

**RAFAEL ÁLVAREZ CÁCERES**

**ESTADÍSTICA APLICADA  
A LAS CIENCIAS  
DE LA SALUD**



© Rafael Álvarez Cáceres, 2007

Reservados todos los derechos.

«No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del Copyright.»

Ediciones Díaz de Santos  
E-mail: ediciones@diazdesantos.es  
Internet://<http://www.diazdesantos.es>

ISBN: 978-84-7978-823-0  
Depósito legal: M. 21.952-2007

Diseño de cubierta: Ángel Calvete y Rafael Álvarez  
Fotocomposición e impresión: Fernández Ciudad  
Encuadernación: Rústica - Hilo

Impreso en España

---

# AGRADECIMIENTOS

La elaboración de un libro tan complejo y extenso como este ha necesitado mucho tiempo, sin la comprensión y el estímulo de mi mujer y de mis hijos no hubiera podido realizarlo.

A mis alumnos que me han enseñado y me siguen enseñando muchas cosas; en cada curso aprendo algo nuevo.

Al doctor Patricio Alonso Sacristán por haber leído parte del original y por sus interesantes sugerencias.

---

# ÍNDICE

PRÓLOGO .....	XXIII
PREFACIO .....	XXV
GUÍA DE LECTURA .....	XXIX
1. INTRODUCCIÓN A LA ESTADÍSTICA .....	1
1.1. Estadística .....	1
1.2. Historia .....	3
1.3. Población y muestra .....	4
1.4. Estadística aplicada .....	7
1.5. Aplicaciones estadísticas a las ciencias de la salud .....	10
1.6. Variables estadísticas: escalas de medida .....	10
1.7. Variables en estadística aplicada .....	11
1.8. Bibliografía .....	13
2. ESTADÍSTICA DESCRIPTIVA .....	15
2.1. Estadística descriptiva .....	15
2.2. Variables aleatorias .....	16
2.3. Descripción de variables cualitativas .....	17
2.4. Descripción de variables cuantitativas .....	21
2.5. Medidas de tendencia central .....	28
2.6. Medidas de dispersión .....	42
2.7. Medidas de posición: n-tilas .....	51
2.8. Valores atípicos ( <i>outliers</i> ) .....	57
2.9. Momentos respecto al origen .....	59
2.10. Momentos respecto a la media .....	59

2.11. Medidas de forma .....	60
2.12. Estadística descriptiva con SPSS .....	63
2.13. Ejercicios .....	74
2.14. Bibliografía .....	76
<b>3. FUNDAMENTOS MATEMÁTICOS DE LA PROBABILIDAD .....</b>	<b>77</b>
3.1. Sucesos .....	77
3.2. Probabilidad .....	78
3.3. Cuantificación de la probabilidad .....	78
3.4. Técnicas de contar .....	80
3.5. Espacio muestral .....	88
3.6. Álgebra de sucesos .....	88
3.7. Espacio de probabilidad .....	89
3.8. Axiomas de la probabilidad .....	90
3.9. Regla general de la adición .....	92
3.10. Ejercicios .....	93
3.11. Bibliografía .....	94
<b>4. PROBABILIDAD CONDICIONADA. TEOREMA DE BAYES .....</b>	<b>95</b>
4.1. Probabilidad condicionada .....	95
4.2. Teorema de la multiplicación .....	98
4.3. Independencia de sucesos .....	98
4.4. Teorema de Bayes .....	101
4.5. Ejercicios .....	105
4.6. Bibliografía .....	106
<b>5. APLICACIONES DEL CÁLCULO DE PROBABILIDADES A LAS CIENCIAS DE LA SALUD .....</b>	<b>107</b>
5.1. El riesgo. Factores de riesgo y de protección .....	107
5.2. Medidas de riesgo .....	108
5.3. Diferencia de riesgos .....	108
5.4. Riesgo relativo (RR) .....	109
5.5. Reducción relativa del riesgo, RRR .....	113
5.6. Predominio. Razón de predominio (OR) .....	113
5.7. Diagnóstico .....	115
5.8. Normalidad, anormalidad y patología de los datos clínicos .....	116
5.9. Características probabilísticas de las pruebas diagnósticas .....	117
5.10. Sensibilidad y proporción de falsos negativos .....	118
5.11. Especificidad y falsos positivos .....	118
5.12. Valor predictivo positivo .....	119
5.13. Valor predictivo negativo .....	120
5.14. Determinación de los valores de la sensibilidad y de la especificidad .....	121

5.15. Determinación de los valores predictivos: teorema de Bayes ...	122
5.16. Deducción de las expresiones para el cálculo de los valores predictivos mediante el teorema de Bayes .....	127
5.17. Ejercicios .....	131
5.18. Bibliografía .....	132
6. VARIABLE ALEATORIA .....	133
6.1. Variable aleatoria .....	133
6.2. Propiedades .....	134
6.3. Variable aleatoria discreta .....	136
6.4. Función probabilidad de una variable aleatoria discreta .....	136
6.5. Función de distribución acumulativa de una variable aleatoria discreta .....	140
6.6. Función probabilidad en variables aleatorias continuas .....	142
6.7. Propiedades de la función probabilidad en variables aleatorias continuas .....	145
6.8. Función de distribución acumulativa en variable aleatoria continua .....	146
6.9. Valor esperado de una variable aleatoria .....	146
6.10. Propiedades del valor esperado .....	148
6.11. Varianza de una variable aleatoria .....	150
6.12. Covarianza de dos variables aleatorias .....	151
6.13. Propiedades de la varianza .....	152
6.14. Teorema de Tchebychev .....	153
6.15. Ejercicios .....	154
6.16. Bibliografía .....	155
7. DISTRIBUCIÓN DE VARIABLES ALEATORIAS DISCRETAS: BINOMIAL, POISSON, HIPERGEOMÉTRICA, GEOMÉTRICA, MULTINOMIAL. BINOMIAL INVERSA .....	157
7.1. Distribuciones teóricas .....	157
7.2. Ensayos o pruebas de Bernouilli .....	158
7.3. Distribución binomial .....	161
7.4. Distribución multinomial .....	174
7.5. Distribución geométrica .....	176
7.6. Distribución binomial negativa .....	177
7.7. Distribución de Poisson .....	178
7.8. Distribución hipergeométrica .....	181
7.9. Ejercicios .....	184
7.10. Bibliografía .....	185

8. DISTRIBUCIONES DE VARIABLES ALEATORIAS CONTINUAS: NORMAL, T DE STUDENT, CHI-CUADRADO, F DE SNEDECOR .....	187
8.1. Distribución normal .....	187
8.2. Propiedades de la curva normal .....	189
8.3. Tipificación de la variable .....	194
8.4. Función de distribución acumulativa de una variable aleatoria normal .....	196
8.5. Cálculo de probabilidades mediante tablas .....	198
8.6. Teorema central del límite .....	205
8.7. Aproximación de una distribución binomial a una normal .....	205
8.8. Aproximación de la distribución de Poisson a la normal .....	207
8.9. Distribución gamma .....	208
8.10. La distribución Chi cuadrado ( $\chi^2$ ) .....	208
8.11. Distribución T de Student .....	211
8.12. Distribución F de Snedecor .....	214
8.13. Ejercicios .....	217
8.14. Bibliografía .....	218
9. INFERENCIA ESTADÍSTICA: TÉCNICAS DE MUESTREO .....	219
9.1. Poblaciones estadísticas .....	219
9.2. Inferencia estadística .....	221
9.3. Muestras estadísticas .....	222
9.4. Representatividad de la muestra .....	223
9.5. Fracción muestral .....	224
9.6. Técnicas de muestreo no probabilístico .....	225
9.7. Muestreos probabilísticos .....	227
9.8. Muestreo aleatorio simple .....	227
9.9. Muestreo sistemático aleatorio .....	229
9.10. Muestreo estratificado aleatorio .....	232
9.11. Muestreo por conglomerados .....	238
9.12. Ejercicios .....	240
9.13. Bibliografía .....	241
10. INFERENCIA ESTADÍSTICA: ESTIMACIÓN DE PARÁMETROS .....	243
10.1. Estimaciones .....	243
10.2. Variable aleatoria estimada $X$ : fracción muestral .....	246
10.3. Estimadores .....	247
10.4. Estimación de la media aritmética poblacional .....	250
10.5. Predeterminación del tamaño de la muestra en la estimación de medias .....	268
10.6. Estimación de proporciones .....	275
10.7. Predeterminación del tamaño de la muestra para estimar proporciones .....	282

10.8.	Estimación del parámetro de Poisson .....	288
10.8.1.	Predeterminación del tamaño de la muestra .....	290
10.9.	Estimación por el método de máxima verosimilitud .....	292
10.9.1.	El método de máxima verosimilitud .....	292
10.10.	Estimación de parámetros con SPSS .....	295
10.11.	Ejercicios .....	296
10.12.	Bibliografía .....	297
11.	INFERENCIA ESTADÍSTICA: CONTRASTES DE HIPÓTESIS ...	299
11.1.	Introducción .....	299
11.2.	Extrapolación de los resultados de un contraste de hipótesis estadísticas .....	300
11.3.	Estructura de un contraste de hipótesis .....	302
11.4.	Errores aleatorios en un contraste de hipótesis .....	310
11.5.	Relación entre alfa y beta .....	325
11.6.	Predeterminación del tamaño de la muestra .....	327
11.7.	¿Qué significa estadísticamente significativo? .....	333
11.8.	Interpretación errónea de contrastes de hipótesis .....	338
11.9.	Contrastes de hipótesis <i>versus</i> intervalos de confianza .....	346
11.10.	Ejercicios .....	348
11.11.	Bibliografía .....	348
12.	RELACIONES ENTRE VARIABLES .....	349
12.1.	Relaciones entre variables .....	349
12.2.	Asociación entre variables .....	350
12.3.	Tipos de asociación .....	351
12.4.	Estudio simultáneo entre dos o más variables: estadística bivariable y multivariable .....	353
12.5.	Sesgo de confusión .....	354
12.6.	Interacción .....	355
12.7.	Bibliografía .....	356
13.	COMPARACIÓN DE DOS PROPORCIONES: PRUEBAS PARAMÉTRICAS .....	357
13.1.	Introducción .....	357
13.2.	Contraste de hipótesis sobre proporciones: contraste respecto a un valor de referencia .....	359
13.3.	Predeterminación del tamaño de la muestra .....	362
13.4.	Comparación de dos proporciones: datos dependientes .....	365
13.5.	Comparación de dos proporciones: datos independientes .....	367
13.6.	Ejercicios .....	372
13.7.	Bibliografía .....	373



14. PRUEBAS BASADAS EN LA DISTRIBUCIÓN CHI-CUADRA- DO: BONDAD DEL AJUSTE, TABLAS DE CONTINGENCIA. PRUEBA DE FISHER, PRUEBA DE MCNEMAR .....	
14.1. Pruebas basadas en la distribución Chi-cuadrado: bondad del ajuste de datos experimentales a distribuciones teóricas .....	375
14.2. Pruebas de independencia y homogeneidad: asociación entre variables cualitativas .....	382
14.3. Tablas de contingencia .....	385
14.4. Estadística analítica mediante tablas de contingencia $2 \times 2$ : contrastes de hipótesis e intervalos de confianza: pruebas ba- sadas en la distribución chi-cuadrado .....	389
14.5. Análisis de tablas de contingencia $K \times R$ .....	405
14.6. Significación estadística y fuerza de la asociación .....	414
14.7. Análisis del riesgo mediante tablas de contingencia .....	416
14.8. Análisis estratificado .....	418
14.9. Análisis del sesgo de confusión, e interacción entre variables cualitativas .....	418
14.10. Bondad del ajuste y análisis de tablas de contingencia con SPSS .....	432
14.11. Ejercicios .....	442
14.12. Bibliografía .....	444
15. COMPARACIÓN DE DOS MEDIAS, PRUEBAS PARAMÉTRI- CAS .....	445
15.1. Comparación de dos varianzas .....	445
15.2. Comparación de dos medias .....	450
15.3. Comparación de dos medias, datos dependientes o pareados ...	451
15.4. Comparación de dos medias con datos independientes .....	458
15.5. Comparación de dos medias con SPSS .....	469
15.6. Ejercicios .....	474
15.7. Bibliografía .....	475
16. ANÁLISIS DE LA VARIANZA .....	477
16.1. Análisis de la varianza unifactorial .....	477
16.2. Tipos de análisis de la varianza .....	479
16.3. Fundamentos del análisis de la varianza .....	479
16.4. Aplicaciones experimentales del análisis de la varianza .....	484
16.5. Modelo matemático .....	488
16.6. Hipótesis de ANOVA .....	493
16.7. Asunciones del análisis de la varianza .....	498
16.8. Comparación de $K$ varianzas .....	500
16.9. Modelos de análisis de la varianza de una vía .....	502
16.10. Comparaciones múltiples .....	507
16.11. Predeterminación del tamaño de la muestra .....	517

16.12.	Análisis de la varianza unifactorial con SPSS .....	519
16.13.	Ejercicios .....	525
16.14.	Bibliografía .....	526
17.	ANÁLISIS DE LA VARIANZA: BLOQUES, MEDIDAS REPETIDAS .....	527
17.1.	Análisis de la varianza con bloques aleatorizados. Modelos con efectos fijos y aleatorios .....	527
17.2.	Modelo matemático .....	530
17.3.	Variabilidad cuadrática: suma de cuadrados .....	532
17.4.	Cuadrados medios .....	533
17.5.	Hipótesis de ANOVA bloques .....	534
17.6.	Comparaciones múltiples .....	538
17.7.	Estudio de la interacción: prueba de no aditividad de Tukey ..	539
17.8.	Medidas repetidas .....	541
17.9.	ANOVA medidas repetidas con SPSS .....	552
17.10.	Ejercicios .....	558
17.11.	Bibliografía .....	559
18.	CORRELACIÓN .....	561
18.1.	Introducción .....	561
18.2.	Cálculo del coeficiente de correlación lineal de Pearson .....	562
18.3.	Contraste de hipótesis sobre $\rho$ .....	566
18.4.	Intervalos de confianza .....	574
18.5.	Coeficiente de correlación de Spearman .....	575
18.6.	Correlación con SPSS .....	579
18.7.	Ejercicios .....	586
18.8.	Bibliografía .....	588
19.	REGRESIÓN LINEAL SIMPLE .....	589
19.1.	Introducción .....	589
19.2.	Tipos de análisis de regresión .....	591
19.3.	Regresión lineal simple .....	592
19.4.	Coeficientes de regresión estandarizados .....	595
19.5.	Variabilidad cuadrática. Relación entre el coeficiente de regresión y el de correlación .....	595
19.6.	Valores observados, valores esperados y residuos .....	597
19.7.	Modelo matemático .....	599
19.8.	Consistencia de la asociación lineal .....	602
19.9.	Hipótesis en regresión lineal simple .....	604
19.10.	Regresión y análisis de la varianza .....	607
19.11.	Intervalos de confianza de los coeficientes de la recta de regresión .....	609
19.12.	Estimaciones en regresión lineal simple: predicciones .....	611

19.13.	Asunciones del análisis de regresión .....	614
19.14.	Análisis de residuos .....	614
19.15.	Comparación de dos coeficientes de regresión .....	616
19.16.	Análisis de un modelo de regresión simple: errores más frecuentes .....	619
19.17.	Ejercicios .....	621
19.18.	Bibliografía .....	622
20.	REGRESIÓN LINEAL MÚLTIPLE .....	623
20.1.	Regresión lineal múltiple: cálculo de los coeficientes de regresión .....	623
20.2.	Coefficientes de regresión estandarizados .....	630
20.3.	Variabilidad cuadrática .....	631
20.4.	Coefficientes de correlación binarios .....	633
20.5.	Valores observados, valores esperados y residuos .....	634
20.6.	Modelo matemático .....	635
20.7.	Consistencia de la asociación lineal: coeficiente de correlación múltiple, coeficiente de determinación .....	640
20.8.	Hipótesis general en regresión lineal múltiple. Tabla de ANOVA .....	642
20.9.	Intervalos de confianza de los coeficientes de regresión .....	647
20.10.	Estimaciones en regresión lineal múltiple: predicciones .....	648
20.11.	Asunciones del análisis de regresión múltiple .....	650
20.12.	Interacción .....	650
20.13.	Colinealidad .....	654
20.14.	Correlación parcial y semiparcial .....	657
20.15.	Confusión en regresión múltiple .....	659
20.16.	Modelos de regresión con variables cualitativas: variables ficticias o Dummy .....	661
20.17.	Análisis de residuos en regresión múltiple .....	670
20.18.	Construcción de un modelo de regresión múltiple .....	673
20.19.	Análisis de un modelo de regresión múltiple: errores más frecuentes .....	674
20.20.	Análisis de regresión lineal con SPSS .....	676
20.21.	Ejercicios .....	697
20.22.	Bibliografía .....	698
21.	REGRESIÓN LOGÍSTICA .....	699
21.1.	Introducción .....	699
21.2.	Estimación de los coeficientes de regresión logística .....	702
21.3.	Contraste de hipótesis de los coeficientes de regresión logística .....	705
21.4.	Intervalos de confianza de los coeficientes .....	712
21.5.	Interacción .....	713

21.6.	Confusión .....	716
21.7.	VARIABLES FICTICIAS O <i>DUMMY</i> .....	716
21.8.	Cálculo de probabilidades: riesgo .....	722
21.9.	Predominio y razón de predominio (Odds y Odds ratio); probabilidad relativa y riesgo relativo, RR .....	725
21.10.	Residuos en regresión logística .....	734
21.11.	Validez de los modelos de regresión logística .....	735
21.12.	Bondad del ajuste: prueba de Hosmer-Lemeshow .....	739
21.13.	Regresión logística con SPSS .....	741
21.14.	Ejercicios .....	760
21.15.	Bibliografía .....	761
22.	ESTADÍSTICA NO PARAMÉTRICA .....	763
22.1.	Estadística paramétrica y no paramétrica .....	763
22.2.	Pruebas para una sola muestra .....	764
22.2.1.	Prueba binomial .....	764
22.2.2.	Bondad del ajuste: prueba $\chi^2$ .....	770
22.2.3.	Pruebas de Kolmogorov-Smirnov, Kolmogorov-Smirnov-Lilliefors y Shapiro Wilks .....	771
22.2.4.	Pruebas de aleatoriedad: prueba de las rachas .....	778
22.3.	Pruebas no paramétricas con dos variables relacionadas .....	782
22.3.1.	Prueba de los signos para dos variables relacionadas ..	783
22.3.2.	La prueba de Wilcoxon .....	786
22.4.	Pruebas no paramétricas para dos muestras independientes .....	790
22.4.1.	Prueba de la mediana para dos muestras independientes .	791
22.4.2.	La prueba de Mann-Whitney .....	794
22.4.3.	La prueba de Kolmogorov-Smirnov para dos variables independientes .....	798
22.4.4.	La prueba de las rachas de Wald-Wolfowitz para dos variables independientes .....	801
22.4.5.	La prueba de los valores extremos de Moses .....	804
22.5.	Pruebas para $k$ variables relacionadas .....	808
22.5.1.	Prueba de Friedman .....	809
22.5.2.	Coefficiente de concordancia de Kendall .....	815
22.5.3.	La prueba de la $Q$ de Cochran .....	819
22.6.	Pruebas no paramétricas para $k$ variables independientes .....	822
22.6.1.	La prueba de Kruskal Wallis .....	823
22.6.2.	La prueba de la mediana para $k$ variables .....	829
22.7.	Ejercicios .....	831
22.8.	Bibliografía .....	832
23.	FUNCIÓN DE LA ESTADÍSTICA EN EL PROCESO DE INVESTIGACIÓN .....	833
23.1.	Investigación científica .....	833

23.2.	Características principales de una investigación científica .....	835
23.3.	Objetivos e hipótesis .....	835
23.4.	Tipo de estudios .....	837
23.5.	Poblaciones de referencia .....	837
23.6.	Selección de la muestra .....	839
23.7.	Variables: mediciones .....	839
23.8.	Plan estadístico .....	842
23.9.	Validez interna .....	845
23.10.	Validez externa .....	847
23.11.	Bibliografía .....	850
24.	<b>CARACTERÍSTICAS ESTADÍSTICAS DE LOS ESTUDIOS OBSERVACIONALES: SERIES DE CASOS, TRANSVERSALES, COHORTES, CASOS Y CONTROLES .....</b>	<b>851</b>
24.1.	Comunicaciones de un caso .....	851
24.2.	Series de casos .....	851
24.3.	Estudios transversales .....	853
24.4.	Estudios de cohortes .....	857
24.5.	Estudios de casos y controles .....	864
24.6.	Bibliografía .....	868
25.	<b>CARACTERÍSTICAS ESTADÍSTICAS DE LOS ESTUDIOS EXPERIMENTALES Y CUASIEXPERIMENTALES: ENSAYOS CLÍNICOS .....</b>	<b>869</b>
25.1.	Estudios experimentales .....	869
25.2.	Estudios cuasiexperimentales: ensayos clínicos .....	871
25.3.	Fases de los ensayos clínicos .....	871
25.4.	Factores que influyen en la evolución de las enfermedades ...	875
25.5.	Control de los factores que pueden influir en la evolución de las enfermedades .....	876
25.6.	Poblaciones en un ensayo clínico. Selección de los participantes en el ensayo .....	883
25.7.	Estudios controlados con asignación aleatoria .....	885
25.8.	Análisis de los resultados .....	890
25.9.	Análisis estadísticos más utilizados en los ensayos clínicos ...	891
25.10.	Validez de los ensayos clínicos .....	898
25.11.	Bibliografía .....	902
26.	<b>ERRORES MÁS FRECUENTES EN LA APLICACIÓN DE LA ESTADÍSTICA A LAS CIENCIAS DE LA SALUD .....</b>	<b>903</b>
26.1.	Confundir la población diana o poblaciones de interés en investigación con la población estadística del estudio .....	903
26.2.	Realizar conclusiones inferenciales en muestras no aleatorias .	904
26.3.	El extraño caso de las muestras representativas .....	905

26.4. Confundir los errores tipo I y tipo II de los contrastes de hipótesis con sus probabilidades .....	905
26.5. Considerar demostrada la hipótesis nula cuando no se ha rechazado en un contraste de hipótesis .....	906
26.6. Considerar que la significación estadística es el parámetro fundamental para evaluar las conclusiones de un estudio .....	906
26.7. Otro extraño caso: la media seguida de $\pm$ la desviación típica o el error estándar de la media .....	907
26.8. Uso de modelos matemáticos como ecuaciones determinísticas .	908
26.9. El extraño caso de las constantes vitales .....	909
<b>A1. ÁLGEBRA DE BOOLE .....</b>	<b>911</b>
A1.1. Conjuntos .....	911
A1.2. Subconjuntos .....	912
A1.3. Conjunto de las partes de un conjunto .....	913
A1.4. Operaciones con conjuntos .....	913
A1.5. Relaciones entre conjuntos-aplicaciones .....	917
<b>A2. SOLUCIONES A LOS EJERCICIOS .....</b>	<b>921</b>
A2.1. Capítulo 2 .....	921
A2.2. Capítulo 3 .....	923
A2.3. Capítulo 4 .....	926
A2.4. Capítulo 5 .....	928
A2.5. Capítulo 6 .....	930
A2.6. Capítulo 7 .....	931
A2.7. Capítulo 8 .....	933
A2.8. Capítulo 9 .....	936
A2.9. Capítulo 10 .....	937
A2.10. Capítulo 11 .....	941
A2.11. Capítulo 13 .....	943
A2.12. Capítulo 14 .....	945
A2.13. Capítulo 15 .....	948
A2.14. Capítulo 16 .....	950
A2.15. Capítulo 17 .....	952
A2.16. Capítulo 18 .....	956
A2.17. Capítulo 19 .....	958
A2.18. Capítulo 20 .....	960
A2.19. Capítulo 21 .....	961
A2.20. Capítulo 22 .....	964
<b>A3. TABLAS ESTADÍSTICAS .....</b>	<b>967</b>
<b>ÍNDICE ANALÍTICO .....</b>	<b>989</b>

---

# PRÓLOGO

Siempre he admirado a los buenos docentes no sólo por su pasión por la transmisión del saber, sino por su curiosidad intelectual y rigor. Conocí a Rafael Álvarez Cáceres durante mi etapa en la Escuela Nacional de Sanidad (Madrid), donde trabajé primero como responsable del Departamento de Epidemiología y Estadística (1991-1995), del que Rafael Álvarez era profesor, y después como director (1995-2000). Desde el primer momento fui testigo de su capacidad, polivalencia y atención a las necesidades reales de los estudiantes, además de su cuidada técnica pedagógica. Así, siempre se destacó por transmitir de forma rigurosa y al mismo tiempo amena, las ideas fundamentales del análisis estadístico.

Materia en la que pocos profesores son capaces de triunfar, la estadística constituye sin duda uno de los pilares de la ciencia y su aplicación al campo de la salud ha progresado exponencialmente en los últimos años. En la actualidad, es difícil comprender la medicina moderna, la epidemiología y salud pública, la metodología de la investigación e incluso la gestión sanitaria sin sólidos conocimientos de estadística. Esto último tuve ocasión de comprobarlo durante mi experiencia como Director General de Salud Pública del Ministerio de Sanidad y Consumo (2002-2004). En esa etapa profesional pude comprender hasta qué punto la estadística me ayudó al análisis del riesgo asociado a la exposición a los residuos del *Prestige*, cómo caracterizar y predecir la tendencia de la neumonía asiática (SARS)... o cómo tras una leve temporada de gripe (con baja mortalidad invernal), la ola de calor había hecho que durante el verano hubiese aumentado la mortalidad de forma significativa... El análisis estadístico tiene consecuencias prácticas evidentes si queremos proceder con básica «inteligencia sanitaria» para la consiguiente acción. Dicho análisis, tan importante para caracterizar factores de riesgo o de protección, está basado en el cálculo de probabilidades; las características de las pruebas diagnósticas (sensibilidad, especificidad y valores predictivos) son probabilidades; la significación estadística de la que tanto se habla y tanto se desconoce es simplemente una probabilidad; la diferencia estadísticamente significativa entre dos tratamientos es una decisión basada en la probabi-

lidad; el control de calidad se fundamenta en cálculos probabilísticos. Sería interminable la lista de circunstancias fundamentales en las ciencias de la salud que serían muy difíciles de entender sin conocimientos sólidos de estadística.

Aunque la lista de textos de estadística aplicada a las ciencias de la salud es extensa, el que ahora tengo el privilegio de prologar no es simplemente uno más. En primer lugar llama la atención la cantidad de información que contiene; con cientos de ejemplos que permiten aclarar los conceptos estadísticos que a veces son muy complejos. Es destacable el tratamiento de la inferencia distinguiendo muy bien entre significación técnica y estadística, que es lo que realmente diferencia la estadística inferencial matemática y la aplicada. Además, se abordan temas que no se tratan frecuentemente en los libros de estadística aplicada a las ciencias de la salud, como el análisis de la varianza de medidas repetidas, la regresión logística y la regresión múltiple, y todo ello con rigor y profundidad.

Dado lo convencido que estoy de la importancia de la materia y de la capacidad del autor para explicarlo, creo que estamos ante un texto singular con valor añadido con respecto a lo que ya existe en la bibliografía. Deseo que las personas que tienen en sus manos esta obra disfruten con su estudio y tengan el mayor éxito en sus trabajos.

JOSÉ MARÍA MARTÍN MORENO

Catedrático de Medicina Preventiva y Salud Pública de la Facultad de Medicina  
y Hospital Clínico Universitario de Valencia & Member, European Advisory  
Committee on Health Research, World Health Organization



---

# PREFACIO

La aplicación de la estadística a las ciencias de la salud y a las ciencias sociales está aumentando rápidamente en los últimos años. Pocos artículos se publican sin que incluyan estudios estadísticos, al menos descriptivos. La estadística es una herramienta muy útil y poderosa para describir y analizar datos, también como apoyo a la toma de decisiones. Debido a su rápido desarrollo, no ha sido todavía debidamente implementada a las técnicas de investigación propias de cada disciplina.

La estadística aplicada tiene grandes diferencias conceptuales respecto a la estadística matemática, aunque sus fundamentos son los mismos. En estadística matemática se trabaja con números que no tienen errores de medida, mientras que en estadística aplicada las poblaciones de números que son los valores de las variables se obtienen después de haber realizado observaciones y medidas; debido a ello, si las mediciones no pueden ser exactas, lo que ocurre en la mayoría de las circunstancias, habrá que tener en cuenta, en los cálculos estadísticos, los errores de medida; sin embargo esto es raro que se haga, y una vez obtenidos los resultados de investigación se tratan como si procedieran de poblaciones de números. Las primeras ciencias a las que se empezaron a aplicar técnicas estadísticas fueron a la física, la química, y a sus aplicaciones tecnológicas: la ingeniería. En general, las mediciones que se realizan en estas disciplinas tienen pocos errores y, además, la mayoría de las variables tienen variabilidades pequeñas, por eso el éxito en la aplicación de la estadística ha sido enorme. En la actualidad, no se podría entender la física moderna sin el uso de la estadística; teorías como la mecánica estadística y la mecánica cuántica no sólo están basadas en la estadística, son teorías estadísticas con muy buenos resultados en la aplicación práctica. Mediante la estimación de parámetros, los contrastes de hipótesis y el control de calidad aplicados a estas disciplinas se suelen obtener magníficos resultados debido a la pequeña varianza de la mayoría de las variables a las que se aplican.

La aplicación de la estadística a las ciencias de la salud y sociales, se ha realizado y se realiza sin tener en cuenta, en muchos casos, que las mediciones no se

pueden hacer con mucha exactitud y que las variables en muchos casos tienen varianzas relativamente grandes. Por eso cuando las mediciones pueden hacerse con cierta exactitud y las varianzas son pequeñas se obtienen grandes éxitos, como ocurre, en general, en bioquímica, genética y fisiología; sin embargo en medicina clínica, administración sanitaria y ciencias sociales se cometen importantes errores.

Los errores en la estadística aplicada están muy generalizados, y no sólo debido a la aplicación de métodos complejos, es muy frecuente aplicar intervalos de confianza y realizar contrastes estadísticos con muestras no probabilísticas, lo cual no tiene ningún fundamento y las tomas de decisiones realizadas de esta manera no tienen la precisión ni el rigor que parecen tener. Un ejemplo muy conocido es el de los estudios de casos y controles, muy útiles en algunas ocasiones; sin embargo, en la mayoría de los casos los datos no se obtienen mediante muestreos probabilísticos, pero se estudian como tales.

La toma de decisiones basadas en la significación estadística parece muy cómoda y además no hay que pensar mucho. Se coloca un nivel de decisión, frecuentemente 0,05, y si la probabilidad obtenida en el contraste de hipótesis es menor se rechaza la hipótesis nula, en caso contrario no se rechaza. El problema es que en estadística aplicada la significación estadística es un parámetro secundario en la toma de decisiones. El parámetro principal es la significación técnica, es decir, la importancia clínica, psicológica, sociológica o fisiológica del valor calculado de los parámetros, y sólo si estos son relevantes tiene sentido preguntarse la probabilidad de haber obtenido los resultados por azar, que es lo único que contesta la significación estadística, y esto si el estudio se basa en un muestreo probabilístico. Sin embargo, es muy frecuente que la discusión de los resultados de un experimento se hagan tomando como parámetro principal la significación estadística, muchas veces sin mencionar el valor de los parámetros clínicos o sociológicos calculados y, en muchos casos, a partir de muestras no probabilísticas. Si las muestras son grandes la significación estadística está casi garantizada.

El poder político y económico necesita apoyo a sus decisiones. En la antigüedad se consultaban los oráculos, que se consideraban la voz de la verdad porque provenían de los dioses o de fuerzas superiores que rara vez se equivocaban. Si el consultante era poderoso, las predicciones casi siempre apoyaban sus deseos; si fallaban se achacaba a errores de interpretación o a ofensas a las divinidades realizadas después de las profecías. Aunque muchos usuarios poderosos y sacerdotes sabían que los oráculos eran una patraña, les interesaba mantenerla: los poderosos porque recibían un respaldo divino a sus decisiones, y los sacerdotes de todos los rangos porque vivían muy bien de este trabajo.

En la actualidad el poder político, en lugar de oráculos consulta encuestas, y en el caso de las ciencias de la salud, el poder económico consulta estudios de investigación; curiosamente los resultados apoyan casi siempre a los poderosos, como ocurría en la antigüedad. Parece que este sistema es cómodo para casi todos los implicados en él, y a pocos preocupa los graves errores que hay en su aplicación. La gran diferencia entre los oráculos y el método científico es que este último permite obtener información acertada cuando se utiliza correctamente.

Si los estudios se realizan con el rigor científico y la precisión que los expertos dicen tener: ¿cómo es posible que se cometan tantos errores y que con tanta frecuencia los resultados, apoyados por los mejores expertos, muchas veces se compruebe que eran erróneos?

Mención especial merecen los ensayos clínicos. Los tratamientos médicos se basan en ellos, y la técnica de estos estudios no ha variado sustancialmente en los últimos treinta años. Sin embargo, muchos estudios realizados hace quince o veinte años, sobre tratamientos que parecían estupendos, la práctica ha demostrado que no eran acertados o que sus riesgos eran mucho mayores de lo que parecía, y que el error aleatorio no es suficiente para explicar tantos desatinos. Sin tener en cuenta el fraude, que puede explicar una parte de los errores, el problema principal es la aplicación incorrecta de técnicas estadísticas y la interpretación inadecuada de los resultados basándose en la significación estadística en lugar de la significación clínica. Las consecuencias pueden ser dramáticas: tratamientos inadecuados, fallecimientos y secuelas por efectos secundarios no previstos, etc.

Es necesario revisar la aplicación de la estadística a las ciencias de la salud, y su implementación con los métodos de investigación, a fin de optimizar los resultados de las investigaciones, lo que sin duda será beneficioso para la mayoría de los ciudadanos.

En Villa Libertad, Benalmádena a 11 de marzo de 2007.

RAFAEL C. ÁLVAREZ CÁCERES

ralvarez@nacom.es

---

# GUÍA DE LECTURA

Este libro puede ser útil como libro de aprendizaje; es decir, como libro de texto y como libro de consulta. Se incluyen casi todos los casos que pueden necesitarse en estadística aplicada, huyendo del simplismo generalizado de aproximar a la normal casi todos los ejemplos, cometiendo importantes errores en muchas ocasiones.

En la actualidad, los cálculos estadísticos se realizan mediante programas estadísticos. Uno de los más utilizados en ciencias de la salud y en ciencias sociales es SPSS<sup>1</sup>; por eso en este libro se explican las salidas de resultados más utilizadas de este programa, aunque como los resultados correspondientes a la mayoría de las aplicaciones (estadística descriptiva, comparación de dos medias, tablas de contingencia, correlación, regresión, regresión logística, etc.) son muy similares a la mayoría de los programas estadísticos, los comentarios acerca de las tablas de resultados obtenidas mediante SPSS pueden ser útiles para los usuarios de otros muchos programas como G-STAT, SAS y STAT GRAPHICS, entre otros. Existe la creencia muy extendida de que los programas estadísticos lo hacen todo respecto a los cálculos estadísticos, y que no es necesario tener grandes conocimientos estadísticos para obtener buenos resultados. Esto es inexacto. Obtener resultados mediante programas estadísticos es relativamente sencillo, pero interpretar los parámetros obtenidos y elegir los adecuados no es tarea fácil y exige amplios conocimientos de estadística para no cometer errores importantes. Además, en el caso de cálculos estadísticos de cierto nivel es necesario saber los parámetros necesarios para realizar los cálculos que muchas veces no se obtiene de manera inmediata; por ejemplo, el cálculo del riesgo relativo mediante un modelo de regresión logística no se obtiene de manera explícita; tampoco es posible comparar de manera inmediata dos coeficientes de regresión o de correlación. Cierto es que la mayoría de los grandes paquetes estadísticos como SPSS, además

---

<sup>1</sup> SPSS es marca registrada de SPSS INC Chicago USA.

de los menús permiten obtener resultados mediante programación, pero ello exige importantes conocimientos informáticos y estadísticos. Cada vez es más necesario tener conocimientos amplios de estadística para utilizar eficientemente los programas. Este libro expone con detalle las principales técnicas utilizadas en estadística aplicada en general, y en particular las de uso cotidiano en las ciencias de la salud.

El libro se puede dividir en dos partes: los primeros once capítulos incluyen la materia fundamental, los conocimientos generales e imprescindibles; la segunda parte incluye técnicas estadísticas concretas.

El Capítulo 1 expone los principios fundamentales estadísticos y científicos. Se recomienda leerlo atentamente porque muchos errores se cometen por interpretaciones inadecuadas de los conceptos de población y muestra, y de un principio elemental en estadística aplicada: la validez de los valores de las variables. Todo estudio estadístico se basa en procesar la información disponible acerca de un conjunto de variables, si los valores de éstas son inexactos, también lo son los cálculos estadísticos.

El Capítulo 2 se refiere a la estadística descriptiva; es decir, tabulación, gráficos y parámetros fundamentales correspondientes a los tipos de variables utilizadas en estadística aplicada.

El Capítulo 3 hace una introducción a los principios elementales de la probabilidad: espacio muestral, axiomas y principales teoremas de la probabilidad son algunos de los temas comentados. Se utiliza como herramienta matemática el álgebra de Boole. Si algún lector necesita recordar los fundamentos de esta técnica matemática, en el apéndice uno se recuerdan sus fundamentos.

El Capítulo 4 aborda los conceptos de sucesos mutuamente excluyentes e independientes; además se estudia la probabilidad condicionada y el teorema de Bayes, fundamental para el cálculo de probabilidades en la aplicación a las ciencias de la salud, como en el caso de ayudas al diagnóstico.

El Capítulo 5 trata de la aplicación del cálculo de probabilidades al cálculo de parámetros muy utilizados en ciencias de la salud, como riesgo, riesgo relativo, razón de predominio, sensibilidad, especificidad y valores predictivos.

El Capítulo 6 es más técnico desde el punto de vista estadístico, pero es muy interesante como base para abordar las principales distribuciones de probabilidad en los capítulos siguientes. Se recomienda estudiarlo detenidamente.

El Capítulo 7 estudia las principales distribuciones de probabilidad aplicables a variables aleatorias discretas. La más importante es la binomial, pero también se utilizan con frecuencia las distribuciones de Poisson, multinomial, hipergeométrica, geométrica y binomial negativa.

En el Capítulo 8 se estudian las distribuciones de probabilidad más importantes aplicables a variables aleatorias continuas. La más utilizada es la distribución normal o de Gauss, pero también son de uso frecuente las distribuciones de la *t* de Student, la Chi-cuadrado y la *F* de Snedecor. En este capítulo se dan las instrucciones necesarias para el manejo de las tablas que permiten el cálculo de probabilidades mediante las distribuciones antedichas.

El Capítulo 9 introduce los principios de la estadística analítica. Es fundamental distinguir entre inferencia estadística e inferencia técnica; también se estudian las principales técnicas de muestreo, con especial atención al muestreo probabilístico.

El Capítulo 10 estudia uno de los temas que aborda la estadística analítica, la estimación de parámetros puntual y por intervalo. Se analiza el cálculo de intervalos de confianza para la estimación de medias y de proporciones teniendo en cuenta muchas circunstancias distintas; también se predetermina el tamaño de muestra necesario para estimar parámetros con una precisión determinada.

El Capítulo 11 estudia con detalle el otro tema que aborda la estadística analítica, el contraste de hipótesis. Los contrastes de hipótesis estadísticos siempre han estado rodeados de controversia, incluso hay quien aboga por su eliminación. Es cierto que su interpretación es compleja, pero son imprescindibles en la toma de decisiones en estadística analítica. Como todas las herramientas complejas, si no se saben utilizar se pueden cometer importantes errores. Se analizan con detalle muchos ejemplos y se estudian los errores que se cometen habitualmente en el contraste de hipótesis estadísticas.

El Capítulo 12 analiza algunos principios básicos a tener en cuenta en estadística aplicada, como causalidad, interacción entre variables y fenómenos de confusión.

El Capítulo 13 estudia la comparación de dos proporciones, tanto en el caso de datos pareados como independientes. Se realizan cálculos de intervalos de confianza y contraste de hipótesis en los casos más utilizados en estadística aplicada.

El Capítulo 14 estudia la estadística descriptiva y analítica mediante el uso de tablas de contingencia, que se utilizan muy frecuentemente en estadística aplicada a las ciencias de la salud y a las ciencias sociales. El cálculo de la Chi-cuadrado de Pearson, la corrección de Yates, la prueba de Fisher y la prueba de McNemar, son algunas de las pruebas estadísticas analizadas. Se realiza el cálculo de los parámetros más utilizados en el análisis del riesgo mediante tablas de contingencia.

El Capítulo 15 trata sobre la comparación de dos medias, tanto en el caso de datos apareados como datos independientes. El cálculo de intervalos de confianza, contraste de hipótesis y predeterminación del tamaño muestral particularizados cuando se analizan simultáneamente dos medias, son los temas fundamentales estudiados en este capítulo.

En el Capítulo 16 se estudia el análisis de la varianza unifactorial. Esta interesante técnica estadística permite comparar  $k$  medias de manera simultánea siendo  $k \geq 2$ . Además, el contraste de hipótesis mediante el cociente de dos varianzas y el análisis de la tabla de ANOVA se utilizan en muchas técnicas de estadística avanzada, como el análisis de regresión. Las técnicas de comparación múltiple se utilizan también en otras técnicas estadísticas, como las pruebas de Friedman y de Kruskal Wallis. Se recomienda al lector interesado en las técnicas estadísticas avanzadas estudiar este capítulo, puesto que los principios que en él se exponen son aplicables a otras muchas técnicas.

El Capítulo 17 estudia el análisis de la varianza bloques y el análisis de la varianza de medidas repetidas, que permite comparar  $k$  medias, siendo  $k \geq 2$ , en el caso de datos pareados. También se estudian las pruebas de comparación múltiple aplicables en el caso de rechazar la hipótesis nula de igualdad de medias.

En el Capítulo 18 se estudia la asociación entre variables cuantitativas, fundamentalmente las medidas de correlación de Pearson y de Spearman. También se analiza la comparación de dos coeficientes de correlación.

El Capítulo 19 analiza la regresión lineal simple. Con objeto de fijar conceptos, se recomienda al lector estudiar primero la regresión simple, y después la múltiple, ya que muchos conceptos son similares y aplicables de manera inmediata a esta última.

El Capítulo 20 analiza la regresión lineal múltiple. Se han evitado los complejos análisis matemáticos, aunque se comenta que el modelo matricial no es imprescindible para el seguimiento del capítulo, aunque sí deseable. Entre otros temas, se analiza con detalle la inclusión en los modelos de variables ficticias (Dummy).

En el Capítulo 21 se estudia la regresión logística simple y múltiple; esta es una de las técnicas más utilizadas en ciencias de la salud en los artículos publicados en las revistas más prestigiosas del mundo. En primer lugar se analiza la regresión simple para fijar conceptos, y después la múltiple. Al igual que en el caso de la regresión lineal múltiple, se hace un estudio amplio sobre las variables ficticias (Dummy). Después se estudia el análisis del riesgo mediante modelos de regresión logística.

En el Capítulo 22 se analizan las técnicas estadísticas no paramétricas más utilizadas en estadística aplicada.

El Capítulo 23 tercero trata los aspectos estadísticos generales de la metodología de la investigación aplicada a las ciencias de la salud.

El Capítulo 24 estudia los aspectos estadísticos más importantes de los estudios observacionales.

El Capítulo 25 analiza las características más importantes de los ensayos clínicos y sus peculiaridades estadísticas.

Y como punto final, en el Capítulo 26 se comentan los errores más frecuentes que se suelen cometer en estadística aplicada.

familiarizarse con ciertas expresiones matemáticas. Los cálculos necesarios se realizan consultando tablas, la tabla de la curva normal tipificada que se utiliza habitualmente en este texto está basada en la función de distribución acumulativa. Cuando se determina el valor de un área mediante la tabla se realiza de una manera sencilla un cálculo que de otra manera habría que realizar resolviendo complejas expresiones integrales.

## **8.5. CÁLCULO DE PROBABILIDADES MEDIANTE TABLAS**

Como se ha indicado en apartados anteriores, el cálculo de probabilidades de variables aleatorias normalmente distribuidas puede realizarse resolviendo complicadas expresiones integrales o mediante el uso de tablas, lo cual es mucho más sencillo.

Las tablas pueden ser distintas, no todas están basadas en la función de distribución acumulativa; saber manejar un tipo de tabla no presupone que se sepan manejar todas; se recomienda al lector que si necesita manejar una tabla distinta de la que emplea habitualmente, para evitar errores, antes de utilizarla estudie detenidamente sus fundamentos.

Las tablas que se utilizan en este libro, para el cálculo de probabilidades correspondientes a variables aleatorias normales y, también para el cálculo de probabilidades de variables aleatorias que se distribuyen de manera distinta a la normal, están basadas en la función de distribución acumulativa.

Los problemas a resolver son fundamentalmente de dos tipos:

- a) Conocida la abscisa de la variable aleatoria  $X$ , o de la variable aleatoria tipificada  $z$ , calcular la probabilidad de que la variable tenga un valor dentro de un determinado intervalo.
- b) Calcular la abscisa o abscisas que delimiten una proporción de área determinada.

Mediante el siguiente ejemplo se comenta detalladamente el uso de la tabla. Se recomienda al lector seguirlo con especial atención, puesto que las cuestiones a resolver han sido especialmente diseñadas para comprender el manejo de la tabla.



**EJEMPLO 8.2**

La glucemia basal de los diabéticos atendidos en un centro sanitario puede considerarse como una variable normalmente distribuida, con media 106 mg por 100 ml, y desviación típica 8 mg por 100 ml  $N(106; 8)$ .

Calcular:

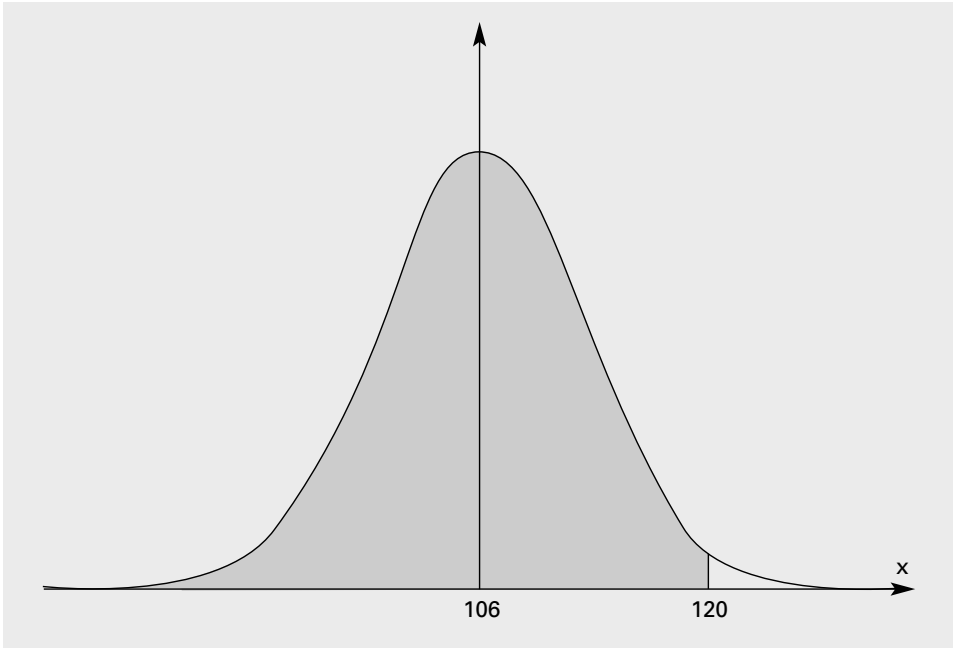
- La proporción de diabéticos con una glucemia basal inferior a 120 mg por 100 ml,  $P(X < 120)$  (recuerde que en variable continua es lo mismo menor que menor o igual).
- La proporción de diabéticos con una glucemia basal comprendida entre 106 y 120 mg por 100 ml.
- La proporción de diabéticos con una glucemia basal mayor de 120 mg por 100 ml.
- El nivel de glucemia basal tal que por debajo de él están el 25% de los diabéticos, es decir, el primer cuartil.

a) El cálculo anterior no puede realizarse directamente, puesto que no se dispone de tablas para los parámetros correspondientes a la variable  $X$ , pero tipificando 120 se obtiene:

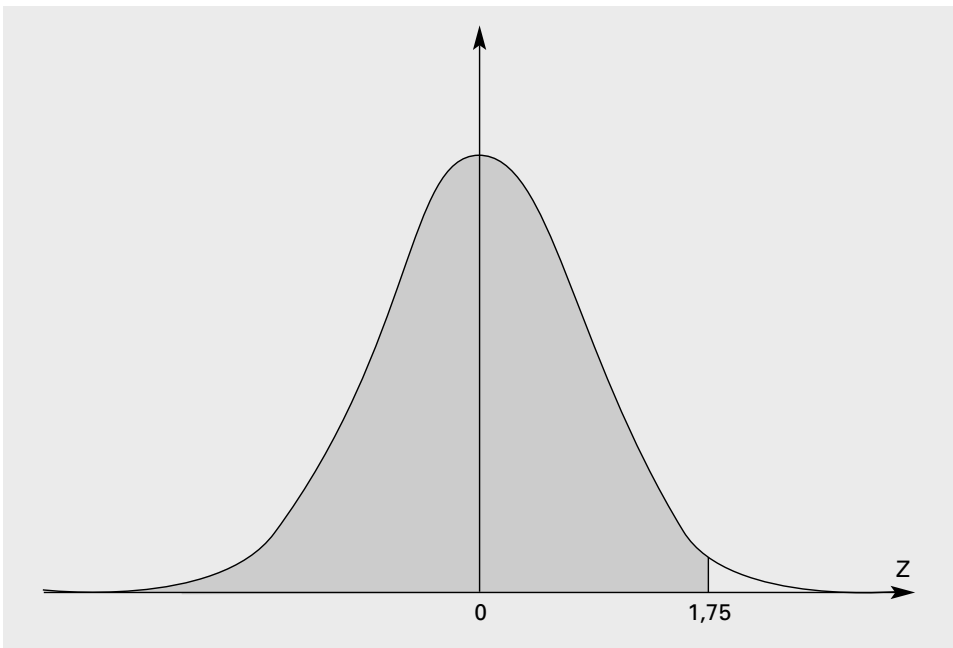
$$Z = \frac{(120 - 106)}{8} = 1,75$$

El valor tipificado tiene la siguiente propiedad:

$$P(X \leq 120) = P(Z \leq 1,75)$$



**FIGURA 8.9.**  $P(X \leq 120)$ .



**FIGURA 8.10.**  $P(Z \leq 1,75)$ .

En los gráficos anteriores el área sombreada representa la probabilidad pedida. Observe que el área bajo la curva  $N(106; 8)$  para  $X \leq 120$  es igual que el área bajo la curva tipificada  $N(0; 1)$  para  $Z \leq 1,75$ . Por lo tanto, con la tabla de la curva normal tipificada se calcula el área necesaria para resolver este punto del ejercicio.

Busque en la Tabla I en la columna encabezada por  $z$ , el valor 1,7, una vez localizado este valor observe que en la parte superior de la tabla hay 10 columnas desde la 0,00 hasta la 0,09, la confluencia de la fila cuyo comienzo es 1,7 y la columna 0,00 corresponde a  $P(X \leq 1,70)$  cuyo valor es 0,9554, la segunda columna  $P(X < 1,71)$  y así sucesivamente hasta llegar a la columna encabezada por 0,05, que corresponde a 1,75 cuya probabilidad es la que se desea calcular; la probabilidad buscada es 0,9599.

Observe que el valor de la columna encabezada por  $Z$  corresponde al valor entero y al primer decimal de la normal tipificada, en la fila correspondiente están los valores de las probabilidades, cada columna añade una centésima al valor que encabeza la fila, en el primer punto del ejemplo se desea calcular la probabilidad que deja a su izquierda el valor de  $Z$  1,75, es decir, la función de distribución acumulativa,  $F(Z = 1,75) = P(Z \leq 1,75)$ , en primer lugar se localiza la fila correspondiente a 1,7, a este valor cada columna añade una centésima a dicho valor, hasta 0,05, que son las centésimas que hay que añadir 1,7 para obtener 1,75; la intersección entre la fila encabezada por 1,7 y la columna encabezada por 0,05, muestra la probabilidad que se quiere calcular.

$$P(X \leq 120) = P(Z \leq 1,75) = 0,9599$$

La proporción de diabéticos con una glucemia basal menor de 120 mg por 100 ml es 0,9599. También se podría decir que la probabilidad de que un diabético seleccionado al azar en esta población tenga una glucemia basal inferior a 120 mg por 100 ml es 0,9599.

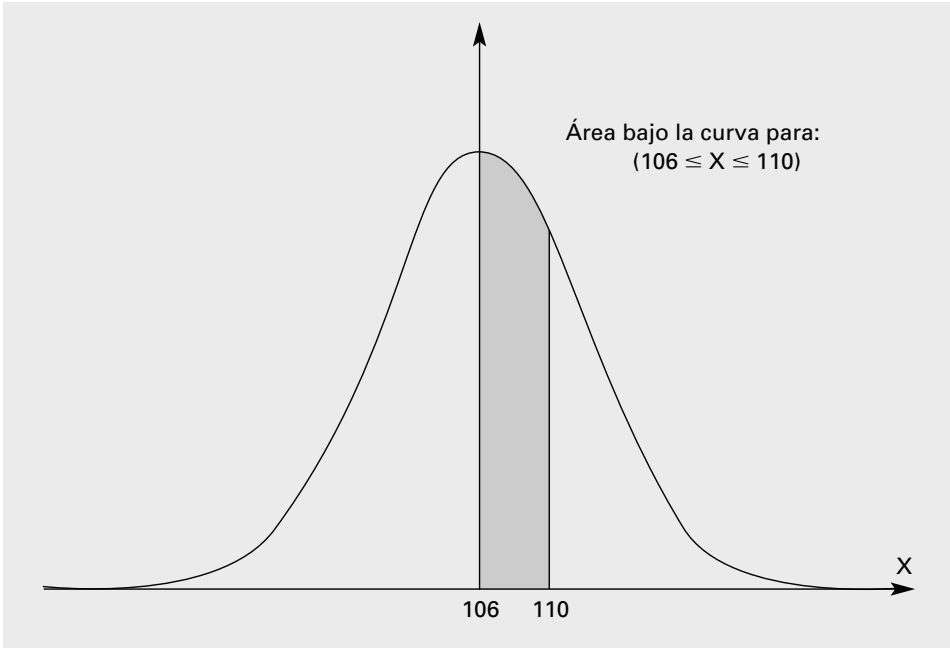
b) Se pide la probabilidad de que el valor de la variable aleatoria esté comprendida entre 106 y 110,  $P(106 \leq X \leq 110)$ . En la Figura 8.11 se representa la probabilidad pedida.

Observe que si al área bajo la curva para  $X \leq 110$ , se le resta el área bajo la curva para  $X \leq 106$  se obtiene la probabilidad pedida, que se corresponde con el área sombreada, esto puede expresarse matemáticamente de la siguiente manera:

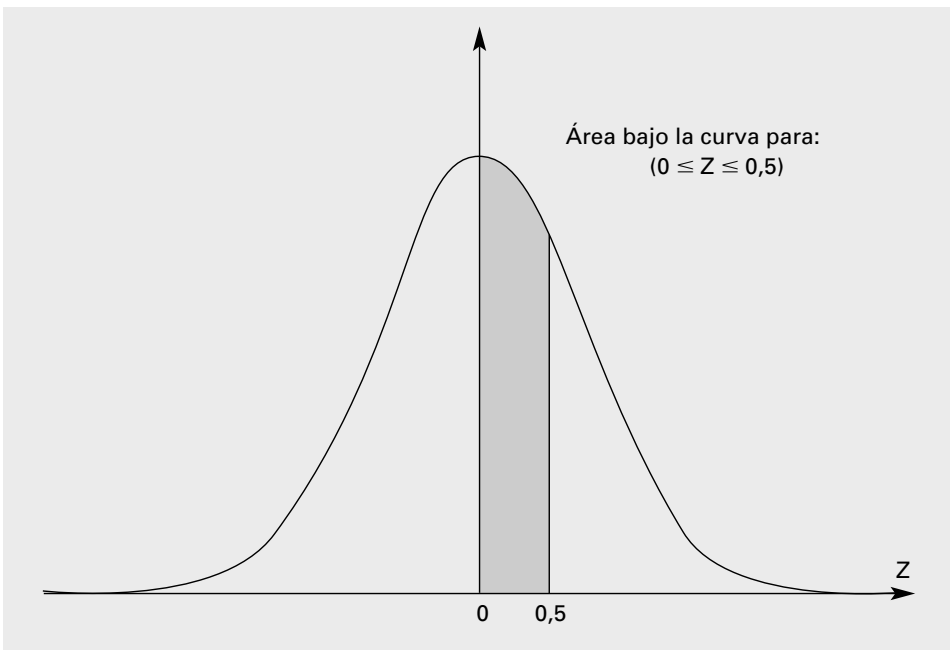
$$P(106 \leq X \leq 110) = P(X \leq 110) - P(X \leq 106)$$

El valor tipificado de 110 es 0,5:  $Z = \frac{110 - 106}{8}$ ; el valor tipificado de 106 es 0:  $Z = \frac{106 - 106}{8}$ . Observe que  $Z$  es la distancia en desviaciones típicas de un valor a la media, puesto que 106 es la media su distancia a 106, o sea, a sí misma es 0.

El área sombreada en la curva normal tipificada es la probabilidad pedida (Figura 8.12).



**FIGURA 8.11.**  $P(106 \leq X \leq 110)$ .



**FIGURA 8.12.**  $P(0 \leq Z \leq 0,5)$ .

$$P(106 \leq X \leq 110) = P(X \leq 110) - P(X \leq 106) = P(Z \leq 0,5) - P(Z \leq 0)$$

$P(Z \leq 0,5)$  consultando en la tabla es igual a 0,6915.

$P(Z \leq 0)$  es 0,5 puesto que es la mitad de la curva.

$$P(X \leq 110) - P(X \leq 106) = P(Z \leq 0,5) - P(Z \leq 0) = 0,6915 - 0,5 = 0,1915$$

La proporción de diabéticos con una glucemia basal comprendida entre 110 y 106 mg por 100 ml es 0,1915.

c) Se pide la proporción de diabéticos con una glucemia basal mayor de 120 mg por 100 ml,  $P(X > 120)$ . En variable continua  $\leq$  que 120 o  $<$  120, es lo mismo, puesto que la única diferencia sería la probabilidad de que la variable tome, exactamente, el valor 120, que es 120 seguido de infinitos ceros, dicha probabilidad es cero.

$$P(X \leq 120) + P(X > 120) = 1$$

Observe que en la expresión anterior están contempladas todas las posibilidades, por lo tanto la probabilidad es 1.

En el punto a se ha calculado  $P(X \leq 120)$ , cuyo valor es 0,9599.

$$P(X > 120) = 1 - P(X \leq 120) = 1 - 0,9599 = 0,0401$$

La proporción de diabéticos con una glucemia basal mayor de 120 mg por 100 ml es 0,0401.

d) En este caso se pide el valor de la glucemia basal  $a$ , que cumpla la siguiente condición:

$$P(X < a) = 0,25$$

Observe que esta cuestión es distinta a las anteriores, en este caso se conoce la probabilidad pero no el valor de  $X$ .

En la Figura 8.13 se representa gráficamente el problema.

El punto  $a$  tiene un valor tipificado  $Z_a$  caracterizado por:  $(Z < z_a) = 0,25$ .

Buscando en la tabla en los valores de las probabilidades, no en la columna  $Z$ , tenga en cuenta que en este caso se conoce la probabilidad, 0,25, y se quiere calcular el valor de  $Z$  que le corresponde. El valor 0,25 exacto no está en la tabla, los valores mayor y menor que el son 0,2514, al que le corresponde el valor  $Z = -0,67$ , y 0,2483 al que corresponde el valor  $Z = -0,68$ ; el valor buscado está entre los dos valores de  $Z$  anteriores y, aproximadamente, en el punto medio de los dos, por lo tanto:  $P(Z \leq -0,675) = 0,25$ , o sea, que  $-0,675$  es el valor que corresponde al primer cuartil de la variable normal tipificada.

En la curva de la Figura 8.14 se representa gráficamente la situación en la curva normal tipificada.

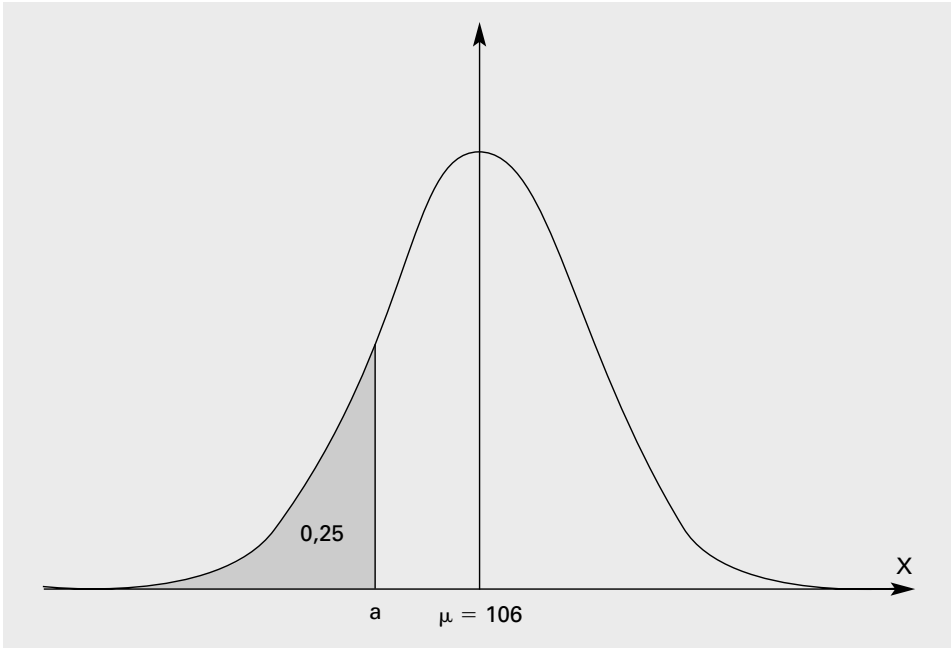


FIGURA 8.13.  $P(X \leq a) = 0.25$ .

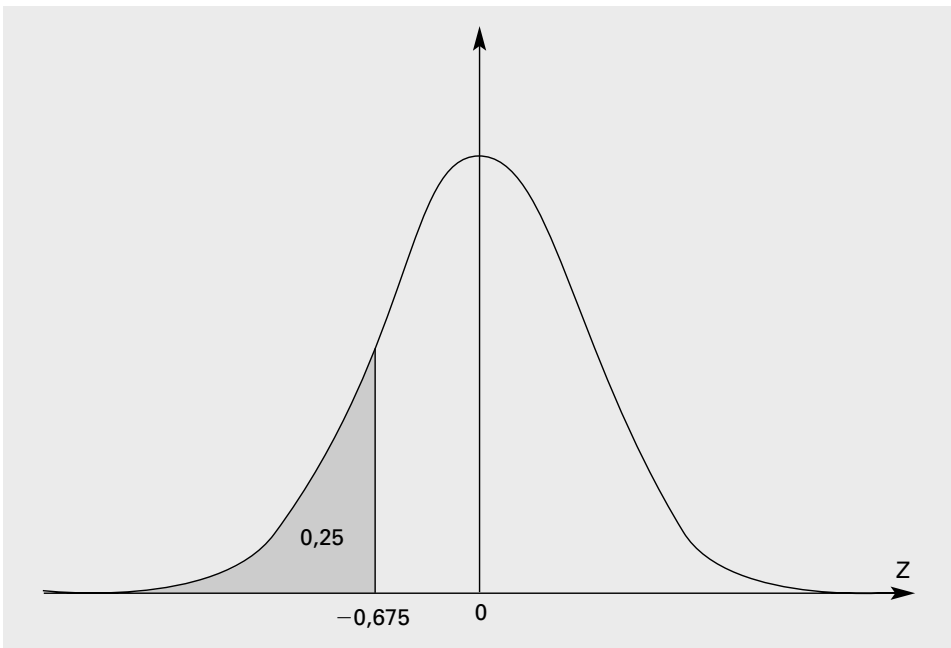


FIGURA 8.14.  $P(Z \leq -0,675) = 0.25$ .

$-0,675$  es el valor tipificado de  $a$ , ahora hay que calcular el valor de la variable  $X$ , denotado mediante  $a$ , que corresponde a ese valor:

$$-0,675 = \frac{a - 106}{8}$$

Despejando  $a$  en la ecuación anterior:

$$a = 100,6$$

El valor pedido es 100,6, esto significa que el 25% de los diabéticos de la población estudiada tienen una glucemia basal inferior a 100,6 mg por 100 ml.

## 8.6. TEOREMA CENTRAL DEL LÍMITE

Si una variable aleatoria  $X$  puede expresarse como suma de  $n$  variables aleatorias independientes, para  $n \geq 30$ ,  $X$  es aproximadamente normal.

El teorema central del límite tiene un gran número de aplicaciones en estadística. Hay muchas variables que pueden descomponerse en la suma de  $n$  variables aleatorias independientes y cuando  $n$  es mayor de 30 se puede considerar que dichas variables son aproximadamente normales. La mayoría de los autores aceptan que con  $n \geq 30$  es suficiente para que la aproximación a la normal sea adecuada, lo cual ha sido comprobado en múltiples experimentos de simulación con ordenador. La aproximación a la normal mejora según aumenta  $n$ .

## 8.7. APROXIMACIÓN DE UNA DISTRIBUCIÓN BINOMIAL A UNA NORMAL

En ocasiones y bajo ciertas condiciones se puede aproximar una distribución binomial a una normal, lo cual puede facilitar notablemente los cálculos. La aproximación no siempre es posible, y si no se tienen en cuenta las condiciones que la permiten pueden cometerse importantes errores de cálculo.

Si una variable aleatoria  $X$  es binomial con parámetros  $n$  y  $p$ ,  $B(n, p)$ , puede aproximarse a una distribución normal con  $\mu = np$  y desviación típica  $\sigma = \sqrt{npq}$  cuando se cumplen simultáneamente las siguientes condiciones:

- a)  $p > 0,05$
- b)  $q > 0,05$  ( $q = 1 - p$ )
- c)  $n \geq 30$

Una distribución binomial de parámetros  $n$  y  $p$ , consiste en  $n$  pruebas independientes, la variable aleatoria binomial  $X$ , puede considerarse como la suma de

$n$  procesos de Bernouille, cada uno de ellos compuesto por una variable aleatoria dicotómica, por lo tanto una variable aleatoria binomial está compuesta por  $n$  variables aleatorias independientes.

Muchos autores consideran que la aproximación binomial a la normal es posible si  $np > 5$  y  $nq > 5$ , pero esta condición puede ser insuficiente en muchos casos, por ejemplo si  $p = 0,5$  y  $n = 11$  se cumple la condición  $np$  y  $nq > 5$ , sin embargo el teorema central del límite exige que  $n$  sea al menos igual a 30.

La variable aleatoria binomial  $X$  es discreta, mientras que las variables normalmente distribuidas son continuas; cuando una variable binomial cumple los requisitos para realizar la aproximación a la normal debe tenerse en cuenta que  $P(X = a) = 0$ , si  $X$  es continua, para evitar errores debe realizarse la corrección por continuidad, esto significa que la equivalencia  $P(X = a)$  cuando  $X$  es binomial y por lo tanto discreta, debe considerarse como  $P(a - 0,5 < X < a + 0,5)$ , al realizar la aproximación a la normal y ser  $X$  continua, calculando la probabilidad de que  $X$  esté contenida en un determinado intervalo.

### EJEMPLO 8.3

Un tratamiento antibiótico es efectivo frente a infecciones pulmonares por legionella en el 25% de los casos. Los pacientes mejoran permaneciendo con buen estado general y afebriles antes de transcurridas 72 horas del comienzo del tratamiento. En una epidemia de infecciones pulmonares por legionella se aplica el tratamiento a 80 pacientes.

Calcular la probabilidad de que antes de 72 horas de iniciado el tratamiento mejoren entre 25 y 35 pacientes, es decir, sea efectivo el tratamiento.

Cada paciente tratado es una prueba independiente de las demás. La variable aleatoria  $X$  número de pacientes que mejoran antes de 72 horas de administrado el tratamiento es discreta y el experimento cumple los requisitos de la distribución binomial con parámetros  $n = 80$  y  $p = 0,25$ ,  $B(80; 0,25)$ .

Se cumplen los requisitos de aproximación a la normal:

- a)  $p = 0,25 \Rightarrow p > 0,05$
- b)  $q = 0,75 \Rightarrow q > 0,05$
- c)  $n = 80 \Rightarrow n > 30$

Podemos pasar de una  $B(80; 0,25)$  a una normal con  $\mu = 20$  y  $\sigma = 3,875$ , los parámetros de la variable normal son:  $N(20; 3,87)$ .

El problema pide calcular:  $P(25 \leq X \leq 30)$ , aproximando a la normal y teniendo en cuenta la corrección por continuidad la probabilidad pedida es:

$$P(24,5 \leq X \leq 30,5)$$

Tenga en cuenta que 25 al pasar de variable discreta a continua es el intervalo 24,5 - 25,5 y 30 el intervalo 29,5 - 30,5.



Resumiendo, se dispone de una variable aleatoria normal con  $\mu = 20$  y  $\sigma = 3,87$   $N(20; 3,87)$ .

$$P(24,5 \leq X \leq 30,5) = P(X \geq 30,5) - P(X \leq 24,5)$$

El valor tipificado de 30,5 es: 2,71, y el de 24,5: 1,16. Por lo tanto:

$$P(X \geq 30,5) - P(X \leq 24,5) = P(Z < 2,71) - P(Z < 1,16)$$

Consultando en las tablas de la curva normal tipificada:

$$P(Z < 2,71) - P(Z < 1,16) = 0,9966 - 0,8770 = 0,1196$$

Hay una probabilidad de 0,1196 de que de los 80 pacientes tratados, entre 25 y 30 mejoren antes de 72 horas de iniciado el tratamiento.

## 8.8. APROXIMACIÓN DE LA DISTRIBUCIÓN DE POISSON A LA NORMAL

La distribución de Poisson es aplicable a variables aleatorias discretas. Cuando el parámetro  $\lambda$  es mayor de 10, se puede aproximar a una distribución normal con  $\mu = \lambda$  y  $\sigma = \sqrt{\lambda}$ .

Teniendo en cuenta que la variable aleatoria  $X$  es discreta, al realizar la aproximación a la normal debe hacerse la corrección por continuidad.

### EJEMPLO 8.4.

En un hospital el número medio de pacientes con dolor abdominal atendidos por día es 16.

Calcular la probabilidad de que un día determinado haya más de 25 pacientes con dolor abdominal.

El número de pacientes con dolor abdominal puede considerarse un suceso de Poisson con  $\lambda = 16$ , teniendo en cuenta que  $\lambda > 10$  se puede hacer una aproximación a una normal con  $\mu = 16$  y  $\sigma = 4$ .

La probabilidad pedida es:

$$P(X > 25) = 1 - P(X \leq 25)$$

Al realizar la aproximación a la normal hay que hacer la corrección por continuidad, por lo tanto, la probabilidad anterior queda de la siguiente manera:

$$P(X > 24,5) = 1 - P(X \leq 24,5); \text{ el valor tipificado de } 24,5 \text{ es: } 2,13 \\ 1 - P(X \leq 24,5) = 1 - P(Z \leq 2,13) = 1 - 0,9834 = 0,0166$$

## 8.9. DISTRIBUCIÓN GAMMA

La distribución o función gamma ( $\Gamma$ ) de Euler es la base de algunas de las más importantes distribuciones de probabilidad utilizadas en estadística.

La función gamma de grado  $k$  está definida por la siguiente expresión:

$$\Gamma(k) = \int_0^{\infty} X^{k-1} e^{-x} dx \quad [8.7]$$

La función anterior está definida para  $k > 0$ .

Se puede demostrar que  $\Gamma(k) = (k - 1)!$

En el caso particular  $k = 1$ :

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx \quad [8.8]$$

$\Gamma(1)$  es la función exponencial negativa para  $x \geq 0$ .

Resolviendo [8.7]:

$$\Gamma(1) = -[e^{-x}] X \text{ entre } 0 \text{ e } \infty$$

$\Gamma(1) = 1$ ; teniendo en cuenta que  $\Gamma(k) = (k - 1)!$ ,  $\Gamma(1) = 0! = 1$ . Observe que  $0! = 1$  es una igualdad matemática y no un convenio. La función exponencial negativa para  $x \geq 0$  es igual a 1 y, por lo tanto, puede definir una distribución de probabilidad. Recuerde que una condición necesaria de una función para poder definir una distribución de probabilidad es que el área entre la curva y el eje de abscisas sea igual a 1.

## 8.10. LA DISTRIBUCIÓN CHI-CUADRADO ( $\chi^2$ )

La distribución  $\chi^2$  es una de las más utilizadas en ciencias de la salud, muchos estimadores de uso corriente en investigación clínica y epidemiológica se distribuyen según esta curva. En este apartado se estudian las características funcionales de la distribución y sus propiedades más importantes, incluyendo la utilización de la tabla.

Una variable aleatoria continua  $X$ , sigue una distribución chi-cuadrado con  $\gamma$  grados de libertad<sup>1</sup>;  $\chi^2_{\gamma}$ , si su función de probabilidad es la siguiente:

<sup>1</sup> Grados de libertad es el número de variables que pueden tomar valores libremente. Las restricciones son los parámetros que tienen que tomar un valor determinado.

$$SS_{\text{Residual}} = \sum_{j=1}^K \sum_{i=1}^b (X_{ji} - \bar{X}_{Gj} - \bar{X}_{Bi} - \bar{\bar{X}})^2 \quad [17.10]$$

Para facilitar los cálculos de la suma de cuadrados es preferible utilizar las siguientes fórmulas:

$$SS_{\text{Total}} = \sum_{j=1}^K \sum_{i=1}^b X_{ji}^2 - \frac{\left( \sum_{j=1}^K \sum_{i=1}^b X_{ji} \right)^2}{Kb} \quad [17.11]$$

$$SS_{\text{Grupos}} = \sum_{j=1}^K \frac{T_{Gj}^2}{b} - \frac{\left( \sum_{j=1}^K \sum_{i=1}^b X_{ji} \right)^2}{Kb} \quad [17.12]$$

En la expresión anterior  $T_{Gj}$  es la suma de todos los valores del  $j$ -ésimo grupo.

$$SS_{\text{Bloques}} = \sum_{i=1}^n \frac{T_{Bi}^2}{K} - \frac{\left( \sum_{j=1}^K \sum_{i=1}^b X_{ji} \right)^2}{Kb} \quad [17.13]$$

En la expresión anterior  $T_{Bi}$  es la suma de todos los valores del  $i$ -ésimo bloque.

$$SS_{\text{Residual}} = SS_{\text{Total}} - SS_{\text{Grupos}} - SS_{\text{Bloques}} \quad [17.14]$$

#### 17.4. CUADRADOS MEDIOS

Los cuadrados medios, que son la estimación puntual de las varianzas, se obtienen dividiendo la suma de cuadrados por los correspondientes grados de libertad. Los grados de libertad,  $GL$ , son aditivos, es decir, los grados de libertad totales, son iguales a los grados de libertad entre grupos más los grados de libertad de bloques más los grados de libertad residuales:

$$\begin{aligned} GL_{\text{Total}} &= GL_{\text{Grupos}} + GL_{\text{Bloques}} + GL_{\text{Residual}} \\ GL_{\text{Total}} &= Kb - 1; \quad GL_{\text{Grupos}} = K - 1; \quad GL_{\text{Bloques}} = b - 1 \end{aligned} \quad [17.15]$$

Despejando en 17.15.

$$\begin{aligned} GL_{\text{Residual}} &= GL_{\text{Total}} - GL_{\text{Grupos}} - GL_{\text{Bloques}} \\ GL_{\text{Residual}} &= (Kb - 1) - (K - 1) - (b - 1) \end{aligned} \quad [17.16]$$

$$GL_{\text{Residual}} = (K - 1)(b - 1) \quad [17.17]$$

$$MS_{\text{Total}} = \frac{SS_{\text{Total}}}{Kb - 1} \quad [17.18]$$

$$MS_{\text{Grupos}} = \frac{SS_{\text{Grupos}}}{K - 1} \quad [17.19]$$

$$MS_{\text{Bloques}} = \frac{SS_{\text{Bloques}}}{b - 1} \quad [17.20]$$

$$MS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{(K - 1) \cdot (b - 1)} \quad [17.21]$$

## 17.5. HIPÓTESIS DE ANOVA BLOQUES

En el análisis de la varianza de bloques la hipótesis principal a contrastar es la igualdad de las medias de los grupos.

$$\begin{aligned} H_0 & \quad \mu_{G1} = \mu_{G2} = \dots \mu_{GK} \\ H_1 & \quad \mu_{Gi} \neq \mu_{Gj} \text{ para algún } i, j \end{aligned}$$

En el ejemplo 17.1, la hipótesis principal a contrastar es si existen diferencias en disminución del colesterol entre los hipolipemiantes, una vez controlado el posible efecto distorsionador de los hipotensores.

Una hipótesis secundaria a contrastar, aunque muchas veces carece de interés, es la igualdad entre las medias de los bloques.

$$\begin{aligned} H_0 & \quad \mu_{B1} = \mu_{B2} = \dots \mu_{BK} \\ H_1 & \quad \mu_{Bi} \neq \mu_{Bj} \text{ para algún } i, j \end{aligned}$$

Lo más frecuente es que haya diferencias entre los bloques, precisamente por eso se controla su efecto para que esta distorsión no afecte al objetivo principal del estudio, que es comparar las medias de los grupos. En cualquier caso, sea o no estadísticamente significativo, el efecto bloques no afecta al resultado del contraste de la hipótesis principal. Lo importante es el diseño, bloquear el posible efecto de una variable extraña a los objetivos del estudio.

En el ejemplo 17.1, el contraste entre las medias de los bloques compararía las medias del colesterol entre los distintos tratamientos hipotensores.

De manera similar que en el análisis de la varianza unifactorial, la hipótesis principal se contrasta comparando el cuadrado medio entre grupos y el cuadrado

medio residual mediante la prueba de la  $F$  de Snedecor, si la significación estadística es menor que el valor  $\alpha$  del contraste de hipótesis se concluye que hay un efecto sobre la variable dependiente debido al grupo, en el ejemplo de los hipolipemiantes, indicaría que el efecto de los tratamientos es diferente.

La hipótesis secundaria se contrasta comparando el cuadrado medio bloques con el cuadrado medio residual, también mediante la prueba de la  $F$  de Snedecor, si la significación estadística es menor que el valor  $\alpha$  del contraste de hipótesis se concluye que hay un efecto sobre la variable dependiente debido al bloque.

Si se sospecha que en las poblaciones de referencia no se cumplen los supuestos de ANOVA, se puede contrastar la hipótesis principal mediante el cociente entre  $MSG$  y  $MSR$ , igual que antes, pero considerando que la  $F$  tiene 1 y  $(b - 1)$  grados de libertad, en lugar de  $(K - 1)$  y  $(K - 1)(b - 1)$  grados de libertad, de esta manera aumenta la probabilidad de cometer un error tipo II, pero disminuye la probabilidad de cometer error tipo I.

### 17.5.1. Tabla de ANOVA bloques

Los parámetros fundamentales para realizar un análisis de la varianza bloques se suelen exponer en una tabla similar a la siguiente:

#### Análisis de la varianza bloques

Fuente de variación	GL	Suma de cuadrados	Cuadrados medios	$F$ cociente	$F$ prob.
Entre grupos	$K - 1$	$SS_{\text{Grupos}}$	$MS_{\text{Grupos}}$	$MS_{\text{Grupos}}/MS_{\text{Residual}}$	$P$
Bloques	$b - 1$	$SS_{\text{Bloques}}$	$MS_{\text{Bloques}}$	$MS_{\text{Bloques}}/MS_{\text{Residual}}$	
Residual	$(k - 1)(b - 1)$	$SS_{\text{Residual}}$	$MS_{\text{Residual}}$		
Total	$Kb - 1$	$SS_{\text{Total}}$			

#### EJEMPLO 17.2

El objetivo de un estudio es evaluar el poder hipocolesterolemiante de tres tratamientos diferentes  $A$ ,  $B$  y  $C$  en pacientes hipertensos; el tipo de tratamiento hipotensor puede influir en los resultados; para controlar este posible efecto se realiza un diseño de bloques. Hay tres pacientes por cada uno de los seis tipos principales de tratamiento hipotensor: diuréticos, betabloqueantes, alfabloqueantes, IECAS, ARAII y calcioantagonistas; se seleccionan tres pacientes de cada tipo de tratamiento hipotensor y se asignan al azar a cada uno de los tres tratamientos hipolipemiantes; tres meses después se analiza el nivel del colesterol total de cada uno de los dieciocho pacientes que participan en el estudio en mg por 100 ml. Los datos obtenidos son los siguientes:

TABLA 17.2. Hipolipemiantes

	A	B	C			
Diuréticos	227	233	210	$T_D = 670$	$\bar{x}_D = 223,33$	$S_D = 11,93$
Betabloq.	231	241	190	$T_{BQ} = 662$	$\bar{x}_{BQ} = 220,67$	$S_{BQ} = 27,02$
Alfabloq.	216	252	186	$T_{AQ} = 654$	$\bar{x}_{AQ} = 218$	$S_{AQ} = 33,05$
IECAS	222	237	175	$T_{IE} = 634$	$\bar{x}_{IE} = 211,33$	$S_{IE} = 32,35$
ARAI	217	242	177	$T_{AR} = 636$	$\bar{x}_{AR} = 212,00$	$S_{AR} = 32,79$
Calcioant.	211	250	168	$T_{CL} = 629$	$\bar{x}_{CL} = 209,67$	$S_{CL} = 41,02$
	$T_A = 1.324$	$T_B = 1.455$	$T_C = 1.106$			
	$\bar{x}_A = 220,67$	$\bar{x}_B = 242,50$	$\bar{x}_C = 184,3$			
	$S_A = 7,45$	$S_B = 7,34$	$S_C = 14,84$			
	$\bar{X} = 215,83$	$\sum_{j=1}^3 \sum_{i=1}^6 X_{ji} = 850.521,00$				

¿Hay diferencias entre los valores del colesterol obtenidos mediante los tratamientos de los tres hipolipemiantes? ¿Hay efecto de bloques? Estas son las preguntas cuya respuesta se quiere obtener, mediante el análisis de la varianza bloques. El contraste se resuelve con  $\alpha = 0,05$ .

A continuación se calculan las sumas de cuadrados. Aplicando la expresión 17.11 se calcula la suma de cuadrados total:

$$SS_{\text{TOTAL}} = 850.521,00 - 838.512,5; SS_{\text{TOTAL}} = 12.008,5$$

Aplicando la expresión 17.12 se calcula la suma de cuadrados grupos:

$$SS_{\text{GRUPOS}} = \left( \frac{1.324^2}{6} + \frac{1.455^2}{6} + \frac{1.106^2}{6} \right) - 838.512,5; SS_{\text{GRUPOS}} = 10.360,32$$

Aplicando la expresión 17.13 se calcula la suma de cuadrados bloques:

$$SS_{\text{BLOQUES}} = \left( \frac{670^2}{3} + \frac{662^2}{3} + \frac{654^2}{3} + \frac{634^2}{3} + \frac{636^2}{3} + \frac{629^2}{3} \right) - 838.512,5$$

$$SS_{\text{BLOQUES}} = 471,82$$

Aplicando la expresión 17.14 se calcula la suma de cuadrados residual:

$$SS_{\text{RESIDUAL}} = 12.008,50 - 10.360,32 - 471,82; SS_{\text{RESIDUAL}} = 1.176,36$$

A continuación teniendo en cuenta los correspondientes grados de libertad se calculan los cuadrados medios:

$$MS_{\text{GRUPOS}} = \frac{10.360,32}{2} \quad MS_{\text{GRUPOS}} = 5.180,16$$

$$MS_{\text{BLOQUES}} = \frac{471,82}{5} \quad MS_{\text{BLOQUES}} = 94,36$$

$$MS_{\text{RESIDUAL}} = \frac{1.176,36}{10} \quad MS_{\text{RESIDUAL}} = 117,63$$

La hipótesis principal a contrastar es si hay diferencias entre las medias de colesterol total correspondientes a los grupos tratados con los tres hipolipemiantes.

$$H_0 \quad \mu_A = \mu_B = \mu_C$$

$$H_1 \quad \mu_i \neq \mu_j \text{ para algún } i, j \quad \alpha = 0,05$$

La hipótesis anterior se resuelve comparando el cuadrado medio entre grupos con el cuadrado medio residual mediante la prueba de la F de Snedecor.

La hipótesis secundaria es comprobar si hay efecto de bloques:

$$H_0 \quad \mu_D = \mu_{BQ} = \mu_{AQ} = \mu_{IE} = \mu_{AR} = \mu_{CL}$$

$$H_1 \quad \mu_i \neq \mu_j \text{ para algún } i, j \quad \alpha = 0,05$$

La hipótesis anterior se resuelve comparando el cuadrado medio entre bloques con el cuadrado medio residual mediante la prueba de la F de Snedecor.

Los resultados de los contrastes se exponen en la siguiente tabla:

#### Análisis de la varianza bloques

Fuente de variación	GL	Suma de cuadrados	Cuadrados medios	F cociente	F prob.
Entre grupos	2	10.360,32	5.180,16	44,14	0,00001
Bloques	5	471,82	94,36	0,80	NS
Residual	10	1.176,36	117,63		
Total	17	12.008,50			

La diferencia entre los colesteroles totales correspondientes a los tres grupos definidos por los tratamientos A, B y C son clínicamente significativos; además, las diferencias son estadísticamente muy significativas, y habrá que dilucidar entre qué medias hay diferencias; esto se hará en los apartados siguientes.

No hay diferencias estadísticamente significativas entre los bloques, es decir, no se ha podido probar que haya efecto bloques; debe tenerse en cuenta que las muestras son pequeñas y, además, haya o no efecto bloques, esto no influye en los resultados principales.

## 17.6. COMPARACIONES MÚLTIPLES

Si al contrastar la hipótesis principal se rechaza la hipótesis nula, hay que dilucidar entre qué medias hay diferencias estadísticamente significativas (evidentemente esto tiene interés práctico si las diferencias son técnicamente significativas), para ello pueden utilizarse las pruebas de comparación múltiple estudiadas en el capítulo anterior, aunque hay que tener en cuenta algunas diferencias. En el caso del análisis de la varianza unifactorial la varianza dentro de grupos es equivalente a la varianza residual en el análisis de la varianza bloques. Cuando en alguna prueba hay que utilizar los grados de libertad correspondientes a la variabilidad dentro de grupos que son  $N - K$ , siendo  $N$  el número total de casos, es decir,  $bk$ , en el análisis de la varianza bloques hay que tener en cuenta que los grados de libertad residual son  $(K - 1)(b - 1)$ . En el análisis de la varianza bloques los grupos tienen el mismo número de casos; una de las pruebas más utilizadas es la de Tukey<sup>1</sup>; en el caso de que las comparaciones se realicen respecto a la media de uno de los grupos se recomienda utilizar la de Dunnett.

En general, haya o no efecto bloques, no se comparan las medias de los bloques, si se quisiera hacerlo habría que considerar a los bloques como grupos.

### EJEMPLO 17.3

Realizar las pruebas de comparación múltiple correspondientes al ejercicio 17.2, con  $\alpha = 0,05$ .

En el ejercicio 17.2 se rechazó la hipótesis nula principal, por lo tanto la conclusión es que hay diferencias estadísticamente significativas entre los tratamientos hipolipemiantes; hay que analizar entre qué medias son las diferencias detectadas. Los colesteroles medios correspondientes a los tres grupos son los siguientes:

$$\bar{x}_A = 220,67 \quad \bar{x}_B = 242,50 \quad \bar{x}_C = 184,33$$

Interesa realizar todas las posibles comparaciones binarias entre las medias. Como todos los grupos tienen el mismo número de casos, seis, se puede aplicar la prueba de Tukey.

Las diferencias entre las medias son:

$$\bar{x}_A - \bar{x}_B = -21,83 \quad \bar{x}_A - \bar{x}_C = 36,37 \quad \bar{x}_B - \bar{x}_C = 58,17$$

Según Tukey, la diferencia mínima que tiene que haber entre dos medias para considerar que hay diferencias estadísticamente significativas, DMS, se calculan mediante la siguiente expresión:

<sup>1</sup> Las características principales de las pruebas de Tukey y de Dunnett se estudian en el Capítulo 16.



$$DMS = q_{\alpha, K, (K-1)(b-1)} \sqrt{\frac{MSR}{b}}$$

En esta caso  $\alpha = 0,05$ ;  $K = 3$ ;  $b = 6$ ;  $MSR = 117,36$ . Por lo tanto:

$$DMS = q_{0,05; 3; 10} \sqrt{\frac{117,36}{6}}$$

Consultando la tabla VIII,  $q_{0,05; 3; 10} = 3,877$ .

$$DMS = 17,15$$

Se considerará que hay diferencias estadísticamente significativas si las diferencias entre dos medias en valor absoluto son mayores que 17,15. En este caso todas las diferencias binarias son mayores que 17,15, la conclusión es que hay diferencias estadísticamente significativas entre todas las medias. El mejor hipocolesterolemiante es el C, seguido del A, y por último del B.

### 17.7. ESTUDIO DE LA INTERACCIÓN: PRUEBA DE NO ADITIVIDAD DE TUKEY

Uno de los supuestos del modelo es que entre el efecto grupos y el efecto bloques, si es que los hay, no hay interacción, es decir, dichos efectos son aditivos. La interacción es un efecto especial, más allá de la simple suma, entre alguno de los grupos y alguno de los bloques, lo que puede ocasionar importantes errores en la interpretación de los resultados.

Estudiar la interacción entre grupos y bloques no es una tarea sencilla porque, entre otras razones, sólo hay un caso por cada una de las posibles combinaciones de grupos y de bloques. Tukey propuso un método denominado *prueba de no aditividad de Tukey*, el cual se expone en la Tabla 17.3.

En dicha tabla, además de los elementos que había en la tabla 17.1, hay unos parámetros que no estaban en ella, los  $d_{Gj}$ , los  $d_{Bi}$  y sus correspondientes cuadrados. Hay un  $d_{Gj}$  por cada grupo, es la diferencia entre la media de su grupo correspondiente y la media global,  $\bar{X}$ :  $d_{Gj} = (\bar{x}_{Gj} - \bar{X})$ ; hay un  $d_{Bi}$  por cada bloque y es la diferencia entre la media de su bloque correspondiente y la media global,  $\bar{X}$ :  $d_{Bi} = (\bar{x}_{Bi} - \bar{X})$ .

Tukey sugiere descomponer la suma de cuadrados residual  $SS_{\text{Residual}}$  en suma de cuadrados debidos a la no aditividad  $SS_{\text{NAD}}$  y la suma de cuadrados correspondiente al resto de circunstancias,  $SS_{\text{Resto}}$ .

$$SS_{\text{Residual}} = SS_{\text{Resto}} + SS_{\text{NAD}} \quad [17.21]$$

1	2	...	...	K				
$X_{11}$	$X_{21}$	...	...	$X_{k1}$	$\bar{x}_{B1}$	$S_{B1}^2$	$d_{B1}$	$d_{B1}^2$
$X_{12}$	$X_{22}$	...	...	$X_{k2}$	$\bar{x}_{B2}$	$S_{B2}^2$	$d_{B2}$	$d_{B2}^2$
...	...	...	...	...	...	...	...	...
$X_{1i}$	$X_{2i}$	...	...	$X_{ki}$	$\bar{x}_{Bi}$	$S_{Bi}^2$	$d_{Bi}$	$d_{Bi}^2$
...	...	...	...	...	...	...	...	...
$X_{1n}$	$X_{2n}$	...	...	$X_{kn}$	$\bar{x}_{Bn}$	$S_{BK}^2$	$d_{Bn}$	$d_{Bn}^2$
$\bar{x}_{G1}$	$\bar{x}_{G2}$	...	...	$\bar{x}_{Gk}$	$S^2$			
$S_{G1}^2$	$S_{G2}^2$	...	...	$S_{GK}^2$				
$d_{G1}$	$d_{G2}$	...	...	$d_{GK}$				
$d_{G1}^2$	$d_{G2}^2$	...	...	$d_{GK}^2$				

**TABLA 17.3. Prueba de no aditividad de Tukey**

La suma de cuadrados de no aditividad se calcula mediante la siguiente expresión:

$$SS_{NAD} = \frac{\left( \sum_{j=1}^K \sum_{i=1}^B X_{ji} d_{Gj} d_{Bi} \right)^2}{\sum_{j=1}^K d_{Gj}^2 \sum_{i=1}^B d_{Bi}^2} \tag{17.22}$$

Despejando en 17.21, se calcula  $SS_{Resto}$  mediante la siguiente expresión:

$$SS_{Resto} = SS_{Residual} - SS_{NAD} \tag{17.23}$$

Los grados de libertad correspondientes a la no aditividad son 1, y los correspondientes al resto  $(K - 1)(B - 1) - 1$ ; recuerde que los grados de libertad residuales son  $(K - 1)(B - 1)$ , y se tiene que cumplir que los grados de libertad residuales son iguales a la suma de los grados de libertad de no aditividad y correspondientes al resto.

Las varianzas correspondientes, es decir, los cuadrados medios, se obtienen dividiendo la suma de cuadrados entre sus correspondientes grados de libertad. El cuadrado medio correspondiente a la no aditividad se obtiene mediante la siguiente expresión:

$$MS_{NAD} = \frac{SS_{NAD}}{1} \tag{17.24}$$

El cuadrado medio correspondiente al resto se obtiene mediante la siguiente expresión:

$$MS_{\text{Resto}} = \frac{SS_{\text{Resto}}}{(K - 1) \cdot (B - 1) - 1} \quad [17.25]$$

El contraste de hipótesis de no aditividad es el siguiente:

$$\begin{array}{ll} H_0 & \text{No interacción entre bloques y grupos.} \\ H_1 & \text{Hay interacción} \quad \alpha. \end{array}$$

Si la variabilidad debida a la no aditividad es significativamente mayor que la variabilidad correspondiente al resto de circunstancias, se rechaza la hipótesis nula y se concluye que hay interacción entre grupos y bloques, lo que indicaría que, en ese caso, no sería aplicable el análisis de la varianza bloques. La evaluación de la variabilidad de no aditividad con respecto a la del resto se realiza mediante la prueba de la F de Snedecor:

$$F_{\alpha, 1, (G-1)(B-1)-1} = \frac{MS_{\text{NAD}}}{MS_{\text{Resto}}} \quad [17.26]$$

Si la  $F$  es estadísticamente significativa al nivel alfa especificado, se rechaza la hipótesis nula y se concluye que hay interacción, en caso contrario se considera que no hay pruebas estadísticas de interacción; consecuentemente, sería aplicable el análisis de la varianza bloques, si se cumplen el resto de supuestos del modelo.

## 17.8. MEDIDAS REPETIDAS

En muchas ocasiones interesa estudiar un determinado parámetro repetidas veces en los mismos elementos de una muestra. En este caso se realiza la primera medida cuando comienza el estudio,  $t_0$ , se hacen una o más medidas antes de acabar el estudio,  $t_i$ , y la última al finalizar el estudio,  $t_f$ , en total  $K$  medidas, siendo  $K > 1$ ; en el caso más sencillo  $K = 2$ , sólo hay dos medidas: al comenzar y al finalizar el estudio, es el caso de la comparación de dos medias con datos pareados o dependientes, que se puede hacer mediante la prueba de la  $t$  de Student (véase Capítulo 15). El análisis de la varianza con medidas repetidas es una comparación de  $K$  medias con datos pareados o dependientes.

A continuación se analizan dos ejemplos teóricos correspondientes a los dos casos más utilizados en la práctica.

Por ejemplo, se quiere estudiar la evolución de un grupo de pacientes hipertensos en relación a un nuevo tratamiento, se selecciona al azar un grupo de  $n$  pa-

cientes, al comenzar se mide la tensión arterial a los  $n$  pacientes participantes, durante el estudio y al finalizar se hacen nuevas medidas de la tensión arterial sistólica, el objetivo es analizar la evolución de la tensión arterial en relación al tiempo de tratamiento para conocer cuánto tiempo tarda en estabilizarse. En este caso interesa comparar todas las medias entre sí; se considerará que la tensión arterial se ha estabilizado cuando no haya diferencias clínica y estadísticamente significativas entre dos medias consecutivas. Se hacen  $K$  medidas de la tensión arterial sistólica en los  $n$  pacientes para estudiar la respuesta en el tiempo a un determinado tratamiento.

El objetivo de un estudio es analizar el efecto hipoglucemiante de  $K$  fármacos en ratas; se seleccionan al azar  $n$  ratas; para evitar las diferencias interindividuales que puedan enmascarar los resultados se prueban los  $k$  fármacos en todas las ratas. Al comenzar el estudio, antes de aplicar ningún tratamiento se mide la glucemia basal, a continuación se aplica el primer tratamiento, y a la semana se vuelve a analizar la glucemia basal; se deja un tiempo sin aplicar ningún tratamiento, tiempo de lavado, y se aplica a los animales un nuevo fármaco; a la semana de tratamiento se vuelve a analizar la glucemia basal, y se vuelve a dejar un tiempo de lavado hasta la aplicación de un nuevo fármaco, y así sucesivamente hasta completar los  $K$  fármacos. En este caso las  $k$  medidas se realizan bajo situaciones experimentales diferentes, una basal y el resto con tratamientos diferentes; se pretende comparar el efecto de los fármacos entre sí y con la situación basal, es decir, interesa realizar todas las comparaciones posibles entre las medias.

Debe tenerse en cuenta que si en un estudio se comprueba que hay diferencias técnica y estadísticamente significativas respecto a un determinado parámetro, no significa que necesariamente sean debidas a la diferencia en tratamientos o a las situaciones que se estén estudiando, los análisis estadísticos indican si existen o no diferencias, pero no las causas de las mismas. Por ejemplo, para comprobar el efecto hipolipemiante de un determinado fármaco se estudia el colesterol LDL basal, es decir, antes de comenzar a tomar el medicamento; al mes y a los seis meses. Las diferencias entre el colesterol LDL basal y los valores al mes y seis meses de tratamiento fueron clínica y estadísticamente significativas; puede ser que las diferencias sean debidas al fármaco o bien que los pacientes diagnosticados de hiperlipemia, al conocer su problema metabólico reduzcan la ingesta de grasas y hagan más ejercicio, de esta manera, al menos en parte, la reducción del colesterol LDL podría ser debida a causas diferentes al tratamiento, el diseño metodológico y la elección de las pruebas adecuadas a cada caso deben solventar estos problemas.

En el caso de medidas repetidas cada individuo es un bloque. En el análisis de la varianza bloques se pretende que los elementos de la muestra sean homogéneos utilizando una variable de bloqueo, en el caso del análisis de la varianza de medidas repetidas los elementos son homogéneos respecto a las  $k$  medidas porque son los mismos en cada serie de observaciones. En estadística aplicada, el análisis de la varianza de medidas repetidas es mucho más utilizado que el análisis de la varianza utilizando variables de bloqueo, sobre todo en ciencias de la salud. En-

tre el análisis de la varianza bloques y el de medidas repetidas hay muchas similitudes, pero también importantes diferencias.

En un estudio que consta de  $n$  elementos y  $K$  medidas a cada uno de ellos los datos y los principales parámetros estadísticos pueden tabularse de la siguiente manera:

1	2	...	...	$K$		
$X_{11}$	$X_{21}$	...	...	$X_{k1}$	$\bar{x}_{E1}$	$S_{E1}^2$
$X_{12}$	$X_{22}$	...	...	$X_{k2}$	$\bar{x}_{E2}$	$S_{E2}^2$
...	...	...	...	...	...	...
$X_{1i}$	$X_{2i}$	...	...	$X_{ki}$	$\bar{x}_{Ei}$	$S_{Ei}^2$
...	...	...	...	...	...	...
$X_{1n}$	$X_{2n}$	...	...	$X_{kn}$	$\bar{x}_{En}$	$S_{En}^2$
$\bar{x}_{M1}$	$\bar{x}_{M2}$	...	...	$\bar{x}_{Mk}$	$\bar{\bar{X}}$	$S^2$
$S_{M1}^2$	$S_{M2}^2$	...	...	$S_{MK}^2$		

**TABLA 17.3**

En la tabla anterior hay algunas diferencias con la Tabla 17.1, correspondiente al análisis de la varianza bloques. Las columnas se corresponden con las medidas, desde 1 hasta  $K$ , en lugar de los grupos utilizados en el análisis de la varianza bloques, por ejemplo, el subíndice  $M2$  significa segunda medida. En las filas los elementos de la muestra se identifican por el subíndice  $E$ , elemento, seguido del orden correspondiente en el muestreo, desde 1 hasta  $n$ .

Igual que en el análisis de la varianza unifactorial y en el análisis de la varianza bloques, si no hay diferencias estadísticamente significativas entre las medidas; entre la varianza entre medidas, la varianza de los elementos de la muestra correspondientes a cada una de las medidas, intramedida, y la varianza global,  $S^2$ , sólo debe haber diferencias explicables por el azar. Si hay diferencias entre las medias debidas a pertenecer a medidas distintas, debe haber diferencias entre la varianza entre medidas y la varianza residual detectables mediante la prueba de la  $F$  de Snedecor. La diferencia fundamental de los análisis de la varianza unifactorial y bloques entre el análisis de la varianza de medidas repetidas, es que en este caso los grupos se corresponden con medidas, es decir, el factor es una variable con  $K$  categorías, siendo cada una de ellas una medida.



FIGURA 22.1

La variable que se quiere contrastar, en este caso Fumar, se pasa a la ventana encabezada por contrastar variables.

En la parte inferior a la izquierda «Definir la dicotomía» se puede optar por que se obtenga la dicotomía de los datos o mediante punto de corte. Obtener de los datos se utiliza cuando la variable que se quiere contrastar es dicotómica, es decir, sólo tiene dos valores definidos. La opción punto de corte se utiliza para variables numéricas con más de dos valores, se consideran pertenecientes a la primera categoría los valores menores o iguales que el valor especificado a la derecha de «Punto de corte», pertenecientes a la otra categoría se consideraran los valores mayores que el punto de corte.

En mitad de la pantalla en la parte inferior está «Contrastar proporción» por defecto el valor indicado es 0,50, este valor puede cambiarlo el usuario.

Dejando el valor 0,50 en contrastar proporción, las hipótesis que se contrastan son las siguientes:

$$\begin{array}{ll}
 H_0 & P(SI) = P(NO) = 0,5 \\
 H_1 & P(SI) \neq P(NO) \quad \alpha = 0,05
 \end{array}$$

En este caso las dos proporciones son iguales, cuando son diferentes, la proporción especificada en «Contrastar proporción» es la correspondiente a la primera categoría.

Pulsando en «Aceptar» se obtienen los siguientes resultados:

## Pruebas no paramétricas

### Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. exacta (bilateral)
Fumar	Grupo 1	Sí	8	0,33	0,50	0,152
	Grupo 2	No	16	0,67		
	Total		24	1,00		

En la tabla anterior se muestran los valores observados en cada categoría: 8 para SÍ y 16 para NO, la proporción de cada una de ellas, 0,33 y 0,67 respectivamente, la proporción de prueba que es la asignada para realizar el contraste, y la significación estadística que en este caso es 0,152, como es mayor que 0,05, no se rechaza la hipótesis nula y se concluye que no hay pruebas estadísticas de que la proporción de la primera categoría sea diferente de 0,5. Recuerde que no rechazar la hipótesis nula no quiere decir que se haya demostrado, de hecho, tampoco se rechazaría para muchos otros valores; pruebe con 0,33, 0,40, 0,38..., verá que tampoco se rechaza la hipótesis nula para estos valores, sería absurdo considerar que se ha demostrado que la proporción es igual a todas ellas.

En la parte inferior derecha de la pantalla 22.1, hay dos posibilidades: «Exactas» y «Opciones».

La primera no se incluye en todas las versiones y permite realizar los cálculos mediante el método de Montecarlo, puede ser útil para hacer simulaciones.

Pulsando opciones aparece la siguiente pantalla:

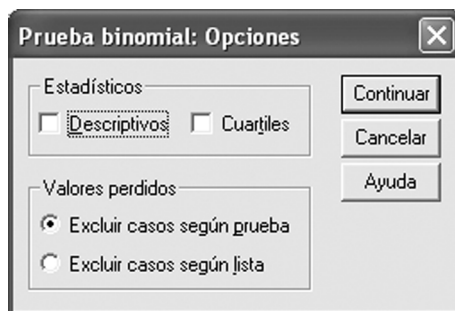


FIGURA 22.2

Si se marca «Descriptivos», en los resultados aparece una tabla con la media, la desviación típica, el número de casos, el máximo y el mínimo. Si se marca

«Cuartiles» se obtienen los valores correspondientes a los cuartiles. Tanto los descriptivos como los cuartiles tienen sentido si la variable es cuantitativa continua, o discreta con muchas categorías. Esta opción puede ser útil cuando se usan variables dicotomizadas mediante punto de corte.

En valores perdidos hay dos posibilidades: «Excluir casos según prueba» y «Excluir casos según lista». La opción por defecto es la primera; en este caso si se especifican varias variables para hacer contrastes, se excluyen los casos con valores ausentes al realizar cada uno de ellos; si se marca «Excluir casos según lista», se excluyen de la realización de todas las pruebas los casos que tengan valores ausentes en alguna de las variables.

### 22.2.2. Bondad del ajuste: prueba $\chi^2$

Una de las pruebas más utilizadas para analizar el ajuste de datos experimentales a distribuciones teóricas es la basada en la distribución  $\chi^2$ . En el Capítulo 14 se analiza esta prueba con detalle y se realizan varios ejemplos; también se estudia cómo realizar ajustes de datos a distribuciones teóricas con SPSS.

La prueba compara las frecuencias observadas con las esperadas, bajo la hipótesis de que en la población muestreada la variable se ajusta a una distribución teórica determinada. El estadístico de contraste es el siguiente:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad [22.3]$$

En la expresión anterior,  $O_i$  representa las frecuencias relativas observadas y  $E_i$  a las frecuencias relativas esperadas,  $K$  es el número de categorías que se comparan. Los grados de libertad del estadístico son  $K - 1$ .

La prueba Chi-cuadrado no es aplicable si más del 20% de las frecuencias esperadas son menores que 5.

Las hipótesis se plantean de la manera siguiente:

- $H_0$  Los datos se ajustan a la distribución teórica.  
 $H_1$  Los datos no se ajustan a la distribución teórica.

El problema de estos contrastes de hipótesis es que si no se rechaza la hipótesis nula <sup>1</sup>, no puede considerarse demostrado que los datos se ajusten a la distribución teórica; sin embargo, si se rechaza la hipótesis nula sí puede considerarse que los datos no se ajustan a la distribución propuesta. Es frecuente que en artículos publicados en revistas e incluso en libros consideren demostrada la hipótesis nula si no se ha rechazado, y se consideran los datos como si se ajustaran

<sup>1</sup> En el Capítulo 11 se analizan las características de los contrastes de hipótesis.



a la distribución teórica propuesta, esto es un error importante y muchas veces trascendente.

### 22.2.3. Pruebas de Kolmogorov-Smirnov; Kolmogorov-Smirnov-Lilliefors y Shapiro Wilks

Esta prueba, desarrollada por Kolmogorov para una muestra y junto con Smirnov para dos muestras, se conoce en ambos casos como prueba de Kolmogorov-Smirnov, y originalmente se desarrolló para comprobar si la distribución empírica de una variable cuantitativa, es decir, la distribución observada de una variable, se ajusta a una distribución teórica conocida; por ejemplo, distribución normal, distribución exponencial, etc. Después se comenzó a aplicar a distribuciones discretas.

La prueba de Kolmogorov-Smirnov se basa en comparar los valores absolutos de las diferencias entre las frecuencias relativas acumulativas experimentales u observadas,  $F_0(x)$ , y las frecuencias relativas acumulativas teóricas o esperadas,  $F_e(x)$ . Las hipótesis son:

$$\begin{array}{ll} H_0 & F_0(x) = F_e(x) \\ H_1 & F_0(x) \neq F_e(x) \end{array} \quad \alpha$$

El estadístico de contraste es el siguiente:

$$\text{Máx } (D = |F_0(x) - F_e(x)|) \quad [22.4]$$

Entre las frecuencias observadas y esperadas se permiten pequeñas diferencias explicables por el azar. El punto crítico, PC, del contraste se obtiene de una tabla si  $n \leq 35$ ; si  $n > 35$  el punto crítico para  $\alpha = 0,05$ , se obtiene mediante la siguiente expresión:

$$\text{PC} = \frac{1,36}{\sqrt{n}} \quad [22.5]$$

Si  $n$  es mayor de 35 y  $\alpha = 0,01$  el punto crítico, PC, se obtiene mediante la siguiente expresión:

$$\text{PC} = \frac{1,63}{\sqrt{n}} \quad [22.6]$$

Si la máxima diferencia observada es mayor que el punto crítico se rechaza la hipótesis nula, en caso contrario no se rechaza.

El principal problema de esta prueba es que se considera que los datos se ajustan a la distribución teórica si no se rechaza la hipótesis nula. Ya se ha comentado en múltiples ocasiones en este libro que las hipótesis nulas no se pueden demostrar estadísticamente; pero si se rechaza la hipótesis nula se considera demostrada la alternativa, es decir, se considera demostrado que no se ajusta a la distribución teórica propuesta, aunque indicando la significación estadística. La prueba de Kolmogorov-Smirnov ha sido muy utilizada para evaluar el ajuste de variables continuas a una distribución normal, en este caso las hipótesis son las siguientes:

$H_0$	La variable se distribuye normalmente.
$H_1$	La variable no se distribuye normalmente.

Es relativamente frecuente encontrar artículos «científicos», incluso libros de estadística, que digan que se ha demostrado que los datos se distribuyen normalmente o que se ajustan a una normal en la población muestreada; esto es claramente erróneo, se podría decir que no se ha rechazado la hipótesis de normalidad, pero nunca que se ha demostrado la normalidad de los datos. Otra cosa es que si no se rechaza la hipótesis de normalidad y el tamaño de la muestra no es muy pequeño se apliquen pruebas que exigen la normalidad, teniendo en cuenta que no se ha rechazado dicha hipótesis.

La prueba de Kolmogorov-Smirnov se considera muy conservadora, es decir, es difícil rechazar la hipótesis nula, sobre todo si la variable es discreta y, aunque sea continua, si hay que estimar los parámetros de la distribución, como la media y la desviación típica en el caso de la distribución normal.

### EJEMPLO 22.3

Un determinado tumor pulmonar se clasifica en cinco tipos distintos, en cuanto a la diferenciación celular, se cree que las cinco se presentan en la misma proporción, es decir, un veinte por ciento, un quinto. Se selecciona una muestra al azar de veinte tumores, obteniéndose las siguientes frecuencias absolutas.

Tipo celular:	1	2	3	4	5
Frecuencia:	4	8	2	2	4

Resolver el contraste con  $\alpha = 0,05$ .

$H_0$	Los cinco tipos celulares tienen la misma proporción.
$H_1$	Al menos una proporción es distinta de las demás.

La tabla de frecuencias relativas acumuladas es la siguiente:

Tipo celular	$F_0(x)$	$F_e(x)$	$D =  F_0(x) - F_e(x) $
1	$1/5 = 0,2$	$4/20 = 0,2$	0
2	$2/5 = 0,4$	$12/20 = 0,6$	0,2
3	$3/5 = 0,6$	$14/20 = 0,7$	0,1
4	$4/5 = 0,8$	$16/20 = 0,8$	0
5	$5/5 = 1$	$20/20 = 1$	0

La máxima diferencia entre las frecuencias relativas acumuladas es 0,2, consultando en la tabla XI de los Anexos correspondiente a la prueba de Kolmogorov, el punto crítico para  $n = 20$  y  $\alpha = 0,05$ , es 0,294, como el valor experimental es menor, no se puede rechazar la hipótesis de que las proporciones de los tipos celulares sean iguales.

### La prueba de Kolmogorov-Smirnov con SPSS

Seleccione en el menú análisis pruebas no paramétricas y en el listado de pruebas disponible: K-S de 1 muestra. Aparece la pantalla siguiente:

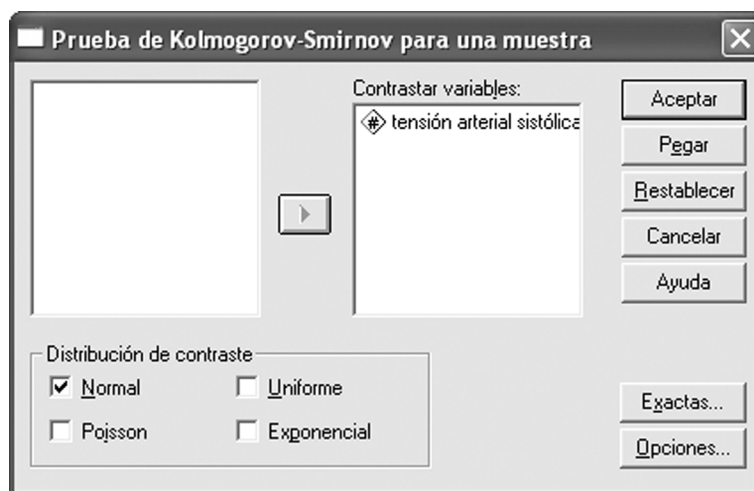


FIGURA 22.3

En la parte inferior a la izquierda en «Distribución de contraste» se puede realizar el ajuste mediante la prueba de Kolmogorov-Smirnov para las distribuciones normal, uniforme, Poisson y exponencial.

Se va a realizar un ajuste a la distribución normal con los 31 datos siguientes que son tensiones arteriales sistólicas, TAS, medidas en milímetros de Hg:

TAS: 115 120 125 125 130 132 132 136 139 139 141 142 142 142 144  
145 146 146 151 152 152 154 155 160 161 162 162 164 165 171

Una vez introducidos los datos, pase la variable TAS a la ventana encabezada por «contrastar variables» y pulse Aceptar, se obtienen los siguientes resultados:

## Pruebas no paramétricas

### Prueba de Kolmogorov-Smirnov para una muestra

		Tensión arterial
<i>N</i>		31
	Media	144,90
Parámetros normales <sup>a, b</sup>	Desviación típica	14,032
Diferencias más extremas	Absoluta	0,085
	Positiva	0,082
	Negativa	-0,085
Z de Kolmogorov-Smirnov		0,472
Sig. asintót. (bilateral)		0,979

<sup>a</sup> La distribución de contraste es la Normal.

<sup>b</sup> Se han calculado a partir de los datos.

En la tabla de resultados en la primera fila se muestra el tamaño de la muestra, que es 31; en la segunda y tercera filas la media y la desviación típica de los datos, que son los parámetros que se utilizan para la distribución normal a la que se ajustan los datos; en las tres filas siguientes se muestran las máximas diferencias entre las frecuencias relativas acumuladas en valor absoluto, la mayor positiva y la mayor negativa; la que se usa para realizar el contraste es la máxima en valor absoluto, es decir 0,85; en la fila siguiente se muestra la Z de Kolmogorov-Smirnov y la significación, que como es mayor que 0,05 no permite rechazar la hipótesis nula que es la de normalidad.

Las teclas virtuales: Exactas y Opciones, ofrecen las mismas posibilidades que se han comentado en la prueba binomial en el apartado 22.2.1.

### ***Ajustes de normalidad: las pruebas de Kolmogorov-Smirnov Lilliefors y de Shapiro Wilks***

En la actualidad las pruebas más aplicadas para el ajuste de los datos a una distribución normal son la de Kolmogorov-Smirnov, con la modificación de Lilliefors si el tamaño de la muestra es mayor que 30 y la de Shapiro Wilks si el tamaño de la muestra es menor o igual que 30.

Lilliefors realizó una modificación en la prueba de Kolmogorov-Smirnov para el ajuste de datos de una variable cuantitativa a una distribución normal, mejorando su potencia estadística. Observe que la prueba de Kolmogorov-Smirnov se desarrolló para ajustar los datos a cualquier distribución de variables continuas, también se utilizó para el ajuste a distribuciones de variables discretas, mientras que la modificación de Lilliefors es válida únicamente para el ajuste a distribuciones normales.

Las hipótesis que contrastar son:

$H_0$	La variable se distribuye normalmente.	
$H_1$	La variable no se distribuye normalmente.	$\alpha$

El estadístico de contraste es el mismo que en la prueba de Kolmogorov-Smirnov:  $\text{Máx} (D = |F_0(x) - F_e(x)|)$ . Cambia la tabla de puntos críticos. La estimación de los parámetros de la curva normal: media y desviación típica, se realiza a partir de los parámetros muestrales. Es la prueba más adecuada si  $n > 30$ .

La prueba de Shapiro-Wilks es la más utilizada si  $n \leq 30$ . Su fundamento es comparar cuantil <sup>1</sup> a cuantil, el valor esperado bajo la hipótesis de que los datos se distribuyen según una normal con la media y la desviación típica de los datos, con el cuantil observado:

$$E\left(\frac{x_i - \mu}{\sigma}\right) = c_{i,n} \quad [22.7]$$

Despejando en la ecuación anterior y teniendo en cuenta que el valor esperado de una constante es ella misma:

$$E(x_i) = \mu + c_{i,n}\sigma \quad [22.8]$$

La expresión anterior es la ecuación de una recta en la que la variable dependiente es el valor esperado bajo la hipótesis de normalidad y la variable independiente el cuantil observado.

El estadístico de contraste se basa en el cuadrado del coeficiente de correlación entre la variable dependiente y la independiente de la ecuación anterior.

<sup>1</sup> Cuantil a cuantil quiere decir dato a dato.

Con SPSS se va a estudiar el ajuste a la normal mediante las dos pruebas anteriores con los datos ajustados en la prueba de Kolmogorov-Smirnov con SPSS.

Las pruebas de Shapiro-Wilks y de Kolmogorov-Smirnov Lilliefors con SPSS se obtienen seleccionando en analizar estadísticos descriptivos y dentro de las opciones «Explorar» pulsando se obtiene la pantalla siguiente.

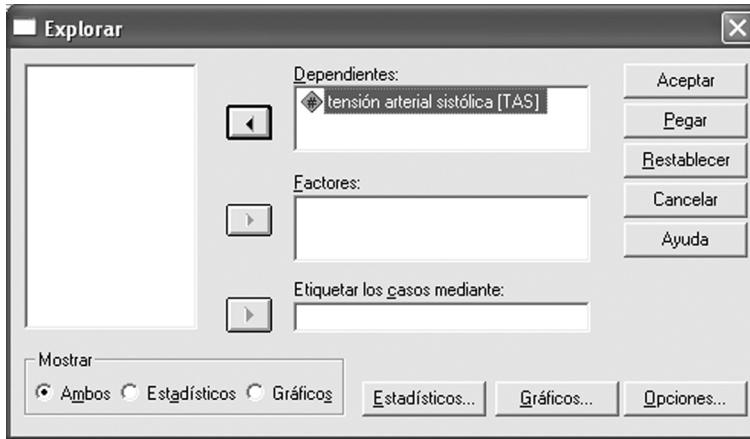


FIGURA 22.5

En la ventana «Dependientes» debe estar la variable que se quiere ajustar. Pulsando la tecla «Gráficos» se obtiene la pantalla siguiente:

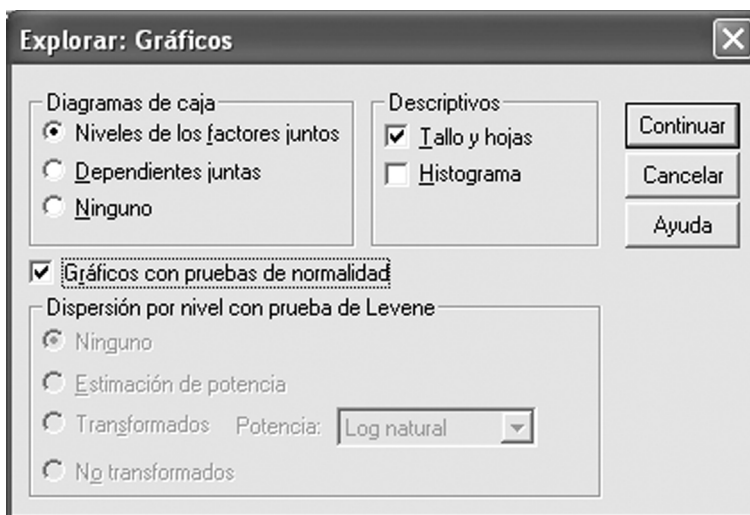


FIGURA 22.6

Si se quiere realizar el ajuste a la normal hay que marcar «Gráficos con pruebas de normalidad».

Se obtienen tablas de estadísticos descriptivos y un histograma. Además, que es lo que se va a comentar en este apartado, una tabla con las pruebas que se quieren realizar y dos gráficos de normalidad:

### Pruebas de normalidad

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Tensión arterial sistólica	0,085	31	0,200*	0,979	31	0,781

\* Este es un límite inferior de la significación verdadera.

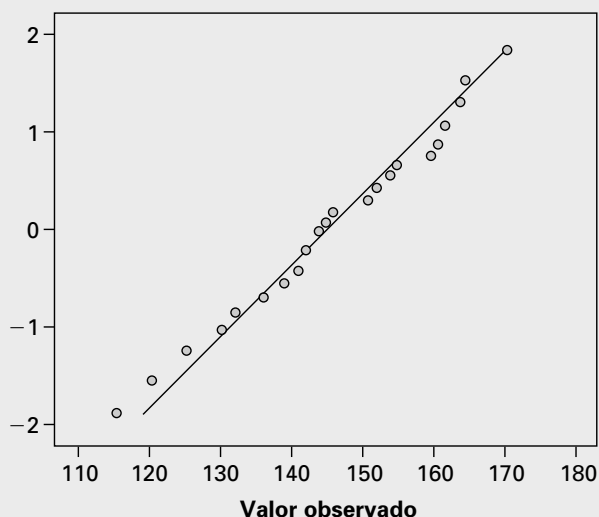
<sup>a</sup> Corrección de la significación de Lilliefors.

Como hay 31 casos, la prueba aplicable es la de Kolmogorov-Smirnov Lilliefors. La máxima diferencia observada entre las frecuencias relativas acumuladas observada y esperada bajo la hipótesis de normalidad es 0,085, que, según la tabla de Lilliefors, corresponde a una significación estadística de 0,2; como es mayor que 0,05 no se rechaza la hipótesis nula, que es la de normalidad.

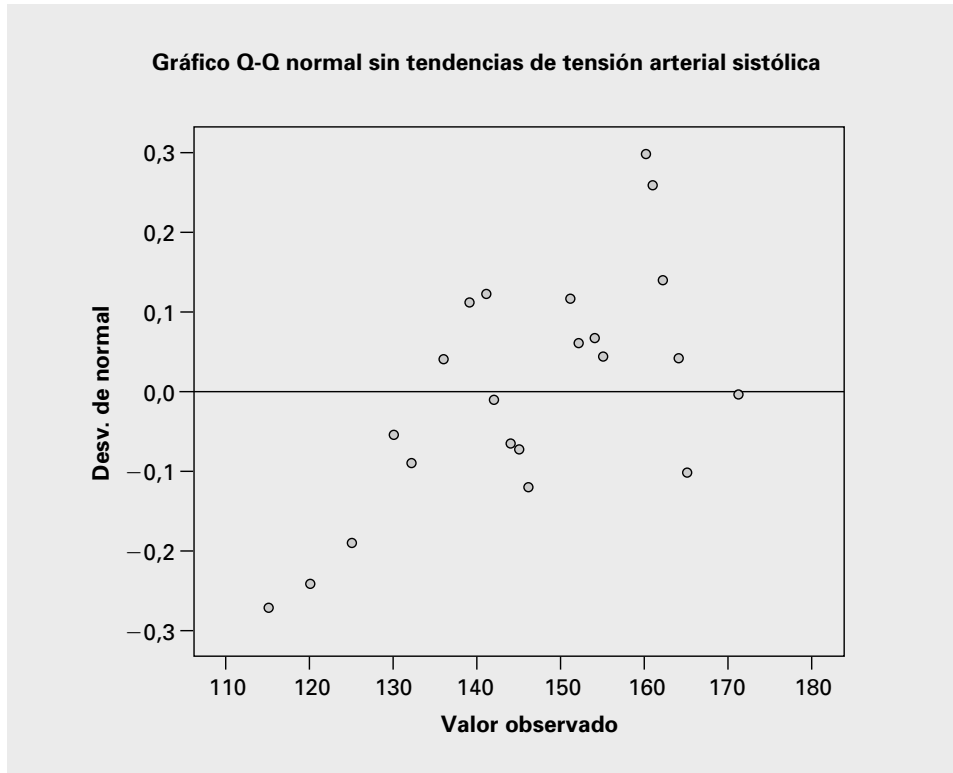
El estadístico de Shapiro-Wilks es 0,979, que se corresponde con una significación de 0,781, que es mayor que 0,05, consecuentemente, no se puede rechazar la hipótesis de normalidad.

No siempre coinciden las dos pruebas en el rechazo o no de la hipótesis nula, se debe utilizar la que corresponda al tamaño de la muestra.

Gráfico Q-Q normal de tensión arterial sistólica



En el caso de un ajuste perfecto a una distribución normal los puntos estarían en la recta.



En una normal perfecta los datos deben distribuirse por encima y debajo de la línea cero al azar.

#### 22.2.4. Pruebas de aleatoriedad: prueba de las rachas

Las tres pruebas analizadas anteriormente analizan la bondad del ajuste de un conjunto de datos a distribuciones teóricas; tienen en cuenta si las frecuencias observadas difieren significativamente de las frecuencias esperadas en el supuesto de que la variable siga una determinada distribución estadística.

La prueba de las rachas analiza si el orden del muestreo es compatible con la aleatoriedad, tiene en cuenta las frecuencias y el orden de los valores en el muestreo. Podría ocurrir que de 20 observaciones 10 tuvieran un valor y 10 otro, pero su orden de observación no fuera el adecuado para considerar la aleatoriedad del muestreo. Para poder realizar esta prueba, es necesario conocer el orden de observación de los datos. Si se ha alterado este orden por haber clasificado los datos