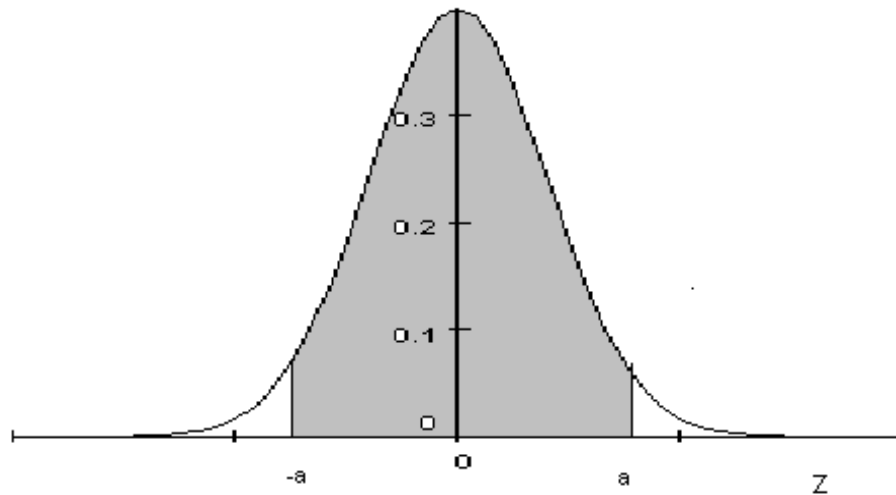
Tipificando la variable  $\bar{X}$ , se obtiene

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in N(0, 1)$$

Buscamos un valor  $a$  que cumpla

$$P(-a < Z < a) = 1 - \alpha$$

Gráficamente, considerando la función de densidad de la  $N(0, 1)$ , tendríamos la siguiente situación:



La zona gris de la figura debería tener un área  $1 - \alpha$ , así que cada una de las zonas laterales tiene un área  $\frac{\alpha}{2}$ ,

Al ser  $Z \in N(0, 1)$ , el valor de  $a$  se busca en las tablas de la normal estándar, considerando que ha de cumplir:

$$P(Z < a) = 1 - \alpha/2$$

Sustituyendo  $a$  por el valor hallado en las tabla, que denotamos<sup>1</sup> por  $z_{1-\alpha/2}$ , y teniendo en cuenta la simetría de la función de densidad, se obtiene:

$$P(-z_{1-\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1 - \alpha$$

Despejando  $\mu$  en ambas desigualdades, se obtiene:

$$P(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}) = 1 - \alpha$$

Así que el intervalo de confianza para la media poblacional  $\mu$ , al nivel de significación  $\alpha$ , será:

$$(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2})$$

**Ejemplo 31** Calculemos un intervalo de confianza al 90%, para la media de una población normal a partir de la siguiente muestra extraída de ella.

6. 411, 4. 324, 5. 282 5, 3. 268 9, 3. 705, 4. 973 6, 1. 897 7, 4. 468 1, 3. 796 1, 4. 966 6  
Suponemos conocida la desviación típica de la población  $\sigma = 1.2$ .

<sup>1</sup>Frecuentemente se usa también la notación  $z_{\alpha/2}$

Obtenemos la media muestral que resulta ser:  $\bar{X} = 4.3094$

El intervalo de confianza para la media de la población, con una confianza del 90%, es:

$$(4.3094 - \frac{1.2}{\sqrt{10}}1.6449, 4.3094 + \frac{1.2}{\sqrt{10}}1.6449) = (3.6852, 4.9336)$$

El valor de  $z_{1-\alpha/2} = z_{1-0.10/2} = z_{0.95}$ , se ha calculado con la condición:  $P(z < z_{0.95}) = 0.95$ . Mirando en la tabla de la  $N(0,1)$  resulta que el valor de  $z_{0.95} = 1.6449$ .

### 5.7.2 Con varianza desconocida

Supongamos que la población se distribuye según una variable

$$X \in N(\mu, \sigma)$$

con  $\sigma$  desconocida. Si tomamos una muestra de tamaño  $n$ , y calculamos el estadístico  $\bar{X}$ , se tiene que

$$\bar{X} \in N(\mu, \sigma/\sqrt{n})$$

de donde, tipificando, se obtiene

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in N(0, 1)$$

Pero ahora, al no conocer el valor de  $\sigma$  no podemos calcular exactamente el valor de  $Z$ .

Estimando el valor de  $\sigma$  por su estimador  $S_c$ , se obtiene el estadístico

$$T = \frac{\bar{X} - \mu}{S_c/\sqrt{n}} \quad (5.1)$$

que ya no se distribuiría según una normal sino que, según el teorema 3, se distribuye como una  $t$  de Student con  $n - 1$  grados de libertad.

Si queremos obtener un intervalo que tenga una probabilidad  $1 - \alpha$  de contener al parámetro  $\mu$ , esto equivaldría a

$$P(-t_{1-\alpha/2} < T < t_{1-\alpha/2}) = 1 - \alpha \quad (5.2)$$

Al ser  $T$  una  $t$  de Student con  $n - 1$  grados de libertad, el valor de  $t_{1-\alpha/2}$  se busca en las tablas de la citada distribución. Considerando que

$$P(T < t_{1-\alpha/2}) = 1 - \alpha/2$$

y sustituyendo la expresión 5.1 de  $T$  en la 5.2, se obtiene:

$$P(-t_{1-\alpha/2} < \frac{\bar{X} - \mu}{S_c/\sqrt{n}} < t_{1-\alpha/2}) = 1 - \alpha$$

Despejando  $\mu$  en la doble desigualdad deducimos que:

$$P\left(\bar{X} - \frac{S_c}{\sqrt{n}}t_{1-\alpha/2} < \mu < \bar{X} + \frac{S_c}{\sqrt{n}}t_{1-\alpha/2}\right) = 1 - \alpha$$

Así que el intervalo de confianza para la media poblacional  $\mu$  con una confianza  $100(1 - \alpha)\%$  es:

$$\left(\bar{X} - \frac{S_c}{\sqrt{n}}t_{1-\alpha/2}, \bar{X} + \frac{S_c}{\sqrt{n}}t_{1-\alpha/2}\right) \quad (5.3)$$

**Ejemplo 32** Hallar un intervalo de confianza al 95% para la media basado en la misma muestra del ejemplo anterior:

6. 411, 4. 324, 5. 282 5, 3. 268 9, 3. 705, 4. 973 6, 1. 897 7, 4. 468 1, 3. 796 1, 4. 966 6

Se supone ahora que la desviación típica de la población de la que procede dicha muestra es desconocida.

Calculamos en primer lugar la media y la cuasidesviación de la población:

$$\bar{X} = 4.3094 \quad S_c = 1.2378$$

El valor de  $t_{1-\alpha/2} = t_{1-0.05/2} = t_{0.975}$  se ha calculado con la condición:  $P(t < t_{0.975}) = 0.975$ . Por tanto, mirando en la tabla de la  $t$  de Student con con 9 grados de libertad obtenemos que su valor es:  $t_{0.975} = 2.2622$ .

Sustituyendo en la expresión 5.3, los valores obtenidos y teniendo en cuenta que el número de elementos de la muestra es 10, obtenemos el siguiente intervalo de confianza para la media poblacional:

$$\left(4.3094 - \frac{1.2378}{\sqrt{10}}2.2622, 4.3094 + \frac{1.2378}{\sqrt{10}}2.2622\right) = (3.4239, 5.1949)$$

## 5.8 Intervalos de confianza para la varianza

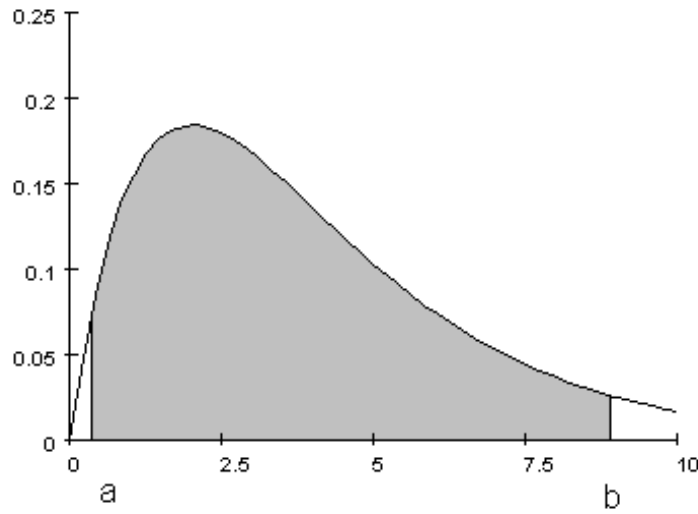
Los intervalos de confianza para la varianza se basan en la distribución del estadístico  $\frac{(n-1)S_c^2}{\sigma^2}$  que según el teorema 17.3, sigue una distribución chi-cuadrado con  $n - 1$  grados de libertad, si la variable de partida es normal..

Para obtener un intervalo de confianza partimos de la relación

$$P\left(a < \frac{(n-1)S_c^2}{\sigma^2} < b\right) = 1 - \alpha$$

Considerando la siguiente gráfica de la distribución chi-cuadrado, asignamos al área central de color gris el valor  $1 - \alpha$ , de modo que

$$P(\chi_{n-1}^2 < a) = \frac{\alpha}{2} \quad \text{y} \quad P(\chi_{n-1}^2 < b) = 1 - \frac{\alpha}{2}$$



Por tanto, si denotamos por  $a$  y  $b$  los valores de chi-cuadrado con  $n - 1$  grados de libertad, que dejan delante las áreas  $\frac{\alpha}{2}$  y  $1 - \frac{\alpha}{2}$  respectivamente:

$$a = \chi_{n-1, \frac{\alpha}{2}}^2 \quad y \quad b = \chi_{n-1, 1-\frac{\alpha}{2}}^2$$

obtenemos:

$$\chi_{n-1, \frac{\alpha}{2}}^2 < \frac{(n-1)S_c^2}{\sigma^2} < \chi_{n-1, 1-\frac{\alpha}{2}}^2$$

Despejando  $\sigma^2$

$$\frac{(n-1)S_c^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1)S_c^2}{\chi_{n-1, \frac{\alpha}{2}}^2}$$

De ello deducimos que un intervalo de confianza para la varianza poblacional  $\sigma^2$ , con nivel de confianza  $(1 - \alpha)\%$  es

$$\left( \frac{(n-1)S_c^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)S_c^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right) \quad (5.4)$$

y para la desviación típica

$$\left( S_c \sqrt{\frac{n-1}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}}, S_c \sqrt{\frac{n-1}{\chi_{n-1, \frac{\alpha}{2}}^2}} \right)$$



**Ejemplo 33** Hallar un intervalo de confianza para la desviación típica de la población, usando la información de la muestra de los dos ejemplos anteriores.

$$\left( 1.2378\sqrt{\frac{9}{19.023}}, 1.2378\sqrt{\frac{9}{2.7004}} \right) = (0.8514, 2.2597)$$

siendo los denominadores:

$$\chi_{n-1, \frac{\alpha}{2}}^2 = \chi_{10-1, \frac{0.05}{2}}^2 = \chi_{9, 0.025}^2 = 2.7004$$

$$\chi_{n-1, 1-\frac{\alpha}{2}}^2 = \chi_{10-1, 1-\frac{0.05}{2}}^2 = \chi_{9, 0.975}^2 = 19.023$$

Estos valores se obtienen a partir de la tabla de la distribución Chi-cuadrado.

## 5.9 Contrastes de hipótesis

### 5.9.1 Introducción:

Proponemos, como paso previo, un ejemplo sencillo que nos ayudará a entender los conceptos fundamentales que se ponen en juego a la hora de construir un contraste de hipótesis (también llamado prueba de hipótesis o test de hipótesis).

Supongamos que tratamos de comprobar si una moneda está bien construida, es decir si cara y cruz tienen la misma probabilidad de aparición. Para ello realizamos una prueba con esta moneda consistente en arrojarla 50 veces. Si esta experiencia ha dado como resultado 49 veces cara y 1 vez cruz. ¿Qué concluiríamos? Seguramente no aceptaríamos que la moneda esté bien construida, a pesar de que el resultado obtenido es posible para una moneda bien construida. Incluso sería posible obtener 50 caras. Entonces ¿por qué la moneda no nos parece aceptable? Intuitivamente percibimos que el resultado que hemos obtenido sería “demasiado raro” para una moneda simétrica. Esperábamos que el resultado estuviera más cerca del promedio: 25 caras y 25 cruces.

Antes de tomar una decisión deberíamos adoptar un criterio razonable para aceptar o rechazar la simetría de la moneda. ¿Qué diferencia con respecto al promedio estaríamos dispuestos a aceptar? ¿Qué número de caras y cruces nos parecerían valores “demasiado raros”?

La respuesta sólo podemos darla en términos de probabilidad, ya que sabemos que los resultados dados en el ejemplo, y considerados “raros”, son posibles. Es decir, debemos aceptar un cierto riesgo: El de catalogar como mal construida una moneda aceptable.

Supongamos que decidimos no admitir que la moneda es simétrica cuando, como en el ejemplo, la diferencia en número de caras o cruces con respecto al valor medio, 25, sea mayor o igual que 24. ¿Cuál sería la probabilidad de declararla erróneamente como defectuosa?

Rechazaríamos una moneda bien construida si se diera uno de los siguientes sucesos:

- a) si salen 49 caras y 1 cruz.
- b) si salen 50 caras.
- c) si salen 49 cruces y una cara.
- d) si salen 50 cruces.

La probabilidad de rechazar erróneamente la moneda sería entonces:

$$\binom{50}{1}0.5^{49} \times 0.5 + 0.5^{50} + \binom{50}{1}0.5^{49} \times 0.5 + 0.5^{50} = 2 \times \left( \binom{50}{1}0.5^{49} \times 0.5 + 0.5^{50} \right) =$$

$$= 9.0594 \times 10^{-14}$$

Si la moneda fuera buena este suceso es muy poco probable. Por eso cuando obtenemos ese resultado, 49 caras y una cruz, pensamos que ha ocurrido algo “demasiado raro” para una moneda bien balanceada. Pensamos que lo normal sería desconfiar y decidimos que la moneda es defectuosa.

Con el objeto de ayudarnos a decidir cuando el suceso nos parece “muy poco probable” obtenemos la probabilidad de los sucesos siguientes:

- a) Que el número de caras o de cruces sea menor o igual que cinco

$$2 \sum_{i=0}^5 \left( \binom{50}{i} 0.5^i \times 0.5^{50-i} \right) = 4.2099 \times 10^{-9}$$

- b) Que el número de caras o de cruces sea menor o igual que diez

$$2 \sum_{i=0}^{10} \left( \binom{50}{i} 0.5^i \times 0.5^{50-i} \right) = 2.3861 \times 10^{-5}$$

- c) Que el número de caras o de cruces sea menor o igual que quince

$$2 \sum_{i=0}^{15} \left( \binom{50}{i} 0.5^i \times 0.5^{50-i} \right) = 6.6004 \times 10^{-3}$$

- d) Que el número de caras o de cruces sea menor o igual que diecisiete.

$$2 \sum_{i=0}^{17} \left( \binom{50}{i} 0.5^i \times 0.5^{50-i} \right) = 3.2839 \times 10^{-2}$$

Naturalmente que la decisión depende del riesgo que queramos correr (rechazando una moneda buena). Hay que resaltar que si disminuye el riesgo de rechazar una moneda buena, aumenta el riesgo de aceptar como buena una moneda mal equilibrada, por lo que debemos llegar a un compromiso entre ambas opciones.

Llamamos nivel de significación,  $\alpha$ , a la probabilidad de rechazar una moneda aceptable. Así, si aceptamos para  $\alpha$  el valor 0.032839, que corresponde al último suceso considerado, admitiríamos la moneda como buena

si el número de caras o de cruces estuviera entre 17 y 33. Este intervalo,  $[17, 33]$ , sería la *región de aceptación* de la prueba y el resto de los valores formarían la *región de rechazo*. Los valores de separación, 17 y 33, recibirían el nombre de *valores críticos* para esta prueba.

En resumen: Se pretende diseñar una prueba que nos permita dar un criterio para aceptar una hipótesis de partida: La moneda está bien construida, que se conoce con el nombre de *Hipótesis nula*, o rechazarla y aceptar la hipótesis contraria: la moneda no está bien construida, que se suele conocer con el nombre de *Hipótesis alternativa*.

El modo de actuar para aceptar o rechazar la moneda en cuestión podría ser el siguiente: La prueba consistiría en arrojar la moneda 50 veces. Se aceptaría que la moneda es buena si el número de caras o de cruces estuviera entre 17 y 33 y se rechazaría en caso contrario. El nivel de significación de la prueba (probabilidad de rechazar la hipótesis nula siendo verdadera) sería  $\alpha = 0.032839$ .

### 5.9.2 Conceptos generales

Una *hipótesis estadística* es una proposición referente a una o varias poblaciones. A menudo se refieren a su distribución de probabilidad o al valor de sus parámetros. Por ejemplo: La distribución de probabilidad de la moneda es uniforme discreta, el valor de la media de una variable aleatoria es 34, etc.

Un *test, prueba o contraste de hipótesis* es un conjunto de reglas para decidir cual de dos hipótesis,  $H_0$ , (*Hipótesis nula*) o  $H_1$ , (*hipótesis alternativa*), debe aceptarse en base a la información obtenida con una muestra. Una de estas hipótesis es la negación de la otra. Por lo general se supone que la hipótesis nula es algo que se ha admitido durante un cierto periodo de tiempo, que está vigente y que se mantendrá, salvo que haya pruebas que favorezcan claramente a la hipótesis alternativa.

Para realizar un test de hipótesis hay que decidir un valor para  $\alpha$  (*nivel de significación o probabilidad de error tipo I*), que es la probabilidad de rechazar la hipótesis nula siendo cierta.

Para realizar el test se supone que se cumple la hipótesis nula,  $H_0$ . La decisión se basa en un estadístico (*Estadístico de contraste*) que comprime la información relevante de la muestra), y que, si se cumple la hipótesis nula, se rige por una ley de probabilidad conocida. Para realizar la decisión se obtiene un intervalo de aceptación para el estadístico de contraste (*región de aceptación*). La región complementaria, formada por los valores que no pertenecen a la región de aceptación, suele llamarse *Región crítica* o de rechazo.

Si la muestra da al estadístico un valor dentro de la región de aceptación se acepta  $H_0$ , y se dice que el test es *estadísticamente no significativo*. En caso

contrario se acepta  $H_1$  y se dice que el test es *estadísticamente significativo*.

En un test de hipótesis pueden cometerse dos tipos de errores:

	$H_0$ es cierta	$H_0$ es falsa
Se acepta $H_0$	Acierto	Error tipo II
Se rechaza $H_0$	Error tipo I	Acierto

La decisión por  $H_1$  viene acompañada de una probabilidad de error dado por  $\alpha$  (Probabilidad de cometer un error tipo I o *nivel de significación* del test) que es la probabilidad de decidirse por la hipótesis alternativa siendo cierta la hipótesis nula.

La decisión por  $H_0$  viene acompañada de una probabilidad de error dado por  $\beta$  (Probabilidad de cometer el error tipo II) que es la probabilidad de decidirse por la hipótesis nula siendo cierta la alternativa. Este error no está controlado de antemano, por eso la decisión de aceptar  $H_0$  no es de fiar. Se denomina *Potencia* del test al valor de  $1 - \beta$  (probabilidad de decidirse por la hipótesis alternativa, si es cierta).

Un test de hipótesis lleva asociado un intervalo de confianza para los parámetros implicados.

Para aclarar estos conceptos teóricos proponemos el siguiente ejemplo:

**Ejemplo 34** *Se pretende diseñar una prueba de hipótesis con una muestra de 74 automóviles para comprobar su capacidad de frenado. Para ello se medirá en todos ellos la distancia de frenado si el automóvil parte de una velocidad inicial de 100 Km/h. Se quiere saber si, tras un frenazo brusco, la distancia media recorrida antes de pararse es de 110 metros. Se supone que la distancia de frenado sigue una distribución normal con desviación típica conocida  $\sigma = 3$  m.*

Consideramos las dos hipótesis:

$H_0 \equiv$  La media de la distribución es 110 m.

$H_1 \equiv$  La media de la distribución no es 110 m.

Para decidir el estadístico de contraste suponemos, momentáneamente, que la hipótesis nula es cierta (la media de la distribución es 110 m). Por tanto se admite que la distancia de frenado sigue una distribución  $N(110, 3)$ . Bajo esta hipótesis y según el teorema 17.3, el estadístico

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 110}{3/\sqrt{74}} \in N(0, 1)$$

Seleccionamos este estadístico como estadístico de contraste. Ahora debemos seleccionar una región de aceptación o de no rechazo para este estadístico. Tomamos como región de aceptación la que contiene los valores más

plausibles para este estadístico, que parece que deberían ser los valores más cercanos a su media, que coincide con la parte central de su distribución, el intervalo  $(-a, a)$ , que tomamos como región de aceptación. Los puntos de separación entre las regiones de aceptación y de rechazo,  $-a$  y  $a$ , se suelen llamar puntos críticos. Si imponemos que el error tipo I o nivel de significación del test sea  $\alpha$  (probabilidad de rechazar la hipótesis nula siendo cierta), entonces la probabilidad de aceptar la hipótesis nula sería

$$P(-a < Z < a) = 1 - \alpha \quad \implies a = z_{1-\frac{\alpha}{2}}$$

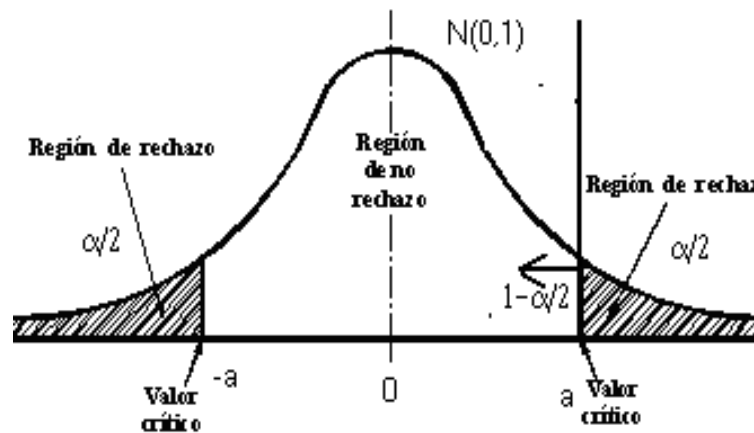


Figura 5.1: Test de dos colas.

Si tomamos para  $\alpha$  el valor estándar 0.05, obtenemos  $a$  de la condición

$$P(Z < a) = 1 - \alpha/2 = 1 - 0.05/2 = 0.975$$

El valor de  $a$  correspondiente lo obtenemos de la tabla de la  $N(0,1)$ , que resulta 1.96, lo que nos da una región de aceptación para  $Z$

$$-1.96 < \frac{\bar{X} - 110}{3/\sqrt{74}} < 1.96$$

La prueba de hipótesis a realizar es la siguiente: Se calcula la media de las distancias recorridas por los 74 automóviles y luego el valor de  $Z = \frac{\bar{X} - 110}{3/\sqrt{74}}$ . Si el valor resultante queda dentro del intervalo de aceptación se aceptaría la hipótesis  $H_0$  (la media de la distribución es 110 m.) con un nivel de significación  $\alpha = 0.05$ . Si quedara fuera de este intervalo se rechazaría esta hipótesis, ya que la muestra no apoyaría suficientemente la hipótesis nula.

Supongamos ahora que hemos realizado efectivamente la prueba a los 74 automóviles y hemos obtenido las siguientes distancias de frenado.

Distancias	98	102	105	113	123	126	
Num. de autos	15	10	12	8	16	13	Total 74

¿Se acepta la hipótesis de que la distancia media de frenado es de 110 m, con un nivel de significación  $\alpha = 0.05$ ?

En este caso el valor de  $\bar{X}$  de la muestra es

$$\bar{X} = \frac{15 * 98 + 10 * 102 + 12 * 105 + 8 * 113 + 16 * 123 + 13 * 126}{15 + 10 + 12 + 8 + 16 + 13} = 111.62$$

así que

$$Z = \frac{\bar{X} - 110}{3/\sqrt{74}} = \frac{111.62 - 110}{3/\sqrt{74}} = 4.65$$

Como este valor no entra dentro de la región de aceptación  $(-1.96, 1.96)$  nos decidimos por la hipótesis alternativa  $H_1$ . Concluimos que la media de frenado no es 110 m.

Este test de hipótesis lleva asociado un intervalo de confianza para la media. El intervalo de confianza para la media al 95%, sería

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right)$$

$$\left(111.62 - \frac{3}{\sqrt{74}}1.96, 111.62 + \frac{3}{\sqrt{74}}1.96\right) = (110.94, 112.3)$$

Dicho intervalo no contiene el valor 110 m que es el valor que estábamos probando para la media de las distancias de frenado. Concluimos que la media no es 110m. Esta comprobación por medio de un intervalo de confianza es totalmente equivalente a la realizada con el contraste de hipótesis.

En resumen, una prueba de hipótesis consta de los siguientes pasos:

1. Concretar la hipótesis nula  $H_0$ .
2. Formular una hipótesis alternativa  $H_1$ .

3. Decidir el estadístico de contraste (con función de distribución conocida si se verifica  $H_0$ ).
4. Fijar el nivel de significación deseado  $\alpha$ . Usar este valor para construir las regiones de aceptación y rechazo del estadístico de contraste.
5. Calcular el valor del estadístico a partir de la muestra.
6. Si el valor del estadístico pertenece a la región crítica o de rechazo, entonces rechazar  $H_0$ . En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar  $H_0$ .

Los contrastes de hipótesis, atendiendo al tipo de hipótesis alternativa, pueden ser bilaterales, como el correspondiente a la figura 5.1 de la página 173, o unilateral como el que corresponde a la figura 5.2 de la página 177.

## 5.10 Prueba de hipótesis para la media

### 5.10.1 Poblaciones normales

Para las pruebas de hipótesis que se describen en este epígrafe, se supone que las muestras proceden de poblaciones normalmente distribuidas.

#### Test de dos colas con varianza conocida

1.  $H_0 \equiv$  la media de la distribución es  $\mu = \mu_0$
2.  $H_1 \equiv$  la media de la distribución ,  $\mu \neq \mu_0$
3. Usamos como estadístico de contraste

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ , que sigue una distribución  $N(0, 1)$  ( $n$  es el tamaño de la muestra,  $\bar{X}$  es la media de la muestra)

4. Si el nivel de significación es  $\alpha$ . La región de aceptación es

$$-z_{1-\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2} \quad (5.5)$$

5. Calcular  $\bar{X}$  a partir de la muestra y evaluar  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ .
6. Si el valor de  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$  no cumple la relación 5.5, rechazar  $\mu = \mu_0$ , y por tanto aceptar  $\mu \neq \mu_0$ . En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea  $\mu_0$ .

**Test de dos colas con varianza desconocida**

1.  $H_0 \equiv$  la media de la distribución es  $\mu = \mu_0$
2.  $H_1 \equiv$  la media de la distribución ,  $\mu \neq \mu_0$
3. Usamos como estadístico de contraste

$T_{n-1} = \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$ , que sigue una distribución  $t$  de Student con  $n - 1$  grados de libertad ( $n$  es el tamaño de la muestra,  $\bar{X}$  es la media de la muestra y  $S_c$  es la cuasidesviación de la muestra)

4. Si el nivel de significación es  $\alpha$ . La región de aceptación es

$$-t_{n-1, 1-\alpha/2} < \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} < t_{n-1, 1-\alpha/2} \quad (5.6)$$

5. Calcular  $\bar{X}$  y  $S_c$  a partir de la muestra y evaluar  $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$ .
6. Si el valor de  $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$  no cumple la relación 5.6, rechazar  $\mu = \mu_0$ , y por tanto aceptar  $\mu \neq \mu_0$ . En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea  $\mu_0$ .

**Test de hipótesis con una cola**

**Cola de la izquierda con varianza conocida** Este test compara las dos hipótesis siguientes:

1.  $H_0 \equiv$  la media de la distribución es  $\mu = \mu_0$
2.  $H_1 \equiv$  la media de la distribución es  $\mu < \mu_0$
3. Usamos como estadístico de contraste

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ , que sigue una distribución  $N(0, 1)$  ( $n$  es el tamaño de la muestra,  $\bar{X}$  es la media de la muestra)

4. Si el nivel de significación es  $\alpha$ . La región de aceptación es

$$-z_{1-\alpha} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (5.7)$$

5. Calcular  $\bar{X}$  a partir de la muestra y evaluar  $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ .



6. Si el valor de  $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$  no cumple la relación 5.7, rechazar  $\mu = \mu_0$ , y por tanto aceptar  $\mu < \mu_0$ . En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea  $\mu_0$ .

**Ejemplo 35** Supongamos que en el ejemplo de los automóviles enunciado en la página 172 queremos contrastar, al 95% de confianza, las dos hipótesis siguientes: a)  $H_0 \equiv$  la distancia media de frenado es 112.5 metros. b)  $H_1 \equiv$  la distancia media de frenado es menor que 112.5 metros.

Calculamos en primer lugar el valor del estadístico:

$$Z = \frac{\bar{X} - 112.5}{3/\sqrt{74}} = \frac{111.62 - 112.5}{3/\sqrt{74}} = -2.5233$$

La región de aceptación es:

$$-z_{1-\alpha} < Z \quad (5.8)$$

Ahora  $-z_{1-\alpha} = -z_{0.95} = -1.6449$ . Como el valor de  $Z$  no pertenece a la región de aceptación, porque pertenece a la cola izquierda, se rechaza la hipótesis nula con una confianza del 95% y se acepta la alternativa: La distancia media de frenado es menor que 112.5 metros.

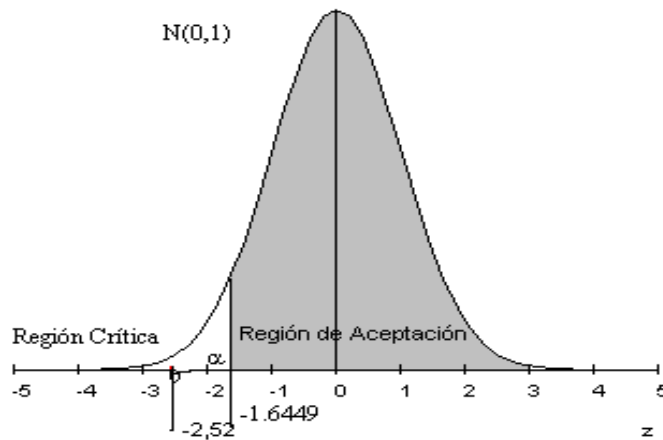


Figura 5.2: Test de una cola.

### Test de la cola izquierda con varianza desconocida

1.  $H_0 \equiv$  la media de la distribución es  $\mu = \mu_0$

2.  $H_1 \equiv$  la media de la distribución es  $\mu < \mu_0$

3. Usamos como estadístico de contraste

$T_{n-1} = \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$ , que sigue una distribución  $t$  de Student con  $n - 1$  grados de libertad ( $n$  es el tamaño de la muestra,  $\bar{X}$  es la media de la muestra y  $S_c$  es la cuasidesviación de la muestra)

4. Si el nivel de significación es  $\alpha$ . La región de aceptación es

$$-T_{n-1, 1-\alpha} < \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \quad (5.9)$$

5. Calcular  $\bar{X}$  y  $S_c$  a partir de la muestra y evaluar  $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$ .

6. Si el valor de  $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$  no cumple la relación 5.9, rechazar  $\mu = \mu_0$ , y por tanto aceptar  $\mu < \mu_0$ . En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea  $\mu_0$ .

### Test de la cola derecha

Sirve para comparar la hipótesis  $\mu = \mu_0$  contra la alternativa  $\mu > \mu_0$ . Se realizan con los mismos estadísticos que los de la cola izquierda con las siguientes modificaciones en el punto 4:

a) Si la varianza es conocida la región de aceptación de la hipótesis nula,  $\mu = \mu_0$ , es:

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$$

b) Si la varianza fuera desconocida la región de aceptación sería:

$$\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} < t_{n-1, 1-\alpha}$$

## 5.11 Distribuciones no normales

¿Como se reaalizan las pruebas de hipótesis si las poblaciones no son normales?

Los procedimientos anteriores se pueden aplicar para contrastar las hipótesis sobre las medias de variables aleatorias que se rijan por una función de distribución arbitraria siempre que las muestras utilizadas para realizar los contrastes sean suficientemente grandes.

a) Cuando se conoce  $\sigma$  basta con  $n > 30$ .

b) Si no se conoce  $\sigma$  se debe cumplir  $n > 100$ .

## 5.12 Prueba de hipótesis para una proporción

Se supone una población cuyos elementos son susceptibles de tomar dos caracteres cualitativos. Considérese, por ejemplo, en una ciudad los habitantes adultos que están desempleados y los que no lo están. Se desea contrastar la hipótesis  $H_0$  de que la proporción de elementos con una de esas características es  $p = p_0$  contra la hipótesis alternativa de que  $p \neq p_0$ . Para ello se usa una muestra de  $n$  elementos y se cuenta el número de ellos  $x$  que posea la citada característica. Así, si se trata el ejemplo de los habitantes de la citada ciudad se seleccionaría una muestra de  $n$  personas adultas y se contaría el número de entre ellas,  $x$ , que estén desempleadas. Tomamos este valor como el estadístico de contraste. Si se cumple la hipótesis nula, la variable aleatoria  $x$  se distribuye como una binomial  $B(n, p_0)$ . Según el teorema central del límite esta distribución puede aproximarse con una distribución  $N(np_0, \sqrt{np_0q_0})$  y la aproximación es razonable si  $np_0, nq_0 = n(1 - p_0) > 5$  y  $p_0, q_0 = 1 - p_0 > 0.05$ .

$$P(np_0 - z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0} < x < np_0 + z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0}) = 1 - \alpha$$

Por tanto el intervalo de aceptación para  $x$  es

$$(np_0 - z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0}, np_0 + z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0}) \quad (5.10)$$

Luego la prueba estadística que corresponde es la siguiente: Se cuenta el número de elementos,  $x$ , que posee la característica en cuestión. Si este valor quedara fuera del intervalo de aceptación dado en 5.10 se rechazaría la hipótesis nula y se concluiría que  $p \neq p_0$ . Si queda dentro del intervalo de aceptación se concluiría que no hay bastante evidencia para rechazar la hipótesis nula.

Otra forma de realizar este test es usar la proporción muestral  $\frac{x}{n}$  como estadístico de contraste. Sin más que dividir por  $n$  las relaciones anteriores concluimos que el intervalo de aceptación para la proporción muestral  $\frac{x}{n}$  es

$$\left( p_0 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0q_0}{n}}, p_0 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0q_0}{n}} \right)$$

**Ejemplo 36** *Constrastar la hipótesis de que la proporción de personas de una población con RH negativo es del 15% contra la hipótesis alternativa de que no es el 15%, con una confianza del 95%, sabiendo que se ha analizado la sangre de 400 personas elegidas aleatoriamente de esta población, obteniéndose que 72 de ellas tenía el RH negativo.*

Considerando como estadístico de contraste  $x = n^\circ$  de personas con Rh negativo, que si se cumple la hipótesis nula se distribuye aproximadamente como una

$$N(400 \times 0.15, \sqrt{400 \times 0.15 \times 0.85})$$

La aproximación es razonable, ya que:

$$\begin{aligned} np_0 &= 400 \times 0.15 = 60 > 5 \quad \text{y} \quad nq_0 = 400 \times 0.85 = 340 > 5 \\ p_0 &= 0.15 > 0.05 \quad \text{y} \quad q_0 = 0.85 > 0.05 \end{aligned}$$

El intervalo de aceptación de la hipótesis nula, viene dado por la expresión 5.10, que en este caso tomaría la forma:

$$(72 - 1.96\sqrt{400 \times 0.15 \times 0.85}, 72 + 1.96\sqrt{400 \times 0.15 \times 0.85}) = (58.003, 85.997)$$

Por lo tanto, con una confianza de 95%, se acepta la hipótesis nula de que el porcentaje de personas con RH negativo es de 15%, ya que el valor experimental, 72 personas, pertenece al intervalo de aceptación del test.

### 5.13 Prueba de hipótesis para la varianza

Frecuentemente, cuando se analizan variables cuantitativas, es importante sacar conclusiones en cuanto a la mayor o menor dispersión de la variable tratada. En este caso, lo que interesa es llegar a conclusiones acerca de la desviación estándar o la varianza de la población.

Si se intenta llegar a conclusiones en cuanto a la dispersión de la población, primero se debe determinar qué prueba estadística se puede emplear para representar la distribución de la dispersión de los datos de la muestra: Si la variable tiene distribución normal, según el teorema 17.3, el estadístico  $\frac{(n-1)S_c^2}{\sigma^2}$  sigue una distribución chi-cuadrado con  $(n-1)$  grados de libertad.

Suponiendo que la población de partida es normal de varianza desconocida  $\sigma^2$ , se formulan las siguientes hipótesis:

$$H_0 : \sigma^2 = \sigma_0^2.$$

$$H_1 : \sigma^2 \neq \sigma_0^2.$$

A partir de una muestra de  $n$  elementos se halla una estimación del estadístico  $\frac{(n-1)S_c^2}{\sigma_0^2}$ . Podemos tomar como intervalo de aceptación de la hipótesis nula, relativa a este estadístico, el intervalo de confianza dado en la expresión 5.4 de la página 168. Es decir, no rechazamos que  $\sigma^2 = \sigma_0^2$ , para un nivel de significación  $\alpha$ , si se cumple que

$$\frac{(n-1)S_c^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} < \sigma_0^2 < \frac{(n-1)S_c^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \quad (5.11)$$

Las pruebas referidas a varianzas son generalmente de cola superior, es decir que se contrasta la hipótesis nula,  $\sigma^2 = \sigma_0^2$ , contra la alternativa  $\sigma^2 > \sigma_0^2$ , ya que normalmente la preocupación está en el hecho de que la varianza

pueda ser demasiado grande. En este caso se tomará como intervalo de aceptación al nivel de significación  $\alpha$  :

$$\frac{(n-1)S_c^2}{\chi_{n-1,1-\alpha}^2} < \sigma_0^2$$

## 5.14 Prueba de bondad de ajuste

Los contrastes de hipótesis que hemos tratado hasta el momento son contrastes paramétricos: Se supone que la muestra pertenece a una población que se distribuye con arreglo a una cierta distribución conocida y se trata de estimar algunos parámetros de ésta. En esta ocasión, estudiaremos un test no paramétrico: No se supone que los datos de la muestra procedan de ninguna distribución concreta. Al contrario, nos preguntamos cuál podrá ser la distribución de la población de donde se han extraído éstos.

Una forma de decidirse por un modelo concreto es aplicar el *test chi-cuadrado de bondad de ajuste a distribuciones*. Este test realiza una comparación entre la distribución muestral y el modelo de distribución teórica a la que queremos probar si se ajustan los datos de la muestra. Puede aplicarse de la forma siguiente: Si se dispone de una muestra de  $n$  elementos (al menos 20, aunque este número puede variar según la precisión deseada), se agrupan en  $k$  clases, como en los histogramas de frecuencias. Si  $n_i$  es la frecuencia observada en la clase  $i$  y  $p_i$  la probabilidad que correspondería a este intervalo en la distribución teórica que deseamos contrastar,  $np_i$  sería el número de elementos que teóricamente debería caer en esta clase.

Para aceptar la validez del test suele exigirse que el valor asignado a cada clase teórica ( $np_i$ ) sea al menos 5.

Si se cumplen estas restricciones, y se cumple la hipótesis nula de que los datos proceden de la distribución teórica que estamos probando, el estadístico

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \quad (5.12)$$

sigue (aproximadamente) una distribución  $\chi^2$  con  $k-p-1$  grados de libertad, donde  $p$  es el número de parámetros de la distribución propuesta que se han estimado a partir de la muestra. El valor de  $\chi^2$  dado por la expresión 10.24 no puede ser negativo. Será nulo si hay un acuerdo perfecto entre los valores experimentales y teóricos. Es obvio que los valores experimentales se aproximan mejor a los teóricos cuanto más cerca de cero esté el valor  $\chi^2$ , esto es, cuanto menor sea la discrepancia entre los valores empíricos y los teóricos.

Rechazamos la distribución propuesta, al nivel de significación  $\alpha$ , si el estadístico es demasiado grande:  $\chi^2 > \chi_{k-p-1,1-\alpha}^2$ , siendo este último valor

el de chi-cuadrado con  $n - p - 1$  grados de libertad que deja a la derecha un área  $\alpha$  ( $P(\chi^2 < \chi_{k-p-1, 1-\alpha}^2) = 1 - \alpha$ ). La gráfica de la figura 5.3, muestra un esquema de la regiones de aceptación y de rechazo de este tipo de tests.

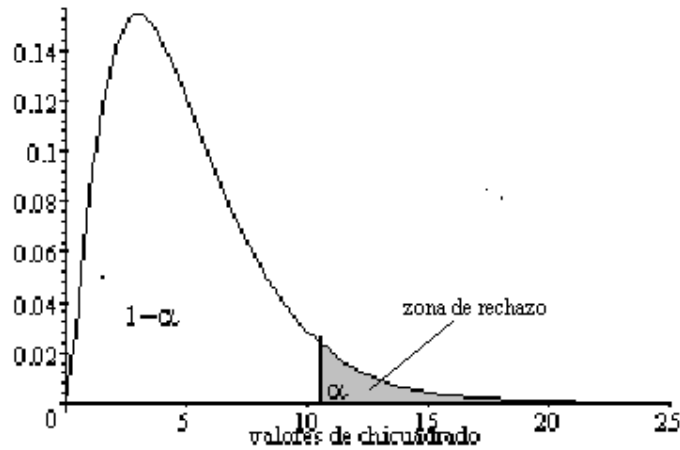


Figura 5.3: Test Chi-Cuadrado.

Si el número de clases,  $k$ , es menor o igual que 4, es preciso realizar en el estadístico  $\chi^2$  la corrección siguiente:

$$\chi^2 = \sum_{i=1}^k \frac{(|n_i - np_i| - 0.5)^2}{np_i}$$

Si el número de elementos no es suficiente para realizar este test puede emplearse el test de bondad de ajuste de *Kolmogorov-Smirnov* que, como el test chi-cuadrado, puede encontrarse implementado en la mayoría de los paquetes estadísticos.

**Ejemplo 37** *Se desea determinar si el número de errores cometen los secretarios de una cierta oficina en las hojas mecanografiadas sigue una distribución de Poisson con un número medio de 3 errores por hoja. Para ello se han contado los errores de 440 hojas seleccionadas al azar. Los valores observados se especifican en la siguiente tabla.*

Errores por hoja	0	1	2	3	4	5	6	7	8	9	Total
Frec. observada	18	53	103	107	82	46	18	10	2	1	440

A continuación mostramos la tabla que compara las frecuencias obser-

vadas con las frecuencia esperadas o teóricas si se supone que la distribución fuese realmente una Poisson de parámetro 3.

N. de errores	$n_i$ Frecuencias observadas	$p_i = \frac{3^i e^{-3}}{i!}$	$440 \times p_i$ Frecuencias esperadas
0	18	$4.9787 \times 10^{-2}$	21.906
1	53	0.14936	65.719
2	103	0.22404	98.578
3	107	0.22404	98.578
4	82	0.16803	73.934
5	46	0.10082	44.36
6	18	$5.0409 \times 10^{-2}$	22.18
7	10	$2.1604 \times 10^{-2}$	9.5058
8	2	$8.1015 \times 10^{-3}$	3.5647
9	1	$2.7005 \times 10^{-3}$	1.1882

En la tabla de  $np_i$  las dos últimas casillas tienen valores menores que 5. Sumamos entonces sus valores a la antepenúltima, para que no haya ninguna con valores menores que 5, teniendo ahora las dos últimas columnas el aspecto siguiente: Ahora sólo hay 8 clases, es decir  $k = 8$ .

N. de errores	$n_i$ Frecuencias observadas	$440 \times p_i$ Frecuencias esperadas
0	18	21.906
1	53	65.719
2	103	98.578
3	107	98.578
4	82	73.934
5	46	44.36
6	18	22.18
7 o más	13	14.2587

$$\chi^2 = \left[ \frac{3.906^2}{21.906} + \frac{12.719^2}{65.719} + \frac{4.422^2}{98.578} + \frac{8.422^2}{98.578} + \frac{8.066^2}{73.934} + \frac{1.64^2}{44.36} + \frac{4.18^2}{22.18} + \frac{1.2587^2}{14.2587} \right]$$

$$= 5.2813$$

Como  $\chi^2 = 5.2813$  es menor que  $\chi_{k-p-1, 1-\alpha}^2 = \chi_{8-0-1, 0.95}^2 = \chi_{7, 0.95}^2 = 14.0671$ , no se puede rechazar la hipótesis de que la distribución sea de Poisson al nivel de significación 0.05. Es decir que hay un buen acuerdo entre los datos experimentales y los teóricos que proceden de la Poisson de media 3.

*Nota:* Hemos tomado para  $p$  el valor 0, ya que es el único parámetro de la distribución de Poisson  $\lambda$ , número medio de errores por hoja. Este parámetro no se ha estimado a partir de la muestra, ya que venía dado como dato en el enunciado del ejemplo ( $\lambda = 3$ ).

## 5.15 Contrastes de hipótesis para dos poblaciones

### 5.15.1 Test para comparar la igualdad entre las varianzas de dos poblaciones

#### Muestras independientes

Queremos contrastar la hipótesis nula de que  $\sigma_1^2 = \sigma_2^2$

Si partimos de dos muestras independientes de tamaño  $n_1$  y  $n_2$  procedentes de dos distribuciones normales, el estadístico de contraste usado suele ser:

$$F = \frac{S_1^2}{S_2^2}$$

cociente entre las cuasivarianzas de estas dos muestras, que se distribuye como una  $F$  de Fisher-Snedecor con  $n_1 - 1$ ,  $n_2 - 1$  grados de libertad si se cumple la hipótesis nula de igualdad entre las varianzas poblacionales. Para contrastar, por ejemplo, la hipótesis nula  $\sigma_1^2 = \sigma_2^2$ , contra la alternativa  $\sigma_1^2 \neq \sigma_2^2$ , la región de aceptación de la hipótesis nula corresponde al intervalo

$$F_{\frac{\alpha}{2}} < \frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}$$

siendo  $F_p$  el percentil  $p$  de la distribución  $F$  de Fisher-Snedecor con  $n_1 - 1$ ,  $n_2 - 1$  grados de libertad cumpliendo

$$P(F < F_p) = p.$$

**Ejemplo 38** *Para contrastar la igualdad entre las varianzas de dos poblaciones normales consideramos el test cuya hipótesis nula es que ambas varianzas son iguales. La hipótesis alternativa supone que las varianzas son distintas. Para realizar el test se parte de dos muestras, una de cada población. Los datos sobre el número de elementos, la cuasidesviación y la media de las muestras de cada población son*

	número de elementos	media muestral	cuasidesviación
Muestra 1	$n_1 = 10$	$\bar{X}_1 = 6$	$S_1 = 0.1$
Muestra 2	$n_2 = 8$	$\bar{X}_2 = 5.2$	$S_2 = 0.3$



Para realizar el test se calcula el valor muestral del estadístico de contraste  $F = \frac{S_1^2}{S_2^2} = \frac{0.1^2}{0.3^2} = 0.111111$

Se concluirá, al nivel 0.05, que puede admitirse la igualdad de las varianzas si este valor,  $F = 0.111111$ , no es demasiado grande ni demasiado pequeño, es decir si está dentro de la región de aceptación del test:

$$F_{\frac{0.05}{2}} < F < F_{1-\frac{0.05}{2}} \quad (5.13)$$

donde se toma el valor de  $F$  con 9, 7 grados de libertad.

Con un programa de ordenador, se ha obtenido que

$$F_{\frac{0.05}{2}} = F_{0.025} = 0.23826$$

$$F_{1-\frac{0.05}{2}} = F_{0.975} = 4.8232$$

Como no se cumple la relación 5.13, concluimos que no puede aceptarse la igualdad entre las varianzas.

### Muestras pareadas

Las muestras pareadas son datos formados por pares de observaciones que guardan algún tipo de relación entre sí. Por ejemplo: Se desea comparar la eficacia de un plan de adelgazamiento. Con este objetivo se han pesado  $n$  personas antes de seguir el citado plan y después de un mes de haberlo seguido. En este caso no esperamos que las observaciones sobre el peso realizadas antes y después de seguir el plan de adelgazamiento no serán independientes, ya que son medidas del peso de la misma persona.

El estadístico de contraste para la hipótesis nula de igualdad de las varianzas es, para el caso de muestras pareadas:

$$T = \frac{|F-1|}{2} \sqrt{\frac{n-2}{F(1-r^2)}}$$

siendo  $F = \frac{S_1^2}{S_2^2}$ ,  $r$  el coeficiente de correlación entre ambas variables y  $n$  el número de pares de datos. Este estadístico se distribuye como una  $t$  de Student con  $n - 2$  grados de libertad si se cumple la hipótesis de igualdad entre las varianzas y las poblaciones de partida son normales .

La región de aceptación al nivel  $\alpha$  es  $T < t_{1-\alpha}$  donde  $t_p$  es el valor de la  $t$  de Student con  $n - 2$  grados de libertad que cumple :

$$P(t < t_p) = p.$$

**Ejemplo 39** La siguiente tabla nos da el peso en Kg. de 9 personas antes y después de haber seguido una misma dieta de adelgazamiento:

	Antes	Después
Persona 1	155	142
Persona 2	147	139
Persona 3	123	110
Persona 4	107	100
Persona 5	105	96
Persona 6	93	87
Persona 7	100	95
Persona 8	123	110
Persona 9	106	93

Comparar las varianzas de las poblaciones de procedencia de estas muestras (peso antes y después de la dieta).

$$F = \frac{S_1^2}{S_2^2} = \frac{455.25}{396} = 1.1496$$

$$r = 0.9829$$

$$T = \frac{|F-1|}{2} \sqrt{\frac{n-2}{F(1-r^2)}} = \frac{|1.1496-1|}{2} \sqrt{\frac{9-2}{1.1496(1-0.9829^2)}} = 1.0024$$

Este estadístico se distribuye como una  $t$  de Student con 7 grados de libertad si se cumple la hipótesis de igualdad entre las varianzas y las poblaciones son normales .

La región de aceptación al nivel 0.05 es  $(0, t_{7, 0.95}) = (0, 1.8946)$  que contiene el valor obtenido de la muestra  $T = 1.0024$ , por lo que no podemos rechazar la igualdad entre las varianzas.

El resultado del test no depende del orden de las muestras. Si tomamos el orden contrario obtenemos

$$F = \frac{S_2^2}{S_1^2} = \frac{396}{455.25} = 0.86985$$

y el estadístico  $T$  toma exactamente el mismo valor:

$$T = \frac{|0.86987-1|}{2} \sqrt{\frac{9-2}{0.86987(1-0.9829^2)}} = 1.0024$$

### 5.15.2 Comparación entre las medias de dos poblaciones normales

#### Muestras independientes (varianzas conocidas)

Para comparar las medias de dos poblaciones, de medias y desviaciones típicas  $\mu_1, \sigma_1$  y  $\mu_2, \sigma_2$  respectivamente, conociendo los valores de  $\sigma_1$  y  $\sigma_2$ , por

medio de sendas muestras independientes de tamaños  $n_1$ ,  $n_2$  y con medias muestrales  $\bar{X}_1$  y  $\bar{X}_2$ , se contrastan las hipótesis  $H_0 \equiv \mu_1 - \mu_2 = d$  y  $H_1 \equiv \mu_1 - \mu_2 \neq d$  usando el estadístico de contraste

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

El estadístico  $Z$  se distribuye con arreglo a una distribución normal estándar. La región de aceptación de la hipótesis nula,  $\mu_1 - \mu_2 = d$ , con una confianza de  $100(1 - \alpha)\%$  es

$$-z_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha/2}$$

donde el valor de  $z_{1-\alpha/2}$  se calcula en la tabla de la distribución  $N(0,1)$ . Es el valor que cumple:

$$P(Z < z_{1-\alpha/2}) = 1 - \alpha/2$$

#### Muestras independientes (varianzas desconocidas e iguales).

En este caso se comparan las medias de dos poblaciones, de medias y desviaciones típicas  $\mu_1, \sigma$  y  $\mu_2, \sigma$  respectivamente, por medio de sendas muestras independientes de tamaños  $n_1$ ,  $n_2$  y con medias y cuasidesviaciones muestrales  $\bar{X}_1, S_1$ , y  $\bar{X}_2, S_2$ . Considerando varianzas poblacionales desconocidas, pero iguales, se contrastan las hipótesis  $H_0 \equiv \mu_1 - \mu_2 = d$  y  $H_1 \equiv \mu_1 - \mu_2 \neq d$  usando el estadístico de contraste

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ siendo } S^2 = \frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1 + n_2 - 2}$$

El estadístico  $T$  se distribuye en esta ocasión como una  $t$  de Student de  $n_1 + n_2 - 2$  grados de libertad. La región de aceptación de la hipótesis nula,  $\mu_1 - \mu_2 = d$ , con una confianza de  $100(1 - \alpha)\%$  es

$$-t_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - d}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\alpha/2}$$

donde el valor de  $T_{1-\alpha/2}$  es el de la  $t$  de Student con  $n_1 + n_2 - 2$  grados de libertad que cumple:

$$P(t < t_{1-\alpha/2}) = 1 - \alpha/2$$

**Muestras independientes (varianzas desconocidas y desiguales)**

Si las varianzas no pueden suponerse iguales se usa el estadístico

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que se supone distribuida como una  $t$  de Student con los grados de libertad dados por el número entero más próximo al resultado que se obtenga en la expresión:

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1-1}\right)^2 + \left(\frac{S_2^2}{n_2-1}\right)^2} \quad (5.14)$$

**Ejemplo 40** Disponemos de dos muestras procedentes de poblaciones normales, independientes y con los siguientes datos:

	número de elementos	media muestral	cuasidesviación
Muestra 1	$n_1 = 10$	$\bar{X}_1 = 6$	$S_1 = 0.1$
Muestra 2	$n_2 = 8$	$\bar{X}_2 = 5.2$	$S_2 = 0.3$

Se desea emplear estos datos para contrastar si son iguales las medias de las poblaciones de partida.

En estos contrastes debe estudiarse en primer lugar si puede admitirse la igualdad de las varianzas. Para ello puede usarse el test dado en la sección 5.15.1. Este ejemplo ya se ha analizado en dicha sección habiendo concluido que, al nivel  $\alpha = 0.05$ , las varianzas se consideraban diferentes.

Constratemos ahora la igualdad entre las medias de ambas distribuciones, al 95% de confianza. Para ello se calcula el valor del estadístico de contraste:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(5.2 - 6) - 0}{\sqrt{\frac{0.01}{10} + \frac{0.09}{8}}} = -7.2281$$

que se compara con la  $t$  de Student cuyos grados de libertad se calculan con la fórmula dada por la expresión 5.14.

$$\frac{\left(\frac{0.01}{10} + \frac{0.09}{8}\right)^2}{\left(\frac{0.01}{9}\right)^2 + \left(\frac{0.09}{7}\right)^2} = \frac{1.5006 \times 10^{-4}}{1.6654 \times 10^{-4}} = 0.90106 \approx 1$$

Como  $t_{1,1-0.05/2} = t_{1,0.975} = 12.706$  y  $-12.706 < -7.2281 < 12.706$ , no se puede rechazar la igualdad entre las medias al 95% de confianza.

**Muestras pareadas**

Como ya hemos aclarado en el apartado 5.15.1 (página 185), las muestras pareadas son datos formados por pares de observaciones que guardan algún tipo de relación entre sí. En aquel caso se probaba la igualdad entre las varianzas. En este apartado contrastaremos los valores medios de ambas muestras. Por ejemplo: Se desea comparar la eficacia de un plan de adelgazamiento. Con este objetivo se han pesado  $n$  personas antes de seguir el citado plan y después de un mes de haberlo seguido. Para cada una de estas personas calculamos la diferencia,  $d_i =$  peso de la persona  $i$  antes de la dieta  $-$  peso de la persona  $i$  después de la dieta. Si suponemos que esta variable sigue una distribución normal, podemos contrastar la hipótesis de que la diferencia de los promedios de los pesos antes y después es  $\mu_1 - \mu_2 = d$  contra la alternativa de que esta diferencia es distinta de  $d$ .

El estadístico de contraste es el correspondiente a contrastar la media de la variable  $d_i$ . Suponiendo varianza desconocida y usando el criterio dado en la expresión 5.6 de la página 176, obtenemos la región de aceptación para la hipótesis  $\mu_1 - \mu_2 = d$

$$-t_{n-1, 1-\alpha/2} < \frac{\bar{d}_i - d}{S_c/\sqrt{n}} < t_{n-1, 1-\alpha/2} \quad (5.15)$$

Siendo  $\bar{d}_i$  la media de las diferencias de pesos de las  $n$  personas y  $S_c$  la cuasidesviación típica de las diferencias de pesos  $d_i$ .

**Ejemplo 41** *Deseamos comprobar si la eficacia de la dieta de adelgazamiento (ejemplo de la página 185) se refleja en una pérdida de peso de 7 Kg por término medio, o si por el contrario, los datos parecen apoyar la alternativa de que la pérdida media de peso ha sido mayor que 7 Kg.*

Las hipótesis a contrastar son:  $H_0 : d = 7$ ,  $H_1 : d > 7$

En este caso usamos el test de la cola derecha especificado en la sección 5.10.1. Si el nivel de significación es 0.05, la región de aceptación de la hipótesis nula es ahora

$$\frac{\bar{d}_i - d}{S_c/\sqrt{n}} < t_{9-1, 1-0.05} = 1.8595$$

En el ejemplo de la página 185, las diferencias de peso son los valores indicados en la columna  $d_i$ . Se supone que esta variable se distribuye normalmente.

	Antes	Después	$d_i$
Persona 1	155	142	13
Persona 2	147	139	8
Persona 3	123	110	13
Persona 4	107	100	7
Persona 5	105	96	9
Persona 6	93	87	6
Persona 7	100	95	5
Persona 8	123	110	13
Persona 9	106	93	13

El valor del estadístico de contraste para la muestra cumple:

$$\frac{\bar{d}_i - d}{S_c/\sqrt{n}} = \frac{9.66667 - 7}{3.3541/\sqrt{9}} = 2.3851$$

Este valor está en la zona de rechazo. Por tanto, entre las dos hipótesis se decide aceptar la hipótesis alternativa y se concluye que la dieta de adelgazamiento disminuye por término medio más de 7 Kg. de peso al cabo de un mes.

### 5.15.3 Test para la diferencia entre dos proporciones

Un ejemplo de este tipo de test se daría si quisiéramos comparar la proporción de parados en Andalucía con la proporción de parados en la Comunidad Europea. Para ello se seleccionarían aleatoriamente una muestra de personas en Andalucía y otra en Europa.

Sean  $\hat{p}_1$  la proporción de parados detectados en la muestra, de  $n_1$  personas seleccionadas en Andalucía y  $\hat{p}_2$  la proporción de parados detectados en la muestra, de  $n_2$  personas seleccionadas en toda la Comunidad Europea.

Para contrastar la hipótesis de que la proporciones verdaderas,  $p_1$  y  $p_2$  difieren en  $\mu = p_1 - p_2$  usamos el estadístico

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \mu}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

que supondremos distribuida como una  $N(0, 1)$  si el tamaño de la muestra es amplio.

### 5.15.4 Test de independencia de dos variables cualitativas

El test chi-cuadrado de bondad de ajuste, ya tratado en la sección 5.14, puede aplicarse para contrastar la hipótesis de que dos variables de clasificación de una población sean independientes contra la hipótesis alternativa de que no

lo sean. En este caso no se suponen conocidas las distribuciones de las muestras, ni se trata de determinar ningún parámetro. Se trata de un test no paramétrico.

Los datos de la muestra en que se va a basar el test se presentan en una tabla de contingencia.

**Ejemplo 42** *Se desea estudiar si el nivel de estudios realizados por las personas guarda alguna relación con su preferencia por pasar sus vacaciones en el campo o en la playa. Para ello se han entrevistado a 500 personas y se han clasificado con arreglo a estos criterios en la forma indicada en la siguiente tabla de contingencia*

OBSERVACIONES	Primarios	Medios	Universitarios
$C = \text{Campo}$	33	30	145
$Pl = \text{Playa}$	60	25	355
	93	55	Total=500

Realizamos un test de hipótesis tipo chi-cuadrado similar al descrito en el apartado 5.14. Consideraremos la hipótesis nula consistente en que ambas clasificaciones (estudios y preferencia para las vacaciones) no guardan relación alguna, es decir que los sucesos correspondientes a una y otra son independientes: Por ejemplo los sucesos consistentes en “haber realizado estudios primarios (Pr)” y “preferir el campo (C)” son independientes. La hipótesis alternativa es que ambas clasificaciones guardan algún tipo de relación o que son dependientes.

Nombramos cada suceso con sus iniciales, estimando la probabilidad de los sucesos por su frecuencia relativa y teniendo en cuenta su independencia (hipótesis nula) tendremos que, por ejemplo, para el suceso correspondiente a la primera casilla:

$$P(\text{Pr} \cap C) = P(\text{Pr}) \times P(C) = \frac{93}{500} \times \frac{145}{500} = 0.05394$$

Entonces, los valores esperados,  $np_i$ , para cada una de las 6 clases serían:

$$\begin{aligned} 500 \times P(\text{Pr} \cap C) &= 500 \times P(\text{Pr}) \times P(C) = 500 \times \frac{93}{500} \times \frac{145}{500} = 26.97 \\ 500 \times P(M \cap C) &= 500 \times P(M) \times P(C) = 500 \times \frac{55}{500} \times \frac{145}{500} = 102.08 \\ 500 \times P(U \cap C) &= 500 \times P(U) \times P(C) = 500 \times \frac{352}{500} \times \frac{145}{500} = 15.95 \\ 500 \times P(\text{Pr} \cap Pl) &= 500 \times P(\text{Pr}) \times P(Pl) = 500 \times \frac{93}{500} \times \frac{355}{500} = 66.03 \\ 500 \times P(M \cap Pl) &= 500 \times P(M) \times P(Pl) = 500 \times \frac{55}{500} \times \frac{355}{500} = 249.92 \\ 500 \times P(U \cap Pl) &= 500 \times P(U) \times P(Pl) = 500 \times \frac{352}{500} \times \frac{355}{500} = 39.05 \end{aligned}$$

En la siguiente tabla se resumen estos valores:

ESPERADOS	Primarios	Medios	Universitarios	
C = Campo	26.97	102.08	15.95	145
Pl = Playa	66.03	249.92	39.05	355
	93	352	55	Total=500

Procedemos ahora a calcular el valor de  $\chi_{\text{exp}}^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$ , que en este caso se compara con la  $\chi^2$ , cuyos grados de libertad se calculan de la siguiente forma: Si  $f$  y  $c$  son el número de características de cada criterio de clasificación, en este caso  $f = 2$  preferencias para las vacaciones y  $c = 3$  tipos de niveles de estudios, entonces el número de grados de libertad se calculan como  $(f - 1)(c - 1) = 1 \times 2 = 2$ .

$$\chi_{\text{exp}}^2 = \frac{(33-26.97)^2}{26.97} + \frac{(82-102.08)^2}{102.08} + \frac{(30-15.95)^2}{15.95} + \frac{(60-66.03)^2}{66.03} + \frac{(270-249.92)^2}{249.92} + \frac{(25-39.05)^2}{39.05} = 24.894$$

La región crítica, o de rechazo, es

$$\chi_{\text{exp}}^2 > \chi_{(f-1)(c-1), 1-\alpha}^2$$

$$\chi_{(f-1)(c-1), 1-\alpha}^2 = \chi_{2, 0.95}^2 = 5.9915$$

Como el valor experimental, 24.894, es mayor que el teórico, 5.9915, se rechaza la hipótesis nula, y se concluye que la muestra parece indicar que hay alguna relación entre el nivel de estudios y las preferencias sobre el lugar de vacaciones.

En el caso de las tablas con 4 clases hay que realizar una corrección en la fórmula empleada para la chi-cuadrado, ya indicada en el apartado 5.14 de la página 181, tal como se muestra en el ejemplo siguiente:

**Ejemplo 43** *Supongamos que se han seleccionado 90 personas adultas y se han clasificado por dos características: sexo y hábito de fumar. Los resultados de la clasificación están resumidos en la siguiente tabla:*

OBSERVACIONES	Hombre	Mujer	
Fuma	15	29	44
No fuma	27	19	46
	42	48	90

Se desea saber si estos datos apoyan la hipótesis nula de que no hay relación entre el sexo y el hábito de fumar o más bien apoyan la hipótesis contraria de que el sexo de una persona influye de alguna forma en el hábito de fumar. Los valores teóricos obtenidos bajo la hipótesis nula de independencia son:



$$\begin{aligned}
90 \times P(H \cap F) &= P(H) P(F) = 90 \times \frac{42}{90} \frac{44}{90} = 20.533 \\
90 \times P(M \cap F) &= P(M) P(F) = 90 \times \frac{48}{90} \frac{44}{90} = 23.467 \\
90 \times P(H \cap F') &= P(H) P(F') = 90 \times \frac{42}{90} \frac{46}{90} = 21.467 \\
90 \times P(M \cap F') &= P(M) P(F') = 90 \times \frac{48}{90} \frac{46}{90} = 24.533
\end{aligned}$$

ESPERADOS	Hombre	Mujer	
Fuma	20.534	23.467	44
No fuma	21.467	24.533	46
	42	48	90

Usando la corrección adecuada para este caso

$$\chi_{\text{exp}}^2 = \sum_{i=1}^k \frac{(|n_i - np_i| - 0.5)^2}{np_i} \quad (5.16)$$

$$\begin{aligned}
\chi_{\text{exp}}^2 &= \frac{(|15-20.534|-0.5)^2}{20.534} + \frac{(|29-23.467|-0.5)^2}{23.467} + \frac{(|27-21.467|-0.5)^2}{21.467} + \frac{(|19-24.533|-0.5)^2}{24.533} = \\
&= 4.5261
\end{aligned}$$

Los grados de libertad se calculan como el producto del número de niveles de una de las características menos uno y el número de niveles de la otra característica menos uno,  $(2-1)(2-1) = 1$ . El valor de la chi-cuadrado con 1 grado de libertad correspondiente a una probabilidad 0.95 es  $\chi_{1, 0.95}^2 = 3.8415$ .

En este caso la chi-cuadrado experimental, 4.5261, supera al valor teórico, 3.8415. Se rechaza la hipótesis de independencia. Concluimos por tanto que el sexo tiene alguna relación con el hábito de fumar. Los datos parecen apoyar la idea de que las mujeres fuman más que los hombres.

## 5.16 EJERCICIOS PROPUESTOS

**Ejercicio 62** La media de una muestra de 36 elementos de una distribución normal es 4.1. Hallar un intervalo de confianza al 95% para la media. (La desviación típica de la población es 3).

**Ejercicio 63** Se ha repetido un experimento físico 9 veces obteniéndose una media de los valores medidos de 42.319 y una cuasi-desviación típica de 5.0. Estimar el valor real de la magnitud con una confianza del 95 por 100.

**Ejercicio 64** Para probar si una moneda es defectuosa (la cara y la cruz no tienen la misma probabilidad) se recurre al siguiente ensayo. Se tira la moneda 100 veces y se declara defectuosa si el número de caras es un número fuera del intervalo  $[40, 60]$ .

1. Calcular la probabilidad de declarar la moneda como defectuosa una moneda correcta (error tipo I del test de hipótesis)
2. Calcular la probabilidad de declararla correcta si la probabilidad de sacar cara fuera: a) 0.6, b) 0.65, c) 0.70, d) 0.80.

**Ejercicio 65** Diseñar una prueba de hipótesis (al 95% de confianza) para la longitud media de una serie de tornillos basada en muestras de 9 elementos, que permita rechazar los lotes cuya longitud media no sea 5 mm. La longitud de estos tornillos se distribuye según una normal de desviación típica  $\sigma = 2$  mm.

**Ejercicio 66** Un vendedor de bandas elásticas afirma que resisten un estiramiento promedio de 180Kg. Se ha hecho una prueba con 5 de estas banda observandose una resistencia promedio de 169.51Kg. con una cuasi desviación de 5.7 kg

1. ¿Se rechazaría al 99% de confianza la media de resistencia indicada por el vendedor.
2. ¿Cual es la región de rechazo para la resistencia promedio de la muestra? ¿Y el valor Crítico?

**Ejercicio 67** Un tipo de botes de pintura esta declarada como apta para pintar un promedio de 80 m<sup>2</sup> con una desviación típica de 8.4 m<sup>2</sup>. Se desea comprobar si puede aceptarse este valor promedio. Con este objetivo se ha decidido probar 100 de estos botes y rechazar la pintura si el promedio de superficie pintada resultará menor que 78 m<sup>2</sup> Se aceptará el valor de la desviación típica.

1. Calcular el nivel de confianza y la significación de esta prueba.
2. Si la pintura pintara ralmente un promedio de 79 m<sup>2</sup> cual sería la probabilidad de no rechazar la media indicada por el fabricante.
3. ¿Y si el promedio fuera de 75 m<sup>2</sup>

**Ejercicio 68** Un vendedor de neumáticos dice que la vida media de sus neumáticos es de 28000 Km. Admitiendo para la desviación típica el valor 1348 Km. diseñar un test de hipótesis al 99% de confianza, basado en muestras de 40 elementos que permita contrastar la hipótesis nula de ser  $\mu = 28000$ Km usando como hipótesis alternativa  $\mu < 28000$ Km

**Ejercicio 69** Si de un total de 100 personas entrevistadas 36 han afirmado que conocen una cierta marca de detergente

1. Hallar un intervalo de confianza al 95% para la proporción real de personas que conocen este detergente.
2. ¿Cuántas personas se precisan entrevistar para que el intervalo de confianza para la proporción tenga una amplitud de 0.1?

**Ejercicio 70** Se desea saber la proporción de personas de una gran ciudad que encuentran adecuado el transporte público. ¿Cuántas personas hay que entrevistar si se desea estimar esta proporción con un intervalo de confianza de 95% y un error de precisión menor del 6%?

**Ejercicio 71** Encuestadas 267 personas ha resultado que 114 de ellas encuentran satisfactorio el transporte público. Dar un intervalo de confianza para la proporción de personas que encuentran satisfactorio este tipo de transporte. (95% de confianza)

**Ejercicio 72** 32 medidas del punto de ebullición del azufre tienen una cuasidesviación de 0.83 grados. Calcular un intervalo de confianza para la varianza con una confianza del 98%

**Ejercicio 73** Las piezas de una maquina deben ser del mismo tamaño, por eso se exige que la desviación típica de la población sea 0.05 mm. Diseñar un test al 95% de confianza para contrastar la hipótesis de que  $\sigma = 0.05$  mm. con muestras de 15 elementos

**Ejercicio 74** Se ha llevado a cabo un estudio para determinar si hay diferencia entre el tiempo que tardan los hombres y las mujeres en hacer determinada maniobra en una línea de ensamble. Los valores obtenidos en el estudio se resumen en la siguiente tabla

	Nº de elementos	media muestral	Varianza poblacional
hombres	50	42 seg.	18 seg <sup>2</sup>
mujeres	50	38 seg	14 seg <sup>2</sup>

¿Es significativa la diferencia de rendimiento entre hombres y mujeres?

**Ejercicio 75** Un fabricante asegura que sus fusibles, con una sobrecarga del 20%, se fundiran por promedio al cabo de 12.40 min. Una muestra de 20 fusibles se sobrecarga un 20%, obteniendose una media de 10.63 y una cuasidesviación de 2.48 min. ¿Confirma la muestra la afirmación del fabricante para el promedio?

**Ejercicio 76** Se han recogido muestras de aire para estudiar su contaminación, obteniéndose las siguientes cantidades de impurezas en  $\frac{Kg}{m^3}$

2.2; 1.8; 3.1; 2.0; 2.4; 2.0; 2.1; 1.2

Dad un intervalo de confianza al 95% para la media de impurezas contenidas en el aire

**Ejercicio 77** El director de un colegio quiere saber el tiempo medio que tardan los alumnos en cambiar de clase, con una confianza del 99% y un error que no sobrepase 0.25 minutos. Si se puede suponer que el valor de  $\sigma$  es 1.40 minutos, ¿Cuál debe ser el tamaño de la muestra?

**Ejercicio 78** Se realizó un muestreo para decidir si los sueldos de los peones de albañil de una ciudad A y de otra B son iguales por promedio o no. Para ello se consulto a 100 peones de la ciudad A y a 150 de la ciudad B. Analizadas la respuestas realizadas por dichos operarios se determino que la media de los sueldos de los 100 operarios de la ciudad A era de 760 € y la de los 150 empleados de ciudad B era de 720 €. Suponiendo que la desviación típica poblacional de los sueldos de A es 12€ y la de B 9 €, decidir si el sueldo medio en ambas ciudades es igual o distinto.

**Ejercicio 79** Se desea comparar el gasto medio mensual en alimentación entre las familias de dos barrios. Para ello se seleccionaron 20 familias de cada barrio, observando sus gastos mensuales en alimentación. Se determino la media y las cuasidesviaciones típicas, obteniéndose los siguientes resultados muestrales:  $(\bar{X}_1 = 200, S_1 = 20, \bar{X}_2 = 175, S_2 = 17)$ . Suponiendo que los gastos se distribuyen normalmente decidir sobre la cuestión planteada. Los gastos medios en alimentación entre ambos barrios, ¿pueden considerarse iguales?

**Ejercicio 80** Mendel sembró 532 plantas de guisantes usando semillas del mismo tipo y los frutos resultantes los clasificó atendiendo al color en: verde, verde amarillento y amarillo y atendiendo a la forma: redondo, levemente rugoso y rugoso. Obtuvo los siguientes datos:

	Verde	Verde-Amarillo	Amarillo	
Redondo	35	68	38	141
Levemente Rugoso	67	138	60	265
Rugoso	30	68	28	126
	132	274	126	532

¿Había alguna relación de dependencia entre la forma y el color de esos guisantes?

**Unidad Temática II**

**CONTROL DE CALIDAD**



## Tema 6

# Control de Calidad. Control por atributos.

### 6.1 Introducción.

Entendemos por calidad la adaptación de un producto a su uso. El producto ha de ser fiel a su diseño. El objetivo de esta unidad es el control estadístico de la calidad como un elemento útil en la mejora de ésta.

El control de calidad se realiza no solo a la entrada y salida de los productos de la fábrica, sino también durante todo el proceso de fabricación. Suele realizarse observando no toda la producción, sino muestras de ésta. En cada elemento se controla bien una característica medible: longitud, peso, proporción de impurezas etc.. que se comparan con medidas estándar, o simplemente se clasifica el elemento en defectuoso o no. En el primer caso se llama *control por variables* y en el segundo *control por atributos*. Este segundo procedimiento es más rápido, pero da menos información.

El *control de recepción* se aplica a los materiales de entrada para comprobar que cumplen las especificaciones establecidas. El *control de fabricación* se realiza en intervalos de tiempo regulares a lo largo de todo el proceso. Los resultados se registran para poder realizar estudios posteriores. Esto no solo permite desechar partidas defectuosas, que no merece la pena que continúen el proceso de fabricación completo, con lo cual se consigue evitar un gasto suplementario e inútil de tiempo, material y mano de obra, sino también mejorar la calidad, pues de esta forma se pueden detectar fallos en el sistema y proceder a la correcciones oportunas. Por paradójico que parezca, resulta que lo más barato es *no producir* artículos defectuosos.

En un proceso de fabricación siempre hay una variabilidad en los productos. Esta variabilidad es atribuible a una gran cantidad de causas: ligeras variaciones en los materiales, desajuste de la maquinaria, destreza de

los empleados, cambios de temperatura. . . Algunas de las causas de la variabilidad pueden ser inevitables o no merece la pena corregirlas por motivos económicos, de seguridad y simplemente las asumimos. Se llaman causas no asignables se caracterizan por ser muchas, de poca importancia y producir una variabilidad aceptada y que ha de ser constante a lo largo del proceso. Por lo general si una pieza es defectuosa por alguna de estas causas que hemos llamado no asignables no tienen porqué ser defectuosas las siguientes. Por el contrario otras causas de la variabilidad producen variaciones importantes en los productos: rotura de una maquina, indisposición o cansancio de un operario. . . Estas causas se llaman asignables y por lo general producen una pérdida de calidad continuada.

Se dice que *el proceso está bajo control* si las únicas causa que producen variabilidad son las no asignables, y por tanto la variabilidad es constante y coincide con la aceptada. Mantener el proceso bajo control es una ocupación constante en un proceso industrial y es uno de los objetivos del control de calidad, que en realidad tiene como objetivo *obtener un producto de la máxima calidad al menor costo y lo antes posible*

## 6.2 Control por atributos. Capacidad

Se aplica principalmente cuando se quiere controlar si los productos tienen o no tienen ciertos atributos: Una lampara se enciende o no, dos piezas encajan bien o no. . . , pero también puede utilizarse como primer paso en el control de variables detectando errores muy grandes en la longitud, peso. . . de los productos indicando si es aceptable o no, en una apreciación poco fina. Tiene la ventaja de ser un control más rápido y barato.

En el control por atributos se revisan muestras de  $n$  elementos de la producción, clasificando estos elementos como defectuosos o no.

Se llama capacidad del proceso a  $1 - p$ , siendo  $p$  la proporción de defectuosos fabricados con el proceso bajo control: El valor de  $p$  es estable y la aparición de piezas defectuosas consecutivas es independiente de las que hayan aparecido antes.

La distribución del número de artículos defectuosos en muestras de tamaño  $n$  es la *Binomial*( $n, p$ )

Una muestra concreta nos va a indicar salida de control si el número de elementos defectuosos está fuera del intervalo  $(np - 3\sqrt{npq}, np + 3\sqrt{npq})$  . Los extremos de este intervalo se llaman límites de control. Este intervalo tiene un nivel de confianza de 99.74% ya que  $P(|z| \leq 3) = 0.9974$  (usamos la aproximación de la binomial por la normal). Este es el valor estandar que se utiliza, pero puede elegirse otros niveles personales usando la distribución normal si es posible ó la distribución binomial en caso contrario. A veces



en lugar de considerar el número de defectos de la muestra se considera la proporción de elementos defectuosos. En este caso los límites de control para la proporción de elementos defectuosos se obtiene del anterior dividiendo por  $n$ :

$$\left( p - 3\sqrt{\frac{pq}{n}}, p + 3\sqrt{\frac{pq}{n}} \right) \quad (6.1)$$

Para establecer la capacidad de un proceso que está bajo control se procede de la forma siguiente:

1. Se toman  $k$  muestras de  $n$  elementos cada una (se suele recomendar que  $K$  sea al menos 25 y  $n$  al menos 50). Las muestras se toman a intervalos regulares durante un periodo amplio de tiempo para que puedan influir todas las causas de variabilidad no asignable. Con estos valores se calcula una estimación de  $p$ :

$$\hat{p} = \frac{d_1 + d_2 + \dots d_k}{kn}$$

siendo  $d_i$  el número de defectuosos en la muestra  $i$ . Sustituyendo este valor estimado en la  $p$  del intervalo de confianza dado en la fórmula 6.1.

2. Se comprueba si las muestras han permanecido bajo control en todo el proceso. Si no es así se eliminan las muestras que se salen de estos límites y se vuelve a calcular  $p$  y así sucesivamente hasta que todas las muestras estén bajo control.

**Ejemplo 44** (*proporción de elementos defectuosos*). Se muestrea durante cuatro horas un proceso que produce transistores. Se seleccionan 24 muestras. Cada muestra consta de 50 transistores. El número de artículos defectuosos en la muestra viene expresado por  $x_i$  en la siguiente tabla. Elaborar una gráfica de control basado en estas muestras.

Muestra	$x_i$	$x_i/n$	Muestra	$x_i$	$x_i/n$
1	3	0.06	13	1	0.02
2	1	0.02	14	2	0.04
3	4	0.08	15	0	0.00
4	2	0.04	16	3	0.06
5	0	0.00	17	2	0.04
6	2	0.04	18	2	0.04
7	3	0.06	19	4	0.08
8	3	0.06	20	1	0.02
9	5	0.10	21	3	0.06
10	4	0.08	22	0	0.00
11	1	0.02	23	2	0.04
12	1	0.02	24	3	0.06

El valor estimado de  $p$  es en este caso:

$$\hat{p} = \frac{3 + 1 + 4 + 2 + \dots + 2 + 3}{50 \times 24} = 0.04$$

Un intervalo de confianza para la proporción sería:

$$\left( p - 3\sqrt{\frac{pq}{n}}, p + 3\sqrt{\frac{pq}{n}} \right) = \left( 0.04 - 3\sqrt{\frac{0.04 \times 0.96}{50}}, 0.04 + 3\sqrt{\frac{0.04 \times 0.96}{50}} \right) = (-0.04, 0.12)$$

Tomamos como intervalo de control para la proporción  $(0, 0.12)$

Si se desea decir el intervalo para el número de elementos defectuosos se multiplica por 50 obteniéndose que el número de artículos defectuosos de cada 50 deben ser 6 por término medio. Como en este caso ninguna de las muestras queda por encima de los límites de control hemos acabado el proceso.

## 6.3 Gráficos de control.

### 6.3.1 Fracción de defectos ó número de defectos. Interpretación

Todo este proceso puede condensarse en una gráfica que se puede realizar con el programa Statgraphics. El gráfico puede hacerse por proporción o por número de defectos. En el ejemplo siguiente se comparan las muestras sucesivas con el valor estándar para la media de elementos defectuosos.

### 6.3.2 Un ejemplo de gráficos de Control con Statgraphics

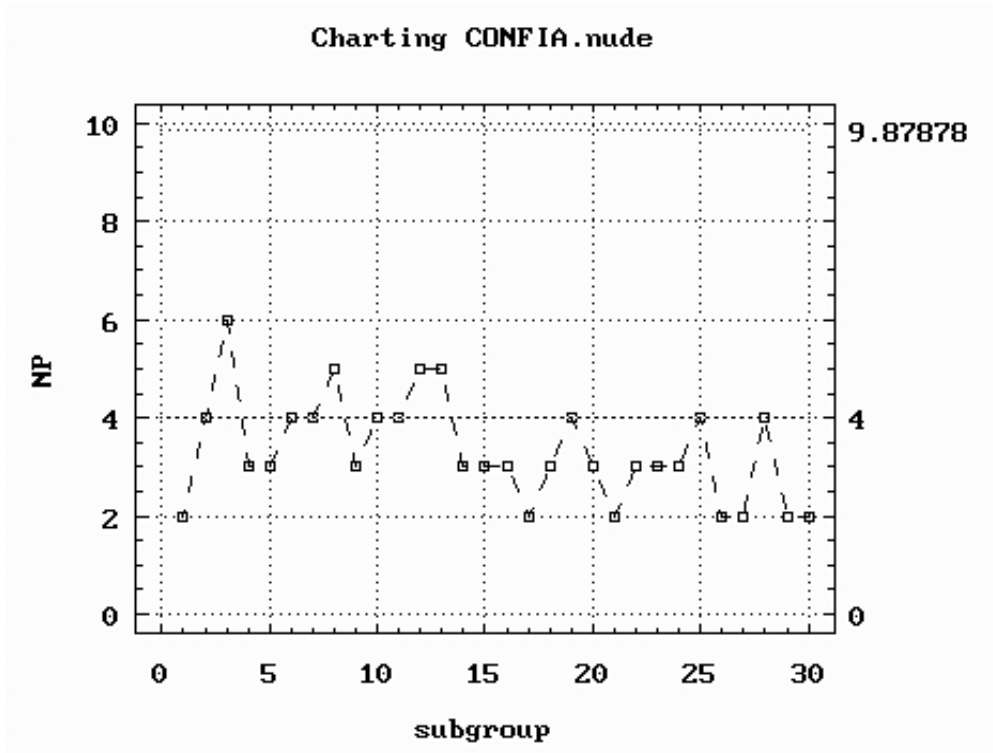
**Ejemplo 45** *En la fabricación de ciertas hojas de madera se admite un promedio de 4 hojas defectuosas. Se han seleccionado 30 muestras de estas hojas de tamaño 100, y se ha contado el número de hojas defectuosas de cada una de estas muestras. (Los resultados se encuentran en la variable CONFIA.nude). Constrúyase un diagrama para realizar el control de calidad de la fabricación.*

Seleccionamos los menús:

ESPECIAL → QUALITY CONTROL → ATTRIBUTES CONTROL  
CHARTS

En la pantalla siguiente seleccionamos con la barra espaciadora **np counts** y **control to standard** y damos F6. tras esto se nos piden las observaciones

Introducimos la variable nude, el número de elementos de cada muestra (100) en el siguiente campo. Con análisis Options elegimos control to estandar que es 4. Con ello nos aparece un gráfico similar al siguiente:



En el diagrama se observa que el proceso está bajo control porque el número de defectos está siempre por debajo del límite de control, que es 9.87878

Este límite de control puede obtenerse con la expresión:

$$np + 3\sqrt{npq} = 4 + 3\sqrt{100 \times 0.04 \times 0.96} = 9.87878$$

A veces se controla el número de defectos por pieza. Se suele emplear la distribución de Poisson y los límites de control son entonces

$$\left( \lambda - 3\sqrt{\lambda}, \lambda + 3\sqrt{\lambda} \right)$$

## 6.4 EJERCICIOS PROPUESTOS

**Ejercicio 81** Se tomaron 25 muestras con 100 lamparas cada una, conteniendo los siguientes números de defectos:

3, 4, 6, 4, 0, 5, 2, 3, 0, 2, 3, 5, 3, 9, 1, 2, 4, 4, 1, 8, 4, 6, 5, 3, 2

1. ¿Se puede aceptar que este proceso produce un 2% de defectuosos por promedio como afirma el fabricante?
2. Con los datos anteriores calcular unos nuevos límites de control para la fracción de lamparas defectuosas
3. ¿Cual debe ser la proporción media de defectos que debe dar el fabricante?

**Ejercicio 82** Durante la fabricación de piezas de un aparato eléctrico se han tomado muestras de 50 elementos cada 4 horas. Se han registrado la cantidad de elementos defectuosos entre estas 50 piezas: 3, 3, 2, 0, 6, 1, 1, 1, 2, 1, 2, 3, 3, 0, 8, 0, 6, 5, 5, 0, 3, 3, 2, 1, 3, 4, 5, 3, 4, 5, 4, 6, 1. Utilizar estos datos para establecer unos valores de control para la proporción de elementos defectuosos que produce este proceso.

**Ejercicio 83** Los siguientes datos son el número de soldaduras defectuosas encontradas en sucesivas muestras de 500 juntas soldadas: 106, 116, 164, 89, 99, 40, 112, 36, 69, 74, 42, 37, 25, 88, 101, 64, 51, 74, 71, 43, 80. ¿Está el proceso bajo control.?

**Ejercicio 84** En una planta industrial se encapsulan las botellas de una bebida refrescante. Cada 5 horas se seleccionan 64 de estas botellas para comprobar si la operación se ha realizado correctamente, resultando que la proporción media de botellas defectuosas ha sido del 2 %.

1. Suponiendo que no hay muestras fuera de control, definir los límites de control para el número de botellas defectuosas de cada muestra de 64 botellas.
2. Si una persona compra 12 de estas botellas ¿Cuál es la probabilidad de que ninguna de ellas sea defectuosa?
3. Un cliente de esta fabrica, que no sabe cuantas defectuosas se producen decide que no las adquirirá si en una muestra de 100 de ellas el número de defectuosas es mayor que 1. ¿Cual es la ordenada, para  $p = 0.05$ , de la curva característica correspondiente.

4. Además este cliente decide que si durante la siguiente semana (7 días) hay más de dos días en que no puede adquirir las botellas dejará de confiar en este proveedor y ya no le comprará más. Calcular la probabilidad que tiene la fábrica de perder este cliente.

**Ejercicio 85** 25 muestras sucesivas de 200 interruptores tomadas de una línea de producción contuvieron respectivamente 6, 7, 13, 7, 0, 9, 4, 6, 0, 4, 5, 11, 6, 8, 18, 1, 4, 9, 8, 2, 17, 9, 12, 10, 5 piezas defectuosas. Se quiere mantener la fracción de piezas defectuosas en 0.02. Elabore un diagrama con estos datos para indicar si se cumple esta norma o no.

**Ejercicio 86** Los siguientes valores

18, 15, 23, 9, 27, 19, 22, 21, 25, 14,  
19, 26, 11, 28, 22, 14, 25, 17, 23, 18

son el número de unidades defectuosas en muestras de 200 componentes electrónicos de los producidos en un cierto proceso.

1. Estimar a partir de estas muestras una estimación para la proporción de componentes electrónicos defectuosos que produce este proceso.
2. Calcular los límites de control para esta proporción a partir de estas muestras
3. Calcular la capacidad del proceso

**Ejercicio 87** Se inspeccionan las botellas de plástico para un detergente líquido. Se toman 20 muestras cada una con 100 botellas, notificándose la fracción de defectuosas de cada muestra. Los datos aparecen a continuación

Muestra	Fracción de defectuosas	Muestra	Fracción de defectuosas
1	0.12	11	0.13
2	0.15	12	0.07
3	0.18	13	0.12
4	0.10	14	0.08
5	0.12	15	0.09
6	0.11	16	0.15
7	0.05	17	0.10
8	0.09	18	0.06
9	0.13	19	0.12
10	0.10	20	0.13

1. Estimar a partir de estas muestras un valor para la proporción de botellas defectuosas que se fabrican en este proceso

206 TEMA 6. CONTROL DE CALIDAD. CONTROL POR ATRIBUTOS.

2. Calcular los límites de control a partir de estas muestras
3. Calcular la capacidad del proceso

**Ejercicio 88** Dentro de un proyecto de mejora de la calidad, una industria textil decide controlar el número de imperfecciones encontradas en cada pieza de tela. Se estima que el número promedio de imperfecciones por cada pieza de tela es de 12. Calcular la probabilidad de que en una de estas piezas de tela fabricada se encuentren.

1. Entre 10 y 12 imperfecciones.
2. Menos de 8 y más de 16 imperfecciones.
3. Inspeccionada un lote de 25 piezas de tela, se han encontrado los siguientes números de defectos:  
13, 15, 9, 7, 12, 8, 4, 10, 3, 5, 8, 14, 10, 11, 14, 15, 7, 16, 8, 8, 9, 14, 17, 13, 9. ¿Se mantiene el número promedio de defectuosos 12?  
Realizar la gráfica de control

## Tema 7

# Control por variables

### 7.1 Control por variables según una distribución Normal.

Suponemos, para empezar, que el proceso está bajo control. Es decir, que se fabrican productos en que la característica de interés tiene media  $\mu$  y de desviación típica  $\sigma$ . Por lo general hay que comprobar que ambos parámetros se mantienen. El control se realiza tomando muestras de  $n$  elementos cada cierto tiempo. Se supone que el proceso está bajo control si la media de cada muestra esta dentro de un cierto intervalo de confianza  $\left(\mu - 3\frac{\sigma}{\sqrt{n}}, \mu + 3\frac{\sigma}{\sqrt{n}}\right)$ . Estos serían los límites de control teóricos. El nivel de confianza es, como en el caso del diagrama de Control por atributos, 99.74%.

Para calcular los parámetros de control a partir de muestras sucesivas de un proceso que se supone que esta bajo control se actúa del siguiente modo:

Paso 1 Se seleccionan  $k$  muestras (se recomienda  $k \geq 20$ ) de  $n$  elementos en intervalos de tiempo regulares.

Paso 2 Para cada una de ellas se calcula la media,  $\bar{x}_j$  y la cuasidesviación muestral,  $s_j$ , o la desviación típica de la muestra,  $S_j$ , o el recorrido,  $r_j$ , (diferencia entre los valores mayor y menor de la muestra). El valor de  $\mu$  se estima como

$$\hat{\mu} = \bar{X} = \frac{\sum_{j=1}^k \bar{x}_j}{k}$$

Si se han calculado las cuasidesviaciones muestrales,  $\sigma$  se estima con

$$\hat{\sigma} = \frac{1}{c_4} \bar{s} = \frac{1}{c_4} \frac{\sum_{j=1}^k s_j}{k} \quad (7.1)$$

Si se han calculado las desviaciones típicas de la muestra, como se cumple la relación:

$$(n-1)s^2 = nS^2 \implies s = S\sqrt{\frac{n}{n-1}} \quad (7.2)$$

obtenemos de las expresiones 7.1 y 7.2, la siguiente estimación de  $\sigma$ , usando los valores de las desviaciones típicas de las muestras:

$$\hat{\sigma} = \frac{1}{c_4} \frac{\sum_{j=1}^k s_j}{k} = \frac{1}{c_4} \frac{\sum_{j=1}^k \sqrt{\frac{n}{n-1}} S_j}{k} = \frac{1}{c_4} \sqrt{\frac{n}{n-1}} \frac{\sum_{j=1}^k S_j}{k} = \frac{1}{c_4 \sqrt{\frac{n-1}{n}}} \bar{S} = \frac{1}{c_2} \bar{S}$$

por tanto  $c_2 = c_4 \sqrt{\frac{n-1}{n}}$  y  $c_4 = c_2 \sqrt{\frac{n}{n-1}}$

Si se hubieran calculado los recorridos muestrales la estimación de la desviación típica muestral es:

$$\hat{\sigma} = \frac{1}{d_2} \bar{r} = \frac{1}{d_2} \frac{\sum_{j=1}^k r_j}{k}$$

Los valores de  $c_4$ ,  $c_2$  y  $d_2$ , que dependen del tamaño de la muestra, sirven para transformar  $\bar{s}$ ,  $\bar{S}$  y  $\bar{r}$ , en estimadores centrados de la desviación típica. Estos factores, así como otros que irán apareciendo a lo largo del capítulo pueden evaluarse usando la tabla 7.1 que aparece en la página 220

En el primer caso los límites de control para la media se calculan con la siguiente expresión:

$$\left( \bar{X} - 3 \frac{\bar{s}}{c_4 \sqrt{n}}, \bar{X} + 3 \frac{\bar{s}}{c_4 \sqrt{n}} \right) = (\bar{X} - A_3 \bar{s}, \bar{X} + A_3 \bar{s})$$

siendo  $A_3 = 3 \frac{1}{c_4 \sqrt{n}}$ .

Para el segundo caso, en el que ha estimado  $\sigma$  a partir de las desviaciones típicas de la muestra los límites de control para las medias muestrales son:

$$\left( \bar{X} - 3 \frac{\bar{S}}{c_2 \sqrt{n}}, \bar{X} + 3 \frac{\bar{S}}{c_2 \sqrt{n}} \right) = (\bar{X} - A_1 \bar{S}, \bar{X} + A_1 \bar{S}) \quad (7.3)$$

siendo  $A_1 = 3 \frac{1}{c_2 \sqrt{n}}$

Si se han calculado los recorridos muestrales los límites de control para las medias de la muestra serán:

$$\left( \bar{X} - 3 \frac{\bar{r}}{d_2 \sqrt{n}}, \bar{X} + 3 \frac{\bar{r}}{d_2 \sqrt{n}} \right) = (\bar{X} - A_2 \bar{r}, \bar{X} + A_2 \bar{r})$$



7.1. CONTROL POR VARIABLES SEGÚN UNA DISTRIBUCIÓN NORMAL.209

Del mismo modo el valor  $A_2 = 3\frac{1}{d_2\sqrt{n}}$ .

Como ya se indicó previamente, todos estos parámetros pueden obtenerse de la tabla 7.1, eligiendo la fila correspondiente al valor  $n$  que representa el número de elementos de cada una de las  $k$  muestras.

Paso 3 Si han quedado muestras fuera de los límites de control calculados, se excluyen estas muestras. Una vez rechazadas la muestras que han quedado fuera de control se vuelve repetir el cálculo de los nuevos parámetros y se repite el proceso hasta que no haya muestras fuera de control.

**Ejemplo 46** *Deseamos iniciar una gráfica de control para una maquina nueva que llena (por peso) cajas de cereal. Se hacen observaciones cada dos horas de la cantidad de llenado hasta obtener 20 muestras. Los resultados aparecen en la tabla. Se pide calcular los límites inferior y superior de control de la media. Se supone que los precios tienen una distribución normal.*

Muestra	Pesos					Media	desviación	Recorrido
1	16.1	16.2	15.9	16	16.1	16.06	0.1020	0.3
2	16.2	16.4	15.8	16.1	16.2	16.14	0.1860	0.6
3	16.0	16.1	15.7	16.3	16.1	16.04	0.1960	0.6
4	16.1	16.2	15.9	16.4	16.6	16.24	0.2417	0.7
5	16.5	16.1	16.4	16.4	16.2	16.32	0.1470	0.4
6	16.8	15.9	16.1	16.3	16.4	16.30	0.3033	0.9
7	16.1	16.9	16.2	16.5	16.5	16.44	0.2800	0.8
8	15.9	16.2	16.8	16.1	16.4	16.28	0.3059	0.9
9	15.7	16.7	16.1	16.4	16.8	16.34	0.4030	1.1
10	16.2	16.9	16.1	17.0	16.4	16.52	0.3655	0.9
11	16.4	16.9	17.1	16.2	16.1	16.54	0.3929	1.0
12	16.5	16.9	17.2	16.1	16.4	16.62	0.3868	1.1
13	16.7	16.2	16.4	15.8	16.6	16.34	0.3200	0.9
14	17.1	16.2	17.0	16.9	16.1	16.66	0.4224	1.0
15	17.0	16.8	16.4	16.5	16.2	16.58	0.2856	0.8
16	16.2	16.7	16.6	16.2	17.0	16.54	0.3072	0.8
17	17.1	16.9	16.2	16.0	16.1	16.46	0.4499	1.1
18	15.8	16.2	17.1	16.9	16.2	16.44	0.4841	1.3
19	16.4	16.2	16.7	16.8	16.1	16.44	0.2728	0.7
20	15.4	15.1	15.0	15.2	14.9	15.12	0.1720	0.5

Los valores que tenemos son los de la desviación típica de cada muestra. Por tanto usamos la expresión 7.3:

$$(\bar{X} - A_1\bar{S}, \bar{X} + A_1\bar{S}) = (16.32 - 1.5956 \times 0.3017, 16.32 + 1.5956 \times 0.3017) = (15.839, 16.801).$$

Si la tabla no tiene los valores de  $A_1$ , podemos emplear la primera expresión de 7.3. Como nuestra tabla no trae los valores de  $c_2$ , usamos la relación ya obtenida  $c_2 = c_4\sqrt{\frac{n-1}{n}}$

$$\begin{aligned} \left( \bar{X} - 3\frac{\bar{S}}{c_2\sqrt{n}}, \bar{X} + 3\frac{\bar{S}}{c_2\sqrt{n}} \right) &= \left( \bar{X} - 3\frac{\bar{S}}{c_4\sqrt{\frac{n-1}{n}}\sqrt{n}}, \bar{X} + 3\frac{\bar{S}}{c_4\sqrt{\frac{n-1}{n}}\sqrt{n}} \right) = \\ &= \left( \bar{X} - 3\frac{\bar{S}}{c_4\sqrt{n-1}}, \bar{X} + 3\frac{\bar{S}}{c_4\sqrt{n-1}} \right) = \\ &= \left( 16.32 - 3\frac{0.3017}{0.9400 \times \sqrt{4}}, 16.32 + 3\frac{0.3017}{0.9400 \times \sqrt{4}} \right) = \\ &= \left( 16.32 - \frac{3}{0.9400 \times \sqrt{4}} 0.3017, 16.32 + \frac{3}{0.9400 \times \sqrt{4}} 0.3017 \right) = \\ &= (16.32 - 1.5957 \times 0.3017, 16.32 + 1.5957 \times 0.3017) = (15.839, 16.801) \end{aligned}$$

En este caso una muestra queda fuera de control, la muestra 20. Eliminando y repitiendo el proceso anterior para las restantes se obtiene para límites de control (15.89, 16.87).

Si el control se realizara por recorrido se hace la media del recorrido en lugar de la media de las desviaciones. La expresión para el intervalo de control para la media de las muestras es:

$$(\bar{X} - A_2\bar{r}, \bar{X} + A_2\bar{r}) = (16.32 - 0.577 \times 0.82, 16.32 + 0.577 \times 0.82) = (15.847, 16.793)$$

Estos límites de control para la media también excluyen la muestra 20. Eliminando esta muestra obtenemos para límites de control los valores (15.90, 16.86) que no difieren grandemente de los que hemos obtenido usando las desviaciones típicas y son más fáciles de calcular.

Suele usarse el recorrido para muestras pequeñas. Para muestras de más de 10 elementos es mejor usar la desviación típica o la cuasidesviación.

## 7.2 Control de la variabilidad del sistema

### 7.2.1 Límites de control para la varianza. Comparación con un estándar

También suele controlarse la variabilidad del sistema ya que, cuando el proceso está bajo control, además de conservarse en media, no debe presentar más variabilidad de la aceptada. Este control lo hacemos estableciendo también límites de control para la varianza o para el recorrido.

Para establecer unos límites de control teórico para la varianza puede usarse un intervalo de varianza basado en el siguiente teorema

**Teorema 4** Si  $s^2$  es la cuasivarianza de una muestra aleatoria de tamaño  $n$  procedente de una población normal, entonces  $\frac{(n-1)s^2}{\sigma^2} = \frac{nS^2}{\sigma^2}$  es una variable aleatoria que sigue una distribución  $\chi^2$  (chi-cuadrado) con  $n-1$  grados de libertad.

Este teorema puede servir para establecer unos límites de confianza para la varianza o cuasivarianza de muestras de tamaño  $n$ . Como en esta caso, la chi-cuadrado no es simétrica. Debemos dar un límite inferior y otro inferior para el nivel de significación.

Otra forma de construir intervalos de control teóricos es emplear una aproximación normal para la distribución de  $s$ .

La esperanza de  $s$  es  $c_4\sigma$  y la desviación típica de  $s$  es  $\sigma\sqrt{1-c_4^2}$ . Por lo tanto actuando de forma similar al establecimientos de límites de control para la media de la población estableceremos como límites de control para el valor de  $s$  de cada muestra el intervalo siguiente:

$$\left( c_4\sigma - 3\sigma\sqrt{1-c_4^2}, c_4\sigma + 3\sigma\sqrt{1-c_4^2} \right) = (\sigma B_5, \sigma B_6)$$

### Establecimientos de límites de control para $s$ a partir de una muestra

Si no se dispone del valor de  $\sigma$  hay que estimarlo. De una relación anterior se deduce que un estimador centrado de  $\sigma$  puede ser

$$\hat{\sigma} = \frac{1}{c_4} \frac{\sum_{j=1}^k s_j}{k}$$

con lo que los límites de control para la cuasidesviación  $s$  viene establecido por el intervalo

$$\left( \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2}, \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2} \right) = \left( \bar{s}\left(1 - \frac{3}{c_4}\sqrt{1-c_4^2}\right), \bar{s}\left(1 + \frac{3}{c_4}\sqrt{1-c_4^2}\right) \right) = (\bar{s}B_3, \bar{s}B_4)$$

Los límites de control para la desviación típica  $S$  son  $(\bar{S}B_3, \bar{S}B_4)$

### 7.2.2 Límites de control para el recorrido

Los límites de control para el rango a partir de valores muestrales son  $(\bar{r}D_3, \bar{r}D_4) = (\hat{\sigma}D_1, \hat{\sigma}D_2)$

### 7.2.3 Resumen:

**Estimaciones de los parámetros:**

$$\hat{\mu} = \bar{X} = \frac{\sum_{j=1}^k \bar{x}_j}{k}, \quad \hat{\sigma} = \frac{1}{c_4} \bar{s}, \quad \hat{\sigma} = \frac{1}{c_2} \bar{S}, \quad \hat{\sigma} = \frac{1}{d_2} \bar{r}$$

**Límites de control para la media y el recorrido:**

	media	s
Teóricos	$\left(\mu - 3\frac{\sigma}{\sqrt{n}}, \mu + 3\frac{\sigma}{\sqrt{n}}\right)$	$(\sigma B_5, \sigma B_6)$
Muestrales	$\begin{pmatrix} (\bar{X} - A_3\bar{s}, \bar{X} + A_3\bar{s}) \\ (\bar{X} - A_2\bar{r}, \bar{X} + A_2\bar{r}) \\ (\bar{X} - A_1\bar{S}, \bar{X} + A_3\bar{S}) \end{pmatrix}$	$(\bar{s}B_3, \bar{s}B_4)$

**7.3 Límites de tolerancia**

La tolerancia es una medida de la adaptación del producto al fin para el que está concebido. Los límites de control dan intervalos de confianza para la media y para la desviación típica. La información que nos suministran estos límites de control nos orientan sobre lo que es capaz de realizar el proceso de fabricación, pero a veces las especificaciones se dan con medidas de tolerancia. Es frecuente que el ingeniero incluya en las especificaciones la proporción de piezas que se espera que estén dentro de unos límites de tolerancia.

Existen varias formas de establecer los límites de tolerancia. Nosotros vamos a exigir que haya una proporción dada de elementos dentro de estos límites. Si se conocen los parámetros de la población podemos asegurar que el 95% de los elementos de la población están entre  $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ . Estos son los *límites de tolerancia naturales del proceso*. Pero si no conocemos los verdaderos valores de  $\mu$  y  $\sigma$ , los límites de tolerancia para que el 95% de la población esté en el intervalo de tolerancia,  $(LIT, LST)$ , se estiman a partir de una muestra. El desconocimiento de los verdaderos parámetros de la distribución no nos permite asegurar el porcentaje de defectuosos dentro del intervalo de tolerancia. En este caso estableceremos el porcentaje esperado de defectuosos dentro del intervalo de tolerancia con una “cierta confianza”. Por ejemplo, se podría dar la información sobre el intervalo de tolerancia en los siguientes términos: El 95% de los producto de la población está contenida en el intervalo de tolerancia  $(LIT, LST)$ , con el 90% de confianza.

Se calculan los intervalos de tolerancia como  $(\bar{X} - ks, \bar{X} + ks)$  donde los valores de  $k$  se encuentran tabulados en la tabla 7.2 de la página 221.

**Ejemplo 47** *Un fabricante quiere establecer unos límites de tolerancia para un proceso de forma que el 95% de los artículos caiga dentro de estos límites con una confianza del 99%. Para ello dispone de 25 observaciones cuya media resultó  $\bar{x} = 40.75$ . Su cuasivarianza fué  $s^2 = 1.87$*

$$(\bar{X} - ks, \bar{X} + ks) = (40.75 - 2.972 \times 1.37, 40.75 + 2.972 \times 1.37) = (36.69, 44.81)$$

Es conveniente resaltar la diferencia entre límites de control y límites de tolerancia:

Los límites de control ( $LIC, LSC$ ) dan intervalo de confianza para *el valor medio de las medidas* (con un cierto nivel de confianza). Queremos decir que hay un cierto grado de confianza, por ejemplo del 99%, de que la *media* de los productos que se fabrican tengan medidas entre  $LIC$  y  $LSC$ , es decir que esté contenida en  $\left(\bar{X} - 3\frac{\bar{s}}{c_4\sqrt{n}}, \bar{X} + 3\frac{\bar{s}}{c_4\sqrt{n}}\right)$  o en  $\left(\bar{X} - 3\frac{\bar{r}}{d_2\sqrt{n}}, \bar{X} + 3\frac{\bar{r}}{d_2\sqrt{n}}\right)$ .

Los límites de tolerancia ( $LIT, LST$ ) nos dan un intervalo donde se encuentra *una proporción de elementos* de la población (el 95% por ejemplo) con un cierto grado de confianza (por ejemplo el 99%). En el ejemplo anterior, afirmamos que el 95% de los artículos fabricados tienen medidas dentro del intervalo de tolerancia (36.69, 44.81) y que esta afirmación la realizamos con el 99% de confianza.

### Intervalo de tolerancia empírica

Para establecer el intervalo de tolerancia del anterior epígrafe, se da por supuesto que la población se distribuye normalmente. En el caso de que esta propiedad no pueda darse por supuesto, podemos utilizar el siguiente procedimiento, que permite dar un intervalo de tolerancia para una distribución arbitraria de las medidas de la variable que estamos observando.

Dada una muestra de  $n$  elementos, sea  $c$  el menor de ellos y  $g$  el mayor. Daremos el intervalo  $(c, g)$  como intervalo de tolerancia. La confianza con la que se afirma que este intervalo de tolerancia contiene una proporción  $p$  de elementos de los elementos puede calcularse con la siguiente expresión:

$$P[(c, g) \text{ cubra al menos una proporción } p \text{ de artículos}] = 1 - np^{n-1} + (n-1)p^n$$

**Ejemplo 48** *Midiendo la vida de 50 lamparas se ha encontrado que la primera que ha fallado ha durado 2150 horas y la última en fallar ha durado más 2610 horas ¿Con que confianza puede decirse que el 90% de las lamparas falla entre estos valores?*

Para calcular la confianza con la que se realiza la afirmación de que el 90% de las lamparas está en el intervalo de tolerancia (2150, 2610) se calcula por medio de la expresión:

$$1 - np^{n-1} + (n-1)p^n = 1 - 50 \times 0.90^{49} + 49 \times 0.90^{50} = 0.96621$$

La afirmación se realiza con un nivel de confianza del 97%.

#### 7.3.1 Capacidad de un proceso

La capacidad del proceso se define como  $6\sigma$ . Los índices de capacidad son distintos parámetros que tratan de reflejar la adaptación del proceso de fabricación a las características deseables para el producto fabricado. Por lo

general se desea que la mayoría de los productos cumplan las especificaciones. Estas especificaciones se identifican a veces con los límites de tolerancia. Se dan varios parámetros que resuman esta información. Damos aquí uno de los más usados:

$$K = \frac{LST - LIT}{6\sigma}$$

El parámetro así definido es útil si los límites de tolerancia son simétricos respecto de la media. En este caso, si  $K < 1$  no se cumplen las especificaciones fijadas. Hay más del 3 por mil de artículo fuera de las especificaciones. Si  $K = 1$  se fabricarán el 3 por mil de defectuosos. Es lo que se espera del proceso y para esto lo hemos diseñado. Si  $K > 1$ , la tolerancia es mayor que la diferencia permitida por los límites de control y se producen menos del 3 por mil de defectuosos.

**Ejemplo 49** *Se probó una muestra aleatoria de 45 resistores con una resistencia promedio de 498 ohms. La cuasidesviación típica de esas resistencias resultó ser de 4 ohms. Determinar un intervalo de tolerancia de 95% para el 90% de la población de los resistores. Estimar el índice de tolerancia*

Suponer que la población es normal.

Para  $\alpha = 0.95$ ,  $p = 0.90$ ,  $n = 45$  el valor de  $k$  es 2.021, así que:

$$(\bar{X} - ks, \bar{X} + ks) = (498 - 2.021 \times 4, 498 + 2.021 \times 4) = (489.916, 506.084).$$

Hay un 95% de confianza de que el 90% de las resistencias tenga valores comprendidos entre 489.92 ohmios y 506.08 ohmios.

El índice de tolerancia se puede estimar sustituyendo el valor de  $\sigma$  por la cuasidesviación:

$$K = \frac{LSC - LIC}{6\sigma} = \frac{506.084 - 489.916}{6 \times 4} = \frac{16.168}{6 \times 4} = 0.673667.$$
 A menudo el proceso rebasa la tolerancia. Por tanto hay demasiadas resistencias fuera de los límites de tolerancia.

## 7.4 EJERCICIOS PROPUESTOS

**Ejercicio 89** *En un proceso industrial se controla la resistencia a la tensión de ciertas piezas metálicas. Para ello se ha medido la resistencia ( $x_i$ ) de 30 muestras de 6 elementos cada una. obteniéndose que la suma de las medias de las 30 muestras es 6000 y la suma de sus cuasidesviaciones 150.*

1. *Calcular, a partir de estas muestras los límites de control para la media y para la cuasidesviación.*
2. *Se ha concluido que el proceso está bajo control. Determinar el índice de capacidad si los límites de tolerancia son  $200 \pm 5$ .*

3. ¿Cuántas piezas defectuosas produce este proceso? (Se entiende que una pieza es defectuosa si sobrepasa los límites de tolerancia).
4. En un momento dado se desajusta el proceso y fabrica piezas con media 199, conservándose no obstante la varianza. ¿Cuál es la probabilidad de detectar el desajuste en la siguiente muestra de 6 elementos que se tome?

**Ejercicio 90** Si la media del peso de unas latas de conservas es 41.5 gr y la desviación típica es 0.5 gr. Se pide:

1. Hallar los límites de control teóricos para las medias muestrales si el número de elementos de cada muestra es  $n = 5$
2. Hallar los límites de control para la cuasidesviación.
3. La siguiente tabla nos da los valores obtenidos para la media y la desviación típica de 20 muestras de tamaño  $n = 5$  del mismo proceso

$\bar{x}$	41.9	41.3	42.1	41.6	41.8	42.3	41.4	41.6	41.8	42
$s$	0.8	0.2	0.3	0.7	0.9	0.1	0.4	0.5	0.6	0.2
$\bar{x}$	42	41.8	41.3	42.0	42.0	41.7	41.5	41.49	41.6	41.4
$s$	0.3	0.3	0.2	0.5	0.6	0.2	0.4	0.4	0.3	0.6

¿Estos valores indican que el proceso está bajo control en media? ¿Y en varianza?

**Ejercicio 91** Los diámetros de las arandelas que salen de una línea de fabricación siguen una distribución normal de media 0.5 cm y una desviación típica de 0.1 cm. Se pide:

1. Hallar los límites de control teóricos 3-sigma para la media de muestras con 10 elementos.
2. Intervalo de control teórico para la varianza basado en la distribución chi-cuadrado
3. Se ha observado en un instante una muestra de 10 de estas arandelas. Las dimensiones de sus diámetros han sido: 0.4; 0.43; 0.6; 0.42; 0.7; 0.51; 0.61; 0.44; 0.62; 0.49. ¿Confirman estos valores el estado de control del proceso?
4. Calcular los límites de control para  $s$  basados en muestras de 10 elementos.

5. Si se analizan 100 muestras de 10 elementos cada una, ¿cuál es la probabilidad de que haya alguna fuera de control en media?

**Ejercicio 92** Se ha observado en un instante una muestra de 10 chapas metálicas. Sus pesos en gramos han sido: 40, 43, 60, 42, 70, 51, 61, 44, 62, 49. Hallar un intervalo de tolerancia que contenga el 90% de las piezas fabricadas con (95% de confianza).

**Ejercicio 93** Los diámetros de las varillas fabricadas en una máquina es una característica importante en su calidad. La siguiente tabla muestra los valores de  $\bar{y}$  y de  $R$  para 20 muestras de 5 varillas cada una. Las especificaciones de las varillas son  $0.5035 \pm 0.0010$  pulgadas. Los valores dados en la tabla son las últimas tres cifras de la medida. Es decir 34.2 significa 0.50342.

Muestra	$\bar{x}$	$R$	Muestra	$\bar{x}$	$R$
1	0.50342 34.2	0.0003 3	11	35.4 0.50354	0.0008 8
2	31.6	4	12	34.0	6
3	31.8	4	13	36.0	4
4	33.4	5	14	37.2	7
5	35.0	4	15	35.2	3
6	32.1	2	16	33.4	10
7	32.6	7	17	35.0	4
8	33.8	9	18	34.4	7
9	34.8	10	19	33.9	8
10	38.6	4	20	34.0	4

- Hallar la media de las muestras, el rango medio y revisar si fuera necesario los límites de control.
- Calcular índice de capacidad del proceso.
- ¿Qué porcentajes de defectos está produciendo este proceso?

**Ejercicio 94** Supongamos que se usa un diagrama de control 3-sigma para un proceso distribuido normalmente. Cada 2 horas se toman muestras de 30 elementos y se marca el punto correspondiente en el diagrama. Hallar el número esperado de muestras que se habrán inspeccionado hasta que un punto esté fuera de los límites de control.

**Ejercicio 95** Un proceso de fabricación varillas de aluminio está bajo control. Las especificaciones indican que los diámetros de las varillas siguen una distribución normal  $N(1.25, 0.01)$ . Para realizar el control de calidad se toman muestras de 5 elementos cada hora.



1. ¿Cuáles serán los límites de control 3-sigma.?
2. ¿Cuál es la probabilidad de obtener una muestra fuera de control en una prueba?
3. ¿Cuál es la probabilidad de obtener 3 muestras fuera de control en 10 pruebas?
4. ¿Cuál es la probabilidad de obtener 1 muestras fuera de control en 100 pruebas?

**Ejercicio 96** Si la media de la medidas del diametro de unas varillas es  $\mu = 4.2$  y la desviación típica es  $\sigma = 0.05$

Se pide:

1. Hallar los límites de control teóricos si el número de elementos de cada muestra es  $n=6$
2. Hallar los límites de control 3-sigma para la desviación típica
3. La siguiente tabla nos da los valores obtenidos para la media y la desviación típica de 20 muestras de tamaño  $n=6$  del mismo proceso

$\bar{x}$	4.24	4.18	4.26	4.21	4.18	4.23
$s$	0.008	0.002	0.003	0.007	0.009	0.001
4.19	4.21	4.18	4.20	4.25	4.23	4.18
0.004	0.005	0.006	0.002	0.003	0.003	0.002
4.25	4.25	4.22	4.20	4.19	4.21	4.19
0.005	0.006	0.002	0.004	0.004	0.003	0.006

¿Estos valores indican que el proceso está bajo control en media? ¿ Y en varianza? Hacer ambas gráficas de control.

**Ejercicio 97** Se supone que un proceso produce piezas cuyas medidas sigue una distribución normal con  $\mu = 100$ ,  $\sigma = 2$ .

1. Calcular los límites de control de calidad teóricos, para muestras de tamaño 5.
2. Se ha realizado un control de calidad para verificar la validez de estos parámetros, obteniéndose los resultados siguientes con muestras de

tamaño 5

Muestra	media	recorrido	Muestra	media	recorrido
1	99.7	3.1	11	98.4	3.1
2	99.8	3.4	12	98.5	2.8
3	100.0	3.3	13	97.9	2.9
4	99.8	3.6	14	98.5	3.2
5	99.9	3.0	15	100.8	3.1
6	99.7	3.2	16	100.5	3.3
7	100.1	3.1	17	99.4	3.4
8	100.2	2.9	18	99.9	3.0
9	99.3	1.9	19	97.5	3.5
10	99.7	3	20	99.2	3.3

¿Se deben cambiar los límites de control?

- Si se consideran defectuosas las piezas que están fuera del intervalo  $(94, 106)$  ¿Qué proporción de piezas defectuosas produce el proceso?

**Ejercicio 98** La longitud del encendedor de cigarrillos de un automóvil es controlada mediante el empleo de gráficos de control para la media y para el recorrido. La siguiente tabla proporciona las medidas de la longitud para 20 muestras de tamaño 4.

	Observaciones					Observaciones			
$n^\circ$	1	2	3	4	$n^\circ$	1	2	3	4
1	5.15	5.10	5.08	5.09	11	5.13	5.08	5.09	5.05
2	5.14	5.14	5.10	5.06	12	5.10	5.15	5.08	5.10
3	5.09	5.10	5.09	5.11	13	5.08	5.12	5.14	5.09
4	5.08	5.06	5.09	5.13	14	5.15	5.12	5.14	5.05
5	5.14	5.08	5.09	5.12	15	5.13	5.16	5.09	5.05
6	5.09	5.10	5.07	5.13	16	5.14	5.08	5.08	5.12
7	5.15	5.10	5.12	5.12	17	5.08	5.10	5.16	5.09
8	5.14	5.16	5.11	5.10	18	5.08	5.14	5.10	5.09
9	5.11	5.07	5.16	5.10	19	5.13	5.15	5.10	5.08
10	5.11	5.14	5.11	5.12	20	5.09	5.07	5.15	5.08

- Hallar los límites superiores e inferiores de control, para la media y el rango de cada muestra y eliminando, si es necesario, las muestras fuera de control.
- Si el intervalo de tolerancia es  $(5.05, 5.15)$  calcula el índice de capacidad y estima la proporción de encendedores que quedarían fuera de este intervalo de tolerancia.

**Ejercicio 99** *En una empresa envasadora de cervezas se han realizado gráficos de control, basados en muestras de 4 botellas, para la media y el rango del contenido en cerveza de las botellas que comercializa. Estos diagramas han dado lugar a los siguientes datos: a) Para el diagrama de la media  $LSC=330.99$ , línea central 328,  $LIC=325.01$  b) Para el diagrama del rango  $LSC = 9.4$  Línea central 4.1,  $LIC=0$ . Se ha confirmado que el proceso está bajo control.*

1. *Calcular la desviación típica estimada por el proceso.*
2. *Si usamos estos los datos del diagrama de rango para realizar un diagrama de control para la desviación típica, ¿Que aspecto presentaría?*
3. *Si se supone que estas botellas quieren venderse como botellas de un tercio y que las especificaciones que desea dar el fabricante para el contenido de las mismas son  $333 \pm 8 \text{ cm}^3$  ¿Que porcentaje de botellas resultarían fuera de estas especificaciones? ¿Que corrección se debería hacer en el proceso de envasado para disminuir este porcentaje?*

Observaciones en la muestra, n	Diagrama para medias						Diagrama para desviaciones estándares						Diagrama para amplitudes					
	Factores para límites de control			Factores para línea central			Factores para límites de control			Factores para línea central			Factores para límites de control			Factores para línea central		
	A	A <sub>2</sub>	A <sub>3</sub>	c <sub>4</sub>	1/c <sub>4</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	d <sub>2</sub>	1/d <sub>2</sub>	d <sub>3</sub>	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>		
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267		
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.574		
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282		
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114		
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004		
7	1.134	0.419	1.182	0.9594	1.04230	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924		
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864		
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816		
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777		
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744		
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717		
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693		
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672		
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653		
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637		
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622		
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608		
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597		
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585		
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575		
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566		
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557		
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548		
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541		

Para n > 25

$$A = \frac{3}{\sqrt{n}}, A_3 = \frac{3}{c_4\sqrt{n}}, c_4 = \frac{4(n-1)}{4n-3}$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}}, B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}}$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}}, B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}$$

Figura 7.1:

Factores para límites normales de tolerancia bilaterales									
$n$	Confianza de 90% de que el porcentaje de la población entre los límites es			Confianza de 95% de que el porcentaje de la población entre los límites es			Confianza de 99% de que el porcentaje de la población entre los límites es		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
2	15.98	18.80	24.17	32.02	37.67	48.43	160.2	188.5	242.3
3	5.847	6.919	8.974	8.380	9.916	12.86	18.93	22.40	29.06
4	4.166	4.943	6.440	5.369	6.370	8.299	9.398	11.15	14.53
5	3.494	4.152	5.423	4.275	5.079	6.634	6.612	7.855	10.26
6	3.131	3.723	4.870	3.712	4.414	5.775	5.337	6.345	8.301
7	2.902	3.452	4.521	3.369	4.007	5.248	4.613	5.448	7.187
8	2.743	3.264	4.278	3.136	3.732	4.891	4.147	4.936	6.468
9	2.626	3.125	4.098	2.967	3.532	4.631	3.822	4.550	5.966
10	2.535	3.018	3.959	2.829	3.379	4.433	3.582	4.265	5.594
11	2.463	2.933	3.849	2.737	3.259	4.277	3.397	4.045	5.308
12	2.404	2.863	3.758	2.655	3.162	4.150	3.250	3.870	5.079
13	2.355	2.805	3.682	2.587	3.081	4.044	3.130	3.727	4.893
14	2.314	2.756	3.618	2.529	3.012	3.955	3.029	3.608	4.737
15	2.278	2.713	3.562	2.480	2.954	3.878	2.945	3.507	4.605
16	2.246	2.676	3.514	2.437	2.903	3.812	2.872	3.421	4.492
17	2.219	2.643	3.471	2.400	2.858	3.754	2.808	3.345	4.393
18	2.194	2.614	3.433	2.366	2.819	3.702	2.753	3.279	4.307
19	2.172	2.588	3.399	2.337	2.784	3.656	2.703	3.221	4.230
20	2.152	2.564	3.368	2.310	2.752	3.615	2.659	3.168	4.161
21	2.135	2.543	3.340	2.286	2.723	3.577	2.620	3.121	4.100
22	2.118	2.524	3.315	2.264	2.697	3.543	2.584	3.078	4.044
23	2.103	2.506	3.292	2.244	2.673	3.512	2.551	3.040	3.993
24	2.089	2.489	3.270	2.225	2.651	3.483	2.522	3.004	3.947
25	2.077	2.474	3.251	2.208	2.631	3.457	2.494	2.972	3.904
26	2.065	2.460	3.232	2.193	2.612	3.432	2.469	2.941	3.865
27	2.054	2.447	3.215	2.178	2.595	3.409	2.446	2.914	3.828
28	2.044	2.435	3.199	2.164	2.579	3.388	2.424	2.888	3.794
29	2.034	2.424	3.184	2.152	2.554	3.368	2.404	2.864	3.763
30	2.025	2.413	3.170	2.140	2.549	3.350	2.385	2.841	3.733
35	1.988	2.368	3.112	2.090	2.490	3.272	2.306	2.748	3.611
40	1.959	2.334	3.066	2.052	2.445	3.213	2.247	2.677	3.518
50	1.916	2.284	3.001	1.996	2.379	3.126	2.162	2.576	3.385
60	1.887	2.248	2.955	1.958	2.333	3.066	2.103	2.506	3.293
80	1.848	2.202	2.894	1.907	2.272	2.986	2.026	2.414	3.173
100	1.822	2.172	2.854	1.874	2.233	2.934	1.977	2.355	3.096
200	1.764	2.102	2.762	1.798	2.143	2.816	1.865	2.222	2.921
500	1.717	2.046	2.689	1.737	2.070	2.721	1.777	2.117	2.783
1000	1.695	2.019	2.654	1.709	2.036	2.676	1.736	2.068	2.718
$\infty$	1.645	1.960	2.576	1.645	1.960	2.576	1.645	1.960	2.576

Figura 7.2:



## Tema 8

# Control de Recepción

### 8.1 Introducción.

El control de recepción se aplica al recibir materias primas, o también a productos intermedios que tienen que continuar el proceso de fabricación, para comprobar que cumplen las condiciones deseadas. El objetivo es disminuir el número de artículos defectuosos que recibe el cliente, o las siguientes fases de producción si es un artículo que aún va a seguir en la línea de producción. Por lo general este tipo de control, frecuentemente realizado utilizando técnicas de muestreo, es un compromiso de calidad entre el vendedor y el comprador. En este muestreo se decidirá aceptar, o no, un lote en función de los resultados de calidad obtenidos en una muestra.

El control de recepción puede ser también por variables y por atributos, aunque este último es el más utilizado. Nos restringimos en este tema al estudio del control por atributos.

### 8.2 El control simple por atributos

Suponemos que se recibe un lote grande de productos. Normalmente ni el comprador ni el vendedor conocen exactamente la proporción de defectuosos en este lote.

El vendedor y el comprador acuerdan un plan de muestreo (simple) que consiste en seleccionar  $n$  elementos del lote y un número de aceptación  $c$ . Si  $x$  (número de defectos) resulta ser menor o igual que  $c$ , se acuerda que el comprador aceptará el lote. En caso contrario lo rechazaría.

Para cada plan de muestreo (donde hay que concretar los valores de  $n$  y  $c$ ) se puede calcular la probabilidad de aceptar el lote según los valores de  $p$  (probabilidad real de defectuosos en el lote). La representación gráfica de esta relación da lugar a la llamada curva característica ( $CO$ ) del plan de muestreo.

En esta gráfica el eje horizontal contiene la variable  $p$  y en el eje vertical se representa  $P_A(p)$  = Probabilidad de aceptar el lote si la probabilidad real de defectuosos fuera  $p$ .

**Ejemplo 50** *Un plan de inspección para muestreo requiere  $n=10$ ,  $c=1$ . Calcular*

*Probabilidad de aceptar un lote si la proporción real de defectuosos fuera  $p=0$ ,  $p=0.1$ ,  $p=0.2$ ,  $p=0.3, \dots$ . Suponemos que el lote es suficientemente grande para que sea adecuada la distribución binomial. La distribución más adecuada para lotes pequeños es la hipergeométrica.*

$$P_A(0) = 1,$$

$$P_A(0.1) = \binom{10}{0} 0.1^0 0.9^{10} + \binom{10}{1} 0.1^1 0.9^9 = .0.736099$$

$$P_A(0.2) = \binom{10}{0} 0.2^0 0.8^{10} + \binom{10}{1} 0.2^1 0.8^9 = 0.37581$$

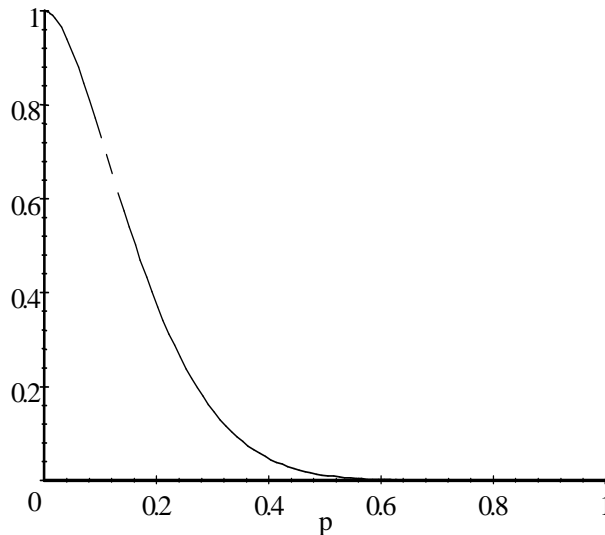
En general:

$$P_A(p) = \sum_{i=0}^c \binom{n}{i} p^i q^{n-i}$$

Este valor es el de la función de distribución de  $B(n, p)$  en  $i$ . Los valores anteriores pueden encontrarse en las tablas de la distribución binomial.

**Ejemplo 51** *Diseñar un boceto de la curva característica de este plan de muestreo.*

La curva característica presenta el siguiente aspecto

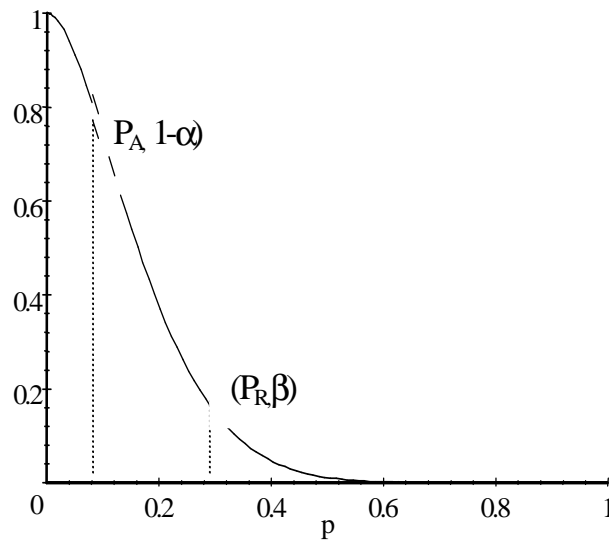




Los valores de las ordenadas pueden obtenerse en la tabla de la distribución binomial

### 8.3 Diseño de un plan de muestreo

Por lo general el comprador considera que si la proporción de defectuosas es  $p_A$  el lote debe aceptarse ( $p_A = NCA$ , Nivel de calidad aceptable o  $AQL$ ) y que si la proporción de defectuosas es  $p_R$  ( $p_R = NCR$ , Nivel de calidad rechazable) debe rechazarse. Supongamos que la proporción real de defectuosos fuera  $p$ . Se define  $\alpha =$  riesgo del vendedor = probabilidad de rechazar un lote con  $p = p_A$  y  $\beta =$  riesgo del comprador = probabilidad de aceptar un lote siendo  $p = p_R$ .



Para unos valores prefijados de  $(p_A, \alpha)$  y  $(p_R, \beta)$  los valores de  $n$  y  $c$  están definidos.

**Ejemplo 52** Diseñar un plan de muestreo con  $p_A=0.05$ ,  $\alpha=0.05$  y  $p_R=0.15$   $\beta=0.10$

$$P(r > c / P = p_A = 0.05) = \alpha = 0.05$$

Suponemos que sea aplicable la aproximación normal de la  $B(n, 0.05)$

$P(\text{aceptación} / p = p_R = 0.15) = \beta = 0.10 \implies P(r \leq c / p = 0.15) = 0.10$

usamos la aproximación normal de la  $B(n, 0.15)$

$$\frac{c - n \cdot 0.05}{\sqrt{n \cdot 0.05 \cdot 0.95}} = 1.64$$

$$\frac{c - n \cdot 0.55}{\sqrt{n \cdot 0.15 \cdot 0.85}} = -1.28$$

la solución es  $n = 66.338$  y  $c = 6.22$ . Podemos tomar los valores  $n = 67$ ,  $c = 7$

Si no es lícito usar la aproximación normal para la distribución del número de defectos el problema es bastante más complicado.

El planteamiento sería en este caso:

$$1 - P_A(p_A) = \sum_{i=c-1}^n \binom{n}{i} p_A^i (1 - p_A)^{n-i} = 0.05$$

$$P_A(p_R) = \sum_{i=1}^c \binom{n}{i} p_R^i (1 - p_R)^{n-i} = 0.10$$

donde las incógnitas son  $n$  y  $c$ .

## 8.4 Planes de muestreos tabulados

Para no tener que realizar problemas de este tipo, que a veces resultan sumamente complicados existen tablas de planes de muestreo donde pueden seleccionarse los valores adecuados para nuestras exigencias. Entre los planes de muestreo tabulados se encuentran los siguientes.

*El plan japonés JIS Z 9002*

Permite hallar una solución aproximada del problema anterior. Un tabla para este plan de muestreo es la que aparece en la página 227:

La solución más próxima que suministra esta tabla para los datos del problema anterior es  $n = 60$ ,  $c = 6$

*Plan Military -Standard (MIL-STD-105D, UNE 66020)*

Desarrollados por el ejercito estadounidense. Solo tienen en cuenta  $P_A$  y  $\alpha$ , pero a cambio tienen en cuenta el precio de la inspección, la calidad prevista de la muestra y el tamaño del lote.

*Planes de control rectificativo Dodge -Roming*

Se basan en que todos los lotes rechazados se inspeccionan al 100% y los elementos defectuosos son sustituidos por buenos, garantizando que la calidad media de entrada en almacén será alta.

La proporción de aceptables (AOQ) en el almacén se calculará (suponiendo que la proporción de defectuosos es en total  $p$ ) de la forma siguiente:

$AOQ(p) =$  Proporción de defectuosos en el almacén



= Probabilidad de aceptar el lote de esta calidad  $\times p$  + la probabilidad de rechazarlo  $\times 0$  =

$$P_A(p) \times p$$

En resumidas cuentas: *para calcular la proporción de elementos defectuosos en el almacén se multiplica las ordenadas de la curva característica por  $p$  (proporción de defectuosos en el lote). Por lo tanto en el almacén habrá menos defectuosos que en el lote.*

Es interesante el parámetro  $AOQL = \max de AOQ(p)$ , que nos indica la proporción máxima de defectuosas que se puede encontrar en el almacén. El valor de  $p$  que da lugar al  $AOQL$  se puede obtener igualando a cero la derivada del  $AOQ(p)$  con respecto a  $p$

Información más detallada sobre planes de muestreo puede encontrarse en **Estadística. Modelos y métodos. 1. Fundamentos, Daniel Peña Sánchez de Rivera, Alianza Editorial**

## 8.5 Muestreo doble y múltiple

El muestreo doble consiste en la inspección sucesiva de dos muestras de productos extraídas del lote. Se realiza de la forma siguiente:

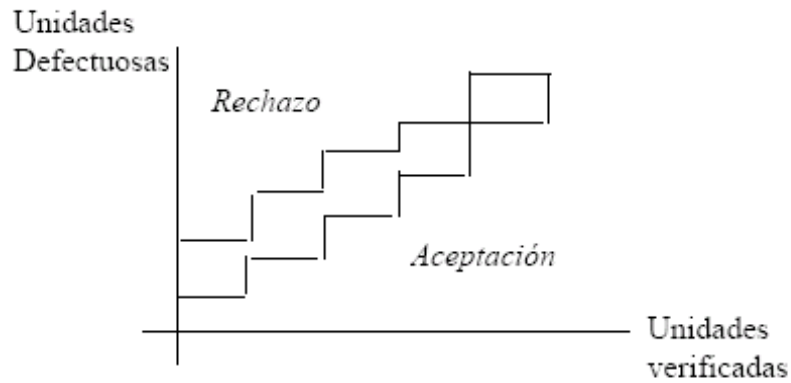
El plan está definido por medio de los valores  $n_1, c_1, r_1$ , para la inspección de la primera muestra, y  $n_2, r_2$  que resultan necesarios para tomar la decisión después de haber extraído la segunda muestra.

Se extrae del lote una primera muestra de tamaño  $n_1$ . Si el número de piezas defectuosas de esta muestra  $d_1$  es menor o igual que  $c_1$  se acepta el lote, si es mayor o igual que  $r_1$  se rechaza el lote. Si está comprendido entre  $c_1$  y  $r_1$  se pasa a la segunda fase del muestreo que consiste en extraer una segunda muestra de  $n_2$  elementos. Sea  $d_2$  el número de piezas defectuosas en esta segunda fase. Si  $d_1 + d_2$  es menor o igual que  $r_2$  se acepta el lote. En caso contrario se rechaza.

Un esquema gráfico de este plan de muestreo puede ser el siguiente:



En el muestreo múltiple este procedimiento se repite sucesivamente un número finito de veces. Un esquema gráfico de este tipo de muestreo es el siguiente:



Estos muestreos múltiples presentan ventajas e inconvenientes con respecto al simple.

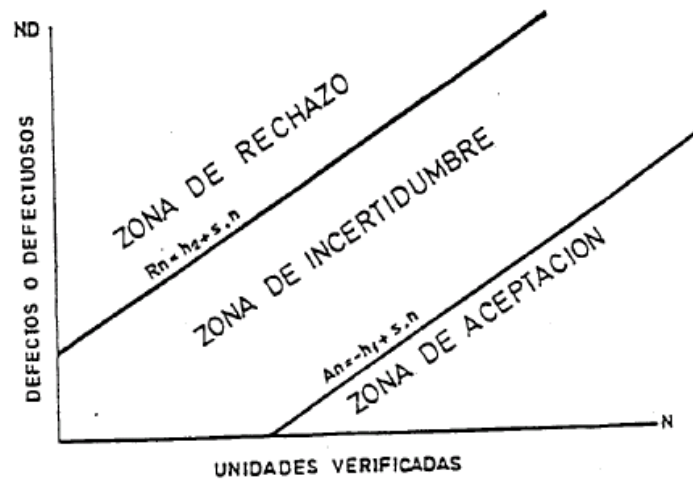
El número medio de elementos que hay que inspeccionar hasta tomar una decisión es menor en el muestreo doble que en el simple, en el triple que en el doble etc... con la misma protección contra el error, lo cual tiene mucho interés económico, ya que se reduciría el coste de la inspección.

Entre los inconvenientes de los muestreos múltiples se cuentan:

- a) Se ignora a priori la cantidad de elementos que vamos a tener que inspeccionar
- b) Es de más complicada realización
- c) Su duración puede ser larga, lo que, en ocasiones, requiere la inmovilización de las partidas con el consiguiente perjuicio en la producción.

## 8.6 Muestreo Secuencial

El muestreo secuencial es una extensión del muestreo múltiple. Consiste en decidir para cada elemento que se incorpora a la muestra si tomamos o no un siguiente elemento o ya la muestra extraída es suficiente para tomar una decisión. En cada elemento que se inspecciona se puede tomar la decisión de aceptar o rechazar. Depende de la posición en que nos situemos en cada momento con respecto a dos rectas paralelas. El criterio de decisión con respecto a estas dos rectas está indicado en la gráfica siguiente.



La forma de determinar la ecuación de las rectas decisorias es la siguiente:

La recta de aceptación es  $y = -h_1 + Sx$

La recta de rechazo es paralela a la anterior:  $y = h_2 + Sx$ .

Los valores de  $h_1$ ,  $h_2$  y  $S$  dependen de las características del plan de muestreo y se calculan con las expresiones:

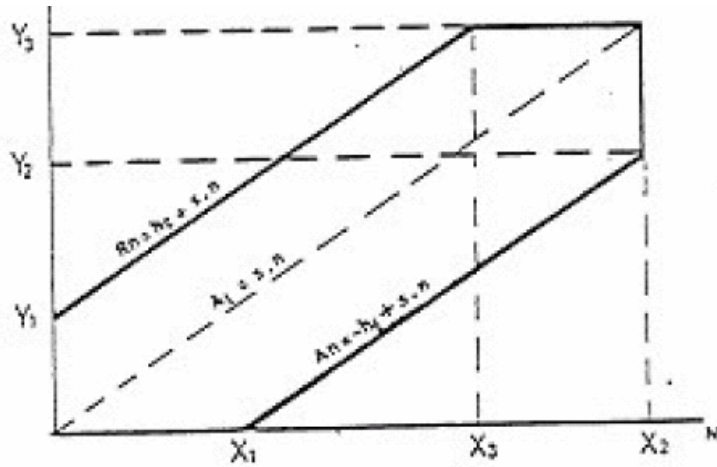
$$h_1 = \frac{\log\left(\frac{1-\alpha}{\beta}\right)}{\log\frac{P_R}{P_A} + \log\frac{1-P_A}{1-P_R}}, \quad h_2 = \frac{\log\left(\frac{1-\beta}{\alpha}\right)}{\log\frac{P_R}{P_A} + \log\frac{1-P_A}{1-P_R}}, \quad S = \frac{\log\frac{1-P_A}{1-P_R}}{\log\frac{P_R}{P_A} + \log\frac{1-P_A}{1-P_R}}$$

Cada vez que seleccionamos un elemento de la muestra seleccionamos un punto del diagrama. Sus coordenadas son:

$x$  = número de unidades inspeccionadas hasta el momento

$y$  = Número de unidades declaradas defectuosas hasta el momento.

El número medio de elementos que hay que inspeccionar hasta tomar una decisión es menor en el muestreo secuencial que en el simple, que en el doble, y en general que en cualquier muestreo múltiple con un número finito de muestras, si se exige la misma protección contra el error. El inconveniente es que su duración puede ser larga, ya que ni siquiera esta asegurada su terminación, ya que podemos quedarnos siempre en la zona de no decisión, la zona intermedia entre las dos rectas. Para remediar esta situación es conveniente usar un plan de muestreo secuencial truncado, cuyo diseño aparece en la siguiente gráfica:



Las rectas de truncamiento son  $x = k\frac{h_1}{S}$ ,  $y = kh_1$ , siendo  $3 \leq k \leq 10$ .

## 8.7 EJERCICIOS PROPUESTOS

**Ejercicio 100** *Un producto se produce en lotes de 25 piezas. El procedimiento de inspección consiste en seleccionar una muestra de 5 elementos de cada lote. El lote se acepta sólo si no aparece ningún elemento defectuoso en la muestra.*

1. *¿Cual es la probabilidad de aceptar un lote que contenga 3 elementos defectuosos?*
2. *¿Sería adecuado usar la aproximación binomial para calcular esta probabilidad? ¿Y si los lotes fueran de 200 piezas.*
3. *Supongamos que un lote de 25 piezas contiene 5 defectuosas. El procedimiento de muestreo va a consistir en seleccionar una muestra de algunas muestras de este lote. Si la muestra contiene algún elemento defectuoso se rechazará el lote. Si se quiere que la probabilidad de rechazar este lote sea al menos 0.95 ¿Cuántos elementos ha de tener la muestra?*

**Ejercicio 101** *Entre un vendedor y un cliente se ha acordado un plan de muestreo que consiste en que el comprador aceptará los lotes si en una muestra de 100 piezas el número de defectuosas no es mayor de 10. En caso contrario los rechazará.*

1. ¿Cuál es la probabilidad de aceptar un lote con un número de defectuosas del 15%? ¿Y del 25%?
2. Da la expresión analítica para la curva característica de este plan de muestreo.

**Ejercicio 102** Un plan de muestreo doble consiste en dos fases. En la primera se inspeccionan 10 elementos de una muestra. Si no hay defectuosos se acepta el lote, Si el número de defectuosos es 3 o más de 3 se rechaza. En otro caso se toma una segunda muestra de 20 elementos, Si entre los 30 elementos inspeccionados en total el número de defectuosos no supera los 3, el lote se acepta, en otro caso se rechaza. ¿Cuál es la probabilidad de aceptar un lote con una proporción de piezas defectuosas de 10%? ¿Y con una proporción de elementos defectuosos del 20%?

**Ejercicio 103** Se ha establecido un plan de muestreo en recepción para lotes de una gran cantidad de arandelas. Las características de este plan son las siguientes: Se inspeccionará una muestra de 100 de estas arandelas. Se aceptará el lote si el número de piezas defectuosas no es superior a 7. En caso contrario se rechazará el lote.

1. Si porcentaje real de defectos de un lote es del 5%, ¿Cuál es la probabilidad de aceptar el lote?. ¿Cuál es la probabilidad de rechazar el lote si el porcentaje de defectos fuera del 10%?
2. Calcular las ordenadas de la curva característica del plan para los porcentajes de defectos 5%, 20% y 25% .
3. Como se sabe un plan de muestreo se establece usando 4 parámetros ( $P_A$ ,  $\alpha$ ,  $P_R$ ,  $\beta$ ), donde  $\alpha$  es el riesgo del vendedor y  $\beta$  es el riesgo del comprador. Si  $P_A$  es 0.06 y  $\beta$  es 0.05, calcular los valores de  $\alpha$ ,  $P_R$  que corresponden a este plan de muestreo
4. Da la expresión y la representación gráfica de la curva característica de este plan de muestreo.

**Ejercicio 104** Diseñar un plan de muestreo secuencial truncado siendo  $p_A = 0.05$ ,  $\alpha = 0.05$  y  $p_R = 0.15$   $\beta = 0.10$  y  $k = 5$ .

Indicar qué decisión hay que tomar en las siguientes situaciones:

1. Se han inspeccionado 30 piezas y se han encontrado 2 defectuosas
2. Se han inspeccionado 50 piezas y se han encontrado 3 defectuosas
3. Se han inspeccionado 61 piezas y se han encontrado 6 defectuosas



## Unidad Temática III

# FIABILIDAD



## Tema 9

# Fiabilidad y Fallos

### 9.1 Introducción. Fallos y clases de fallos

Las tecnologías RAMS agrupan cuatro conceptos distintos pero relacionados entre sí, que son: Fiabilidad (Reliability), Disponibilidad (Availability), Mantenibilidad (Maintainability) y Seguridad (Safety). En español, la agrupación de estas cuatro materias suele designarse con el nombre conjunto de *Confiabilidad*. La Fiabilidad está relacionada con la duración sin fallos de los productos. La Disponibilidad es la capacidad del sistema para funcionar en un determinado instante, bien porque no ha fallado previamente, o porque habiendo fallado sus componentes defectuosos han sido reparado o sustituidos por otros. La Mantenibilidad es la capacidad de ser mantenido o reparado preventiva y correctivamente con objeto de mejorar su disponibilidad. La Seguridad es la capacidad de operar sin producir daño.

En Ingeniería, se dice que un aparato o componente es fiable si realiza adecuadamente la tarea para la que ha sido diseñado a lo largo de su vida útil. El estudio de la Calidad, que se ha tratado en la unidad anterior, pretende garantizar que el producto sale de cada una de las fases de su fabricación, y por supuesto de la fábrica, en buenas condiciones. La Fiabilidad intenta garantizar que el producto realizará adecuadamente su labor durante un periodo razonable de tiempo.

En esta unidad tratamos la Fiabilidad desde un punto de vista estadístico. En este contexto, se define la *fiabilidad* como la probabilidad de que un dispositivo funcione satisfactoriamente durante un tiempo dado bajo ciertas condiciones definidas.

Damos algunas notas sobre esta definición:

La fiabilidad se define como una *probabilidad*, porque el tiempo de uso y las condiciones de funcionamiento no definen determinísticamente la duración de un dispositivo. Por eso el tiempo de duración sin fallos de éste se con-

sidera una variable aleatoria. Por ejemplo, como cualquiera de nosotros ha podido observar, no todas las bombillas que salen de la misma fabrica lucen el mismo tiempo, aunque se usen en la misma habitación y estén conectadas en la misma lámpara, por lo tanto la duración de estas bombillas tiene un comportamiento aleatorio.

Por *funcionamiento satisfactorio* se entiende el cumplimiento de ciertas actuaciones específicas. El cese de este funcionamiento satisfactorio se llama *fallo*. Las *condiciones* de uso son importantes para definir la fiabilidad. No duran lo mismo unos neumáticos si se usan en una autopista que si se usan en un camino rural

La fiabilidad, según la definición dada previamente, es función del tiempo. Sin embargo a veces el tiempo no es la mejor medida de la exposición al fallo. Para un coche los kilómetros recorridos podrían ser mejores indicadores para su fiabilidad que el tiempo que haya transcurrido desde la compra del vehículo, para un interruptor eléctrico el número de veces que se usa, etc... No obstante el tiempo es la variable que más frecuentemente influye en los fallos de los dispositivos, y es la que usaremos en esta unidad.

Es usual *clasificar los fallos* atendiendo a diferentes criterios:

Atendiendo a la forma en que se produce el fallo se clasifican en: *fallos catastróficos* (que acontecen de forma súbita) y *fallos por degradación*, que están relacionadas con el desgaste y deterioro de los materiales.

Atendiendo al momento en que se produce el fallo pueden clasificarse en:

*Fallos infantiles*: Se producen en los productos recién salidos de fábrica. Generalmente son defectos de fabricación. Este tipo de fallos pueden disminuirse por el control de calidad y son los que normalmente cubre la garantía.

*Fallos por azar* o aleatorios: averías accidentales, debidas a una conjunción de circunstancias adversas. Estos fallos pueden darse durante toda la vida de un dispositivo

*Fallos por desgaste*: Son los producidos tras un largo periodo de uso y se producen, generalmente, al final de la vida útil.

Atendiendo a la relación entre los fallos se llaman *fallos primarios* los que se producen sin que ningún otro fallo los haya provocado y *fallos secundarios* que son provocados o inducidos por otros fallos primarios.

## 9.2 Distribución de los fallos y función de fiabilidad

Consideramos que el origen del tiempo se toma en 0 y que el *tiempo de vida* o duración de un dispositivo es una variable aleatoria a la que llamamos  $\tau$  que puede tomar cualquier valor real entre 0 e infinito. Sea  $f(t)$  y  $F(t)$  las

funciones de densidad y de distribución de esta variable aleatoria. Se define la *función de fiabilidad* como

$$R(t) = P(\tau > t) = 1 - F(t) \quad (9.1)$$

Por este motivo se llama a  $F(t)$  función de in fiabilidad.

Se tienen, por tanto, las relaciones siguientes:

$$f(t) = F'(t) = -R'(t) \quad (9.2)$$

### 9.3 Vida media y tasa de fallo

Se llama *vida media* de un dispositivo a la esperanza matemática de la variable aleatoria “tiempo de vida”:

$$\mu = E(t) = \int_0^{\infty} t f(t) dt \quad (9.3)$$

La probabilidad de que un dispositivo falle en el intervalo  $(t, t + \Delta t)$  si no ha fallado antes es:

$$P(t < \tau \leq t + \Delta t / \tau > t) = \frac{P(t < \tau \leq t + \Delta t)}{P(\tau > t)}. \quad (9.4)$$

Dividiendo por  $\Delta t$  se halla la *tasa media de fallo* en el intervalo  $(t, t + \Delta t)$

$$h(t, \Delta t) = \frac{P(t < \tau \leq t + \Delta t)}{P(\tau > t)} : \Delta t = \frac{\frac{F(t+\Delta t) - F(t)}{\Delta t}}{R(t)} \quad (9.5)$$

(Una forma aproximada de calcular la tasa media de fallos en un intervalo es dividir número medio de fallos por unidad de tiempo en el intervalo, por el número de supervivientes en el punto medio de dicho intervalo)

Hallando límite cuando  $\Delta t \rightarrow 0$  en la expresión 9.5 se obtiene la función tasa de fallo instantánea, o simplemente, la *tasa de fallo*.

$$h(t) = \frac{f(t)}{R(t)} \quad (9.6)$$

que también puede tomar la forma

$$h(t) = \frac{-R'(t)}{R(t)} \quad (9.7)$$

La tasa de fallo,  $h(t)$ , es una medida de la variación de la fiabilidad en el tiempo.

De la última expresión (9.7) se obtiene, mediante su integración, una expresión de la fiabilidad en función de la tasa de fallo. En efecto:

$$\int_0^t h(t) dt = \int_0^t \frac{-R'(t)}{R(t)} dt \quad (9.8)$$

$$\int_0^t h(t) dt = -\ln R(t)|_0^t = -\ln R(t) + \ln R(0) = -\ln R(t) + \ln 1 = -\ln R(t)$$

y por tanto

$$R(t) = \exp\left(-\int_0^t h(t) dt\right) \quad (9.9)$$

**Ejemplo 53** Calcular las funciones de fiabilidad, in fiabilidad y de densidad que corresponden a una función tasa de fallo  $h(t) = 2t$  con  $t \geq 0$ .

$$R(t) = \exp\left(-\int_0^t h(t) dt\right) = \exp\left(-\int_0^t 2t dt\right) = e^{-t^2}, t \geq 0$$

$$F(t) = 1 - R(t) = 1 - e^{-t^2}, t \geq 0$$

$$f(t) = F'(t) = \frac{d}{dt}(1 - e^{-t^2}) = 2te^{-t^2}, t \geq 0$$

## 9.4 La vida media en función de la fiabilidad

Si  $\lim_{t \rightarrow \infty} h(t) \neq 0^1$ , la vida media puede expresarse en función de la fiabilidad en la forma:

$$\mu = E(t) = \int_0^{\infty} R(t) dt \quad (9.10)$$

En efecto, realizando por partes la integral de la expresión 9.10, tomando  $u = R(t)$ , y  $dv = dt$ , se tiene:

$$\int_0^{\infty} R(t) dt = |t(1 - F(t))|_0^{\infty} + \int_0^{\infty} t f(t) dt = \lim_{t \rightarrow \infty} t(1 - F(t)) - 0 + \mu \quad (9.11)$$

ya que  $\mu$  representa la vida media (tiempo medio de duración del producto) y viene dada por la expresión:

$$\mu = E(t) = \int_0^{\infty} t f(t) dt$$

---

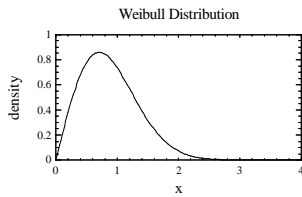
<sup>1</sup>En la práctica, las funciones empleadas en fiabilidad cumplen la condición  $\lim_{t \rightarrow \infty} h(t) \neq 0$ . Desde el punto de vista práctico la condición  $\lim_{t \rightarrow \infty} h(t) = 0$ , supondría que los dispositivos en cuestión tendrían una tasa de fallo cada vez más cercana a cero, así que tenderían a ser cada vez mejores hasta la perfección, lo que no parece un modelo adecuado para sistemas reales.

Realizando el límite por la regla de L'Hôpital se obtiene:

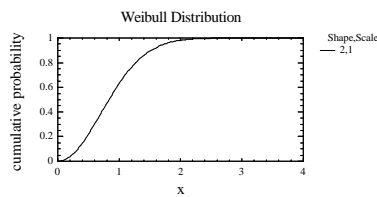
$$\lim_{t \rightarrow \infty} t(1 - F(t)) = \lim_{t \rightarrow \infty} \frac{t}{\frac{1}{1-F(t)}} = \lim_{t \rightarrow \infty} \frac{1}{\frac{f(t)}{R^2(t)}} = \lim_{t \rightarrow \infty} \frac{R(t)}{h(t)} = \lim_{t \rightarrow \infty} \frac{1 - F(t)}{h(t)} \tag{9.12}$$

El numerador tiende a 0, así que este límite tenderá a 0 siempre que  $\lim_{t \rightarrow \infty} h(t) \neq 0$ .

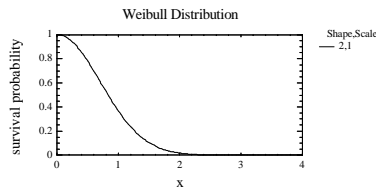
A continuación mostramos representaciones gráficas de las funciones de densidad, infiabilidad, fiabilidad y tasa de fallo de la distribución cuya función de densidad,  $f(t) = 2te^{-t^2}$ , ha sido tratada en el ejemplo 53.



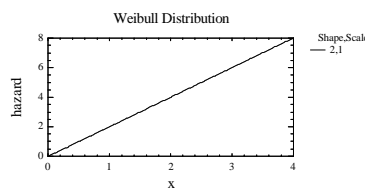
Función densidad



Función de distribución



Función de fiabilidad



Función tasa de fallo

**Ejemplo 54** La función de distribución de tiempo de fallos (en horas) de cierto motor para aviones se ajusta al modelo de distribución exponencial:

$$F(t) = 1 - e^{-\frac{t}{10000}}$$

Hallar:

- a) la funciones de fiabilidad, de densidad, la tasa de fallo y la vida media de estos motores.
- b) Calcular la probabilidad de que estos motores duren, al menos, 100 horas.
- c) Calcular la probabilidad de que duren, al menos, 10100 horas si ya han durado 10000.

a)

$$R(t) = 1 - F(t) = e^{-\frac{t}{10000}},$$

$$f(t) = \frac{dF(t)}{dt} = \frac{1}{10000} e^{-\frac{t}{10000}},$$

$$h(t) = \frac{f(t)}{R(t)} = \frac{1}{10000} = 0.0001$$

$$\begin{aligned} \mu &= \int_0^{\infty} R(t) dt = \int_0^{\infty} e^{-\frac{t}{10000}} dt = -10000 e^{-\frac{t}{10000}} \Big|_0^{\infty} = \\ &= 0 - \left( -10000 e^{-\frac{0}{10000}} \right) = 10000 \text{ horas.} \end{aligned}$$

b)

$$R(100) = e^{-\frac{100}{10000}} = 0.99005 \quad \implies \quad F(100) = 1 - 0.99005 = 0.00995$$

c)

$$\begin{aligned} P(\tau < 10100 / \tau > 10000) &= \frac{P(10000 < \tau < 10100)}{P(\tau > 10000)} = \\ &= \frac{(1 - e^{-\frac{10100}{10000}}) - (1 - e^{-\frac{10000}{10000}})}{e^{-\frac{10000}{10000}}} = 0.00995. \end{aligned}$$

La coincidencia de estos dos últimos resultados confirman la propiedad de “falta de memoria” que tiene la distribución exponencial. Se observa que la probabilidad de durar 100 horas más es la misma para los motores nuevos que para los que ya han durado 10000 horas. Este hecho se traduce, desde el punto de vista práctico, por la idea de que los elementos no sufren desgaste. Este comportamiento suele darse en los dispositivos electrónicos.

## 9.5 EJERCICIOS PROPUESTOS

**Ejercicio 105** Comprueba la propiedad de falta de memoria, ( $P(\tau > t + h / \tau > t) = P(\tau > h)$ ), de la distribución exponencial y que esta propiedad no se cumple si la distribución es uniforme con función de densidad:

$$\begin{cases} f(t) = 0.1 & \text{si } 0 \leq t \leq 10 \\ f(t) = 0 & \text{en el resto} \end{cases}$$

**Ejercicio 106** Calcular las funciones de fiabilidad, in fiabilidad y de densidad que corresponden a una función tasa de fallo  $h(t) = t$  con  $t \geq 0$ .

**Ejercicio 107** El tiempo de vida de unos dispositivos sigue una distribución Normal de media 5000 horas y desviación típica 500 horas.

1. Calcular la función de fiabilidad y la probabilidad de que uno de estos dispositivos dure al menos 4500 horas.



2. Si se sabe que uno de estos dispositivos ya ha durado 4500 horas, ¿Cuál es la probabilidad de que dure por lo menos 500 horas más?

**Ejercicio 108** El tiempo de vida de unos dispositivos sigue una distribución exponencial de media 5000 horas

1. Calcular la función de fiabilidad y la probabilidad de que uno de estos dispositivos dure al menos 4500 horas.
2. Si se sabe que uno de estos dispositivos ya ha durado 4500 horas, ¿Cuál es la probabilidad de que dure por lo menos 500 horas más?

**Ejercicio 109** La función de densidad (en horas) de la duración en funcionamiento de ciertos componentes es  $f(t) = 9te^{-3t}$  para  $t > 0$ .

1. Hallar la vida media de dicho componente
2. Hallar la probabilidad de que un componente dure al menos 1 hora.
3. Si al principio había 10000 elementos, ¿cuantos se esperan que sobrevivan despues de la primera hora
4. Hallar la probabilidad de que un componente que haya durado ya 1 hora, resista todavía sin fallar por lo menos una hora más.
5. Hallar la función tasa de fallo y su valor para  $t = 1$  por hora y por minuto
6. Calcular la probabilidad de que sobrevivan un minuto los que ya han durado una hora.

**Ejercicio 110** Suponiendo que la distribución del tiempo de duración sin fallos del disco duro de un ordenador sigue una distribución uniforme en el intervalo de tiempo de 100 horas a 1500 horas se pide:

1. Encontrar las funciones de distribución de fiabilidad y tasa de fallo.
2. Representar gráficamente la función tasa de fallo en dicho intervalo. ¿Se puede deducir que sufre desgaste este tipo de disco duro?
3. ¿Cuál es la probabilidad de que un disco de estas características dure 500 horas?
4. ¿Cuál es la probabilidad de que un disco de estas características y que ya haya durado 500 horas, dure todavía 500 horas más ?

**Ejercicio 111** Un componente electrónico tiene una función tasa de fallo constante e igual a 0.005 fallos/hora. Calcular:

1. Su función de fiabilidad
2. Su vida media
3. La probabilidad de que este componente dure más de 125 horas
4. Si un componente de este tipo ya ha durado 125 horas, ¿Cuál es la probabilidad de que dure 125 horas más?

**Ejercicio 112** El tiempo en horas que la batería de una calculadora mantiene su carga es una variable aleatoria  $T$ . Suponemos que esta variable aleatoria sigue una distribución cuya función densidad es  $f(t) = 0.02 t e^{-0.01t^2}$

1. ¿Cuál es la función de fiabilidad? ¿Cuál es la fiabilidad para  $t = 12$  horas?
2. ¿Cuál es la probabilidad de que la batería dure al menos 3 horas?
3. Calcular la función tasa de fallo, indicando si es una función creciente o decreciente

**Ejercicio 113** Suponiendo que una distribución de tiempo de fallo esta dado por una distribución uniforme:

$$\begin{cases} f(t) = \frac{1}{5} & \text{si } 0 \leq t \leq 5 \\ f(t) = 0 & \text{en el resto} \end{cases}$$

1. Determinar la función de in fiabilidad
2. Determinar la función de fiabilidad
3. Calcular la probabilidad de que las unidades que se ajusten a esta distribución duren entre 3 y 4 horas
4. Determinar la función tasa de fallo e indicar si el modelo sería adecuado para piezas que sufran desgaste.

**Ejercicio 114** A partir de los datos de la siguiente tabla, correspondiente al momento en que han fallado ciertos dispositivos, calcular un valor aproximado para la tasa de fallo por minuto, correspondiente a cada intervalo

Intervalo de tiempo	1ª hora	2ª hora	3ª hora	4ª hora
nº de fallos en el intervalo	30	20	15	10
Elementos supervivientes al principio del intervalo	500	470	450	435

## Tema 10

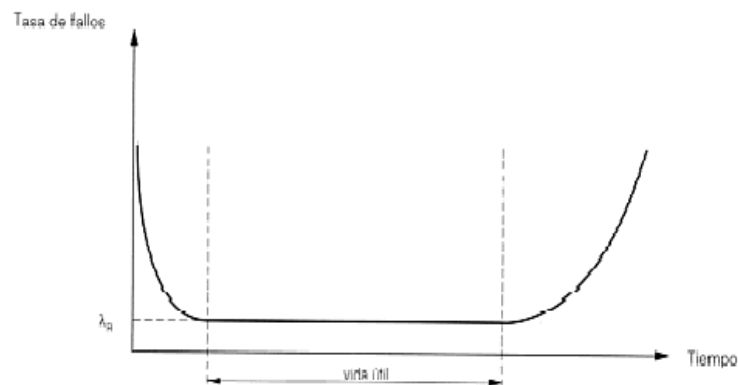
# Distribuciones de tiempos de fallos

### 10.1 Introducción. La curva de bañera.

El tiempo hasta el fallo de un cierto dispositivo es una variable aleatoria. Para aplicar un modelo aleatorio se ha de usar alguna distribución concreta. Entre las distribuciones que se aplican más frecuentemente en fiabilidad se cuentan exponenciales, normales, gamma, log-normal, etc. Estudiaremos en este tema algunas de estas distribuciones, haciendo especial hincapié en su uso en fiabilidad.

### 10.2 La curva de bañera

La curva tasa de fallo tiene frecuentemente forma de bañera, tal como la representada en la figura :



La primera parte, región con tasa de fallo decreciente, corresponde a los

fallos infantiles. Se suele llamar *periodo de mortalidad infantil*. Al siguiente periodo, donde la tasa es más o menos constante se le suele dar el nombre de *periodo de vida útil*. Al último periodo, con tasa de fallo creciente, se le conoce con el nombre de *periodo de desgaste*. Por lo general se emplea una función diferente para cada una de estas fases. Para el periodo infantil hay que usar funciones con tasa de fallo decreciente como, por ejemplo, algunos tipos de la distribución gamma o de Weibull. Para el periodo de vida útil lo más frecuente es usar una exponencial. Por último, en el periodo de desgaste hay una tasa de fallo creciente. Se emplean las normales y ciertas formas de las distribuciones de Weibull y de la log-normal.

### 10.3 La distribución exponencial

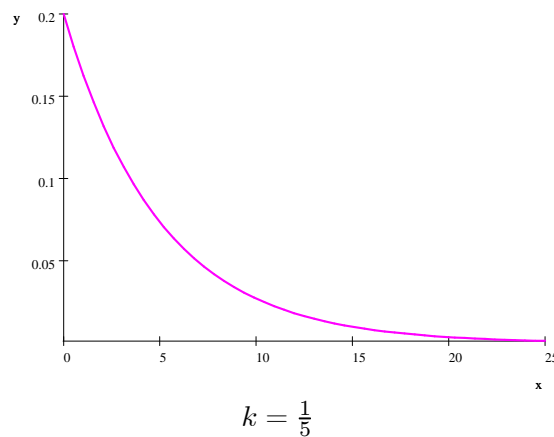
Si el dispositivo está en periodo de vida útil es adecuado suponer que la tasa de fallo es constante ( $h(t) = k$ ). En este caso

$$R(t) = \exp\left(-\int_0^t h(t) dt\right) = \exp\left(-\int_0^t k dt\right) = \exp(-kt) \quad (10.1)$$

y

$$f(t) = F'(t) = (1 - R(t))' = (1 - \exp(-kt))' = k \exp(-kt) \text{ y } t > 0 \quad (10.2)$$

que corresponde a la función densidad de la distribución exponencial. En la siguiente figura se muestra la representación gráfica de una exponencial con el valor  $k = \frac{1}{5}$ .



El calculo de la vida media es

$$\mu = E(\tau) = \int_0^{\infty} t f(t) dt = \int_0^{\infty} R(t) dt = \int_0^{\infty} \exp(-kt) dt = \frac{1}{k} \quad (10.3)$$

la tasa de fallo de esta distribución es constante:

$$h(t) = \frac{f(t)}{R(t)} = \frac{k \exp(-kt)}{\exp(-kt)} = k$$

Esta propiedad nos indica que los componentes que se ajusten a este modelo no sufren desgaste, es decir que tienen la misma probabilidad de fallar en un instante dado los componentes nuevos que los usados, independientemente del tiempo que lleven funcionando.

Otra forma de manifestación analítica de esta característica es la propiedad de “carencia de memoria” propiedad, que se concreta en la siguiente relación:

$$P(\tau > t_0 + h/\tau > t_0) = P\tau > h)$$

es decir la probabilidad de sobrevivir un intervalo de tiempo  $h$  es igual para un componente nuevo que para otro que ya ha durado hasta  $t_0$ .

En efecto:

$$P(\tau > t_0 + h/\tau > t_0) = \frac{R(t_0+h)}{R(t_0)} = \frac{\exp(-k(t_0+h))}{\exp(-kt_0)} = \exp(-kh) = P(t > h)$$

La coincidencia de ambas probabilidades confirma la propiedad de “falta de memoria” de la distribución exponencial.

Para que sea aplicable esta distribución el dispositivo ha de ser insensible a la edad y al uso. Se puede aplicar en los casos siguientes

1. En componentes cuya fiabilidad se está estudiando en el periodo de vida que corresponde a la parte central de la curva de bañera.
2. En componentes purgados (que ya han pasado el periodo infantil o que han sido sometidos a ensayos previos) y de vida útil muy larga. En este caso están bastantes componentes electrónicos.
3. En componentes purgados a los que se someten a reemplazos preventivos antes de que lleguen al periodo de desgaste.

## 10.4 La distribución normal

Esta distribución puede usarse en los periodos de desgaste, por ejemplo en la mortalidad humana en la edad senil. También es un modelo adecuado para modelar el tiempo de vida de las lámparas eléctricas.

En este caso

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \quad (10.4)$$

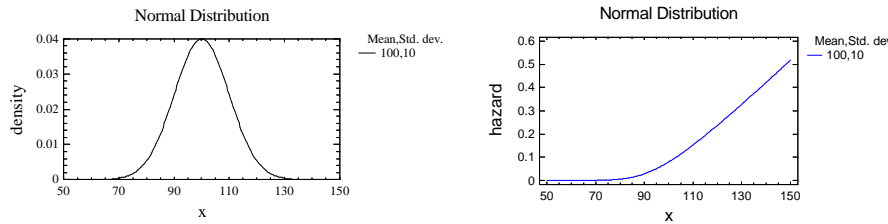
$$F(t) = \int_{-\infty}^t \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx ; \quad R(t) = \int_t^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (10.5)$$

La vida media es  $\mu$ , y la tasa de fallo es

$$h(t) = \frac{f(t)}{R(t)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\int_t^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt} \quad (10.6)$$

que es una función creciente.

Las siguientes gráficas representan la función densidad y la función tasa de fallo de una distribución normal con  $\mu = 100$ ,  $\sigma = 10$ .



El empleo de la distribución normal para modelar tiempos de fallos presenta la anomalía de que admite tiempo de fallos negativos. Esto no es un inconveniente grave con tal que el origen de los tiempos este suficientemente a la izquierda del valor medio, de modo que la cola de la distribución a la izquierda del origen sea suficientemente pequeña. Para ello es necesario que la media sea mayor que 3 o 4 veces la desviación típica y usamos la normal como un modelo aproximado.

**Ejemplo 55** Suponiendo que la vida de unas lamparas está normalmente distribuida con una media de 10000 horas y una desviación típica de 1000 horas.

a) Hallar la probabilidad de que una lámpara que ya ha durado 9000 horas dure, al menos, 500 horas más.

b) Idem si ya ha durado 10000 horas.

$$\begin{aligned} \text{a) } P(\tau > 9500 / \tau > 9000) &= \frac{P(\tau > 9500)}{P(\tau > 9000)} = \frac{P(z > \frac{9500-10000}{1000})}{P(z > \frac{9000-10000}{1000})} = \\ &= \frac{P(z > -0.5)}{P(z > -1)} = \frac{0.69146}{0.84134} = 0.82186 \end{aligned}$$

es la probabilidad que tienen las lámparas que ya han durado 9000 horas de sobrevivir 500 horas más.

$$\begin{aligned} \text{b) } P(\tau > 10500 / \tau > 10000) &= \frac{P(\tau > 10500)}{P(\tau > 10000)} = \frac{P(z > \frac{10500-10000}{1000})}{P(z > \frac{10000-10000}{1000})} = \\ &= \frac{P(z > 0.5)}{P(z > 0)} \\ &= \frac{1-F(0.5)}{0.5} = \frac{0.30854}{0.5} = 0.61708. \end{aligned}$$

es la probabilidad que tienen de sobrevivir 500 horas más las lámparas que ya han durado 10000 horas.

Como puede observarse la probabilidad de vivir 500 horas más disminuye con la edad. Este comportamiento es propio del periodo de desgaste y corresponde a tasa de fallo creciente.

A veces se considera únicamente como función de densidad la parte positiva del eje de los tiempos de la normal. En este caso se requiere renormalizar la función para que el área que queda bajo su curva sea la unidad, condición requerida para poder ser considerada una función de densidad de probabilidad. Hablamos entonces de la *distribución normal truncada*. Si el truncamiento se hace en el valor 0 la función de densidad es

$$f(t/t > 0) = \frac{f(t)}{1-F(0)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{1 - \int_{-\infty}^0 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt}$$

La media de la distribución normal truncada inferiormente en 0 es:

$$\mu + \sigma \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu)^2}{2\sigma^2}}}{1 - \int_{-\infty}^0 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt} = \mu + \sigma h(0)$$

donde  $h(0)$  es el valor de la función tasa de fallo de la  $N(\mu, \sigma)$  completa en el origen.

La varianza de la normal truncada inferiormente en el origen puede calcularse como:

$$\sigma^2 \left[ 1 - \frac{\mu}{\sigma} h(0) - (h(0))^2 \right].$$

## 10.5 La distribución log-normal

La distribución log-normal es la distribución de una variable  $t$  cuyo logaritmo neperiano  $x = \ln t$  sigue una distribución normal. Si la media de la normal es  $\mu'$  y la desviación típica  $\sigma'$ , haciendo el cambio de variable  $x = \ln t$  en la función de distribución de la normal se obtiene la función de densidad de la lognormal que resulta ser:

$$f(t) = \frac{1}{t\sigma'\sqrt{2\pi}} e^{-\frac{(\ln t - \mu')^2}{2\sigma'^2}}, \quad t > 0, \mu', \sigma' > 0 \quad (10.7)$$

La media de la distribución log-normal, es decir de la variable  $t$ , es:

$$\mu = \exp\left(\mu' + \frac{\sigma'^2}{2}\right) \quad (10.8)$$

y su varianza

$$\sigma^2 = (\exp \sigma'^2 - 1) \exp(2\mu' + \sigma'^2) \quad (10.9)$$

Suele emplearse para modelar los fallos ocurridos en la primera fase de la vida de un componente (periodo Infantil). Es útil para estudiar la fatiga de componentes metálicos, la duración de los aislamientos eléctricos, y en especial para la implantación de los programas de mantenimiento, ya que se usa también para modelar el tiempo de duración de las reparaciones.

**Ejemplo 56** Si los parámetros de una distribución lognormal son  $\mu' = 4$  y  $\sigma' = 2$ , calcular su media y hallar la probabilidad de que un componente, cuya duración sin fallos se rige por esta distribución dure más que su vida media

$$\begin{aligned} \mu &= \exp\left(\mu' + \frac{\sigma'^2}{2}\right) = \exp\left(4 + \frac{2^2}{2}\right) = e^6 = 403.43 \\ P(t \geq 403.43) &= P(\ln t \geq \ln 403.43) = P\left(\frac{\ln t - \mu'}{\sigma'} \geq \frac{\ln 403.43 - 4}{2}\right) = \\ &= 1 - P(Z \leq 1.0) = 1 - 0.84 = 0.16 \end{aligned}$$

## 10.6 La distribución de Weibull

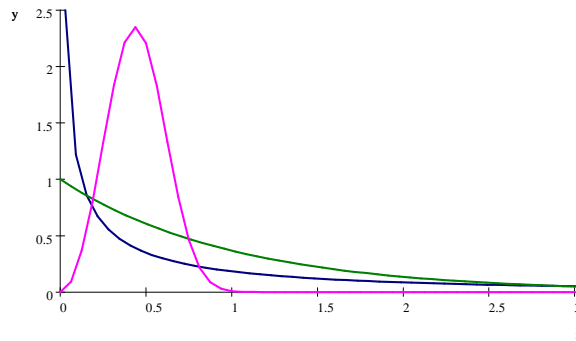
El investigador sueco Weibull propuso esta función para el estudio de la fatiga de los metales (1939). Posteriormente se ha aplicado (H.J. Kao de la universidad de Cornell, 1950) al estudio de tiempo de vida de tubos electrónicos). Esta distribución se emplea mucho en fiabilidad por su versatilidad (puede aplicarse a las tres partes de la curva de bañera) y, además, porque es fácil su manejo mediante el llamado *papel probabilístico de Weibull*.



Su función de densidad es

$$f(t) = \frac{\beta}{\eta^\beta} (t - \gamma)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta}, \eta, \gamma > 0, t > \gamma \quad (10.10)$$

Los parámetros de esta distribución  $\gamma$ ,  $\eta$ , y  $\beta \geq 0$  se conocen, respectivamente, como parámetro de posición, de escala y de forma. El parámetro  $\gamma$  suele ser el origen de los tiempos y frecuentemente toma el valor cero. En la figura pueden verse ejemplos de gráficas de la función densidad de la Weibull para distintos valores de los parámetros



Funciones de densidad de Weibull  
 $(\beta, \eta, \gamma) = (0.5, 1, 0), (1, 1, 0), (3, 0.5, 0)$

La función de distribución es

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad (10.11)$$

así que la función de fiabilidad es

$$R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad (10.12)$$

La vida media y la varianza de esta distribución son

$$\mu = \gamma + \eta \Gamma\left(1 + \frac{1}{\beta}\right) \quad \sigma^2 = \eta^2 \Gamma\left(1 + \frac{2}{\beta}\right) - \mu^2 \quad (10.13)$$

donde  $\Gamma$  indica la función *gamma completa* que puede ser obtenida numéricamente con algún programa de ordenador o recurriendo a tablas, que están disponibles a veces en el mismo papel de Weibull, y que se define como:

$$\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$$

donde  $a$  debe ser real y mayor o igual que 1.

No obstante es fácil de calcular en algunos casos particulares: Si  $a = 1$ ,  $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$ . En el caso de que  $a$  sea un número entero  $n$ , descomponiendo esta integral por partes se obtiene la fórmula de recurrencia:

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx = (n-1) \int_0^\infty e^{-x} x^{n-2} dx = (n-1) \Gamma(n-1)$$

de donde se deduce que

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx = (n-1)! \quad (10.14)$$

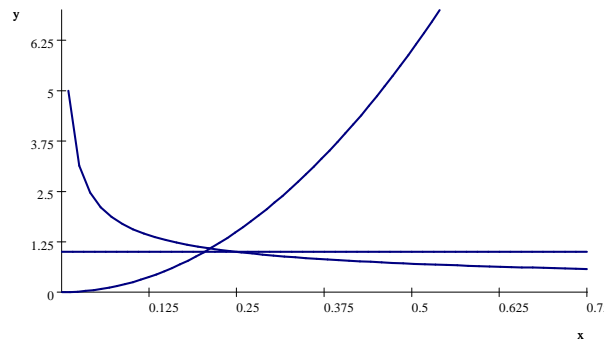
La tasa de fallo de la distribución de Weibull toma la forma:

$$h(t) = \frac{\beta}{\eta^\beta} (t - \gamma)^{\beta-1} \quad (10.15)$$

Para estudiar su variación, calculamos su función derivada:

$$h'(t) = \frac{\beta}{\eta^\beta} (\beta - 1) (t - \gamma)^{\beta-2} \quad (10.16)$$

Como el origen de los tiempos se toma en  $\gamma$ ,  $t - \gamma$  es positivo. Se observa que si  $\beta > 1$  la derivada de la tasa de fallos sería positiva y por tanto la tasa de fallo sería creciente. Si  $\beta < 1$  la tasa de fallo sería decreciente y si  $\beta = 1$  la tasa de fallo sería constante. En este último caso lo que obtenemos es una exponencial. En la gráfica están representadas las funciones tasa de fallo correspondientes a las funciones de Weibull de la gráfica de la página 249, y que corresponden a cada uno de los modelos con tasa de fallo creciente, decreciente o constante.



Tasas de fallo de Weibull  
 $(\beta, \eta, \gamma) = (0.5, 1, 0), (1, 1, 0), (3, 0.5, 0)$

**Ejemplo 57** *El tiempo de vida de un componente se distribuye como una variable de Weibull de parámetros  $\gamma=0$ ,  $\beta=2$ . Se ha observado que entre los componentes que sobrepasan las 90 horas el 15% fallaba antes de las 100 horas. Estimar, a partir de estos datos, el valor del restante parametro de la distribución.*

$$P(90 < t < 100 / t > 90) = 0.15$$

$$0.15 = \frac{P(90 < t < 100)}{P(t > 90)} = \frac{1 - e^{-\left(\frac{100}{\eta}\right)^2} - 1 + e^{-\left(\frac{90}{\eta}\right)^2}}{e^{-\left(\frac{90}{\eta}\right)^2}} = 1 - e^{-\left(\frac{100}{\eta}\right)^2 + \left(\frac{90}{\eta}\right)^2}$$

$$\ln 0.85 = -\left(\frac{100}{\eta}\right)^2 + \left(\frac{90}{\eta}\right)^2 = -\frac{1900}{\eta^2}$$

Resulta

$$\frac{1}{\eta^2} = \frac{\ln 0.85}{-1900} = 8.5536 \times 10^{-5}$$

La solución es  $\eta = 108.12$ .

### 10.6.1 Uso del papel probabilístico de Weibull

De la función de distribución de Weibull

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad (10.17)$$

se deduce que

$$e^{\left(\frac{t-\gamma}{\eta}\right)^\beta} = \frac{1}{1 - F(t)} \quad (10.18)$$

Aplicando dos veces logaritmos se obtiene

$$\ln \ln \left( \frac{1}{1 - F(t)} \right) = \beta \ln(t - \gamma) - \beta \ln \eta \quad (10.19)$$

Si  $\gamma = 0$ , llamando  $Y = \ln \ln \left( \frac{1}{1 - F(t)} \right)$  y  $X = \ln t$  obtenemos la ecuación de una recta

$$Y = \beta X - \beta \ln \eta \quad (10.20)$$

Gracias a esta transformación, la distribución de Weibull puede representarse por medio de una recta en papel de Weibull, que es un papel con

una cuadrícula en el que la escala vertical es  $\ln \ln \left( \frac{1}{1-F(t)} \right)$  y la horizontal es  $\ln t$ . Puede verse un papel con estas características en la página 260. Obsérvese que a la derecha aparecen algunos valores de la función gamma que puede emplearse para calcular el valor medio de la Weibull.

De las expresiones 10.19 y 10.20 se deduce que  $\beta$  es la pendiente de la recta y que el parámetro  $\eta$  es el valor de  $t$  en el corte de la recta con la abcisas, ya que si  $Y = \ln \ln \left( \frac{1}{1-F(t)} \right) = 0$ , es decir  $F(t) \approx 0.63$ , entonces  $X = \ln \eta$ ,  $t = \eta$ . Si tenemos representada la recta podemos emplearla para calcular un valor aproximado de  $F(t)$  para cada  $t$ .

Si  $\gamma \neq 0$ , pero conocido, puede emplearse todo el razonamiento anterior realizando un cambio de origen en los datos de tiempo, tal como se hace en el siguiente ejemplo:

**Ejemplo 58** *Los siguientes valores son los tiempo en que fallaron unos dispositivos en el laboratorio. Ajustar gráficamente una distribución de Weibull sabiendo que: 52, 62, 70, 78, 86, 94, 104, 115, 130. Tomad  $\gamma = 30$ .*

Lo primero que hacemos es representar en el papel de Weibull los puntos que corresponden a la función de distribución empírica de la muestra. Dichos valores se han representado en la gráfica 10.2 de la página 261:

$t - 30$	22	32	40	48	56	64	74	85	100
Nº de orden de la avería	1	2	3	4	5	6	7	8	9
$F(t - 30)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$

Haciendo un desplazamiento del origen de coordenadas pueden representarse en papel de Weibull los puntos de coordenadas  $(t - 30)$  y  $F(t - 30)$ . Para efectos del dibujar los puntos en el papel consideraremos  $t - 30$  como si fuera  $t$  y  $F(t - 30)$  como si fuera  $F(t)$ . Se representan estos puntos mirando los valores que aparecen en la escala horizontal inferior (escala de  $t$ ) y vertical derecha (escala para  $F(t)$  en porcentaje). Los valores de la variable  $F(t - 30)$  corresponden a la función de distribución empírica. En este ejemplo hemos empleado como aproximación de esta función la expresión  $\frac{i}{n+1}$ , donde  $i$  es el número de orden del valor correspondiente dentro de la muestra y  $n$  el número total de elementos de ésta (a veces se usa como aproximación la expresión  $\frac{i-0.3}{n+0.4}$  que se conoce con el nombre de *Rango medio*) De esta forma obtenemos una serie de puntos, que si la distribución es de Weibull, deben ajustarse visualmente a una recta.

El valor de  $\beta$  es la pendiente de la recta y  $\ln \eta$  es la abcisa del punto de intersección de la recta con el eje de abcisas. Para obtener estos valores hay que relacionar la recta con su propio sistema de coordenadas,  $X$  e  $Y$ .

Tal sistema está representado en la escala derecha y superior del papel. Las coordenadas X e Y de los puntos representados, aunque no sea necesario calcularlas, están en la siguiente tabla para facilitar su interpretación.

$\ln(t - 30)$	3.09	3.47	3.69	3.87	4.03	4.16	4.30	4.44	4.61
$\ln \ln \left( \frac{1}{1-F(t-30)} \right)$	-2.25	-1.51	-1.03	-0.67	-0.37	-0.09	0.19	0.48	0.83

estos valores corresponden a los mismos puntos representados anteriormente, aunque referidos al sistema de coordenadas X, Y. Considerando la gráfica de la recta sobre estos ejes, se ha estimado que la pendiente de la recta es  $\beta \approx 2$ , y que  $\eta \approx 65$ .

Si  $\gamma \neq 0$  y desconocido, la distribución de Weibull se representaría en dicho papel como la función

$$Y = \beta \ln(e^X - \gamma) - \beta \ln \eta \tag{10.21}$$

Como  $\beta > 0$

$$\lim_{x \rightarrow \ln \gamma} Y = -\infty \tag{10.22}$$

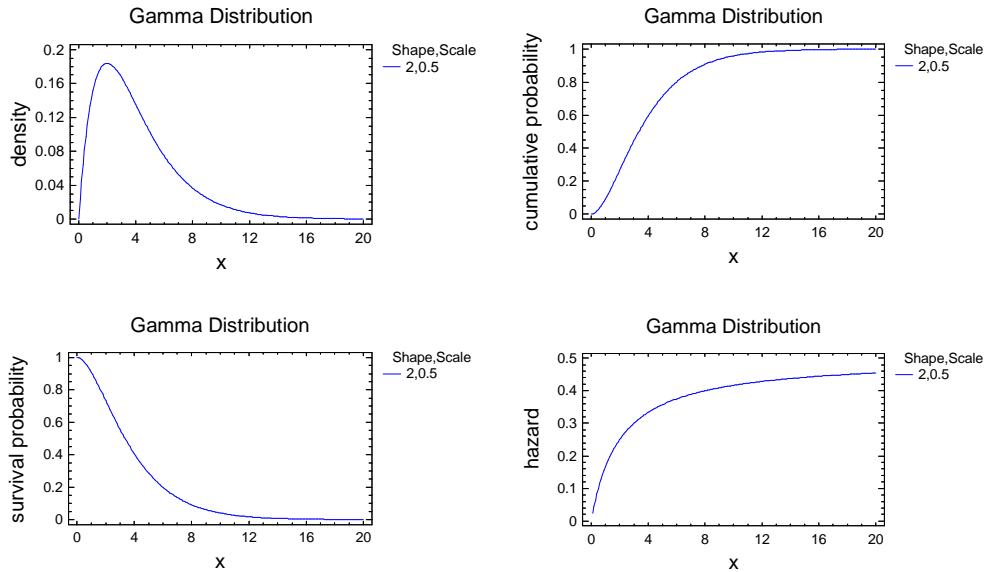
aprovechamos este hecho para hallar un valor aproximado para  $\gamma$ , ya que  $x = \ln \gamma$  será una asíntota vertical. Esto se hará si se observa que la representación de los muestras sobre el papel de Weibull parece seguir el trazado de una curva y no de una recta. En este caso el valor de X correspondiente a la asíntota vertical de la curva, que se dibujará a mano alzada, sería  $\ln \gamma$ . Una vez estimado el valor de  $\gamma$ , se puede operar como en el caso del ejemplo 58, en el que el valor de  $\gamma$  era conocido.

### 10.7 La distribución gamma

La distribución gamma tiene por función de densidad

$$f(t; a, \lambda) = \frac{\lambda}{\Gamma(a)} (\lambda t)^{a-1} e^{-\lambda t}, \quad a, \lambda > 0, \quad t > 0 \tag{10.23}$$

siendo  $\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt$ . Los parámetros  $a$  y  $\lambda$  se llaman respectivamente parámetros de forma y de escala. La *media* de esta distribución es  $\frac{a}{\lambda}$  y la *varianza* es  $\frac{a}{\lambda^2}$ . Si  $a < 1$  la tasa de fallo es decreciente, si  $a = 1$  la tasa de fallo es constante y si  $a > 1$  la tasa de fallo es creciente. Las siguientes gráficas son las representaciones de la función de densidad, de infijabilidad, de fiabilidad y la tasa de fallo de la distribución gamma de parámetros  $a = 2$ ,  $\lambda = \frac{1}{2}$



Si el parámetro de forma es entero se conoce con el nombre de *distribución de Erlang*.

La distribución de Erlang se utiliza para modelar el tiempo que transcurre desde que sucede un acontecimiento hasta que suceden los  $n$  acontecimientos siguientes si el tiempo transcurrido entre dos acontecimientos consecutivos se rige por una distribución exponencial. Por ejemplo el tiempo que pasa desde que ocurre un fallo hasta que ocurre el siguiente se rige por una exponencial (Erlang con  $n = 1$ ). Si consideramos un sistema donde se va sustituyendo sucesivamente un componente que ha fallado por otro componente idéntico (repuesto), y queremos modelar el tiempo transcurrido entre fallos no consecutivos, podemos usar una distribución de Erlang ( $n > 1$ ). El parámetro  $n$  es el número de componentes fallados contando a partir del momento inicial que se haya considerado.

## 10.8 El Test Chi cuadrado de bondad de Ajuste

A veces nos preguntaremos cual es el modelo más adecuado para una serie de datos que representen el tiempo hasta el fallo de un conjunto de dispositivos idénticos. Una forma de decidirse por un modelo es el *test chi-cuadrado de bondad de ajuste a distribuciones*, ya tratado en el epígrafe 5.14, que puede aplicarse de la forma siguiente:

Si se dispone de una muestra de  $n$  elementos (al menos 20, aunque este número depende de la precisión deseada) se agrupan en  $k$  clases, como en los histogramas de frecuencias. Si llamamos  $n_i$  a la frecuencia observada en la

clase  $i$ ,  $p_i$  a la probabilidad que correspondería a esta clase en la distribución teórica que deseamos contrastar, y, por tanto,  $np_i$  al número de elementos que teóricamente debería caer en dicha clase. (el número esperado de elementos contenidos en cada clase,  $np_i$ , debe ser al menos 5), entonces el estadístico

$$\chi_{\text{exp}}^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \quad (10.24)$$

sigue (aproximadamente) una distribución  $\chi^2$  con  $k-p-1$  grados de libertad, donde  $p$  es el número de parámetros de la distribución propuesta que se hayan estimado por medio de estadísticos muestrales. El valor de  $\chi_{\text{exp}}^2$  no puede ser negativo. Será nulo si hay un acuerdo perfecto entre los valores experimentales y teóricos. Es obvio que los valores teóricos se aproximan mejor a los experimentales mientras más cerca de cero este el valor  $\chi_{\text{exp}}^2$ .

Rechazamos la distribución propuesta, al nivel de significación  $\alpha$ , si el estadístico  $\chi^2(\text{exp}) > \chi_{\alpha, k-p-1}^2$ , siendo este último valor el de chi-cuadrado con  $n-p-1$  grados de libertad que corresponde al nivel de significación  $\alpha$ , es decir  $F(\chi_{\alpha, k-p-1}^2) = 1 - \alpha$

Si el número de elementos no es suficiente para realizar este test puede emplearse el test de bondad de ajuste de *Kolmogorov-Smirnov* que, como el test chi-cuadrado, puede encontrarse implementado en la mayor parte de los paquetes estadísticos.

**Ejemplo 59** *Se someten a ensayo hasta el fallo a 95 componentes y se obtiene el tiempo hasta el fallo en horas. Se han resumido los datos en la tabla siguiente:*

<i>Horas</i>	<i>Marcas de clase</i>	<i>Frecuencia absoluta</i>
<i>4000-4500</i>	<i>4250</i>	<i>3</i>
<i>4500-5000</i>	<i>4750</i>	<i>6</i>
<i>5000-5500</i>	<i>5250</i>	<i>18</i>
<i>5500-6000</i>	<i>5750</i>	<i>20</i>
<i>6000-6500</i>	<i>6250</i>	<i>24</i>
<i>6500-7000</i>	<i>6750</i>	<i>16</i>
<i>7000-7500</i>	<i>7250</i>	<i>7</i>
<i>7500-8000</i>	<i>7750</i>	<i>1</i>
		<i>95= Total</i>

*Usa el test chi- cuadrado de bondad de ajuste para comprobar si los datos pueden considerarse procedentes de una distribución normal.*

La media de la muestra es

$$\bar{x} = \frac{4250 \times 3 + 4750 \times 6 + 5250 \times 18 + 5750 \times 20 + 6250 \times 24 + 6750 \times 16 + 7250 \times 7 + 7750 \times 1}{95} = 5971.05$$

y la cuasidesviación:

$$s = \sqrt{\frac{(4250-5971.05)^2 \times 3 + (4750-5971.05)^2 \times 6 + \dots + (7250-5971.05)^2 \times 7 + (7750-5971.05)^2 \times 1}{94}} = 760.44$$

En la tabla siguiente aparecen agrupados los dos primeros intervalos y los dos últimos, para que en ninguna de las casillas haya una frecuencia menor que 5. La table incluye las frecuencias experimentales y las teóricas si la distribución de los fallos fuera efectivamente una  $N(5971.05, 760.44)$ .

Horas	Frecuencia teórica	Frecuencia absoluta
4000-5000	9.1223	9
5000-5500	15.865	18
5500-6000	23.501	20
6000-6500	22.941	24
6500-7000	14.756	16
7000-8000	7.998	8
		95= Total

Detallamos como se obtiene la columna de frecuencias teóricas:

$$Pr(4000 < x < 5000) = pr\left(\frac{4000-5971.05}{760.4} < z < \frac{5000-5971.05}{760.4}\right) = pr(-2.5921 < z < -1.277) = F(-1.277) - F(-2.5921) = 0.09603$$

$$np_1 = 0.09603 \times 95 = 9.123$$

$$Pr(5000 < x < 5500) = F(5500) - F(5000; 5971.05, 760.4) = 0.167$$

$$np_2 = 0.167 \times 95 = 15.865$$

$$Pr(5500 < x < 6000) = F(6000) - F(5500) = 0.24738$$

$$np_3 = 0.24738 \times 95 = 23.501$$

$$Pr(6000 < x < 6500) = F(6500) - F(6000) = 0.24148$$

$$np_4 = 0.24148 \times 95 = 22.941$$

$$Pr(6500 < x < 7000) = F(7000) - F(6500) = 0.15533$$

$$np_5 = 0.15533 \times 95 = 14.756$$

$$Pr(7000 < x < 8000) = F(8000) - F(7000) = 8.4189 \times 10^{-2}$$

$$np_6 = 8.4189 \times 10^{-2} \times 95 = 7.9980$$

Observamos que se cumple la propiedad requerida de que ninguna frecuencia esperada (teórica) sea menor que 5.

El valor de la chi experimental se obtiene

$$\chi_{6-3}^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} = \frac{(9 - 9.123)^2}{9.808} + \frac{(18 - 15.865)^2}{15.7681} + \frac{(20 - 23.501)^2}{22.7059} + \frac{(24 - 22.941)^2}{22.207} + \frac{(16 - 14.756)^2}{14.7428} + \frac{(8 - 7.998)^2}{8.67934} = 0.98591$$

La chi cuadrado con 3 grados de libertad (6 clases - 2 parámetros extraídos de la muestra - 1) correspondiente al nivel de significación 0,05 es 7.815.

$$\chi_{\alpha, k-p-1}^2 = \chi_{0.05, 6-2-1}^2 = 7.815$$



Como el valor experimental es menor que el teórico ( $0.98591 < 7.815$ ), no puede rechazarse que la distribución sea normal.

Los programas de ordenador suelen dar lo que se llama *p-value*, que en este caso toma el valor:

$$p - value = 1 - \chi_3^2(0.98591) = 1 - 0.19534 = 0.80466$$

Se acepta la hipótesis nula (la distribución de los fallos es de Poisson) si el *p-value* es mayor que  $\alpha$  (nivel de significación del test) Por tanto como  $0.80466 > 0.05$  no se puede rechazar la hipótesis de que la distribución sea Normal al nivel de significación 0.05 (confianza del 95%). Por tanto concluimos que hay un buen acuerdo entre los datos experimentales y los teóricos, si estos procedieran de una distribución Normal, por lo que aceptamos que los tiempos de fallo de los componentes, recogidos en la tabla, parecen regirse por una distribución de este tipo.

## 10.9 EJERCICIOS PROPUESTOS

**Ejercicio 115** *La función de densidad del tiempo de vida de un componente es exponencial:*

$$f(t) = 0.5e^{-0.5t}, \quad t \geq 0, \quad (t \text{ en meses})$$

1. *Calcula la vida media del componente, así como la tasa de fallo a los 3 meses.*
2. *Si ponemos en funcionamiento un lote de 1000 componentes simultáneamente, ¿Cuántos de estos se espera sobrevivan más de dos meses?*
3. *Cálcula y representa gráficamente las funciones de densidad, fiabilidad y tasa de fallo de esta distribución.*

**Ejercicio 116** *El nº de kilometros recorridos por un modelo de automovil antes que los parachoques resulten inservibles sigue un modelo de distribución lognormal. Se ha observado que el 5% de los parachoques fallan antes de que el vehículo haya recorrido 120000 km y que otro 5% falla despues de que el vehículo haya recorrido más de 180000 km.*

1. *Estimar la media y la desviación típica de la distribución lognormal.*
2. *Hallar el valor de la tasa de fallo a los 150000 km.*
3. *Si se han fabricado 9000 unidades de este automovil. ¿ Cuántos de ellos tendrán los parachoques rotos cuando hayan recorrido 150000 km.?*

**Ejercicio 117** La fiabilidad de los alternadores de unos automóviles es:

$$R(x) = e^{-\left(\frac{x}{180000}\right)^3}$$

El número de vehículos en que se ha instalado estos alternadores es de 100000. La variable  $x$  es el número de kilómetros recorridos antes de la avería.

Se pide:

1. ¿Cuántos alternadores se puede esperar que tengan averías antes de que hayan recorrido 60000Km.
2. Calcula la tasa de fallo a los 60000 km y a los 120000 km
3. Calcular la vida media de estos alternadores.
4. Si la garantía cubre las averías producidas en los primeros 15000 km. ¿Cuántos alternadores puede esperarse que habrá que reparar en garantía?

**Ejercicio 118** Cien unidades se han sometido a una prueba de vida hasta que han fallado. Se han obtenido los datos siguientes:

$t$  = tiempo en horas de duración de las unidades,

$n$  = Número de unidades que han durado este tiempo

$t$	0-100	100-200	200-300	300-400	400-500	Más de 500
$n$	50	18	17	8	4	3

¿Se puede admitir que una distribución exponencial con valor medio 160 horas representa razonablemente los tiempo de fallo del modelo del que proceden estos datos?

**Ejercicio 119** Los tiempos, en horas, de duración en funcionamiento de 20 baterías ha sido:

26, 32, 34, 39, 56, 71, 84, 88, 89, 95, 98, 113, 118, 119, 123, 127, 160, 219, 224, 242

Ajustar una distribución de Weibull a estos datos usando los procedimientos siguientes:

1. Por medio de papel de Weibull.
2. Numéricamente, realizando un ajuste de regresión lineal
3. Por medio del procedimiento **Distribution Fitting** de Statgraphics.

**Ejercicio 120** Los elementos fabricados por un cierto proceso tienen una duración (en meses) cuya función tasa de fallo viene dada por  $h(t) = t^2$  para  $t > 0$ . Hallar:

1. la función de densidad, de fiabilidad y de in fiabilidad
2. La probabilidad de que uno de estos dispositivos dure más de 1 mes
3. La producción de un mes es de 10000 elementos. Si hay que reponer todos los dispositivos que duren menos de medio mes. ¿ Cuántos elementos se puede esperar que haya que reponer de estos 10000?
4. Da una expresión para la vida media de estos dispositivos

**Ejercicio 121** El tiempo de fallo (en horas) de un dispositivo sigue una distribución de probabilidad cuya función de densidad es  $f(t) = \alpha^2 t e^{-\alpha t}$ ,  $t > 0$ ,  $\alpha > 0$ .

1. Calcular, en función de  $\alpha$  la probabilidad de que un componente que ya haya durado 100 horas dure 100 horas más.
2. El coste de producir un componente es proporcional al cuadrado de su vida media ( $k\mu^2$ ) y se estima que la ganancia obtenida por cada uno de estos componentes es de 48 euros por cada hora que funciona sin fallar. Calcular una expresión en función de  $\alpha$  y  $k$  para el beneficio medio obtenido con estos componentes.
3. Demostrar que el beneficio máximo que se obtiene corresponde a un valor de  $\alpha = \frac{K}{12}$  con un valor medio de  $\frac{576}{K}$  euros por unidad.

**Ejercicio 122** Un submarinista, que ha de reparar una plataforma petrolí-fica, puede elegir entre dos equipos de buceo. La reparación se realiza en condiciones peligrosas y el equipo de buceo puede fallar. La distribución del tiempo de fallo de los estos equipos sigue una distribución de Weibull de parámetros a)  $\eta = 1$ ,  $\beta = 2$  b)  $\eta = 2$ ,  $\beta = 1$

1. ¿Qué equipo debe usar si la reparación dura una hora?
2. Responder a la misma pregunta si ambos equipos han sido ya usados, sin fallos, durante 3.17 horas.

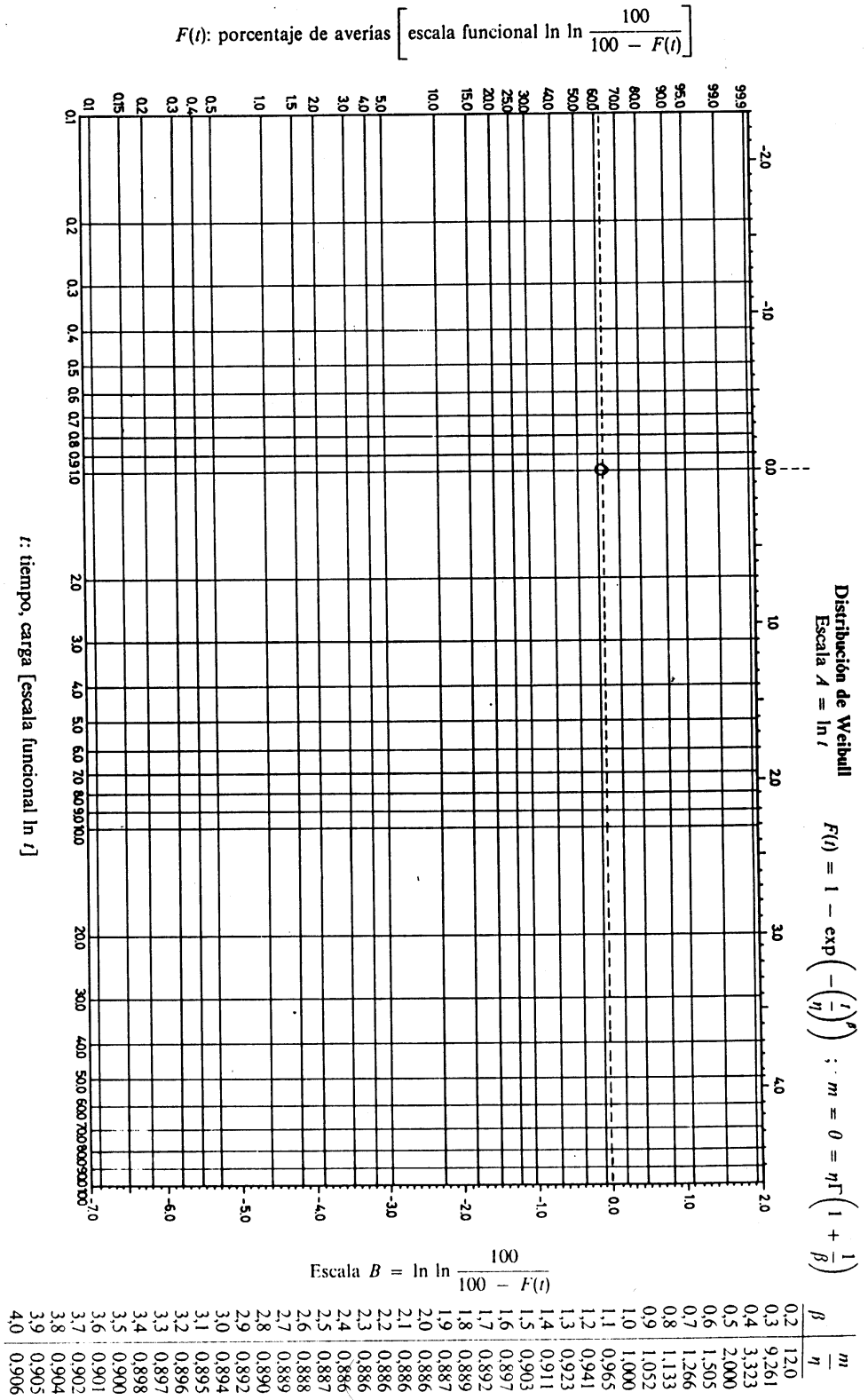


Figura 10.1:

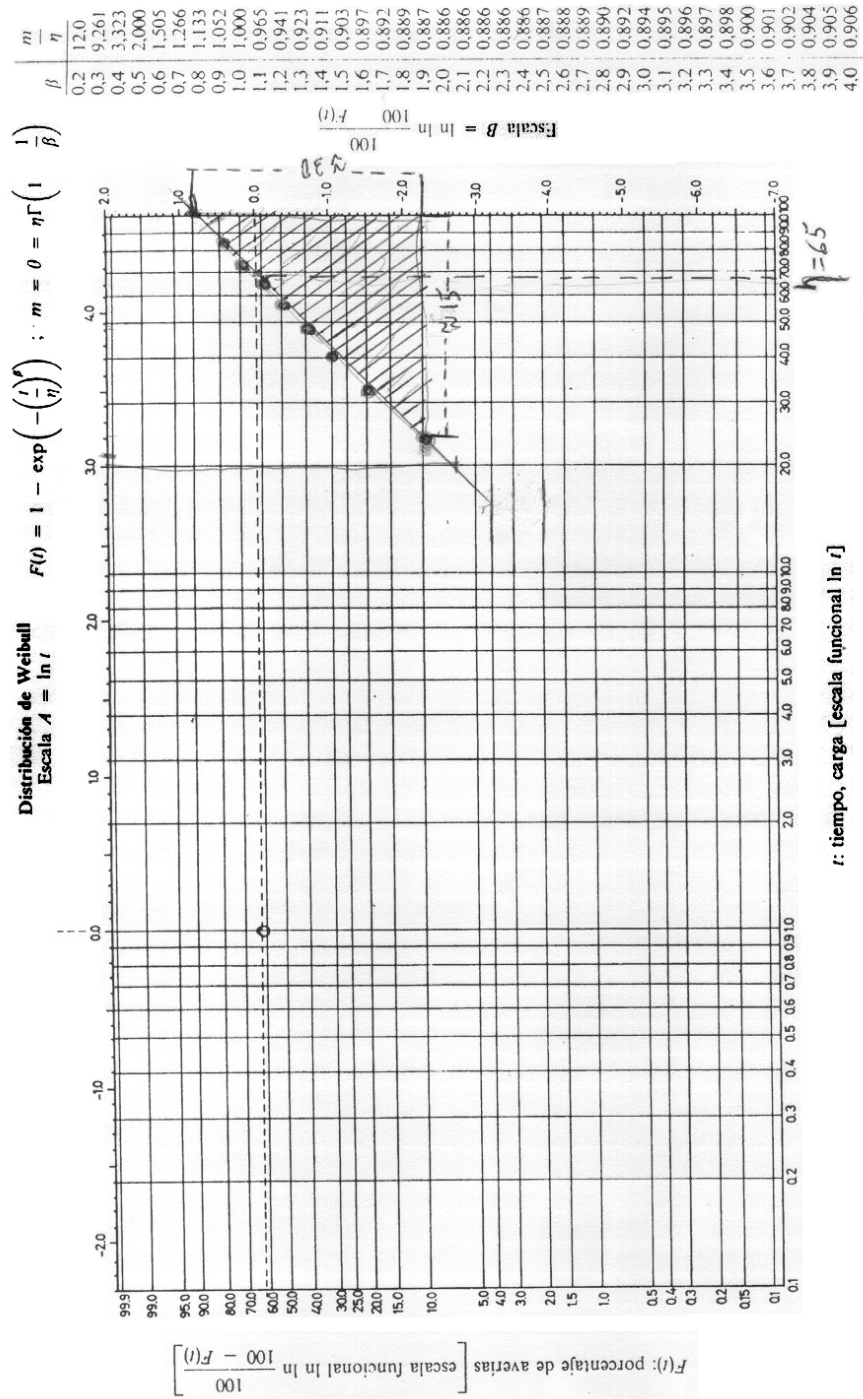


Figura 10.2:



## Tema 11

# Modelos para sistemas. Redundancia

### 11.1 Introducción. Modelo matemático.

Entendemos que un sistema es un conjunto de componentes conectados entre sí. El sistema está diseñado para realizar una o varias tareas. Es razonable suponer que la fiabilidad del sistema depende de la fiabilidad de sus componentes y de la estructura o forma en que han sido conectados entre sí dentro del sistema con el objetivo de que éste realice la función para la que haya sido diseñado.

Los modelos matemáticos para sistemas tienen por objeto el cálculo de la fiabilidad de estos a partir de la de sus componentes.

En los modelos que vamos a tratar admitimos las condiciones siguientes:

- a) El modelo consta de una serie de bloques. Cada bloque representa un componente o combinación de componentes que realiza una función.
- b) Cada bloque tiene solo dos estados mutuamente excluyentes: Funciona o no funciona
- c) La función representada por un bloque es necesaria para el funcionamiento del sistema. No obstante el fallo de un bloque no tiene por qué implicar forzosamente el fallo del sistema, siempre que existan otros bloques que realicen la misma función.
- d) Interpretamos que los fallos son siempre independientes. No se tiene en cuenta la relación de dependencia de unos fallos con otros.
- e) Conocemos la probabilidad de fallo de cada uno de los componentes.

## 11.2 Redundancia

La Redundancia se define como la existencia de más de un medio para realizar una función dada. Por ejemplo, un avión de tres motores (A, B, C) que funciona aunque falle uno de sus motores.

Hay cuatro formas de realizar la función

- Que funcionen los tres motores A, B y C.
- Que funcionen sólo A y B
- Que funcionen sólo A y C.
- Que funcionen sólo B y C.

Entre los tipos de redundancia se considera la *redundancia activa total* que consiste en que todos los elementos en redundancia están en funcionamiento, aunque el sistema funciona con tal de que uno de ellos sobreviva. La *redundancia activa parcial*, consiste en que todos los elementos ( $n$ ) funcionan simultáneamente, pero sólo es necesario que funcionen por lo menos  $k$  de ellos. Se dice que el sistema está en redundancia activa  $k/n$ . Es el caso del avión trimotor anteriormente citado, que sería un sistema en redundancia activa parcial  $2/3$ . Otro tipo de redundancia se da cuando la función es realizada por un solo componente, aunque hay otro en reserva, que entran en funcionamiento sólo si falla el primero. Este tipo de redundancia se llama en *Reserva, Secuencial* o también redundancia en "*Standby*". Puede que haya un tercer componente que sustituya a este segundo, y así sucesivamente.

## 11.3 Sistemas en serie.

Un sistema en serie se caracteriza porque el fallo de cualquiera de sus bloques implica el fallo del sistema, es decir que el sistema solo funciona cuando funcionan todos sus componentes.

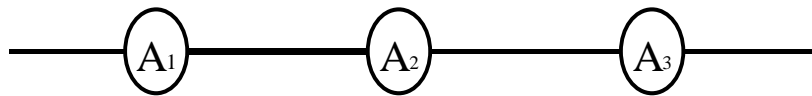


Figura 11.1

Una cadena compuesta por un cierto número de eslabones es un ejemplo de un sistema en serie. Normalmente, basta que uno de los eslabones de una cadena se abra para que la cadena no cumpla su función.



Sea  $A_i$  el suceso que se cumple cuando el componente  $i$  funciona.

La probabilidad de que el sistema en serie funcione es:

$$P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2)P(A_n/A_1 \cap A_2 \cap \dots \cap A_n) \quad (11.1)$$

si los fallos son dependientes o

$$P(A_1).P(A_2).P(A_3)P(A_n) \quad (11.2)$$

si los fallos son independientes.

Si lo expresamos en términos de fiabilidad (fallos independientes) tenemos el siguiente teorema:

**Teorema 5 (Teorema del producto de las fiabilidades)** *La fiabilidad,  $R(t)$ , de un sistema en serie es el producto de las fiabilidades de sus componentes.*

$$R(t) = R_1(t)R_2(t)\dots R_n(t) = \prod_{i=1}^n R_i(t) \quad (11.3)$$

Derivando obtenemos

$$R'(t) = \frac{d}{dt} \left( \prod_{i=1}^n R_i(t) \right) = \sum_{k=1}^n \left( \frac{\prod_{i=1}^n R_i(t)}{R_k(t)} \right) R'_k(t) \quad (11.4)$$

$$f(t) = - \sum_{k=1}^n \left( \frac{\prod_{i=1}^n R_i(t)}{R_k(t)} \right) R'_k(t) \quad (11.5)$$

Esta última es la función de densidad del sistema en serie y rige el tiempo de vida del elemento que primero falle entre los  $n$  considerados. Puede comprenderse fácilmente que esta expresión es también válida como función densidad de la distribución que sigue el menor elemento de una muestra de  $n$  elementos para cualquier variable aleatoria.

Dividiendo la igualdad 11.5 por el producto de las fiabilidades (fórmula 11.3) se obtiene:

$$h(t) = h_1(t) + h_2(t)\dots + h_n(t) = \sum_{i=1}^n h_i(t) \quad (11.6)$$

Así que en un sistema en serie la tasa de fallo  $h(t)$  del sistema es la suma de las tasas de fallo de sus componentes.

Se comprueba fácilmente que un sistema en serie de  $n$  bloques exponenciales es también exponencial y la tasa de fallo es la suma de la tasa de fallo de cada una de sus componentes. Si todos los componentes exponenciales son idénticos, la vida media del sistema se obtiene dividiendo por  $n$  la vida media de uno de sus componentes. Comprobemos estas afirmaciones.

1. Un sistema en serie con bloques exponenciales es exponencial:

$$R(t) = R_1(t)R_2(t)\dots R_n(t) = e^{-\lambda_1 t} e^{-\lambda_2 t} \dots e^{-\lambda_n t} = e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)t}$$

Por lo tanto es una función de fiabilidad exponencial.

2. Su tasa de fallo es la suma de la tasa de fallo de sus componentes:

Este resultado está ya contenido en 11.6. Lo particularizamos para el caso exponencial.

$$h(t) = \frac{f(t)}{R(t)} = \frac{-R'(t)}{R(t)} = \frac{(\lambda_1 + \lambda_2 + \dots + \lambda_n)e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)t}}{e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)t}} = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

3. Si los bloques son idénticos, su vida media se obtiene dividiendo la vida media de cada componente por  $n$  (número de estos que están en serie):

Como el sistema es exponencial, la vida media del sistema es la inversa de su tasa de fallo. Por tanto se obtiene:

$$\mu_{sistema} = \frac{1}{\lambda_1 + \lambda_2 + \dots + \lambda_n}$$

Si todos los componentes son idénticos

$$\mu_{sistema} = \frac{1}{n\lambda} = \frac{1}{n} \frac{1}{\lambda} = \frac{1}{n} \mu = \frac{\mu}{n}$$

#### 11.4 Sistemas en paralelo.

El esquema de un sistema en paralelo podría ser el siguiente:

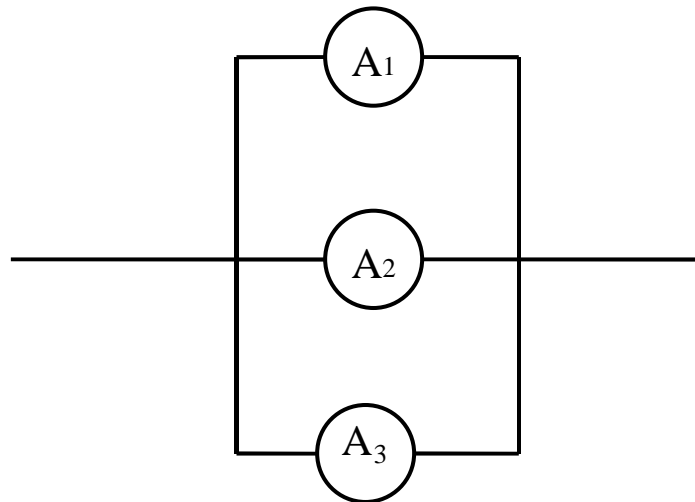


Figura 11.2

Un sistema en paralelo es aquel que falla si y solo si fallan todos sus componentes o bloques. Es un sistema de redundancia activa total del tipo  $1/n$ . El sistema funciona si funciona al menos uno de sus componentes.

Para que un sistema de este tipo no funcione es necesario que fallen todos los componentes. Por lo tanto, la probabilidad de que el sistema falle antes de  $t$  se obtiene según se enuncia en el siguiente teorema:

**Teorema 6 (Teorema del producto de las infiabilidades)** *La infiabilidad,  $F(t)$ , de un sistema en paralelo es el producto de las infiabilidades de sus componentes.*

$$F(t) = F_1(t)F_2(t) \dots F_n(t) = \prod_{i=1}^n F_i(t) \quad (11.7)$$

Es decir,

$$1 - R(t) = (1 - R_1(t))(1 - R_2(t)) \dots (1 - R_n(t)) = \prod_{i=1}^n (1 - R_i(t)) \quad (11.8)$$

Por lo tanto la fiabilidad del sistema es:

$$R_s(t) = 1 - \prod_{i=1}^n (1 - R_i(t)) \quad (11.9)$$

Si los bloques son idénticos

$$R_s(t) = 1 - \prod_{i=1}^n (1 - R_i(t)) = 1 - F^n(t) \quad (11.10)$$

Derivando obtenemos la función densidad de los sistemas en paralelo:

$$f_s(t) = nF^{n-1}(t) f(t) \quad (11.11)$$

que corresponde a la función densidad de la distribución del tiempo hasta el fallo del elemento que más tiempo dure. Esta expresión es también válida como función densidad de la distribución que sigue el mayor elemento de una muestra de  $n$  elementos de cualquier variable aleatoria.

Si los bloques siguen distribuciones exponenciales,

$$R_s(t) = 1 - \prod_{i=1}^n (1 - e^{-\lambda_s t}) \quad (11.12)$$

En este caso el sistema resultante no sigue la ley exponencial

Si todos los bloques son idénticos, la fiabilidad sería:

$$R_s(t) = 1 - (1 - e^{-\lambda t})^n \quad (11.13)$$

y

$$f_s(t) = n \left(1 - e^{-\lambda t}\right)^{n-1} \lambda e^{-\lambda t} \quad (11.14)$$

La vida media puede calcularse integrando la función de fiabilidad:

$$\int_0^{\infty} R_s(t) dt = \int_0^{\infty} \left[1 - \left(1 - e^{-\lambda t}\right)^n\right] dt \quad (11.15)$$

Es mejor hacerlo por recurrencia, hallando la diferencia entre la vida media de los sistemas de  $n$  bloques y de  $n - 1$  bloques

$$\begin{aligned} \mu_n - \mu_{n-1} &= \int_0^{\infty} \left[ \left(1 - (1 - e^{-\lambda t})^n\right) - \left(1 - (1 - e^{-\lambda t})^{n-1}\right) \right] dt = \\ &= \int_0^{\infty} (1 - e^{-\lambda t})^{n-1} e^{-\lambda t} dt = \left| \frac{(1 - e^{-\lambda t})^n}{n\lambda} \right|_0^{\infty} = \frac{1}{n\lambda} \end{aligned}$$

de donde se deduce que la vida media de un sistema en paralelo con  $n$  elementos es:

$$\mu_n = \frac{1}{\lambda} + \frac{1}{2\lambda} + \dots + \frac{1}{n\lambda} \quad (11.16)$$

Por tanto en los sistemas en paralelo la vida media del sistema va aumentando según aumenta el número de bloques del sistema, aunque la importancia de esta mejoría va decelerando con el número de bloques.

$$\mu_n = \frac{1}{\lambda} + \frac{1}{2\lambda} + \dots + \frac{1}{n\lambda} = \mu \left(1 + \frac{1}{2} + \dots + \frac{1}{n}\right) \quad (11.17)$$

## 11.5 Redundancia activa parcial

En los sistemas en paralelo basta que sobreviva un bloque. Si es necesario que sobrevivan  $k$  bloques el sistema se llama de redundancia activa parcial  $k/n$ .

¿Como se halla la fiabilidad en este caso? Podemos usar los esquemas de Venn, o las tablas de verdad, ú otros procedimientos basados en la función de estructura, que no tratamos aquí. Ponemos como ejemplo el caso del avión trimotor comentado antes, que es un sistema en redundancia activa  $2/3$ .

La tabla de verdad correspondiente es la que sigue (1 significa funciona, 0 significa falla)

A	B	C	sistema
1	1	1	1
1	1	0	1
1	0	1	1
1	0	0	0
0	1	1	1
0	1	0	0
0	0	1	0
0	0	0	0

Para calcular la fiabilidad del sistema, hallamos primero la de los casos en que el sistema funciona (que son los señalados con 1 en la columna *sistema*). La fiabilidad de cada uno de estos cuatro casos se calcula con un producto de tres factores. Si aparece un 1 en la columna de una componente, el factor es la fiabilidad de esta componente. Si aparece un 0 el factor es la infiabilidad. La fiabilidad del avión trimotor se obtiene sumando la fiabilidad correspondiente a los cuatro casos en que el sistema funciona.

De este modo obtenemos que la fiabilidad del sistema de motores del avión trimotor es:

$$\begin{aligned}
 &R_A(t)R_B(t)R_C(t) + R_A(t)R_B(t)F_C(t) + R_A(t)F_B(t)R_C(t) + \\
 &F_A(t)R_B(t)R_C(t) = \\
 &R_A(t)R_B(t)R_C(t) + R_A(t)R_B(t)(1 - R_C(t)) + R_A(t)(1 - R_B(t))R_C(t) + \\
 &+ (1 - R_A(t))R_B(t)R_C(t) = \\
 &R_A(t)R_B(t) + R_A(t)R_C(t) + R_B(t)R_C(t) - 2R_A(t)R_B(t)R_C(t) \quad (11.18)
 \end{aligned}$$

Si todos los bloques son iguales puede usarse el siguiente procedimiento ( $p$ = fiabilidad,  $q$ = infiabilidad)

$$(p + q)^n = \sum_{i=1}^n \binom{n}{i} p^i q^{n-i} \quad (11.19)$$

La probabilidad de que funcione el sistema (funcionen al menos  $k$  de los componentes) sería

$$\sum_{i=k}^n \binom{n}{i} p^i q^{n-i} \quad (11.20)$$

Es decir, que la fiabilidad de un sistema  $k/n$  con componentes idénticos es:

$$\sum_{i=k}^n \binom{n}{i} p^i q^{n-i} = \sum_{i=k}^n \binom{n}{i} R(t)^i F(t)^{n-i} \quad (11.21)$$

## 11.6 Combinaciones serie-paralelo

Las combinaciones *serie-paralelo* están formados por  $n$  sistemas en serie, constando cada uno de ellos de  $m_j$  dispositivos en paralelo. La fiabilidad de un sistema como este es:

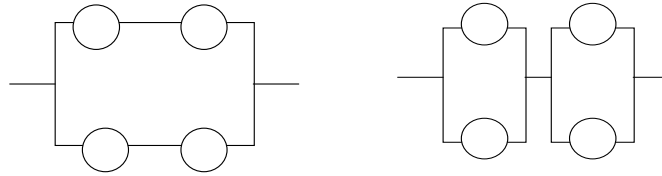
$$R_{sp}(t) = \prod_{i=1}^n R_i(t) = \prod_{i=1}^n \left( 1 - \prod_{j=1}^{m_j} (1 - R_{ij}(t)) \right) \quad (11.22)$$

Las combinaciones *paralelo-serie* constan de  $n$  sistemas en paralelo cada uno de ellos con  $m_j$  dispositivos en serie.

$$R_{ps}(t) = 1 - \prod_{i=1}^n \left( 1 - \prod_{j=1}^{m_j} R_{ij}(t) \right) \quad (11.23)$$

**Ejemplo 60** Consideremos un dispositivo con dos componentes de la misma fiabilidad en serie. Sea  $p$  la probabilidad de que cada uno de los componentes funcione. La fiabilidad del dispositivo es, por tanto,  $p^2$ . Para mejorar la fiabilidad se van a colocar dispositivos redundantes en paralelo. ¿Es mejor duplicar el dispositivo o duplicar cada uno de los componentes?

Los sistemas que se desean comparar están esquematizados en las siguientes figuras



a) Repitiendo el sistema      b) Repitiendo cada elemento

Para responder a la pregunta formulada calculamos cada una de las funciones de fiabilidad y las representamos gráficamente (ver figura)

- $p^2$  (curva 1) es la probabilidad de que sobreviva el sistema simple.
- a)  $1 - (1 - p^2)^2$  (curva 2) es la probabilidad de que funcione el sistema paralelo-serie.
- b)  $\left( 1 - (1 - p)^2 \right)^2$  (curva 3) es la probabilidad de que funcione el sistema serie-paralelo.

**Comprobación gráfica:**

En la gráfica se representa cada una de estas fiabilidades en función de  $p$ . Se observa que el último caso es el de mayor fiabilidad y que en ambos casos se mejora la fiabilidad del sistema primitivo.

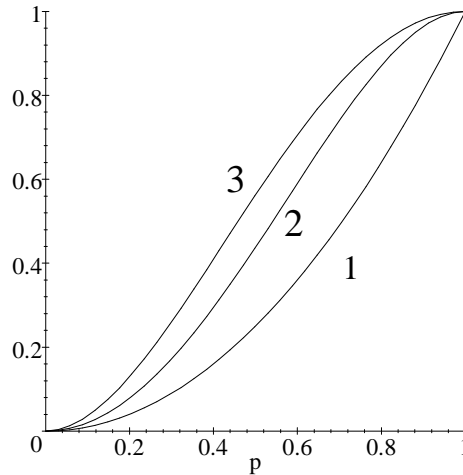


Figura 11.3 Comparación de los valores de la fiabilidad

Por ejemplo, para  $p = 0.9$  resulta

$$p^2 = 0.81, 1 - (1 - p^2)^2 = 0.9639, \left(1 - (1 - p)^2\right)^2 = .9801$$

**Comprobación analítica:**

$$\left(1 - (1 - p)^2\right)^2 - \left(1 - (1 - p^2)^2\right)^2 = 4p^2 - 4p^3 + p^4 - (2p^2 - p^4) = 2p^2 - 4p^3 + 2p^4 = 2p^2(p - 1)^2 \geq 0$$

Así que el primer término  $\left(1 - (1 - p)^2\right)^2$  es mayor que el segundo y por tanto es preferible repetir cada elemento que repetir el dispositivo.

## 11.7 Fiabilidad de sistemas Complejos

Sobrentendemos que un sistema complejo es el que no es exactamente combinación de sistemas serie-paralelo o paralelo-serie. Un ejemplo de este tipo es el de la gráfica de la figura siguiente, que se conoce con el nombre de *sistema puente*. Una forma de calcular su fiabilidad es representar esta fiabilidad como la suma de la fiabilidad de sistemas que sean serie-paralelo o paralelo-serie.

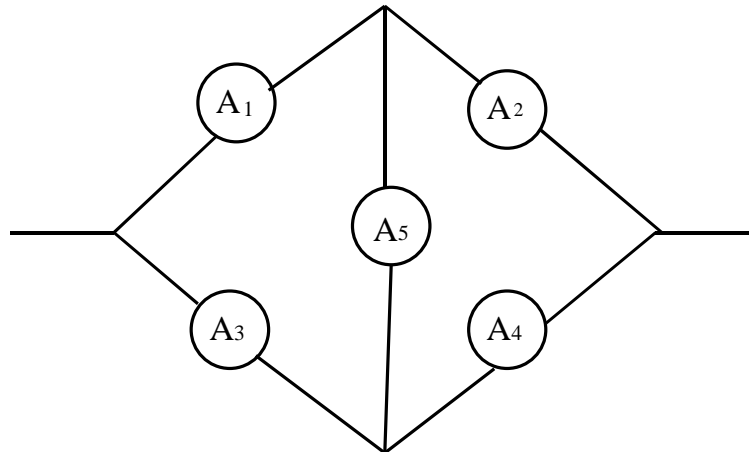


Figura 11.4

Así, para calcular la fiabilidad del sistema puente se consideran dos casos mutuamente excluyentes e incompatibles: Sea  $A_5$  el suceso que se verifica cuando el elemento  $A_5$  funciona, y  $\bar{A}_5$ , el que se verifica si no funciona. La probabilidad de que el sistema funcione puede expresarse, aplicando el teorema de la probabilidad total, como:

$$P(\text{el sistema funcione}) = P(A_5) \times P(\text{el sistema funcione si } A_5) + P(\bar{A}_5) \times P(\text{el sistema funcione si } \bar{A}_5)$$

Si ocurre  $A_5$ , el sistema equivale al siguiente, que es del tipo serie-paralelo:

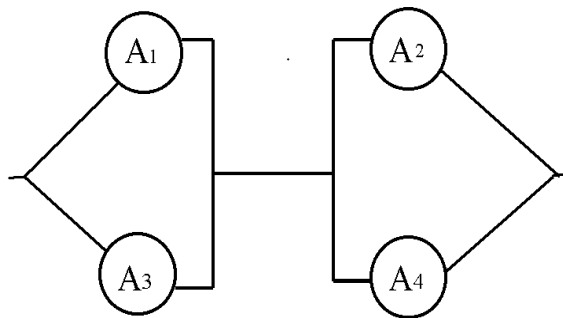


Figura 11.5

Y si ocurre  $\bar{A}_5$ , el sistema equivale al siguiente, que es en sistema paralelo-serie:



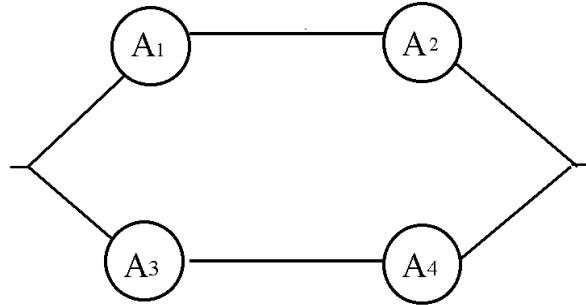


Figura 11.6

Llamando  $R(A_i) = R_i$ , a la fiabilidad del elemento  $i$  y  $F(A_i) = F_i$  a su infiabilidad, la fiabilidad del sistema puente podría expresarse en la forma siguiente:

$$R_5 \times [(1 - F_1 F_3) (1 - F_2 F_4)] + F_5 \times [1 - (1 - R_1 R_2) (1 - R_3 R_4)]$$

### 11.8 Redundancia secuencial

En este caso los bloques no están todos en funcionamiento como en el caso de los sistemas en paralelo. Los bloques no funcionan simultáneamente, sino que hay un dispositivo que detecta el fallo del elemento primario y conecta el secundario. Igualmente puede haber redundancia en este segundo elemento. Además estos sistemas pueden dejar de funcionar por fallo en el mecanismo de conmutación entre el elemento primario y secundario o también por fallo del elemento de reserva.

Calculamos ahora la fiabilidad de un dispositivo en redundancia secuencial:

El sistema funcionará hasta el momento  $t$  si no falla el dispositivo primario(A) hasta ese instante o si, habiendo fallado éste, funciona correctamente el dispositivo conmutador, conecta el dispositivo secundario(B) y este funciona bien hasta  $t$ . En los cálculos que siguen se supone que el dispositivo conmutador no ha fallado antes del fallo del primario y que no puede conmutarse erróneamente sin haber fallado éste.

Bajo estas suposiciones, dividiendo el intervalo  $(0, t)$  en  $n$  intervalos iguales con los puntos  $x_1, x_2, \dots, x_{n-1}$ , siendo  $x_0 = 0, x_n = t$  y  $\Delta x_i = x_{i+1} - x_i$ ,

$$R_s(t) = R_A(t) + \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} [F_A(x_i + \Delta x_i) - F_A(x_i)] p_{dc} R_B(t - x_i) = \quad (11.24)$$

$$\begin{aligned} R_A(t) + \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} \frac{F_A(x_i + \Delta x_i) - F_A(x_i)}{\Delta x_i} p_{dc} R_B(t - x_i) \Delta x_i = \\ = R_A(t) + \int_0^t f_A(x) p_{dc} R_B(t - x) dx \end{aligned}$$

siendo  $f_A(x)$  la función de densidad del tiempo hasta el fallo del dispositivo  $A$  y  $p_{dc}$  la probabilidad de que el dispositivo conmutador funcione satisfactoriamente.

## 11.9 Redundancia secuencial con bloques exponenciales.

Si los dos bloques son exponenciales y el dispositivo conmutador es perfecto obtenemos:

$$R_s(t) = e^{-\lambda_A t} + \int_0^t \lambda_A e^{-\lambda_A x} e^{-\lambda_B(t-x)} dx = \frac{\lambda_A e^{-\lambda_B t} - \lambda_B e^{-\lambda_A t}}{\lambda_A - \lambda_B} \quad (11.25)$$

En el caso de que ambos dispositivos sean idénticos

$$R_s(t) = e^{-\lambda t} + \int_0^t \lambda e^{-\lambda x} e^{-\lambda(t-x)} dx = e^{-\lambda t} + \lambda e^{-\lambda t} \int_0^t dx = e^{-\lambda t} (1 + \lambda t)$$

Integrando la expresión 11.25 se puede comprobar que la vida media de un sistema de dos componentes exponenciales, dispuestos en redundancia secuencial y con dispositivo conmutador perfecto, es la suma de la vida media de cada uno de los dispositivos:

$$\begin{aligned} \int_0^\infty \frac{\lambda_A e^{-\lambda_B t} - \lambda_B e^{-\lambda_A t}}{\lambda_A - \lambda_B} dt &= \frac{\lambda_A}{\lambda_A - \lambda_B} \int_0^\infty e^{-\lambda_B t} dt - \frac{\lambda_B}{\lambda_A - \lambda_B} \int_0^\infty e^{-\lambda_A t} dt = \\ \frac{\lambda_A}{\lambda_A - \lambda_B} \frac{-1}{\lambda_B} e^{-\lambda_B t} \Big|_0^\infty - \frac{\lambda_B}{\lambda_A - \lambda_B} \frac{-1}{\lambda_A} e^{-\lambda_A t} \Big|_0^\infty &= \\ = \frac{\lambda_A}{\lambda_A - \lambda_B} \frac{-1}{\lambda_B} (0 - 1) - \frac{\lambda_B}{\lambda_A - \lambda_B} \frac{-1}{\lambda_A} (0 - 1) &= \\ \frac{\lambda_B + \lambda_A}{\lambda_B \lambda_A} = \frac{1}{\lambda_A} + \frac{1}{\lambda_B} = \mu_A + \mu_B \end{aligned}$$

En este caso (dispositivo conmutador perfecto) no importa en que orden se pongan los dispositivos. Pero si el dispositivo conmutador no es absolutamente fiable es mejor poner primero el más seguro como se muestra a continuación:

$$R_s(t) = e^{-\lambda_A t} + p \int_0^t \lambda_A e^{-\lambda_A x} e^{-\lambda_B(t-x)} dx = e^{-\lambda_A t} + p \lambda_A e^{-\lambda_B t} \int_0^t e^{(\lambda_B - \lambda_A)x} dx =$$

$$e^{-\lambda_A t} + \frac{p \lambda_A e^{-\lambda_B t}}{\lambda_B - \lambda_A} \left| e^{(\lambda_B - \lambda_A)x} \right|_0^t = e^{-\lambda_A t} + \frac{p \lambda_A e^{-\lambda_B t}}{\lambda_B - \lambda_A} \left( e^{(\lambda_B - \lambda_A)t} - 1 \right)$$

Si intercambiamos los dos componentes, y restamos ambas fiabilidades resulta

$$\begin{aligned}
 R(AB) - R(BA) &= \\
 e^{-\lambda_A t} + \frac{p\lambda_A e^{-\lambda_B t}}{\lambda_B - \lambda_A} \left( e^{(\lambda_B - \lambda_A)t} - 1 \right) - \left( e^{-\lambda_B t} + \frac{p\lambda_B e^{-\lambda_A t}}{\lambda_A - \lambda_B} \left( e^{(\lambda_A - \lambda_B)t} - 1 \right) \right) &= \\
 = \frac{e^{-\lambda_A t} \lambda_A - e^{-\lambda_A t} \lambda_B - p\lambda_A e^{-\lambda_A t} + p\lambda_A e^{-\lambda_B t}}{\lambda_A - \lambda_B} - & \\
 - \frac{e^{-\lambda_B t} \lambda_A - e^{-\lambda_B t} \lambda_B + p\lambda_B e^{-\lambda_B t} - p\lambda_B e^{-\lambda_A t}}{\lambda_A - \lambda_B} &= \\
 = -e^{-\lambda_B t} + e^{-\lambda_A t} - p e^{-\lambda_A t} + p e^{-\lambda_B t} = -e^{-\lambda_B t} + e^{-\lambda_A t} + p \left( e^{-\lambda_B t} - e^{-\lambda_A t} \right) &= \\
 = (p - 1) \left( e^{-\lambda_B t} - e^{-\lambda_A t} \right) &
 \end{aligned}$$

Ya que hemos supuesto que el dispositivo conmutador no es absolutamente fiable ( $p < 1$ ) esta diferencia será positiva si  $e^{-\lambda_A t} > e^{-\lambda_B t}$ . Es decir que  $R(AB) \geq R(BA)$  si la fiabilidad del componente instalado primero ( $A$ ) es mayor que la del componente en reserva ( $B$ ).

Cuando se tienen en redundancia secuencial  $n$  bloques exponenciales e idénticos, el dispositivo conmutador sea perfecto y las tasas de fallo de todos ellos sea  $\lambda$ , conviene tener en cuenta lo siguientes resultados:

La **distribución del tiempo,  $t$ , entre dos fallos consecutivos** (desde que falla un elemento hasta que falla el siguiente) sigue una distribución exponencial de parámetro  $\lambda$ .

$$f(t) = \lambda e^{-\lambda t} \quad (11.26)$$

La **distribución del número de fallos ( $r$ )** que ocurren durante el tiempo  $t$  sigue una Poisson de parámetro  $\lambda t$ .

$$P(n^\circ \text{ de fallos} = r) = \frac{(\lambda t)^r}{r!} e^{-\lambda t} \quad (11.27)$$

La **distribución del tiempo ( $t$ ) transcurrido hasta el  $n$ -simo fallo** es una gamma con  $n$  entero que recibe el nombre de distribución de Erlang y que es la distribución de la suma de  $n$  variables aleatorias exponenciales.

$$f(t) = \frac{\lambda}{(n-1)!} (\lambda t)^{n-1} e^{-\lambda t}$$

La **Fiabilidad de un sistema en redundancia secuencial con  $n$  componentes exponenciales idénticos y con conmutadores perfectos** sería la probabilidad de que el tiempo transcurrido hasta el  $n$ -simo fallo sea mayor que  $t$ , que como se ha indicado, seguiría una distribución de Erlang. Esta fiabilidad puede obtenerse realizando (iterativamente por partes) la siguiente integral.

$$R_s(t) = \int_t^{\infty} \frac{\lambda}{(n-1)!} (\lambda t)^{n-1} e^{-\lambda t} dt = e^{-\lambda t} \sum_{i=0}^{n-1} \frac{(\lambda t)^i}{i!} \quad (11.28)$$

La vida media del sistema es, en este último caso caso, la media de la distribución gamma que como ya notamos en el tema anterior es:

$$n\mu = \frac{n}{\lambda} \quad (11.29)$$

Comparando esta expresión con la 11.17 se deduce que es mejor colocar  $n$  elementos en redundancia secuencial que en paralelo si la conmutación (paso de un elemento que ha fallado al siguiente en reserva) fuera perfecta. Pero como esto no suele ser así, ha de estudiarse con detalle cada caso.

## 11.10 EJERCICIOS PROPUESTOS

**Ejercicio 123** La duración (en horas) de unos dispositivos se rige por una distribución cuya función densidad viene dada por:

$$\begin{aligned} f(t) &= \frac{1}{8}t & \text{si } t \in [0, 4] \\ f(t) &= 0 & \text{en el resto} \end{aligned}$$

1. Calcular la función de Infiabilidad, la función de Fiabilidad y la vida media.
2. Calcular la probabilidad de que uno de estos dispositivos dure mas de 3 horas
3. Probabilidad de que un sistema formado por dos de estos dispositivos en paralelo dure más de 3 horas

**Ejercicio 124** Tres componentes con tiempo de fallo exponencial y tasa de fallo 0.03, 0.06 y 0.04 se han dispuesto formando un sistema en serie.

1. Hallar  $R(6)$  para el sistema
2. Calcular la vida media del sistema

3. Calcular la probabilidad de que el sistema permanezca funcionando al menos 4 horas

**Ejercicio 125** La fiabilidad (tiempo hasta el fallo en horas) de un dispositivo viene dada por la función:  $R(t) = e^{-\frac{t}{10}}$

1. Calcular la función de densidad, la tasa de fallos y la vida media de estos dispositivos.
2. Calcular la probabilidad de que uno de estos dispositivos dure más de 9 horas
3. Para aumentar la fiabilidad del sistema se han colocado 4 de estos dispositivos en paralelo. ¿Cuál es la probabilidad de que este sistema dure más de 9 horas?
4. Calcular la vida media de este sistema

**Ejercicio 126** Un sistema con tres componentes independientes trabaja correctamente si al menos uno de ellos funciona. Las tasas de fallo de cada uno de ellos son: 0.01, 0.02, 0.03. Suponiendo que el tiempo de vida de estos componentes sigue una distribución exponencial, calcular:

1. La función de Fiabilidad del sistema
2. La probabilidad de que el sistema funcione al menos 100 horas.
3. La tasa de fallo del sistema

**Ejercicio 127** Se tienen tres componentes A, B, C en serie con fiabilidad 0.5, 0.8, 0.85. Se desea mejorar la fiabilidad del sistema añadiendo redundancia activa componente a componente.

1. Con un solo elemento
2. Con dos elementos
3. Con tres elementos

¿Cuál es la mejor composición del sistema en cada caso?

**Ejercicio 128** El tiempo de vida de unos ciertos dispositivos sigue una distribución Normal de media 10000 horas y desviación típica 1000 horas.

1. Calcular la función de fiabilidad y la probabilidad de que uno de estos dispositivos dure al menos 9000 horas.

2. Si se sabe que uno de estos dispositivos ya ha durado 9000 horas, ¿Cuál es la probabilidad de que dure por lo menos 500 horas más?
3. Formamos un sistema en serie con dos dispositivos usados. El primero ha sido usado 9000 horas y el segundo 11000 horas. ¿Cuál es la probabilidad de que este sistema dure 500 horas?
4. ¿Cuál es la probabilidad de que este sistema dure 500 horas si los dos dispositivos anteriores los colocamos en paralelo?

**Ejercicio 129** *Calcula:*

1. La fiabilidad en un instante de un sistema como el siguiente si cada componente tiene en ese instante una fiabilidad de 0.4:

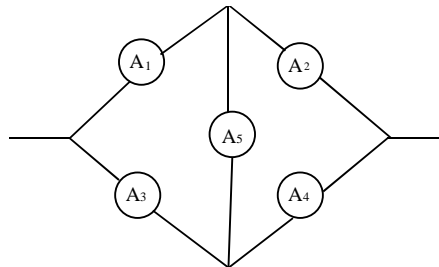


Figura 11.7

2. Fiabilidad de un sistema formado por dos elementos en serie cada uno de ellos como el de la figura
3. Idem si se montan en paralelo

**Ejercicio 130** *Hallar la fiabilidad de un flash con 3 pilas en redundancia secuencial con distribución de fallo exponencial.*

1. Si se supone que las tres pilas son idénticos y la conmutación de un dispositivo a otro es perfecta:
2. Si se supone que las tasas de fallo son diferentes y la conmutación de un dispositivo a otro es perfecta:

**Ejercicio 131** *Un sistema está compuesto por dos componentes en serie con tiempo de vida exponencial con una vida media de 200 horas y 500 horas respectivamente.*

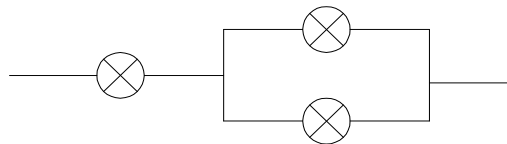
1. Hallar la función de fiabilidad del sistema.
2. Si ponemos ambos componentes en paralelo ¿Cual sería la función de fiabilidad del sistema?
3. Si añadimos en paralelo al sistema del apartado anterior un componente que sigue una distribución uniforme entre 0 y 150 horas ¿Aumenta la fiabilidad del sistema?
4. Si ponemos los tres elementos en paralelo ¿Cuál es la probabilidad de que este sistema de tres elementos en paralelo dure menos de 100 horas?

**Ejercicio 132** El tiempo de duración de ciertos componentes siguen una distribución exponencial con tasa de fallo de 0.005 fallos por hora.

Se pide:

1. Hallar la función de fiabilidad
2. Probabilidad que el componente dure menos de 300 horas.
3. Hallar la probabilidad de que dos de estos componentes no fallen antes de las 300 horas
4. Hallar la función de fiabilidad de tres de estos componentes colocados en paralelo

**Ejercicio 133** Tres componentes idénticos con fiabilidad exponencial y tiempo medio de vida de 2000 horas están conectados formando el sistema de la figura



1. Hallar la función de fiabilidad de cada uno de sus componentes
2. Hallar la probabilidad de que cada componente dure al menos 1000 horas
3. Hallar la función de fiabilidad del sistema
4. ¿Cuál es la probabilidad de que el sistema dure al menos 1000 horas.

**Ejercicio 134** Se supone que el vuelo de un avión es un sistema que consta de tres componentes principales:  $A$  (avión),  $B$  (tripulación) y  $C$  (aeropuerto), además el componente  $B$  puede considerarse como un subsistema en paralelo formado por un capitán ( $B_1$ ) y un suboficial ( $B_2$ ). También el aeropuerto consta de dos pistas ( $C_1, C_2$ ) y el avión debe usar por lo menos una de ellas. Para que el vuelo se realice tienen que estar disponibles los tres componentes principales. La probabilidad de que cada uno de los elementos del sistema realice su función satisfactoriamente es la siguiente:

$$P(A) = 0.9999, P(B_1) = 0.995, P(B_2) = 0.8, P(C_1) = 0.95, P(C_2) = 0.85$$

1. ¿Cuál es la fiabilidad del sistema?
2. ¿Cuál sería si se añadiera una nueva pista de aterrizaje con probabilidad de estar utilizable el 50% de las veces?
3. ¿Y si se suprimiera el suboficial?

**Ejercicio 135** Cuatro unidades idénticas permanecen en un sistema en redundancia activa con fallos independientes. Al menos tres de las unidades deben permanecer activas para que el sistema pueda cumplir su misión.

1. Si las unidades tienen función de fiabilidad exponencial con tasa de fallo 0.02, calcular la función de fiabilidad del sistema y su vida media.
2. ¿Cuál sería la función de fiabilidad si solo se precisará una unidad para el funcionamiento del sistema? ¿Cuál sería en este caso la vida media del sistema?

**Ejercicio 136** Hallar la función de fiabilidad, de distribución, de densidad y la vida media del sistema formado por dos componentes idénticos

1. Si ambos son exponenciales, están colocados en paralelo y cada uno de ellos tiene una vida media de 2 horas.
2. Si están en paralelo y cada uno de ellos se rige por una distribución uniforme en el intervalo entre 0 y 4 horas.

**Ejercicio 137** Un sistema está compuesto por dos componentes en serie con tiempo de vida exponencial con una vida media de 100 horas y 400 horas respectivamente.

1. Hallar la función de fiabilidad del sistema y su vida media
2. Si ponemos ambos componentes en paralelo ¿Cuál sería la función de fiabilidad del sistema?



3. Si añadimos al sistema del apartado anterior un componente en serie que sigue una distribución uniforme cuya vida está entre 0 y 150 horas ¿Aumenta la fiabilidad del sistema?
4. Si ponemos los tres elementos en paralelo ¿Cuál es la probabilidad de que el sistema dure menos de 100 horas?

**Ejercicio 138** Un sistema consta de dos componentes idénticos con función de densidad exponencial y conectados en paralelo. La tasa de fallo en horas de cada componente es  $9 \times 10^{-4}$ .

1. Calcular la función de fiabilidad de cada componente.
2. Calcular la función de fiabilidad de un sistema con dos componentes en paralelo.
3. Calcular la probabilidad de que este sistema de dos componentes en paralelo dure al menos 1200 horas
4. ¿Cuántos componentes como mínimo habría que colocar en paralelo para que la vida media del sistema sea al menos de 2400 horas?

**Ejercicio 139** Tres componentes con tiempo de fallo exponencial y tasa de fallo 0.02, 0.04 y 0.05 (tiempo en horas) se han dispuesto formando un sistema en serie.

1. Cual la tasa de fallo del sistema.
2. Hallar la función de fiabilidad del sistema.
3. Calcular la probabilidad de que el sistema permanezca funcionando al menos 10 horas.

**Ejercicio 140** Cinco unidades idénticas permanecen en un sistema en redundancia activa con fallos independientes. Al menos dos de las unidades deben permanecer activas para que el sistema pueda cumplir su misión.

1. Si las unidades tienen función de fiabilidad exponencial con tasa de fallo 0.02, calcular la función de fiabilidad del sistema.
2. ¿Cual sería la la función de fiabilidad si solo se precisará una unidad para el funcionamiento del sistema? ¿Y su vida media?
3. Si las cinco unidades estuvieran en redundancia secuencial y el dispositivo de conmutación fuese perfecto ¿Cual sería la vida media del sistema?

**Ejercicio 141** *Calcular la función de fiabilidad de dos componentes en redundancia secuencial con dispositivo conmutador perfecto si la duración de cada una de ellas se rige por una Distribución Uniforme en el intervalo  $[0,4]$*

**Ejercicio 142** *La vida media de un componente sigue una distribución exponencial de media 0.2 meses. Cuando este componente falla se reemplaza inmediatamente por otro idéntico, por lo que tiene que haber suficientes elementos de repuesto, ya que el suministrador sólo atiende la demanda una vez al mes. ¿Cuántos elementos hay que tener en stock si no se desea que el riesgo de quedarnos sin repuestos supere el 5%?*

**Ejercicio 143** *Trabajamos con un componente exponencial de vida media 0.2 meses. Cuando se rompe este componente debe sustituirse otro para poder seguir trabajando. Ocurre que la periodicidad del reparto de ese componente es mensual y por tanto solo podemos adquirir repuestos nuevos una vez al mes. Por ese motivo queremos tener en stock al menos el número de un número de elementos suficientes para ir reponiendo de modo que la probabilidad de quedarnos sin repuestos, y por tanto tener que detener la producción en medio del mes sea menor que 0,01. ¿De cuántos de estos componentes debemos disponer al principio de este ciclo mensual?*

**Ejercicio 144** *Cuando hay dos componentes en paralelo, parece razonable suponer que si uno de ellos falla el segundo está sometido a unas condiciones más duras de trabajo, y por tanto tendrá más posibilidades de fallar:*

*Supongamos dos componentes idénticos con función de fiabilidad exponencial y colocados en paralelo. La tasa de fallo (en fallos cada mil horas) de cada elemento funcionando juntos es  $\lambda_1 = 5$  y si sólo funciona uno de ellos es  $\lambda_2 = 7$ .*

1. *Hallar la función de fiabilidad de un sistema de estas características.*
2. *Comparar la fiabilidad en el instante  $t = 100$  horas de este sistema y de otro sistema en paralelo en forma convencional, en que la tasa de fallo del elemento que sobrevive continúe siendo 5.*

**Ejercicio 145** *Se disponen de 7 elementos idénticos dispuestos en redundancia secuencial. La vida media de estos elementos es 1000 horas. Calcular, si no hay ningún problemas de conmutación:*

1. *La Fiabilidad del sistema para 3000 días de funcionamiento y para 5000 días.*
2. *La vida media del sistema de 7 elementos.*

## Tema 12

# Inferencia con pruebas de vida

### 12.1 Pruebas o ensayos de vida

La validez de los pronósticos sobre fiabilidad depende de la buena elección de su función de distribución. A veces la experiencia con dispositivos análogos puede justificar esta elección. Suponemos en este tema que conocemos el tipo de distribución, Exponencial, Weibul, Normal,... que sigue el tiempo de duración hasta el fallo de los dispositivos. Este tipo de estimación se conoce en Estadística con el nombre de *Estimación paramétrica*. En este caso es preciso estimar únicamente los parámetros de la distribución. También existen estudios donde no se tiene en cuenta ningún tipo de distribución previa para la función de fiabilidad. En este caso se habla de *Estimación no paramétrica*, que no trataremos aquí. Para estudiar este tipo de estimación puede consultarse el tema 6 del texto de Nachlas en la bibliografía básica recomendada. También en la página [http://www.uoc.edu/in3/emath/docs/Fiab\\_4.pdf](http://www.uoc.edu/in3/emath/docs/Fiab_4.pdf) se tratan algunos de estos métodos. También pueden consultarse los textos de Lawless, J.F y Nelson, Wayne (Statistical Models And Methods for Lifetime Data y Applied Life Data Analysis, ambos de John Wiley & Sons, Inc., New York, 1982).

Para estimar los parámetros de la distribución, que se supone conocida, necesitamos recoger muestras de la duración real de estos dispositivos en funcionamiento. Estos datos pueden recogerse después de la introducción del producto en el mercado, aunque las conclusiones de estos estudios llegan demasiado tarde, pues ya el proceso de fabricación está en marcha, aunque pueden ser útiles para diseñar futuras modificaciones. Muchos contratos obligan hoy en día a una demostración de la fiabilidad previa a la entrega del producto, por lo que frecuentemente se realizan pruebas de vida en fábrica

antes de la salida del producto al mercado.

Como los tiempos de vida de algunos productos son demasiado largos se suelen usar las llamadas *Pruebas Aceleradas de Vida*, que consisten en realizar las pruebas en condiciones más duras de las que van a ser las de funcionamiento habitual del producto. Así, para unos monitores, los ingenieros pueden ponerlos a prueba en condiciones que quedan fuera de los límites de tolerancia especificados de calor, frío, vibración, impactos y caídas, para garantizar que ofrecen una durabilidad más que suficiente como para mantener un óptimo rendimiento en condiciones reales.

Por lo general estas pruebas de vida no se realizan hasta que dejan de funcionar todos los elementos que intervienen en los ensayos, sino solamente algunos de éstos, por lo que las estimaciones que habitualmente se usan en Estadística no son adaptables y es necesario emplear unos modelos especiales, con censura, que estudiamos en este tema.

### 12.1.1 Tipos de pruebas de vida

Las pruebas de vida que trataremos aquí son realizadas con datos censurados *tipo I*, que quiere decir que se observan los productos durante un tiempo  $T$  prefijado de antemano, o datos censurados *tipo II* que son obtenidos observando los productos hasta que se produce el fallo  $r$ -ésimo, siendo  $r$  un número de fallos determinado previamente.

Trataremos los cuatro tipos de pruebas que se describen a continuación:

#### **Ensayos terminados a $r$ fallos, con reposición.**

En el instante inicial se ponen a funcionar  $n$  unidades del tipo de dispositivo ensayado. Cada vez que una de ellas falla se sustituye por otra nueva. Cuando se produce el  $r$ -ésimo fallo, siendo  $r$  un número prefijado, se termina la prueba. En este tipo de pruebas hay siempre  $n$  elementos funcionando. El único dato que se registra es el tiempo  $t$  transcurrido desde el comienzo hasta el fallo  $r$ -ésimo.

#### **Ensayos terminados a $r$ fallos, sin reposición.**

En este caso sólo se utilizan las  $n$  unidades de partida. Cada unidad que falla es retirada, quedando por tanto una menos en el ensayo. Al fallar la  $r$ -ésima se termina la prueba. En este caso se registran los tiempos  $t_1, t_2, t_3, \dots, t_r$  en que se produce cada fallo.

#### **Ensayos terminados a tiempo $T$ , con reposición**

Se efectúa el ensayo como en el primer caso, pero en vez de terminarse el ensayo cuando se han producido un determinado número de fallos ahora se termina cuando ha transcurrido un cierto tiempo  $T$ . Se registra el número de fallos  $k$  ocurridos en el tiempo  $T$ .

#### **Ensayos terminados a tiempo $T$ , sin reposición**

Este tipo de ensayo se realiza como el segundo caso, pero en vez de terminar el ensayo cuando se han producido un determinado número de fallos ahora se termina cuando ha transcurrido un cierto tiempo  $T$ . Se registra el número de fallos  $k$  ocurridos en el tiempo  $T$  y los tiempos  $t_1, t_2, t_3, \dots, t_k$  en que se produce cada fallo.

## 12.2 Estimadores de máxima verosimilitud.

Para estimar los parámetros de la función de fiabilidad se usan frecuentemente estimadores de máxima verosimilitud. Un estimador de máxima verosimilitud del parámetro de una distribución se obtiene a partir de una muestra y es el valor del parámetro que da a la muestra la máxima posibilidad de ocurrencia.

Si  $f(x, \theta)$  es la función de densidad poblacional y  $\theta$  el parámetro desconocido, se llama *función de verosimilitud* de la muestra  $(x_1, x_2, \dots, x_n)$  a la función de densidad conjunta de los valores muestrales

$$L(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) = \prod_1^n f(x_i; \theta)$$

El estimador de máxima verosimilitud de  $\theta$  es el valor que haga máxima la función de verosimilitud de la muestra, por lo cual debe anular su derivada (o la derivada de su logaritmo) respecto de  $\theta$ :

$$\frac{d \ln L(x_1, x_2, \dots, x_n; \theta)}{d\theta} = 0$$

**Ejemplo 61** Calcular un estimador de máxima verosimilitud para el parámetro  $\mu$  de una distribución normal basado en muestras con 3 elementos

En este caso la función de densidad es:  $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ , así que la función de verosimilitud será:

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta) &= f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) = \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_3-\mu}{\sigma}\right)^2} \end{aligned}$$

El logaritmo neperiano de esta función de verosimilitud es:

$$3 \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2} \left(\frac{x_1-\mu}{\sigma}\right)^2 - \frac{1}{2} \left(\frac{x_2-\mu}{\sigma}\right)^2 - \frac{1}{2} \left(\frac{x_3-\mu}{\sigma}\right)^2$$

La derivada con respecto a  $\mu$  igualada a 0 es:

$$0 - \frac{2}{2} \left(\frac{x_1-\mu}{\sigma}\right) \left(-\frac{1}{\sigma}\right) - \frac{1}{2} \left(\frac{x_2-\mu}{\sigma}\right) \left(-\frac{1}{\sigma}\right) - \frac{2}{2} \left(\frac{x_3-\mu}{\sigma}\right) \left(-\frac{1}{\sigma}\right) = 0$$

Despejando de esta ecuación  $\mu$ , obtenemo su estimación máximo-verosímil:

$$\hat{\mu} = \frac{x_1+x_2+x_3}{3}$$

que, como se ve, es la media de la muestra.

### 12.3 Estimación de la vida media útil

En el periodo de vida media útil la función más frecuentemente usada es la exponencial, que es la que emplearemos en los siguientes apartados.

#### 12.3.1 Ensayos terminados a $r$ fallos con reposición.

En este caso el estimador de máxima verosimilitud para la vida media  $\mu$  es

$$\hat{\mu} = \frac{nt_r}{r} = \frac{T_{ac}}{r} \quad (12.1)$$

donde  $n$  es el tamaño de la muestra,  $r$  el número de fallos total y  $t_r$  el tiempo hasta el  $r$ -ésimo fallo.  $T_{ac} = nt_r$  suele llamarse *tiempo acumulado del ensayo*.

Para comprobar que la expresión 12.1 es el estimador de máxima verosimilitud, podemos usar el hecho de que el tiempo transcurridos entre dos fallos consecutivos,  $t_i - t_{i-1}$  se distribuye exponencialmente con tasa de fallo  $n\lambda$  (la distribución del tiempo hasta que falle el primer elemento entre estos  $n$  elementos, tiene la misma distribución que los sistemas en serie) así que la función de densidad de estos tiempos son  $n\lambda e^{-n\lambda(t_i - t_{i-1})}$  por tanto la función de verosimilitud es, tomando  $t_0 = 0$

$$L(t_1, t_2, \dots, t_r; \theta) = n\lambda e^{-n\lambda(t_1 - 0)} \cdot n\lambda e^{-n\lambda(t_2 - t_1)} \dots n\lambda e^{-n\lambda(t_r - t_{r-1})} = (n\lambda)^r e^{-n\lambda t_r}$$

$$\begin{aligned} \frac{d \ln L(t_1, t_2, \dots, t_r; \theta)}{d\theta} &= \frac{d \ln ((n\lambda)^r e^{-n\lambda t_r})}{d\lambda} = \frac{d (r \ln(n\lambda) - n\lambda t_r)}{d\lambda} \\ &= \frac{d (r \ln n + r \ln \lambda - n\lambda t_r)}{d\lambda} = \frac{r}{\lambda} - nt_r = 0 \end{aligned}$$

Por lo tanto

$$\hat{\lambda} = \frac{r}{nt_r} = \frac{r}{T_{ac}} \quad \text{y} \quad \hat{\mu} = \frac{nt_r}{r} = \frac{T_{ac}}{r} \quad (12.2)$$

Los intervalos de confianza para  $\lambda$  se construyen teniendo en cuenta que  $2\lambda T_{ac}$  es una chi-cuadrado con  $2r$  grados de libertad

$$2\lambda T_{ac} = \frac{2T_{ac}}{\mu} = \chi_{2r}^2$$

Si el nivel de significación es  $\alpha$ , se verifica

$$\chi_{2r, \frac{\alpha}{2}}^2 < 2\lambda T_{ac} < \chi_{2r, 1 - \frac{\alpha}{2}}^2$$

donde los subíndices  $\frac{\alpha}{2}$  e  $1 - \frac{\alpha}{2}$  indican el área a la izquierda de la curva de Chi-cuadrado con  $2r$  grados de libertad.

Los límites bilaterales para  $\lambda$  son, dividiendo por  $2T_{ac}$  :

$$\frac{\chi_{2r, \frac{\alpha}{2}}^2}{2T_{ac}} < \lambda < \frac{\chi_{2r, 1-\frac{\alpha}{2}}^2}{2T_{ac}}$$

y por tanto el intervalo de confianza para  $\mu$  es

$$\left( \frac{2T_{ac}}{\chi_{2r, 1-\frac{\alpha}{2}}^2}, \frac{2T_{ac}}{\chi_{2r, \frac{\alpha}{2}}^2} \right)$$

Frecuentemente se usa el límite unilateral inferior para  $\mu$ : Si consideramos un nivel de significación del  $\alpha\%$ , el intervalo unilateral de confianza para la vida media es

$$\mu > \frac{2T_{ac}}{\chi_{2r, 1-\alpha}^2}$$

Estos intervalos de confianza pueden usarse para realizar contrastes de hipótesis referentes a la vida media. Se rechazará la hipótesis nula sobre el valor de la vida media si el parámetro muestral cae fuera del intervalo de confianza correspondiente.

**Ejemplo 62** *Se desea estimar la vida media de un tipo de condensadores a base de un ensayo con reposición terminado al  $r$ -ésimo fallo. Para ello se toma una muestra de 400 condensadores y se fija  $r = 10$ . El tiempo transcurrido hasta que acontece el décimo fallo resulta ser de 65 días.*

1. *Calcular el estimador de máxima verosimilitud para la vida media*
2. *Calcular un intervalo bilateral de confianza al 90% para este parámetro.*
3. *Estimaciones correspondientes para la tasa de fallo.*

El tiempo acumulado es  $T_{ac} = nt_r = 400 \times 65 = 26000.0$  días.

1. Por lo tanto la estimación de la vida media será  $\hat{\mu} = \frac{nt_r}{r} = \frac{26000}{10} = 2600.0$  días  $\simeq 7.12$  años
2. Para hallar el intervalo de confianza tenemos que calcular  $\chi_{2r, \frac{\alpha}{2}}^2$  y  $\chi_{2r, 1-\frac{\alpha}{2}}^2$ , puesto que en este caso, al contrario de lo que ocurre en la distribución normal, no hay simetría.

$$\begin{aligned} \chi_{2r, 1-\frac{\alpha}{2}}^2 &= \chi_{20, 0.95}^2 = 31.41 \\ \chi_{2r, \frac{\alpha}{2}}^2 &= \chi_{20, 0.05}^2 = 10.851 \end{aligned}$$

Por lo que el intervalo de confianza (en años) pedido para la vida media de estos condensadores resulta:

$$\left( \frac{2T_{ac}}{\chi_{2r, 1-\frac{\alpha}{2}}^2}, \frac{2T_{ac}}{\chi_{2r, \frac{\alpha}{2}}^2} \right) = \left( \frac{2 \times 71.2}{31.41}, \frac{2 \times 71.2}{10.851} \right) \simeq (4.5, 13.1)$$

3. Un estimador de la tasa de fallo (anual) es

$$\hat{\lambda} = \frac{1}{\hat{\mu}} = \frac{1}{7.12} = 0.14$$

La idea intuitiva de esta tasa de fallo es la estimación de la proporción de elementos (en tanto por uno) que fallan por unidad de tiempo, en este caso por año.

El intervalo de confianza para la tasa de fallo es

$$\left( \frac{\chi_{2r, \frac{\alpha}{2}}^2}{2T_{ac}}, \frac{\chi_{2r, 1-\frac{\alpha}{2}}^2}{2T_{ac}} \right) = \left( \frac{10.851}{2 \times 71.2}, \frac{31.41}{2 \times 71.2} \right) = (0.076, 0.22)$$

### 12.3.2 Ensayos terminados a $r$ fallos sin reposición

En este caso, el tiempo acumulado del ensayo es

$$T_{ac} = (n - r) t_r + \sum_1^r t_i$$

y el estimador de máxima verosimilitud para  $\lambda$ , resulta ser

$$\hat{\lambda} = \frac{r}{T_{ac}}$$

por tanto

$$\hat{\mu} = \frac{T_{ac}}{r} = \frac{(n - r) t_r + \sum_1^r t_i}{r}$$

es el estimador de máxima verosimilitud para la vida media.

Se pueden aplicar las mismas fórmulas que en el caso anterior, salvo la interpretación de  $T_{ac}$ , ya que también ahora  $\frac{2T_{ac}}{\mu}$  es una chi-cuadrado con  $2r$  grados de libertad

Naturalmente  $n$  ha de ser mayor o igual que  $r$  ya que no hay reposición.



**Ejemplo 63** Se ensaya una muestra de 200 resistencias eléctricas, sin reponer las que fallan, hasta producirse el séptimo fallo. Los tiempos en horas transcurridos hasta que han fallado cada uno de las 7 resistencias han sido:

916, 11064, 63296, 74628, 94913, 123408, 164115

1. Calcular la estimación de máxima verosimilitud para la vida media de estas resistencias, suponiendo que la distribución es exponencial
2. Calcular un intervalo de confianza al 80% para la vida media.

$$(n - r) t_r = 193 \times 164115$$

$$1. \quad \sum_1^r t_i = 916 + 11064 + 63296 + 74628 + 94913 + 123408 + 164115 = 532\,340$$

$$T_{ac} = (n - r) t_r + \sum_1^r t_i = 31\,674\,195 + 532\,340 = 32\,206\,535$$

$$\hat{\mu} = \frac{T_{ac}}{r} = \frac{32\,206\,535}{7} = 4.600\,9 \times 10^6$$

$$2. \quad \frac{2T_{ac}}{\mu} = \chi_{2r}^2 \quad \text{El intervalo de confianza es } \chi_{2r,0.10}^2 < \frac{2T_{ac}}{\mu} < \chi_{2r,0.90}^2$$

$$\chi_{2r,0.10}^2 = \chi_{14,0.10}^2 = 7.7895$$

$$\chi_{2r,0.90}^2 = \chi_{14,0.90}^2 = 21.064$$

$$2T_{ac} = 2 \times 32\,206\,535 = 64\,413\,070$$

$$7.7895 < \frac{2T_{ac}}{\mu} < 21.064, \quad 7.7895 < \frac{64\,413\,070}{\mu} < 21.064$$

Por lo tanto un intervalo de confianza para la vida media es

$$\frac{64\,413\,070}{21.064} < \mu < \frac{64\,413\,070}{7.7895}; \quad 3.058 \times 10^6 < \mu < 8.2692 \times 10^6$$

### 12.3.3 Ensayos terminados a tiempo T con reposición

En este caso  $T_{ac} = nT$ , que no es una variable sino una constante. El estimador de máxima verosimilitud para la media es

$$\hat{\mu} = \frac{T_{ac}}{k}$$

siendo  $k$  el número de fallos registrados en ese tiempo.

Los límites superior e inferior de confianza para la media con una confianza de  $100 \times (1 - \alpha)\%$  son

$$\left( \frac{2T_{ac}}{\chi_{2(k+1),1-\frac{\alpha}{2}}^2}, \frac{2T_{ac}}{\chi_{2k,\frac{\alpha}{2}}^2} \right)$$

y el límite unilateral :

$$\mu > \frac{2T_{ac}}{\chi_{2(k+1),1-\alpha}^2}$$

### 12.3.4 Ensayos terminados a tiempo T sin reposición

El tiempo acumulado del ensayo es

$$T_{ac} = (n - k)T + \sum_1^k t_i$$

y la estimación de máxima verosimilitud para la vida media:

$$\hat{\mu} = \frac{T_{ac}}{k}$$

No se disponen de métodos exactos para calcular los intervalos de confianza. Obsérvese que en este caso tanto el numerador como el denominador son variables aleatorias. Suele usarse la siguiente aproximación: Se calculan los intervalos de confianza con las expresiones usadas en el caso anterior (ensayos terminados a tiempo T con reposición), pero sustituyendo el valor de  $T_{ac}$  por  $(n - k)T + \sum_1^k t_i$  que es el valor del tiempo acumulado en el presente caso.

### 12.3.5 Contraste con hipótesis alternativa. Planes de muestreo. Curva característica.

Diseñar un plan de muestreo consiste en especificar una forma de actuación y un criterio de decisión que permita decidir cuando se acepta y cuando se va a rechazar un lote, en función del resultado obtenido en una muestra de elementos extraídas de dicho lote.

Para limitar los riesgos, tanto del productor como del comprador, se parte de 4 parámetros específicos, y cuyos valores son acordados entre ambos.

**Parámetros impuestos por el productor:**  $\mu_0$  y  $\alpha$ .  $\mu_0$  es el valor para la vida media ofrecido por éste y  $\alpha$ , llamado *riesgo del productor*, se define como la probabilidad de que, una vez realizado el plan de muestreo diseñado, resulte rechazado un lote cuya vida media real sea  $\mu_0$ , que es la vida media ofrecida por el productor.

**Parámetros impuestos por el comprador:**  $\mu_1$  y  $\beta$ . El comprador no desea aceptar un lote cuya vida media sea menor o igual que  $\mu_1$ .  $\beta$ , llamado *riesgo del comprador* se define como la probabilidad de aceptar un lote con media verdadera  $\mu_1$ . En función de estos componentes se diseña el plan de muestreo.

A modo de ejemplo, detallamos como se determina un plan de muestreo para el caso del *ensayo terminado a r fallos con reposición*. Es necesario concretar valores de  $n$  (número de elementos que entraran en la muestra del ensayo),  $r$  (número de fallos que se van a observar) y determinar el tiempo  $C$  (mínimo valor aceptado de la vida media muestral) que será el criterio de

decisión. Se aceptará el lote si la estimación de la vida media obtenida de la muestra es mayor o igual que  $C$ , y se rechaza en caso contrario.

**Formalización de las condiciones del productor:**

Riesgo del productor =  $\alpha$  = (Probabilidad de rechazar una partida con media verdadera  $\mu_0$ ) sería:

$$P(\hat{\mu} < C / \mu = \mu_0) = \alpha$$

Teniendo en cuenta la expresión del estimador de la vida media en este tipo de pruebas dado en la expresión 12.1 de la página 286 se tiene:

$$\begin{aligned} \alpha &= P\left(\frac{T_{\alpha c}}{r} < C / \mu = \mu_0\right) = P\left(\frac{2T_{\alpha c}}{2r} < C / \mu = \mu_0\right) = P\left(\frac{\chi_{2r}^2}{\mu} < C / \mu = \mu_0\right) = \\ &= P\left(\frac{\chi_{2r}^2}{\frac{2r}{\mu_0}} < C\right) = P\left(\chi_{2r}^2 < \frac{2rC}{\mu_0}\right) = \alpha \implies \boxed{\frac{2rC}{\mu_0} = \chi_{2r, \alpha}^2} \end{aligned} \quad (12.3)$$

(Tomamos el valor de chi cuadrado que deja a la izquierda el área  $\alpha$ )

**Formalización de las condiciones del comprador:**

Riesgo del comprador =  $\beta$  = (Probabilidad de aceptar una partida con media verdadera  $\mu_1$ ). Por lo tanto la probabilidad de aceptar será

$$P(\hat{\mu} \geq C / \mu = \mu_1) = \beta \quad (12.4)$$

De esta condición se deduce que:

$$P\left(\chi_{2r}^2 \geq \frac{2rC}{\mu_1}\right) = \beta \implies \boxed{\frac{2rC}{\mu_1} = \chi_{2r, (1-\beta)}^2} \quad (12.5)$$

Usando las relaciones obtenidas para el vendedor y el comprador se deduce:

$$\frac{\mu_0 \chi_{2r, \alpha}^2}{2r} = C$$

y

$$\frac{\mu_1 \chi_{2r, 1-\beta}^2}{2r} = C$$

Igualando ambas expresiones de  $C$ , concluimos que

$$\frac{\mu_1}{\mu_0} = \frac{\chi_{2r, \alpha}^2}{\chi_{2r, 1-\beta}^2} \quad (12.6)$$

El valor de  $r$  se puede obtener por tanteo en una tabla de la chi-cuadrado. Como la igualdad 12.6 es raro que se cumpla exactamente, se suele elegir el valor de  $r$  más próximo que cumpla:

$$\frac{\mu_1}{\mu_0} > \frac{\chi_{2r, \alpha}^2}{\chi_{2r, 1-\beta}^2}$$

Como el muestreo es con reposición, el número  $n$  no está sujeto a ninguna restricción, ya que todos los elementos que se rompan van a reponerse. En la práctica, la distribución exponencial se usa en el periodo de vida útil, por lo tanto se debe tomar  $n$  lo suficientemente grande para que los componentes no puedan llegar al periodo de desgaste. Como

$$\hat{\mu} = \frac{nt_r}{r}$$

será

$$n = \frac{\hat{\mu}r}{t_r}$$

Normalmente se toma para  $\hat{\mu}$  el valor de una estimación muestral de la vida media obtenida en ensayos anteriores y para  $t_r$  el tiempo que consideremos adecuado esperar. Hay que tener en cuenta que si se quiere acabar pronto ( $t_r$  pequeño) hay que emplear muchas unidades en el ensayo.

El valor de  $C$  puede determinarse por cualquiera de las dos expresiones anteriores. Por ejemplo

$$C = \frac{\mu_0 \chi_{2r, \alpha}^2}{2r}$$

La **curva característica (OC) de un plan de muestreo** (en el que se han fijado los valores de  $n$ ,  $r$ , y  $C$ ) es una representación de la probabilidad de aceptar un lote (en ordenadas) en función del valor real de  $\mu$  (en abscisas).

$$OC(\mu) = \text{Probabilidad de aceptar} = P\left(\chi_{2r}^2 > \frac{2rC}{\mu}\right)$$

De las expresiones 12.3 y 12.5 se deduce que la curva característica debe pasar por los puntos  $(\mu_0, 1 - \alpha)$  y  $(\mu_1, \beta)$ .

### Ejemplo 64

1. Diseñar un plan de muestreo para un ensayo terminado a  $r$  fallos con reposición, que se ajuste a los siguientes datos

Datos del vendedor:  $\mu_0 = 3 \times 10^6$  horas  $\alpha = 0.10$

Datos del comprador:  $\mu_1 = 10^6$  horas  $\beta = 0.10$

Se desea que el tiempo del ensayo sea 4000 horas aproximadamente. De una prueba anterior se sabe que la vida media es aproximadamente  $2 \times 10^6$  horas

2. Dar la expresión de la curva OC de este plan de muestreo

3. Calcular las ordenadas de la curva característica si la vida media fuese realmente  $2 \times 10^6$  horas

1. Hallamos en primer lugar el valor de  $r$  que cumpla más aproximadamente la expresión.

$$\frac{\mu_1}{\mu_0} = 0.33 > \frac{\chi_{2r,0.10}^2}{\chi_{2r,0.90}^2} \quad (12.7)$$

$$\begin{aligned} \frac{\text{ChiSquareInv}(0.10;2)}{\text{ChiSquareInv}(0.90;2)} &= \frac{0.21072}{4.6052} = 4.5757 \times 10^{-2} \\ \frac{\text{ChiSquareInv}(0.10;4)}{\text{ChiSquareInv}(0.90;4)} &= \frac{1.0636}{7.7794} = 0.13672 \\ \frac{\text{ChiSquareInv}(0.10;6)}{\text{ChiSquareInv}(0.90;6)} &= \frac{2.2041}{10.645} = 0.20706 \\ \frac{\text{ChiSquareInv}(0.10;8)}{\text{ChiSquareInv}(0.90;8)} &= \frac{3.4895}{13.362} = 0.26116 \\ \frac{\text{ChiSquareInv}(0.10;10)}{\text{ChiSquareInv}(0.90;10)} &= \frac{4.8652}{15.987} = 0.30432 \\ \frac{\text{ChiSquareInv}(0.10;12)}{\text{ChiSquareInv}(0.90;12)} &= \frac{6.3038}{18.549} = 0.33984 \end{aligned}$$

Aunque el valor que más próximo a  $\frac{\mu_1}{\mu_0} = \frac{1}{3}$  se da para  $r = 6$ , el valor  $r = 5$  es el más cercano a la igualdad que cumple la relación 12.7. Tomando este valor para  $r$ , obtenemos:

$$C = \frac{\mu_0 \chi_{2r,\alpha}^2}{2r} = \frac{3 \times 10^6 \times \text{ChiSquareInv}(0.10; 10)}{10} = 1.4596 \times 10^6$$

Si usamos el valor de  $C$  de la otra expresión obtenemos:

$$C = \frac{\mu_1 \chi_{2r,1-\beta}^2}{2r} = \frac{10^6 \times \text{ChiSquareInv}(0.90; 10)}{10} = 1.5987 \times 10^6$$

que no es exactamente igual que el anterior, ya que la igualdad de la ecuación 12.7 no se cumple exactamente. Se puede acordar, por ejemplo, dar para  $C$  el valor de la media aritmética de ambos valores o quedarnos con el menor de ambos que da una ligera ventaja al productor. Es lo que vamos a decidir en esta ocasión. Suponiendo que queremos limitar la duración del ensayo a 4000 horas, debemos de tomar

$$n = \frac{\hat{\mu}r}{t_r} = \frac{2 \times 10^6 \times 5}{4000} = 2500.0$$

La prueba consistiría en lo siguiente: Tomar 2500 unidades del producto, y ponerlas en funcionamiento. Cuando falle alguna de estas se cambiará por otra nueva, y así hasta que fallen 5 de ellas. Se estimará la vida media de estas unidades como  $\hat{\mu} = \frac{nt_r}{r} = \frac{2500t_5}{5}$ , siendo  $t_5$  el tiempo total del ensayo. Si este valor estimado para la vida media es mayor o igual que  $C = 1.4596 \times 10^6$  el comprador deberá aceptar el lote. En caso contrario lo rechazará.

2. La curva característica se calcula del siguiente modo

$$\begin{aligned} \text{probabilidad de aceptación} &= P\left(\chi_{2r}^2 > \frac{2rC}{\mu}\right) = 1 - P\left(\chi_{2r}^2 \leq \frac{2rC}{\mu}\right) = \\ &= 1 - P\left(\chi_{10}^2 \leq \frac{2 \times 5 \times 1.4596 \times 10^6}{\mu}\right) = 1 - \text{ChiSquareDist}\left(\frac{2 \times 5 \times 1.4596 \times 10^6}{\mu}; 10\right). \end{aligned}$$

Por tanto la curva característica toma la siguiente forma

$$OC(\mu) = 1 - P\left(\chi_{10}^2 \leq \frac{2 \times 5 \times 1.4596 \times 10^6}{\mu}\right) = 1 - P\left(\chi_{10}^2 \leq \frac{1.4596 \times 10^7}{\mu}\right)$$

3. La ordenada correspondiente al valor  $\mu = 2000000$  es

$$\begin{aligned} f(2000000) &= 1 - P\left(\chi_{10}^2 \leq \frac{1.4596 \times 10^7}{2000000}\right) = 1 - P(\chi_{10}^2 \leq 7.298) \\ &= 1 - \text{ChiSquareDist}(7.298; 10) = 1 - 0.30296 = 0.69704. \end{aligned}$$

## 12.4 Nota sobre planes de muestreo tabulados para pruebas de vida.

Ya que no es sencillo diseñar planes de muestreo para las distintas situaciones que pueden presentarse, existen algunas tablas con planes de muestreo para fiabilidad, como por ejemplo el *DOD Handbook H 108* para todos los tipos de ensayos en que sea aplicable la exponencial, los planes *MIL-STD-690* que están ligados a la clasificación de la calidad de componentes electrónicos de alta fiabilidad y los Planes *AGREE 2*, *AGREE 3* y *MIL-STD-781* que son planes secuenciales truncados.

## 12.5 Pruebas de vida para el periodo de desgaste

### 12.5.1 Ensayo hasta el fallo de todas las unidades. Estimación de la vida media en el periodo de desgaste. Estimación de la varianza.

Para estimar los parámetros de un dispositivo cuya vida en el periodo de desgaste puede considerarse normalmente distribuida conviene realizar el ensayo hasta el fallo de todas las unidades y procurar no tener en cuenta aquellas unidades cuyo fallo no puede atribuirse a desgaste, sino que puede ser debido al azar o a mortalidad infantil.

### 12.5.2 Intervalos de confianza y contraste de hipótesis para la vida media y la varianza

Si suponemos que la distribución es normal y medimos las  $n$  vidas de desgaste, llamando a los tiempos registrados hasta el fallo de todas estas unidades,

$t_1, t_2, \dots, t_n$ , el estimador de máxima verosimilitud para la vida media es la media muestral

$$\hat{\mu} = \frac{\sum_{i=1}^n t_i}{n}$$

El estimador de máxima verosimilitud de la varianza de la población es la varianza de la muestra, pero se suele usar como estimador la cuasivarianza muestral porque es un estimador sin sesgo

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (t_i - \hat{\mu})^2}{n - 1}$$

Para calcular los intervalos de confianza para la vida media recurrimos a las distribuciones de los estadísticos muestrales usuales de la distribución Normal:

$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$  se distribuye como una  $N(0, 1)$

$\frac{\hat{\mu} - \mu}{s/\sqrt{n}}$  se distribuye como una  $t$  de Student con  $n - 1$  grados de libertad

$\frac{(n-1)s^2}{\sigma^2}$  se distribuye como una chi-cuadrado con  $n - 1$  grados de libertad

Así que los intervalos de confianza para la media son:

a) Si se conoce  $\sigma$

$$\left( \hat{\mu} - z_{1-\frac{\alpha}{2}} \sigma / \sqrt{n} \leq \mu \leq \hat{\mu} + z_{1-\frac{\alpha}{2}} \sigma / \sqrt{n} \right)$$

b) Si, como ocurre más frecuentemente, no se conoce  $\sigma$

$$\left( \hat{\mu} - t_{n-1, 1-\frac{\alpha}{2}} s / \sqrt{n} \leq \mu \leq \hat{\mu} + t_{n-1, 1-\frac{\alpha}{2}} s / \sqrt{n} \right)$$

Los intervalos de confianza para la varianza son:

$$\frac{(n-1) s^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1) s^2}{\chi_{n-1, \frac{\alpha}{2}}^2}$$

y para la desviación típica

$$s \sqrt{\frac{n-1}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}} \leq \sigma \leq s \sqrt{\frac{n-1}{\chi_{n-1, \frac{\alpha}{2}}^2}}$$

Si interrumpimos el ensayo en un tiempo  $T$  o decidimos observar los datos a partir de un cierto instante, es decir que no observamos el periodo de vida completo es más adecuado usar la llamada distribución Normal truncada, ya que no tendríamos completa la muestra. Estudios de este tipo pueden verse en *Statistical Theory*, Hald, 1952.

## 12.6 Parámetros de la distribución de Weibull.

La distribución de Weibul

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad (12.8)$$

puede aplicarse a cualquier periodo de la curva de bañera, lo que unido a su facilidad de manejo hace que su uso sea frecuente en fiabilidad. Ya hemos tratado previamente el caso en que el parametro de origen  $\gamma$  es cero o conocido. En este caso la estimación de los parametros  $\eta$  y  $\beta$  se podía realizar por medio de una regresión lineal. Además se podía hacer una estimación gráfica por medio del papel de Weibul.

En el caso de que el parametro de origen sea desconocido y no sea nulo, la representación gráfica de los puntos en el papel de Weibul no sería una recta, sino una curva. Se puede obtener también un valor aproximado para este parámetro ( $\gamma$ ) en el papel de Weibul como se ha detallado en un tema anterior. No obstante tambien podemos estimar los parámetros en este caso general usando, por ejemplo, el paquete estadístico STATGRAPHICS. En la resolución del siguiente problema detallamos el procedimiento a seguir.

**Ejemplo 65** *Los siguientes valores son los tiempo en que fallaron unos dispositivos en el laboratorio: 52, 62, 70, 78, 86, 94, 104, 115, 130 (estos datos coinciden con los del ejemplo 58 de la página 252). Ajustad, con ayuda de Statgraphics, una distribución de Weibul a los datos.*

La siguiente tabla corresponde a la función de distribución empírica:

tiempo (variable independiente)	52	62	70	78	86	94	104	115	130
fdt (variable dependiente)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Tenemos unos valores aproximados para los parámetros que obtuvimos con el papel de Weibul:

$$\gamma = 30, \eta = 65, \beta = 2$$

Se pueden mejorar estos valores con Statgraphics PLUS 5.0:

Se introducen los datos de la tabla en un fichero del programa y se usa el procedimiento

*SPECIAL → ADVANCED REGRESSION →  
NONLINEAR REGRESION*



Usamos los valores  $\gamma = 30$ ,  $\eta = 65$ ,  $\beta = 2$  como parámetros iniciales. Si no tuviéramos estos valores podemos usar otros, pero es preferible que sean valores cercanos a los verdaderos. La variable dependiente es la función de distribución empírica o experimental variable (fdt) y la variable independiente es el tiempo hasta el fallo. La función de ajuste es, naturalmente, la función de distribución de Weibul

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta}$$

Es más conveniente introducirla en la forma

$$F(t) = 1 - \frac{1}{e^{\left(\frac{t-\gamma}{\eta}\right)^\beta}}$$

obteniéndose el siguiente los siguientes valores mejorados para los parámetros

$$\gamma = 29.3520694, \eta = 67.6825549, \beta = 2.0455598$$

A continuación presentamos parte de la salida de Statgraphics plus 5.0 para este problema:

---



---

Nonlinear Regression

Dependent variable: fdt

Independent variables: tiempo

Function to be estimated:  $1 - 1/\exp(((\text{tiempo}-a)/b)^c)$

Initial parameter estimates:

a = 30.0

b = 65.0

c = 2.0

Estimation method: Marquardt

Estimation stopped due to convergence of parameter estimates.

Number of iterations: 4

Number of function calls: 17

Estimation Results

---

Asymptotic 95.0%

Asymptotic Confidence Interval

Parameter	Estimate	Standard Error	Lower	Upper
a	29.3521	1.59428	25.451	33.2531
b	67.6826	1.66753	63.6023	71.7629
c	2.04556	0.0631164	1.89112	2.2

R-Squared = 99.9868 percent

R-Squared (adjusted for d.f.) = 99.9825 percent.

## 12.7 EJERCICIOS PROPUESTOS

**Ejercicio 146** *Calcular un estimador de máxima verosimilitud para el parámetro  $\sigma$  de una distribución normal, suponiendo que se conozca el parámetro  $\mu$ . Emplear muestras con 3 elementos.*

**Ejercicio 147** *Supongamos que se observan muestras de 50 elementos hasta que se obtenga el octavo fallo.*

*Los tiempos de fallo han sido: 91, 145, 221, 285, 315, 328, 411, 496.*

*Estima el valor de la vida media de estos elementos bajo la hipótesis de distribución exponencial*

1. *Si la prueba de vida es con reposición*
2. *Si la prueba de vida es sin reposición*
3. *Hallar en cada caso un intervalo de confianza bilateral y unilateral al nivel 80%*

**Ejercicio 148** *Un fabricante nos informa que sus productos duran por término medio 10000 horas. Hemos instalado 50 unidades de este producto en nuestra empresa y al cabo de 990 horas hemos apreciado que habíamos tenido que reponer siete de ellas. ¿Podemos aceptar la información del fabricante al 95% de confianza?*

**Ejercicio 149** *Se someten 50 unidades a un ensayo censurado por número de fallos sin reposición. El ensayo se terminó al producirse el decimo fallo. Los tiempos hasta el fallo de los 10 elementos observados fueron:*

*65, 110, 380, 420, 505, 580, 650, 840, 910, 950.*

*Hallar un estimador para la vida media y un intervalo bilateral de confianza al 95% para este parámetro.*

**Ejercicio 150** *Se someten 20 unidades a una prueba de vida hasta 10 fallos con reemplazamiento. El decimo fallo se ha producido a las 80 horas. Estimar la vida media, dando un intervalo de confianza bilateral al 95%*

**Ejercicio 151** *Se someten 20 unidades a una prueba de vida sin reemplazamiento durante un tiempo de 600 horas. En este intervalo de tiempo han fallado 18 de ellas. La duración de las unidades falladas (en horas) son:*

*0.69, 0.94, 1.12, 6.79, 9.28, 9.31, 9.95, 12.9, 12.93, 21.33, 64.56, 69.66, 108.38, 124.88, 157.02, 190.19, 250.55, 552.87*

*Estimar la vida media, dando un intervalo de confianza bilateral al 95%.*

**Ejercicio 152** *Se someten 2000 unidades a una prueba de vida sin reemplazamiento durante un tiempo de 600 horas. En este intervalo de tiempo han fallado 18 de ellas. La duración de las unidades falladas (en horas) son:*

*0.69, 0.94, 1.12, 6.79, 9.28, 9.31, 9.95, 12.9, 12.93, 21.33, 64.56, 69.66, 108.38, 124.88, 157.02, 190.19, 250.55, 552.87*

*Estimar la vida media, dando un intervalo de confianza bilateral al 95%.*

**Ejercicio 153** *Se someten 20 unidades a una prueba de vida con reemplazamiento durante un tiempo de 600 horas. En este intervalo de tiempo han fallado 18 de ellas. La duración de las unidades falladas (en horas) son:*

*0.69, 0.94, 1.12, 6.79, 9.28, 9.31, 9.95, 12.9, 12.93, 21.33, 64.56, 69.66, 108.38, 124.88, 157.02, 190.19, 250.55, 552.87*

*Estimar la vida media, dando un intervalo de confianza bilateral al 95%.*

**Ejercicio 154** *Se ha realizado una prueba de vida con reemplazamiento de acuerdo con un plan de muestreo consistente en poner en funcionamiento 2500 elementos hasta que se produzca el 5º fallo. Se ha acordado que si la estimación de la vida media obtenida con esta prueba fuese mayor que 1500, se aceptaría el lote. Si el tiempo medio de vida ofrecido por el productor es de 3000 horas y el valor mínimo exigible por el comprador es de 1000 horas, calcular:*

1. *El riesgo del comprador y del productor.*
2. *El valor de la ordenada de la curva característica que corresponde al valor de 2500 horas para la vida media del producto.*

**Ejercicio 155** *Se desea estimar la duración de un cierto tipo de lámparas. Para ello se ha observado la duración de 15 de ellas, hasta que han fallado todas. Se supone que los tiempos de vida se ajustan bien a una distribución normal. Los tiempos de vida de estas lámparas, en horas, resultaron ser*

*848, 932, 938, 959, 961, 993, 1120, 1126, 1012, 1013, 1035, 1066, 1085, 1123, 1166.*

*Calcular:*

1. *Estimación de la vida media e intervalos unilateral y bilateral de confianza al 95%.*

2. *Estimación de la desviación típica y un intervalo de confianza para ésta al 90%.*

**Ejercicio 156** *En nuestra empresa empleamos unos dispositivos electrónicos cuya función de densidad de probabilidad de su duración sin fallos (en horas) es  $f(t) = 7 \times 10^{-4}e^{-7 \times 10^{-4}t}$ ,  $t > 0$ .*

*Por otra parte hemos hecho una test de vida con otros dispositivos del mismo tipo que nos ofrece un nuevo proveedor. Dicho test ha consistido en poner en funcionamiento 20 de estas unidades hasta que fallarán diez de ellos, en una prueba sin reposición. El registro del tiempo de fallos hasta que ha ocurrido el décimo fallo ha sido: 940, 950, 951, 970, 982, 1007, 1021, 1050, 1079, 1154. ¿Pueden considerarse estos nuevos dispositivos más fiables al 95% de confianza?*

**Ejercicio 157** *En nuestra empresa empleamos unos dispositivos electrónicos cuya función de densidad de probabilidad de su duración sin fallos (en horas) es  $f(t) = 7.5 \times 10^{-4}e^{-7.5 \times 10^{-4}t}$ ,  $t > 0$ .*

*Por otra parte hemos hecho una test de vida con otros dispositivos del mismo tipo que nos ofrece un nuevo proveedor. Dicho test ha consistido en poner en funcionamiento durante 1200 horas 20 de estas unidades en una prueba sin reposición. El registro del tiempo de fallos hasta que ha transcurrido las 1200 horas fué: 940, 950, 951, 970, 982, 1007, 1021, 1050, 1079, 1154. ¿Pueden considerarse estos nuevos dispositivos más fiables al 95% de confianza?*

Unidad Temática IV

**ANÁLISIS DE LA  
VARIANZA**



## Tema 13

# Análisis de la varianza con un factor

### 13.1 Generalidades sobre el diseño de experimentos

En el mundo de la Ingeniería es frecuente tener que elegir entre diferentes opciones: tomar decisiones sobre la implantación de determinadas técnicas de fabricación, seleccionar trabajadores para diferentes empleos, adoptar los métodos más convenientes para insertar publicidad en los medios de comunicación que faciliten la venta de la producción, contactar con los mejores suministradores de la materia prima para la fabricación de nuestros productos... En todos estos casos necesitamos establecer comparaciones entre las características de las diferentes opciones a nuestro alcance. Una técnica estadística adecuada para este propósito es el Análisis de Varianza. A veces disponemos de datos previos para realizar este análisis y por tanto no hay forma de influir en la forma en que estos han sido obtenidos. No obstante en muchos otros casos podemos decidir sobre la forma en que van a recogerse estos datos y podemos diseñar un experimento que sea lo más apropiado posible para conseguir nuestro objetivo, la elección entre las diferentes opciones que se nos presentan, de la manera más eficiente posible.

Un *experimento* es una prueba o conjuntos de pruebas en los que se realizan cambios deliberados en ciertos factores de forma que sea fácil detectar los cambios producidos en la variable de interés o variable de respuesta en función del valor de dichos factores.

**Ejemplo 66** *Supongamos que un ingeniero desea comparar la velocidad de producción cuatro máquinas (A, B, C y D). Para ello selecciona, por ejemplo, 12 operarios que trabajarán durante una hora con esas máquinas. Cada una*

*de ellas será empleada por tres obreros que se seleccionarán al azar entre los 12 empleados seleccionados. La decisión se va a basar en el registro del número de unidades producidas por las máquinas.*

Dispondremos de 12 datos (3 valores por máquina). En este experimento, como en muchos otros, el resultado va a depender de la forma en que se realice el experimento. Normalmente en el número de unidades producidas pueden intervenir diferentes factores como, por ejemplo, la habilidad de los operarios, la temperatura ambiente, la hora del día, la calidad del material con que se realicen las pruebas, etc. Será por tanto difícil decidir si la mayor o menor producción registrada por cada una de las máquinas dependen de las características de la propia máquina o de los otros factores influyentes. Por este motivo hay que planear el experimento usando principios básicos estadísticos. *El diseño estadístico de experimentos* es el proceso de planear un experimento para obtener datos adecuados para ser analizados mediante métodos estadísticos.

## **13.2 Análisis de varianza con un factor.**

### **13.2.1 Introducción**

El Análisis de Varianza es una técnica estadística para determinar si varias muestras aleatorias proceden o no de la misma población. Se suele denominar con el nombre de ANOVA (acrónimo de ANalysis Of VAriance). En el ejemplo 66 hay que considerar cuatro muestras, cada una de ellas con tres elementos, las unidades fabricadas por los tres obreros que trabajan con cada máquina. Se consideran en este ejemplo dos variables. Una de ellas es la variable máquina, que es una variable cualitativa y puede tomar los valores A, B, C y D. La otra variable es el número de unidades producidas por las máquinas que, en el ejemplo, toma 12 valores. La primera variable, la máquina empleada, es denominada, en el contexto del Análisis de Varianza *factor* en estudio, que normalmente es controlada por el experimentador. La otra variable, las unidades producidas, es el resultado del experimento, y como tal no puede ser controlada por el experimentador. Se conoce con el nombre de variable *respuesta*.

Esta técnica fué creada por Ronald A. Fisher (1925) durante los años que estuvo al cargo de la estación agrícola experimental de Rothamsted (Inglaterra). Se pretendía saber si la cantidad empleada de fertilizante tenía alguna influencia en la cantidad de cosecha obtenida. Para hacer este estudio se parceló el terreno, aplicando distinta cantidad de fertilizante a cada parcela. La cantidad de fertilizante que se aplicaba a cada una de las parcelas se decidió al azar.



Este estudio considera dos variables: la cantidad de fertilizante aplicada a cada parcela y la cantidad de cosecha obtenida en ella. La variable independiente, que puede ser controlada por el investigador, es la cantidad de fertilizante, por tanto el factor es la cantidad de fertilizante. La cantidad de cosecha recogida en las distintas parcelas es la variable respuesta. Se desea saber si la variable respuesta es o no independiente del factor en estudio, es decir, si la cantidad de cosecha obtenida depende o no de la cantidad de fertilizante aplicada al terreno.

Normalmente los posibles valores de este factor se llaman *tratamientos*, *niveles* o *grupos*. Esta última denominación es frecuente cuando el factor toma niveles cualitativos. En el análisis de la varianza, el factor cuya influencia sobre la variable respuesta se desea estudiar, se introduce en forma discreta, aunque sea una variable aleatoria continua. Por ejemplo, la cantidad de fertilizante es, potencialmente, una variable aleatoria continua pero, para realizar el experimento, se seleccionan ciertas cantidades de fertilizante, unas pocas, para aplicar a las parcelas y se establecerá la comparación entre estas pocas cantidades seleccionadas.

### 13.2.2 Diseño completamente aleatorizado

Este tipo de diseño de experimentos se puede realizar de la siguiente forma: Se selecciona una muestra aleatoria e independiente de la población con un número de elementos prefijado, por ejemplo 12 elementos. Una forma de realizar este diseño consiste en numerar los elementos de la muestra. Si tenemos cuatro niveles del factor, A, B, C, D, elegimos al azar 3 de los 12 elementos de la muestra y se le aplica el tratamiento A, otros 3 elegidos también al azar le aplicamos el B, y así sucesivamente hasta el último de los tratamientos. Es decir que se elige al azar el elemento al que se le va a aplicar un cierto nivel del factor o tratamiento. Un enfoque similar es aplicar los tratamientos aleatoriamente a las distintas muestras.

A continuación mostramos un modelo para el diseño de un experimento completamente aleatorizado, preparado con el programa Statgraphics (seleccionando en el menú principal *Special* y después *Experimental Design*). Esta aplicado al caso del ingeniero que desea decidir entre varias varias maquinas, la de mayor velocidad de producción (ejemplo 66). Las maquinas se han asignados al azar a cada uno de los operarios.

Operario	Máquina	Producción
1	D	54
2	C	
3	B	
4	D	
5	A	
6	B	
7	A	
8	A	
9	C	
10	D	
11	B	
12	C	

En la última columna se recogerá la variable respuesta. A modo de ejemplo hemos rellenado la primera fila: El primer obrero trabajando con la máquina D ha producido 54 unidades. Así se procederá con los siguientes resultados hasta rellenar la columna de la variable respuesta.

**Ejemplo 67** Una empresa desea enseñar a los operarios una cierta técnica y dispone de tres métodos de enseñanza (A, B, C). Para determinar si alguno de estos métodos es mejor, se seleccionan al azar 14 personas entre sus empleados. Con el método A enseña a 4, con el B a 5 y con el C a los 5 restantes. Posteriormente somete a estos empleados a una prueba para ver si han aprendido la técnica. Las calificaciones de los empleados según el método de enseñanza están dadas en la siguiente tabla:

Método A	6	7	8	7	
Método B	9	8	9	8	9
Método C	9	8	7	8	9

En este caso el factor es el método de enseñanza que tiene tres grupos, niveles o tratamientos, A, B y C. La variable respuesta es la calificación. Deseamos saber si el método (factor) influye en la calificación (respuesta).

El modelo matemático adecuado para el análisis estadístico de este tipo de experimentos es el análisis de la varianza con un solo factor, que se detalla a continuación.

### 13.3 El modelo del análisis de la varianza

Suele adaptarse a este tipo de problemas el siguiente modelo matemático:

Se parte de  $g$  muestras independientes procedentes de distribuciones normales de la misma varianza (la hipótesis de igual varianza entre las distribuciones se llama *Hipótesis de Homocedasticidad*). En el ejemplo 67 hay 3 muestras (las calificaciones de los obreros para cada método de enseñanza).

Se realiza el contraste de hipótesis siguiente:

**Hipótesis nula:** La media de las  $g$  distribuciones es la misma.

**Hipótesis alternativa:** Al menos una de las medias es distinta de las restantes.

En la tabla siguiente especificamos la notación que se usará en lo sucesivo

Factor	Elementos de las muestras	Media muestral	Suma de cuadrados de los errores dentro de cada grupo
Nivel 1	$y_{11}, y_{12} \dots y_{1n_1}$	$\bar{y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j}$	$\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2$
Nivel 2	$y_{21}, y_{22} \dots y_{2n_2}$	$\bar{y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j}$	$\sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2$
...			
Nivel $g$	$y_{g1}, y_{g2} \dots y_{gn_g}$	$\bar{y}_g = \frac{1}{n_g} \sum_{j=1}^{n_g} y_{gj}$	$\sum_{j=1}^{n_g} (y_{gj} - \bar{y}_g)^2$

Considerando todas las muestras como una sola, el número de elementos de la muestra es

$$n = n_1 + n_2 + \dots + n_g$$

y la media total es:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^g \sum_{j=1}^{n_i} y_{ij} \quad (13.1)$$

### 13.3.1 Suma de cuadrados

La desviación de una observación respecto de la media total puede expresarse:

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i) \quad (13.2)$$

El primer paréntesis es la diferencia entre la media del grupo  $i$  (al que pertenece la observación  $y_{ij}$ ) y la media total. Esta desviación se conoce como *desviación explicada por el factor*. El segundo es la diferencia entre el

valor de la observación y la media de la muestra a la que pertenece. Se llama *desviación dentro del grupo o desviación residual*.

Elevando al cuadrado ambos miembros de la expresión 13.2 se puede demostrar que la suma de las desviaciones cuadráticas respecto de la media puede descomponerse de la siguiente forma:

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (13.3)$$

Cada uno de estos términos se conoce con el nombre siguiente

$STC = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 =$  *Suma total de cuadrados* (refleja la variabilidad total)

$SCF = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2 =$  *Suma de cuadrados debida al factor* (refleja la variabilidad explicada por el factor o derivada de las diferencias entre los niveles de éste)

$SCR = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 =$  *Suma de los cuadrados de los residuales* (refleja la variabilidad no explicada por el factor, residual o de dentro de los grupos)

La expresión 13.3 suele abreviarse como

$$STC = SCF + SCR \quad (13.4)$$

### 13.3.2 Grados de libertad de las sumas de cuadrados y medias cuadráticas

Establecemos ahora el número de grados de libertad asociado a cada suma de cuadrados. Como regla general puede establecerse que el número de grados de libertad es el número de datos menos el número de restricciones o relaciones independientes entre estos datos.

$SCT$  emplea  $n$  datos. Pero la media total impone una restricción a los datos. El número de grados de libertad es  $n - 1$ .

$SCF$  tiene  $g$  datos (las medias de cada muestra) y 1 restricción (la media total), es decir  $g - 1$  grados de libertad.

$SCR$  tiene  $n$  datos y  $g$  restricciones (las medias parciales), así que sus grados de libertad son  $n - g$ .

Se debe verificar que la suma de grados de libertad ( $gl$ ) de ambos miembros de la igualdad 13.3 coincida

$$gl(SCT) = gl(SCF) + gl(SCR) = (g - 1) + (n - g) = n - 1$$

Una *media cuadrática* se define como la suma de cuadrados dividido por sus grados de libertad:

$$MTC = \frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{n-1}$$

$$MCF = \frac{1}{g-1} \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2, \quad MCR = \frac{1}{n-g} \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

### 13.3.3 Construcción del test de hipótesis

Se puede demostrar que, si se cumple la hipótesis nula, el estadístico

$$F_{\text{exp}} = \frac{MCF}{MCR} = \frac{\frac{1}{g-1} \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-g} \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

sigue una distribución  $F$  de Fisher -Snedecor con  $g - 1$ , grados de libertad en el numerador y  $n - g$  grados de libertad en el denominador. En el caso de que se cumpla la hipótesis nula el numerador y el denominador son estimadores centrados de la varianza poblacional. En cambio si la hipótesis nula es falsa se espera que el numerador sea mayor que el denominador. Por eso imponemos la condición de que el valor experimental de  $F$  no sea demasiado grande, para que apoye la veracidad de la hipótesis nula. En concreto imponemos la condición de que la  $F_{\text{exp}}$  no exceda de la  $F$  teórica, cuya función de distribución toma el valor  $1 - \alpha$ , siendo  $\alpha$  es el nivel de significación del test.

**Ejemplo 68** Como parte de la investigación del derrumbe del techo de un edificio, un laboratorio prueba todos los pernos disponibles que conectaban la estructura de acero en tres distintas posiciones del techo. Las fuerzas requeridas para cortar cada uno son las siguientes

	Nombre de la variable	Valores muestrales
Posición 1	$X$	90, 82, 79, 98, 83, 91
Posición 2	$Y$	105, 89, 93, 104, 89, 95, 86
Posición 3	$Z$	83, 89, 80, 94

Efectuar el análisis de la varianza para comprobar si las diferencias de fuerza requeridas para las tres distintas posiciones del techo son significativas

Calculamos en primer lugar las medias de las fuerzas correspondientes a cada posición y también la media total

$$\text{Media de } X = 87.167 = \bar{y}_1, \quad \text{Media de } Y = 94.429 = \bar{y}_2, \quad \text{Media de } Z = 86.5 = \bar{y}_3,$$

$$\text{Media Total} = 90.00 = \bar{y}$$

Calculamos a continuación la suma total de cuadrados

$$\begin{aligned} STC &= \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \\ &= \sum_{j=1}^6 (y_{1j} - 90)^2 + \sum_{j=1}^7 (y_{2j} - 90)^2 + \sum_{j=1}^4 (y_{3j} - 90)^2 = \\ &(90 - 90)^2 + \dots + (91 - 90)^2 + (105 - 90)^2 + \dots + (86 - 90)^2 + (83 - 90)^2 + \\ &(94 - 90)^2 = 938 \end{aligned}$$

La suma de cuadrados correspondientes al factor Posición sería

$$\begin{aligned} SCF &= \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2 = \\ &= 6(87.167 - 90)^2 + 7(94.429 - 90)^2 + 4(86.5 - 90)^2 = 234.468 \end{aligned}$$

La suma de cuadrados de los residuos se puede deducir de la relación 13.4:

$$SCR = STC - SCF = 938 - 234.468 = 703.532$$

También se puede realizar directamente el cálculo:

$$\begin{aligned} SCR &= \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \\ &\sum_{j=1}^6 (y_{1j} - 87.167)^2 + \sum_{j=1}^7 (y_{2j} - 94.429)^2 + \sum_{j=1}^4 (y_{3j} - 86.5)^2 = \\ &(90 - 87.167)^2 + \dots + (91 - 87.167)^2 + (105 - 94.429)^2 + \dots + (86 - 94.429)^2 \\ &+ (83 - 86.5)^2 + (94 - 86.5)^2 = 703.532 \end{aligned}$$

Los resultados parciales obtenidos suelen resumirse en una tabla como la siguiente, que se conoce con el nombre de *Tabla de Análisis de la Varianza* o *Tabla ANOVA*.

Fuente de variación	g. libertad	Suma de Cuadrados
Posiciones	$g - 1 = 2$	$SCF = 234.47$
Residuos	$n - g = 14$	$SCR = 703.53$
Total	$n - 1 = 16$	$STC = 938$
Medias Cuadráticas		$F_{\text{exp}}$
$MCF = \frac{234.47}{2} = 117.235$		$\frac{MCF}{MCR} = 2.33$
$MCR = \frac{703.53}{14} = 50.2521$		

$F_{2,14}(2.3329) = 0.866473 < 0.95$ , por lo que no se puede rechazar la hipótesis de igualdad de medias al 95% de confianza.

Otra forma de realizar el contraste, más apropiada si se emplean tablas, es calcular valor de  $F$  que corresponde al nivel de significación 0.05:

$$F_{2,14}^{-1}(0.95) = 3.73889.$$

Como  $F_{\text{exp}} = 2.33$  es menor que este valor se concluye que no se puede rechazar la hipótesis nula.

Es decir nos inclinamos a aceptar que no existe diferencia entre las tres posiciones.

### 13.4 Comparación de dos muestras

Si solo se comparan dos muestras, la igualdad entre las medias poblacionales ( $\mu_1 - \mu_2 = 0$ ) puede estudiarse analizando el estadístico

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{MCR \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

que se distribuye como una  $t$  de Student con  $n - 2$  grados de libertad (ver párrafo 5.15.2 de la página 188).

### 13.5 Validación del modelo

Este modelo sólo es aplicable si se cumplen las hipótesis de partida del test, que son: Normalidad de los datos, igualdad entre las varianzas de las poblaciones de procedencia de las muestras e independencia entre las observaciones de estas muestras. Comentamos ahora las formas de comprobación de estas hipótesis. El incumplimiento de algunas de estas condiciones no tiene por que invalidar todas las conclusiones. En lo que sigue comentaremos también las condiciones en que el análisis es aún válido, aunque haya alguna perturbación en las hipótesis del test.

**Normalidad de los datos:** Pueden realizarse con el test Chi-cuadrado o con el de Kolmogorov-Smirnov. También son útiles los gráficos que algunos autores llaman Q-Q que representan la función de distribución empírica de los residuales del modelo frente a los valores que se esperarían para esta función de distribución si los residuos se comportaran como una variable normal. Si las muestras son grandes, la falta de Normalidad de los datos no afecta de forma importante a la validez del test, ya que las medias que se comparan siguen asintóticamente una distribución Normal. Por lo general la técnica es bastante *Robusta* ante las perturbaciones en la normalidad de los datos.

Algunas veces los datos no son normales, pero sí sus logaritmos. Podremos usar entonces esta técnica realizando previamente una transformación logarítmica de los datos.

**Homocedasticidad:** La igualdad entre las varianzas es también una de las hipótesis del modelo. El contraste no se ve muy afectado por la falta de homocedasticidad siempre que las muestras sean de igual o similar tamaño. Si no fuera así para comprobar la homocedasticidad de los datos pueden usarse los test de Cochran, Barlett y Box.

Algunos autores afirman que puede usarse el contraste de Análisis de la Varianza, aunque no pueda suponerse la igualdad de las varianzas, si la

muestra con más elementos no llega a tener el doble de elementos que la muestra más pequeña.

**Independencia de las muestras:** Se supone que se han elegido muestras aleatoria e independientes. Puede realizarse el estudio de la independencia usando la representación gráfica de los residuos dentro de cada grupo. Si siguen alguna tendencia, en lugar de estar distribuidos de forma aleatoria, se rechaza la hipótesis de independencia de las muestras, aunque también se usan test de hipótesis específicos para este propósito: test de las rachas, análisis de las autocorrelaciones muestrales, test de Durbin-Watson, etc.

**Efecto del incumplimiento de las hipótesis de partida:**

La falta de normalidad afecta poco a los contrastes de igualdad de medias, pero sí afecta a la estimación de la varianza.

La heterocedasticidad influye poco en los contrastes resultantes si no hay una gran diferencia entre los tamaños muestrales.

La falta de independencia en las observaciones es la perturbación más grave, pues puede tener mucha influencia en los resultados.

### 13.5.1 Coeficiente de Determinación

Es la relación entre la suma de cuadrados debida al factor y la suma de total de cuadrados.

$$R^2 = \frac{SCF}{STC}$$

siendo  $0 \leq R^2 \leq 1$ .

En efecto:

$$1 = \frac{STC}{STC} = \frac{SCF+SCR}{STC} = \frac{SCF}{STC} + \frac{SCR}{STC} = R^2 + \frac{SCR}{STC}$$

y como ambos sumandos son positivos ambos estarán comprendidos entre 0 y 1.

Si se cumple la hipótesis de igualdad entre las medias el valor de  $R^2$  ha de ser próximo a 0. En cambio un valor próximo a 1 indica que la mayor parte de la variabilidad es achacable al factor. Éste representa la mayor parte de la variabilidad total. En el ejemplo anterior

$$R^2 = \frac{SCF}{STC} = \frac{234}{938} = 0.249467.$$

Por lo tanto las distintas posiciones explican muy poca proporción de la variabilidad total.



## 13.6 Comparaciones parciales entre las medias

En los casos en que se rechace la hipótesis de igualdad entre las medias, podemos investigar cuales son los grupos responsables de este rechazo. Con este análisis vamos a tratar de determinar en que medida difieren las medias de las distintas poblaciones. Esto puede realizarse por dos procedimientos:

- a) Hallando los intervalos de confianza individuales para la media de cada muestra.
- b) Realizando comparaciones multiples, adoptando una visión de conjunto.

### 13.6.1 Intervalos de confianza individuales para la media de cada grupo

En las hipótesis del modelo se incluye la igualdad de las varianzas. Por este motivo se pueden usar todas las muestras para estimar la varianza de la muestra total y de cada una de las muestras. Emplearemos la media de los cuadrados de los residuos ( $MCR$ ) para estimar la varianza. Usamos el estadístico de contraste:

$$t_{n-g} = \frac{\bar{y}_i - \mu_i}{\sqrt{\frac{MCR}{n_i}}}$$

que se distribuye como una  $t$  de Student con  $n - g$  grados de libertad. Con este estadístico se pueden obtener intervalos de confianza individuales, con un nivel de significación  $\alpha$  para la media de cada muestra

$$\left( \bar{y}_i - t_{n-g, 1-\frac{\alpha}{2}} \sqrt{\frac{MCR}{n_i}} < \mu_i < \bar{y}_i + t_{n-g, 1-\frac{\alpha}{2}} \sqrt{\frac{MCR}{n_i}} \right)$$

Se concluye que dos muestras proceden de poblaciones iguales si sus respectivos intervalos de confianza tienen parte común.

### 13.6.2 Comparaciones multiples

En este caso los distintos intervalos de confianza se establecen simultáneamente, es decir se dan intervalos de confianzas, con nivel  $\alpha$  de significación, para comparar todas las medias simultáneamente. Hay distintos métodos para establecer estas comparaciones multiples: Tukey, Scheffé, Bomferroni. Especificamos este último, que se basa en una idea muy sencilla: la llamada *Desigualdad de Bomferroni*:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) \leq P(A_1) + P(A_2) + \dots + P(A_n)$$

Supongamos que hay  $g$  muestras y queremos realizar comparaciones entre sus medias de dos en dos. El número total de comparaciones que hay que realizar es  $\binom{g}{2} = \frac{g(g-1)}{2} = n_c$ . Sea  $A$  el suceso que corresponde a rechazar la igualdad de todas las medias siendo estas iguales. Esta hipótesis resulta rechazable si al menos resultará rechazable la igualdad entre dos de ellas. Si denotamos por  $A_i$  al suceso que consiste en rechazar la igualdad de dos de estas muestras siendo iguales se tiene que:

$$A = A_1 \cup A_2 \cup \dots \cup A_{n_c}$$

Por tanto el nivel de significación de este test conjunto sería

$$\alpha = P(A) = P(A_1 \cup A_2 \cup \dots \cup A_{n_c}) \leq P(A_1) + P(A_2) + \dots + P(A_{n_c})$$

Si todas las comparaciones individuales toman el mismo nivel de significación  $\alpha^*$  entonces

$$\alpha \leq n_c \alpha^* \implies \alpha^* \geq \frac{\alpha}{n_c}$$

Tomando la cota inferior mínima para  $\alpha^*$  ( $\alpha = n_c \alpha^*$ ), cada una de las comparaciones individuales habrá que realizarla al nivel de significación

$$\alpha^* = \frac{\alpha}{n_c} = \frac{\alpha}{\frac{g(g-1)}{2}}$$

Los contrastes individuales, sobre la diferencia entre dos de las medias, se pueden realizar con el estadístico

$$t_{n-g} = \frac{|\bar{y}_p - \bar{y}_q|}{\sqrt{MCR \left( \frac{1}{n_p} + \frac{1}{n_q} \right)}}$$

al nivel de significación  $\alpha^*$

**Ejemplo 69** *Supongamos que se ensayan tres formulas para un pegamento y que los tiempos de secado observados para cada uno de ellos se recogen en la siguiente tabla.*

Fórmula A	13, 10, 8, 11, 8
Fórmula B	13, 11, 14, 14
Fórmula C	4, 1, 3, 4, 2, 4

*Se desea saber si estos datos indican o no una diferencia significativa en los tiempos de secado de los pegamentos.*

En este caso las medias de cada grupo son:

Media de secado para la fórmula A =  $\bar{y}_1 = 10$

Media de secado para la fórmula B =  $\bar{y}_2 = 13$

Media de secado para la fórmula C =  $\bar{y}_3 = 3$

Media total de secado:  $\bar{y} = \frac{5\bar{y}_1 + 4\bar{y}_2 + 6\bar{y}_3}{5 + 4 + 6} = 8$

La suma de los cuadrados resulta ahora:

$$STC = 302, \quad SCF = 270, \quad SCR = 32$$

y la tabla ANOVA:

Fuente de Variación	g. de libertad	Suma de Cuadrados
Factores	$g - 1 = 2$	$SCF = 270$
Residuos	$n - g = 12$	$SCR = 32$
Total	$n - 1 = 14$	$STC = 302$

Medias cuadráticas	$F_{\text{exp}}$
$MCF = \frac{270}{2} = 135$	$F = \frac{135}{2.667} = 50.62$
$MCR = \frac{32}{12} = 2.67$	

$F_{2,12}(50.62) \approx 1 > 0.95$ , por lo que se rechaza la hipótesis de igualdad entre las medias al 95% de confianza. Otra forma es hallar la F que corresponde al 95% de confianza (nivel de significación 0.05):

$$F_{2,12}^{-1}(0.95) = 3.8853 < 50.62.$$

Concluimos que los tiempos de secado de los distintos pegamentos no son todos iguales.

Los intervalos de confianza individuales para las medias de secados son:

$$\begin{aligned} \text{Fórmula A} & \left( 10 - t_{12,1-0.025} \sqrt{\frac{2.667}{5}}, 10 + t_{12,1-0.025} \sqrt{\frac{2.667}{5}} \right) = \\ & = \left( 10 - 2.178 \sqrt{\frac{2.667}{5}}, 10 + 2.178 \sqrt{\frac{2.667}{5}} \right) = (8.4093, 11.591) \end{aligned}$$

$$\begin{aligned} \text{Fórmula B} & \left( 13 - t_{12,1-0.025} \sqrt{\frac{2.667}{4}}, 13 + t_{12,1-0.025} \sqrt{\frac{2.667}{4}} \right) = \\ & = \left( 13 - 2.178 \sqrt{\frac{2.667}{4}}, 13 + 2.178 \sqrt{\frac{2.667}{4}} \right) = (11.222, 14.778) \end{aligned}$$

$$\begin{aligned} \text{Fórmula C} & \left( 3 - t_{12,1-0.025} \sqrt{\frac{2.667}{6}}, 3 + t_{12,1-0.025} \sqrt{\frac{2.667}{6}} \right) = \\ & = \left( 3 - 2.178 \sqrt{\frac{2.667}{6}}, 3 + 2.178 \sqrt{\frac{2.667}{6}} \right) = (1.5479, 4.4521) \end{aligned}$$

Se observa que los dos primeros intervalos tienen una zona común, por lo que se puede aceptar la igualdad entre los tiempos de secado de los pegamentos correspondientes a las dos fórmulas A y B. En cambio el último no tiene ninguna zona en común con los primeros. Concluimos que este último tiene un tiempo de secado diferente a los dos primeros.

Realizamos ahora comparaciones múltiples usando el método de Bonferroni al nivel 0.05 y usamos en este caso intervalos de confianza para la diferencia entre las medias:

El número  $n_c$  de comparaciones es  $\binom{3}{2} = 3$ . El valor de  $\alpha^*$  es  $0.05/3 = 0.016667$  y  $\frac{\alpha^*}{2} = 0.00833$

### Contraste AB

$$t_{n-3} = \frac{|\bar{y}_p - \bar{y}_q|}{\sqrt{MCR \left( \frac{1}{n_p} + \frac{1}{n_q} \right)}} = t_{12} = \frac{|10 - 13|}{\sqrt{2.667 \left( \frac{1}{5} + \frac{1}{4} \right)}} = 2.7384$$

$t_{12}^{-1}(1 - 0.00833) = 2.7797 > 2.732$ . Por tanto la diferencia entre las dos primeras fórmulas no es significativa.

El intervalo de confianza para la diferencia de medias es

$$\left( |10 - 13| - 2.7797 \sqrt{2.667 \left( \frac{1}{5} + \frac{1}{4} \right)}, |10 - 13| + 2.7797 \sqrt{2.667 \left( \frac{1}{5} + \frac{1}{4} \right)} \right) = (-4.5199 \times 10^{-2}, 6.0452)$$

Este intervalo contiene el valor 0. Es decir admitimos la igualdad de las medias de A y B

### Contraste AC

$$t_{12} = \frac{|10 - 3|}{\sqrt{2.667 \left( \frac{1}{5} + \frac{1}{6} \right)}} = 7.0787$$

Ahora  $t_{12}^{-1}(1 - 0.00833) = 2.7797 < 7.0787$ . Por tanto la diferencia entre los pegamentos de fórmulas A y C es significativa, es decir pueden considerarse diferentes los tiempos de secado de A y C.

El intervalo de confianza para las diferencias de medias es

$$\left( |10 - 3| - 2.7797 \sqrt{2.667 \left( \frac{1}{5} + \frac{1}{6} \right)}, |10 - 3| + 2.7797 \sqrt{2.667 \left( \frac{1}{5} + \frac{1}{6} \right)} \right) = (4.2512, 9.7488)$$

El intervalo no contiene el valor 0.

### Contraste BC

$$t_{12} = \frac{|13 - 3|}{\sqrt{2.667 \left( \frac{1}{4} + \frac{1}{6} \right)}} = 9.4862$$

Ahora  $t_{12}^{-1}(1 - 0.00833) = 2.7797 < 9.4862$ . Por tanto la diferencia entre los pegamentos de fórmulas B y C es significativa, es decir pueden considerarse diferentes los tiempos de secado de B y C.

El intervalo de confianza para las diferencias de medias es

$$\left( |13 - 3| - 2.7797 \sqrt{2.667 \left( \frac{1}{4} + \frac{1}{6} \right)}, |13 - 3| + 2.7797 \sqrt{2.667 \left( \frac{1}{4} + \frac{1}{6} \right)} \right) =$$

$$= (7.0698, 12.93)$$

El intervalo no contiene el valor 0, por tanto se concluye que hay diferencia entre los tiempos medios de secado de estos dos pegamentos.

Por este procedimiento se concluye que los pegamentos A y B pueden considerarse iguales a efectos de la rapidez de secado, en cambio C es distinto. El pegamento C es el de secado más rápido.

## 13.7 EJERCICIOS PROPUESTOS

**Ejercicio 158** <sup>1</sup>Un vendedor de refrescos está considerando la importancia del color del bote en la cantidad de ventas. El registro del número de unidades vendidas en diferentes tiendas de la ciudad elegidas al azar es el siguiente:

Azul (X)	93, 85, 89
Rojo (Y)	102, 86, 90, 100, 89, 94
Amarillo (Z)	81, 82, 80, 84

1. ¿Se debe concluir que el color tiene alguna influencia sobre la cantidad promedio de unidades vendidas?
2. Hallar un intervalo de confianza (al 95%) para la diferencia de las medias de ventas entre los botes rojos y amarillos

**Ejercicio 159** Los siguientes datos dan el consumo de electricidad diario por habitante realizado en 4 barrios de una ciudad. Los distintos datos provienen de 6 mediciones seleccionados al azar entre las realizadas en los días de un

---

<sup>1</sup>En todos los problemas se supondrá que se cumplen las hipótesis de partida válidas para aplicar el análisis de varianza.

año .

Barrio A	Barrio B	Barrio C	Barrio D
13.1	11.4	10.6	11.5
13.4	12.1	11.1	12.0
13.8	12.1	11.4	12.9
14.4	12.6	12.5	13.4
14.0	12.8	11.7	12.6
14.8	13.4	13.0	14.0
13.9167	12.4	11.7167	12.7333

1. ¿ Se puede considerar diferente el consumo medio por barrio.?
2. Calcular intervalos de confianza para la diferencia entre las medias de consumo entre los barrios usando el método de Bomferroni
3. Se observa que la media en el barrio A es superior a la del barrio C. Es esta diferencia significativa al nivel de significación 0.05?

**Ejercicio 160** El beneficio obtenido (en millones de pesetas) por cinco supermercados en distintos años viene dado en la siguiente tabla

Super. 1	Super. 2	Super. 3	Super. 4	Super. 5
222	196	204	305	128
220	235	190	351	109
170	188	182	351	112
175		190	348	139
155		104		70

1. Hacer la tabla de Análisis de Varianza.
2. ¿Hay evidencia suficiente para concluir que el beneficio es distinto en algunos de estos supermercado?
3. Si es así, indica cuales son y por qué motivo.

**Ejercicio 161** Los datos siguientes se refieren a las pérdidas de peso de ciertas piezas mecánicas debidas a la fricción cuando la usaron tres fabricantes diferentes

Fabricante A	12.2, 11.8, 13.1, 11.0, 3.9, 4.1, 10.3, 8.4
Fabricante B	10.9, 5.7, 13.5, 9.4, 11.4, 15.7, 10.8, 14.0
Fabricante C	12.7, 19.9, 13.6, 11.7, 18.3, 14.3, 22.8, 20.4

Probar al nivel de significación 0.01 si las diferencias entre las medias de desgaste entre los fabricantes es significativa.

**Ejercicio 162** La siguiente tabla recoge el número de disquetes defectuosos fabricados usando diferentes sistemas de fabricación durante seis meses consecutivos.

Sistema A	Sistema B	Sistema C
6	14	10
14	9	12
10	12	7
8	10	15
11	14	11
8	12	11

1. Puede detectarse alguna diferencia significativa entre el número de defectuosos que produce cada sistema de fabricación
2. Hallar un intervalo de confianza para la media de defectos mensuales obtenidos con el Sistema A
3. Hallar un intervalo de confianza para la diferencia entre las media de defectos mensuales obtenidos con el Sistema A Y B.

**Ejercicio 163** Un experimento consiste en determinar el efecto de las burbujas de aire en la resistencia del asfalto. Las burbujas de aire se controlan en tres niveles: Bajo(2% - 4%), Medio(4% - 6%) y Alto(6%-8%).

Los datos medidos sobre la resistencia del asfalto en los distintos niveles son:

Burbujas	Resistencia del asfalto
Bajo	106 90 103 90 79 88 92 95
Medio	80 69 94 91 70 83 87 83
Alto	74 80 62 69 76 85 69 85

¿Los niveles de burbujas de aire influyen en la resistencia del asfalto a un nivel de significación 0.01?

**Ejercicio 164** La siguiente tabla recoge el número de errores cometidos por cuatro cajeras de un supermercado en cinco meses consecutivos.

Cajera A	Cajera B	Cajera C	Cajera D
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

1. Hacer la tabla de análisis de la varianza ¿Puede atribuirse al azar la diferencia de errores entre estas cajeras?

2. Halla un intervalo de confianza para la media de errores cometidos por la primera cajera
3. Halla un intervalo de confianza para la diferencia promedio de errores cometidos por las cajeras A y B

**Ejercicio 165** Se estudia la valoración de los estudiantes de distintos lugares de procedencia sobre la calidad de la residencia universitaria donde habitan, sus valoraciones se recogen en la siguiente tabla:

Procedencia	Valoración
Sevilla	7,5,6,8
Resto de España	6, 8,7,7
Europa	5,4,4,5
América	7,4,4,7

1. Hallar la tabla de análisis de varianza
2. ¿La valoración media depende del lugar de origen?

**Ejercicio 166** Para ver si el precio de un producto alimenticio depende del barrio en que se adquiere, se ha seleccionado al azar un número de tiendas de cada barrio y el precio de este producto en cada una de estas tiendas se ha registrado en la siguiente tabla

Barrio A	Barrio B	Barrio C
210	182	226
192	200	198
183	187	185
227	182	237
242		237
212		

1. Construir la tabla de análisis de la varianza
2. Contrastar la igualdad entre los precios medios del producto en los tres barrios



## Tema 14

# Análisis de la varianza con varios factores

### 14.1 Introducción

**Ejemplo 70** Una empresa fabricante de automóviles ha realizado un estudio sobre la valoración de tres modelos de automóviles,  $M1$ ,  $M2$ ,  $M3$ , dependiendo del lugar donde habitan los posibles clientes. Se han considerado tres tipos de comprador según el entorno en que vive: rural, pequeña ciudad o gran ciudad. Las valoraciones medias recogidas según el modelo y tipo de comprador consultando en seis ubicaciones diferentes para cada uno de los tipos de población vienen resumidas en la siguiente tabla:

	$M1$	$M2$	$M3$
Rural	2.7, 2.9, 3.3	1.9, 2.3, 2.8	3.1, 3.3, 2.9
	1.7, 2.3, 2.8	2.3, 3.1, 1.8	2.7, 2.6, 3.2
Pequeña ciudad	5.1, 4.3, 5.3	3.9, 4.3, 4.8	4.9, 5.3, 5.1
	3.7, 5.6, 5.3	5.1, 5.3, 4.7	6.0, 5.2, 4.9
Gran ciudad	4.1, 3.7, 4.6	2.8, 3.7, 4.2	4.7, 5.6, 5.1
	3.2, 3.9, 4.0	2.9, 3.5, 2.3	4.3, 4.6, 4.4

Planteamos las siguientes preguntas: ¿Se valoran por igual los distintos modelos o, por el contrario, hay modelos más valorados y otros menos valorados? ¿El tipo de entorno en que viven los posibles clientes influye en la valoración dada a los modelos de automóviles? ¿Hay interacción<sup>1</sup> entre el modelo de automovil y el lugar donde habitan los posibles compradores.

En este caso estamos considerando dos factores que podrían tener influencia en la valoración de los automóviles: modelo y lugar de residencia.

<sup>1</sup>Se entiende por interacción el efecto combinado de ambos factores que no está justificado por la influencia que ejerce cada uno de ellos por separado.

El factor *modelo* tiene tres niveles, tratamientos o grupos M1, M2 y M3. El factor *lugar de residencia* se aplica en tres niveles: rural, pequeña ciudad y gran ciudad. El experimento tiene 6 réplicas o elementos en cada una de las 9 combinaciones de tratamientos considerados.

Cuando hay dos factores asignamos los elementos de la muestra al azar a cada casilla. El modelo matemático que se suele emplear es el de dos factores con interacción, que describiremos en el siguiente apartado. Si se concluye que no hay interacción se debe repetir el análisis considerando el modelo de dos factores sin interacción. Si hay interacción, sobre todo si es bastante significativa, se debería estudiar la influencia conjunta de ambos factores (como si cada casilla fuera un nivel de un único factor), porque la interacción oscurece la influencia de los factores por separado y la conclusión del análisis con respecto a los factores por separado puede ser poco útil.

## 14.2 Modelo de Análisis de Varianza con dos factores.

### 14.2.1 Descripción del modelo con interacción

Consideremos dos factores A y B, el primero de ellos con  $a$  niveles y el segundo con  $b$  niveles. Aunque en el ejemplo anterior el número de réplicas por cada una de las nueve combinaciones de tratamientos es el mismo (6 réplicas), desarrollaremos el modelo general, donde no es obligatorio que el número de réplicas, o elementos en cada casilla, sea el mismo. La notación que usamos es la siguiente:

Niveles	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>b</sub>
A <sub>1</sub>	$y_{111}, y_{112}, \dots, y_{11n_{11}}$	$y_{121}, y_{122}, \dots, y_{12n_{12}}$	...	$y_{1b1}, y_{1b2}, \dots, y_{1bn_{1b}}$
A <sub>2</sub>	$y_{211}, y_{212}, \dots, y_{21n_{21}}$	$y_{221}, y_{222}, \dots, y_{22n_{22}}$	...	$y_{2b1}, y_{2b2}, \dots, y_{2bn_{2b}}$
	...	...	...	...
A <sub>a</sub>	$y_{a11}, y_{a12}, \dots, y_{a1n_{a1}}$	$y_{a21}, y_{a22}, \dots, y_{a2n_{a2}}$	...	$y_{ab1}, y_{ab2}, \dots, y_{abn_{ab}}$

14.2. MODELO DE ANÁLISIS DE VARIANZA CON DOS FACTORES.323

En la tabla siguiente se especifica la notación para el número de elementos y para las medias por fila, columna, casilla y total

	Factor B (nivel 1)	Factor B (nivel 2)	...	Factor B (nivel b)	Marginal A
Factor A (nivel 1)	$\bar{y}_{11}, n_{11}$	$\bar{y}_{12}, n_{12}$	...	$\bar{y}_{1b}, n_{1b}$	$\bar{y}_{1.}, n_{1.}$
Factor A (nivel 2)	$\bar{y}_{21}, n_{21}$	$\bar{y}_{22}, n_{22}$	...	$\bar{y}_{2b}, n_{2b}$	$\bar{y}_{2.}, n_{2.}$
	...	...	...	...	
Factor A (nivel a)	$\bar{y}_{a1}, n_{a1}$	$\bar{y}_{a2}, n_{a2}$	...	$\bar{y}_{ab}, n_{ab}$	$\bar{y}_{a.}, n_{a.}$
Marginal B	$\bar{y}_{.1}, n_{.1}$	$\bar{y}_{.2}, n_{.2}$		$\bar{y}_{.b}, n_{.b}$	$n, \bar{y}$

donde

$$n = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \quad y \quad \bar{y} = \frac{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} y_{ijk}}{n}$$

Si consideremos dos factores puede demostrarse que la suma total de cuadrados puede descomponerse de la forma siguiente:

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2 = \sum_{i=1}^a n_{i.} (\bar{y}_{i.} - \bar{y})^2 + \sum_{i=1}^b n_{.j} (\bar{y}_{.j} - \bar{y})^2 + \sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$$

que escribimos abreviadamente en la forma

$$STC = SCF_A + SCF_B + SCF_{AB} + SCR$$

Los grados de libertad son en este caso

$$n - 1 = (a - 1) + (b - 1) + (ab - a - b + 1) + (n - ab)$$

Dividiendo cada suma por sus grados de libertad obtenemos la medias correspondientes.

$$MCF_A = \frac{SCF_A}{a - 1}; \quad MCF_B = \frac{SCF_B}{b - 1};$$

$$MCF_{AB} = \frac{SCF_{AB}}{ab - a - b + 1}; \quad MCR = \frac{SCR}{n - ab}$$

### 14.2.2 Descripción del test de hipótesis

En el modelo teórico se supone que todas las muestras son independientes y proceden de poblaciones normales con igual varianza.

La hipótesis nula consta de tres partes:

1. Las medias correspondientes a todos los niveles del factor A son iguales (el factor A no tiene ningún efecto).
2. Las medias correspondientes a todos los niveles del factor B son iguales (el factor B no tiene ningún efecto).
3. No existe interacción (no hay efecto combinado de ambos factores).

La hipótesis alternativa, que también consta de tres partes, es la negación de cada una de las partes de la hipótesis nula.

Para comprobar cada una de estas hipótesis se usan, respectivamente, los estadísticos:

$F_A(\text{exp}) = \frac{MCF_A}{MCR}$ , con  $a - 1$  g.l. en el numerador y  $n - ab$  g.l. en el denominador

$F_B(\text{exp}) = \frac{MCF_B}{MCR}$ , con  $b - 1$  g.l. en el numerador y  $n - ab$  g.l. en el denominador

$F_{AB}(\text{exp}) = \frac{MCF_{AB}}{MCR}$ , con  $ab - a - b + 1$  g.l. en el numerador y  $n - ab$  g.l. en el denominador

Se rechaza la hipótesis nula correspondiente a una de las partes del contraste si el valor correspondiente estadístico,  $F$  experimental, es mayor que la  $F$  teórica (con los grados de libertad y nivel de significación especificados)

Lo mejor es efectuar primero la prueba de interacción, ya que si ésta es significativa el efecto de los factores no tiene sentido práctico sino que lo que interesa es la interacción, procediendo, en este caso, a estudiar cada una de las casillas por separado.

La siguiente tabla es la salida de Statgraphics correspondiente al problema del ejemplo.

Variabilidad	g.l	Sum. cuad.	media cuad.	$F_{\text{exp}}$	$p - \text{value}$
tipo de población	2	47.338	23.67	84.165	0.0000
modelo de vehículo	2	7.3526	3.6763	13.073	0.0000
Interaction	4	1.9085	0.47713	1.6966	0.167
Residual	45	12.655	0.28122		
	53	69.254	1.6966		

Los programas de ordenador suelen suministrar los  $p - \text{values}$ .

El cálculo de los  $p - \text{values}$  se realiza de la siguiente forma:

$$\begin{aligned}1 - F_{2,45}(84.165) &= 0.0000 \\1 - F_{2,45}(13.073) &= 3.3412 \times 10^{-5} \\1 - F_{2,45}(1.6966) &= 0.167\end{aligned}$$

Se acepta la hipótesis nula al nivel de confianza 95%,  $\alpha = 0.05$ , si el  $p$ -value es mayor que  $\alpha$ . Se concluye por tanto que no hay interacción al nivel  $\alpha = 0.05$ , ya que  $0.1674 > 0.05$ . Se acepta la hipótesis de que no existe interacción, pero se rechazan las otras dos hipótesis ( $0.0000 < 0.05$ ), concluyendo por tanto que hay influencia de ambos factores. La valoración de los automóviles depende por tanto del modelo considerado y del lugar de residencia.

### 14.2.3 Descripción del modelo sin interacción

Cuando no existe interacción suele repetirse el análisis considerando también como residual el efecto de la interacción, que se suma al residual anterior. En este caso el modelo empleado sería el llamado *Modelo con dos factores sin interacción*. Este modelo se aplica también cuando sabemos, por experiencias previas, que no hay interacción entre los factores. También se usa este modelo cuando uno de los factores actúa como bloque. El concepto de factor bloque se trata en el epígrafe 14.3.

El modelo de análisis de varianza de dos factores sin interacción se realiza haciendo la descomposición:

$$STC = SCF_A + SCF_B + SCR'$$

siendo  $SCR' = SCF_{AB} + SCR$ , aunque es más fácil si se despeja de la relación anterior:

$$SCR' = STC - (SCF_A + SCF_B)$$

Los cálculos a realizar son los siguientes:

$$\begin{aligned}STC &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2, \quad SCF_A = \sum_{i=1}^a n_i (\bar{y}_i - \bar{y})^2, \\SCF_B &= \sum_{j=1}^b n_{.j} (\bar{y}_{.j} - \bar{y})^2, \quad SCR' = STC - (SCF_A + SCF_B)\end{aligned}$$

Los grados de libertad de estas sumas son, respectivamente:

$$n - 1, \quad a - 1, \quad b - 1 \quad \text{y} \quad n - b - a + 1$$

Dividiendo cada suma por sus grados de libertad obtenemos la medias correspondientes.

$$MCF_A = \frac{SCF_A}{a - 1}; \quad MCF_B = \frac{SCF_B}{b - 1};$$

$$MCR' = \frac{SCF_{AB} + SCR}{(ab - a - b + 1) + (n - ab)} = \frac{SCR'}{n - a - b + 1}$$

El esquema de la tabla ANOVA para este modelo es el siguiente:

326 TEMA 14. ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES

Fuente de variación	gr. de lib	Suma de cuadrados
A	a-1	$SCF_A$
B	b-1	$SCF_B$
Residual	n-a-b+1	$SCR'$
Total	n-1	$STC$

Fuente de variación	Medias de cuadrados	$F_{exp}$
A	$MCF_A = \frac{SCF_A}{a-1}$	$\frac{MCF_A}{MCR'}$
B	$MCF_B = \frac{SCF_B}{b-1}$	$\frac{MCF_B}{MCR'}$
Residual	$MCR' = \frac{SCR'}{n-a-b+1}$	
Total		

Se omite de esta forma el efecto de la interacción, teniendo únicamente en cuenta la influencia de los factores A y B.

**Ejemplo 71** Los valores de la siguiente tabla indican el número medio de errores por página cometidos por dos secretarías trabajando con diferentes máquinas. Los valores se han tomado durante cuatro días elegidos al azar.

	Maquina 1	Maquina 2	Maquina 3
Secr. 1	10.96, 11.03	10.95, 11.00	11.07, 11.01
	11.08, 11.01	11.04, 10.97	10.97, 11.03
Secr. 2	10.97, 10.96	10.97, 10.96	11.02, 11.00
	10.94, 10.95	10.97, 10.98	11.01, 11.01

Considerando un modelo de dos factores con interacción, se pide:

1. Hallar la tabla de análisis de la varianza
2. ¿Es la interacción significativa?
3. Analizar si hay diferencia de rendimiento entre las secretarías.
4. ¿Tiene el tipo de máquina algún efecto en el rendimiento de las secretarías?

La tabla de las medias es

	Maquina 1	Maquina 2	Maquina 3	Medias
Secr. 1	11.02	10.99	11.02	Secr. 1=11.01
Secr. 2	10.955	10.97	11.01	Secr. 2= 10.98
Medias	10.9875	10.98	11.015	Media total=10.99

$$SCF_{(A = \text{sec})} = \sum_{i=1}^a n_i (\bar{y}_i - \bar{y})^2 = 12(11.01 - 10.99)^2 + 12(10.98 - 10.99)^2 = 0.006$$

$$SCF_{(B = \text{maq})} = \sum_{j=1}^b n_j (\bar{y}_j - \bar{y})^2 = 8(10.9875 - 10.99)^2 + 8(10.98 - 10.99)^2 + 8(11.015 - 10.99)^2 = .00585$$

14.2. MODELO DE ANÁLISIS DE VARIANZA CON DOS FACTORES.327

$$\begin{aligned}
 SCF_{AB} &= \sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2 = \\
 &= 4(11.02 - 11.01 - 10.9875 + 10.99)^2 + 4(10.99 - 11.01 - 10.98 + 10.99)^2 \\
 &+ 4(11.02 - 11.01 - 11.015 + 10.99)^2 + 4(10.955 - 10.98 - 10.9875 + 10.99)^2 \\
 &+ 4(10.97 - 10.98 - 10.98 + 10.99)^2 + 4(11.01 - 10.98 - 11.015 + 10.99)^2 \\
 &= .00405
 \end{aligned}$$

$$SCR = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2 = (10.96 - 11.02)^2 + (11.03 - 11.02)^2 + (11.08 - 11.02)^2 + \dots + (11.01 - 11.01)^2 = 0.0181$$

Por tanto la suma total de cuadrados es:

$$STC = 0.006 + 0.00585 + 0.00405 + 0.0181 = 0.034 = 0.034.$$

Las operaciones realizadas también pueden ordenarse como en la siguientes tablas, que además pueden servir de guía para confeccionar una hoja de cálculo, por ejemplo con Excel, que sirva para realizar este tipo de cálculos.

$y_{ijk}$	sec ( $i$ )	maq ( $j$ )	$\bar{y}$	$\bar{y}_{i.}$	$\bar{y}_{.j}$	$\bar{y}_{ij}$	$(\bar{y}_{i.} - \bar{y})^2$
10.96	1	1	10.99	11.01	10.9875	11.02	0.0004
11.03	1	1	10.99	11.01	10.9875	11.02	0.0004
11.08	1	1	10.99	11.01	10.9875	11.02	0.0004
11.01	1	1	10.99	11.01	10.9875	11.02	0.0004
10.95	1	2	10.99	11.01	10.98	10.99	0.0004
11	1	2	10.99	11.01	10.98	10.99	0.0004
11.04	1	2	10.99	11.01	10.98	10.99	0.0004
10.97	1	2	10.99	11.01	10.98	10.99	0.0004
11.07	1	3	10.99	11.01	11.015	11.02	0.0004
11.01	1	3	10.99	11.01	11.015	11.02	0.0004
10.97	1	3	10.99	11.01	11.015	11.02	0.0004
11.03	1	3	10.99	11.01	11.015	11.02	0.0004
10.97	2	1	10.99	10.98	10.9875	10.955	1E-04
10.96	2	1	10.99	10.98	10.9875	10.955	1E-04
10.94	2	1	10.99	10.98	10.9875	10.955	1E-04
10.95	2	1	10.99	10.98	10.9875	10.955	1E-04
10.97	2	2	10.99	10.98	10.98	10.97	1E-04
10.96	2	2	10.99	10.98	10.98	10.97	1E-04
10.97	2	2	10.99	10.98	10.98	10.97	1E-04
10.98	2	2	10.99	10.98	10.98	10.97	1E-04
11.02	2	3	10.99	10.98	11.015	11.01	1E-04
11	2	3	10.99	10.98	11.015	11.01	1E-04
11.01	2	3	10.99	10.98	11.015	11.01	1E-04
11.01	2	3	10.99	10.98	11.015	11.01	1E-04
			SUMAS	DE	CUADRADOS		0.006

328TEMA 14. ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES

$(\bar{y}_{.j} - \bar{y})^2$	$(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$	$(y_{ijk} - \bar{y}_{ij})^2$	$(y_{ijk} - \bar{y})^2$
6.25E-06	0.00015625	0.0036	0.0009
6.25E-06	0.00015625	1E-04	0.0016
6.25E-06	0.00015625	0.0036	0.0081
6.25E-06	0.00015625	1E-04	0.0004
1E-04	0.0001	0.0016	0.0016
1E-04	1E-04	1E-04	1E-04
1E-04	1E-04	0.0025	0.0025
1E-04	1E-04	0.0004	0.0004
0.000625	0.000225	0.0025	0.0064
0.000625	0.000225	1E-04	0.0004
0.000625	0.000225	0.0025	0.0004
0.000625	0.000225	1E-04	0.0016
6.25E-06	0.00050625	0.000225	0.0004
6.25E-06	0.00050625	2.5E-05	0.0009
6.25E-06	0.00050625	0.000225	0.0025
6.25E-06	0.00050625	2.5E-05	0.0016
1E-04	3.15544E-30	3.15544E-30	0.0004
1E-04	0	1E-04	0.0009
1E-04	0	0	0.0004
1E-04	0	1E-04	1E-04
0.000625	2.5E-05	1E-04	0.0009
0.000625	2.5E-05	1E-04	1E-04
0.000625	2.5E-05	0	0.0004
0.000625	2.5E-05	0	0.0004
0.00585	0.00405	0.0181	0.034

Los grados de libertad correspondientes son

$$a - 1 = 1, \quad b - 1 = 2, \quad ab - a - b + 1 = 2, \quad n - ab = 24 - 6 = 18$$

Analizamos en primer lugar si la interacción es significativa:

$$F_{AB}(\text{exp}) = \frac{MCF_{AB}}{MCR} = \frac{SCF_{AB}}{ab - a - b + 1} = \frac{0.00405}{\frac{SCR}{n-ab}} = \frac{2}{\frac{0.0181}{18}} = 2.0138$$

El valor de la  $F$  teórica con 2, 18 grados de libertad correspondiente al nivel 0.05 es 3.55. Por lo tanto se admite la hipótesis nula respecto a la interacción, puesto que  $2.0138 < 3.55$ , concluyendo que no hay interacción entre los dos factores.

Analizamos si existe diferencia entre las secretarías:



$$F_A(\text{exp}) = \frac{MCF_A}{MCR} = \frac{\frac{SCF_A}{b-1}}{\frac{SCR}{n-ab}} = \frac{\frac{0.006}{1}}{\frac{0.0181}{18}} = 5.9669 = 6.0663$$

Este valor experimental sobrepasa el valor teórico  $F_{1,18}^{-1}(0.05) = 4.41$  Por tanto se rechaza la hipótesis nula, concluyendo por tanto que existe diferencia en el rendimiento entre las secretarias.

Analizamos ahora si existe diferencia entre las maquinas

$$F_A(\text{exp}) = \frac{MCF_B}{MCR} = \frac{\frac{SCF_B}{a-1}}{\frac{SCR}{n-ab}} = \frac{\frac{0.00585}{2}}{\frac{0.0181}{18}} = 2.90884$$

Este valor experimental es menor que el valor teórico  $F_{2,18}(0.05) = 3.55$ , por lo que se acepta la hipótesis nula respecto a las maquinas, concluyendo que no hay diferencia entre ellas. Las maquinas son causantes del mismo número de errores por promedio.

La tabla de análisis de varianza (Tabla ANOVA) correspondiente al modelo sería:

Fuente de variación	f.l	Suma de cuadrados	Medias de cuadrados
secretarias	1	0.006	$\frac{0.006}{1} = 0.006$
maquinas	2	0.00585	$\frac{0.00585}{2} = 2.925 \times 10^{-3}$
Interactions AB	2	0.00405	$\frac{0.00405}{2} = 2.025 \times 10^{-3}$
Residual	18	0.0181	$\frac{0.0181}{18} = 1.0056 \times 10^{-3}$
Total	23	0.034	

Fuente de variación	$F_{\text{exp}}$	p-value
secretarias	$\frac{0.006}{1.0056 \times 10^{-3}} = 5.9666$	0.025
maquinas	$\frac{2.925 \times 10^{-3}}{1.0056 \times 10^{-3}} = 2.9087$	0.08
Interactions AB	$\frac{2.025 \times 10^{-3}}{1.0056 \times 10^{-3}} = 2.0137$	0.16247
Residual		
Total		

El cálculo de los P-values se realiza de la siguiente forma:

$$1 - \text{FDist}(2.9087; 2, 18) = 8.0427 \times 10^{-2} = 0.08$$

$$1 - \text{FDist}(5.9666; 1, 18) = 2.5121 \times 10^{-2} = 0.025$$

$$1 - \text{FDist}(2.0137; 2, 18) = 0.16247$$

Valores del P-value por debajo del 5% nos harían rechazar la hipótesis nula. Como cabía esperar usando los p-valies se llega a las mismas conclusiones. Se acepta la igualdad entre las máquinas ( $0.08 > 0.05$ ), se rechaza la igualdad entre las secretarias ( $0.025 < 0.05$ ), se acepta que no hay interacción entre las máquinas y las secretarias ( $0.16247 > 0.05$ ).

Si optamos por el modelo sin interacción, ya que hemos aceptado en el análisis anterior que no existe, la tabla ANOVA quedaría:

Fuente de variación	f.l	Suma de cuadrados
secretarias	1	0.006
maquinas	2	0.00585
Residual	18+2=20	0.0181 + 0.00405 = .02215
Total	23	0.034

Fuente de variación	Medias de cuadrados	F <sub>exp</sub>	p-value
secretarias	$\frac{0.006}{1} = .006$	$\frac{0.006}{1.1 \times 10^{-3}} = 5.4545$	0.030
maquinas	$\frac{0.00585}{2} = 2.9 \times 10^{-3}$	$\frac{2.9 \times 10^{-3}}{1.1 \times 10^{-3}} = 2.6411$	0.096
Residual	$\frac{0.02215}{20} = 1.1 \times 10^{-3}$		
Total			

La conclusión, en este caso, es la misma para las secretarias y las maquinas que usando el modelo con interacción.

### 14.3 Diseño por bloques completos al azar

**Ejemplo 72** *Se realiza un experimento para estudiar si la proporción de un cierto aditivo que se añade al gasoil influye en el consumo de este combustible. Se sospecha que el tipo de autobus utilizado puede tener también influencia en el consumo y por ello se han realizado pruebas con 20 autobuses de cuatro modelos diferentes, cinco de cada modelo. Para cada una de las proporciones de aditivos añadidas al gasoil se han utilizado los cuatro modelos, pero el camión concreto que se ha asignado para probar cada proporción de aditivo se ha elegido al azar entre los cinco de dicho modelo. Los valores de la tabla indican el consumo en litros por 100 Km de recorrido.*

% de aditivo	Modelo 1	Modelo 2	Modelo 3	Modelo 4
0%	15.4	10.6	17.8	14.6
1%	10.3	5.5	10.9	8.9
2%	7.4	1.2	8.1	5.5
3%	10.7	6.5	9.6	8.9
4%	13.5	11.6	15.5	13.5

En este caso la variable que se quiere controlar es la influencia del contenido de aditivo en la gasoil. El modelo de autobús es una variable controladas por el experimentador, que da una clasificación de la muestra en

distintas categorías. Suele darse a cada una de estas categorías el nombre de *Bloques*.

Cada uno de los  $a$  niveles del factor (% de aditivo), cuya influencia en la variable respuesta (el consumo de gasoil) se desea investigar se mide una vez en cada uno de los  $b$  bloques (modelos de camión). El orden de aplicación de los tratamientos dentro de los bloques es aleatorio. Un experimento completo dentro de este modelo consta de  $ab$  medidas. La eficacia de este diseño depende de que los efectos de los niveles del factor que actúa como bloque sean importantes. Es decir que las medias de, por lo menos, alguno de los niveles del factor bloque deben ser diferentes (por promedio, al menos alguno de los modelos de camión debe consumir diferente cantidad de gasolina). Si no es así es preferible utilizar el modelo anterior (diseño completamente aleatorizado).

El diseño de este experimento del estudio de la mejor proporción de aditivo en el gasoil es un *Diseño por bloques completos al azar*.

Explicamos ahora, por medio del ejemplo, el motivo de la denominación de este diseño: Se dice que el *diseño es por bloques* porque se realiza una clasificación (los bloques) de los 20 valores de consumo de gasoil por el modelo de autobús. Se llama diseño en bloques *completos*, porque la muestra tiene representación en todos los bloques (modelos de autobús) y niveles del factor que se estudia (proporción de aditivo). En otras palabras: No hay ninguna casilla en blanco. La última denominación (*al azar*) está justificada porque la elección de los autobuses a los que se va a añadir una cierta proporción de aditivo se realiza al azar dentro de cada modelo. El Diseño en Bloques Completos al azar se realiza considerando que no hay interacción entre el factor bajo estudio y los bloques, por lo que el modelo a aplicar es el de dos factores sin interacción.

El diseño siguiente está generado por Statgraphics para un modelo en bloques completos al azar. El factor en estudio consta de cuatro niveles y la muestra viene clasificada en tres niveles a causa del factor bloque.

## 332TEMA 14. ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES

run	FACTOR	BLOQUE	Respuesta
1	1	1	
2	1	3	
3	1	2	
4	2	2	
5	2	1	
6	2	3	
7	3	2	
8	3	1	
9	3	3	
10	4	1	
11	4	3	
12	4	2	

Con este diseño se podía haber realizado el siguiente experimento:

**Ejemplo 73** *Se diseñó un experimento para estudiar el rendimiento de 4 detergentes diferentes. Las lecturas de blancura se obtuvieron con un equipo diseñado para 12 cargas de lavado y distribuidos en tres modelos de lavadora, ya que se sospechaba que la clase de lavadora también tenía influencia en la blancura conseguida. Usando el diseño anterior se han obtenido las siguientes lecturas de blancura:*

run	Detergente	lavadora	Respuesta
1	1	1	45
2	1	3	51
3	1	2	43
4	2	2	46
5	2	1	47
6	2	3	52
7	3	2	50
8	3	1	48
9	3	3	55
10	4	1	42
11	4	3	49
12	4	2	37

*Esta forma de recoger los datos es adecuada para analizarlos con el programa statgraphics. Para realizar los cálculos a mano es más adecuada la*

siguiente organización.

	Lavadora 1	Lavadora 2	Lavadora 3
Detergente A	45	43	51
Detergente B	47	46	52
Detergente C	48	50	55
Detergente D	42	37	49

Considerando los detergentes como tratamientos y las lavadoras como bloques, se pide:

1. Hallar la tabla de análisis de la varianza
2. Contrastar al nivel 0.01 si existe diferencia entre los detergentes o entre las lavadoras

Tabla de las medias

	Lavadora 1	Lavadora 2	Lavadora 3	medias detergentes
Detergente A	45	43	51	46.334
Detergente B	47	46	52	48.334
Detergente C	48	50	55	51.000
Detergente D	42	37	49	42.667
medias lavadoras	45.5	44	51.75	media total=47.083

Suma de los cuadrados debidos al bloqueo (lavadora)

$$SCF_{lav} = 4(45.5 - 47.083)^2 + 4(44 - 47.083)^2 + 4(51.75 - 47.083)^2 = 135.17$$

Suma de los cuadrados debidos al factor en estudio (detergente)

$$SCF_{det} = 3(46.334 - 47.083)^2 + 3(48.334 - 47.083)^2 + 3(51 - 47.083)^2 + 3(42.667 - 47.083)^2 = 110.91$$

Suma total de cuadrados

$$STC = (45 - 47.083)^2 + (43 - 47.083)^2 + (51 - 47.083)^2 + (47 - 47.083)^2 + (46 - 47.083)^2 + (52 - 47.083)^2 + (48 - 47.083)^2 + (50 - 47.083)^2 + (55 - 47.083)^2 + (42 - 47.083)^2 + (37 - 47.083)^2 + (49 - 47.083)^2 = 264.92$$

$$SCR = 264.92 - (110.91 + 135.17) = 18.84$$

Tabla de análisis de la varianza

Fuente de variación	g.l.	Sum. cuadr.	medias cuadr.	F
Detergente	3	110.91	$\frac{110.9}{3} = 36.967$	$\frac{36.967}{3.14} = 11.773$
Lavadora	2	135.17	$\frac{135.17}{2} = 67.585$	$\frac{67.585}{3.14} = 21.524$
Residual	6	18.84	$\frac{18.84}{6} = 3.14$	
total	11	264.92		

Comparando la F experimental, correspondiente al efecto de los detergentes con  $F_{3,6}(0.01) = 9.78$

Se comprueba que la F experimental supera a la teórica por lo que concluimos que hay diferencia entre las medidas de blancura conseguidas por los distintos detergentes.

Consideremos ahora, aunque no sea motivo de estudio, el efecto de las lavadoras:

$$F_{2,6}(0.01) = 10.925, 21.524 > 10.925$$

Así que concluimos que también las lavadoras son diferentes, al 99% de confianza. El hecho de que las lavadoras sean diferentes entre sí, justifica la decisión de considerar este factor (lavadora) como bloque para el estudio de la calidad de los detergentes.

Si no hubiéramos tenido en cuenta la influencia de la lavadora, es decir que se aplicará el modelo de análisis de varianza con un solo factor, se obtendría la siguiente tabla de análisis de varianza:

F. de variación	g.l.	Sum. cuadr.	medias cuadr.	F
Detergente	3	110.91	$\frac{110.9}{3} = 36.967$	$\frac{36.967}{19.25} = 1.92$
Residual	8=6+2	154=18.84+135.17	$\frac{154}{8} = 19.25$	
total	11	264.92		

$F_{3,8}(0.01) = 7.5910$ . En este caso la F experimental es menor que la teórica, por lo que la conclusión sería que la diferencia entre los detergentes no es significativa. En este caso la variación que hay entre las lavadoras ha oscurecido la influencia de los detergentes y no nos ha permitido detectar la diferencia entre ellos.

## 14.4 Principios básicos para el diseño de experimentos

Como ya se ha tenido ocasión de observar, las estrategias básicas del diseño de experimentos son: la aleatorización, la replicación, y la clasificación por bloques.

**La aleatorización**

El principio de aleatorización exige que los factores no controlados por el experimentador se asignen al azar a los niveles del factor a estudiar. Es decir que debe haber una asignación al azar tanto del material que se usa en la experimentación como el orden en que se realizan las mediciones. Esto ayuda a cancelar la influencia de los diferentes factores que por su incidencia en el error experimental podría contribuir al oscurecimiento de los resultados. Además, por lo general, y así es en el caso de la técnica de Análisis de la Varianza, los modelos estadísticos utilizan la hipótesis de independencia y aleatorización en la obtención de los datos de la muestra.

**La replicación**

Consiste en la repetición del experimento básico. En el ejemplo 66 de la página 303, sobre la velocidad de producción de cuatro máquinas, el experimento básico puede consistir en realizar cuatro pruebas, una con cada máquina. Repitiendo este experimento básico tres veces se obtendrían las 12 medidas. Cada réplica consiste de cuatro ensayos: uno para cada máquina que se desea comparar. En el caso tratado en este ejemplo habría por tanto tres réplicas. De este modo se puede conseguir una estimación del error experimental, ya que las producciones obtenidas por los tres operarios que emplean la misma máquina serán diferentes, permitiendo estos 3 valores obtener una estimación de la varianza de la variable en estudio, lo que va permitir dar un elemento comparativo para determinar lo que puede haber de error experimental o de diferencia real en el número de unidades producidas que sea asignable a las distintas máquinas.

**El bloqueo**

Es una técnica que se utiliza para disminuir el error experimental, es decir que se trata de obtener medidas lo más homogéneas que sea posible. Para ello las medidas se realizan en idénticas condiciones experimentales. Volvemos a considerar el ejemplo 66: se podría pensar que la habilidad del operario que trabaja en cada máquina tiene influencia en la producción. Para minimizar la confusión que podría inducir en las medidas de la producción el operario concreto que ha realizado la tarea podríamos poner a cada empleado a trabajar con cada una de las cuatro máquinas. Por ejemplo seleccionamos solamente 3 empleados y cada uno de ellos trabajará con cada una de las cuatro máquinas. Aquí también tendríamos un total de 12 datos, y tres medidas para cada una de las máquinas, pero ya no puede decirse que la diferencia en la producción sea achacable a la diferencia entre los obreros, puesto que cada máquina ha sido manejada por los tres obreros. Tendremos la producción de las máquinas clasificadas según el operario que la haya obtenido. En este caso el operario actúa como factor bloque. En resumen, el diseño por bloques es un plan para reunir valores de la variable respuesta (unidades producidas) obtenidos cuando cada uno de los tratamientos en

estudio(maquina) se aplica una vez en cada uno de los bloques (operario).

## 14.5 Otros diseños de experimentos

**Modelos de efectos aleatorios:** En los diseños experimentales descritos en este tema los niveles del factor son fijos. Se llaman **modelos de efectos fijos**. Por ejemplo las cuatro maquinas consideradas para estudiar la producción. A veces el factor considerado puede tener una gran cantidad de niveles y es demasiado costoso experimentar con todos ellos. Si el experimentador selecciona aleatoriamente algunos de estos niveles se dice que el **factor es aleatorio**. Estos diseños experimentales suelen llamarse **Modelos de efectos aleatorios**

**Diseño en bloques completos al azar y medidas repetidas:** Es un caso particular de diseño de bloques completos al azar en el que cada bloque esta formado por un solo individuo al que se asignan en un orden aleatorio los niveles del factor que se investiga.

En el caso del ejemplo 14.3, del aditivo para el gasoil, este diseño consistiría emplear un sólo autobús de cada modelo. Con este único autobus se probarían sucesivamente, en un orden aleatorio, todas las proporciones de aditivo. A continuación se haría lo mismo con el único autobús del modelo 2, 3 y 4. También se harían 20 pruebas, pero usando únicamente cuatro autobuses. En cambio en el diseño de bloques completos al azar del ejemplo 14.3, el número de autobuses que participan en el experimento es de veinte.

**Diseño con más de dos factores:** Puede que deseemos conocer la influencia de más de dos factores sobre la variable en estudio. En este caso se precisa una muestra aleatoria independiente para cada una de las combinaciones de los factores. Así si hay tres factores con  $a$ ,  $b$ ,  $c$  niveles cada uno de ellos, precisamos como mínimo  $abc$  medidas. En este caso solo habría un dato para cada combinación de los niveles de los tres factores (cada casilla). En este caso, pueden estudiarse las interacciones entre dos factores ( $AB$ ,  $AC$  y  $BC$ ). Si se replica el experimento habría varios elementos en cada casilla, y podría estudiarse también la interacción entre los tres factores  $ABC$ .

**Diseños  $2^k$ :** Este diseño es útil en las primeras fases de la investigación donde seguramente hay que considerar muchos factores que pudieran tener influencia en la variable respuesta. El objetivo del estudio sería reducir el número de factores en estudio desechando los que no sean de mucha influencia y reteniendo los más influyentes para un examen más detallado. Para simplificar lo más posible el número de medidas necesarias solo se consideran dos niveles para cada uno de los  $k$  factores. Una réplica completa para un experimento de este tipo contiene  $2^k$  observaciones.

**El diseño en cuadrado latino:** Un cuadrado latino de  $4 \times 4$  podría ser



el siguiente

b	d	a	c
d	b	c	a
c	a	d	b
a	c	b	d

Un cuadrado latino de  $n \times n$  elementos tiene las mismas  $n$  letras en cada fila o columna. En cada fila o columna debe aparecer una sola vez cada letra

El diseño en cuadrado latino puede considerar tres factores con el mismo número de niveles cada uno. Supongamos que queremos analizar tres factores con cuatro niveles cada uno. El diseño factorial más simple (una sola réplica) precisa  $4^3 = 64$  medidas. Un diseño alternativo que precisa solamente 16 medidas es el diseño en cuadrado latino que se puede emplear cuando existan tres factores en estudio con el mismo número de niveles cada uno y no se espera que haya interacción entre los factores. El número mínimo de medidas necesarias para el diseño en cuadrado latino es  $n^2$ . Este diseño se aplica cuando solo hay un factor de interés y los otros dos actúan como bloques.

Mostramos a continuación un esquema del experimento correspondiente al anterior cuadrado latino. A y B son los factores de cuatro niveles cada uno que actúan como bloques y C es el factor cuya influencia quiere estudiarse. Los niveles 1, 2, 3, 4 del factor C se asignan a las letras a, b, c, d del cuadrado latino anterior.

El esquema del experimento será el siguiente

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>
B <sub>1</sub>	C <sub>2</sub>	C <sub>4</sub>	C <sub>1</sub>	C <sub>3</sub>
B <sub>2</sub>	C <sub>4</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>1</sub>
B <sub>3</sub>	C <sub>3</sub>	C <sub>1</sub>	C <sub>4</sub>	C <sub>2</sub>
B <sub>4</sub>	C <sub>1</sub>	C <sub>3</sub>	C <sub>2</sub>	C <sub>4</sub>

Por ejemplo la medida correspondiente a la primera casilla se realizara con los niveles 1 para el factor A, 1 para el factor B y 2 para el factor C. Análogamente se procede con las siguientes medidas. De esta forma para cada nivel de C hay cuatro medidas en las que están representados todos los niveles de los bloques A y B.

**El diseño en cuadrado grecolatino:** Consiste en la superposición de dos cuadrados latinos uno con letras griegas y otro con letras latinas. Además cada letra griega aparece una sola vez con cada letra latina. Por ejemplo, el cuadrado grecolatino siguiente

a $\alpha$	b $\beta$	c $\gamma$	d $\delta$
b $\delta$	a $\gamma$	d $\beta$	c $\alpha$
c $\beta$	d $\alpha$	a $\delta$	b $\gamma$
d $\gamma$	c $\delta$	b $\alpha$	a $\beta$

se usa cuando hay cuatro factores tres de los cuales actúan como bloques y solo un factor en estudio. Todos los factores tienen el mismo número de niveles. El diseño factorial con 4 factores de  $n$  niveles cada uno de ellos, requeriría como mínimo  $n^4$  medidas, una por cada combinación de los  $n$  niveles de los cuatro factores. En cambio con este diseño el experimento puede hacerse con  $n^2$  medidas.

En concreto, el diseño en cuadrado grecolatino correspondiente al cuadrado del ejemplo correspondería al estudio de la influencia del factor D, que tiene 4 niveles, en la variable respuesta. Actuarían como bloques los factores A, B, C, cada uno de ellos con cuatro niveles. No se espera que haya interacción entre los cuatro factores. El diseño está esquematizado en la siguiente tabla. Se tomarían 16 medidas de la variable respuesta. Por ejemplo la medida correspondiente a la casilla recuadrada se realizaría aplicando, los niveles 3, 3, 1 de los factores A, B, C, que actúan como bloques, y el 4 del factor D, que es el que está en estudio. En este diseño se realiza una medida de la variable respuesta en los cuatro niveles del factor en estudio para cada uno de los cuatro niveles de los tres factores que actúan como bloque.

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>
B <sub>1</sub>	C <sub>1</sub> D <sub>1</sub>	C <sub>2</sub> D <sub>2</sub>	C <sub>3</sub> D <sub>3</sub>	C <sub>4</sub> D <sub>4</sub>
B <sub>2</sub>	C <sub>2</sub> D <sub>4</sub>	C <sub>1</sub> D <sub>3</sub>	C <sub>4</sub> D <sub>2</sub>	C <sub>3</sub> D <sub>1</sub>
B <sub>3</sub>	C <sub>3</sub> D <sub>2</sub>	C <sub>4</sub> D <sub>1</sub>	C <sub>1</sub> D <sub>4</sub>	C <sub>2</sub> D <sub>3</sub>
B <sub>4</sub>	C <sub>4</sub> D <sub>3</sub>	C <sub>3</sub> D <sub>4</sub>	C <sub>2</sub> D <sub>1</sub>	C <sub>1</sub> D <sub>2</sub>

Con Statgraphics se pueden construir diseños de cuadrados latinos y grecolatinos con diversos número de niveles en cada factor, así como analizar estos modelos y otros muchos diseños experimentales que no se enumeran en este breve resumen.

Para una información más detallada, incluyendo los modelos matemáticos apropiados para analizar una gran cantidad de diseños experimentales, puede consultarse, por ejemplo, el libro *Diseño y Análisis Estadístico de Experimentos* de Douglas C. Montgomery editado por Grupo Editorial Iberoamérica.

## 14.6 EJERCICIOS PROPUESTOS

**Ejercicio 167** <sup>1</sup>En un experimento para investigar la calidad de un plástico se ha medido la resistencia del material en diferentes condiciones de temperatura y de humedad, Se han realizado dos medidas de la resistencia del material obtenido para cada combinación de temperatura y humedad. Los valores obtenidos se muestran en la tabla siguiente:

	temperatura 1	temperatura 2
Humedad 1	13, 11	8, 9
Humedad 2	11, 10	6, 7

1. Halla la tabla del análisis de varianza usando un modelo con interacción.
2. ¿Es la interacción significativa?
3. ¿Tiene influencia en la resistencia del plástico los cambios de temperatura?
4. ¿Tienen influencia los cambios de humedad?

**Ejercicio 168** Se realiza un experimento para estudiar la influencia de los neumáticos de los coches en el desgaste de las pastillas de freno. Para ello se realizan recorridos con 5 marcas de neumáticos. Como se sospecha que el desgaste debe tener relación con el tipo de suelo, se realizan recorridos idénticos por suelo asfaltado de autopista, carretera comarcal y camino rural. El resultado de la prueba se resume en la siguiente tabla donde la variable respuesta es una medida del desgaste.

	Comarcal	Autopista	Rural	Medias por aditivo
Neumático 1	14.4	10.6	18.8	14.6
Neumático 2	11.3	5.5	9.9	8.9
Neumático 3	7.4	2.2	7.1	5.6
Neumático 4	10.7	5.5	10.6	8.9
Neumático 5	13.5	11.6	15.5	13.5
Medias por modelos	11.5	7.1	12.4	10.3

1. ¿Influye la marca de los neumáticos en el desgaste? Utiliza el modelo de bloques completos al azar

---

<sup>1</sup>En todos los problemas se supondrá que se cumplen las hipótesis de partida válidas para aplicar el análisis de varianza.

340TEMA 14. ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES

2. ¿Es aceptable utilizar el tipo de carretera como bloque?

**Ejercicio 169** Para estudiar el consumo de aceite de un motor se prueban 4 motores distintos con 3 tipos de aceite obteniéndose 12 medidas de consumo. Se han obtenido los resultados siguientes :

Suma de los cuadrados del factor aceite=100

Suma de los cuadrados del factor motor=80

Suma total de cuadrados=220

Se pide:

1. Escribe la tabla ANOVA, considerando que no hay interacciones entre los factores.
2. ¿Se puede considerar que los tipos de aceite no tienen influencia en el consumo?, ¿Y el tipo de motor? (Utiliza un nivel de significación 0.05).

**Ejercicio 170** El beneficio obtenido (en millones de pesetas) por cinco supermercados en cinco años viene dado en la siguiente tabla

	Super. 1	Super. 2	Super. 3	Super. 4	Super. 5
1990	222	196	204	305	128
1991	220	235	190	351	109
1992	170	188	182	351	112
1993	175	199	190	348	139
1994	155	108	104	205	70

1. Hacer la tabla de Análisis de Varianza usando el año como bloque.
2. ¿Hay evidencia suficiente para concluir que hay algún supermercado con beneficio significativamente mayor que los demás? ¿Y menor?

**Ejercicio 171** Los siguientes datos son los tiempos empleados por tres trabajadores usando tres tipos de maquinaria diferente en dos días

	Trabajador 1	Trabajador 2	Trabajador 3
Maquinaria A	37, 43	38, 44	38, 40
Maquinaria B	31, 36	40, 44	43, 41
Maquinaria C	36,40	33, 37	41, 39

Usando un modelo de dos factores (maquinaria y trabajador) con interacción

1. Hacer la tabla de Análisis de la varianza

2. ¿Algun tipo de maquinaria es más rápida que las demás? ¿Algun trabajador es más rápido?
3. ¿Hay interacción entre los factores?

**Ejercicio 172** Se desea adquirir un nuevo equipo de maquinas para una fábrica. Para comparar las velocidades de la nueva maquinaria con la antigua. se encarga un cierto trabajo a 5 empleados que ya lo han realizado con el equipo antiguo. La tabla siguiente resume los tiempos en minutos.

Empleado	1	2	3	4	5
Equipo nuevo	115	205	147	121	186
Equipo antiguo	124	212	151	132	195

¿A qué conclusión puede llegarse

Usa el modelo de dos factores sin interacción (El factor bloque es el empleado) Tomese para alfa el valor 0.05

**Ejercicio 173** Se desea comparar el funcionamiento de cuatro dispositivos eléctricos A, B, C, D bajo distintos niveles de tensión  $T_1, T_2, T_3$ . Los niveles de eficiencia son:

	$T_1$	$T_2$	$T_3$
A	4	3	9
B	7	9	10
C	2	7	8
D	8	8	4

Considerando la tensión como bloque (no hay interacción entre dispositivo y tensión), se pide:

1. Construir la tabla de análisis de la varianza
2. Decir si hay diferencia significativa entre las eficiencias de estos dispositivos
3. ¿Esta justificado considerar el factor tensión como bloque?

**Ejercicio 174** Tres agentes inmobiliarios fueron interrogados acerca del precio de cinco viviendas de un barrio. Las valoraciones de estas viviendas segun los agentes se dan a continuación:

	Agente A	Agente B	Agente C
vivienda 1	210	218	226
vivienda 2	192	190	198
vivienda 3	183	187	185
vivienda 4	227	223	237
vivienda 5	242	240	237

342TEMA 14. ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES

1. Construir la tabla de análisis de la varianza
2. Contrastar la igualdad entre las valoraciones medias de las viviendas.
3. Intervalo de confianza para el valor medio de la vivienda 5.

**Ejercicio 175** Un ingeniero que ha de diseñar una batería ha probado su duración para distintos materiales y soportando distintas temperaturas. La duración en horas viene dada en la tabla siguiente:

Tipo de Material	Temperatura en grados Farenheit		
	15°	70°	125°
M1	150, 188	136, 122	25, 70
	159, 126	106, 115	58, 45
M2	130, 74	34, 80	20, 82
	155, 80	40, 75	70, 58
M3	138, 110	174, 120	96, 104
	168, 160	150, 139	82, 60

1. Hallar la suma de total de cuadrados, la suma de cuadrados correspondientes al factor temperatura, al factor material y a la interacción de ambos factores.
2. Indica si son significativos los efectos de la temperatura, del material y de la interacción entre ambos factores sobre la duración de las baterías.

**Unidad Temática V**

**ANÁLISIS**  
**MULTIVARIANTE**





## Tema 15

# Análisis multivariante. Regresión

### 15.1 Generalidades

El análisis multivariante es el conjunto de técnicas estadísticas que permiten analizar datos de elementos que poseen diversas características. Los datos de cada una de estas características de cada elemento vienen registrados en una variable. Por tanto las variables tratadas en el análisis multivariante son variables aleatorias multidimensionales.

Los estudios económicos, así como los relacionados con la investigación de mercados y con otros muchos aspectos del ámbito empresarial, requieren una considerable cantidad de variables. Los estudios citados son de interés para un ingeniero, y el Análisis Multivariante es una herramienta útil, por no decir imprescindible, para el tratamiento de este tipo de datos multidimensionales. La información que suministran las técnicas multivariantes sobre los datos es mucho más completo que el que se hace en el caso univariante, y más cercano a la realidad. Pero hay que pagar un precio: el tratamiento matemático es mucho más complejo. Por este motivo la difusión del Análisis Multivariante es paralela a la de los ordenadores.

Uno de los primeros trabajos dentro de este campo es el de Karl Pearson (1901). En él se establecen las primeras ideas sobre el *Análisis de Componentes Principales*. Pero el desarrollo de la mayoría de los modelos multivariantes se produce en los años treinta con las aportaciones de Hotelling, Wilks, Fisher, Mahalanobis y Bartlett.

En un principio los trabajos eran principalmente teóricos. Las aplicaciones comenzaron más tarde (Rao, 1952). En los años sesenta comienzan a utilizarse los ordenadores en el análisis de datos. Entonces deja de tener importancia la complejidad de los cálculos y el Análisis Multivariante em-

pieza a aplicarse a la Psicología, Educación, Biología, Medicina, Economía, etc. En nuestro país, salvo escasas excepciones, los ordenadores no empiezan a emplearse en investigación hasta los años setenta. Pero es en los ochenta cuando empieza su gran difusión, y con ello también se extiende el uso de las técnicas Multivariantes.

**Economía:** Los datos pueden ser indicadores de la actividad económica en varios países por ejemplo: renta per cápita, déficit del Estado, desempleo, tipos de interés, etc. *El Análisis Factorial* puede usarse para reducir esta información en un menor número de variables.

**Biología:** Los seres vivos se distinguen unos de otros por diversas características. Las especies agrupan seres vivos de características similares. Esta clasificación se realiza a veces con ayuda de una técnica de análisis multivariante: *El análisis Cluster o de Conglomerados*.

Existen diversas técnicas de Análisis de datos multivariante. Se usarán unas u otras dependiendo del objetivo del estudio que se vaya a realizar. Estas técnicas se suelen clasificar en tres grandes grupos: métodos de dependencia, métodos de interdependencia y métodos estructurales.

En este tema detallaremos el modelo de *Regresión Lineal Múltiple*.

## 15.2 Regresión Múltiple

### 15.2.1 Introducción

Los modelos de regresión múltiple tratan de cuantificar la influencia que ejercen variables explicativas de distinto tipo sobre una o varias variables dependientes de ella. Esta relación se suele usar con fines predictivos. Ejemplos:

Predecir la puntuación de un alumno (variable dependiente) en las pruebas de selectividad basándose en el promedio de las notas de BUP y COU, el coeficiente intelectual, el autoconcepto y el dominio de las técnicas de estudio.

Relación entre la presión sanguínea (variable dependiente) y la edad, el peso, la talla y el nivel de actividad física.

Predecir los gastos anuales en vestido en función de la edad, el sueldo anual y el número de personas con las que se convive.

Número de victorias de un equipo de fútbol en una temporada en función de los puntos ganados, los goles a favor, los goles en contra, partidos ganados la temporada anterior, el número de jugadores extranjeros, etc.

El modelo de regresión lineal simple (ver párrafo 1.4.4 de la página 45) consistía en obtener una recta que permitiera obtener valores de una variable

dependiente a partir de los valores de otra variable dependiente o variable predictora. Esta recta se obtenía a partir de una muestra bidimensional, imponiendo la condición de mínimos cuadrados. En el caso de que haya más de una variable predictora hablamos de regresión múltiple, como ocurre en el siguiente ejemplo: Sea  $Y$  la variable que toma como valores los precios actuales de una serie de casas. La consideramos como variable dependiente, ya que puede pensarse que estos precios guardan relación con otras variables como, por ejemplo  $X_1 =$  metros cuadrados de la casa,  $X_2 =$  índice de localización,  $X_3 =$  precio de la casa en el año anterior,  $X_4 =$  índice de calidad. Estas cuatro variables serán las variables independientes o predictoras, en cambio  $Y$ , precio actual de la casa es la variable dependiente.

### 15.2.2 Modelo de Regresión lineal

El modelo de regresión clásico establece que el valor medio de  $Y$  correspondiente a unos ciertos valores de  $X$  depende linealmente de estos valores de  $X$  :

$$E(Y/X_1 = x_1, X_2 = x_2, \dots, X_r = x_r) = b_0 + b_1x_1 + b_2x_2 + \dots + b_rx_r$$

Los valores de  $X$  ( $x_1, x_2, \dots, x_r$ ) son considerados como fijos. Para cada elemento de la muestra el valor de la variable  $Y$  (indicamos los valores de esta variable por  $y$ ), puede expresarse en la forma

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_rx_r + \varepsilon$$

donde  $\varepsilon$  representa una cantidad aleatoria que va recoger la influencia de los factores no incluidos en  $X$ . El modelo de regresión lineal que consideramos es el siguiente:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_rx_r + \varepsilon$$

donde se supone que  $\varepsilon$ , llamado residuo o error es una variable normal de valor medio 0, y varianza  $\sigma^2$ , que es constante independientemente del valor de  $X$  considerado. La covarianza de los valores de  $\varepsilon$  para cada par de puntos es nula.

Si tenemos  $n$  valores muestrales para  $(X, Y)$  se tiene:

$$y_1 = b_0 + b_1x_{11} + b_2x_{12} + \dots + b_rx_{1r} + \varepsilon_1$$

$$y_2 = b_0 + b_1x_{21} + b_2x_{22} + \dots + b_rx_{2r} + \varepsilon_2$$

...

...

$$y_n = b_0 + b_1x_{n1} + b_2x_{n2} + \dots + b_rx_{nr} + \varepsilon_n$$

con las condiciones

$$E(\varepsilon_i) = 0, E(\varepsilon_i, \varepsilon_i) = \sigma^2, E(\varepsilon_i, \varepsilon_j) = 0 \quad (15.1)$$

En forma matricial :

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1r} \\ 1 & x_{21} & x_{22} & \dots & x_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} \dots & \dots & x_{nr} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \text{ó} \quad \mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

$$\text{con } E(\boldsymbol{\varepsilon}) = \mathbf{0}, E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$$

**Ejemplo 74** Si los valores muestrales son los de la tabla siguiente

x1	1	3	2	3	4
x2	0	1	3	8	9
y	1	4	3	8	9

Obtenemos:

$$\begin{pmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 8 \\ 1 & 4 & 9 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

### 15.3 Estimación mínimo cuadrática

Consiste en seleccionar valor para  $\mathbf{b}$  de modo que

$$\sum_{j=1}^n \widehat{\varepsilon}_j^2 = \sum_{j=1}^n (y_j - b_0 - b_1 x_{j1} - b_2 x_{j2} - \dots - b_r x_{jr})^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b})$$

sea mínimo. Los coeficientes obtenidos usando este criterio son estimaciones de  $\mathbf{b}$   $(\widehat{b}_0, \widehat{b}_1, \widehat{b}_2, \dots, \widehat{b}_r)$

$$\widehat{y}_j = \widehat{b}_0 + \widehat{b}_1 x_{j1} + \widehat{b}_2 x_{j2} + \dots + \widehat{b}_r x_{jr}$$

es una estimación de la media de  $Y$  en el valor de  $X$  correspondiente, y

$$\widehat{\varepsilon}_j = y_j - (\widehat{b}_0 * + \widehat{b}_1 * x_{j1} + \widehat{b}_2 * x_{j2} + \dots + \widehat{b}_r * x_{jr})$$

son estimaciones para los residuos. Estos estimadores cumplen el siguiente teorema:

**Teorema 7** Si rango de la matriz de las  $x$  es  $r + 1$  (menor que el número de datos  $n$ ) entonces:

1. El estimador de mínimos cuadrados de la matriz columna de los coeficientes puede obtenerse de la expresión

$$\hat{\mathbf{b}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

**Teorema 8** 2. Puede comprobarse que se cumple, al igual que en la regresión simple, que:

$$\sum_{j=1}^n (y_j - \bar{y})^2 = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n (y_j - \hat{y}_j)^2 = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 \quad (15.2)$$

**Ejemplo 75** El plano de regresión del ejemplo anterior se hallaría

$$\hat{\mathbf{b}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

$$(\mathbf{x}'\mathbf{x}) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 3 & 4 \\ 0 & 1 & 3 & 8 & 9 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 8 \\ 1 & 4 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 13 & 21 \\ 13 & 39 & 69 \\ 21 & 69 & 155 \end{pmatrix}$$

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} \frac{321}{175} & -\frac{283}{350} & \frac{39}{350} \\ -\frac{283}{350} & \frac{167}{350} & -\frac{18}{175} \\ \frac{39}{350} & -\frac{18}{175} & \frac{13}{350} \end{pmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \begin{pmatrix} \frac{321}{175} & -\frac{283}{350} & \frac{39}{350} \\ -\frac{283}{350} & \frac{167}{350} & -\frac{18}{175} \\ \frac{39}{350} & -\frac{18}{175} & \frac{13}{350} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 3 & 4 \\ 0 & 1 & 3 & 8 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{pmatrix} =$$

$$\begin{pmatrix} -0.414286 \\ 1.22857 \\ 0.528571 \end{pmatrix}$$

Así que el plano de regresión es en este caso

$$y = \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix} \begin{pmatrix} -0.414286 \\ 1.22857 \\ 0.528571 \end{pmatrix} =$$

$$= -0.414286 + 1.22857x_1 + 0.528571x_2$$

En la tabla siguiente se indican los valores predichos y los residuos para cada valor de  $x_1, x_2$

$x_1$	1	3	2	3	4
$x_2$	0	1	3	8	9
$y$	1	4	3	8	9
$\hat{y}$	0.81426	3.8	3.62857	7.5	9.25
$\hat{\varepsilon}$	0.1857	0.2	-0.6285	0.5	-0.25714

El coeficiente de determinación es suma de cuadrados explicados por la regresión dividida por la suma total de cuadrados.

$$R^2 = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2} = \frac{45.2143}{46.000} = 0.98292$$

e indica la proporción de variación explicada por la regresión. De la relación dada en 15.2, se deduce que el valor de  $R^2$  está comprendido entre cero y uno. El ajuste es lineal es tanto mejor cuanto más cercano a 1 sea el valor de  $R^2$ .

## 15.4 Test de hipótesis para la regresión lineal múltiple

La prueba de hipótesis que detallamos a continuación requiere que los términos de error sigan distribuciones normales y cumplan las condiciones 15.1 del modelo descrito en el párrafo 15.2.2. Es un test de Hipótesis para determinar si existe relación lineal entre la variable respuesta o dependiente y las variables independientes o predictoras.

La hipótesis nula es:  $b_1 = b_2 = \dots = b_r = 0$

La hipótesis alternativa es algún  $b_j \neq 0$ .

Para realizar esta prueba se usa el estadístico  $F$  con  $r$ ,  $n - r - 1$  grados de libertad. El rechazo de la hipótesis nula se interpreta como que alguna de las variables es significativa en el modelo, es decir, que nos da alguna información sobre la variable dependiente.

Para realizar esta prueba se usa el estadístico

$$F = \frac{\text{media de cuadrados explicados por la regresión}}{\text{media de cuadrados no explicados por la regresión}} = \frac{\frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{r}}{\frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n - r - 1}}$$

que sigue, si se cumple la hipótesis nula una distribución  $F$  con  $r$ ,  $n - r - 1$  grados de libertad. Si el valor experimental supera el valor de la  $F_{r, n-r-1}$

teórica, cuya función de distribución toma el valor  $\alpha$ , se rechaza la hipótesis nula de que los coeficientes  $b_1, b_2, \dots, b_r$  son ceros, y por tanto se concluye que al menos una de las variables independientes contribuye a la explicación de la variable dependiente.

**Ejemplo 76** *Se efectua un estudio sobre el desgaste de unos cojinetes ( $y$ ) y su relación con la viscosidad del aceite ( $x_1$ ) y la carga ( $x_2$ ), obteniéndose los siguientes datos*

$x_1$	1.6	15.5	22	43	33	40
$x_2$	851	816	1058	1201	1357	1115
$y$	193	230	172	91	113	125

Las estimaciones mínimo cuadráticas de los coeficientes de regresión se obtienen de la expresión

$$\begin{aligned} \hat{\mathbf{b}} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\ (\mathbf{x}'\mathbf{x}) &= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1.6 & 15.5 & 22 & 43 & 33 & 40 \\ 851 & 816 & 1058 & 1201 & 1357 & 1115 \end{pmatrix} \begin{pmatrix} 1 & 1.6 & 851 \\ 1 & 15.5 & 816 \\ 1 & 22 & 1058 \\ 1 & 43 & 1201 \\ 1 & 33 & 1357 \\ 1 & 40 & 1115 \end{pmatrix} \\ &= \begin{pmatrix} 6.0 & 155.1 & 6398.0 \\ 155.1 & 5264.81 & 1.7831 \times 10^5 \\ 6398.0 & 1.7831 \times 10^5 & 7.0365 \times 10^6 \end{pmatrix}, \\ (\mathbf{x}'\mathbf{x})^{-1} &= \begin{pmatrix} 8.59535 & 8.09653 \times 10^{-2} & -9.86711 \times 10^{-3} \\ 8.09653 \times 10^{-2} & 2.1026 \times 10^{-3} & -1.269 \times 10^{-4} \\ -9.86711 \times 10^{-3} & -1.269 \times 10^{-4} & 1.23296 \times 10^{-5} \end{pmatrix} \\ \hat{\mathbf{b}} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \\ &= \begin{pmatrix} 8.59535 & 8.09653 \times 10^{-2} & -9.86711 \times 10^{-3} \\ 8.09653 \times 10^{-2} & 2.1026 \times 10^{-3} & -1.269 \times 10^{-4} \\ -9.86711 \times 10^{-3} & -1.269 \times 10^{-4} & 1.23296 \times 10^{-5} \end{pmatrix} \begin{pmatrix} 193 \\ 230 \\ 172 \\ 91 \\ 113 \\ 125 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1.6 & 15.5 & 22 & 43 & 33 & 40 \\ 851 & 816 & 1058 & 1201 & 1357 & 1115 \end{pmatrix} \begin{pmatrix} 193 \\ 230 \\ 172 \\ 91 \\ 113 \\ 125 \end{pmatrix} = \begin{pmatrix} 350.995 \\ -1.27217 \\ -0.153908 \end{pmatrix} \end{aligned}$$

Así que el plano de regresión es en este caso

$$y = \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix} \begin{pmatrix} 350.995 \\ -1.27217 \\ -0.153908 \end{pmatrix} = 350.995 - 1.27217x_1 - 0.153908x_2$$

Realizamos ahora el test de hipótesis de significación de la regresión. Para ello evaluamos para cada punto los valores predichos por la regresión

$x_1$	1.6	15.5	22	43	33	40
$x_2$	851	816	1058	1201	1357	1115
$y$	193	230	172	91	113	125
$\hat{y}$	217.987	205.692	160.18	111.46	100.171	128.511

Media de  $y = 1233.3$

$$F = \frac{\frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{n - r - 1}}{\frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n - r - 1}};$$

$$\sum_{j=1}^6 (\hat{y}_j - \bar{y})^2 = (217.987 - 154)^2 + \dots + (128.511 - 154)^2 = 12161.5$$

$$\sum_{j=1}^6 (y_j - \hat{y}_j)^2 = (217.987 - 193)^2 + \dots + (128.511 - 125)^2 = 1950.5$$

Por tanto

$$F = \frac{\frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{n - r - 1}}{\frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n - r - 1}} = \frac{\frac{12161.5}{2}}{\frac{1950.5}{3}} = 9.3526$$

Como la  $F_{2,3}^{-1}(0.95) = 9.55209$  que es mayor que la experimental no se rechaza la hipótesis nula al nivel 0.05, y por tanto se concluye que no hay influencia de la carga ni de la viscosidad en el desgaste de los cojinetes a este nivel. Sin embargo al nivel de confianza del 90% si se admitiría la existencia de relación lineal, ya que  $F_{2,3}(0.10) = 5.46233$ , siendo en este caso el valor experimental mayor que el teórico por lo que al nivel de significación 0.10 se rechazaría la hipótesis nula, concluyéndose por tanto que existe alguna dependencia lineal entre la  $y$  y las variables  $x_1$  y  $x_2$ .

El coeficiente de determinación es:

$R^2 = \frac{12161.5}{12161.5 + 1950.5} = 0.861784$  que indica la proporción de la variación total que esta explicada por la regresión.

## 15.5 Intervalos de confianza para los coeficientes

Se basan en el estadístico

$$\frac{\hat{b}_j - b_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad (15.3)$$



siendo  $\hat{b}_j$  el valor estimado para este coeficiente a partir de la muestra  $\hat{\sigma}^2$  una estimación de la varianza de los residuos

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^n (y_j - \hat{y})^2}{n - r - 1}$$

que coincide con el denominador de  $F$  y  $C_{jj}$  es el término que ocupa el lugar  $j + 1$  en la diagonal de la matriz  $(\mathbf{x}'\mathbf{x})^{-1}$ . Este estadístico, 15.3, se distribuye como una  $t$  de Student con  $n - r - 1$  grados de libertad. Por tanto el intervalo de confianza para  $b_j$  será:

$$\left( \hat{b}_j - t_{n-r-1}\left(\frac{\alpha}{2}\right) \sqrt{\hat{\sigma}^2 C_{jj}} < b_j < \hat{b}_j + t_{n-r-1}\left(\frac{\alpha}{2}\right) \sqrt{\hat{\sigma}^2 C_{jj}} \right)$$

**Ejemplo 77** Hallar los intervalos de confianza para los coeficientes del ajuste del ejemplo 75

$$\begin{aligned} \hat{b}_0 &= -0.414289, & t_{n-r-1}\left(\frac{\alpha}{2}\right) &= t_{5-2-1}\left(\frac{0.05}{2}\right) = \text{TInv}(0.975; 2) = 4.30265, \\ \hat{\sigma}^2 &= \frac{\sum_{j=1}^n (y_j - \hat{y})^2}{n - r - 1} = \frac{0.7857}{2} = 0.39285, & C_{00} &= \frac{321}{175} = 1.83429 \end{aligned}$$

Por lo tanto el intervalo de confianza para el término independiente es:  
 $-0.414289 \pm 4.30265 \times \sqrt{0.39285 \times 1.83429}$ .

Es decir:

$$(-4.06673 < b_0 < 3.23815)$$

$$\begin{aligned} \hat{b}_1 &= 1.22857, & t_{n-r-1}\left(\frac{\alpha}{2}\right) &= t_{5-2-1}\left(\frac{0.05}{2}\right) = \text{TInv}(0.975; 2) = 4.30265, \\ \hat{\sigma}^2 &= \frac{\sum_{j=1}^n (y_j - \hat{y})^2}{n - r - 1} = \frac{0.7857}{2} = 0.39285, & C_{11} &= \frac{167}{350} = 0.477143 \end{aligned}$$

Por lo tanto el intervalo de confianza para el coeficiente de  $x_1$  es:

$$1.22857 \pm 4.30265 \times \sqrt{0.39285 \times 0.477143} = 1.22857 \pm 4.30265 \times 0.43295$$

Por tanto

$$(-0.634262 < b_1 < 3.0914)$$

El intervalo de confianza para el otro coeficiente  $b_2$ , cuyo valor estimado era 0.528571 resulta

$$(0.00883, 1.04832)$$

## 15.6 Predicción. Intervalos de confianza

Los modelos de regresión lineal se formulan muy a menudo con el propósito de estimar o predecir el valor de la variable dependiente  $y$  para determinados valores de las variables independientes  $x_i$ . El valor estimado o pronosticado

para  $y$  correspondiente a los valores  $x_{j1}, x_{j2}, \dots, x_{jr}$  de las variables independientes es:

$$\hat{y}_j = \hat{b}_0 + \hat{b}_1 x_{j1} + \hat{b}_2 x_{j2} + \dots + \hat{b}_r x_{jr}$$

Este estimador se emplea tanto para estimar la media de los valores de  $y$  que correspondan a los valores  $x_{j1}, x_{j2}, \dots, x_{jr}$  de las variables independientes, como para predecir uno cualquiera de estos valores individuales de  $y$ . La diferencia entre ambos enfoques es la amplitud del correspondiente intervalo de confianza.

Si lo que se pretende es estimar la media,  $\hat{E}(Y/X_1 = x_{j1}, X_2 = x_{j2}, \dots, X_r = x_{jr})$ , entonces el intervalo de confianza, con significación  $\alpha$ , viene dado por:

$$\left( \hat{y}_j - t_{n-r-1}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{x_j(x'x)^{-1}x'_j}, \left( \hat{y}_j + t_{n-r-1}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{x_j(x'x)^{-1}x'_j} \right) \right)$$

En cambio, la amplitud del intervalo de confianza para la predicción puntual es mayor:

$$\left( \hat{y}_j - t_{n-r-1}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{1 + x_j(x'x)^{-1}x'_j}, \hat{y}_j + t_{n-r-1}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{1 + x_j(x'x)^{-1}x'_j} \right)$$

Es conveniente resaltar que no es conveniente realizar predicciones fuera del rango de los datos, pues los intervalos de confianza dados se van agrandando conforme el punto  $x_{j1}, x_{j2}, \dots, x_{jr}$  se aleja del centro de gravedad de la nube de puntos que forman los datos en el espacio de las variables independientes.

## 15.7 Modelos de regresión polinomial

El modelo de regresión lineal puede usarse en los modelos de regresión que sean lineales en los parámetros  $b_i$

Esto incluye a los modelos polinomiales, como por ejemplo

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 + \varepsilon$$

**Ejemplo 78** Ajustar un polinomio de segundo grado  $y = b_0 + b_1 x + b_2 x^2$  a los datos siguientes

$y$	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18
$x$	20	25	30	35	40	50	60	65	70	75	80	90

Previamente construimos la tabla siguiente, que incluye, como una nueva variable  $x_2$ , los cuadrados de  $x$

$x_1 = x$	20	25	30	35	40	50	60	65	70	75	80	90
$x_2 = x^2$	400	625	900	1225	1600	2500	3600	4225	4900	5625	6400	8100
$y$	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18

al que ajustamos el modelo  $y = b_0 + b_1x_1 + b_2x_2$ .resultando

$$y = 2.19827 - 0.02252x_1 + 0.00012x_2.$$

Por lo tanto el polinomio de ajuste resulta

$$y = 2.19827 - 0.02252x + 0.00012x^2.$$

## 15.8 Regresión paso a paso.

En el principio de una investigación podemos tener una gran cantidad de variables. A veces sospechamos que puedan estar relacionadas entre sí. Quizá queremos considerar una de ellas como dependiente de las restantes. Una forma de actuar consistiría en estudiar todas las regresiones posibles y analizar entre todas ellas cual es el modelo más adecuado, teniendo en cuenta además que, por motivos de economía, conviene que el número de variables elegidas no sea demasiado grande. Esta opción, el estudio de todas las regresiones posibles, resulta demasiado laboriosa incluso para un número no demasiado amplio de variables. Por ejemplo para ocho variables deberíamos de comparar  $2^8 = 256$  modelos y para doce variables  $2^{12} = 4096$  modelos. Por ello se han propuesto distintos algoritmos para aligerar esta selección. Uno de ellos es el de *Regresión Paso a Paso* o *Regresión por etapas* (Stepwise Regression), que es una técnica para elegir las variables más adecuadas para predecir la variable dependiente. Se emplean dos modelos. El modelo de selección hacia delante (Forward stepwise regression) y el modelo de regresión hacia detrás (Backward stepwise regression). Ambos construyen una sucesión de modelos de regresión mediante la incorporación o eliminación de una variable en cada paso del algoritmo.

El modelo de selección hacia delante comienza seleccionando para tomar parte del modelo de regresión la variable más explicativa (la que tenga el estadístico F más alto o el menor p-value). Supongamos que las posible variables independientes o explicativas son  $\{x_1, x_2, x_3, x_4\}$  y la dependiente es la  $y$ . Para seleccionar la primera variable se comparan los valores de F de los modelos de regresión de  $y$  con respecto a cada una de las variables dependientes:

$$\{y = b_{10} + b_{11}x_1, y = b_{20} + b_{21}x_2, y = b_{30} + b_{31}x_3, y = b_{40} + b_{41}x_4, \}.$$

La variable de entrada es la que tenga un valor de  $F$  más alto. Supongamos que fuera  $x_1$ . Ahora comprobamos si el valor de  $F$  es mayor que uno seleccionado de antemano, que indicaremos como  $F$  de entrada ( $F_{en}$ ). Si es así aceptamos  $x_1$ , si no es así no admitimos ninguna variable en la regresión.

En cada nuevo paso, se intenta incluir una nueva variable, aquella cuya  $F$  parcial, condicionada a que las variables seleccionadas en el paso anterior estén en el modelo, sea más grande, siempre que sea mayor que ( $F_{en}$ ). Si una nueva variable ha entrado en el modelo se reconsidera el mantener las que ya se había añadido previamente, es decir que no sólo puede entrar una nueva variable en cada paso sino que puede salir alguna de las que se habían incluido en el paso previo. Si todos los valores de las  $F$  parciales de cada variable en el modelo son mayores que un valor seleccionado también de antemano que llamamos  $F$  de salida ( $F_{sal}$ ) retenemos todas las variables. En caso contrario se remueve la variable inspeccionada del modelo.

El proceso termina cuando ninguna variable que no esté seleccionada pueda entrar porque las  $F$  sean menores que  $F_{en}$ , ni ninguna de las seleccionadas pueda salir del modelo porque sus valores de  $F$  parciales sean mayores que  $F_{sal}$ . Por lo general se toma  $F_{en} = F_{sal}$ .

El conjunto de variables que finalmente quede incluido en la ecuación de regresión puede depender del camino seguido a la hora de seleccionarlas, salvo en el caso de que se evalúen todos los modelos de regresión posibles que obviamente sólo tiene una conclusión.

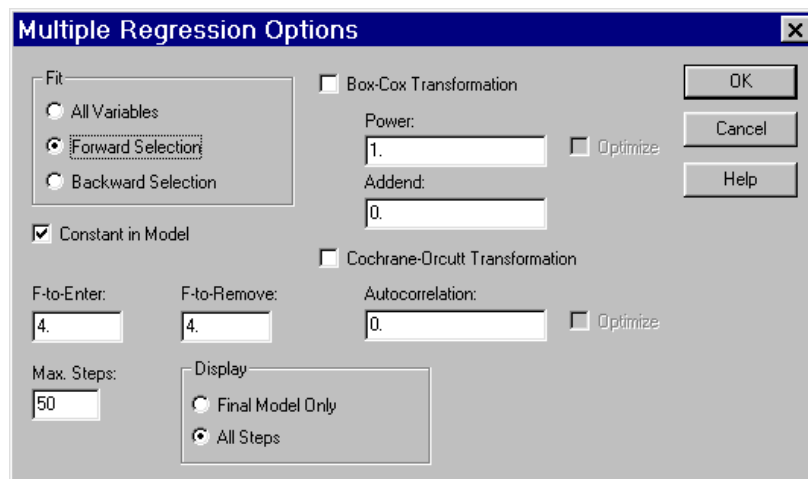
Las formulas necesarias para calcular las  $F$  parciales pueden encontrarse en el texto *Probabilidad y estadística aplicadas a la Ingeniería* de Douglas C. Montgomery y George C. Runger editado por McGraw-Hill.

El modelo de regresión hacia atrás comienza incluyendo en el modelo todas las variables y va extrayéndolas de una en una por un procedimiento similar al anterior, acabando como antes cuando se cumplen las condiciones de terminación (ninguna variable pueda entrar porque las  $F$  sean menores que  $F_{en}$ , ni salir porque sean mayores que  $F_{sal}$ ).

Afortunadamente esta técnica se aplica usando programas de ordenador. Indicamos aquí la aplicación de esta técnica con Statgraphics Plus 5.0 al fichero *obesity.sf* del directorio de datos de este programa.

La variable dependiente va ser *U1\_volume* y las independientes todas las que empiezan por  $x$ .

Realizando la siguiente selección



el programa nos selecciona, usando el procedimiento de regresión paso a paso hacia adelante las variables  $x2\_coeff$  y  $x7\_creatnn$ . El modelo de regresión más adecuado resulta ser :

$$U1\_volume = 535.473 - 43.7279 \times (x2\_coeff) - 56.7447 \times x7\_creatnn$$

Los detalles más importantes del desarrollo del método vienen resumidos a continuación

Method: forward selection

F-to-enter: 4.0

F-to-remove: 4.0

Step 0:

---

0 variables in the model. 44 d.f. for error.

R-squared = 0.00%      Adjusted R-squared = 0.00%      MSE = 11729.3

Step 1:

---

Adding variable  $x7\_creatnn$  with F-to-enter = 46.5084

1 variables in the model. 43 d.f. for error.

R-squared = 51.96%      Adjusted R-squared = 50.84%      MSE = 5765.82

Step 2:

---

Adding variable  $x2\_coeff$  with F-to-enter = 17.6021

2 variables in the model. 42 d.f. for error.

R-squared = 66.15%    Adjusted R-squared = 64.54%    MSE = 4159.76  
 Final model selected.

## 15.9 Estudio de un caso práctico

En un laboratorio de Metrología se ha realizado diversas mediciones que han arrojado los siguientes datos. Se esperaba que una expresión del tipo

$R = C e^{\sum_{i=1}^2 \sum_{j=1}^2 k_{ij} a^i v^{-0.27j}}$  fuera aceptable para ajustar estos datos.

v	a	R
43	0.05	17
43	0.1	17
43	0.2	21.2
43	0.3	28.3
64	0.05	16.3
64	0.1	17
64	0.2	17.8
64	0.3	26.3
85	0.05	18.5
85	0.1	17.8
85	0.2	18.5
85	0.3	21.5
127	0.05	18.3
127	0.1	15
127	0.2	18
127	0.3	17
170	0.05	15
170	0.1	16.8
170	0.2	16.5
170	0.3	17.5

Con objeto de linealizar la expresión aplicamos logaritmos:

$$\ln R = \ln C + \sum_{i=1}^2 \sum_{j=1}^2 k_{ij} a^i v^{-0.27j} = \ln C + k_{11} a v^{-0.27} + k_{12} a v^{-0.54} + k_{21} a^2 v^{-0.27} + k_{22} a^2 v^{-0.54}$$

Utilizando el Paquete Statgraphic Plus 5.1 se han obtenido los resultados siguientes para un ajuste de este tipo:

El coeficiente de determinación del ajuste es del 85.73%. Las pruebas de hipótesis se han realizado al 95%: La prueba  $F$  de Análisis de Varianza

de la Regresión reconoce que hay relación entre las variable dependientes  $R$  y las independientes. No obstante la prueba  $t$  de Student no reconoce la significación de algunos de los coeficientes de este ajuste. Por este motivo se ha empleado el procedimiento de Regresión Paso a Paso, que permite eliminar del modelo las variables no significativas con una pérdida mínima de la calidad del ajuste.

El procedimiento de Regresión Paso a Paso hacia adelante (Forward) con un valor de  $F = 4$  para aceptar o rechazar una variable nos suministra el siguiente modelo:

$$\ln R = 2.87 - 10.49av^{-0.27} + 23.33av^{-0.54} + 62.05a^2v^{-0.54}$$

El coeficiente de determinación resultante es 85.16 %. La desviación estándar de los residuales es 0.067. y su error absoluto medio es 0.05. El estadístico de Durbin Watson y el primer coeficiente de autocorrelación dan valores consistentes con la hipótesis de independencia de los residuos. También los valores del estadístico  $t$  de Student permiten aceptar que los valores de los coeficientes del ajuste son significativos.

El procedimiento de Regresión Paso a Paso hacia detrás (Backward) con un valor de  $F = 4$  para aceptar o rechazar una variable nos suministra el siguiente modelo:

$$\ln R = 2.80 - 33.99a^2v^{-0.27} + 142.74a^2v^{-0.54}$$

El coeficiente de determinación del ajuste es 85.04 %. La desviación estándar de los residuales es 0.067. y su error absoluto medio es 0.05. El estadístico de Durbin Watson y el primer coeficiente de autocorrelación dan valores consistentes con la hipótesis de independencia de los residuos. También ahora los valores del estadístico  $t$  permiten aceptar que los valores de los coeficientes del ajuste son significativos.

*Conclusión:* Cualquiera de estos dos últimos ajustes es aceptable (95% de confianza).

La diferencia en el coeficiente de determinación así como en los errores de los residuales es tan pequeña que no hay, desde el punto de vista de estos parámetros, razones para decantarse por uno u otro. En estos casos se suele seleccionar el que contenga un menor número de variables regresoras. Actuando de esta forma seleccionaríamos el último modelo analizado. El ajuste seleccionado tomaría la forma:

$$R = e^{2.80 - 33.99a^2v^{-0.27} + 142.74a^2v^{-0.54}} = 16.45e^{-33.99a^2v^{-0.27} + 142.74a^2v^{-0.54}}$$

### 15.10 EJERCICIOS PROPUESTOS

**Ejercicio 176** *Los tabla siguiente indica la edad, los años de experiencia y los ingresos mensuales (en miles de pesetas) de 5 ingenieros.*

Edad	37	45	38	42	31
Experiencia	4	0	5	2	4
Ingresos	512	468	550	503	454

*El modelo de regresión lineal que relacione los ingresos con las otras dos variables es:*

$$\text{ingresos} = 37.21 + 9.61 \text{ edad} + 29.76 \text{ experiencia}$$

1. *Calcular el coeficiente de determinación*
2. *¿Es la regresión significativa?*
3. *Emplea este ajuste para predecir cuanto ganan por promedio los ingenieros de 40 años de edad y 4 de experiencia*

**Ejercicio 177** *En un estudio de consumo se estimo la siguiente ecuación, obtenida con 200 datos:*

$\log(y) = -0.243 - 0.562 \log x_1 + 0.327 \log x_2 + 0.219 \log x_3 - 0.127 \log x_4$ ,  
*siendo el coeficiente de determinación del ajuste  $R^2 = 0.853$ , y 0.219, 0.161, 0.157, 0.082 los errores estandar correspondientes a los coeficientes del ajuste.*

*y = Cantidad de carne de cerdo comprada*

*$x_1 =$  Precio de la carne de cerdo,  $x_2 =$  Precio de la carne de ternera,  $x_3 =$  Precio de la carne de pollo,  $x_4 =$  ingreso medio por familia.*

1. *Interpretar los coeficientes del modelo de regresión*
2. *Indicar que variables son significativas.*

**Ejercicio 178** *Para 11 provincias españolas se conocen los siguientes datos:*

*$Y =$  número de mujeres conductoras dividido por el número de hombres conductores*

*$X_1 =$  Porcentaje de mujeres trabajadoras sobre el total de trabajadores de la provincia*

*$X_2 =$  Porcentaje de población que trabaja en el sector agrícola.*

*para obtener el modelo de regresión lineal donde la primera variable es dependiente y las dos restantes independientes, se ha obtenido:*

$$(X'X)^{-1} = \begin{pmatrix} 5.1 & -0.12 & -0.05 \\ -0.12 & 30.8 & 0.08 \\ -0.05 & 0.08 & 0.001 \end{pmatrix}, \quad (X'Y) = \begin{pmatrix} -0.06 \\ 0.05 \\ -9.45 \end{pmatrix}$$

$$\widehat{\sigma^2} = 0.003; \quad \sum (y - \bar{y})^2 = 0.0645$$



1. Estimar el modelo de regresión y los contrastes individuales para los coeficientes.
2. Calcular el coeficiente de determinación.

1. **Ejercicio 179** Con los datos de los 12 meses del año 1973 de la encuesta de presupuestos familiares se han probado seis distintos modelos de regresión lineal (sin constante) en los que la variable dependiente es *GTINE* (Gasto Total según el INE) y las variables explicativas son las siguientes:

*IT* = Ingreso Total

*G6* = Gasto en transporte y comunicaciones

*G7* = Gasto en esparcimiento y enseñanza.

Los coeficientes de los modelos estudiados y sus errores estándar (entre paréntesis se indican en la tabla siguiente:

	<i>M1</i>	<i>M2</i>	<i>M3</i>	<i>M4</i>	<i>M5</i>	<i>M6</i>
<i>IT</i>	0.79 (0.09)			0.69 (0.11)	0.59 (0.08)	0.57 (0.09)
<i>G6</i>		3.04 (0.59)		0.79 (0.62)		0.18 (0.50)
<i>G7</i>			3.33 (0.43)		2.35 (0.35)	2.33 (0.36)
$R^2$	49.40	26.16	44.49	50.53	68.61	68.67

Comentar los resultados y elegir el mejor modelo.

- Ejercicio 180** Los siguientes datos se refieren a seis pisos que pone a la venta una agencia inmobiliaria. Los datos son un índice de valoración del barrio en el que se ubica, la distancia desde cada piso al centro escolar más próximo en km y el precio por metro cuadrado de la vivienda. Estudia el modelo de regresión lineal múltiple cuya variable dependiente es el precio y las otras dos las variables independientes o regresoras.

barrio	Distancia	precio
4	1.5	1600
3	2.2	1120
1.6	1.0	690
1.2	2.0	900
3.4	0.8	1230
4.8	1.6	1860

**Ejercicio 181** *Dados los puntos de coordenadas  $(20,22)$   $(16,41)$   $(10,120)$   $(11,89)$   $(14,56)$  ajusta una parábola a estos puntos considerando la segunda variable como variable dependiente usando los procedimientos siguientes:*

- 1. Utilizando las ecuaciones normales de ajuste polinomial.*
- 2. Usando un modelo de regresión múltiple y el enfoque matricial.*
- 3. Usando algún paquete estadístico.*

## Tema 16

# Diversas técnicas de Análisis Multivariante .

En este tema vamos a limitarnos a describir el objetivo de algunas técnicas de Análisis Multivariante como son el Análisis Discriminante, el Análisis Cluster, el Análisis Factorial y el de Componentes Principales, ilustrándolas con algunos ejemplos, que resolveremos con ayuda de Statgraphics Plus 5.0.

### 16.1 El Análisis Discriminante

Uno de las primeras aplicaciones del análisis discriminante en un trabajo de Fisher que trata sobre la clasificación en especies (Virgínica, Setosa y Versicolor) de las flores del género Iris. La clasificación se basaba en los tamaños (longitud y anchura) de los sépalos y los pétalos de las flores. El análisis discriminante es una técnica de clasificación y asignación de un individuo a un grupo usando como criterio de clasificación ciertas características conocidas de este individuo. Se dispone de una serie de grupos ya establecidos, con una serie de observaciones para cada individuo referidas a un conjunto de variables relevantes. Sobre la base de esta información se llega a calcular una función discriminante para hacer predicciones futuras, permitiendo asignar individuos al grupo formado por los individuos que poseen características más parecidas a las suyas propias.

Los objetivos de este tipo de análisis son:

Determinar si existen diferencias significativas (perceptible en las variables consideradas) entre los grupos ya establecidos.

Detectar el conjunto de variables que expliquen mejor la diferencia entre los grupos.

Establecer reglas de clasificación que permitan la asignación de un nuevo dato a uno de los grupos.

Por ejemplo se puede aplicar el método de Análisis Discriminante en el caso de un banco que desea hacer un pronóstico sobre si un cliente nuevo que ha solicitado un préstamo va a pagarlo o no. En este caso pueden asignarse dos valores a la variable dependiente: 1 si el préstamo se paga, 0 si no se paga. Las variables independientes suelen ser en este caso características del cliente, tales como ingresos, patrimonio, deudas pendientes, etc.

En Ingeniería estos métodos son aplicables al problema de reconocimiento de patrones (*Pattern Recognition*), para diseñar máquinas capaces de realizar clasificaciones de una manera automática.

Las variables que mejor discriminan son las que sirven para determinar las llamadas variables canónicas, que son combinaciones lineales de las variables originarias y vienen expresadas por una función discriminante.

**Ejemplo 79** *Vamos a considerar el caso tratado por Fisher: el de las flores del género Iris usando el fichero mvdata.sf del directorio de datos de STAT-GRAPHICS PLUS para DOS.*

Las variables contenidas en este fichero contienen datos sobre 150 flores pertenecientes a tres especies diferentes. La variable *species* contiene la especie de la flor clasificada por medio de un número (1, 2 o 3), la variable *lsepal* es la medida de la longitud de su sépalo, la variable *wsepal* la medida de su anchura. Las variables *lpetal* y *wpetal* indican la longitud y la anchura del pétalo. En la siguiente tabla mostramos algunos de los datos del fichero:

species	lsepal	wsepal	lpetal	wpetal
1	5.1	3.5	1.4	0.2
1	4.9	3	1.4	0.2
1	4.7	3,2	1.3	0.2
....				
2	7	3.2	4.7	1.4
2	6.4	3.2	4.5	1.5
2	6.9	3.1	4.9	1.5
....				
3	6.3	3.3	6	2.5
3	5.8	2.7	5.1	1.9
....				

Comprobaremos que las variables que recogen las medidas de las flores, dadas en el fichero, son adecuadas para clasificar las especies de flores entrando en el procedimiento *Special, Multivariate Analysis, Discriminant Analysis*.

Trás introducir las variables obtenemos distintos resultados. Interpretamos algunos de ellos.

En la siguiente tabla un *p-value* menor que 0.05 nos indican que las dos funciones discriminante obtenidas por el programa son significativas.

Functions Derived	Wilks Lambda	Chi-Square	DF	P-Value
1	0.0234386	546.1153	8	0.0000
2	0.7779730	36.5297	3	0.0000

La siguiente tabla de clasificación nos indica que las funciones discriminantes han clasificado correctamente el 98% de las flores, especificando los errores cometidos:

Classification Table

Actual species	Group Size	Predicted species		
		1	2	3
1	50	50 (100.00%)	0 ( 0.00%)	0 ( 0.00%)
2	50	0	48 ( 96.00%)	2 ( 4.00%)
3	50	0 ( 0.00%)	0 ( 0.00%)	50 (100.00%)

Percent of cases correctly classified: 98.00%

Por tanto las variables usadas sirven bastante bien para clasificar las flores en las tres especies.

## 16.2 El Análisis Cluster o de conglomerados

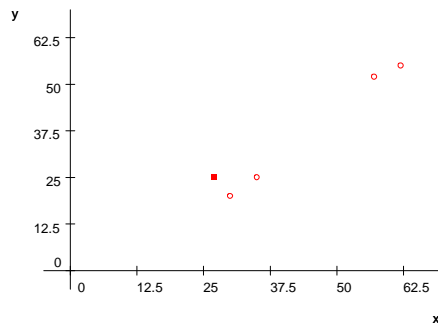
Es una técnica para formar grupos con los elementos de un conjunto. La diferencia entre este procedimiento y el Análisis Discriminante es que aquí no se conocen "a priori" el grupo de pertenencia, ni tampoco el número de grupos. Los grupos son sugeridos por la propia esencia de los datos. En cada grupo o cluster se incluyen elementos tan parecidos entre sí y al mismo tiempo tan diferentes al resto de grupos como sea posible basándose en los datos de las variables observadas de los elementos que se desean clasificar.

Para aclarar esta idea proponemos el ejemplo siguiente: Supongamos los siguientes datos que nos dan la altura y el peso de cuatro perros.

	perro 1	perro 2	perro 3	perro 4
altura	30	35	57	62
peso	20	25	52	55

Notamos inmediatamente que hay dos perros pequeños y dos perros grandes. Los clasificamos en dos grupos. ¿A qué grupo asignamos un perro de altura 27 y peso 25?

Representando los valores como puntos se ve aún más fácilmente.



El parecido entre los elementos puede medirse con una medida de proximidad entre los puntos, que en este contexto suelen llamarse similaridad (proximidad) o disimilaridades (distancia). Entre las disimilaridades puede considerarse por ejemplo, la distancia Euclídea, Norma  $r$ , la distancia de Mahalanobis...

El análisis cluster puede servir también para clasificar variables.

Los procedimientos de formación de cluster pueden ser jerárquicos o no jerárquicos. Los clusters jerárquicos pueden ser aglomerativos. Comienzan designando un grupo distinto a cada uno de los elementos. Luego se van formando agrupaciones con los clusters anteriores hasta llegar a una clasificación final de los elementos primitivos. Los clusters divisivos realizan el proceso en orden inverso. La representación gráfica de todos los pasos de este proceso se recoge en el Dendrograma.

Mostramos como ejemplo de aplicación de esta técnica a los datos del fichero *europa.sf3* que contiene porcentajes de empleados en diferentes actividades en algunos países europeos en 1979. Se pretende clasificar los países en grupos, compuestos por países que sigan patrones de empleos similares.

El fichero contiene datos de 26 países en las siguientes variables:

*Country*: Nombre del país

*Agr*: Porcentaje de empleos en agricultura

*Min*: Porcentaje de empleos en minería

*Man*: Porcentaje de empleos en manufacturas

*PS*: Porcentaje de empleos en industria

*Con*: Porcentaje de empleos en industrias suministradoras de energía

*SI*: Porcentaje de empleos en industrias de servicios

*Fin*: Porcentaje de empleos en el mundo de las finanzas

*SPS*: Porcentaje de empleos en servicios sociales.

*TC*: Porcentaje de empleos en comunicación y transporte.

Mostramos, a modo de ejemplo, los datos de los primeros países.

Country	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Belgium	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Denmark	0.1	9.2	21.8	0.6	8.3	14.6	6.5	32.2	7.1
France	0.8	10.8	27.5	0.9	8.9	16.8	6.0	22.6	5.7
W. Germany	6.7	1.3	35.8	0.9	7.3	14.4	5.0	22.3	6.1
.....									

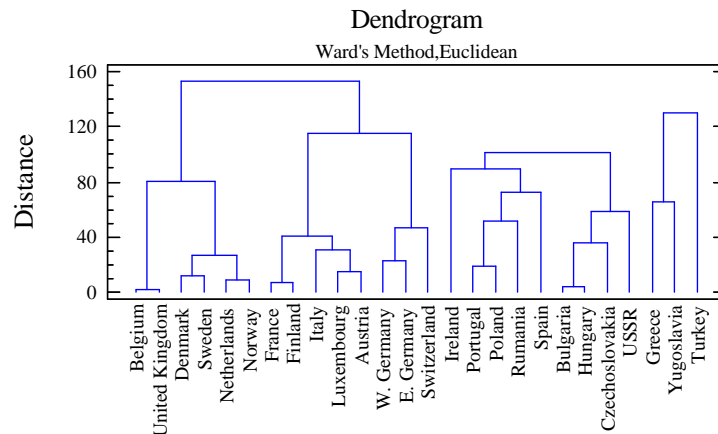
Interpretamos algunos resultados obtenidos seleccionando las siguientes opciones; tres clusters, método de Wards, distancia euclídea y datos sin tipificar (no se ha seleccionado la opción *Standarsize*).

La tabla de centroides es la siguiente:

	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
1	8.23571	0.9857	29.0857	0.9571	8.4286	16.2786	5.0	24.1714	6.8643
2	25.0222	1.7778	28.0778	0.9333	8.7222	9.5444	2.1333	17.1111	6.6889
3	52.3	0.9333	14.1	0.6	5.2667	7.7	4.9333	9.4	4.6333

Los centroides, centro de gravedad de los puntos de cada uno de los cluster y por tanto los valores medios de cada grupo, nos informan de que los países del cluster 1 se caracterizan por tener poco empleo en agricultura, mucho en industria de servicios, en finanzas y en servicios sociales. Son los más ricos e industrializados (en 1979). Los países del grupo 2 tienen las siguientes características: Los parámetros toman valores intermedios. Destacan en los empleos en Minería. Los clasificamos como países en vía de desarrollo. Los del tercer grupo son eminentemente agrícolas con poca industria y servicios. Son los menos desarrollados.

En el dendograma los países que están conectados pertenecen al mismo cluster.



Podemos ver los nombres de los países en el eje horizontal del dendrograma. El cluster 1 comienza en Bélgica y termina en Suiza, España se encuentra en el cluster 2 y el cluster 3 está formado por Grecia, Yugoslavia y Turquía, con lo que podemos reafirmarnos en la primera impresión en cuanto al tipo de países que forman cada Cluster. El dendrograma representa el orden en que se han ido formando los grupos, así el primer agrupamiento ha sido de Bélgica con el Reino Unido que serían los dos países más similares en los datos analizados. La siguiente pareja más parecida es la formada por Bulgaria y Hungría.

### 16.3 El Análisis de componentes Principales

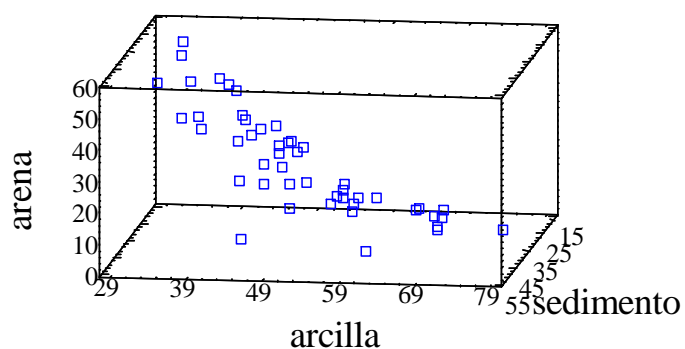
El análisis de Componentes principales trata de transformar un conjunto de variables (variables originales) en un conjunto con menor número de variables (componentes principales), con la particularidad de que las nuevas variables están incorreladas entre sí. El Análisis de componentes principales es una técnica de reducción de la dimensión. Esto quiere decir que se pretende conservar la mayor cantidad posible de información contenida en los datos originales, en un número menor de variables, las componentes principales, que son combinación lineal de las variables primitivas. De este modo sería posible obtener tantas componentes como variables originales, siempre que no haya dependencia lineal perfecta entre los datos, aunque esto no se hace en la práctica ya que lo que se persigue es disminuir el número de variables a considerar. Si a pesar de todo conservamos el número de variables originales lo que obtendremos es un cambio de ejes de coordenadas. La dirección de la primera componente principal se obtiene de forma que las proyecciones de los



datos primitivos sobre ella tengan la máxima dispersión. En el Análisis de Componentes Principales, la primera componente sería aquella que explique una mayor parte de la varianza total, la segunda componentes ha de ser perpendicular a la primera y explicar la mayor parte de la varianza restante, es decir, de la que no explicaba la primera y así sucesivamente.

Para el ejemplo usamos el fichero *soil.sf* de *statgraphicsplus 5.0*, aunque lo hemos traducido al español como *suelo.sf3*. Sus variable recogen la cantidad de arcilla, sedimento y arena contenida en muestras de tierra obtenidas en distintas localizaciones y profundidades. En primer lugar, observamos la representación de los datos de estas tres variables.

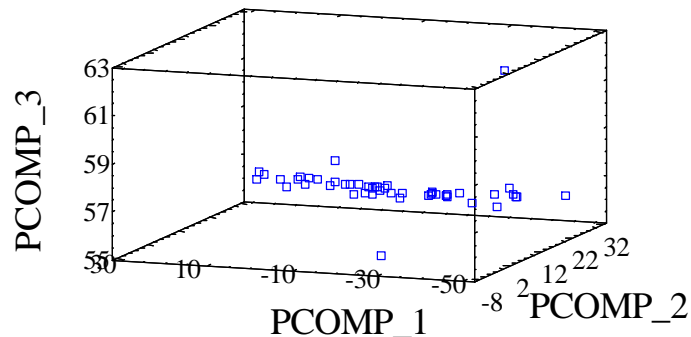
Plot of arena vs sedimento and arcilla



Observamos que los puntos parecen estar aproximadamente en un plano, por lo que una representación por dos variables va a ser razonable.

Representando estos puntos con respecto a las coordenadas que se obtendrían usando como referencia las componentes principales obtendríamos la representación siguiente.

Plot of PCOMP\_3 vs PCOMP\_2 and PCOMP\_1



Se observa que casi todos los puntos han quedado a la misma altura, por lo que la varianza de la componente 3 es bastante pequeña, así que suprimir la tercera componente no va a suponer una gran pérdida de información.

Comenzamos ahora la aplicación del método de componentes principales detallando como se obtienen estas componentes: Usamos el procedimiento *Principal Components* de Statgraphicsplus 5.0 seleccionando 2 como número de componentes y usando los datos primitivos tal como se dan en el fichero (es decir que no se selecciona la opción *standarsize*, pues en ese caso se tipificarían las variables).

Comentamos e interpretamos algunas salidas del programa:

En primer lugar obtenemos la siguiente tabla:

Principal Components Analysis

<i>Component number</i>	<i>eigenvalue</i>	<i>Percent of Variance</i>	<i>Cumulative Percentage</i>
1	314.186	90.275	90.275
2	33.1924	9.537	99.813
3	0.652113	0.187	100.000

En la columna *Percent of Variance* esta indicado el porcentaje de varianza total explicada por cada una de las componentes, así la primera componente retiene el 90.275% de la información contenida en los datos primitivos. La segunda componente retiene el 9.5375% de la información restante, con lo que ambas componente retienen el 99.81% de la información que se da en las variables primitivos. Por lo tanto la reducción de dimensionalidad se hace a costa de solo un 0.187% de pérdida de información.

A continuación damos la relación entre las nuevas variables (las componentes principales ) y las primitivas:

	Component 1	Component 2
arena	0.775069	0.268923
sedimento	-0.17735	-0.775955
arcilla	-0.606477	0.57059

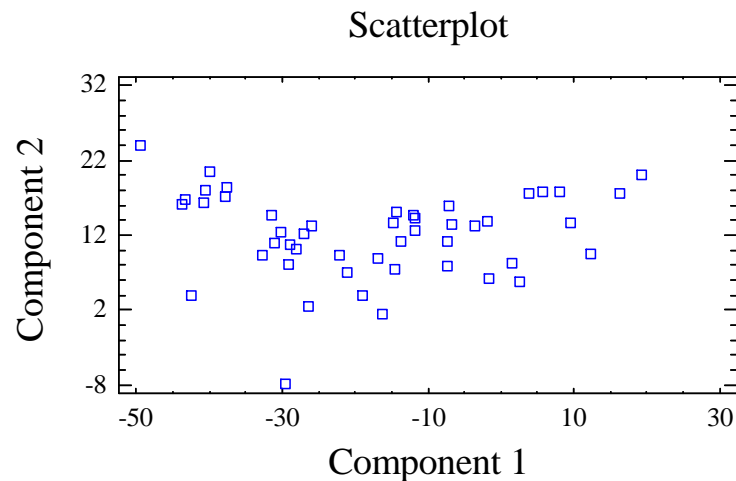
de modo que los valores para el primer componente principal se obtienen con la expresión:

$$0.775069 \times \text{arena} - 0.17735 \times \text{sedimento} - 0.606477 \times \text{arcilla}$$

A continuación se dan las coordenadas para las primeras filas del fichero de datos con respecto a las dos componentes seleccionadas

row	Component 1	Component 2
1	-12.0746	14.7559
2	-6.64362	13.4448
3	-14.6943	13.6992
...	...	...

La siguiente gráfica representa las proyecciones de los puntos de partida en el plano de las dos primeras componentes.



## 16.4 El Análisis Factorial

Es una técnica que, como el análisis de componentes principales, trata de concentrar la información contenida en los datos en algunas variables, ocultas e inobservables, que se llaman factores. En el análisis factorial exploratorio no se conoce a priori el número de factores, sino que se determinan con el propio modelo factorial. En el Análisis Factorial la variable primitiva es la dependiente y las ocultas o factores son las variables independientes. Cada uno de estos factores explica una parte de la varianza de cada variable primitiva. El número de factores es menor que el de variables. La parte de varianza de cada variable explicada por los factores recibe el nombre de *comunalidad* y la no explicada se conoce con el nombre de *especificidad* o *unicidad* ya que no puede ser atribuida a ningún factor.

El Análisis Factorial confirmatorio se utiliza para confirmar que un número de factores determinado de antemano es suficiente para la estimación del modelo. A veces los primeros factores comunes extraídos no tienen una fácil interpretación y puede utilizarse rotaciones ortogonales u oblicuas para obtener soluciones válidas que al mismo tiempo faciliten la interpretación de estos factores en función de las variables primitivas.

Ilustramos la aplicación de este método con un ejemplo. Los datos son características económicas, sociales y políticas de 14 países que se dan en el siguiente fichero *naci55.sf3* y que se recogen en la tabla siguiente:

Nación (1955)	rpc	trade_million	powran	stability	freedom
Brazil	91	2729	7	0	2
Burma	51	407	4	0	1
China	58	349	11	0	0
Cuba	359	1169	3	0	1
Egypt	134	923	5	1	1
India	70	2689	10	0	2
Indonesia	129	1601	8	0	1
Israel	515	415	2	1	2
Jordan	70	83	1	0	1
Netherlands	707	5395	6	1	2
Poland	468	1852	9	0	0
USSR	749	6530	13	1	0
UK	998	18677	12	1	2
US	2334	26839	14	1	2

	for_ con	agre_ usa	defense	gnp_ def	acep_ law
	0	69.1	148	2.8	0
	0	-9.5	74	6.9	0
	1	-41.7	3054	8.7	0
	0	64.3	53	2.4	0
	1	-15.4	158	6	1
	0	-28.6	410	1.9	1
	0	-21.4	267	6.7	0
	1	42.9	33	2.7	1
	1	8.3	29	25.7	0
	0	52.3	468	6.1	1
	1	-41.7	220	1.5	0
	1	-41.7	34000	20.4	0
	1	69	3934	7.8	0
	1	100	40641	12.2	1

El significado de las variables es el siguiente:

*rpc* = Renta per Capita.

*trade\_million* = Dinero en circulación (en millones de dólares).

*powran* = Índice de gasto energético.

*stability* = 0 (Inestable por que se dan alguna de estas circunstancias: guerrillas, revueltas de la población, golpe de estado, huelgas frecuentes...), 1 (estable).

*freedom* = 0 (prohibida la oposición política), 1 (permitida, pero no puede hacer campañas de control del gobierno), 2 (aceptada).

*for\_con* = 1 (malas relaciones con países extranjeros), 0 = relaciones normales

*agre\_usa* = grado de acuerdo con los Estados Unidos en las Naciones Unidas.

*defense* = presupuesto de Defensa en millones de dólares.

*gnp\_def* = porcentaje de la Renta per Capita empleada en Defensa.

*acep\_law* = 0 (no suscribe el estatuto de la Corte de Justicia Internacional), 1 (lo suscribe con restricciones)

Interpretamos algunos de los resultados obtenidos usando el procedimiento Factor Analysis de Stagraphicsplus 5.0 eligiendo las opciones *Standarsize*, *método: Principal componentes, rotación varimax y número de factores 3*.

La siguiente tabla nos indica el porcentaje de varianza (información) sobre los datos retenida en cada uno de los factores, así el factor número 1 retiene el 42.164% de la varianza. Como hemos decidido obtener tres factores, los que

374TEMA 16. DIVERSAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE .

retengan más información, observamos que entre los tres primeros retienen el 78.059% de información. La reducción de la dimensionalidad de diez, que hay en los datos primitivos, a tres (los tres factores) se hace a costa de una pérdida aproximada del 22% de la información.

<i>Factor Number</i>	<i>Eigenvalue</i>	<i>Percent of Variance</i>	<i>Cumulative Percentage</i>
1	4.21642	42.164	42.164
2	2.43332	24.333	66.497
3	1.15621	11.562	78.059
4	0.952816	9.528	87.588
5	0.53895	5.389	92.977
6	0.29764	2.976	95.954
8	0.0840885	0.841	99.436
9	0.0417072	0.417	99.853
10	0.0146737	0.147	100.000

La siguiente tabla, obtenida después de la rotar los factores, nos da información sobre la influencia que tienen las variables primitivas en cada uno de los factores. Esta tabla se usa en ocasiones para dar una interpretación de los factores (en este caso sería el tipo de información sobre las naciones contenida en los factores)

Factor	1	2	3
rpc	0.844909*	0.442061	0.13107
trade _million	0.907236*	0.351939	0.0172263
powran	0.815629*	-0.194869	0.135147
Stability	0.332096	0.736393*	0.406301
freedom	0.138658	0.73446*	-0.556049
for _con	0.197234	0.151275	0.833105*
agre _usa	0.41044	0.59383*	-0.439799
defense	0.81275*	0.0936149	0.409729
gnp _def	0.202335	-0.131529	0.670016*
acep _law	-0.0396941	0.841814*	0.0174381

Para facilitar la interpretación de los factores se han señalado, para cada variable, el factor en que tiene más peso. De esta forma obtenemos que en el factor 1 influyen sobre todo las variables Renta per Capita, gasto energético y presupuesto de Defensa. Como todas estas variables están relacionadas con la riqueza del país podríamos dar a este factor la interpretación de *índice de riqueza*. En el segundo factor tienen importancia las variables stability,

freedom, agre\_usa y acep law. Como estas características corresponden a los países democráticos occidentales, le llamaremos *índice de democracia*. El tercer factor se caracteriza por la variable que da el índice de conflicto con países extranjeros y el gasto en defensa por persona. Este factor podría estar relacionado con el *peligro de conflicto bélico*.

La tabla siguiente nos presenta la comunalidad de cada variable (la comunalidad indica la parte de información, en tanto por 1, sobre cada variable que ha sido retenida por los factores seleccionados).

Variable	Communality
rpc	0.926469
trade_million	0.947235
powran	0.721489
Stability	0.817644
freedom	0.867847
for_con	0.755849
agre_usa	0.714518
defense	0.837205
gnp_def	0.507161
acep_law	0.710531

En este caso la variable mejor recogida por los tres factores es trade\_million y la peor representada es gnp\_def.

También se pueden calcular los índices de las naciones respecto de las nuevas variables (los tres factores). Vienen dados en la tabla siguiente:

Nación (datos de 1955)	Factor 1 Riqueza	Factor 2 Democracia	Factor 3 Peligro bélico
Brazil	-1.27175	-0.318898	-3.07361
Burma	-2.88742	-2.24222	-1.41219
China	-1.34271	-3.58445	1.67618
Cuba	-2.09238	-0.976246	-2.4537
Egypt	-1.6379	1.16175	1.05761
India	-1.61656	0.068466	-2.15807
Indonesia	-1.95626	-2.45678	-1.17135
Israel	-1.20536	3.22486	-0.482575
Jordan	-2.42183	-1.94288	1.72317
Netherlands	0.223803	3.15535	-1.67452
Poland	-1.37471	-3.01757	0.9392
USSR	3.55669	-1.47333	4.73718
UK	4.17215	2.468	0.315048
US	9.85423	5.93394	1.97763

Concluimos que los países con mayor índice de riqueza serían (en 1955) Estados Unidos y la URSS, los de mayor índice de democracia son Estados Unidos e Israel y los de mayor peligro bélico URSS y Estados Unidos.

## 16.5 EJERCICIOS PROPUESTOS

**Ejercicio 182** Las siguientes tablas suministran los gastos por distintos conceptos en algunas comunidades autónomas españolas (en el fichero *coaut.sf3*):

*AL*=Alimentación, bebidas y tabaco

*VES*=Vestido y calzado

*VIV*=Vivienda, calefacción y alumbrado

*SER*=Artículos de mobiliario, menaje y conservación del hogar

*MED*=Servicios médicos y sanitarios

*TRANS*=Transportes y comunicaciones

*ESP*=Esparcimiento, enseñanza y cultura

*OBIEN*=Otros bienes y servicios

*OGAS*=Otros gastos

COMUN	AL	VES	VIV	SER	MED
ANDALUCIA	605	222	183	121	50
ARAGON	548	255	202	126	50
ASTURIAS	587	281	233	131	58
BALEARES	550	227	206	144	87
CANARIAS	572	186	180	134	74
CANTABRIA	588	289	261	118	64
CASTILLA LA MANCHA	543	221	201	126	53
CASTILLA Y LEON	547	218	191	119	41
CATALUÑA	686	262	283	163	93
CEUTA Y MELILLA	683	193	134	81	27
COM. VALENCIANA	542	218	177	132	62
EXTREMADURA	470	211	138	100	44
GALICIA	615	248	199	132	51
LA RIOJA	602	210	196	127	55
MADRID	674	254	253	146	87
MURCIA	604	210	189	128	47
NAVARRA	643	325	251	221	81
PAIS VASCO	636	267	232	158	65



TRANS	ESP	OBIEN	OGAS	COMUN
255	115	281	83	ANDALUCIA
247	111	263	82	ARAGON
336	151	311	121	ASTURIAS
357	151	334	131	BALEARES
305	159	280	98	CANARIAS
302	117	277	106	CANTABRIA
244	96	253	101	CASTILLA LA MANCHA
252	110	260	108	CASTILLA Y LEON
361	228	363	107	CATALUÑA
142	84	235	64	CEUTA Y MELILLA
280	123	281	99	COM. VALENCIANA
208	87	223	71	EXTREMADURA
291	128	256	104	GALICIA
263	126	313	122	LA RIOJA
370	216	433	130	MADRID
319	104	310	114	MURCIA
407	186	409	155	NAVARRA
343	174	395	122	PAIS VASCO

Calcular con *Statgraphics* las tres primeras componentes principales por medio de su relación con las variables primitivas y el porcentaje de varianza explicada por éstas componentes.

**Ejercicio 183** Una compañía aseguradora ha realizado una investigación sobre la siniestralidad en los vehículos asegurados con el objeto de obtener un criterio para la admisión de nuevos clientes. Para ello ha seleccionado aleatoriamente 40 pólizas, clasificadas en dos grupos, separando a los asegurados que han tenido un siniestro grave de los restantes. Se desea tomar en consideración la información sobre la edad del conductor, la antigüedad del coche y su potencia. Los datos obtenidos son los del fichero *siniestr.sf3*. Los vehículos con siniestro grave vienen codificados en la variable *siniestr* con un 1 y los restantes con un 2:

1. Aplicar la técnica de Análisis discriminante para estudiar qué información dan las tres variables sobre si un asegurado va a tener o no un siniestro grave.
2. Con esta información, ¿Cuál sería el pronóstico para un posible cliente de 30 años de edad, que desea asegurar un vehículo de 5 años de antigüedad y con valor 150 para la potencia?

**Ejercicio 184** Se dispone de 6 observaciones y dos variables y se trata de reunir las observaciones en dos grupos, en función de la semejanza entre los

378TEMA 16. DIVERSAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE .

valores de estas dos variables a) Utilizar como opciones la distancia euclídea al cuadrado y para la vinculación intergrupos el método del vecino más próximo.

Observación	1	2	3	4	5	6
$x_1$	1	2	2	2.5	6	9
$x_2$	6	6.5	7	3	4	6

**Ejercicio 185** Usando los datos de las variables acel, cilindros, litros/cien, peso, precio, y potencia del fichero de datos coches.sf.

1. Realiza un análisis factorial para seleccionar un número de factores que parezca adecuado para representar las características de estos vehículos reteniendo al menos el 90% de la información contenida en los datos.
2. Realiza una rotación de estos factores para facilitar su interpretación.
3. Especifica los pesos de las variables primitivas en estos factores.
4. Da la ecuación expresión del primer factor en función de las variables primitivas.
5. Da los valores de estos factores para el primer vehículo.
6. Intenta dar un nombre a cada uno de estos factores, ya que pretendemos usarlos para describir los vehículos.

Unidad Temática VI

**SERIES TEMPORALES**



## Tema 17

# Series temporales . Modelos Clásicos.

### 17.1 Introducción y ejemplos de series temporales

Una **serie temporal, cronológica o histórica** es una sucesión de valores de una variable observada durante un periodo de tiempo. Son ejemplos de series temporales el registro de la temperatura máxima durante todos los días del año en una ciudad, el precio de los automóviles de una cierta marca registrados mensualmente, las cotizaciones diarias de las acciones en Bolsa. Con el estudio de estas series se pretende tener conocimiento de la evolución de los valores de la variable tratada, con el objeto de hacer predicciones en el supuesto de que no se produzcan cambios de estructuras en el fenómeno estudiado. Por ejemplo, si la serie que se esta considerando son los gastos mensuales de una familia con el propósito de realizar un pronóstico sobre el gasto en los meses sucesivos, un cambio de estructura podría ser que la familia obtuviese un gran premio en la Loteria en el mes siguiente del último registrado en la serie.

En Economía y en general en el mundo empresarial hay que tomar decisiones, invertir, contratar más empleados, arreglar la maquinaria o adquirir una nueva. Estas decisiones hay que tomarlas en una en situación de incertidumbre, porque no hay seguridad sobre la evolución del mercado de trabajo, de los precios de las materias primas, de las perspectivas de ventas... El análisis y predicción de series temporales pretende reducir el grado de incertidumbre sobre los valores en que van a situarse los parámetros de interés. Con las técnicas de previsión se trata de hacer pronósticos los más acertadamente posible sobre sucesos que todavía no han tenido lugar. Las previsiones se basan en la información proporcionada por los sucesos ocurridos en un pasado más o menos remoto.

Una herramienta muy usada y también muy útil para el estudio de las series temporales es su representación gráfica: En el eje horizontal se representa el registro temporal. Por lo general los datos suelen estar tomados en intervalos regulares, por ejemplo cada día, cada mes... En el eje vertical suelen representarse los valores de la serie. Los puntos resultantes suelen unirse por medio de una línea poligonal, ya que de esta forma se facilita su visibilidad y por lo tanto su análisis.

La gráfica de la figura 17.1 representa la serie temporal, con datos mensuales, de ventas de botellas de una cierta marca de refresco. En ella se observa que los momentos de más venta se repiten periódicamente cada año, coincidiendo posiblemente con los periodos de verano. También se aprecia una cierta tendencia a la periodicidad anual, aunque el volumen de ventas experimenta un cierto aumento según pasan los años.

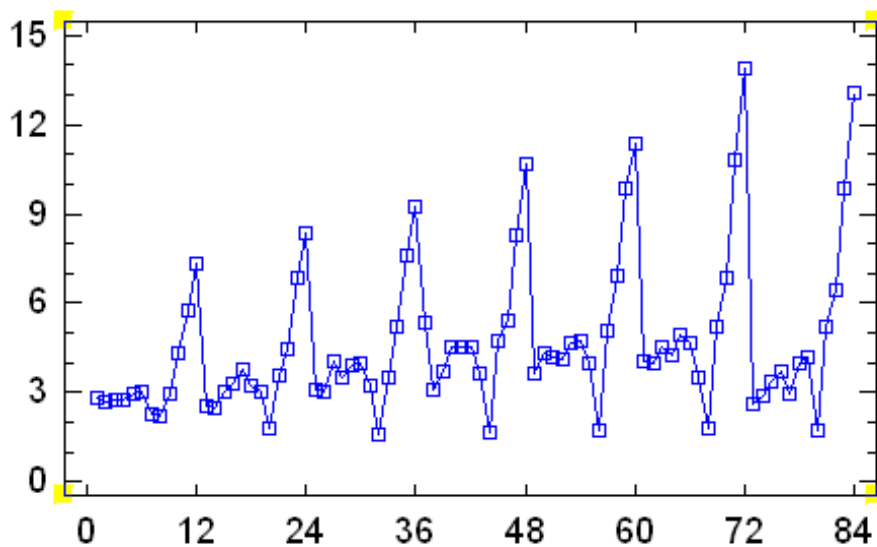


Figura 17.1:

La gráfica de la figura 17.2 representa la serie temporal del número de nacimientos en España desde 1946 a 1996. También ahora, a la vista de la gráfica se pueden sacar algunas conclusiones obvias, como la caída de la natalidad a partir de mediados de los setenta.

Para hacer predicciones en base a los datos de una serie temporal **es necesario que esta serie contenga información**. Para poner de manifiesto este hecho consideremos dos series temporales de características bien distintas: la primera registra la hora de la pleamar en Cádiz durante 60 días

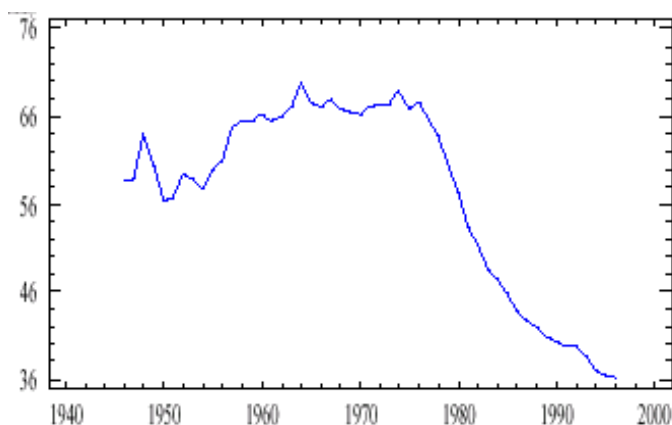


Figura 17.2:

consecutivos, la segunda los números que van saliendo consecutivamente durante una partida de Bingo. Con la primera se pretende predecir la hora de la pleamar del siguiente día. Con la segunda el número que va salir a continuación en la partida de Bingo. En el primer caso se puede realizar un pronóstico acertado. En el segundo no se puede hacer ninguna predicción sobre el futuro. Son ejemplos extremos. La primera serie es totalmente determinista y por tanto se puede predecir muy bien. En el segundo no se puede predecir nada. La serie es puramente aleatoria. Por lo general no estamos en estos casos extremos, sino que, para muchas series temporales, puede considerarse que tienen una parte determinista y otra aleatoria. Por ejemplo, si consideramos una serie que este formada por las temperaturas diarias registradas a las 12 del mediodía durante varios años, esta serie tiene una componente determinista y otra aleatoria. De esta serie se puede obtener alguna información sobre la evolución de la temperatura, que empleada adecuadamente permite a los servicios meteorológicos hacer predicciones más o menos acertadas sobre la temperatura diaria. Pero con esta información no es suficiente para predecir *con exactitud* la temperatura que hará mañana. En este caso decimos que la serie tiene componentes deterministas y aleatorios.

En el estudio de series temporales se pueden considerar dos enfoques alternativos: El **análisis univariante** en el que la predicción se realiza por medio de la información contenida en una única serie, y al **análisis causal** que realiza la predicción por medio del estudio de varias series temporales que están relacionadas entre sí.

En el análisis univariante de series temporales se pueden considerar distintos modelos: métodos de descomposición, métodos de alisado, modelos

ARIMA y modelos espectrales

A) **Métodos de descomposición**

En este modelo la serie temporal se suele considerar compuesta de todos o algunos de los componentes siguientes, que se combinan como una suma (modelo aditivo) o como un producto (modelo multiplicativo):

1) Tendencia. Refleja las variaciones a largo plazo

2) Movimiento estacional. Componente que tiene un periodo constante. Por ejemplo las temperaturas tienen una periodicidad anual.

3) Componente Cíclica. Este factor cíclico recoge los movimientos oscilatorios por encima y por debajo de la tendencia. La duración de un ciclo en las series económicas suele ser de más de un año y su periodo no se mantiene constante. Su estudio es difícil y es frecuente considerar la componente cíclica conjuntamente con la tendencia.

4) Componente irregular o Residual. Recoge todo lo que no es explicable con la tendencia, variación estacional y factor cíclico. Se compone de dos partes. Una de estas partes no es predecible, pero fácil de detectar y explicar a posteriori, ya que normalmente es producida por circunstancias excepcionales. La otra es un componente aleatorio que es completamente impredecible.

B) **Métodos de alisado exponencial**. Los modelos de alisado exponencial simple también calculan los valores de la tendencia, pero se trata de una tendencia de carácter local calculando los valores en función de los datos más cercanos. La predicción se realiza en base a la media ponderada de un cierto número de estos elementos. Entre estos métodos se encuentran los de medias móviles, y los de Holt(1957), Winters(1960) y Brown (1956, 1959, 1961, 1963).

C) **Modelos ARIMA**. Se considera que la serie temporal objeto de estudio ha sido generada por un proceso estocástico. Este enfoque se popularizó a partir de 1970 merced a la obra de los estadísticos Box y Jenkins. Los métodos de alisado exponencial se pueden considerar un caso particular de los modelos ARIMA. También se han desarrollado modelos que incluyen en un solo enfoque los modelos de descomposición y los ARIMA con el nombre de UCARIMA. Estos modelos han sido desarrollados por Harvey (1984)

D) **Modelos Espectrales**, basados en el análisis de Fourier. Son particularmente útiles para el estudio de la estacionalidad. Cuando se realiza el estudio de series temporales por estos modelos se suele decir que el análisis se realiza en el llamado *Dominio de la frecuencia*, porque no se estudian directamente los valores de la serie temporal sino los coeficientes de su serie transformada de Fourier.

La importancia del análisis de series temporales se ha puesto de manifiesto recientemente por la concesión del Premio Nobel de Economía 2003 a Robert F. Engle y Clive W. J. Granger por sus trabajos sobre series económicas temporales, que son aplicables a los análisis de los mercados financieros, a la



evolución de los precios y los tipos de interés.

## 17.2 Software para el análisis de series temporales

La selección y adaptación de modelos matemáticos apropiados para el estudio, análisis y previsión de series temporales requiere una gran cantidad de cálculo. Afortunadamente en la actualidad existe una gran cantidad de programas de ordenador que facilitan enormemente esta tarea, por lo que por lo que normalmente se usa algún software estadístico para realizar estos cálculos. Entre los programas que tienen más difusión, al menos en nuestro entorno más inmediato pueden destacarse los siguientes:

**SPSS:** Es un paquete estadístico de propósito general y muy extendido y, que incluye entre sus opciones un amplio abanico de modelos para el estudio de series temporales.

**STATGRAPHICS PLUS-** También de propósito general, con poderosas opciones gráficas y facilidades de información. Distribuido por módulos: Base (para estadísticas Básicas), series temporales, diseño experimental, control de calidad, métodos multivariantes y técnicas de regresiones avanzadas.

**EVIEWS:** Orientado principalmente a la Econometría, ofrece un amplio abanico de modelos para el análisis de series temporales y publicación de gráficos de alta calidad.

Otro software también disponible para el estudio de las series temporales puede ser:

**AMBER:** Amber es un entorno de visualización y análisis de series temporales usando redes neuronales y métodos estadísticos.

**AUTOBOX:** Realiza análisis tipo ARIMA.

**FORESCSTPRO:** Incluye predicción para métodos extrapolativos, modelos Box-Jenkins y regresiones dinámicas.

**RATS:** paquete para el análisis de series temporales con buenos gráficos, manejo flexible de datos y varias técnicas estadísticas.

**SCA:** Familia de programas para la predicción y análisis de series temporales y modelos econométricos. Incluye algoritmos para la identificación automática de modelos ARIMA.

**STAMP:** Software diseñado para la predicción y el modelaje de estructuras de series temporales que no están incluidos en otros programas como por ejemplo, el Filtrado de Kalman.

**UNISTAT:** Sistema estadístico integrado con gráficos excepcionales, incluye la mayoría de las técnicas estadísticas incluyendo series temporales.

### 17.3 Parámetros para el análisis de series temporales

Denotaremos una serie temporal por  $\{X_t, t \in N\}$  donde  $X_t$  son variables. Los valores de una serie concreta los interpretamos como muestras de estas variables. Supongamos que la serie muestral contiene  $N$  elementos y designemos sus valores con la letra minúscula correspondiente. Llamamos *serie retardada con retardo  $m$*  la que resulta de la anterior de suprimiendo sus  $m$  primeros elementos, y que contendrá  $N - m$  elementos.

Para el estudio de las series temporales son útiles los siguientes parámetros.

Media muestral:

$$\bar{x} = \frac{1}{N} \sum_{t=1}^N x_t$$

Varianza:

$$S_0^2 = \frac{1}{N} \sum_{t=1}^N (x_t - \bar{x})^2$$

Autocovarianza muestral de orden  $m$

$$S_m = \frac{1}{N-m} \sum_{t=1}^{N-m} (x_t - \bar{x})(x_{t+m} - \bar{x})^2$$

Coefficiente de autocorrelación muestral de orden  $m$

$$r_m = \frac{\sum_{t=1}^{N-m} (x_t - \bar{x})(x_{t+m} - \bar{x})^2}{\sum_{t=1}^N (x_t - \bar{x})^2}$$

Es de notar que cuando aumenta el valor de  $m$  el número de elementos del numerador decrece. Por este motivo, en la práctica debe partirse de series temporales con un número de elementos,  $N$ , suficientemente grande. Además, no suelen tomarse en cuenta coeficientes de autocorrelación cuando el valor de  $m$  es mayor que la cuarta parte de la longitud de la serie

Se llama *Correlograma muestral* a la representación gráfica de los valores del coeficiente de autocorrelación muestral  $r_m$  frente a  $m$ , que es el orden del retardo correspondiente. La función correspondiente se llama *Función de Autocorrelación Muestral*. Lo característico de una serie temporal, al contrario de lo que es más habitual en casi todos los muestreos estadísticos, es que existan relaciones de dependencia entre los valores de la serie temporal y la serie con retardo  $m$ , que es la que resulta de suprimir los  $m$  primeros elementos de ésta. Los coeficientes de autocorrelación muestrales nos dan información sobre esas relaciones de dependencia. No obstante no está excluido del modelo el caso en que las variables sean independientes entre sí, como es el caso de una variable de ruido blanco.

**Ejemplo 80** Calcular la media, la varianza, los dos primeros coeficientes de autocovarianza y de autocorrelación de la serie (ruido\_1 del fichero ejemst.sf3) cuyos valores son:

$$-0.22, 0.27, -0.37, 0.15, 0.28, \\ 0.15, 0.06, -0.34, 0.24, 0.02, 0.06$$

Media muestral:

$$\bar{x} = \frac{1}{N} \sum_{t=1}^N x_t,$$

$$\bar{x} = \frac{-0.22+0.27-0.37+0.15+0.28+0.15+0.06-0.34+0.24+0.02+0.06}{11} = \frac{0.3}{11} = 0.027273$$

Varianza muestral:

$$S_0^2 = \frac{1}{N} \sum_{t=1}^N (x_t - \bar{x})^2 = \\ \frac{(-0.22-0.027273)^2 + (0.27-0.027273)^2 + \dots + (0.02-0.027273)^2 + (0.06-0.027273)^2}{11} = \\ \frac{0.55318}{11} = 0.050289$$

Autocovarianza muestral de orden 1 :

$$S_1 = \frac{1}{11-1} \sum_{t=1}^{10} (x_t - \bar{x})(x_{t+1} - \bar{x})^2$$

$$\sum_{t=1}^{10} (x_t - \bar{x})(x_{t+1} - \bar{x})^2 = (-0.22 - 0.027273)(0.27 - 0.027273) + \\ + (0.27 - 0.027273)(-0.37 - 0.027273) + \dots + \\ + (0.24 - 0.027273)(0.02 - 0.027273) + \\ + (0.02 - 0.027273)(0.06 - 0.027273) = -0.23110;$$

$$S_1 = \frac{-0.23110}{10} = -0.02311$$

Autocovarianza muestral de orden 2

$$S_2 = \frac{1}{11-2} \sum_{t=1}^9 (x_t - \bar{x})(x_{t+m} - \bar{x})^2$$

$$\sum_{t=1}^9 (x_t - \bar{x})(x_{t+m} - \bar{x})^2 = (-0.22 - 0.027273)(-0.37 - 0.027273) + \\ + (0.27 - 0.027273)(0.15 - 0.027273) + \dots + (0.24 - 0.27273)(0.06 - 0.27273) \\ = 0.17858$$

$$S_2 = \frac{0.17858}{9} = 0.019842$$

Coefficiente de Autocorrelacion de orden 1

$$r_1 = \frac{\sum_{t=1}^{10} (x_t - \bar{x})(x_{t+m} - \bar{x})^2}{\sum_{t=1}^{11} (x_t - \bar{x})^2} = \frac{-0.23110}{0.5318} = -0.43456$$

Coefficiente de Autocorrelación de orden 2

$$r_2 = \frac{\sum_{t=1}^9 (x_t - \bar{x})(x_{t+2} - \bar{x})}{\sum_{t=1}^{11} (x_t - \bar{x})^2} = \frac{0.001788}{0.5318} = 3.3622 \times 10^{-3} = 0.003362$$

La representación gráfica de la función de autocorrelación muestral o correlograma muestral presenta el siguiente aspecto:

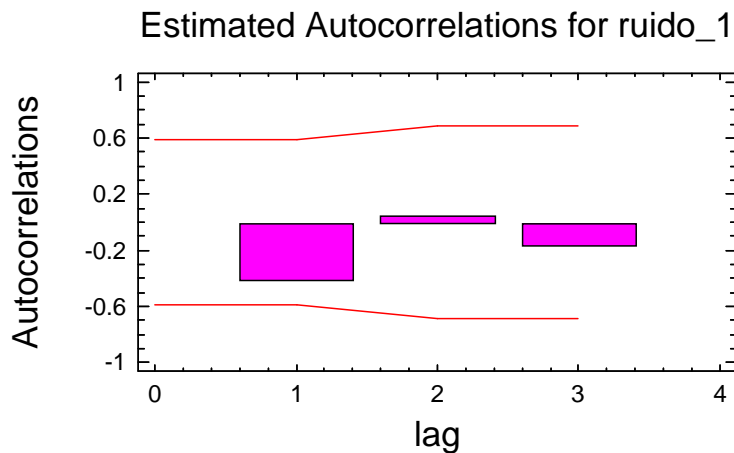


Figura 17.3:

Las líneas que rodean los coeficientes de correlación indican los respectivos intervalos de confianza que corresponderían al caso en que el correspondiente coeficiente de correlación sea nulo. Los intervalos de confianza para los coeficientes de autocorrelación muestral de cualquier orden, bajo la hipótesis de que la variable de la serie sea de ruido blanco (ver la definición en el epígrafe 17.4), pueden obtenerse, siempre que  $N$  sea suficientemente grande mediante la expresión:

$$\left( -\frac{1}{\sqrt{N}} F_{N(0,1)}^{-1}\left(1 - \frac{\alpha}{2}\right), \frac{1}{\sqrt{N}} F_{N(0,1)}^{-1}\left(1 - \frac{\alpha}{2}\right) \right)$$

Estos intervalos de confianza son de igual anchura para todos los coeficientes de autocorrelación. Para obtener intervalos de confianzas simultáneos para todos los coeficientes de autocorrelación hay que emplear otros criterios, como por ejemplo el estadístico  $Q^*$  de Ljung-Box, que da lugar a intervalos de confianza de diferente amplitud, como los que aparecen en la figura 17.3.

## 17.4 Variable de ruido blanco

Una variable de ruido blanco es una serie  $\varepsilon_t$  que cumple las condiciones siguientes: Los términos de la sucesión son muestras de variables que se distribuyen normalmente, su media es nula y su varianza constante. Además las autocovarianzas entre las variables de dos términos son nulas. Se cumplen por tanto las siguientes condiciones:

1.  $\varepsilon_t$  es  $N(\mu, \sigma) \forall t$
2.  $\mu = E(\varepsilon_t) = 0, \forall t$
3.  $E(\varepsilon_t, \varepsilon_t) = \sigma^2, \forall t$
4.  $E(\varepsilon_t, \varepsilon_{t+m}) = 0, \forall t, m \neq 0$

No hay relación de dependencia entre esta serie y sus series retardadas, según la condición tercera. Esta última propiedad es la que le da variable de ruido blanco el carácter de impredecible.

Es conveniente que la componente residual de una serie, obtenida cuando se han eliminado las componentes de tendencia y estacionalidad, sea una variable de ruido blanco, ya que esto querría decir que estas dos componentes contienen toda la información que es posible extraer de los valores de la serie. Las estimaciones de la media, la varianza, las autocovarianzas y los coeficiente de autocorrelación teóricos se realizan por medio de los parámetros muestrales dados en las expresiones muestrales del apartado 17.3.

La función de autocorrelación de una variable de ruido blanco es nula para cualquier valor de  $m$ , ya que la autocovarianza es siempre cero según la definición de este tipo de variable, por lo que el correlograma teórico aparecerá en blanco. No quiere decir, sin embargo, que cuando tengamos una serie de ruido blanco (muestra o realización de una variable de ruido blanco) el correlograma muestral esté en blanco, ya que como ocurre para todas las variables aleatorias, sus valores presentan fluctuaciones. No obstante es razonable esperar que los valores muestrales estén contenidos en los intervalos de confianza de los valores teóricos, que en este caso son nulos. En el ejemplo anterior, y si observamos el correlograma, se observa que los intervalos de confianza obtenidos incluyen dentro de sí a los valores muestrales, lo que quiere decir que, al 95% de confianza, la diferencia con los valores nulos (los teóricos) no es significativa y por tanto los valores muestrales de los coeficientes de autocorrelación de la serie del ejemplo son compatibles con el modelo teórico de ruido blanco.

## 17.5 Estudio de la tendencia

### Modelos aditivos

Considereremos en este análisis que los términos de la serie pueden considerarse como suma de dos componentes, la tendencia y el componente irregular y que esta componente residual es una variable aleatoria de ruido blanco

$$X_t = T_t + \varepsilon_t$$

Los modelos matemáticos que suelen usarse para expresar la componente  $T_t$  son de diversos tipos. Es frecuente considerar un ajuste polinómico de un cierto grado  $k$ :

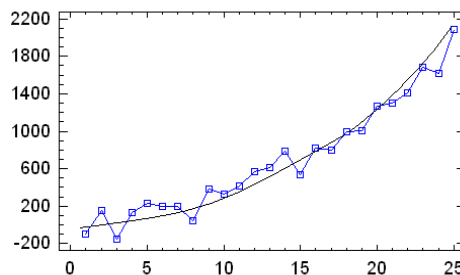
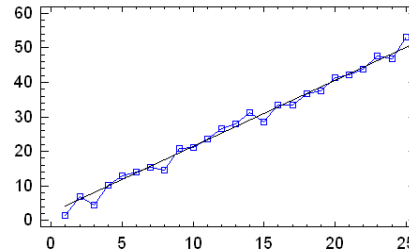
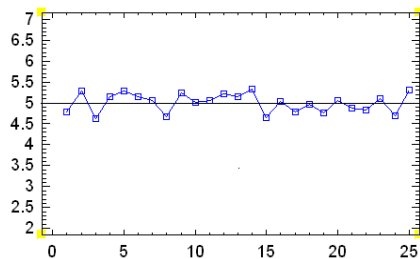
$$T_t = a_0 + a_1t + a_2t^2 + \dots + a_kt^k$$

con lo que la serie temporal tomaría la forma

$$X_t = T_t + \varepsilon_t = a_0 + a_1t + a_2t^2 + \dots + a_kt^k + \varepsilon_t$$

Los modelos con  $k = 0$ , son llamados de *media constante*. Si  $k = 1$  se dice que son de *tendencia lineal* y si  $k = 2$  se dice que son de *tendencia cuadrática o parabólica*. En general se dice que son modelos con *tendencia polinómica*.

**Ejemplo 81** En las siguientes gráficas los puntos son los términos de la serie. En cada caso se ha representado el modelo de tendencia: de media constante, lineal y parabólica o cuadrática.



Para realizar ajustes polinómico para la tendencia se suele usar el método de mínimos cuadrados.

**Ejemplo 82** *Ajustar una tendencia lineal a la serie ((lineal\_1 del fichero ejemst.sf3):*

1.5, 6.9, 4.4, 10.1, 13.0, 14.1, 15.4, 14.6, 20.7, 21.1

El ajuste se realiza aplicando el método de mínimos cuadrados a los puntos dados por la siguiente tabla:

$t$	1	2	3	4	5	6	7	8	9	10
$X_t$	1.5	6.9	4.4	10.1	13.0	14.1	15.4	14.6	20.7	21.1

Ajustando por mínimos cuadrados obtenemos la recta de regresión cuya expresión es:

$$y - \bar{Y} = \frac{S_{XY}}{S_X^2}(x - \bar{X})$$

Aplicada a las variable  $t$  y  $X_t$  resulta

$$X_t = \frac{S_{tX_t}}{S_t^2}(t - \bar{t}) + \bar{X}$$

	$t$	$X_t$	$t^2$	$tX_t$
	1	1.5	1	1.5
	2	6.9	4	13.8
	3	4.4	9	13.2
	4	10.1	16	40.4
	5	13.0	25	65
	6	14.1	36	84.6
	7	15.4	49	107.8
	8	14.6	64	116.8
	9	20.7	81	186.3
	10	21.1	100	211
Sumas	55	121.8	385	840.4
Parámetros	$\bar{t} = \frac{55}{10} = 5.5$	$\bar{X}_t = \frac{121.8}{10}$	$\bar{t}^2 = \frac{385}{10}$	$\overline{tX_t} = \frac{840.4}{10} = 84.04$

$$\frac{S_{tX_t}}{S_t^2} = \frac{\overline{tX_t} - \bar{t}\bar{X}_t}{\bar{t}^2 - \bar{t}^2} = \frac{84.04 - 5.5 \times 12.18}{38.5 - 5.5^2} = 2.0667$$

$$X_t = 2.0667(t - 5.5) + 12.18$$

Así que el modelo lineal ajustado a la serie resulta

$$X_t = 2.067t + 0.813 + \varepsilon_t$$

### Predicciones:

Una vez calculados los estimadores de los coeficientes, las funciones de tendencia ajustadas permiten realizar predicciones para periodos pasados o futuros. Para realizar estas predicciones sobre el futuro basta sustituir en el valor de  $t$  deseado en la recta de ajuste. Las predicciones relativas al pasado lo son en sentido estricto si utilizamos para calcularlas la información sobre la serie en los periodos de tiempo anteriores al considerado. Sin embargo, en los análisis del tipo que nos ocupa es más frecuente emplear como estimación del pasado la obtenida por la expresión global del ajuste, es decir la obtenida empleando todos los datos. También se puede calcular un intervalo de confianza para las predicciones del futuro. Estos intervalos serán mayores cuanto más nos alejemos de los valores conocidos de la serie temporal.

		Predicciones	Residuos o errores
$t$	$x_t$	$\widehat{X}_t = 2.067t + 0.813$	$\widehat{\varepsilon}_t = \widehat{X}_t - x_t$
1	1.5	2.88	-1.38
2	6.9	4.94667	1.95333
3	4.4	7.01333	-2.61333
4	10.1	9.08	1.02
5	13.0	11.1467	1.85333
6	14.1	13.2133	0.886667
7	15.4	15.28	0.12
8	14.6	17.3467	-2.74667
9	20.7	19.4133	1.28667
10	21.1	21.48	-0.38
11		23.5467	
12		25.6133	

Para ver hasta que punto el modelo elegido se ajusta a los datos de la serie se estudian las diferencias entre los valores predichos por la función de ajuste  $\widehat{Y}_t$  y valores observados  $Y_t$ , que es el llamado error residual.

$$\widehat{\varepsilon}_t = \widehat{X}_t - X_t.$$

Los parámetros más empleados para este propósito son los siguientes:  
La raíz cuadrada del error cuadrático medio:

$$RECM = \sqrt{\frac{\sum_{t=2}^t (\widehat{X}_t - X_t)^2}{t-1}} = \sqrt{\frac{\sum_{t=2}^t \widehat{\varepsilon}_t^2}{t-1}}.$$



En el ejemplo anterior el error cuadrático medio toma el valor 1.73.

El error absoluto medio se calcula como la media de los valores absolutos de los residuos:

$$EAM = \frac{\sum_{t=2}^t |\widehat{X}_t - X_t|}{t-1} = \frac{\sum_{t=2}^t |\widehat{\varepsilon}_t|}{t-1}$$

Normalmente entre varios modelos posibles para la tendencia, seleccionamos el que dé valores menores para estos parámetros. También se debe comprobar que los residuos no contengan información, es decir que se puedan considerar como una variable de ruido blanco.

### 17.5.1 Tendencia polinómica.

Para ajustar una tendencia polinómica a una serie temporal  $x_t$ ,  $t = 1, 2, \dots, N$ , se impone la condición de que la suma de los cuadrados de los residuos (diferencia entre los valores de la serie y el valor para el mismo valor de  $t$  del polinomio de ajuste sea mínimo. Es decir el problema consiste en estimar valores de los coeficientes del polinomio  $a_0 + a_1t + a_2t^2 + \dots + a_kt^k$  de modo que:

$$\sum_{t=1}^N \varepsilon_t^2 = \sum_{t=1}^N (x_t - \widehat{x}_t)^2 = \sum_{t=1}^N (x_t - a_0 - a_1t - a_2t^2 - \dots - a_kt^k)^2$$

sea mínimo. Derivando con respecto a cada uno de los  $k + 1$  parámetros  $a_i$  e igualando a cero (por la condición de mínimo)

$$\sum_{t=1}^N 2(x_t - a_0 - a_1t - a_2t^2 - \dots - a_kt^k)(-t^i) = 0,$$

se obtienen  $k + 1$  ecuaciones lineales entre estos parámetros:

$$\left(\sum_{t=1}^N t^i\right) a_0 + \left(\sum_{t=1}^N t^{i+1}\right) a_1 + \left(\sum_{t=1}^N t^{i+2}\right) a_2 + \dots + \left(\sum_{t=1}^N t^{i+k}\right) a_k = \sum_{t=1}^N x_t t^i$$

De este sistema lineal de  $k + 1$  ecuaciones pueden despejarse los valores de los  $k + 1$  coeficientes del polinomio de ajuste.

### 17.5.2 Modelos multiplicativos

A veces las series temporales se ajustan mejor a un esquema multiplicativo:

$$X_t = T_t \times \varepsilon'_t$$

En este caso tomando logaritmos podemos aplicar los mismos métodos del modelo aditivo, considerando la serie

$$Y_t = \ln X_t = \ln T_t + \ln \varepsilon'_t.$$

Si ajustamos un modelo polinómico a la serie  $\ln X_t$  conviene que en este caso el ruido blanco se identifique con  $\ln \varepsilon'_t$ . El modelo tomaría la forma

$$Y_t = \ln X_t = \ln T_t + \ln \varepsilon'_t. = a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k + \varepsilon_t.$$

En este caso la expresión para la serie primitiva sería:

$$X_t = e^{a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k + \varepsilon_t}. = e^{a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k} e^{\varepsilon_t}.$$

Tomamos para la tendencia de la serie la expresión:  $T_t = e^{a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k}$ , es decir que la tendencia sería exponencial. y el componente irregular o aleatorio del tipo  $\varepsilon' = e^{\varepsilon_t}$  que debería ser una variable aleatoria lognormal.

## 17.6 Métodos de suavizado

### 17.6.1 Método de las medias móviles

Este método también puede emplearse para representar la tendencia, aunque está se va construyendo punto a punto. Para determinar el valor que tomará la serie suavizada, que es la que nos dará la tendencia en cada punto, se tienen en cuenta los valores de la serie original que son cercanos en el tiempo al que se está calculando y no, como se hacía en los casos detallados en el apartado anterior, todos los valores de la serie. El orden de la serie viene determinado por el número de elementos que intervienen para calcular los valores de cada punto de la serie suavizada.

En concreto se llama media móvil simétrica de orden  $2p + 1$ , que es un número impar, centrada en  $t$  a

$$MM_t(2p + 1) = \frac{X_{t-p} + X_{t-p+1} + \dots + X_t + X_{t+1} + \dots + X_{t+p}}{2p+1}$$

Por ejemplo la media móvil de orden 5 centrada en 4 se obtendría por medio de la expresión

$$MM_4(5) = \frac{X_2 + X_3 + X_4 + X_5 + X_6}{5}$$

Para calcular la media móvil de orden  $2p$ , un número par, se calculan dos medias móviles de orden  $2p$  y luego se halla la media de ambas.

Por ejemplo, para calcular la media móvil de orden 4 centrada en 3 se calcularía primero las medias

$$MM_{2.5}(4) = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

$$MM_{3.5}(4) = \frac{X_2 + X_3 + X_4 + X_5}{4}$$

y luego la media de ambas.

$$MM_3(4) = \frac{\frac{X_1+X_2+X_3+X_4}{4} + \frac{X_2+X_3+X_4+X_5}{4}}{2} = \frac{\frac{1}{2}X_1+X_2+X_3+X_4+\frac{1}{2}X_5}{4}$$

En general una media simétrica de orden par,  $2p$ , centrada en  $t$  se define como

$$MM_t(2p) = \frac{\frac{1}{2}X_{t-p} + \dots + X_{t-1} + X_t + X_{t+1} + \dots + \frac{1}{2}X_{t+p}}{2p}$$

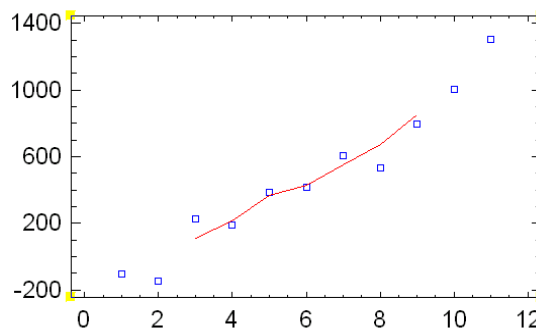
La serie  $MM_t(s)$  obtenida por este procedimiento de las medias móviles es la serie suavizada de orden  $s$  de  $X_t$ . El orden conveniente para calcular las medias móviles depende del propósito perseguido. Un número alto para el orden disminuye el efecto de la componente irregular, ya que queda compensada al realizar el promedio. En cambio, un orden pequeño reflejará mejor los valores primitivos de la serie, por lo que será más fácil detectar los cambios. Es conveniente resaltar que la nueva serie tiene menos valores que la primitiva.

**Ejemplo 83** Dada la serie (cuad\_1 del fichero ejemst.sf3) cuyos valores son -102, -150, 226, 192, 384, 417, 609, 535, 798, 1007, 1306, hallar una serie suavizada usando medias móviles simétricas de orden 5.

Usando la expresión

$$MM_t(5) = \frac{X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2}}{5}$$

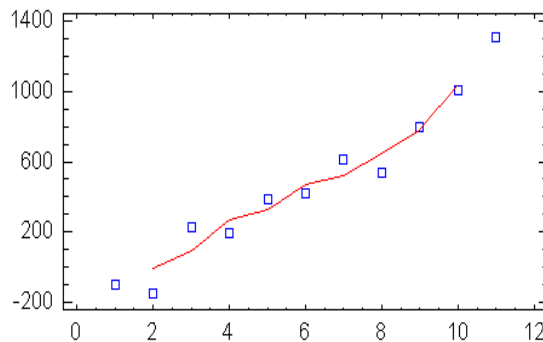
el primer valor que es posible calcular es el que corresponde a  $t = 3$  y el último el que corresponde a  $t = 9$ . La serie primitiva tenía 11 elementos, la suavizada sólo 7 elementos. A continuación mostramos la representación gráfica de la serie suavizada:



La siguiente tabla recoge los valores de las medias móviles de la serie suavizada y las diferencias de estos valores con los primitivos.

$t$	$X_t$	$MM_t(5)$	<i>Diferencias</i>
1	-102.0		
2	-150.0		
3	226.0	110.0	116.0
4	192.0	213.8	-21.8
5	384.0	365.6	18.4
6	417.0	427.4	-10.4
7	609.0	548.6	60.4
8	535.0	673.2	-138.2
9	798.0	851.0	-53.0
10	1007.0		
11	1306.0		
12			

La representación gráfica de la serie móvil de orden 3 presenta el siguiente aspecto:



Por lo general el suavizado aumenta cuando aumenta el número de términos de la serie con el que se calculan las medias móviles.

Se llama media móvil asimétrica de orden  $l$ , asignada al lugar  $t$  a

$$MMA_t(l) = \frac{X_{t-l+1} + X_{t-l+2} + \dots + X_{t-1} + X_t}{l}$$

Por ejemplo la media móvil asimétrica de orden 4 asignada a  $t = 6$  se obtendría por medio de la expresión

$$MMA_6(4) = \frac{X_2 + X_3 + X_4 + X_5}{4}$$

Este tipo de media sirve para predecir valores para el término siguiente. Esto no puede hacerse con las medias móviles centradas, ya que usan información del futuro. La tabla siguiente muestra la previsión para  $\widehat{X}_{t+1}$ , por medio de la media móvil de orden 5 asignada al lugar  $t$  :

$t$	$X_t$	$MMA_t(5)$	$\widehat{X}_t$	<i>Errores</i>
1	-102.0			
2	-150.0			
3	226.0			
4	192.0			
5	384.0	110.0		
6	417.0	213.8	110.0	307
7	609.0	365.6	213.8	395.2
8	535.0	427.4	365.6	169.4
9	798.0	548.6	427.4	370.6
10	1007.0	673.2	548.6	458.4
11	1306.0	851.0	673.2	632.8
12			851.0	

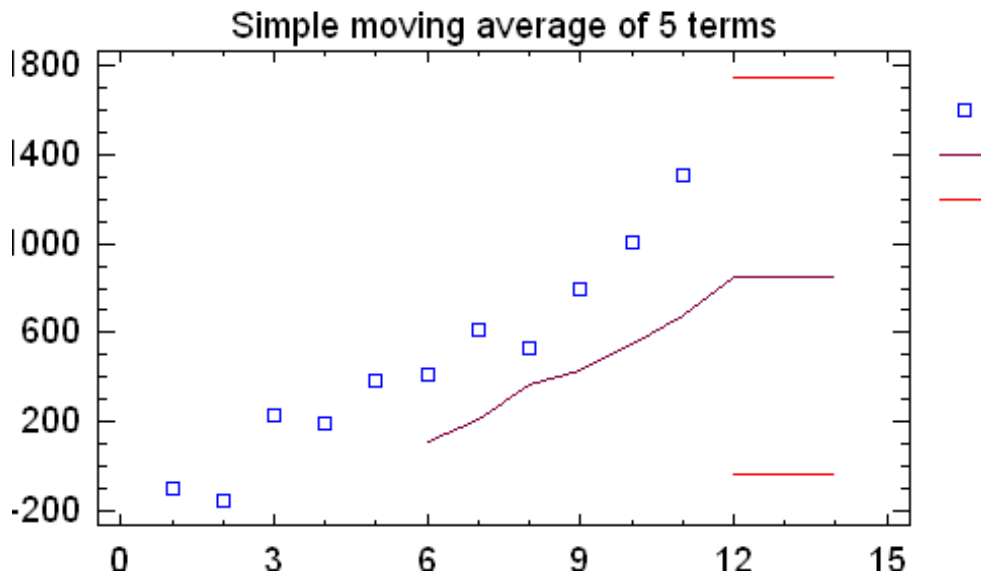
La previsión para  $t = 12$ ,  $\widehat{X}_{12}$ , usando este modelo (medias móviles asimétricas de orden 5) es 851.

En general la previsión para este modelo es:

$$\widehat{X}_{t+m/t} = MMA_t(l), m \geq 1$$

No obstante no se suele usar este procedimiento para predicciones a largo plazo.

La representación gráfica de la serie de medias móviles asimétricas de 5 términos es la de la siguiente figura:



La tabla de errores dada por Statgráphics, para esta serie y la de medias asimétricas de orden 3, es la siguiente:

Modelo	RECM	EAM	EM
5 términos	413.73	388.9	388.9
3 terminos	298.777	268.917	268.917

siendo: RECM (Raiz del Error Cuadrático Medio) la raíz cuadrada del error cuadrático medio , EAM (Error Absoluto Medio) la media de los errores absolutos y EM (Error Medio) la media de los errores. De estos valores se puede deducir que la segunda serie predice mejor que la primera.

### 17.6.2 Método de Alisado exponencial simple

La idea básica de este modelo es suponer que para cada valor del tiempo las observaciones guardan relación con todas las pasadas, pero que esta relación va aminorandose conforme son más lejanas. Esta dependencia se expresa considerando que la serie suavizada (o alisada) es una media *ponderada* de toda los valores históricos de la serie, en lugar de que, como ocurre en el caso de las medias móviles, todos los valores del pasado contribuyan con el mismo peso.

Si llamamos a los elementos de la serie suavizada  $S_t$ , la expresamos como:

$$S_t =$$

$$= (1-a)X_t + (1-a)aX_{t-1} + (1-a)a^2X_{t-2} + (1-a)a^3X_{t-3} + \dots, \quad (17.1)$$

$$0 < a < 1$$

El factor comun  $(1-a)$  es un factor de normalización, que sirve para asegurar que la suma de los pesos es la unidad:

$$\sum_{i=0}^{\infty} (1-a)a^i = (1-a) \sum_{i=0}^{\infty} a^i = (1-a) \frac{1}{1-a} = 1$$

ya que es la suma de los términos de una serie geométrica de razón menor que la unidad.

Como la expresión 17.1 es imposible de evaluar, si como suele ocurrir, contamos con un número finito de valores, se recurre al método iterativo, que se detalla a continuación. Consideramos la expresión 17.1 en un momento anterior y la multiplicamos por  $a$ :

$$\begin{aligned} aS_{t-1} &= \\ &= (1-a)aX_{t-1} + (1-a)a^2X_{t-2} + (1-a)a^3X_{t-3} + (1-a)a^4X_{t-4} + \dots, \quad (17.2) \end{aligned}$$

$$0 < a < 1$$

Restando y despejando  $S_t$ , obtenemos la expresión iterativa:

$$S_t = (1-a)X_t + aS_{t-1}.$$

En realidad esta expresión suele darse en la forma equivalente

$$S_t = \alpha X_t + (1-\alpha)S_{t-1} \text{ con } 0 < \alpha < 1$$

Para comenzar a escribir la serie suavizada se requiere conocer  $S_0$ , ya que

$$S_1 = \alpha X_1 + (1-\alpha)S_0$$

Suele tomarse como valor de partida  $S_0 = X_1$ , o  $S_0$  la media de una parte de los primeros valores disponibles en la serie e incluso la media de todos los valores de la serie si esta no presentará muchas fluctuaciones.

La predicción para los valores siguientes, no incluidos en la serie es el último valor de la serie suavizada:

$$\hat{X}_{t+m/t} = S_t$$

Por supuesto la serie alisada depende del valor de  $\alpha$ . Mientras más grande sea  $\alpha$  menos influencia tendrá el pasado. Se suele seleccionar el valor que suministre un menor error cuadrático medio. Frecuentemente los paquetes estadísticos que tratan las series temporales suministran el valor óptimo para  $\alpha$ .

**Ejemplo 84** Dada la serie temporal (aes del fichero ejemst.sf3): 70, 72, 57, 79, 54, 73, 82, 69, 51, 75, 81.

1. Calcula por el método de Alisado Exponencial Simple, tomando  $\alpha = 0.1$  y  $S_0 = X_1$ , la serie alisada, los errores o residuos, el error cuadrático medio y la previsión para el siguiente periodo de tiempo.
2. Idem tomando  $\alpha = 0.9$  y  $S_0 = X_1$ . ¿Qué método parece mejor?

$t$	$X_t$	$S_t = 0.1X_t + 0.9S_{t-1}$	$\widehat{X}_t$	$\varepsilon_t$	$\varepsilon_t^2$
0		70			
1	70	$0.1 \times 70 + 0.9 \times 70 = 70$			
2	72	$0.1 \times 72 + 0.9 \times 70 = 70.2$	70	2	4
3	57	$0.1 \times 57 + 0.9 \times 70.2 = 68.88$	70.2	-13.2	174.24
4	79	$0.1 \times 79 + 0.9 \times 68.88 = 69.892$	68.88	10.12	102.41
5	54	$0.1 \times 54 + 0.9 \times 69.892 = 68.303$	66.7136	-15.892	252.56
6	73	$0.1 \times 73 + 0.9 \times 68.892 = 68.773$	68.303	4.697	22.06
7	82	$0.1 \times 82 + 0.9 \times 68.773 = 70.096$	68.773	13.227	174.95
8	69	$0.1 \times 69 + 0.9 \times 70.096 = 69.986$	70.096	-1.096	1.20
9	51	$0.1 \times 51 + 0.9 \times 69.986 = 68.087$	69.986	-18.986	360.47
10	75	$0.1 \times 75 + 0.9 \times 68.087 = 68.778$	68.087	6.913	47.79
11	81	$0.1 \times 81 + 0.9 \times 68.778 = 70.0$	68.778	12.222	149.38
12			70	ECM=128.906	

La raíz del error cuadrático medio resulta 11.356

$t$	$X_t$	$S_t = 0.9X_t + 0.1S_{t-1}$	$\widehat{X}_t$	$\varepsilon_t$	$\varepsilon_t^2$
1	70	70			
2	72	$0.9 \times 72 + 0.1 \times 70 = 71.8$	70	2	4
3	57	$0.9 \times 57 + 0.1 \times 71.8 = 58.48$	71.8	-14.8	219.04
4	79	$0.9 \times 79 + 0.1 \times 58.48 = 76.948$	58.48	20.52	421.07
5	54	$0.9 \times 54 + 0.1 \times 76.948 = 56.295$	76.948	-22.948	526.61
6	73	$0.9 \times 73 + 0.1 \times 56.295 = 71.330$	56.295	16.705	279.06
7	82	$0.9 \times 82 + 0.1 \times 71.330 = 80.933$	71.330	10.67	113.85
8	69	$0.9 \times 69 + 0.1 \times 80.933 = 70.193$	80.933	-11.933	142.40
9	51	$0.9 \times 51 + 0.1 \times 70.193 = 52.919$	70.193	-19.193	368.37
10	75	$0.9 \times 75 + 0.1 \times 52.919 = 72.792$	52.919	22.081	487.57
11	81	$0.9 \times 81 + 0.1 \times 72.792 = 80.179$	72.792	8.208	67.37
12			80.179	ECM = 262, 53	



La raíz del error cuadrático medio resulta 16.214

2. La cuestión sobre cuál de los dos valores es preferible para  $\alpha$  nos lleva a preguntarnos: ¿Mejor para qué?. La primera serie consigue una mayor suavización, la segunda serie suavizada toma valores más cercanos a los de la serie primitiva, sin embargo es peor si el propósito es la predicción, ya que suministra un mayor error cuadrático medio que la primera.

Sin otra información se suele preferir la de menor error cuadrático medio, es decir la primera.

### 17.6.3 Método de alisado exponencial doble o de Brown

Es un modelo que adopta la filosofía de adaptar un modelo lineal cambiante a las distintas posiciones de la serie.

$$\widehat{X}_{t+m/t} = a_t + b_t m. + \varepsilon_t.$$

Habrá que estimar los parámetros  $a_t$ ,  $b_t$  para cada valor de  $t$ .

El procedimiento seguido es el siguiente:

$$a_t = 2S_t - S'_t \quad (17.3)$$

$$b_t = \frac{\alpha}{1 - \alpha} (S_t - S'_t) \quad (17.4)$$

Donde  $S_t$  es la primera serie suavizada que se obtiene de la misma forma que en el modelo de alisado exponencial simple y  $S'_t$  es la serie suavizada que se obtiene aplicando a  $S_t$  un nuevo alisado exponencial simple con el mismo valor de  $\alpha$

$$\begin{aligned} S_t &= \alpha X_t + (1 - \alpha)S_{t-1} & 0 < \alpha < 1 \\ S'_t &= \alpha S_t + (1 - \alpha)S'_{t-1} & 0 < \alpha < 1 \end{aligned}$$

Predicciones:

$$\widehat{X}_{t+m/t} = a_t + b_t m.$$

En particular:

$$\widehat{X}_{t+1/t} = a_t + b_t$$

Para iniciar el procedimiento se precisa dar valores a  $S_0$  y  $S'_0$ . Para ello se parte de valores de  $a_0$  y  $b_0$ , sustituirlos en las expresiones 17.3 y 17.4, y despejar  $S_0$  y  $S'_0$ . Hay diversas formas de iniciar el procedimiento. Uno de ellos es tomar  $a_0 = x_1$ ,  $b_0 = 0$ . Otra forma es ajustando una tendencia lineal global (la recta de regresión correspondiente a todos los valores de la serie) La ecuación de la citada recta es

$$X_t = \frac{S_t X_t}{S_t^2} (t - \bar{t}) + \bar{X}$$

que nos servirá para obtener unos valores primitivos

$$a_0 = \frac{S_t X_t}{S_t^2} \bar{t} + \bar{X}, \quad b_0 = \frac{S_t X_t}{S_t^2} \quad (17.5)$$

Sustituyendo en las expresiones 17.3 y 17.4:

$$a_0 = 2S_0 - S'_0 \quad (17.6)$$

$$b_0 = \frac{\alpha}{1 - \alpha} (S_0 - S'_0) \quad (17.7)$$

se pueden obtener los primeros valores de  $S_0$  y  $S'_0$  para comenzar el procedimiento. resolviendo el sistema se obtiene:

$$S_0 = a_0 - \frac{1-\alpha}{\alpha} b_0 \text{ y } S'_0 = a_0 - 2\frac{1-\alpha}{\alpha} b_0$$

**Ejemplo 85** En la siguiente tabla con 12 valores de una serie se aplica el método de alisado exponencial doble o de Brown tomando  $\alpha = 0.2$

Para calcular los valores de inicio para  $S_0$  y  $S'_0$  se realiza un ajuste de regresión lineal a la nube de puntos de la serie. Resulta la recta  $X_t = 52.64 + 1.92t$ . Por tanto  $a_0 = 52.64$  y  $b_0 = 1.92$

A partir de estos valores se calculan  $S_0$  y  $S'_0$  :

$$S_0 = a_0 - \frac{1-\alpha}{\alpha} b_0 = 52.64 - \frac{0.8}{0.2} \times 1.92 = 44.96$$

$$S'_0 = a_0 - 2\frac{1-\alpha}{\alpha} b_0 = 52.64 - 2 \times \frac{0.8}{0.2} \times 1.92 = 37.28.$$

Detallamos los valores que corresponden a  $t = 1$  :

$$\begin{aligned}
 S_1 &= \alpha X_1 + (1 - \alpha)S_0 = 0.2 \times 55 + 0.8 \times 44.96 = 46.968 \\
 S_1 &= \alpha S_1 + (1 - \alpha)S'_0 = 0.2 \times 46.968 + 0.8 \times 37.28 = 39.218 \\
 a_1 &= 2S_1 - S'_1 = 2 \times 46.968 - 39.218 = 54.718 \\
 b_1 &= \frac{\alpha}{1-\alpha}(S_1 - S'_1) = \frac{0.2}{0.8}(46.968 - 39.218) = 1.9376 \\
 \widehat{X}_{2/t=1} &= a_1 + b_1 \times 1 = 54.718 + 1.9376 = 56.656
 \end{aligned}$$

t	X <sub>t</sub>	S <sub>t</sub>	S' <sub>t</sub>	a <sub>t</sub>	b <sub>t</sub>	$\widehat{X}_{t+1}$
0		44,96	37,28	52.64	1.92	
1	55	46,968	39,2176	54,7184	1,9376	54.56
2	58	49,1744	41,2090	57,1398	1,99136	56,656
3	54,5	50,2395	43,0151	57,4640	1,8061	59,1312
4	56	51,39163	44,6904	58,0929	1,6753	59,2701
5	69	55,0133	46,7550	63,2716	2,0646	59,7682
6	68	57,61063	48,9261	66,2952	2,1711	65,3362
7	65,5	59,18853	50,9786	67,3984	2,0525	68,4663
8	62,5	59,85083	52,7530	66,9486	1,7744	69,4509
9	68,5	61,58063	54,5185	68,6427	1,7655	68,7230
10	73,5	63,96453	56,4077	71,52133	1,8892	70,4083
11	75,5	66,27163	58,3805	74,1627	1,9728	73,4104
12	75	68,0173	60,3079	75,7267	1,9274	76,1355
						77,6541

La previsión que da este modelo para el término 13 de la serie es 77.654064. El error cuadrático medio de este ajuste es 4.07

### 17.6.4 Método de Holt

Es similar al de Brown, pero se trabaja con dos constantes de alisado  $\alpha$  y  $\beta$

$$\begin{aligned}
 a_t &= \alpha X_t + (1 - \alpha)(a_{t-1} + b_{t-1}) & 0 < \alpha < 1 \\
 b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} & 0 < \beta < 1
 \end{aligned}$$

Para iniciar el procedimiento se puede partir de la recta de regresión correspondiente a todos los valores de la serie usando los valores dados en las fórmulas 17.5.

Fórmula de predicción:

$$\widehat{X}_{t+m/t} = a_t + b_t m$$

**Ejemplo 86** Se aplica ahora el método de Holt, tomando  $\alpha = 0.2$ ,  $\beta = 0.3$ , a la misma serie del ejemplo 85.

Para calcular los valores de inicio para se realiza un ajuste de regresión lineal. Tomando los valores dados en el ejemplo 85 resultan  $a_0 = 52.64$  y  $b_0 = 1.92$ . Detallamos también los valores que corresponden a  $t = 1$  :

$$a_1 = \alpha X_1 + (1 - \alpha)(a_0 + b_0) = 0.2 \times 55 + 0.8 \times (52.64 + 1.92) = 54.648$$

$$b_1 = \beta(a_1 - a_0) + (1 - \beta)b_0 = 0.3 \times (54.648 - 52.64) + 0.7 \times 1.92 = 1.9464$$

$$\widehat{X}_{2/t=1} = a_1 + b_1 \times 1 = 54.648 + 1.9464 = 56.594$$

t	$X_t$	$a_t$	$b_t$	$\widehat{X}_{t+1}$
0		52.64	1.92	
1	55	54.648	1.9464	54.56
2	58	56.8755	2.0307	56.5944
3	54,5	58.0250	1.7664	58.9063
4	56	59.0331	1.5389	59.7914
5	69	59.0331	2.0446	60.5720
6	68	62.2576	2.2664	64.3021
7	65,5	65.0417	2.1579	67.3081
8	62,5	66.9465	1.7617	69.1045
9	68,5	67.7836	1.6990	69.5452
10	73,5	69.3362	1.8469	71.0352
11	75,5	71.5281	1.9744	73.3750
12	75	73.8000	1.9279	75.7743
13				77.5474

La predicción para el término 13 es 77.5474. La predicción para el término 14, sin conocer el valor de  $X_{13}$  sería:  $\widehat{X}_{t+m/t} = a_t + b_t m = \widehat{X}_{12+2/t=12} = 73.8000 + 1.9279 \times 2 = 77.656$ .

El error cuadrático medio de este ajuste es 3.87, ligeramente inferior al que da el método de Brown con  $\alpha = 0.2$ .

## 17.7 Análisis de la estacionalidad

Uno de los objetivos del análisis de la estacionalidad es retirar esta componente de la serie para hacer comparables entre si datos que pertenezcan a diferentes periodos. Por ejemplo, como todo el mundo sabe la tasa de paro disminuye en España durante el verano. Por lo tanto, si queremos comparar este dato con el de primavera se detecta normalmente una mejoría. ¿Como podríamos saber si esa mejoría es coyuntural, debida exclusivamente a las fechas estivales, o una verdadera mejoría del mercado de trabajo? Incluso podría darse el caso que a pesar de los datos favorables para el empleo durante el verano lo que de verdad se esté produciendo sea una situación de pérdida de empleo. Por este motivo, para poder comparar los datos de verano con los de primavera deberemos separar de la serie de datos la influencia propia del periodo, es decir la componente estacional.

### 17.7.1 El método de la razón a la media móvil

Se aplica en el caso de que el modelo que empleemos sea multiplicativo. Este método mantiene la hipótesis de que la parte estacional correspondiente a cada periodo opera como una proporción constante en ese periodo. Es decir que si  $X_t = T_t \times E_t \times \varepsilon_t$ , se supone que el factor estacional no puede cambiar, por ejemplo, de verano en verano. Si suponemos que el periodo estacional es de un año y los datos estuvieran registrados por mes se verificaría que  $E_t = E_{t+12}$  para cualquier valor de  $t$ .

La aplicación práctica del modelo consta de los siguientes pasos:

a) Se obtiene la serie suavizada  $MM_t(L)$  que corresponde a las medias móviles de orden  $L$ , siendo  $L$  la longitud o amplitud del periodo o intervalo de tiempo a partir del cual se detecta la repetición de una pauta de comportamiento en los valores de la serie. Así si los datos son mensuales, este valor es generalmente 12. Si son trimestrales 4, etc... En estos casos, para obtener la serie suavizada, se utiliza la expresión que corresponde a una media móvil de orden par. Como al obtener esta serie en cada valor han intervenido todos los valores de un periodo, la influencia estacional quedaría eliminada. es decir que podríamos suponer que  $MM_t(L) = \widehat{T}_t$ . Por ejemplo si el periodo es  $L = 3$ ,  $\widehat{T}_4 = MM_4(3) = \frac{X_3+X_4+X_5}{3}$ , y si el periodo es  $L = 4$ ,  $\widehat{T}_5 = MM_5(4) = \frac{\frac{1}{2}X_3+X_4+X_5+X_6+\frac{1}{2}X_7}{4}$

b) Se obtienen los índices brutos de variación estacional, IBVE, dividiendo la serie primitiva por los valores correspondientes para la serie de medias móviles

$$IBVE_t = \frac{X_t}{MM_t(L)} = \frac{T_t \times E_t \times \varepsilon_t}{\widehat{T}_t} = \widehat{E}_t \times \varepsilon_t$$

Estos valores son los llamados índices brutos de variación estacional, IBVE, correspondientes a cada término de la serie.

c) Se calculan los índices IBVE correspondiente a los términos que tienen el mismo número de orden dentro de cada periodo. Así, en una serie de estacionalidad anual que recoja valores mensuales, podemos obtener el índice IBVE correspondiente a los meses de marzo hallando la media de todos los IBVE correspondiente a todos los valores de Marzo .

Si la serie tiene  $N$  elementos distribuidos en  $s$  periodos cada uno de ellos con  $L$  datos,  $s \times L = N$ , obtenemos el índice correspondiente a cada lugar de la serie que tiene el mismo número de orden dentro de cada periodo (por ejemplo, si se supone que los datos fueran mensuales, el IVE de Marzo se calcularía hallando la media de los IBVE correspondientes a ese mes), hallando la media de los índices brutos estacionales, correspondiente a todos los valores de éste.

Así el índice estacional correspondiente a los datos en la posición  $k$  dentro del periodo de longitud  $L$  se calcularía como:

$$IVE_k = \frac{\sum_{i=1}^{s-1} E_{l \times (i-1) + k} \times \widehat{E}_{l \times (i-1) + k}}{s-1}$$

El motivo de que la media se haga con  $s - 1$  valores es que a hallar la media móvil se han perdido datos de la serie.

d) Normalización de los índices:

Consiste en dividir cada uno de estos índices por la media de todos ellos

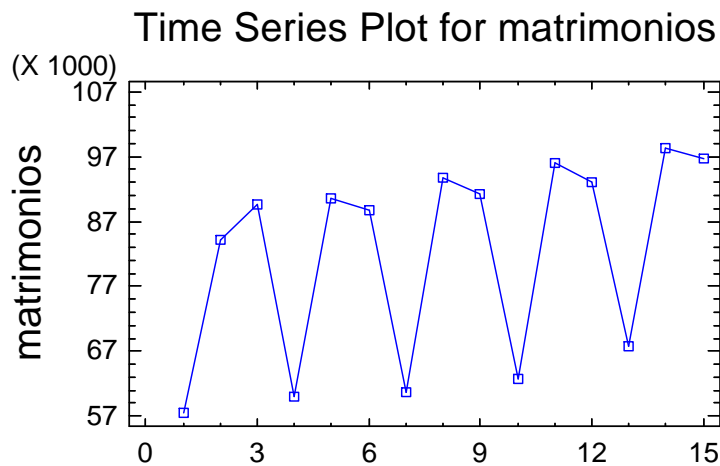
$$IVEN_k = \frac{IVE_k}{\frac{\sum_{i=1}^l IVE}{l}}$$

e) Desestacionalización de la serie. Se dividen los valores de la serie por el correspondiente índices estacionales normalizados.

**Ejemplo 87** La siguiente tabla contiene los matrimonios habidos en España, por cuatrimestres, desde 1968 a 1972 inclusive (matri del fichero ejemst.sf3). Desestacionalizar la serie partiendo de un modelo multiplicativo y hacer una predicción para el número de matrimonios que tendrán lugar en los tres cuatrimestres de 1973.

	1968	1969	1970	1971	1972
1 <sup>er</sup> cuatrim.	57412	59940	60702	62755	67680
2 <sup>o</sup> cuatrim.	84177	90445	93692	95985	98332
3 <sup>o</sup> cuatrim.	89604	88805	91150	92948	96768

La representación gráfica de la serie que aparece en la siguiente figura pone de manifiesto su caracter estacional y una cierta tendencia ascendente.



En la tabla siguiente se recogen los pasos que hay que seguir para aplicar el método de la razón a la media móvil. Detallamos la primera fase del proceso:

$$\widehat{T}_2 = \text{MM}_2(3) = \frac{X_1 + X_2 + X_3}{3} = \frac{57412 + 84177 + 89604}{3} = 77064.03$$

$$IBVE_2 = \frac{X_2}{\widehat{T}_2} = \frac{84177}{77064.03} = \widehat{E}_t \times \varepsilon_t = 1.0923$$

$$IVE_1 = \frac{\sum_{i=1}^4 IBVE_{3 \times (i-1) + 2}}{4} = \frac{IBVE_4 + IBVE_7 + IBVE_{10} + IBVE_{13}}{4} = \frac{0.749284 + 0.748794 + 0.753391 + 0.784059}{4} = 0.75888.$$

El índice correspondiente al segundo cuatrimestre es el único que puede evaluarse con 5 valores

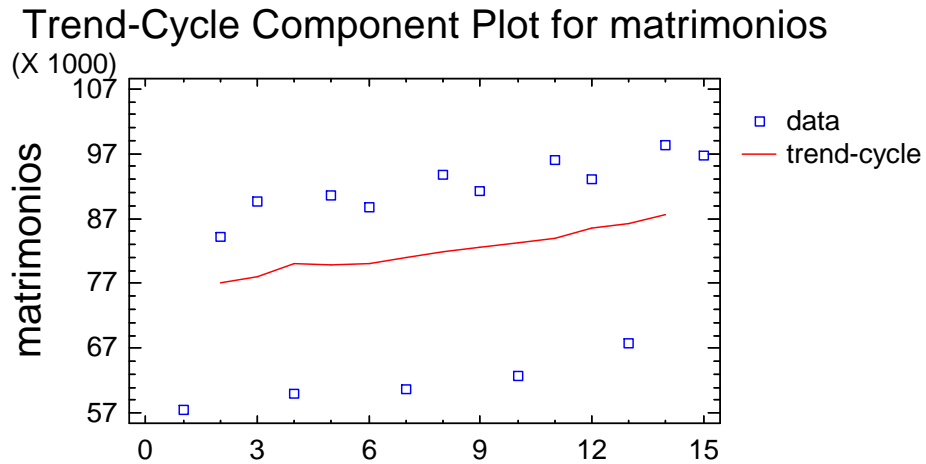
$$IVE_2 = \frac{IBVE_2 + IBVE_5 + IBVE_8 + IBVE_{11} + IBVE_{14}}{5} = \frac{1.0923 + 1.13439 + 1.14471 + 1.1441 + 1.1226}{5} = 1.1276$$

$$IVEN_1 = \frac{IVE_1}{\frac{IVE_1 + IVE_2 + IVE_3}{3}} = \frac{0.7588}{\frac{0.7588 + 1.1276 + 1.1129}{3}} = 0.7590$$

$$\widehat{\varepsilon}_2 = \frac{X_2}{\widehat{T}_2 \times \widehat{E}_2} = \frac{84177}{77064.03 \times 1.1279} = 0.96844$$

$t$	$X_t$	$\widehat{T}_t$ MM $_t(3)$	$IBVE_t$ $\widehat{E}_t \times \varepsilon_t$	IVE	$IVEN$ $\widehat{E}_t$	Residuos $\widehat{\varepsilon}_t$	Serie Desestac. $X_t / IVEN$
1	57412			0.7588	0.7590		75637.4
2	84177	77064.0	1.0923	1.1276	1.1279	0.9684	74634.6
3	89604	77907.0	1.15014	1.1129	1.1131	1.0332	80499.3
4	59940	79996.3	0.749284			0.98714	78967.9
5	90445	79730.0	1.13439			1.0058	80192.1
6	88805	79984.0	1.11028			0.997468	79781.5
7	60702	81066.3	0.748794			0.986499	79971.8
8	93692	81848.0	1.14471			1.01494	83071.0
9	91150	82532.3	1.10442			0.992195	81888.2
10	62755	83296.7	0.753391			0.992555	82676.6
11	95985	83896.0	1.1441			1.0144	85104.0
12	92948	85537.7	1.08663			0.976219	83503.5
13	67680	86320.0	0.784059			1.03296	89165.0
14	98332	87593.3	1.1226			0.995338	87185.0
15	96768						86935.4

Para realizar las predicciones elegimos un modelo teórico para la tendencia. La representación gráfica de la tendencia es:



que parece que se ajusta bien a una recta. Realizando un ajuste lineal global a la serie desestacionalizada se obtiene una recta que vamos a tomar como modelo para la serie de tendencia:

$$T_t = 75554.1 + 813.154t$$

$$X_t = T_t \times E_t$$

Este modelo puede usarse para realizar predicciones para el número de matrimonios registrados en el siguiente año, 1973:

$$\widehat{X}_{16} = \widehat{T}_{16} \times IVEN_{16} = (75554.1 + 813.154 \times 16) \times E_1 = 88565. \times 0.7590 = 67221 \text{ matrimonios en el primer cuatrimestre.}$$

$$\widehat{X}_{17} = \widehat{T}_{17} \times IVEN_{17} = (75554.1 + 813.154 \times 17) \times E_2 = 89378. \times 1.12785 = 100800 \text{ matrimonios en el segundo cuatrimestre.}$$

$$X_{18} = T_{18} \times IVEN_{18} = (75554.1 + 813.154 \times 18) \times E_3 = 90191. \times 1.1131 = 100390 \text{ matrimonios en el tercer cuatrimestre.}$$

### 17.7.2 El método de Holt-Winters

Es similar al de Holt, aunque incluye un triple alisado que le permite tratar la componente estacional. Hay distintas presentaciones de este método. El modelo que detallamos se ajusta a la expresión:  $X_t = T_t \times E_t + \varepsilon_t$ , que no es exactamente aditivo ni multiplicativo y se encuadra dentro de los modelos de *tipo mixto*. En concreto

$$X_t = (a_t + b_t t) E_t + \varepsilon_t.$$



La tendencia,  $a_t + b_t t$ , es lineal y cambiante en el tiempo como en el modelo de Holt. Las expresiones iterativas son casi iguales a las de Holt, salvo que en lugar de usar directamente la serie se emplea su correspondiente valor desestacionalizado. El valor de  $L$  es la longitud de un periodo. En el ejemplo de los matrimonios, sería  $L = 3$ .

$$\begin{aligned} a_t &= \alpha \frac{X_t}{E_{t-L}} + (1 - \alpha)(a_{t-1} + b_{t-1}) & 0 < \alpha < 1 \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} & 0 < \beta < 1 \end{aligned}$$

La componente estacional, que era constante en los distintos periodos en el método de la razón a la media móvil, aquí se considera variable de periodo en periodo.

Las fórmula de recurrencia para esta componente es:

$$E_t = \gamma \frac{X_t}{a_t} + (1 - \gamma)E_{t-L} \quad 0 < \gamma < 1$$

Las predicciones de futuro se realizan con la fórmula:

$$\widehat{X}_{t+m}/t = (a_t + b_t m) E_{t+m-L}$$

Las predicciones de pasado se realizan con la fórmula:

$$\widehat{X}_{t/t-1} = (a_{t-1} + b_{t-1}) E_{t-L}$$

Los valores iniciales para  $a_1$ , y  $b_1$  pueden ser los de la recta de regresión global correspondiente a la serie desestacionalizada. Para los valores iniciales para  $E_1, E_2, \dots, E_L$  puede usarse los *IVEN* que pueden obtenerse por el método de la razón a la media móvil.

A continuación se muestran los resultados obtenidos con el método de Holt-Winters aplicado al problema del ejemplo 87. Los valores iniciales para  $a_t$  y  $b_t$  son los de la recta ajustada a la serie desestacionalizada obtenida en dicho ejemplo,  $T_t = 75554.1 + 813.154t$ , y los valores iniciales de la componente estacional son, igualmente los *IVEN* obtenidos en el mismo ejemplo.

$t$	$X_t$	$a_t$	$b_t$	$E_t$	$\widehat{X}_{t/t-1}$
				0.759	
				1.12785	
		75554.1	813.154	1.1131	
1	57412	76294.69197	805.8977973	0.7831	57962.74579
2	84177	76854.02328	781.2411482	1.115065	86957.90017
3	89604	77921.68857	809.8835627	1.10179	86415.81283
4	59940	78512.60979	787.987328	0.781134424	61654.69414
5	90445	79481.72375	806.0999912	1.117351956	88425.32032
6	88805	80319.10726	809.2283431	1.102176223	88460.32132
7	60702	80786.50766	775.0455489	0.778159767	63372.13573
8	93692	81790.58033	797.948261	1.120167848	91132.96098
9	91150	82599.67812	799.0632143	1.10231012	91027.1125
10	62755	83123.40648	771.5297285	0.775839974	64897.54509
11	95985	84074.24771	789.4608788	1.122318008	93976.41019
12	92948	84809.44756	784.0347759	1.101675389	93546.12482
13	67680	85757.58253	800.444795	0.777176112	66406.84511
14	98332	86663.73419	811.0154814	1.123550047	97145.63284
15	96768	87510.98731	814.6392454	1.102085965	96368.77884

Las predicciones para el siguiente año son:

Primer Cuatrimestre  $(87510.98731 + 814.6392454 \times 1) \times 0.777176112 = 68645$ .

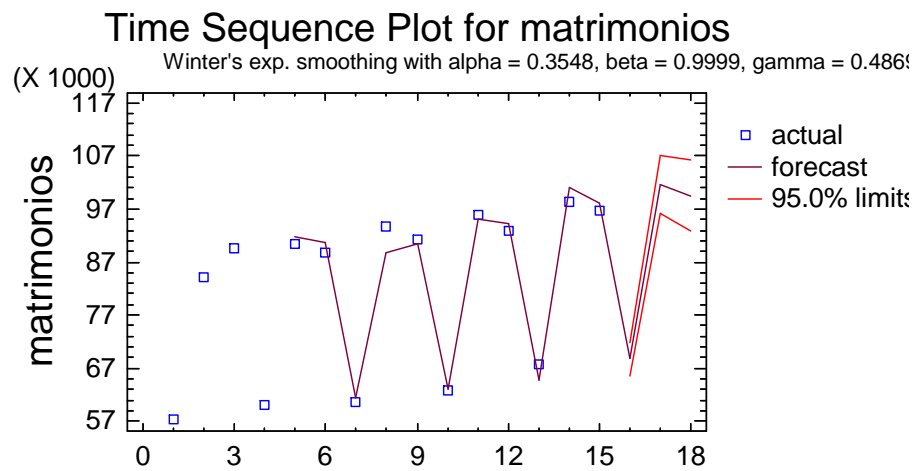
Segundo Cuatrimestre:  $(87510.98731 + 814.6392454 \times 2) \times 1.123550047 = 100500$

tercer Cuatrimestre:  $(87510.98731 + 814.6392454 \times 3) \times 1.102085965 = 99138$ .

El paquete estadístico Statgraphics proporciona un procedimiento que permite calcular los parámetros  $\alpha$ ,  $\beta$  y  $\gamma$  que proporcionan un menor Error Cuadrático Medio.

A continuación mostramos los resultados obtenidos por Statgraphics para esta misma serie, en la que se utilizan parámetros optimizados para obtener el mínimo error cuadrático medio. Los parámetros empleados son:

$$\alpha = 0.3145, \beta = 0.9999, \gamma = 0.4869$$



En la siguiente tabla tenemos los valores de los datos, las estimaciones conseguidas por el método de Holt-Winters y los errores o residuos. En cuanto a las previsiones para el siguiente periodo vienen acompañadas del correspondiente intervalo de confianza al 95%.

$t$	$X_t$	$\widehat{X}_t$	Residuos $\widehat{\varepsilon}_t$
1	57412		
2	84177		
3	89604		
4	59940		
5	90445	92754.2	-2309.2
6	88805	91190.1	-2385.09
7	60702	61976.0	-1274.03
8	93692	88748.0	4943.99
9	91150	89662.1	1487.9
10	62755	62860.0	-105.005
11	95985	94747.9	1237.14
12	92948	94133.6	-1185.56
13	67680	64970.1	2709.89
14	98332	100961.0	2628.66
15	96768	98288.3	-1520.26

Pronósticos sobre el número de matrimonios que se celebrarán durante el año 1973 e intervalos de confianza al 95% para cada uno de estos.

---

(Cuatrimestre.1º, $t = 16$ )	$\widehat{X}_{16} = 68863.4$	(65418.1, 72308.7)
(Cuatrimestre.2º, $t = 17$ )	$\widehat{X}_{17} = 102041$	(96239.9, 107842.0)
(Cuatrimestre.3º, $t = 18$ )	$\widehat{X}_{18} = 99985.5$	(93135.1, 106836.0)

## 17.8 EJERCICIOS PROPUESTOS

**Ejercicio 186** Calcular la media, la varianza, los dos primeros coeficientes de autocovarianza y los dos primeros coeficientes de autocorrelación de la serie cuyos valores son:

$$-0.22, 0.27, -0.37, 0.15, 0.28,$$

$$0.15, 0.06, -0.34, 0.24, 0.02, 0.06$$

Calcular también los intervalos de confianza, al 95%, para estos dos coeficientes y contrasta la hipótesis de que cada uno de ellos sea nulo.

**Ejercicio 187** Las temperaturas medias registradas en una determinada localidad durante los meses de 4 años han sido las siguientes:

MESES	2000	2001	2002	2003
Enero	4	5	5	3
Febrero	10	9	11	12
Marzo	15	15	13	13
Abril	17	17	17	18
Mayo	18	19	18	19
Junio	21	20	22	23
Julio	27	27	27	27
Agosto	27	28	26	28
Septiembre	19	18	19	17
Octubre	12	13	11	10
Noviembre	9	9	8	8
Diciembre	5	6	6	6

Calcular los coeficientes estacionales y predecir la temperatura media en Enero de 2004. Justificar la bondad de la predicción.

**Ejercicio 188** Un laboratorio farmacológico presenta las siguientes cifras de

ventas en las cuatro estaciones de cinco años:

	VENTA EN MILLONES				
Primavera	2.1	2.3	2.2	2.5	2.6
Verano	3.2	3.1	3.6	3.7	3.7
Otoño	2.6	2.9	4.4	4.5	4.9
Invierno	1.4	1.6	1.7	1.8	2.1

Obtener los coeficientes estacionales y hacer una predicción de las ventas en la próxima primavera.

**Ejercicio 189** La tabla adjunta contiene el número de nacimientos habidos en España (en miles) entre los años 1967 a 1971 inclusive, agrupados por cuatrimestres:

	1967	1968	1969	1970	1971
1 <sup>er</sup> cuatrim.	57	59	60	62	67
2 <sup>o</sup> cuatrim.	82	96	107	118	129
3 <sup>o</sup> cuatrim.	80	88	91	92	96

1. Desestacionalizar la serie usando el método de la razón a la media móvil.
2. Aplicando un esquema adecuado, estudiar la tendencia de la serie.
3. Hacer una predicción del número de nacimientos para el segundo cuatrimestre de 1972.

**Ejercicio 190** La tabla adjunta, idéntica a la del ejercicio 189, contiene el número de nacimientos habidos en España (en miles) entre los años 1967 a 1971 inclusive, agrupados por cuatrimestres:

	1967	1968	1969	1970	1971
1 <sup>er</sup> cuatrim.	57	59	60	62	67
2 <sup>o</sup> cuatrim.	82	96	107	118	129
3 <sup>o</sup> cuatrim.	80	88	91	92	96

Aplica el método de Holt-Winters para desestacionalizar la serie y emplea el modelo estimado para hacer una predicción del número de nacimientos para el segundo cuatrimestre de 1972. Tomad para los tres parámetros  $\alpha, \beta, \gamma$  el valor 0.1.

**Ejercicio 191** La tabla siguiente muestra las producciones mensuales medias de maíz, en millones de toneladas para los años 1948-1958.

48	49	50	51	52	53	54	55	56	57	58
50'0	38'5	43'0	44'5	38'9	38'1	32'6	38'7	41'7	41'1	33'8

1. Construir las medias móviles de cuatro años para obtener la tendencia.
2. Representar gráficamente los datos originales y los de las medias móviles en un mismo gráfico.
3. Proponer modelos que parezcan adecuados para representar la tendencia de la serie.

**Ejercicio 192** Hallar un modelo de suavizado por el método de Holt para la serie del ejercicio 191.

**Ejercicio 193** Hallar un modelo de suavizado por el método de Brown para la serie del ejercicio 191 .

**Ejercicio 194** ¿Cuál de los modelos, tratados en los ejercicios 191 y siguientes, sobre las producciones anuales de maíz está mejor ajustado?.

**Ejercicio 195** Se realizó un estudio de seguimiento de la cantidad de insectos de una cierta especie recontados en un espacio natural protegido obteniéndose los siguientes datos.

	Pri.	Ver.	Oto.	Inv.
1999	203	424	82	506
2000	301	501	163	607
2001	342	588	184	669

1. Suponiendo que la serie temporal se ajusta al esquema aditivo, desestacionalizar los valores de la serie.
2. Modelar la serie y predecir el número de insectos que habrá en el verano de 2002.

**Ejercicio 196** Consideraremos la serie temporal de la siguiente tabla, tomada de los datos del Instituto Nacional de estadística dentro de la sección de Hostelería y turismo de la página <http://www.ine.es/inebase/> La tabla registra el número de entradas de personas que visitan nuestro país (datos men-

*suales en miles de personas).*

<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>
1999M02	3728.7	2000M02	3920.1	2001M02	4091.7
1999M03	4613.3	2000M03	4804.1	2001M03	4897.7
1999M04	5627.4	2000M04	6533.2	2001M04	6588
1999M05	6569.8	2000M05	6185.5	2001M05	6453.4
1999M06	6270.6	2000M06	6723.5	2001M06	6972.1
1999M07	9500.9	2000M07	9561	2001M07	9641.5
1999M08	10399.5	2000M08	10325.2	2001M08	10761.3
1999M09	6906.9	2000M09	7688.8	2001M09	7492.8
1999M10	6319.1	2000M10	6230.8	2001M10	6002
1999M11	4227.7	2000M11	4312.6	2001M11	4209.2
1999M12	4300.7	2000M12	4552.6	2001M12	4666.6
2000M01	3624.4	2001M01	3901.9	2002M01	3925.8

<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>
2002M02	4424.8	2003M02	4423.8
2002M03	5785	2003M03	5545.7
2002M04	6039.1	2003M04	6712.6
2002M05	6789.4	2003M05	7378.7
2002M06	7131	2003M06	7510.3
2002M07	9869.8	2003M07	10117
2002M08	12199.3	2003M08	11847.4
2002M09	7629.4	2003M09	7652.8
2002M10	6528.7	2003M10	6791.2
2002M11	4720.1	2003M11	4907.7
2002M12	4982	2003M12	5358.4
2003M01	4279.5	2004M01	4673.9

*Realizar el estudio de la serie usando un modelo clásico.*





## Tema 18

# Series temporales. Modelos Arima.

### 18.1 Procesos estocásticos

Los modelos ARIMA para el estudio de serie temporales consideran que la serie temporal es una muestra o realización de un proceso estocástico.

Definimos los proceso estocástico como una sucesión de variables aleatorias  $\{X_t, t \in N\}$ .

Un proceso estocástico esta perfectamente caracterizado si se conoce la distribución conjunta de cada subconjunto finito de variables del proceso.

Se define como media del proceso a la sucesión o función de  $t$

$$\mu_t = E(X_t)$$

Como momentos de segundo orden respecto de la media se consideran las covarianzas entre cada par de variables aleatorias del proceso

$$\gamma_{t,s} = cov(X_t, X_s) = E[(X_t - \mu_t)(X_s - \mu_s)]$$

Cuando  $t = s$  obtenemos la varianza de  $X_t$

$$\gamma_{t,t} = var(X_t) = E(X_t - \mu_t)^2$$

En la práctica para caracterizar un proceso estocástico se utilizan los coeficientes de autocorrelación

$$\rho_{t,s} = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sqrt{var(X_t)var(X_s)}}$$

conjuntamente con las varianzas.

Por supuesto esta caracterización es más incompleta que si se realiza mediante las funciones de distribución. Ahora bien, si el proceso es normal, queda perfectamente caracterizado por los momentos de primer y segundo orden.

Cada uno de los elementos  $x_t$  de una serie temporal de  $N$  elementos se interpreta como una muestra de un solo elemento de la variable aleatoria correspondiente  $X_t$ . Lo característico de una serie temporal, al contrario de lo que es más habitual en casi todos los muestreos estadísticos, es que existan relaciones de dependencia entre las distintas variables,  $X_t$ , del proceso. Los modelos ARIMA estudian esta dependencia por medio de los coeficientes de autocorrelación total, definidos en el tema anterior y los coeficientes de autocorrelación parcial que definiremos en la siguiente sección

## 18.2 Estacionaridad Funciones de autocorrelación

Un proceso estocástico se dice **estacionario en sentido estricto** cuando la función de distribución conjunta de un subconjunto cualquiera de  $n$  variables no varía cuando todas las variables de la anterior distribución se desplazan  $m$  periodos de tiempo

$$F(X_{t_1}, X_{t_2}, \dots, X_{t_n}) = F(X_{t_1+m}, X_{t_2+m}, \dots, X_{t_n+m}), \forall n, \forall m \in \mathbb{N}$$

Un proceso estocástico se dice **estacionario en sentido amplio** si la media es constante para todo  $t$ , la varianza es constante y finita y la covarianza entre un par de variables solo depende del intervalo de tiempo  $m$  transcurrido entre ellas

Es decir:

$$\mu_t = \mu \quad \forall t$$

$$\gamma_{t,t} = \sigma^2 < \infty, \forall t$$

$$\gamma_{t,s} = \text{cov}(X_t, X_s) = \gamma_{t+m,s+m} = \text{cov}(X_{t+m}, X_{s+m}) = \gamma_m$$

Si la media de la serie y su varianza fueran constantes:

$$E(X_t) = \mu, \text{ y } E(X_t - \mu)^2 = \sigma^2$$

sus valores podrían estimarse con  $\bar{x}$  y  $S_0^2$  respectivamente. Entonces la expresión para la autovarianza teórica tomaría la forma

$$\gamma_m = E(X_t - \mu)(X_{t+m} - \mu)$$

que podría estimarse con  $S_m$ .

En este caso el *coeficiente de autocorrelación teórico* de orden  $m$ ,

$$\rho_m = \frac{\gamma_m}{\sigma^2}$$

podría estimarse por medio de  $r_m$ .

La definición de estos estimadores está dada en 17.3.

Un proceso estacionario en sentido estricto será estacionario en sentido amplio, siempre que la varianza sea finita. El recíproco no es cierto. No obstante, si las variables son normales, la estacionariedad en sentido amplio implica la estacionariedad en sentido estricto.

Desde el punto de vista intuitivo, la estacionariedad significa que las propiedades estadísticas de la serie permanecen invariante a lo largo del tiempo.

Puesto que la media y la varianza permanecen invariantes a lo largo del tiempo parece razonable estimar la media usando todos los valores muestrales de la serie considerada.

$$\hat{\mu} = \frac{\sum_{t=1}^N x_t}{N}$$

Estudiar series estacionarias simplifica el problema original considerablemente. Pero, ¿son las series que nos interesan, por ejemplo las series económicas, estacionarias? Por lo general no lo son, pero con transformaciones sencillas suelen convertirse en series aproximadamente estacionarias. Por este motivo en lo que sigue estudiaremos procesos que sean estacionarios y también trataremos algunas transformaciones que se emplean para convertir las series que no lo sean en series estacionarias.

En un proceso estacionario las autocorrelaciones quedan definidas por

$$\rho_m = \frac{\gamma_m}{\gamma_0} = \rho_{-m}$$

Además de la condición de estacionariedad, impondremos la condición de **ergodicidad**. Para los efectos del análisis de series temporales esta condición se traduce en que los estimadores de los parámetros, medias, varianzas y coeficientes de autocorrelación son consistentes y por tanto las autocorrelaciones muestrales tiende a aproximarse a la función de autocorrelación teórica, así como también el correlograma muestral tiende a aproximarse al correlograma teórico conforme aumentamos el número de elementos,  $N$ , de la serie, lo que implica que haya una cierta concordancia entre ambos correlogramas. De esta forma el análisis del correlograma muestral es una vía para descifrar la estructura de la serie.

Recordamos que cuando aumenta el valor de  $m$  el número de elementos del numerador decrece. Por eso, en la práctica debe partirse de series temporales con un número de elementos,  $N$ , suficientemente grande y no suelen tomarse

en cuenta coeficientes de autocorrelación para  $m$  mayor que la cuarta parte de la longitud de la serie

### 18.3 La función de autocorrelación parcial.

Se define el **coeficiente de correlación parcial** de orden  $m$  como el coeficiente de correlación entre las variables que resultan de eliminar previamente de  $X_t$  y  $X_{t+m}$  el efecto de las variables intermedias  $X_{t+1}, X_{t+2}, \dots, X_{t+m-1}$ . Si consideramos desviaciones respecto de la media,  $Z_t = X_t - \mu_t$ , se sabe que estos coeficientes pueden calcularse ajustando las familias de regresiones

$$\begin{aligned} Z_{t+m} &= \alpha_{11}Z_{t+m-1} + \varepsilon_1 \\ Z_{t+m} &= \alpha_{21}Z_{t+m-1} + \alpha_{22}Z_{t+m-2} + \varepsilon_2 \\ Z_{t+m} &= \alpha_{31}Z_{t+m-1} + \alpha_{32}Z_{t+m-2} + \alpha_{33}Z_{t+m-3} + \varepsilon_3 \\ &\dots\dots\dots \\ &\dots\dots\dots \\ Z_{t+m} &= \alpha_{m1}Z_{t+m-1} + \alpha_{m2}Z_{t+m-2} + \dots + \alpha_{mm}Z_t + \varepsilon_m \end{aligned}$$

donde  $\varepsilon_i$  recoge la parte  $Z_{t+m}$  no explicada por  $Z_{t+m-1}, Z_{t+m-2}, \dots, Z_{t+m-i}$ . Los coeficientes de correlación parcial de orden  $i$  son precisamente  $\alpha_{ii}$  de cada uno de estos ajustes.

En la práctica estos coeficientes suelen obtenerse por un procedimiento iterativo debido a Durbin (1960). **La función de autocorrelación parcial** hace corresponder a los valores de  $t = 1, 2, 3, \dots$  sus correspondientes coeficientes de correlación parcial  $\alpha_{tt}$ . Su representación gráfica es el **Correlograma de Autocorrelación Parcial**.

### 18.4 Procesos lineales

Los procesos lineales que se usan como modelos para el estudio de series temporales estacionarias son de varios tipos:

A) Procesos autorregresivos de orden  $p$  (abreviadamente  $AR(p)$ )

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t$$

B) Procesos de medias móviles de orden  $q$  (abreviadamente  $MA(q)$ )

$$X_t = \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

C) Procesos  $ARMA(p, q)$

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Es conveniente definir el operador retardo:

$$BX_t = X_{t-1}, B^2X_t = X_{t-2}, \dots, B^mX_t = X_{t-m}$$

Con esta notación los modelos anteriores pueden expresarse, *suponiendo que la media es nula*, de la forma siguiente :

*AR(p)* :

$$(1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p)X_t + \varepsilon_t = \Phi_p(B)X_t = \varepsilon_t$$

*MA(q)* :

$$\begin{aligned} X_t &= -\theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \dots - \theta_q\varepsilon_{t-q} + \varepsilon_t = \\ &= (1 - \theta_1B - \theta_2B^2 - \dots - \theta_qB^q)\varepsilon_t = \Theta_q(B)\varepsilon_t \end{aligned}$$

*ARMA(p, q)* :

$$(1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p)X_t = (1 - \theta_1B - \theta_2B^2 - \dots - \theta_qB^q)\varepsilon_t$$

$$\Phi_p(B)X_t = \Theta_q(B)\varepsilon_t$$

En este tema nos limitaremos al estudio teórico detallado del modelo *AR(1)*. No obstante, los modelos más complicados participan de los mismos aspectos teóricos y de las mismas fases de análisis (Identificación del modelo, Estimación de los parámetros, Validación, Predicción) que el modelo *AR(1)*, distanciándose desde luego en las dificultades de cálculo. Confiamos que este estudio teórico particular sea lo suficientemente ilustrativo para acometer el análisis de los modelos más complicados, al menos desde un punto de vista exclusivamente práctico, contando con el auxilio de los procedimientos para series temporales de Statgraphics u otro paquete estadístico que trate los modelos *ARIMA*. Al final del tema ilustraremos con sendos ejemplos el modelo *AR(1)* y un modelo más complejo de serie no estacionaria del tipo *ARIMA(p, d, q)*.

## 18.5 Estudio teórico del modelo AR(1)

En este caso el modelo es

$$X_t = c + \phi X_{t-1} + \varepsilon_t$$

donde  $\varepsilon_t$  es una variable de ruido blanco e independiente de los valores del pasado de la serie.

Partiendo de un valor inicial  $X_0$ , y sustituyendo sucesivamente obtenemos

$$X_1 = c + \phi X_0 + \varepsilon_1$$

$$X_2 = c + \phi(c + \phi X_0 + \varepsilon_1) + \varepsilon_2 = c + \phi c + \phi^2 X_0 + \phi \varepsilon_1 + \varepsilon_2$$

$$X_3 = c + \phi(c + \phi c + \phi^2 X_0 + \phi \varepsilon_1 + \varepsilon_2) + \varepsilon_3 = c + \phi c + \phi^2 c + \phi^3 X_0 + \phi^2 \varepsilon_1 + \phi \varepsilon_2 + \varepsilon_3 = c(1 + \phi + \phi^2) + \phi^3 X_0 + \phi^2 \varepsilon_1 + \phi \varepsilon_2 + \varepsilon_3$$

...

resultando que

$$X_n = c \sum_{i=1}^n \phi^{i-1} + \phi^n X_0 + \sum_{i=1}^n \phi^{n-i} \varepsilon_i$$

y por lo tanto

$$E[X_n] = c \sum_{i=1}^n \phi^{i-1} + \phi^n E[X_0] + E\left[\sum_{i=1}^n \phi^{n-i} \varepsilon_i\right] = c \sum_{i=1}^n \phi^{i-1} + \phi^n E[X_0]$$

Para que el proceso sea estacionario hace falta que  $|\phi| < 1$ , ya que de otra forma  $\lim_{n \rightarrow \infty} E[X_n]$  tendería a infinito en valor absoluto, con lo que el proceso no se mantendría con media constante.

En cambio, si se supone que el proceso ha comenzado en  $-\infty$  y es estacionario ( $|\phi| < 1$ )

$$\begin{aligned} E[X_n] &= \lim_{n \rightarrow \infty} \left[ c \left[ \sum_{i=1}^n (\phi^{i-1}) \right] + \phi^n E[X_0] \right] = \\ &= \lim_{n \rightarrow \infty} \left( c \left[ \frac{\phi^n - 1}{\phi - 1} \right] + \phi^n E[X_0] \right) = \frac{c}{1 - \phi} \end{aligned}$$

Así que

$$\mu = \frac{c}{1 - \phi}$$

cuando el proceso alcanza la estacionaridad. en media.

En cuanto a la varianza de un proceso  $AR(1)$

$$var[X_t] = var[c + \phi X_{t-1} + \varepsilon_t] = \phi^2 var[X_{t-1}] + var[\varepsilon_t]$$

y si el proceso ha de ser estacionario en varianza  $var[X_t]$  y  $var[X_{t-1}]$  toman el mismo valor.

Despejando se obtiene

$$var[X_t] = \frac{var[\varepsilon_t]}{1 - \phi^2} = \frac{\sigma_\varepsilon^2}{1 - \phi^2}$$

Deducimos ahora la función de autocorrelación. Expresamos el proceso por medio de las desviaciones respecto de su media:  $z_t = X_t - \mu$

$$X_t = c + \phi X_{t-1} + \varepsilon_t; \quad z_t + \mu = \mu(1 - \phi) + \phi(z_{t-1} + \mu) + \varepsilon_t = \mu + \phi z_{t-1} + \varepsilon_t,$$

$$z_t = \phi z_{t-1} + \varepsilon_t$$

$$\gamma_m = E[z_t z_{t-m}] = E[(\phi z_{t-1} + \varepsilon_t) z_{t-m}] = \phi E[z_{t-1} z_{t-m}] + E[\varepsilon_t z_{t-m}] = \phi \gamma_{m-1}$$

$$\text{siendo } \gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \phi^2}.$$

Así que obtenemos para la covarianza de orden  $m$

$$\gamma_m = \phi^m \gamma_0 = \phi^m \frac{\sigma_\varepsilon^2}{1 - \phi^2}$$

Por lo tanto la función de autocorrelación viene definida por:

$$\rho_m = \frac{\gamma_m}{\gamma_0} = \frac{\phi^m \gamma_0}{\gamma_0} = \phi^m$$

Los coeficientes de correlación parcial, se deducen de  $z_t = \phi z_{t-1} + \varepsilon_t$  y por tanto  $\alpha_{11} = \phi$  y los restantes coeficientes de autocorrelación parcial son nulos.

Como ejemplo, obtenemos el coeficiente de autocorrelación parcial de segundo orden:

$$\alpha_{22} = \frac{E[z_t, z_{t+2} - \phi z_{t+1}]}{\sigma \sigma_\varepsilon} = \frac{E[z_t, \varepsilon_{t+2}]}{\sigma \sigma_\varepsilon} = 0$$

ya que la variable de ruido esta incorrelacionada con su pasado.

## 18.6 Análisis del modelo AR(1)

### 18.6.1 Identificación del modelo AR(1)

La identificación de la estructura ARMA de un modelo, partiendo de una muestra (una serie temporal concreta) se realiza calculando las funciones de autocorrelaciones total y parcial muestrales y comparándolas con las del modelo teórico.

En el caso del modelo AR(1), como ya hemos indicado, la función de autocorrelación total teórica presentaría un decaimiento exponencial hacia 0 si  $\phi$  es positivo. Si fuera negativo este decaimiento exponencial es en valor

absoluto, aunque habría una alternancia de signo. La función de autocorrelación parcial es más clara, ya que contiene un solo coeficiente distinto de cero.

### Test de hipótesis para los coeficientes de autocorrelación total y parcial

Para identificar una serie temporal como generada por este modelo se estudia si la concordancia entre los coeficientes de correlación muestrales y los teóricos es significativa. Naturalmente esto no suele ser demasiado claro, debido a los errores muestrales. Por este motivo, es necesario construir intervalos de confianza para estos coeficientes.

En la práctica se usan los intervalos de confianza empíricos (al 95%) :

$$\left( -1.96\sqrt{\frac{1}{N}(1 + 2\sum_{i=1}^{k-1} r_i^2)}, 1.96\sqrt{\frac{1}{N}(1 + 2\sum_{i=1}^{k-1} r_i^2)} \right)$$

para contrastar la hipótesis de que el coeficiente de autocorrelación total,  $\rho_k$ , es nulo, y

$$\left( -1.96\sqrt{\frac{1}{N}}, 1.96\sqrt{\frac{1}{N}} \right)$$

para contrastar la hipótesis de que un coeficiente de correlación parcial es nulo. Solo se consideran significativos los coeficientes que queden fuera de los intervalos de confianza indicados anteriormente.

### 18.6.2 Estimación del modelo AR(1)

Una vez que se ha identificado el modelo como del tipo  $AR(1)$ , procede estimar los parámetros de la serie:  $\phi$ ,  $\sigma$  y  $c$ . A partir de una muestra del proceso, los valores de la serie temporal, se pueden obtener estimadores para los parámetros de la serie, aplicando el método de máxima verosimilitud

La función de densidad conjunta de las  $N$  variables de la muestra es

$$f(X_1, X_2, \dots, X_N) = f(X_1) f(X_2/X_1) f(X_3/X_1, X_2) \dots f(X_N/X_1, X_2, \dots, X_{N-1})$$

Tomando logaritmos

$$\ln f(X_1, X_2, \dots, X_N) =$$

$$\ln f(X_1) + \ln f(X_2/X_1) + \ln f(X_3/X_1, X_2) + \dots + \ln f(X_N/X_1, X_2, \dots, X_{N-1})$$



Considerando, para simplificar, desviaciones respecto de la media,  $Z_t = X_t - \mu$ ,  $Z_t = \phi Z_{t-1} + \varepsilon_t$ . Como  $X_1$  se distribuye como una normal de media  $\frac{c}{1-\phi} = \mu$  y varianza  $\frac{\sigma^2}{1-\phi^2}$ , la función de densidad correspondiente a  $Z_1$  es

$$f(Z_1; \mu, \sigma) = \frac{1}{\sqrt{2\pi \frac{\sigma^2}{1-\phi^2}}} e^{-\frac{Z_1^2}{2 \frac{\sigma^2}{1-\phi^2}}}$$

El logaritmo de esta función

$$\ln f(Z_1) = -\frac{1}{2} \ln \left( \frac{2\pi\sigma^2}{1-\phi^2} \right) - \frac{Z_1^2}{2} \frac{1-\phi^2}{\sigma^2} = \frac{1}{2} \left( -\ln 2\pi - \ln \frac{\sigma_\varepsilon^2}{(1-\phi^2)} - \frac{1-\phi^2}{\sigma^2} Z_1^2 \right)$$

La distribución condicional de  $f(Z_t/Z_1, Z_2, \dots, Z_{t-1})$ , sigue una distribución normal de media

$$E[Z_t] = E[\phi Z_{t-1} + \varepsilon_t] = \phi Z_{t-1}, \text{ y } \text{var}[Z_t] = \text{var}[\phi Z_{t-1} + \varepsilon_t] = \text{var}[\varepsilon_t] = \sigma^2$$

$$\ln f(Z_t/Z_1, Z_2, \dots, Z_{t-1}) = \frac{1}{2} \left[ -\ln 2\pi - \ln \sigma^2 - \frac{1}{\sigma^2} (Z_t - \phi Z_{t-1})^2 \right]$$

Por lo tanto el doble de la función del logaritmo de la función de verosimilitud es

$$2\ln f(Z_1, Z_2, \dots, Z_N) = \left( -\ln 2\pi - \ln \frac{\sigma^2}{(1-\phi^2)} - \frac{1-\phi^2}{\sigma_\varepsilon^2} Z_1^2 \right) + \sum_{t=2}^N \left[ -\ln 2\pi - \ln \sigma^2 - \frac{(Z_t - \phi Z_{t-1})^2}{\sigma^2} \right]$$

Así que la función que hay que maximizar es:

$$-\ln \frac{2\pi\sigma^2}{(1-\phi^2)} - \frac{(1-\phi^2) Z_1^2}{\sigma^2} - (N-1) \ln 2\pi - (N-1) \ln \sigma^2 - \frac{\sum_{t=2}^N (Z_t - \phi Z_{t-1})^2}{\sigma^2} \quad (18.1)$$

### Enfoque condicionado

Si adoptamos la simplificación de considerar el valor inicial fijo, y lo estimamos, por ejemplo con la media muestral, el procedimiento se simplifica grandemente, ya que solo nos resta maximizar

$$\sum_{t=2}^N \left[ -(N-1) \ln 2\pi - (N-1) \ln \sigma^2 - \frac{1}{\sigma_\varepsilon^2} (z_t - \phi z_{t-1})^2 \right]$$

Para estimar  $\phi$ , sólo tendríamos que hallar la derivada de la expresión  $\sum_{t=2}^N \left[ -\frac{1}{\sigma_\varepsilon^2} (z_t - \phi z_{t-1})^2 \right]$ , e igualarla a cero, ya que el resto de la función no depende de  $\phi$ . Este enfoque se conoce con el nombre de *Enfoque condicionado* (al valor dado al primer valor de la serie).

$$\begin{aligned} & \frac{\partial}{\partial \phi} \sum_{t=2}^N (z_t - \phi z_{t-1})^2 = \\ & = 2\phi \sum_{t=2}^N z_{t-1}^2 - 2 \sum_{t=2}^N z_t z_{t-1} = 0; \\ & \hat{\phi} = \frac{\sum_{t=2}^N z_t z_{t-1}}{\sum_{t=2}^N z_{t-1}^2} \end{aligned}$$

Una vez obtenido un estimador para  $\phi$  obtenemos el estimador de  $\sigma_\varepsilon^2$

$$\frac{\partial}{\partial \sigma} \sum_{t=2}^N \left[ -(N-1) \ln \sigma_\varepsilon^2 - \frac{(z_t - \phi z_{t-1})^2}{\sigma_\varepsilon^2} \right] = -\frac{2\sigma_\varepsilon^2(N-1) - 2 \sum_{t=2}^N (z_t - \phi z_{t-1})^2}{\sigma_\varepsilon^3} = 0$$

obteniéndose como estimador de  $\sigma_\varepsilon^2$

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=2}^N (z_t - \hat{\phi} z_{t-1})^2}{N-1}$$

### Enfoque no condicionado

Si partimos de la función de verosimilitud exacta, para hallar los valores de  $\phi$  y  $\sigma$  que maximizan la función de verosimilitud igualamos a cero sus derivadas con respecto a estos parámetros de la expresión 18.1, obteniéndose el sistema

$$\begin{aligned} & \phi (\sigma_\varepsilon^2 - Z_1^2) + \phi^3 Z_1^2 + \sum_{t=2}^N (-Z_t + \phi Z_{t-1}) Z_{t-1} - \phi^2 \sum_{t=2}^N (-Z_t + \phi Z_{t-1}) Z_{t-1} = 0 \\ & N\sigma_\varepsilon^2 - Z_1^2 + Z_1^2 \phi^2 - \sum_{t=2}^N (-Z_t + \phi Z_{t-1})^2 = 0 \end{aligned}$$

que conduce a una ecuación de tercer grado en  $\phi$

Este procedimiento es conocido con el nombre de *enfoque no condicionado*, porque considera como aleatorias todas las variables que intervienen en la serie temporal.

### 18.6.3 Predicción en el modelo AR(1)

Designamos por  $X_t(l)$  el predictor óptimo para el en el momento  $t+l$  utilizando la información disponible hasta el periodo  $t$ . El criterio de selección de este predictor óptimo es que la varianza del error de predicción

$$E(X_{t+l} - X_t(l))^2$$

sea mínima. Se demuestra que este estimador es la esperanza de  $X_{t+l}$  condicionada a la información recogida hasta el instante  $t$ .

Para evaluar  $X_t(l) = E(X_{t+l}/X_1, X_2, \dots, X_t)$ , se va a expresar  $X_{t+l}$  en función de  $X_t$ . Considerando desviaciones respecto a la media;

$$Z_{t+l} = \phi^l Z_t + \sum_{i=1}^l \phi^{l-i} \varepsilon_{t+i}$$

$$Z_t(l) = E(Z_{t+l}/Z_1, Z_2, \dots, Z_t) = E\left(\phi^l Z_t + \sum_{i=1}^l \phi^{l-i} \varepsilon_{t+i}/Z_1, Z_2, \dots, Z_t\right) = \phi^l Z_t$$

puesto que los ruidos son posteriores al momento  $t$  y su esperanza es nula, ya que que no se tiene ninguna información sobre su valor.

Veamos como puede obtenerse la varianza del error

$$\begin{aligned} E(Z_{t+l} - Z_t(l))^2 &= E\left(\phi^l Z_t + \sum_{i=1}^l \phi^{l-i} \varepsilon_{t+i} - \phi^l Z_t\right)^2 = \\ &= E\left(\sum_{i=1}^l \phi^{l-i} \varepsilon_{t+i}\right)^2 = \sigma_\varepsilon^2 \sum_{i=1}^l \phi^{2(l-i)} \end{aligned}$$

ya que los residuos de los distintos instantes son incorrelados.

Con estos valores se pueden calcular intervalos de confianza para las predicciones. Como se observa la varianza del error se va ampliando conforme se aleja el horizonte  $l$  de la predicción.

#### 18.6.4 Validación

Antes de emplear el modelo estimado para la predicción se debe comprobar que se cumplen las hipótesis del modelo:

- El modelo es estacionario. En este caso basta con  $|\phi| < 1$
- Los parámetros estimados son significativos
- Los residuos son incorrelados
- Puede aceptarse que los residuos tienen media cero y varianza constante. (Siguen una distribución normal de media 0 si se supone la hipótesis de normalidad de los datos).

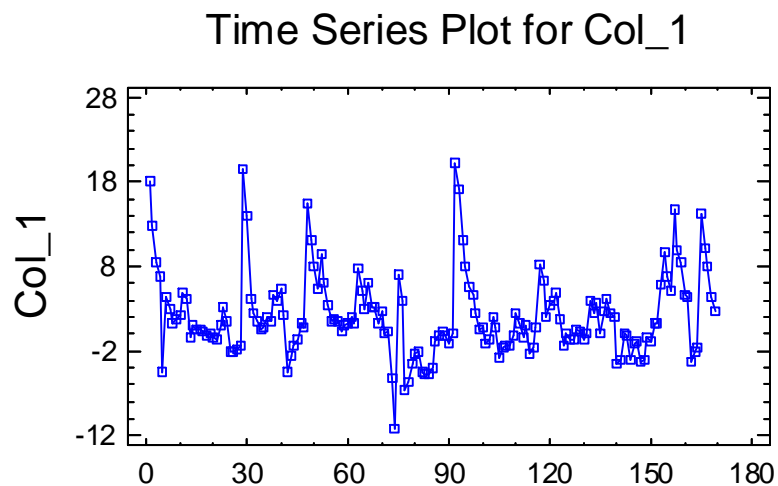
**Ejemplo 88** *Comprobar que el modelo AR(1) es adecuado para la serie dada en la variable ar1 del fichero ejemst.sf3. Estimar los parámetros del modelo y usarlo para predecir los tres siguientes términos de la serie.*

18.19, 12.841, 8.595, 6.858, -4.496, 4.429, 3.068, 1.324, 1.764, 2.221, 4.901, 4.135, -0.461, 1.106, 0.679, 0.441, 0.232, -0.062, -0.003, -0.314, -0.603, 1.028, 3.238, 1.64, -2.11, -1.967, -1.83, -1.444, 19.475, 13.927, 4.185, 2.57, 1.43, 0.49, 0.905, 1.897, 1.433, 4.627, 3.818, 5.345, 2.215, -4.538, -2.453, -1.261, -0.72, 1.365, 0.823, 15.551, 11.023, 7.979, 5.434, 9.346, 6.164, 3.351, 1.459, 1.851, 1.634, 0.213, 1.091, 1.31, 1.965, 1.371, 7.757, 5.037, 3.063, 6.184, 3.253, 3.283, 1.285, 2.838, 0.157, 0.413, -5.246, -11.162, 7.003, 4.039, -6.666, -5.723, -3.403,

-2.243, -2.073, -4.472, -4.632, -4.673, -4.034, -0.784, -0.275, 0.436, -0.096, -1.178, 0.01, 20.21, 17.085, 11.24, 7.9, 5.657, 4.548, 2.409, 0.63, 0.691, -1.102, -0.65, 1.94, 0.789, -2.763, -1.682, -1.259, -1.46, -0.259, 2.43, 1.261, -0.286, 0.994, -2.272, -1.528, 0.887, 8.222, 6.433, 2.031, 3.373, 3.833, 4.807, 1.749, -1.244, 0.088, -0.637, -0.586, 0.457, 0.306, -0.643, 0.194, 3.911, 2.469, 3.722, 0.029, 2.832, 4.189, 2.572, 1.926, -3.479, -3.077, 0.07, -0.221, -3.088, -1.198, -0.894, -3.2, -2.974, -0.313, -0.915, 1.226, 1.282, 5.883, 9.658, 6.782, 5.2, 14.68, 9.938, 8.433, 4.663, 4.367, -3.277, -2.089, -1.603, 14.171, 10.123, 8.03, 4.505, 2.746.

### Fase de identificación.

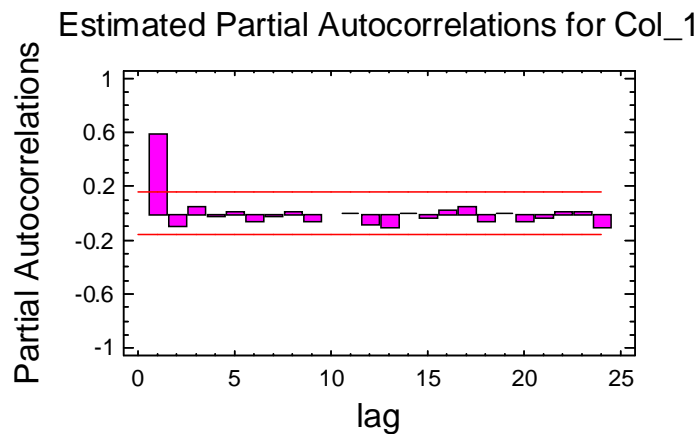
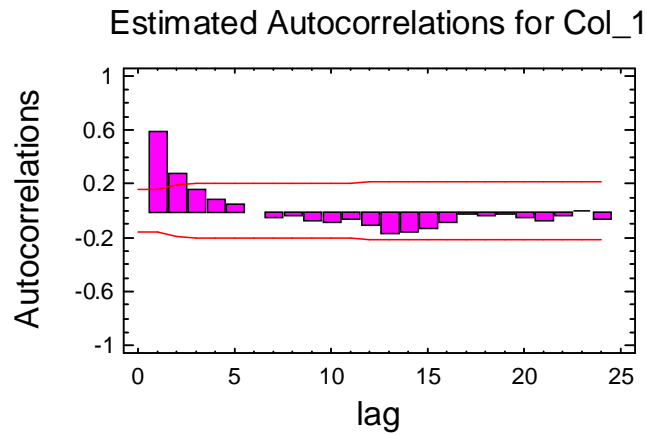
Comenzamos realizando la representación gráfica de la serie



Observamos que es de tendencia constante y que la variación de los valores entre si no parece que sufra cambios importantes a través del tiempo. Por tanto podemos aceptarla razonablemente como estacionaria.

A continuación calculamos sus coeficientes de autocorrelación total y parcial. A continuación se muestran ambos correlogramas. Se observa que los coeficientes de autocorrelación total disminuyen paulatinamente hasta cero, ya que a partir del segundo entran dentro de los intervalos de confianza para 0. El correlograma parcial registra un único coeficiente significativo. Estas

son las características del AR(1). Aceptamos que la serie sigue un modelo AR(1)



### Fase de estimación

Se pasa ahora a calcular los parámetros  $\mu$ ,  $\phi$ ,  $c$ ,  $\sigma_\varepsilon^2$

$\mu$  se estima por medio de  $\bar{X}$ , la media muestral,  $\phi$ , si se usa el enfoque condicionado, puede estimarse, tomando  $z_t = X_t - \bar{X}$  como:

$$\hat{\phi} = \frac{\sum_{t=2}^N z_t z_{t-1}}{\sum_{t=2}^N z_{t-1}^2}$$

Una vez obtenido un estimador para  $\phi$  obtenemos el estimador de  $\sigma_\epsilon^2$

$$\widehat{\sigma_\epsilon^2} = \frac{\sum_{t=2}^N (z_t - \widehat{\phi}z_{t-1})^2}{N-1}$$

De la expresión  $\mu = \frac{c}{1-\phi}$  podemos estimar  $\widehat{c} = \mu(1 - \widehat{\phi}) = \overline{X}(1 - \widehat{\phi})$ .  
Los valores para estas estimaciones obtenidas con Statgraphics son:

$$\widehat{\mu} = 2.3187, \widehat{\phi} = 0.629018, c = 0.860196, \widehat{\sigma_\epsilon} = 3.93834$$

Por lo tanto el modelo ajustado es:

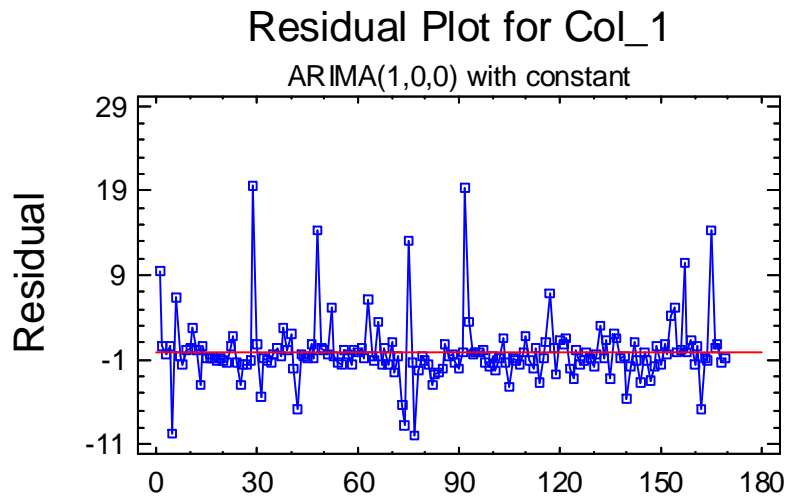
$$X_t = 0.860196 + 0.629018X_{t-1}$$

### Fase de Validación

Debe cumplirse que los residuos sean una variable de ruido blanco.

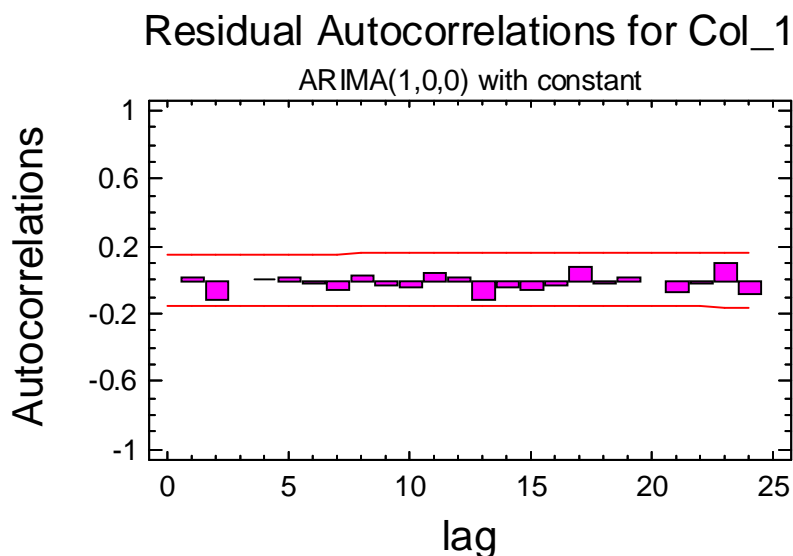
Los recursos son:

a) observar la representación gráfica de los residuales.



que debe ser una serie estacionaria de media cero y varianza constante.

b) La incorrelación de los residuos se observa en el correlograma de los residuos.



Observamos que todos los coeficientes de autocorrelación están contenidos dentro de los intervalos de aceptación de valores nulos. Es decir podemos admitir que los residuos son incorrelados.

También Statgraphics realiza otros test para declarar los residuos como proceso de ruido blanco:

Test de las rachas arriba y abajo, Test de las rachas por arriba y por abajo de la mediana, Test de Box\_Pierce para detectar la excesiva autocorrelación, test para la constancia de la media y de la varianza. En este caso se superan todos estos test.

No se supera en cambio el test de normalidad de los residuos.

Hay que tener en cuenta que el hecho de que no se supere algún test en particular no es motivo de peso para rechazar un modelo, ya que todos los test de hipótesis tienen una probabilidad de error, el nivel de significación. Por ejemplo si se realizan cinco test de Hipótesis con  $\alpha = 0.05$  y se cumplen las respectivas hipótesis nulas, la probabilidad de que todas las hipótesis nulas se declaren como ciertas sería  $0.95^5 = 0.77378$ . Es decir que aproximadamente en el 23% de los casos alguno de los cinco test declarará falsa, sin razón, alguna de las hipótesis nulas. Por lo tanto no es excesivamente raro que esto ocurra.

#### **Fase de Predicción:**

Para obtener los valores predichos para los tres siguientes valores de la serie, usamos el modelo estimado:

$$X_t = 0.860196 + 0.629018X_{t-1}$$

Así obtenemos:

$$X_{170} = 0.860196 + 0.629018 \times X_{169} = 0.860196 + 0.629018 \times 2.746 = 2.5875$$

$$X_{171} = 0.860196 + 0.629018 \times X_{170} = 0.860196 + 0.629018 \times 2.5875 = 2.4878$$

$$X_{172} = 0.860196 + 0.629018 \times X_{171} = 0.860196 + 0.629018 \times 2.4878 = 2.4251$$

Statgraphics da además intervalos de confianza para estos valores:

Period	Forecast	Limit	Limit
170.0	2.58748	-5.18789	10.3629
171.0	2.48777	-6.69792	11.6735
172.0	2.42505	-7.26213	12.1122

## 18.7 Identificación de los modelos ARMA(p,q)

El procedimiento de identificar la serie a partir del Correlograma total y parcial se aplica también al resto de los modelos  $ARMA(p, q)$ . Se precisa conocer como es el correlograma teórico en cada caso. En la siguiente tabla se da un resumen que puede ayudar en la identificación de los distintos modelos

Proceso	Func. de autocorrelación total	Func. de autocorrelación parcial
AR(p)	Muchos coeficientes no nulos que decrecen exponencialmente y/o sinusoidalmente	p primeros coeficientes no nulos y el resto cero
MA(q)	q primeros coeficientes no nulos y el resto cero	Muchos coeficientes no nulos que decrecen exponencialmente y/o sinusoidalmente
ARMA(p,q)	decrecimiento hasta 0 después del retraso q-p	decrecimiento hasta 0 después del retraso p-q

## 18.8 Procesos no estacionarios

### 18.8.1 Modelos ARIMA(p,d,q). Eliminación de la tendencia

Consideremos el proceso  $AR(1)$   $X_t = X_{t-1} + \varepsilon_t$ , que se conoce con el nombre de **Paseo aleatorio**. En este caso  $\phi = 1$ , así que no se cumple la condición de estacionariedad. No obstante si consideramos la serie diferenciada  $\Delta X_t = X_t - BX_t = (1 - B)X_t = X_t - X_{t-1} = \varepsilon_t$ , el proceso  $\Delta X_t$  si que



es estacionario, un proceso de ruido blanco. Esta operación se llama diferenciación, que ha permitido transformar la serie en otra estacionaria. En general si la serie no tiene componente estacional y tiene tendencia polinomial de grado  $n$ , se transforma en estacionaria tras someterla sucesivamente a  $n$  diferenciaciones. Por ejemplo, si la tendencia es un polinomio de segundo grado

$$\begin{aligned}\Delta^2 (at^2 + bt + c) &= (1 - B)^2 (at^2 + bt + c) = \\ &= (1 - B) [(1 - B) (at^2 + bt + c)] = \\ &= (1 - B) \left[ (at^2 + bt + c) - (a(t-1)^2 + b(t-1) + c) \right] = \\ (1 - B) [2at - a + b] &= [2at - a + b] - [2a(t-1) - a + b] = 2a\end{aligned}$$

Esta última serie ya no tendrá tendencia. El proceso de diferenciación transforma una serie no estacionaria en media en otra que si lo es.

Dentro de los modelos no estacionarios se consideran los modelos integrados  $ARIMA(p, d, q)$  que son los que se transforman mediante  $d$  operaciones de diferenciación en el modelo  $ARMA(p, q)$ .

Otra perturbación de la estacionariedad puede ser la falta de homogeneidad de la varianza. En este caso un procedimiento usado, si la varianza guarda con la media una relación aproximada del tipo  $s = k\bar{X}^\alpha$ , es someter a la serie a una transformación de Box-Cox. Es frecuente que la varianza sea proporcional a la media. En este caso la transformación de Box-Cox se reduce a realizar una transformación logarímicamente.

### 18.8.2 Eliminación de la estacionalidad

Del mismo modo que la operación anterior elimina la tendencia si una serie  $X_t$  tiene una componente estacional constante, con un periodo que abarca  $p$  elementos de la serie, la sucesión  $(1 - B^p) X_t$  carecerá de esta componente estacional. Varias diferenciaciones estacionales pueden eliminar la tendencia de la componente estacional. Una vez eliminadas la tendencia y la estacionalidad, la serie resultante será aproximadamente estacionaria, y como tal se podrá intentar estimar por medio de un modelo ARMA.

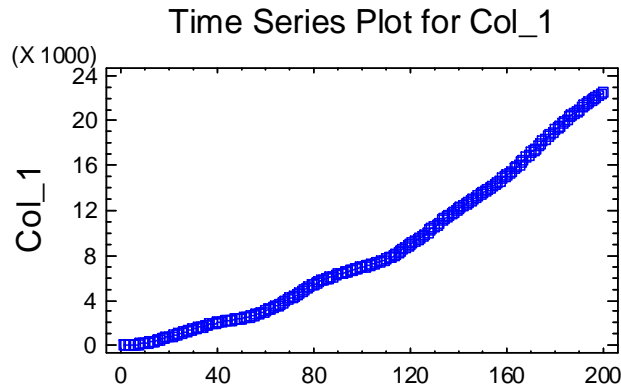
**Ejemplo 89** Realizar el análisis de la serie de la variable arima del fichero `ejemst.sf3` :

2.533221, 9.075923, 19.16261, 33.12857, 52.81117, 76.74696, 103.2003, 132.419, 165.8976, 204.0274, 246.0101, 291.2664, 339.8645, 391.4285, 446.3886, 504.7951, 566.7244, 631.5703, 698.1581, 766.6702, 836.0842, 905.3469, 975.1016, 1044.21, 1111.252, 1177.355, 1242.904, 1307.318, 1371.756, 1436.51, 1499.991, 1561.489, 1620.846, 1679.025, 1736.917, 1794.468, 1851.077, 1906.859, 1960.592, 2009.976, 2055.399, 2097.481, 2136.418, 2173.31, 2210.799, 2250.452, 2290.607,

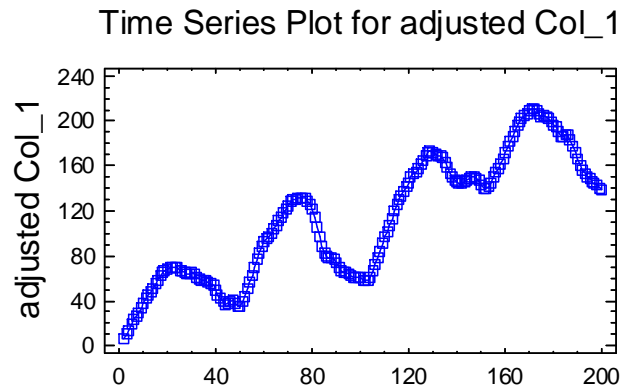
2328.468, 2363.411, 2398.694, 2437.595, 2481.963, 2532.981, 2589.522, 2651.942, 2721.417, 2797.623, 2880.387, 2968.625, 3060.799, 3156.102, 3253.448, 3353.284, 3456.878, 3564.321, 3674.987, 3788.705, 3906.614, 4028.158, 4152.984, 4281.718, 4412.014, 4542.326, 4673.323, 4805.021, 4937.22, 5068.708, 5197.765, 5323.609, 5444.392, 5557.963, 5663.825, 5760.606, 5848.78, 5931.625, 6011.752, 6090.37, 6167.945, 6244.318, 6317.813, 6387.302, 6453.514, 6518.92, 6584.648, 6649.018, 6711.818, 6773.328, 6834.459, 6896.037, 6957.059, 7015.645, 7073.066, 7131.673, 7193.304, 7259.945, 7332.856, 7411.194, 7495.3, 7586.351, 7683.235, 7784.692, 7891.431, 8004.574, 8123.962, 8249.225, 8379.052, 8512.296, 8649.823, 8791.181, 8936.366, 9086.25, 9239.603, 9394.604, 9552.023, 9713.29, 9878.415, 10047.16, 10219.03, 10392.09, 10564.01, 10733.97, 10903.16, 11071.92, 11238.94, 11401.98, 11560.23, 11713.78, 11863.9, 12011.69, 12157.49, 12302.53, 12447.54, 12593.37, 12740.76, 12889.65, 13039.82, 13190.25, 13339.56, 13486.25, 13629.28, 13769.9, 13910.1, 14051.69, 14196.29, 14345.89, 14499.74, 14657.51, 14819.8, 14986.36, 15158.3, 15335.89, 15518.39, 15704.95, 15895.81, 16091.51, 16290.87, 16492.67, 16697.11, 16904.22, 17113.52, 17324.42, 17535.69, 17745.23, 17951.45, 18155.61, 18360.14, 18564.21, 18766.18, 18965.08, 19161.7, 19356.27, 19546.96, 19733.54, 19919.49, 20107.26, 20294.51, 20478.04, 20655.74, 20827.69, 20993.73, 21154.39, 21310.9, 21463.65, 21613.98, 21762.12, 21908.27, 22053.19, 22195.57, 22335.07, 22474.41.

Considerando que los datos son semanales se espera un periodo de longitud 52.

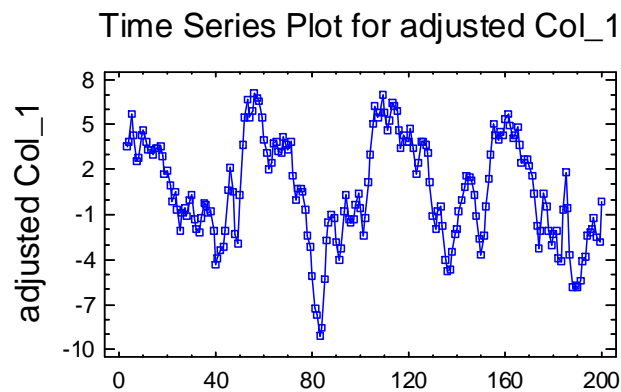
Representando la serie por medio del procedimiento *Descriptive Methods* del análisis de series temporales de Statgraphics se obtiene la gráfica siguiente



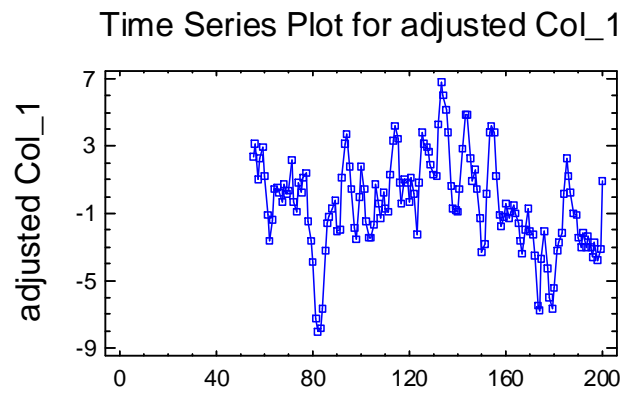
Como la serie tiene tendencia aplicamos una diferenciación, con lo que se obtiene la representación siguiente.



Como sigue siendo ascendente aplicamos una segunda diferenciación, que da lugar a la siguiente representación:

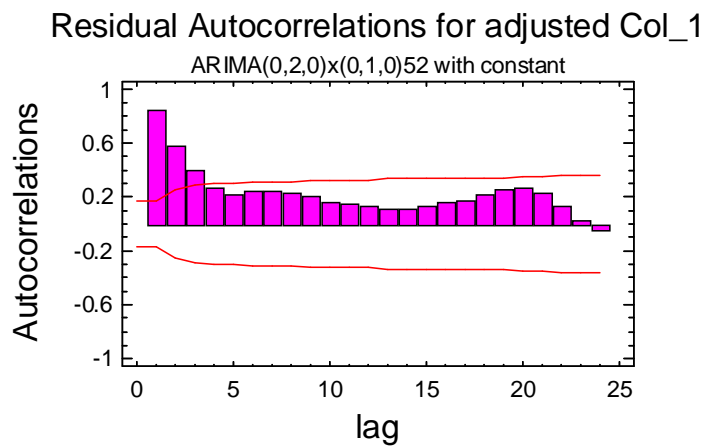


Ahora la serie no tiene tendencia pero percibimos la periodicidad. Realizamos ahora una diferenciación estacional de retardo 52,  $(1 - B^{52})$ , a la serie, dando lugar a la siguiente gráfica. Se observa la pérdida de 54 elementos en el total de la serie debido a las diferenciaciones sucesivas.

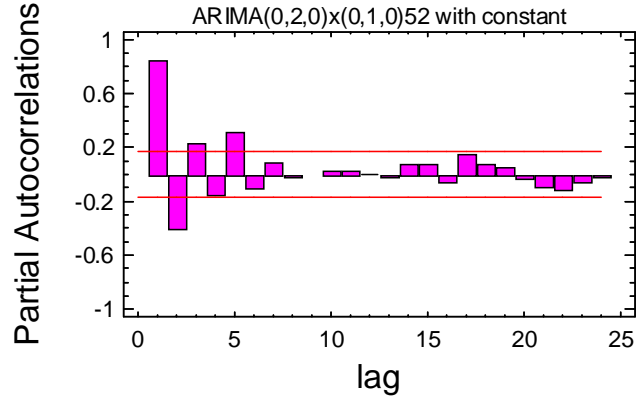


Admitimos que esta serie es ya estacionaria.porque no detectamos suficientes cambios en su media ni en su varianza.

Observamos ahora los correlogramas total y parcial de la última serie diferenciada para proceder a la identificación del modelo:



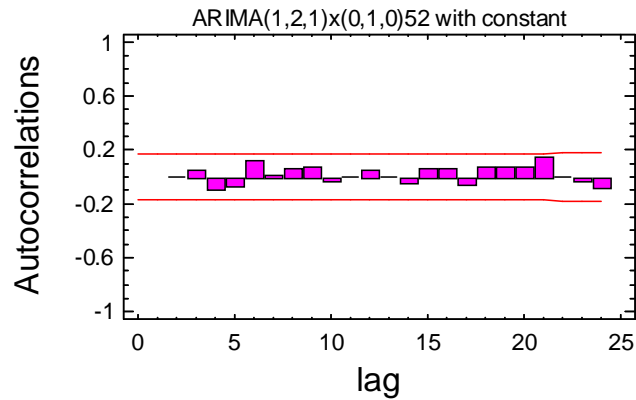
Residual Partial Autocorrelations for adjusted Col\_1



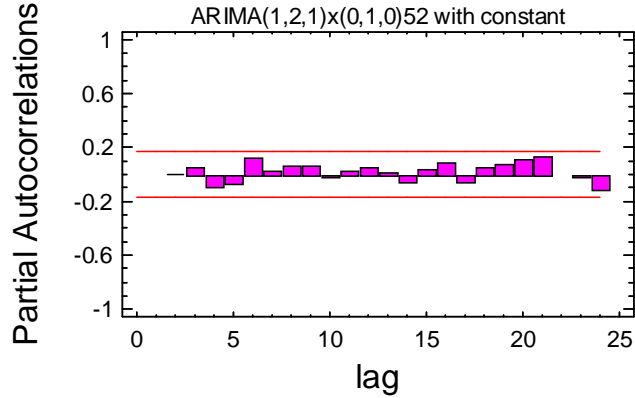
En ambos correlogramas observamos un decrecimiento, por lo que ajustamos un modelo ARMA. Empezamos con un modelo  $ARMA(1, 1)$  con constante.

Los correlogramas de sus residuos son los característicos de un ruido blanco

Residual Autocorrelations for adjusted Col\_1



## Residual Partial Autocorrelations for adjusted Col\_1



el modelo estimado para la serie diferenciada es:

$$X_t = -0.111987 + 0.678289X_{t-1} + 0.348071\varepsilon_{t-1} + \varepsilon_t$$

No obstante se detecta que el término independiente no es significativo.

Eliminando la constante del modelo obtenemos

$$X_t = 0.683488X_{t-1} + 0.828249\varepsilon_{t-1} + \varepsilon_t$$

que resulta tener una desviación típica para el ruido  $\sigma_\varepsilon = 1.24749$ , algo menor que el conseguido con el modelo con constante anterior y que también pasa todos los test de hipótesis al 95%: Test de las rachas arriba y abajo, Test de las rachas por arriba y por abajo de la mediana, Test de Box\_Pierce para detectar la excesiva autocorrelación, test para la constancia de la media y de la varianza. En este caso se superan todos estos test, incluso el de normalidad de los residuos. Por tanto el modelo estimado ya nos parece satisfactorio.

En la tabla siguiente se recogen los 5 últimos valores de la serie y los valores predichos por el modelo

$t$	$x_t$	$\hat{x}_t$	$\varepsilon_t$
196	21908.3	21910.0	-1.68809
197	22053.2	22054.3	-1.12992
198	22195.6	22198.3	-2.74898
199	22335.1	22335.6	-0.534352
200	22474.4	22472.9	1.4665

En la siguiente se registran los valores predichos para los siguientes 5 valores desconocidos de la serie con su correspondiente intervalo de confianza al 95%.

$t$	$\hat{x}_t$	Intervalo al 95%
201	22612.5	(22611.3, 22616.2)
202	22751.2	(22742.2, 22760.2)
203	22887.5	(22867.9, 22907.1)
204	23024.1	(22990.1, 23058.2)
205	23162.8	(23110.6, 23215.0)

No obstante, los resultados del modelo propuesto no son los únicos que se deben considerar. Por ejemplo, usando el modelo  $ARIMA(0, 3, 1)$  se obtiene una estimación del error de 0.998, y se superan todos los test, con lo cual este modelo sería preferible, siendo además más simple. Por lo general, en el análisis de series temporales se suelen proponer varios modelos aceptables, y se selecciona entre ellos el que mejor se adapta a los requisitos que se deseen cumplir. Si preferiéramos minimizar el valor medio de los porcentajes de error, sería mejor elegir el primer método, puesto que el promedio de porcentaje de error es para el primer método 0.015. y para el segundo 0.070.

## 18.9 EJERCICIOS PROPUESTOS

**Ejercicio 197** *Las temperaturas medias registradas en una determinada localidad durante los meses de 4 años han sido las siguientes:*

MESES	2000	2001	2002	2003
Enero	4	5	5	3
Febrero	10	9	11	12
Marzo	15	15	13	13
Abril	17	17	17	18
Mayo	18	19	18	19
Junio	21	20	22	23
Julio	27	27	27	27
Agosto	27	28	26	28
Septiembre	19	18	19	17
Octubre	12	13	11	10
Noviembre	9	9	8	8
Diciembre	5	6	6	6

Los datos están en la variable cuatro del fichero `ejemst.sf3`.

1. Realiza una diferenciación estacional de periodo 12, para conseguir una serie desestacionalizada.
2. Calcula los dos primeros coeficientes de autocorrelación total y parcial.

3. Por medio de algún programa estadístico, haz la representación gráfica de las funciones de autocorrelación total y parcial.
4. ¿Son compatibles estos correlogramas con la identificación del modelo  $AR(1)$  para la serie desestacionalizada?
5. Predecir la temperatura media en Enero de 2004. Justificar la bondad de la predicción.

**Ejercicio 198** Consideraremos la serie temporal de la siguiente tabla, tomada de los datos del Instituto Nacional de estadística dentro de la sección de Hostelería y turismo de la página <http://www.ine.es/inebase/>. La tabla registra el número de entradas de personas que visitan nuestro país (datos mensuales en miles de personas). Los datos están en la variable `totalvisitantes` del fichero `ejemst.sf3`.

<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>
1999M02	3728.7	2000M02	3920.1	2001M02	4091.7
1999M03	4613.3	2000M03	4804.1	2001M03	4897.7
1999M04	5627.4	2000M04	6533.2	2001M04	6588
1999M05	6569.8	2000M05	6185.5	2001M05	6453.4
1999M06	6270.6	2000M06	6723.5	2001M06	6972.1
1999M07	9500.9	2000M07	9561	2001M07	9641.5
1999M08	10399.5	2000M08	10325.2	2001M08	10761.3
1999M09	6906.9	2000M09	7688.8	2001M09	7492.8
1999M10	6319.1	2000M10	6230.8	2001M10	6002
1999M11	4227.7	2000M11	4312.6	2001M11	4209.2
1999M12	4300.7	2000M12	4552.6	2001M12	4666.6
2000M01	3624.4	2001M01	3901.9	2002M01	3925.8

<i>periodo</i>	<i>visitantes</i>	<i>periodo</i>	<i>visitantes</i>
2002M02	4424.8	2003M02	4423.8
2002M03	5785	2003M03	5545.7
2002M04	6039.1	2003M04	6712.6
2002M05	6789.4	2003M05	7378.7
2002M06	7131	2003M06	7510.3
2002M07	9869.8	2003M07	10117
2002M08	12199.3	2003M08	11847.4
2002M09	7629.4	2003M09	7652.8
2002M10	6528.7	2003M10	6791.2
2002M11	4720.1	2003M11	4907.7
2002M12	4982	2003M12	5358.4
2003M01	4279.5	2004M01	4673.9



Realizar el estudio de la serie usando un modelo ARIMA. Esta serie es la misma que la del ejercicio 196.

**Ejercicio 199** Dada la serie temporal siguiente, cuyos datos están recogidos en la variable `plastic` del fichero `ejemst.sf3`, estudia con un paquete estadístico la posibilidad de adaptación a los siguientes modelos: a) ARIMA(2,2,6) sin constante, b) ARIMA(3,2,4) sin constante, c) ARIMA(0,2,6) sin constante, d) ARIMA(0,1,3) con constante, e) ARIMA(0,1,1) sin constante.

5000, 4965, 4496, 4491, 4566, 4585, 4724, 4951, 4917, 4888, 5087, 5082, 5039, 5054, 4940, 4913, 4871, 4901, 4864, 4750, 4856, 4959, 5004, 5415, 5550, 5657, 6010, 6109, 6052, 6391, 6798, 6740, 6778, 7005, 7045, 7279, 7367, 6934, 6506, 6374, 6066, 6102, 6204, 6138, 5938, 5781, 5813, 5811, 5818, 5982, 6132, 6111, 5948, 6056, 6342, 6626, 6591, 6302, 6132, 5837, 5572, 5744, 6005, 6239, 6523, 6652, 6585, 6622, 6754, 6712, 6675, 6882, 7011, 7140, 7197, 7411, 7233, 6958, 6960, 6927, 6814, 6757, 6765, 6870, 6954, 6551, 6022, 5974, 6052, 6033, 6030, 5944, 5543, 5416, 5571, 5571, 5627, 5679, 5455, 5443.

Seleccionar entre los cinco modelos el que creas reúna mejores características, explicando los motivos de esta elección detalla el modelo y halla la previsión para los primeros valores de la serie. Consideramos nulo el error del primer término.