Dividiendo el numerador y el denominador por $2N$, eso se convierte en

$$p = \frac{P + \frac{z_c^2}{2N} \pm z_c\sqrt{\frac{P(1-P)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

(*b*) Para límites de confianza 99.73%, $z_c = 3$. Entonces, usando $P = 0.55$ y $N = 100$ en la fórmula deducida en (*a*), vemos que $p = 0.40$ y 0.69, de acuerdo con el Problema 9.10(*c*).

(*c*) Si N es grande, entonces $z_c^2/(2N)$, $z_c^2/(4N^2)$ y z_c^2/N son todos despreciables y pueden tomarse esencialmente como cero, así que se llega al resultado deseado.

9.13. En 40 lanzamientos de una moneda, han salido 24 caras. Hallar los límites de confianza (*a*) 95% y (*b*) 99.73% para la proporción de caras que se obtendrían en un número ilimitado de lanzamientos de esa moneda.

Solución

(*a*) Al nivel 95%, $z_c = 1.96$. Haciendo $P = 24/40 = 0.6$ y $N = 40$ en la fórmula del Problema 9.12(*a*), hallamos $p = 0.45$ y 0.74. Luego podemos decir que, con 95% de confianza, p está entre 0.45 y 0.74.

Usando la fórmula aproximada $p = P \pm z_c\sqrt{P(1-P)/N}$, deducimos $p = 0.60 \pm 0.15$, que da al intervalo de 0.45 a 0.75.

(*b*) Al nivel 99.73%, $z_c = 3$. Usando la fórmula del Problema 9.12(*a*), hallamos $p = 0.37$ y 0.79.

Mediante la fórmula aproximada $p = P \pm z_c\sqrt{P(1-P)/N}$, hallamos $p = 0.60 \pm 0.23$, que da el intervalo de 0.37 a 0.83.

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.14. Una muestra de 150 lámparas del tipo A ha dado una vida media de 1400 horas (h) y una desviación típica de 120 h. Una muestra de 200 lámparas del tipo B dan vida media de 1200 h y desviación típica de 80 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la diferencia de las vidas medias de las poblaciones de ambos tipos.

Solución

Los límites de confianza para la diferencia en medias de los dos tipos A y B vienen dados por

$$\bar{X}_A - \bar{X}_B \pm z_c\sqrt{\sigma_A^2/N_A + \sigma_B^2/N_B}$$

(*a*) Los límites de confianza 95% son $1400 - 1200 \pm 1.96\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 24.8$. Luego tenemos 95% de confianza de que la diferencia de las medias de las poblaciones está entre 175 y 225 h.

(*b*) Los límites de confianza 99% son $1400 - 1200 \pm 2.58\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 32.6$. Por tanto, tenemos 99% de confianza de que la diferencia de las medias de las poblaciones esté entre 167 y 233 h.

9.15. En una muestra aleatoria de 400 adultos y 600 jóvenes que vieron un cierto programa de televisión, 100 adultos y 300 jóvenes reconocieron que les había gustado. Determinar los límites de confianza (*a*) 95% y (*b*) 99% para la diferencia en proporciones de todos los adultos y jóvenes que vieron con agrado el programa.

Solución

Los límites de confianza para las diferencias en proporciones de los dos grupos vienen dados por

$$P_1 - P_2 \pm z_c\sqrt{p_1q_1/N_1 + p_2q_2/N_2}$$

donde los subindices 1 y 2 se refieren a jóvenes y adultos, respectivamente. Aquí, $P_1 = 300/600 = 0.50$ y $P_2 = 100/400 = 0.25$ son, respectivamente, las proporciones de jóvenes y de adultos a quienes agradó el programa.

(*a*) Los límites de confianza 95% son $0.50 - 0.25 \pm 1.96\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.06$. Luego tenemos 95% de confianza de que la verdadera diferencia en proporciones está entre 0.19 y 0.31.

(*b*) Los límites de confianza 99% son $0.50 - 0.25 \pm 2.58\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.08$. Luego tenemos 99% de confianza de que la verdadera diferencia en proporciones está entre 0.17 y 0.33.

9.16. La fuerza electromotriz media (fem) de las baterías producidas por una empresa es 45.1 voltios (V) y su desviación típica 0.04 V. Si se conectan en serie cuatro de ellas, hallar (*a*) 95%, (*b*) 99%, (*c*) 99.73% y (*d*) 50%.

Solución

Si E_1, E_2, E_3 y E_4 representa la fem de las cuatro baterías, tenemos

$$\mu_{E1+E2+E3+E4} = \mu_{E1} + \mu_{E2} + \mu_{E3} + \mu_{E4} \quad \text{y} \quad \sigma_{E1+E2+E3+E4} = \sqrt{\sigma_{E1}^2 + \sigma_{E2}^2 + \sigma_{E3}^2 + \sigma_{E4}^2}$$

Entonces, como $\mu_{E1} = \mu_{E2} = \mu_{E3} = \mu_{E4} = 45.1$ V y $\sigma_{E1} = \sigma_{E2} = \sigma_{E3} = \sigma_{E4} = 0.04$ V, tenemos $\mu_{E1+E2+E3+E4} = 4(45.1) = 180.4$ y $\sigma_{E1+E2+E3+E4} = \sqrt{4(0.04)^2} = 0.08$.

(*a*) Los límites de confianza 95% son $180.4 \pm 1.96(0.08) = 180.4 \pm 0.16$ V.
(*b*) Los límites de confianza 99% son $180.4 \pm 2.58(0.08) = 180.4 \pm 0.21$ V.
(*c*) Los límites de confianza 99.73% son $180.4 \pm 3(0.08) = 180.4 \pm 0.24$ V.
(*d*) Los límites de confianza 50% son $180.4 \pm 0.6745(0.08) = 180.4 \pm 0.054$ V. El valor 0.054 V se llama el *error probable*.

INTERVALOS DE CONFIANZA PARA DESVIACION TIPICA

9.17. La desviación típica de las vidas medias de una muestra de 200 bombillas es de 100 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de ese tipo de bombillas.

Solución

Los límites de confianza para la desviación típica de la población σ vienen dados por $s \pm z_c\sigma/\sqrt{2N}$, donde z_c indica el nivel de confianza. Usamos la desviación típica muestral para estimar σ.

(*a*) Los límites de confianza 95% son $100 \pm 1.96(100)/\sqrt{400} = 100 \pm 9.8$. Luego tenemos 95% de confianza de que la desviación típica de la población está entre 90.2 y 109.8 h.

(*b*) Los límites de confianza 99% son $100 \pm 2.58(100)/\sqrt{400} = 100 \pm 12.9$. Luego tenemos 99% de confianza de que la desviación típica de la población está entre 87.1 y 112.9 h.

9.18. ¿De qué tamaño ha de tomarse una muestra de las bombillas del Problema 9.17 para tener 99.73% de confianza de que la verdadera desviación típica de la población no difiere de la desviación típica muestral en más de (*a*) 5% y (*b*) 10%?

Solución

Los límites de confianza 99% para σ son $s \pm 3\sigma/\sqrt{2N} = s \pm 3s/\sqrt{2N}$, usando s como estimación de σ. Luego el porcentaje de error en la desviación típica es

$$\frac{3s/\sqrt{2N}}{s} = \frac{300}{\sqrt{2N}}\%$$

(*a*) Si $300/\sqrt{2N} = 5$, entonces $N = 1800$. Luego la muestra ha de ser de al menos 1800 bombillas.
(*b*) Si $300/\sqrt{2N} = 10$, entonces $N = 450$. Por tanto, es necesaria una muestra de 450 o más bombillas.

ERROR PROBABLE

9.19. Los voltajes de 50 baterías del mismo tipo tienen una media de 18.2 V y una desviación típica de 0.5 V. Hallar (*a*) el error probable de la media y (*b*) los límites de confianza 50%.

Solución

(*a*) Error probable de la media $= 0.674\sigma_{\bar{X}} = 0.6745\dfrac{\sigma}{\sqrt{N}} = 0.6745\dfrac{\hat{s}}{\sqrt{N}}$

$$= 0.6745\frac{s}{\sqrt{N-1}} = 0.6745\frac{0.5}{\sqrt{49}} = 0.048 \text{ V}$$

Nótese que si la desviación típica de 0.5 V se toma como $\hat{s}$, el error probable es $0.6745(0.5/\sqrt{50}) = 0.048$ también, de modo que cualquier estimación puede utilizarse cuando N es lo bastante grande.

(*b*) Los límites de confianza 50% son 18 ± 0.048 V.

9.20. Se ha anotado una medida como 216.480 gramos (g) con un error probable de 0.272 g. ¿Cuáles son los límites de confianza 95% para esa medida?

Solución

El error probable es $0.272 = 0.6745\sigma_{\bar{X}}$, es decir, $\sigma_{\bar{X}} = 0.272/0.6745$. Luego los límites de confianza 95% son $\bar{X} \pm 1.96\sigma_{\bar{X}} = 216.480 \pm 1.96(0.272/0.6745) = 216.480 \pm 0.790$ g.

PROBLEMAS SUPLEMENTARIOS

ESTIMACIONES SIN SESGO Y EFICIENTES

9.21. Mediciones de una muestra de masas dieron 8.3, 10.6, 9.7, 8.8, 10,2 y 9.4 kilogramos (kg), respectivamente. Determinar estimaciones sin sesgo y eficientes de (*a*) la media de la población y (*b*) la varianza de la población, y comparar la desviación típica de la muestra con la estimada para la población.

9.22 Una muestra de 10 tubos de televisión procedentes de una cierta empresa dieron una vida media de 1200 h y una desviación típica de 100 h. Estimar (*a*) la media y (*b*) la desviación típica de la población de todos los tubos de esa clase.

9.23. (*a*) Rehacer el Problema 9.22 si los mismos

resultados se hubiesen dado con 30, 50, y 100 tubos.

(*b*) ¿Qué se puede concluir sobre la relación entre desviaciones típicas muestrales y estimaciones de las desviaciones típicas de la población para diferentes tamaños de las muestras?

INTERVALOS DE CONFIANZA PARA MEDIAS

9.24. La media y la desviación típica de las cargas máximas soportadas por 60 cables (véase Prob. 3.59) son 11.09 y 0.73 toneladas, respectivamente. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la media de las cargas máximas soportadas por los cables de ese tipo.

9.25. La media y la desviación típica de los diámetros de una muestra de 250 remaches manufacturados por una empresa, son 0.72642 y 0.00058 in, respectivamente (véase Problema 3.61). Hallar los límites de confianza (*a*) 99%, (*b*) 98%, (*c*) 95% y (*d*) 90% para el diámetro medio de los remaches allí producidos.

9.26. Hallar (*a*) los límites de confianza 50% y (*b*) el error probable de los diámetros del Problema 9.25.

9.27. Si la desviación típica de las vidas medias de los tubos de televisión se estima en 100 h, ¿cómo de grande ha de ser una muestra para tener confianza del (*a*) 95%, (*b*) 90%, (*c*) 99% y (*d*) 99.73% de que el error en la vida media estimada no supera 20 h?

9.28. Idem si el error no debe superar 10 h.

9.29. Una empresa dispone de 500 cables, de los que una muestra de 40 elegidos al azar revela una tensión de ruptura media de 2400 lb y una desviación típica de 150 lb.

(*a*) Hallar los límites de confianza 95% y 99% para la estimación de la tensión media de ruptura de los 460 cables restantes.

(*b*) ¿Con qué grado de confianza se puede decir que la tensión media de ruptura de los 460 restantes es 2400 ± 35 lb?

INTERVALOS DE CONFIANZA PARA PROPORCIONES

9.30. Una urna contiene una proporción desconocida de fichas rojas y blancas. Una muestra aleatoria de 60 fichas, seleccionada con reposición, indicó que el 70% de ellas eran rojas. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la proporción real de fichas rojas en la urna. Presentar los resultados usando tanto la fórmula aproximada como la más exacta del Problema 9.12.

9.31. ¿De qué tamaño ha de ser una muestra de las fichas del Problema 9.30 para tener confianza del (*a*) 95%, (*b*) 99% y (*c*) 99.73% de que la verdadera proporción no difiere de la muestral en más del 5%?

9.32. Se espera que una elección entre dos candidatos sea muy reñida. ¿Cuál es el mínimo número de votantes a sondear si se quiere tener un (*a*) 80%, (*b*) 90%, (*c*) 95% y (*d*) 99% de confianza sobre la decisión a favor de uno u otro?

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.33. De dos grupos similares de pacientes, *A* y *B*, con 50 y 100 individuos respectivamente, se suministró al *A* un nuevo tipo de somnífero y al *B* uno convencional. Para los del grupo *A* el número medio de horas de sueño fue 7.82 con desviación típica de 0.24 h. Para los del grupo *B*, 6.75 h y 0.30 h, respectivamente. Hallar los límites de confianza (*a*) 95% y (*b*) 99%, para la diferencia en media de las horas de sueño inducidas por ambos somníferos.

9.34. Una muestra de 200 tuercas de una cierta máquina probó que 15 eran defectuosas, mientras una muestra de 100 tuercas de otra máquina dio 12 defectuosas. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la diferencia en proporciones de tuercas defectuosas de las dos máquinas. Discutir los resultados obtenidos.

9.35. Una compañía produce bolas de cojinetes de peso medio 0.638 lb y desviación típica de 0.012 lb. Hallar los límites de confianza (*a*)

95% y (*b*) 99% para los pesos de lotes de 100 bolas cada uno.

INTERVALOS DE CONFIANZA PARA DESVIACION TIPICA

9.36. La desviación típica de las tensiones de ruptura de 100 cables probados por una empresa era de 180 lb. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la desviación típica de todos los cables de ese tipo.

9.37. Hallar el error probable de la desviación típica en el Problema 9.36.

9.38. ¿Cómo ha de ser de grande una muestra para tener confianza del (*a*) 95%, (*b*) 99% y (*c*) 99.73% de que la desviación típica de una población no diferirá de la desviación típica muestral en más del 2%?

CAPITULO 10

Teoría estadística de las decisiones

DECISIONES ESTADISTICAS

En la práctica nos vemos obligados con frecuencia a tomar decisiones relativas a una población sobre la base de información proveniente de muestras. Tales decisiones se llaman *decisiones estadísticas*. Por ejemplo, podemos querer decidir, basados en datos muestrales, si un método pedagógico es mejor que otro, o si una moneda está trucada o no.

HIPOTESIS ESTADISTICAS

Al intentar alcanzar una decisión, es útil hacer hipótesis (o conjeturas) sobre la población implicada. Tales hipótesis, que pueden ser o no ciertas, se llaman *hipótesis estadísticas*. Son, en general, enunciados acerca de las distribuciones de probabilidad de las poblaciones.

Hipótesis nula

En muchos casos formulamos una hipótesis estadística con el único propósito de rechazarla o invalidarla. Así, si queremos decidir si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea, $p = 0.5$, donde p es la probabilidad de cara). Análogamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que *no* hay *diferencia* entre ellos (o sea, que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la *misma* población). Tales hipótesis se suelen llamar *hipótesis nula* y se denotan por H_0.

Hipótesis alternativa

Toda hipótesis que difiera de una dada se llamará una *hipótesis alternativa*. Por ejemplo, si una hipótesis es $p = 0.5$, hipótesis alternativas podrían ser $p = 0.7$, $p \neq 0.5$ o $p > 0.5$. Una hipótesis alternativa a la hipótesis nula se denotará por H_1.

CONTRASTES DE HIPOTESIS Y SIGNIFICACION, O REGLAS DE DECISION

Si suponemos que una hipótesis particular es cierta pero vemos que los resultados hallados en una muestra aleatoria difieren notablemente de los esperados bajo tal hipótesis (o sea, esperados sobre la base del puro azar, por teoría de muestreo), entonces diremos que las diferencias observadas son *significativas* y nos veríamos inclinados a rechazar la hipótesis (o al menos a no aceptarla ante la evidencia obtenida). Así, si en 20 tiradas de una moneda salen 16 caras, estaríamos inclinados a rechazar la hipótesis de que la moneda es buena, aunque cabe la posibilidad de equivocarnos.

Los procedimientos que nos capacitan para determinar si las muestras observadas difieren significativamente de los resultados esperados, y por tanto nos ayudan a decidir si aceptamos o rechazamos hipótesis, se llaman *contrastes* (o *tests*) de *hipótesis* o *de significación* o *reglas de decisión*.

ERRORES DE TIPO I Y DE TIPO II

Si rechazamos una hipótesis cuando debiera ser aceptada, diremos que se ha cometido un *error de Tipo I*. Por otra parte, si aceptamos una hipótesis que debiera ser rechazada, diremos que se ha cometido un *error de Tipo II*. En ambos casos, se ha producido un juicio erróneo.

Para que las reglas de decisión (o contrastes de hipótesis) sean buenas, deben diseñarse de modo que minimicen los errores de la decisión. Y no es una cuestión sencilla, porque para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse un compromiso que disminuya el error más grave. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra, que no siempre es posible.

NIVEL DE SIGNIFICACION

Al contrastar una cierta hipótesis, la máxima probabilidad con la que estamos dispuestos a correr el riesgo de cometer un error de Tipo I se llama *nivel de significación* del contraste. Esta probabilidad, denotada a menudo por α, se suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en nuestra elección.

En la práctica, es frecuente un nivel de significación de 0.05 ó 0.01, si bien se usan otros valores. Si, por ejemplo, se escoge el nivel de significación 0.05 (o 5%) al diseñar una regla de decisión, entonces hay unas 5 oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado; es decir, tenemos un 95% *de confianza* de que hemos adoptado la decisión correcta. En tal caso decimos que la hipótesis ha sido rechazada al nivel de significación 0.05, lo cual quiere decir que la hipótesis tiene una probabilidad 0.05 de ser falsa.

CONTRASTES MEDIANTE LA DISTRIBUCION NORMAL

Para ilustrar las ideas presentadas hasta este momento, supongamos que bajo cierta hipótesis la distribución de muestreo de un estadístico S es una distribución normal con media μ_S y desviación

típica σ_S. Así pues, la distribución de la variable tipificada z, dada por $z = (S - \mu_S)/\sigma_S$, es la distribución normal canónica (media 0, varianza 1), como indica la Figura 10.1.

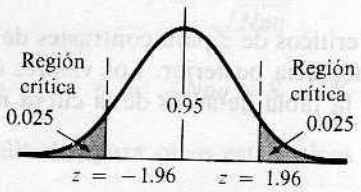


Figura 10.1.

Como se ve en la Figura 10.1, podemos tener 95% de confianza de que si la hipótesis es verdadera, entonces el valor de z para un estadístico muestral S estará entre -1.96 y 1.96 (porque el área bajo la curva normal entre esos valores es 0.95). Sin embargo, si al escoger una sola muestra al azar hallamos que el valor de z de su estadístico está *fuera* de ese rango, debemos concluir que tal suceso podría ocurrir con una probabilidad de sólo 0.05 (el área total sombreada en la figura) si la hipótesis dada fuera cierta. Diremos entonces que esta z difiere *de forma significativa* de lo que sería de esperar bajo la hipótesis, y nos veríamos empujados a rechazar la hipótesis.

El área total sombreada 0.05 es el nivel de significación del contraste. Representa la probabilidad de equivocarnos al rechazar la hipótesis (o sea, la probabilidad de un error de Tipo I). Así pues, decimos que la hipótesis *se rechaza a un nivel de significación 0.05*, o que el valor de z del estadístico muestral dado es *significativo al nivel 0.05*.

El conjunto de z fuera del rango -1.96 a 1.96 se llama *la región crítica de la hipótesis región de rechazo de la hipótesis*, o *región de significación*. El conjunto de z en el rango -1.96 a 1.96 se conoce como *región de aceptación de la hipótesis* o *región de no significación*.

Basados en las anteriores observaciones, podemos formular la siguiente regla de decisión (o contraste de hipótesis o significación):

Rechazar la hipótesis al nivel de significación 0.05 si el valor de z para el estadístico S está fuera del rango -1.96 a 1.96 (o sea, si $z > 1.96$ o $z < -1.96$). Esto equivale a decir que el estadístico muestral observado es significativo al nivel 0.05.

Aceptar la hipótesis en caso contrario (o, si se desea, no tomar decisión alguna).

Dado que z juega tan importante papel en el contraste de hipótesis, se le llama un *estadístico de contraste*.

Hay que hacer notar que se utilizan también otro nivel de significación. Por ejemplo, si se usa el nivel 0.01, debe sustituirse el 1.96 de antes por 2.58 (véase Tabla 10.1). Cabe utilizar asimismo la Tabla 9.1, ya que la suma de los niveles de significación y de confianza es 100%.

CONTRASTES DE UNA Y DE DOS COLAS

En el test precedente estábamos interesados en los valores extremos del estadístico S o en su correspondiente valor de z a *ambos* lados de la media (o sea, en las dos colas de la distribución). Tales tests se llaman *contrastes de dos colas* o *bilaterales*.

Con frecuencia, no obstante, estaremos interesados tan sólo en valores extremos a un lado de la media (o sea, en una de las colas de la distribución), tal como sucede cuando se contrasta la

hipótesis de que un proceso es mejor que otro (lo cual no es lo mismo que contrastar si un proceso es mejor o peor que el otro). Tales contrastes se llaman *unilaterales*, o *de una cola*. En tales situaciones, la región crítica es una región situada a un lado de la distribución, con área igual al nivel de significación.

La Tabla 10.1, que da valores críticos de z para contrastes de una o dos colas en varios niveles de significación, será útil como referencia posterior. Los valores críticos de z para otros niveles de significación se hallan a partir de la tabla de áreas de la curva normal (Apéndice II).

Tabla 10.1

Nivel de significación, α	0.10	0.05	0.01	0.005	0.002
Valores críticos de z para tests unilaterales	-1.28 o 1.28	-1.645 o 1.645	-2.33 o 2.33	-2.58 o 2.58	-2.88 o 2.88
Valores críticos de z para tests bilaterales	-1.645 y 1.645	-1.96 y 1.96	-2.58 y 2.58	-2.81 y 2.81	-3.08 y 3.08

CONTRASTES ESPECIALES

Para grandes muestras, las distribuciones de muestreo de muchos estadísticos son distribuciones normales (o casi normales), y los contrastes anteriores pueden aplicarse a los z correspondientes. Los siguientes casos especiales, tomados de la Tabla 8.1, no son sino unos pocos de los estadísticos de interés práctico. En cada caso los resultados son válidos para poblaciones infinitas o para muestreos con reposición. Para muestreos sin reposición en poblaciones finitas, esos resultados requieren modificación (véase pág. 186).

1. **Medias.** Aquí $S = \bar{X}$, la media muestral; $\mu_S = \mu_{\bar{X}} = \mu$, la media de la población; y $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{N}$, donde σ es la desviación típica de la población y N el tamaño de la muestra. El valor z viene dado por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

Cuando sea necesario, se utilizará la desviación muestral s o $\hat{s}$ como estimación de σ.

2. **Proporciones.** Ahora $S = P$, la proporción de «éxitos» en una muestra; $\mu_S = \mu_P = p$, donde p es la proporción de éxitos de la población y N el tamaño de la muestra; y $\sigma_S = \sigma_P = \sqrt{pq/N}$, donde $q = 1 - p$.

El valor de z viene dado por

$$z = \frac{P - p}{\sqrt{pq/N}}$$

En el caso $P = X/N$, donde X es el número real de éxitos en una muestra, z es

$$z = \frac{X - N_p}{\sqrt{Npq}}$$

Esto es, $\mu_X = \mu = Np$, $\sigma_X = \sigma = \sqrt{Npq}$ y $S = X$.

Análogamente se obtienen los resultados para otros estadísticos.

CURVAS DE OPERACION CARACTERISTICAS; POTENCIA DE UN CONTRASTE

Hemos visto cómo limitar el error de Tipo I eligiendo adecuadamente el nivel de significación. Es posible evitar el riesgo de cometer error de Tipo II simplemente no aceptando nunca hipótesis, pero en muchas aplicaciones prácticas esto es inviable. En tales casos, se suele recurrir a *curvas de operación características*, o *curvas OC*, que son gráficos que muestran las probabilidades de error de Tipo II bajo diversas hipótesis. Proporcionan indicaciones de hasta qué punto un test dado nos permitirá evitar un error de Tipo II; es decir, nos indicará la *potencia de un test* a la hora de prevenir decisiones erróneas. Son útiles en el diseño de experimentos porque sugieren entre otras cosas el tamaño de muestra a manejar.

GRAFICOS DE CONTROL

A menudo adquiere importancia práctica saber cuándo un proceso ha variado tanto que deben adoptarse medidas para remediar la situación. Tales problemas aparecen, por ejemplo, en el control de calidad. Los supervisores del control de calidad han de decidir frecuentemente si los cambios observados se deben simplemente a fluctuaciones de azar o a cambios reales en un proceso de producción por deterioro de la maquinaria, descuidos de los empleados, etc. Los *gráficos de control* ponen a nuestra disposición un método sencillo y eficaz para enfrentarnos a esa clase de problemas (véase Prob. 10.16).

CONTRASTES MEDIANTE DIFERENCIAS MUESTRALES

Diferencias de medias

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales obtenidas en grandes muestras de tamaños N_1 y N_2 tomadas de poblaciones con respectivas medias μ_1 y μ_2, y desviaciones típicas σ_1 y σ_2. Consideremos la hipótesis nula de que *no hay diferencia* entre las medias de las poblaciones (o sea, $\mu_1 = \mu_2$), que es como afirmar que las muestras se han tomado en dos poblaciones que tienen la misma media.

Poniendo $\mu_1 = \mu_2$ en la ecuación (5) del Capítulo 8, vemos que la distribución de muestreo de diferencia en medias está casi normalmente distribuida, con media y desviación típica dadas por

$$\mu_{\bar{X}1-\bar{X}2} = 0 \qquad \text{y} \qquad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{1}$$

donde podemos, si es necesario, usar las desviaciones típicas muestrales s_1 y s_2 (o $\hat{s}_1$ y $\hat{s}_2$) como estimaciones de σ_1 y σ_2.

Usando la variable tipificada z dada por

$$z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} \tag{2}$$

podemos contrastar la hipótesis nula frente a hipótesis alternativas (o la significación de una diferencia observada) a un nivel de significación apropiado.

Diferencias de proporciones

Sean P_1 y P_2 las proporciones muestrales obtenidas en grandes muestras de tamaños N_1 y N_2 tomadas de respectivas poblaciones que tienen proporciones p_1 y p_2. Consideremos la hipótesis nula de que *no hay diferencia* entre los parámetros de las poblaciones (o sea, $p_1 = p_2$) y por tanto que las muestras se han tomado de una misma población.

Poniendo $p_1 = p_2 = p$ en la ecuación (6) del Capítulo 8, vemos que la distribución de muestreo de diferencias en proporciones está casi normalmente distribuida, con media y desviación típica dadas por

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{3}$$

donde

$$p = \frac{N_1P_1 + N_2P_2}{N_1 + N_2}$$

se usa como estimación para la proporción poblacional y donde $q = 1 - p$.

Mediante la variable tipificada

$$z = \frac{P_1 - P_2 - 0}{\sigma_{P1-P2}} = \frac{P_1 - P_2}{\sigma_{P1-P2}} \tag{4}$$

podemos contrastar diferencias observadas a un nivel de significación apropiado y, en consecuencia, contrastar la hipótesis nula.

Contrastes que involucran a otros estadísticos se diseñan de manera similar.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

También cabe diseñar contrastes mediante distribuciones binomiales (u otras distribuciones) de forma parecida a como se ha hecho con la distribución normal; los principios básicos son esencialmente los mismos. Véanse Problemas 10.23 a 10.28.

PROBLEMAS RESUELTOS

CONTRASTES DE MEDIAS Y PROPORCIONES USANDO DISTRIBUCIONES NORMALES

10.1. Hallar la probabilidad de sacar entre 40 y 60 caras inclusive en 100 tiradas de una moneda buena.

Solución

De acuerdo con la distribución binomial, la probabilidad pedida es

$$\binom{100}{40}\left(\frac{1}{2}\right)^{40}\left(\frac{1}{2}\right)^{60} + \binom{100}{41}\left(\frac{1}{2}\right)^{41}\left(\frac{1}{2}\right)^{59} + \cdots + \binom{100}{60}\left(\frac{1}{2}\right)^{60}\left(\frac{1}{2}\right)^{40}$$

Como $Np = 100(\frac{1}{2})$ y $Nq = 100(\frac{1}{2})$ son ambos mayores que 5, la aproximación normal a la distribución binomial es correcta a la hora de evaluar esa suma. La media y la desviación típica del número de caras en 100 tiradas son

$$\mu = Np = 100(\tfrac{1}{2}) = 50 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(100)(\tfrac{1}{2})(\tfrac{1}{2})} = 5$$

En una escala continua, decir entre 40 y 60 inclusive es como decir entre 39.5 y 60.5 caras. Luego

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 50}{5} = -2.10 \qquad 60.5 \text{ en unidades estándar} = \frac{60.5 - 50}{5} = 2.10$$

$$\text{Probabilidad pedida} = \text{área bajo la curva normal entre } z = -2.10 \text{ y } z = 2.10$$
$$= 2(\text{área entre } z = 0 \text{ y } z = 2.10) = 2(0.4821) = 0.9642$$

10.2. Para contrastar la hipótesis de que una moneda es buena, adoptemos la siguiente regla de decisión:

Aceptarla si el número de caras en una sola muestra de 100 tiradas está entre 40 y 60 inclusive.

Rechazarla en caso contrario.

(*a*) Hallar la probabilidad de rechazar la hipótesis cuando en verdad sea correcta.
(*b*) Representar gráficamente la regla de decisión y el resultado de la parte (*a*).
(*c*) ¿Qué conclusiones se desprenden si resultan 53 caras en la muestra de 100 tiradas? ¿Y si salieran 60 caras?
(*d*) ¿Podría ser equivocada su conclusión sobre (*c*)? Explicar la respuesta.

Solución

(*a*) Del Problema 10.1, la probabilidad de no obtener entre 40 y 60 caras inclusive si la moneda es buena, es $1 - 0.9642 = 0.0358$. Luego la probabilidad de rechazar la hipótesis cuando sea correcta es 0.0358.

(*b*) La regla de decisión se ilustra en la Figura 10.2, que muestra las distribuciones de probabilidad de caras en 100 tiradas de una moneda buena. Si una sola muestra de 100 tiradas arroja un z entre -2.10 y 2.10, aceptamos la hipótesis; en caso contrario, la rechazamos y decidimos que la moneda está trucada.

El error de rechazar la hipótesis siendo correcta es el *error de Tipo I* de la regla de decisión; y su probabilidad, 0.0358 según (*a*), está representada por el área sombreada total en la figura. Si una sola muestra de 100 tiradas da un número de caras cuyo z está en las zonas sombreadas, diremos que ese valor de z difiere *de forma significativa* del esperado si la hipótesis fuese verdadera. Por tal razón, el área total sombreada (o sea, la probabilidad de un error de Tipo I) se llama el *nivel de significación* de la regla de decisión y vale 0.0358 en este caso. Así que podemos hablar de que rechazamos la hipótesis al nivel de significación 0.0358 (o sea al 3.58%).

(*c*) De acuerdo con la regla de decisión, tendremos que aceptar la hipótesis de que la moneda es buena en ambos casos. Cabe argumentar que con sólo una cara más ya la hubiésemos rechazado. ¡Siempre tiene uno que enfrentarse a una línea brusca de división al tomar decisiones!

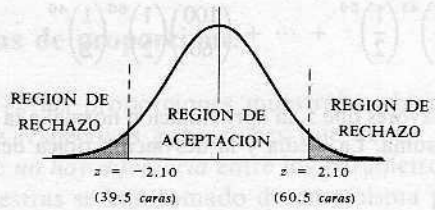


Figura 10.2.

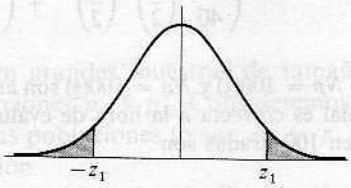


Figura 10.3.

(*d*) Sí. Podríamos aceptar la hipótesis cuando en realidad es rechazable, como sería el caso por ejemplo si la probabilidad de caras es 0.7 en vez de 0.5. El error cometido al aceptar la hipótesis que debiera ser rechazada es el *error de Tipo II* de la decisión. (Para más detalles, véanse Problemas 10.10 a 10.12).

10.3. Diseñar una regla de decisión para contrastar la hipótesis de que una moneda es buena y usar nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

(*a*) *Primer método*

Si el nivel de significación es 0.05, cada área sombreada en la Figura 10.3 es 0.025 por simetría. Entonces el área entre 0 y z_1 es $0.5000 - 0.0250 = 0.4750$, y $z_1 = 1.96$; los valores críticos -1.96 y 1.96 pueden leerse también en la Tabla 10.1. Así pues, una posible regla de decisión es:

Aceptar la hipótesis de que la moneda es buena si z está entre -1.96 y 1.96.

Rechazarla en caso contrario.

Para expresar la regla de decisión en términos del número de caras que se obtendrán en 64 tiradas de la moneda, nótese que la media y la desviación típica de la distribución de caras vienen dadas por:

$$\mu = Np = 64(0.5) = 32 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{64(0.5)(0.5)} = 4$$

bajo la hipótesis de que la moneda es buena. Entonces $z = (X - \mu)/\sigma = (X - 32)/4$. Si $z = 1.96$, entonces $(X - 32)/4 = 1.96$ y $X = 39.84$; si $z = -1.96$, entonces $(X - 32)/4 = -1.96$ y $X = 24.16$. Luego la regla de decisión se convierte en:

Aceptar la hipótesis de que la moneda es buena si el número de caras está entre 24.16 y 39.84 (o sea, entre 25 y 39 inclusive).

Rechazarla en caso contrario.

Segundo método

Con probabilidad 0.95, el número de caras estará entre $\mu - 1.96\sigma$ y $\mu + 1.96\sigma$ (o sea, entre $Np - 1.96\sqrt{Npq}$ y $Np + 1.96\sqrt{Npq}$, es decir, entre $32 - 1.96(4) = 24.16$ y $32 + 1.96(4) = 39.84$, lo que conduce a la regla de decisión precedente.

Tercer método

Como $-1.96 < z < 1.96$ es equivalente a $-1.96 < \frac{1}{4}(X - 32) < 1.96$, entonces $-1.96(4) < (X - 32) < 1.96(4)$, o sea $32 - 1.96(4) < X < 32 + 1.96(4)$ (o sea, $24.16 < X < 39.84$), que también conduce a la anterior regla de decisión.

(*b*) Si el nivel de significación es 0.01, cada área sombreada en la Figura 10.3 es 0.005. Luego el área entre 0 y z_1 es $0.5000 - 0.0050 = 0.4950$ y $z_1 = 2.58$ (más exactamente 2.575); esto puede leerse en la Tabla 10.1. Siguiendo el procedimiento del segundo método de la parte (*a*), vemos que con probabilidad 0.99 el número de caras estará entre $\mu - 2.58\sigma$ y $\mu + 2.58\sigma$, que son $32 - 2.58(4) = 21.68$ y $32 + 2.58(4) = 42.32$. Luego la regla de decisión es:

Aceptar la hipótesis si el número de caras está entre 22 y 42 inclusive.

Rechazarla en caso contrario.

10.4. ¿Cómo diseñaría una regla de decisión en el Problema 10.3 de modo que se evite el error de Tipo II?

Solución

Un error de Tipo II consiste en aceptar una hipótesis falsa, y se puede evitar como sigue: en vez de aceptar la hipótesis, simplemente no la rechazamos, lo que quiere decir que estamos rehusando tomar decisión en ese caso. Por ejemplo, podríamos enunciar la regla de decisión del Problema 10.3(*b*) así:

No rechazar la hipótesis si el número de caras está entre 22 y 42 inclusive.

Rechazarla en caso contrario.

En muchas situaciones prácticas, es importante decidir si una hipótesis dada debe ser aceptada o rechazada. Una discusión completa de tales casos requiere considerar los errores de Tipo II (véanse Probs. 10.10 a 10.12).

10.5. En un experimento sobre percepción extrasensorial (PES), un individuo en una habitación es invitado a adivinar el color (rojo o azul) de una carta elegida de un mazo de 50 cartas bien mezcladas por otro individuo en otra habitación. El no sabe cuántas rojas y cuántas azules hay en el mazo. Si el sujeto identifica 32 cartas correctamente, determinar si el resultado es significativo al nivel (*a*) 0.05 y (*b*) 0.01.

Solución

Si p es la probabilidad de que el sujeto acierte el color de una carta, hemos de decidir entre dos hipótesis:

H_0: $p = 0.5$, y el sujeto está simplemente diciendo colores al azar.

H_1: $p > 0.5$, y el sujeto tiene poderes de PES.

Como no estamos interesados en el caso de que obtenga muy pocos aciertos, sino en el de que

consiga muchos, escogemos un contraste de una cola. Si la hipótesis H_0 es verdadera, la media y la desviación típica del número de cartas acertadas vienen dadas por

$$\mu = Np = 50(0.5) = 25 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{50(0.5)(0.5)} = \sqrt{12.5} = 3.54$$

(*a*) Para un contraste unilateral al nivel de significación 0.05, debemos tomar z_1 en la Figura 10.4 de modo que el área en la región crítica sea 0.05. Entonces, el área entre 0 y z_1 es 0.4500 y $z_1 = 1.645$; lo que puede verse también en la Tabla 10.1. Luego nuestra regla de decisión (o contraste de significación) es:

Si el z observado es mayor que 1.645, el resultado es significativo al nivel 0.05 y el individuo tiene poderes PES.

En caso contrario, el resultado se debe al azar (no es significativo al nivel 0.05) y el sujeto no tiene PES.

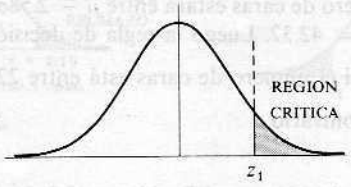


Figura 10.4.

Como 32 en unidades estándar es $(32 - 25)/3.54 = 1.98$, que es mayor que 1.645, concluimos que, al nivel 0.05, el individuo tiene poderes de PES.

Nótese que en realidad deberíamos aplicar una corrección de continuidad, porque 32 en escala continua está entre 31.5 y 32.5. Sin embargo, 31.5 tiene un valor estándar de $(31.5 - 25)/3.54 = 1.84$, y por tanto se alcanza idéntica conclusión.

(*b*) Si el nivel de significación es 0.01, el área entre 0 y z_1 es 0.4900, y $z_1 = 2.33$. Como 32 (o 31.5) en unidades estándar es 1.98 (o 1.84), que es menor que 2.33, concluimos que el resultado *no es significativo* al nivel 0.01.

Algunos estadísticos adoptan la terminología de que los resultados significativos al nivel 0.01 son *altamente significativos*, los que lo son al 0.05 pero no al 0.01 son *probablemente significativos*, y los que ni lo son al 0.05 se dicen *no significativos*. De modo que en el anterior experimento, el resultado es *probablemente significativo*, de manera que sería conveniente una investigación adicional.

Como los niveles de significación sirven de guía al tomar decisiones, algunos estadísticos citan las probabilidades implicadas. Así, como $\Pr\{z \geq 1.84\} = 0.0322$, en este problema, dirían que sobre la base del experimento, la probabilidad de equivocarnos al concluir que el sujeto tiene PES es de alrededor de un 3%. La probabilidad obtenida (0.0322 en este caso) se suele llamar *nivel de significación experimental o descriptivo*.

10.6. Un laboratorio de farmacia sostiene que uno de sus productos es 90% efectivo para reducir una alergia en 8 horas. En una muestra de 200 personas con esa alergia, el medicamento dio buen resultado en 160. Determinar si la afirmación del laboratorio es legítima.

Solución

Sea p la probabilidad de curación mediante ese fármaco. Hemos de decidir entre dos hipótesis:

H_0: $p = 0.9$, y la afirmación es correcta. H_1: $p < 0.9$, y la afirmación es falsa.

Como estamos interesados en determinar si la proporción de personas curadas es demasiado baja, elegimos un contraste de una cola. Si tomamos como nivel de significación el 0.01 (o sea, si el área sombreada en la Figura 10.5 es 0.01), entonces $z_1 = -2.33$, como se ve del Problema 10.5(*b*) por simetría de la curva o de la Tabla 10.1. Por tanto, adoptamos como regla de decisión:

No es legítima si z es menor que -2.33 (en cuyo caso rechazamos H_0).

En caso contrario, es legítima y los resultados observados se deben al azar (en cuyo caso aceptamos H_0).

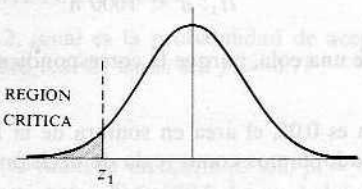


Figura 10.5.

Si H_0 es verdadera, entonces $\mu = Np = 200(0.9) = 180$ y $\sigma = \sqrt{Npq} = \sqrt{(200)(0.9)(0.1)} = 4.24$. Ahora bien, 160 en unidades estándar es $(160 - 180)/4.24 = -4.72$, que es mucho menor que -2.33. Luego, de acuerdo con nuestra regla de decisión, concluimos que la afirmación no es legítima y que los resultados del muestreo son altamente significativos (véase el final del Prob. 10.5).

10.7. La vida media de una muestra de 100 tubos fluorescentes producidos en una empresa es de 1570 h con una desviación típica de 120 h. Si μ es la vida media de todos los productos en esa empresa, contrastar la hipótesis de que $\mu = 1600$ h contra la hipótesis alternativa $\mu \neq 1600$ h, usando nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

Debemos decidir entre dos hipótesis:

$$H_0: \mu = 1600 \text{ h}$$
$$H_1: \mu \neq 1600 \text{ h}$$

Puesto que $\mu \neq 1600$ incluye valores mayores y menores que 1600, usaremos un contraste de dos colas.

(*a*) Para un contraste de dos colas al nivel de significación de 0.05, tenemos la siguiente regla de decisión:

Rechazar H_0 si el z de la media muestral está fuera del rango -1.96 a 1.96.

Aceptar H_0 en caso contrario.

El estadístico bajo consideración es la media muestral $\bar{X}$. La distribución de muestreo de $\bar{X}$ tiene media $\mu_{\bar{X}} = \mu$ y desviación típica $\sigma_{\bar{X}} = \sigma/\sqrt{N}$, donde μ y σ son la media y la desviación típica de toda la población de tubos producidos por la empresa. Bajo la hipótesis H_0, tenemos $\mu = 1600$ y $\sigma_{\bar{X}} = \sigma/\sqrt{N} = 120/\sqrt{100} = 12$, usando la desviación típica muestral como estimación de σ. Como $z = (\bar{X} - 1600)/12 = (1570 - 1600)/12 = -2.50$ está fuera del rango -1.96 a 1.96, rechazamos H_0 al nivel de significación 0.05.

(*b*) Si el nivel de significación es 0.01, el rango pasa a ser -2.58 a 2.58. Así pues, como el valor -2.50 de z cae dentro de ese rango, aceptamos H_0 (o rehusamos tomar decisión al nivel de significación 0.01.

10.8. En el Problema 10.7, contrastar la hipótesis $\mu = 1600$ h frente a la hipótesis alternativa $\mu < 1600$ h con nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

Tenemos que decidir entre las hipótesis:

$$H_0: \mu = 1600 \text{ h}$$
$$H_1: \mu < 1600 \text{ h}$$

Habrá que usar un contraste de una cola, porque la correspondiente figura es idéntica a la Figura 10.5 del Problema 10.6.

(*a*) Si el nivel de significación es 0.05, el área en sombra de la Figura 10.5 es 0.05, y hallamos que $z_1 = -1.645$. Por tanto, adoptamos como regla de decisión:

Rechazar H_0 si z es menor que -1.645.

Aceptarla en caso contrario (o declinar cualquier decisión).

Ya que [como en el Prob. 10.7(*a*)] z es -2.50, menor que -1.645, rechazamos H_0 al nivel 0.05. Nótese que esta decisión es idéntica a la alcanzada en el Problema 10.7(*a*) por medio de un contraste bilateral.

(*b*) Si el nivel de significación es 0.01, el valor z_1 en la Figura 10.5 es -2.33. Por consiguiente, adoptamos la regla de decisión siguiente:

Rechazar H_0 si z es menor que -2.33.

Aceptar H_0 en caso contrario (o declinar cualquier decisión).

Ya que [como en el Prob. 10.7(*a*)] z es -2.50, menor que -2.33, rechazamos H_0 al nivel 0.01. Nótese que esta decisión no es la alcanzada en el Problema 10.7(*b*) por medio de un contraste bilateral.

Se deduce que las decisiones relativas a una cierta hipótesis H_0 que están basadas en contrastes de una o dos colas no siempre concuerdan. Lo cual era de esperar, naturalmente, pues estamos contrastando H_0 frente a alternativas diferentes según el caso.

10.9. Las tensiones de ruptura de los cables fabricados por una empresa tienen media de 1800 lib y una desviación típica de 100 lb. Se desea comprobar si un nuevo proceso de fabricación aumenta dicha tensión media. Para ello se toma una muestra de 50 cables y se encuentra que su tensión media de ruptura es 1850 lb. ¿Se puede afirmar la mejoría del nuevo proceso al nivel de significación 0.01?

Solución

Tenemos que decidir entre dos hipótesis:

H_0: $\mu = 1800$ lb, y no hay realmente cambio en la tensión de ruptura.

H_1: $\mu > 1800$ lb, y hay realmente cambio en la tensión de ruptura.

Hay que usar un contraste de una cola; el diagrama asociado con él es idéntico a la Figura 10.4. Al nivel de significación 0.01, la regla de decisión es:

Si el z observado es mayor que 2.33, el resultado es significativo al nivel 0.01 y rechazamos H_0.

En caso contrario, se acepta H_0 (o se aplaza la decisión).

Bajo la hipótesis de que H_0 es verdadera, vemos que

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.55$$

que es mayor que 2.33. Así que el resultado es altamente significativo y la afirmación puede mantenerse.

CURVAS DE OPERACION CARACTERISTICAS

10.10. Refiriendo al Problema 10.2, ¿cuál es la probabilidad de aceptar la hipótesis de que la moneda es buena cuando la probabilidad real de caras sea $p = 0.7$?

Solución

La hipótesis H_0 de que la moneda es buena (o sea, $p = 0.5$), es aceptada cuando el número de caras en 100 lanzamientos está entre 39.5 y 60.5. La probabilidad de rechazar H_0 cuando debería ser aceptada (o sea, la probabilidad de un error de Tipo I) viene representada por el área total α de la región sombreada de la izquierda en la Figura 10.6. Como calculamos en el Problema 10.2(*a*), esa área, que representa el nivel de significación del contraste de H_0, es igual a 0.0358.

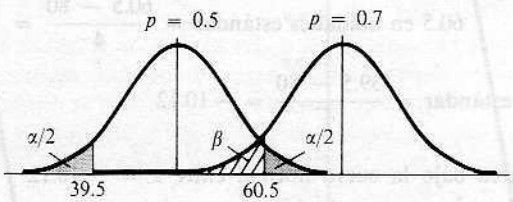


Figura 10.6.

Si $p = 0.7$, la distribución de caras en 100 lanzamientos está representada por la curva normal a la derecha en la Figura 10.6. Del diagrama es claro que la probabilidad de aceptar H_0 cuando en verdad $p = 0.7$ (es decir, la probabilidad de un error de Tipo II) viene dada por el área rayada β de la figura. Para calcularla, observamos que la distribución bajo la hipótesis $p = 0.7$ tiene media y desviación típica dadas por

$$\mu = Np = (100)(0.7) = 70 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.7)(0.3)} = 4.58$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 70}{4.58} = -2.07$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 70}{4.58} = -6.66$$

Entonces $\beta = (\text{área bajo la curva normal entre } z = -6.66 \text{ y } z = -2.07) = 0.0192$

Luego hay poca opción, con la regla de decisión adoptada, de aceptar la hipótesis de que la moneda es buena si tiene en verdad $p = 0.7$.

Nótese que en este problema se nos da la regla de decisión, de la que calculamos α y β. En la práctica, aparecen otras dos posibilidades:

(1) Acordamos un α (tal como 0.05 o 0.01), llegamos a una decisión y entonces calculamos β.
(2) Acordamos α y β, y entonces llegamos a una regla de decisión.

10.11. Resolver el Problema 10.10 si (*a*) $p = 0.6$, (*b*) $p = 0.8$, (*c*) $p = 0.9$ y (*d*) $p = 0.4$.

Solución

(*a*) Si $p = 0.6$, la distribución de caras tiene su media y su desviación típica dadas por

$$\mu = Np = (100)(0.6) = 60 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.6)(0.4)} = 4.90$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 60}{4.90} = 0.102$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 60}{4.90} = -4.18$$

Entonces β = (área bajo la curva normal entre $z = -4.18$ y $z = 0.102$) = 0.5406

Así que con la regla de decisión dada existen muchas posibilidades de aceptar la hipótesis de que la moneda es buena aunque en realidad tiene $p = 0.6$.

(*b*) Si $p = 0.8$, entonces

$$\mu = Np = (100)(0.8) = 80 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.8)(0.2)} = 4$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 80}{4} = -4.88$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 80}{4} = -10.12$$

Entonces β = (área bajo la curva normal entre $z = -10.12$ y $z = -4.88$) = 0.0000 muy aproximadamente.

(*c*) Comparando con la parte (*b*) o por cálculo, vemos que si $p = 0.9$, entonces $\beta = 0$ a efectos prácticos.

(*d*) Por simetría, $p = 0.4$ da el mismo valor de β que $p = 0.6$ (es decir, $\beta = 0.5040$).

10.12. Representar los resultados de los Problemas 10.10 y 10.11 construyendo un gráfico de (*a*) β versus p y (*b*) $(1 - \beta)$ versus p. Interpretar los gráficos obtenidos.

Solución

La Tabla 10.2 muestra los valores de β correspondientes a valores dados de p, tal como se obtienen en el Problema 10.10 y en el 10.11. Aquí β representa la probabilidad de aceptar la hipótesis $p = 0.5$ cuando p es algún otro valor; si en verdad es $p = 0.5$, podemos interpretar β como la probabilidad de aceptar $p = 0.5$ cuando de hecho debía ser aceptada. Esta propiedad es $1 - 0.0358 = 0.9642$ y se ha incluido en la Tabla 10.2.

Tabla 10.2

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.0000	0.0000	0.0192	0.5040	0.9642	0.5040	0.0192	0.0000	0.0000

(*a*) El gráfico de β versus p, que se ve en la Figura 10.7(*a*), se llama la *curva de operación característica*, o *curva OC*, de la regla de decisión (o contraste de hipótesis). La distancia de su máximo a la recta $\beta = 1$ es igual a $\alpha = 0.0358$, el nivel de significación del test.

En general, cuanto más agudo el pico de la curva OC, mejor es la regla de decisión a la hora de rechazar hipótesis incorrectas.

(*b*) El gráfico de $(1 - \beta)$ versus p, Figura 10.7(*b*), se llama la *curva de potencia* de la regla de decisión. Se obtiene sin más que invertir la curva OC; luego ambos gráficos son equivalentes.

La cantidad $(1 - \beta)$ se suele llamar una *función de potencia*, porque indica la *potencia de un test* (o *contraste*) para rechazar hipótesis falsas, rechazables en consecuencia. La cantidad β se llama *función de operación característica* de un test.

10.13. Una compañía produce sogas cuya tensión de ruptura tiene media de 300 lb y desviación típica de 24 lb. Se espera que un nuevo proceso de fabricación haga crecer la media.

(*a*) Diseñar una regla de decisión para rechazar el proceso antiguo al nivel de significación 0.01 con una muestra de 64 sogas.

(*b*) Con esa regla de decisión, ¿cuál es la probabilidad de aceptar el antiguo procedimiento cuando de hecho el nuevo ha aumentado la tensión media de las sogas a 310 lb? Suponemos que la desviación típica es todavía de 24 lb.

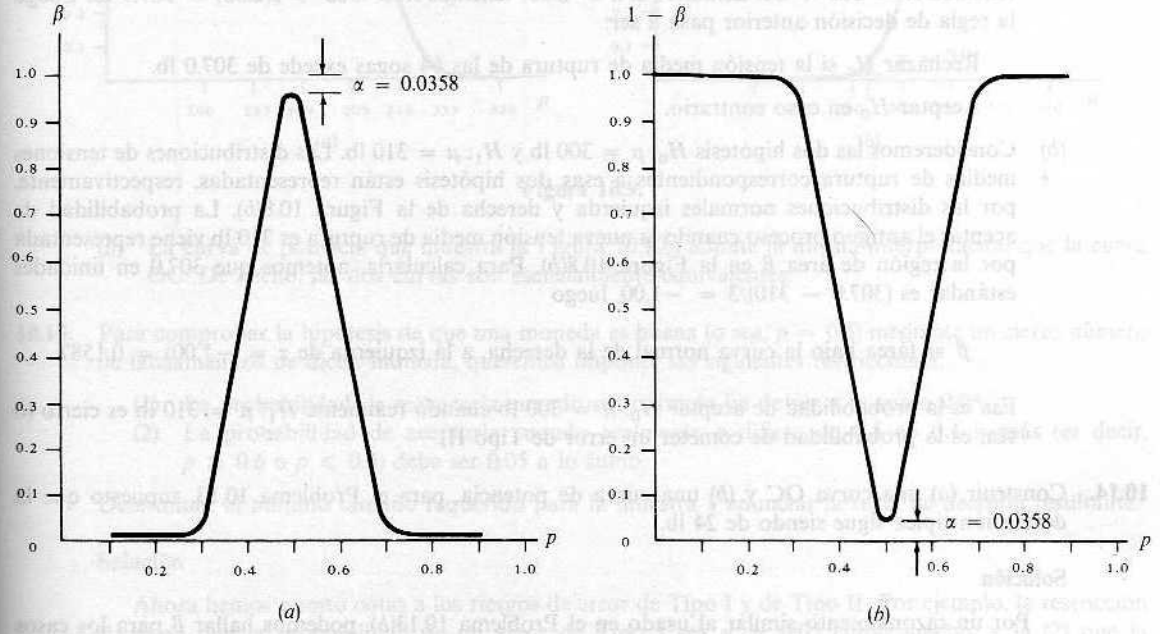


Figura 10.7.

Solución

(*a*) Si μ es la tensión media de ruptura, queremos decidir entre dos hipótesis:

H_0: $\mu = 300$ lb, y el nuevo proceso es como el antiguo.

H_1: $\mu > 300$ lb, y el nuevo proceso es mejor que el antiguo.

Para un contraste de una cola al nivel de significación 0.01, tenemos la siguiente regla de decisión [véase Fig. 10.8(*a*)]:

Rechazar H_0 si el valor z para la tensión media de ruptura es mayor que 2.33.

Aceptar H_0 en caso contrario.

Como $$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{\bar{X} - 300}{24/\sqrt{64}}$$

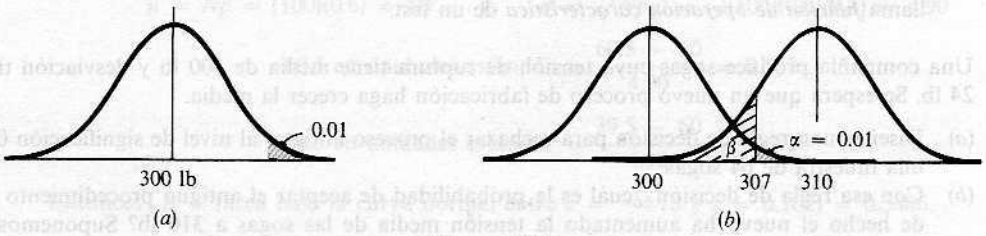


Figura 10.8.

tenemos $\bar{X} = 300 + 3z$. Entonces si $z > 2.33$, tenemos $\bar{X} > 300 + 3(2.33) = 307.7$ lb. Luego la regla de decisión anterior pasa a ser:

Rechazar H_0 si la tensión media de ruptura de las 64 sogas excede de 307.0 lb.

Aceptar H_0 en caso contrario.

(*b*) Consideremos las dos hipótesis H_0: $\mu = 300$ lb y H_1: $\mu = 310$ lb. Las distribuciones de tensiones medias de ruptura correspondientes a esas dos hipótesis están representadas, respectivamente, por las distribuciones normales izquierda y derecha de la Figura 10.8(*b*). La probabilidad de aceptar el antiguo proceso cuando la nueva tensión media de ruptura es 310 lb viene representada por la región de área β en la Figura 10.8(*b*). Para calcularla, notemos que 307.0 en unidades estándar es $(307.0 - 310)/3 = -1.00$, luego

$$\beta = (\text{área bajo la curva normal de la derecha, a la izquierda de } z = -1.00) = 0.1587$$

Esa es la probabilidad de aceptar H_0: $\mu = 300$ lb cuando realmente H_1: $\mu = 310$ lb es cierto (o sea, es la probabilidad de cometer un error de Tipo II).

10.14. Construir (*a*) una curva OC y (*b*) una curva de potencia, para el Problema 10.13, supuesto que la desviación típica sigue siendo de 24 lb.

Solución

Por un razonamiento similar al usado en el Problema 10.13(*b*), podemos hallar β para los casos en que el nuevo proceso de tensiones medias de ruptura μ iguales a 305 lb, 315 lb, etc. Por ejemplo, si $\mu = 305$ lb, entonces 307.0 lb en unidades estándar es $(307.0 - 305)/3 = 0.67$, y por tanto

$$\beta = (\text{área bajo la curva normal de la derecha, a la izquierda de } z = 0.67) = 0.7486$$

De esta forma se obtiene la Tabla 10.3.

Tabla 10.3

μ	290	295	300	305	310	315	320
β	1.0000	1.0000	0.9900	0.7486	0.1587	0.0038	0.0000

(*a*) La curva OC se ve en la Figura 10.9(*a*). En ella apreciamos que la probabilidad de conservar el antiguo proceso si la nueva tensión media de ruptura es menor que 300 lb es casi (excepto para el nivel de significación 0.01 cuando el nuevo proceso da una media de 300 lb). Cae rápidamente a cero, lo cual quiere decir que no hay prácticamente opción de mantener el antiguo proceso cuando la tensión media de ruptura es mayor que 315 lb.

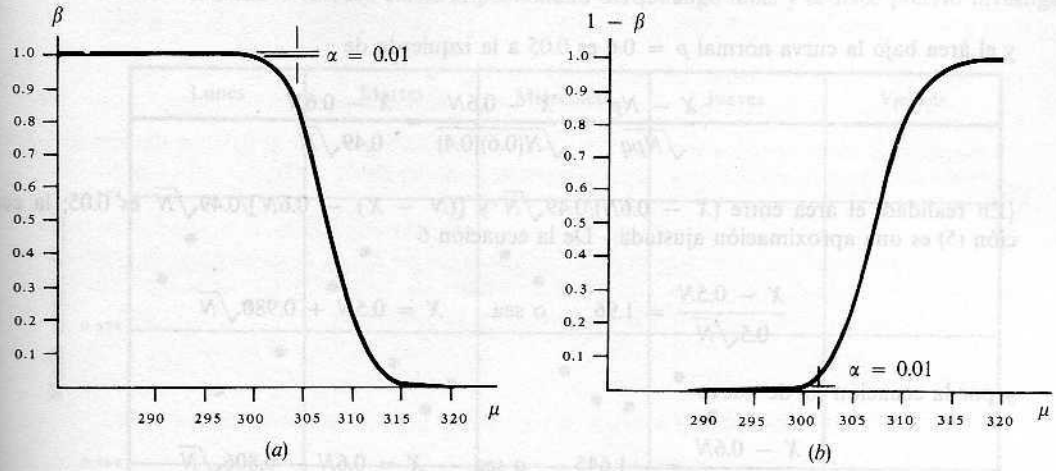


Figura 10.9.

(*b*) La curva de potencia que muestra la Figura 10.9(*b*) admite la misma interpretación que la curva OC. De hecho, las dos curvas son esencialmente equivalentes.

10.15. Para comprobar la hipótesis de que una moneda es buena (o sea, $p = 0.5$) mediante un cierto número de lanzamientos de dicha moneda, queremos imponer las siguientes restricciones:

(1) La probabilidad de rechazarla cuando sea correcta ha de ser a lo sumo 0.05.
(2) La probabilidad de aceptarla cuando realmente p difiera de 0.5 en 0.1 o más (es decir, $p \geqslant 0.6$ o $p \leqslant 0.4$) debe ser 0.05 a lo sumo.

Determinar el mínimo tamaño requerido para la muestra y enunciar la regla de decisión resultante.

Solución

Ahora hemos puesto cotas a los riesgos de error de Tipo I y de Tipo II. Por ejemplo, la restricción (1) exige que la probabilidad de un error de Tipo I sea $\alpha = 0.05$ como mucho, y la (2) que la probabilidad de un error de Tipo II sea $\beta = 0.05$ a lo más. La situación se refleja en la Figura 10.10.

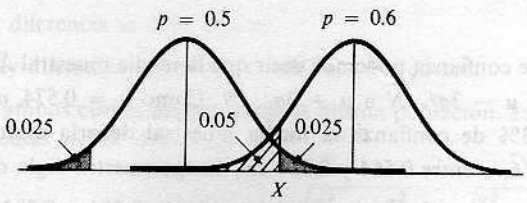


Figura 10.10.

Sea N el tamaño requerido para la muestra y X el número de caras en N tiradas, por encima del cual rechazamos la hipótesis de que $p = 0.5$. De la Figura 10.10, el área bajo la curva normal $p = 0.5$ es 0.025 a la derecha de

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.5N}{\sqrt{N(0.5)(0.5)}} = \frac{X - 0.5N}{0.5\sqrt{N}} \tag{5}$$

y el área bajo la curva normal $p = 0.6$ es 0.05 a la izquierda de

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.6N}{\sqrt{N(0.6)(0.4)}} = \frac{X - 0.6N}{0.49\sqrt{N}} \tag{6}$$

{En realidad, el área entre $(X - 0.6N)/0.49\sqrt{N}$ y $[(N - X) - 0.6N]/0.49\sqrt{N}$ es 0.05; la ecuación (5) es una aproximación ajustada.} De la ecuación 6

$$\frac{X - 0.5N}{0.5\sqrt{N}} = 1.96 \quad \text{o sea} \quad X = 0.5N + 0.980\sqrt{N} \tag{7}$$

y por la ecuación (6) de nuevo

$$\frac{X - 0.6N}{0.49\sqrt{N}} = -1.645 \quad \text{o sea} \quad X = 0.6N - 0.806\sqrt{N} \tag{8}$$

Y de (7) y (8) deducimos $N = 318.98$, luego la muestra ha de ser de 319 al menos (o sea, hay que lanzar al menos 319 veces la moneda). Poniendo $N = 319$ en la ecuación (7) u (8), $X = 177$.

Para $p = 0.5$ se tiene por tanto $X - Np = 177 - 159.5 = 17.5$. En consecuencia, adoptamos la siguiente regla de decisión:

Aceptar la hipótesis de que $p = 0.5$ si el número de caras en 319 lanzamientos está en el rango 159.5 ± 17.5 (o sea, entre 142 y 177).

Rechazarla en caso contrario.

GRAFICOS DE CONTROL

10.16. Se construye una máquina para fabricar bolas de rodamiento con diámetro medio de 0.574 cm y desviación típica de 0.008 cm. Para determinar si funciona correctamente, se toma una muestra de 6 bolas cada 2 horas y se halla para cada una de las muestras el diámetro medio.

(*a*) Diseñar una regla de decisión con la que se esté muy seguro de que la calidad del producto cumple los propósitos exigidos.

(*b*) Ilustrar gráficamente la regla de decisión de (*a*).

Solución

(*a*) Con el 99.73% de confianza podemos decir que la media muestral $\bar{X}$ debe estar entre $\mu_{\bar{X}} - 3\sigma_{\bar{X}}$ y $\mu_{\bar{X}} + 3\sigma_{\bar{X}}$, o sea $\mu - 3\sigma/\sqrt{N}$ a $\mu + 3\sigma/\sqrt{N}$. Como $\mu = 0.574$, $\sigma = 0.008$ y $N = 6$, se sigue que con el 99.73% de confianza la media muestral debería estar entre $0.574 - 0.024/\sqrt{6}$ y $0.574 + 0.024/\sqrt{6}$, o entre 0.564 y 0.584 cm. Luego nuestra regla de decisión es como sigue:

Si una media muestral cae dentro del rango de 0.564 a 0.584, aceptamos que la máquina funciona bien.

Si no, concluimos que no funciona bien e investigamos la razón.

(*b*) Se pueden anotar las observaciones en un gráfico como el de la Figura 10.11, llamado un *gráfico de control de calidad*. Cada vez que se toma una muestra, se representa por un punto concreto. En tanto que los puntos están entre el límite inferior (0.564 cm) y el superior (0.584 cm), el proceso está bajo control. Cuando un punto se sale de esos límites de control (como sucede con la tercera muestra tomada el jueves), existe la posibilidad de que algo falle, y se hace preciso investigarlo.

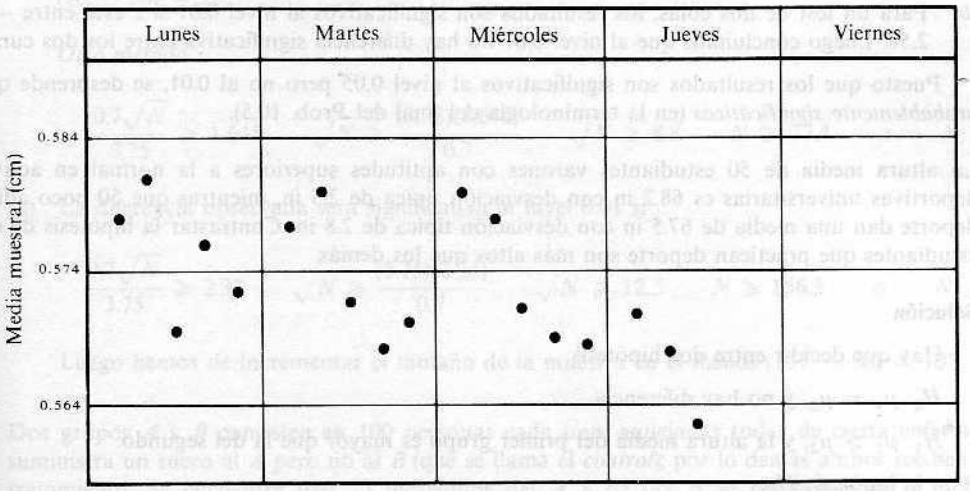


Figura 10.11.

Los límites de control antes especificados se llaman los límites de confianza 9.73%, o más brevemente, los límites 3σ. Otros límites de confianza (tales como 99% o 95%) se determinan del mismo modo. La elección en cada caso depende de las circunstancias particulares.

CONTRASTES MEDIANTE DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.17. En un mismo examen realizado en dos cursos, la nota media del primero fue 74 con desviación típica 8, y en el otro fue 78 con desviación típica 7. ¿Hay diferencia significativa entre las calificaciones de ambos cursos al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Supongamos que los dos cursos provienen de dos poblaciones con medias respectivas μ_1 y μ_2. Hemos de decidir entre las dos hipótesis:

H_0: $\mu_1 = \mu_2$, y la diferencia se debe al azar.

H_1: $\mu_1 \neq \mu_2$, y hay diferencia significativa entre los dos cursos.

Bajo la hipótesis H_0. ambos cursos provienen de la misma población. La media y la desviación típica de la diferencia en medias vienen dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \text{y} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{8^2}{40} + \frac{7^2}{50}} = 1.606$$

donde hemos usado las desviaciones típicas muestrales como estimaciones de σ_1 y σ_2. Así pues

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{74 - 78}{1.606} = -2.49$$

(*a*) Para un test de dos colas, los resultados son significativos al nivel 0.05 si z está entre -1.96 y 1.96. Luego concluimos que al nivel 0.05 hay diferencia significativa y probablemente es mejor el segundo de los cursos.

(*b*) Para un test de dos colas, los resultados son significativos al nivel 0.01 si z está entre -2.58 y 2.58. Luego concluimos que al nivel 0.01 no hay diferencia significativa entre los dos cursos.

Puesto que los resultados son significativos al nivel 0.05 pero no al 0.01, se desprende que son *probablemente significativos* (en la terminología del final del Prob. 10.5).

10.18. La altura media de 50 estudiantes varones con aptitudes superiores a la normal en actividades deportivas universitarias es 68.2 in con desviación típica de 2.5 in, mientras que 50 poco adictos al deporte dan una media de 67.5 in con desviación típica de 2.8 in. Contrastar la hipótesis de que los estudiantes que practican deporte son más altos que los demás.

Solución

Hay que decidir entre dos hipótesis:

H_0: $\mu_1 = \mu_2$, y no hay diferencia.

H_1: $\mu_1 > \mu_2$, y la altura media del primer grupo es mayor que la del segundo.

Bajo la hipótesis H_0,

$$\mu_{\bar{X}1-\bar{X}2} = 0 \quad \text{y} \quad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}} = 0.53$$

donde hemos usado las desviaciones típicas muestrales como estimaciones de σ_1 y σ_2. Luego

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{68.2 - 67.5}{0.53} = 1.32$$

Con un contraste de una cola al nivel de significación 0.05, rechazaríamos H_0 si z fuera mayor que 1.645. Así que no podemos rechazarla a este nivel de significación.

Hay que hacer notar, no obstante, que la hipótesis puede ser rechazada al nivel 0.01 si estamos dispuestos a correr el riesgo de equivocarnos con una probabilidad de 0.10 (un 10%).

10.19. ¿Cuánto hay que aumentar el tamaño de la muestra en cada uno de los grupos del Problema 10.18 al objeto de que la diferencia observada de 0.7 in en las alturas medias sea significativa al nivel (*a*) 0.05 y (*b*) 0.01?

Solución

Sea N el tamaño de la muestra en cada grupo y supongamos que la desviación típica de los grupos sigue siendo la misma. Entonces, bajo la hipótesis H_0 tenemos

$$\mu_{\bar{X}1-\bar{X}2} = 0 \quad \text{y} \quad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}} = \sqrt{\frac{(2.5)^2 + (2.8)^2}{N}} = \sqrt{\frac{14.09}{N}} = \frac{3.75}{\sqrt{N}}$$

Para una diferencia observada en alturas medias de 0.7 in, tenemos pues

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{0.7}{3.75/\sqrt{N}} = \frac{0.7\sqrt{N}}{3.75}$$

(*a*) La diferencia observada será significativa al nivel 0.05 si $0.7\sqrt{N}/3.75 = 1.645$ al menos, de modo que N ha de ser al menos 78. Por tanto debemos aumentar el tamaño de la muestra en al menos $(78 - 50) = 28$.

Otro método

$$\frac{0.7\sqrt{N}}{3.75} \geq 1.645 \qquad \sqrt{N} \geq \frac{(3.75)(1.645)}{0.7} \qquad \sqrt{N} \geq 8.8 \qquad N \geq 77.4 \qquad \text{o} \qquad N \geq 78$$

(*b*) La diferencia observada será significativa al nivel 0.01 si

$$\frac{0.7\sqrt{N}}{3.75} \geq 2.33 \qquad \sqrt{N} \geq \frac{(3.75)(2.33)}{0.7} \qquad \sqrt{N} \geq 12.5 \qquad N \geq 156.3 \qquad \text{o} \qquad N \geq 157$$

Luego hemos de incrementar el tamaño de la muestra en el menos $(157 - 50) = 107$.

10.20. Dos grupos A y B consisten en 100 personas cada uno, aquejadas todas de cierta enfermedad. Se suministra un suero al A pero no al B (que se llama el *control*); por lo demás ambos reciben idéntico tratamiento. Se encuentra que 75 individuos del A y 65 del B se recuperan de la enfermedad. Contrastar la hipótesis de que el suero cura la enfermedad al nivel de significación (*a*) 0.01, (*b*) 0.05 y (*c*) 0.10.

Solución

Sean p_1 y p_2 las proporciones de población curadas (1) con, y (2) sin ese suero. Hemos de decidir entre dos hipótesis:

H_0: $p_1 = p_2$, y la diferencia observada se debe al azar (el suero es ineficaz).

H_1: $p_1 > p_2$, y el suero es eficaz.

Bajo la hipótesis H_0,

$$\mu_{P1-P2} = 0 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{100} + \frac{1}{100}\right)} = 0.0648$$

donde hemos usado como estimación de p la proporción media de curaciones en los dos grupos muestra, dadas por $(75 + 65)/200 = 0.70$, donde $q = 1 - p = 0.30$. Por tanto

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0648} = 1.54$$

(*a*) Con contraste de una cola al nivel de significación 0.01, debemos rechazar H_0 sólo si el valor z es mayor que 2.33. Como z es 1.54, concluimos que los resultados se deben al azar, a este nivel de significación.

(*b*) Con contraste de una cola al nivel de significación 0.05, debemos rechazar H_0 sólo si el valor z

es mayor que 1.645. Por tanto, concluimos que los resultados se deben al azar a este nivel de significación también.

(*c*) Con contraste de una cola al nivel de significación 0.10, debemos rechazar H_0 sólo si el valor z es mayor que 1.28. Como z es 1.54, concluimos que el suero es eficaz a este nivel de significación.

Nótese que estas conclusiones dependen de cuánto estamos dispuestos a arriesgar en equivocarnos. Si los resultados fuesen realmente debidos al azar, pero concluyésemos que el suero es eficaz (error de Tipo I), podríamos proceder a suministrarlo a grupos más grandes de enfermos, y nos convenceríamos finalmente de su ineficacia. Es un riesgo que no siempre se está dispuesto a correr.

Por otro lado, podríamos concluir que el suero no es efectivo, cuando en verdad lo fuese (error de Tipo II). Tal conclusión es muy peligrosa, especialmente si hay vidas en juego.

10.21. Resolver el Problema 10.20 si cada grupo consta de 300 enfermos y se curan 225 del *A* y 195 del *B*.

Solución

En este caso las proporciones de curación son 225/300 = 0.750 y 195/300 = 0.650, iguales que en el Problema 10.20. Bajo la hipótesis H_0,

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{300} + \frac{1}{300}\right)} = 0.0374$$

donde (225 + 195)/600 = 0.70 se usa como estimación de p. Luego

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0374} = 2.67$$

Como este valor de z es mayor que 2.33, podemos rechazar la hipótesis al nivel de significación 0.01; es decir, concluimos que el suero es efectivo con sólo un 1% de probabilidad de equivocarnos.

Esto enseña la importancia del tamaño de la muestra en la fiabilidad de las decisiones. En muchos casos, sin embargo, puede no ser factible aumentar el tamaño. En tal circunstancia, estamos obligados a tomar decisiones sobre la base de la información disponible y arrostrar, por tanto, mayores riesgos de equivocación.

10.22. Un sondeo de 300 votantes del distrito *A* y 200 del *B* dan 56% y 48% respectivamente de votos en favor de un cierto candidato. Al nivel de significación 0.05, contrastar la hipótesis de que (*a*) hay diferencia entre los distritos y (*b*) ese candidato es el preferido en el distrito *A*.

Solución

Sean p_1 y p_2 las proporciones de todos los votantes en los distritos *A* y *B*, respectivamente, que son favorables a ese candidato. Bajo la hipótesis H_0: $p_1 = p_2$, tenemos

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.528)(0.472)\left(\frac{1}{300} + \frac{1}{200}\right)} = 0.0456$$

donde hemos usado como estimaciones para p y q los valores[(0.56)(300) + (0.48)(200)]/500 = 0.528 y (1 − 0.528) = 0.472, respectivamente. Luego

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.560 - 0.480}{0.0456} = 1.75$$

(*a*) Si sólo deseamos averiguar si hay diferencia entre los dos distritos, hemos de decidir entre las hipótesis H_0: $p_1 = p_2$ y H_1: $p_1 \neq p_2$, que implican un test de dos colas. Con él, rechazaríamos H_0 al nivel de significación 0.05 si z cae fuera del intervalo -1.96 a 1.96. Como $z = 1.75$ cae dentro de ese intervalo, no podemos rechazar H_0 a este nivel; esto es, no hay diferencia significativa entre los distritos.

(*b*) Si queremos determinar si el candidato es preferido en el distrito A, debemos decidir entre H_0: $p_1 = p_2$ y H_1: $p_1 > p_2$, lo cual implica un contraste de una cola. Usándolo al nivel de significación 0.05, rechazaremos H_0 si z es mayor que 1.645. Ya que tal es el caso, podemos rechazar H_0 a este nivel y concluir que el candidato es preferido en el distrito A.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

10.23. Un profesor propone a sus alumnos 10 cuestiones verdadero-falso. Para comprobar la hipótesis de que los estudiantes contestan al azar, adopta la siguiente regla de decisión:

Si al menos 7 respuestas son acertadas, el estudiante no ha contestado al azar.
Si hay menos de 7 correctas, ha contestado al azar.

Hallar la probabilidad de rechazar la hipótesis cuando sea correcta.

Solución

Sea p la probabilidad de que una cuestión sea acertada correctamente. La probabilidad de lograr X correctas de las 10 es $\binom{10}{X}p^X q^{10-X}$, con $q = 1 - p$. Bajo la hipótesis $p = 0.5$ (o sea, el estudiante responde al azar),

$$\Pr\{7 \text{ o más correctas}\} = \Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}$$

$$= \binom{10}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3 + \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + \binom{10}{10}\left(\frac{1}{2}\right)^{10} = 0.1719$$

Así que la probabilidad de concluir que no contestaban al azar cuando realmente sí lo hacían, es 0.1719. Nótese que esta es la probabilidad de un error de Tipo I.

10.24. En el Problema 10.23, hallar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad $p = 0.7$

Solución

Bajo la hipótesis $p = 0.7$.

$$\Pr\{\text{menos de 7 correctas}\} = 1 - \Pr\{7 \text{ o más correctas}\} =$$

$$= 1 - \left[\binom{10}{7}(0.7)^7(0.3)^3 + \binom{10}{8}(0.7)^8(0.3)^2 + \binom{10}{9}(0.7)^9(0.3) + \binom{10}{10}(0.3)^{10}\right] =$$

$$= 0.3504$$

10.25. En el Problema 10.23, hallar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando (*a*) $p = 0.6$, (*b*) $p = 0.8$, (*c*) $p = 0.9$, (*d*) $p = 0.4$, (*e*) $p = 0.3$, (*f*) $p = 0.2$ y (*g*) $p = 0.1$.

Solución

(*a*) Si $p = 0.6$,

$$\text{Probabilidad pedida} = 1 - [\Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}]$$

$$= 1 - \left[\binom{10}{7}(0.6)^7(0.4)^3 + \binom{10}{8}(0.6)^8(0.4)^2 + \binom{10}{9}(0.6)^9(0.4) + \binom{10}{10}(0.6)^{10}\right] = 0.618$$

Los resultados de las partes (*b*) hasta (*g*) se pueden obtener de manera análoga, y se recogen en la Tabla 10.4, junto con los valores correspondientes a $p = 0.5$ y $p = 0.7$. Nótese que la probabilidad en la Tabla 10.4 se denota por β (probabilidad de un error de Tipo II); la entrada β para $p = 0.5$ viene dada por $\beta = 1 - 0.1719 = 0.828$ (del Prob. 10.23), y para $p = 0.7$ del Problema 10.24.

Tabla 10.4

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013

10.26. Con ayuda del Problema 10.25, construir el gráfico de β versus p, obteniendo así las curvas de operación características de la regla de decisión del Problema 10.23.

Solución

El gráfico requerido es el de la Figura 10.12; obsérvese el parecido con la curva OC del Problema 10.14. Si hubiésemos representado $(1 - \beta)$ versus p, hubiéramos obtenido la *curva de potencia* de la regla de decisión. El gráfico indica que la regla de decisión es *potente* para rechazar $p = 0.5$ cuando realmente $p \leqslant 0.4$ o $p \geqslant 0.8$.

10.27. Una moneda da 6 caras en 6 tiradas. ¿Podemos concluir el nivel de significación (*a*) 0.05 y (*b*) 0.01 que está trucada? Considerar tanto contraste de una como de dos colas.

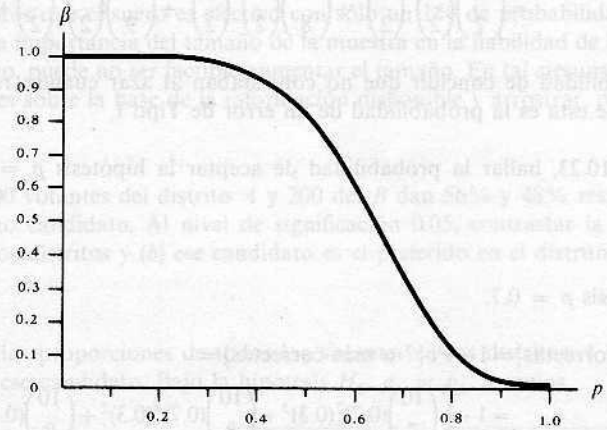


Figura 10.12.

Solución

Sea p la probabilidad de cara en una sola tirada de esa moneda. Bajo la hipótesis H_0: $p = 0.5$ (o sea, la moneda es buena),

$$p(X) = \Pr\{X \text{ caras en 6 tiradas}\} = \binom{6}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{6-X} = \binom{6}{X}\left(\frac{1}{64}\right)$$

Así pues, las probabilidades de 0, 1, 2, 3, 4, 5 y 6 caras vienen dadas, respectivamente, por $\frac{1}{64}$, $\frac{6}{64}$, $\frac{15}{64}$, $\frac{20}{64}$, $\frac{15}{64}$, $\frac{6}{64}$ y $\frac{1}{64}$, representadas en la distribución de probabilidad de la Figura 10.13.

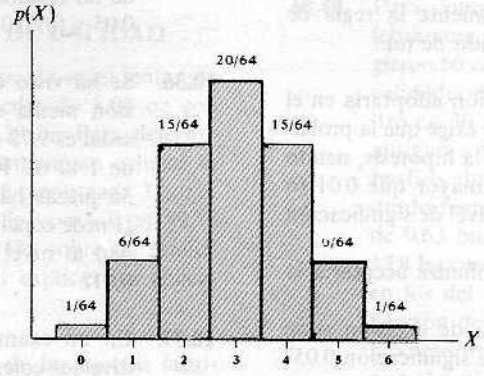


Figura 10.13.

Contraste de una cola

Aquí hay que decidir entre las hipótesis H_0: $p = 0.5$ y H_1: $p > 0.5$. Como Pr{6 caras} $= \frac{1}{64} =$ $= 0.01562$ y Pr{5 ó 6 caras} $= \frac{6}{64} + \frac{1}{64} = 0.1094$, podemos rechazar H_0 al nivel 0.05, pero no al 0.01 (es decir, el resultado observado es significativo al nivel 0.05 pero no al 0.01).

Contraste de dos colas

Ahora hemos de decidir entre H_0: $p = 0.5$ y H_1: $p \neq 0.5$. Ya que Pr{0 ó 6 caras} $= \frac{1}{64} + \frac{1}{64} =$ $= 0.03125$, podemos rechazar H_0 al nivel 0.05 pero no al 0.01.

10.28. Resolver el Problema 10.27 si la moneda diese 5 caras.

Solución

Contraste de una cola

Como Pr{5 ó 6 caras} $= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.1094$, no podemos rechazar H_0 al nivel 0.05 ni al 0.01.

Contraste de dos colas

Como Pr{0 ó 1 ó 5 ó 6 caras} $= 2(\frac{7}{64}) = 0.2188$, no podemos rechazar H_0 al nivel 0.05 ni al 0.01.

PROBLEMAS SUPLEMENTARIOS

CONTRASTES DE MEDIAS Y PROPORCIONES USANDO LA DISTRIBUCION NORMAL

10.29. Una urna contiene fichas rojas y azules. Para comprobar la hipótesis de que hay tantas de un color como del otro, tomamos una muestra de 64 fichas con reposición y adoptamos la siguiente regla de decisión:

Aceptar la hipótesis si se sacan entre 28 y 36 rojas.

Rechazarla en caso contrario.

(*a*) Hallar la probabilidad de rechazar la hipótesis, siendo ésta verdadera.
(*b*) Representar gráficamente la regla de decisión y el resultado de (*a*).

10.30. (*a*) ¿Qué regla de decisión adoptaría en el Problema 10.29 si se exige que la probabilidad de rechazar la hipótesis, siendo ésta cierta, no sea mayor que 0.01 (o sea, si se desea un nivel de significación 0.01)?
(*b*) ¿A qué nivel de confianza aceptaría la hipótesis?
(*c*) ¿Cuál sería la regla de decisión si se adoptara el nivel de significación 0.05?

10.31. Supongamos que en el Problema 10.29 queremos comprobar la hipótesis de que hay mayor proporción de rojas que de azules.
(*a*) ¿Qué tomaría como hipótesis nula y como hipótesis alternativa?
(*b*) ¿Usaría un contraste de una o de dos colas? ¿Por qué?
(*c*) ¿Qué regla de decisión adoptaría para un nivel de significación de 0.05?
(*d*) ¿Cuál es la regla de decisión si el nivel de significación es 0.01?

10.32. Se tira un par de dados 100 veces y se ve que aparece suma 7 en 23 ocasiones. Contrastar la hipótesis de que los dados son buenos al nivel de significación 0.05 mediante un contraste de (*a*) una cola y (*b*) dos colas. Discutir las razones, si las hay, para preferir uno de ellos.

10.33. Rehacer el Problema 10.32 si el nivel de significación es 0.01.

10.34. Un fabricante afirma que al menos el 95% del equipamiento que ha suministrado a un cliente es acorde a las especificaciones. El examen de una muestra de 200 piezas revela que 18 eran defectuosas. Contrastar su afirmación al nivel de significación (*a*) 0.01 y (*b*) 0.05.

10.35. El porcentaje de grados A en un curso de Física de cierta Universidad en un largo período de tiempo fue del 10%. Durante un curso particular hubo 40 grados A entre 300 estudiantes. Contrastar la significación de tal resultado al nivel de significación (*a*) 0.05 y (*b*) 0.01.

10.36. Se ha visto experimentalmente que la tensión media de ruptura de cierta clase de sedal es 9.72 onzas (oz) con desviación típica de 1.40 oz. Recientemente, una muestra de 36 piezas ha dado una media de 8.93 oz. ¿Puede concluirse que ha empeorado la calidad al nivel de significación (*a*) 0.05 y (*b*) 0.01?

10.37. En un examen de muchos estudiantes de diversos colegios, la nota media ha sido 74.5 con desviación típica de 8.0. En un colegio particular, con 200 estudiantes, la nota media es 75.9. Discutir la significación de tal resultado al nivel de significación 0.05 desde el punto de vista de un contraste de (*a*) una cola y (*b*) de dos colas, explicando cuidadosamente qué conclusiones se desprenden de ellos.

10.38. Resolver el Problema 10.37 al nivel de significación 0.01.

CURVAS DE OPERACION CARACTERISTICAS

10.39. Refiriéndonos al Problema 10.29, hallar la probabilidad de aceptar la hipótesis de que haya igual proporción de rojas y azules cuando la proporción real p de fichas rojas es (*a*) 0.6, (*b*) 0.7, (*c*) 0.8, (*d*) 0.9 y (*e*) 0.3.

10.40. Representar los resultados del Problema 10.39 en un gráfico de (*a*) β versus p y (*b*) $1 - \beta$ versus p. Compararlos con los del Problema 10.12, considerando la analogía de fichas rojas y azules con cara y cruz, respectivamente.

10.41. (*a*) Resolver los Problemas 10.13 y 10.14 si se acuerda tomar una muestra de 400 sogas.
(*b*) ¿Qué conclusión se desprende acerca de los riesgos de error de Tipo II cuando se aumenta el tamaño de la muestra?

10.42. Construir (*a*) una curva OC y (*b*) una curva de potencia, para el Problema 10.31. Compararlas con las del Problema 10.14.

GRAFICOS DE CONTROL DE CALIDAD

10.43. En el pasado, cierto tipo de sedal tenía una tensión de ruptura media de 8.64 oz con desviación típica de 1.28 oz. Para determinar si el producto mantiene su calidad se toma una muestra de 16 piezas cada 3 horas. Registrar los límites de control (*a*) 99.73 (o 3σ), (*b*) 99% y (*c*) 95% sobre un gráfico de control de calidad y explicar sus aplicaciones.

10.44. En promedio, un 3% de las tuercas fabricadas por una empresa son defectuosas. Para mantener esa calidad de producción, se toma una muestra de 200 tuercas cada 4 horas. Determinar los límites de control (*a*) 99% y (*b*) 95% para el número de tuercas defectuosas en cada muestra. Nótese que sólo se necesitan *límites superiores de control* en este caso.

CONTRASTES MEDIANTE DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.45. Una muestra de 100 bombillas de la marca *A* dan vida media de 1190 h y desviación típica de 90 h. Una muestra de 75 bombillas de la marca *B* dan vida media de 1230 h y desviación típica de 120 h. ¿Hay diferencia entre las vidas medias de esas dos marcas de bombillas al nivel de significación (*a*) 0.05 y (*b*) 0.01?

10.46. En el Problema 10.45, contrastar la hipótesis de que las bombillas de la marca *B* son de más calidad que las del *A*, usando nivel de significación (*a*) 0.05 y (*b*) 0.01. Explicar las diferencias entre estos resultados y los citados en la última parte del Problema 10.45. ¿Contradicen estos resultados a los del Problema 10.45?

10.47. En un examen de ortografía, la nota media de 32 niños ha sido 72 con una desviación típica de 8, mientras que la nota media de 36 niñas ha sido 75 con una desviación típica de 6. Contrastar la hipótesis de que al nivel de significación (*a*) 0.05 y (*b*) 0.01, las niñas superan a los niños en ortografía.

10.48. Para comprobar los efectos de un nuevo fertilizante en la producción de trigo, se escogieron 60 campos cuadrados de iguales áreas, calidades de tierra, horas de sol, etc. Se utilizó en 30 de ellos el nuevo fertilizante y el antiguo a los demás. El número medio de bushels (bu) de trigo cosechados por cuadrado fueron 18.2 bu con desviación típica de 0.63 bu, en los del nuevo fertilizante, y 17.8 bu con una desviación típica de 0.54 bu, en los del antiguo. Usando nivel de significación de (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que el nuevo fertilizante es mejor que el antiguo.

10.49. Muestras aleatorias de 200 piezas producidas por una máquina *A* y 100 fabricadas por otra *B* dieron 19 y 5 piezas defectuosas, respectivamente. Contrastar las hipótesis de que (*a*) las dos máquinas tienen distinta calidad de producción y (*b*) la *B* es mejor que la *A*. Usar el nivel de significación 0.05.

10.50. Dos urnas *A* y *B* contienen el mismo número de fichas, pero la proporción de rojas y blancas es desconocida en ambas. Una muestra de 50 fichas tomada con reposición en cada una de ellas dio 32 rojas en la urna *A* y 23 en la *B*. Con el nivel de significación 0.05, contrastar las hipótesis de que (*a*) la proporción de rojas es la misma en las dos urnas y (*b*) *A* tiene mayor proporción de rojas que *B*.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

10.51. Con referencia al Problema 10.23, hallar el número mínimo de cuestiones que un estudiante debe contestar correctamente para que el profesor esté seguro con nivel de significación de (*a*) 0.05, (*b*) 0.01, (*c*) 0.001 y (*d*) 0.06 de que no ha sido por azar. Discutir los resultados.

10.52. Construir gráficos similares a los del Problema 10.10 para el Problema 10.24.

10.53. Resolver los Problemas 10.23 al 10.25 cambiando en la regla de decisión el 7 por 8.

10.54. En 8 tiradas una moneda ha dado 7 caras. ¿Podemos rechazar la hipótesis de que la moneda es buena al nivel de significación (*a*) 0.05, (*b*) 0.10 y (*c*) 0.01? Usar un contraste bilateral.

10.55. Repetir el Problema 10.54 con contraste unilateral.

10.56. Repetir el Problema 10.54 si la moneda diera cara las 8 veces.

10.57. Repetir el Problema 10.54 si la moneda diera cara 6 veces.

10.58. Una bolsa contiene un gran número de bolas rojas y blancas. Una muestra de 8 bolas da 6 blancas y 2 rojas. Mediante contrastes y nivel de significación adecuados, discutir la proporción de rojas y blancas en la bolsa.

10.59. Discutir cómo se puede recurrir a la teoría del muestreo para investigar las proporciones de distintos tipos de peces en un lago.

CAPITULO 11

Teoría de pequeñas muestras

PEQUEÑAS MUESTRAS

En capítulos precedentes hemos hecho uso de que para muestras de tamaño $N > 30$, llamadas *grandes muestras*, las distribuciones de muestreo de muchos estadísticos son aproximadamente normales, siendo la aproximación tanto mejor cuanto mayor sea N. Para muestras de tamaño menor que 30, llamadas *pequeñas muestras*, esa aproximación no es buena y empeora al decrecer N, de modo que son precisas ciertas modificaciones.

El estudio de la distribución de muestreo de estadísticos para pequeñas muestras se llama *teoría de pequeñas muestras*. Sin embargo, un nombre más apropiado sería *teoría exacta del muestreo*, pues sus resultados son válidos tanto para pequeñas muestras como para grandes. En ese capítulo analizamos tres distribuciones importantes: la distribución de Student, la distribución ji-cuadrado y la distribución F.

DISTRIBUCION t DE STUDENT

Definamos el estadístico

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}/\sqrt{N}} \tag{1}$$

que es análogo al estadístico z dado por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

(véase pág. 225).

Si consideramos muestras de tamaño N tomadas de una población normal (o casi normal) con media μ y si para cada una calculamos t, usando la media muestral $\bar{X}$ y la desviación típica muestral s o $\hat{s}$, puede obtenerse la distribución de muestreo para t. Esta distribución (véase Figura 11.1) viene dada por

$$Y = \frac{Y_0}{\left(1 + \dfrac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \dfrac{t^2}{\nu}\right)^{(\nu+1)/2}} \tag{2}$$

donde Y_0 es una constante que depende de N tal que el área total bajo la curva es 1, y donde la constante $\nu = (N - 1)$ se llama el *número de grados de libertad* (ν es la letra griega nu). Para una definición de grados de libertad, véase página 255.

La distribución (2) se llama *distribución t de Student* en honor de su descubridor, W. S. Gossett, quien publicó su obra bajo el pseudónimo de «Student» («estudiante») a principios de este siglo.

Para grandes valores de ν o de N (ciertamente $N \geqslant 30$), las curvas (2) se ajustan mucho a la curva normal canónica

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

como se muestra en la Figura 11.1.

INTERVALOS DE CONFIANZA

Al igual que se hizo con la distribución normal, se pueden definir los intervalos de confianza 95%, 99%, u otros, usando la tabla de la distribución t en el Apéndice III. De esta forma podemos estimar la media de la población dentro de límites especificados.

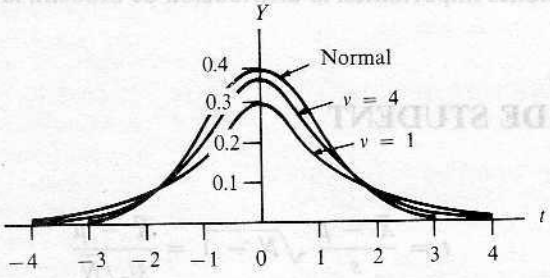


Figura 11.1. Distribución t de Student para varios valores de ν.

Por ejemplo, si $-t_{.975}$ y $t_{.975}$ son los valores de t para los que el 2.5% del área está en cada cola de la distribución t, entonces el intervalo de confianza 95% para t es

$$-t_{.975} < \frac{\bar{X} - \mu}{s}\sqrt{N-1} < t_{.975} \tag{3}$$

de donde vemos que μ se estima que estará en el intervalo

$$\bar{X} - t_{.975}\frac{S}{\sqrt{N-1}} < \mu < \bar{X} + t_{.975}\frac{S}{\sqrt{N-1}} \tag{4}$$

con el 95% de confianza (o sea, probabilidad 0,95).

Nótese que $t_{.975}$ representa el valor 97.5 percentil, mientras que $t_{.025} = -t_{.975}$ representa el valor 2.5 percentil.

En general, podemos representar límites de confianza para medias poblacionales por

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}} \tag{5}$$

donde los valores $\pm t_c$, llamados *valores críticos* o *coeficientes de confianza*, dependen del nivel de confianza deseado y del tamaño de la muestra. Pueden verse en el Apéndice III.

Comparando las ecuaciones (5) con los límites de confianza ($\bar{X} \pm z_c\sigma/\sqrt{N}$) del Capítulo 9, página 211, vemos que para pequeñas muestras debemos sustituir z_c (obtenido de la distribución normal) por t_c (obtenido de la distribución de Student) y σ con $\sqrt{N/(N-1)}s = \hat{s}$, que es la estimación muestral de σ. Cuando N crece, ambos métodos tienden a coincidir.

CONTRASTES DE HIPOTESIS Y SIGNIFICACION

Los contrastes de hipótesis y significación o reglas de decisión (discutidos en el Capítulo 10), se extienden fácilmente a pequeñas muestras. La única diferencia consiste en que el *estadístico z* queda sustituido por el *estadístico t*.

1. **Medias.** Para constrastar la hipótesis H_0 de que una población normal tiene medida μ, usamos el estadístico t

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{N} \tag{6}$$

donde $\bar{X}$ es la media de una muestra de tamaño N. Esto es análogo al uso de

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

para grandes N, excepto que se usa $\hat{s} = \sqrt{N/(N-1)}s$ en lugar de σ. La diferencia es que mientras z está normalmente distribuida, t sigue la distribución de Student. Al crecer N, ambas tienden a coincidir.

2. **Diferencias de medias.** Supongamos que se toman dos muestras aleatorias de tamaños N_1 y N_2 de poblaciones normales cuyas desviaciones típicas son iguales ($\sigma_1 = \sigma_2$). Y supongamos además que estas dos muestras tienen medias $\bar{X}_1$ y $\bar{X}_2$ y desviaciones típicas s_1 y s_2, respectivamente. Para contrastar la hipótesis H_0 de que las muestras provienen de la misma población (o sea, $\mu_1 = \mu_2$ y también $\sigma_1 = \sigma_2$),

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \quad \text{donde} \quad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{7}$$

Su distribución es una distribución de Student con $\nu = N_1 + N_2 - 2$ grados de libertad. El uso de (7) aparece como plausible si se hace $\sigma_1 = \sigma_2 = \sigma$ en el z de la ecuación (2) del Capítulo 10, y se usa entonces como estimación de σ^2 la media ponderada

$$\frac{(N_1 - 1)\hat{s}_1^2 + (N_2 - 1)\hat{s}_2^2}{(N_1 - 1) + (N_2 - 1)} = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}$$

donde $\hat{s}_1^2$ y $\hat{s}_2^2$ son las estimaciones sin sesgo de σ_1^2 y σ_2^2 (véase Propiedad 3 en la página 95).

DISTRIBUCION JI-CUADRADO

Definamos el estadístico

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{\sigma^2} \tag{8}$$

donde χ es la letra griega ji y χ^2 se lee «ji-cuadrado».

Si consideramos muestras de tamaño N tomadas de una población normal con desviación típica σ, y si para cada muestra calculamos χ^2, se obtiene para χ^2 una distribución de muestreo, llamada *distribución ji-cuadrado*, que viene dada por

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \tag{9}$$

donde $\nu = N - 1$ es el *número de grados de libertad*, e Y_0 es una constante que depende de ν tal que el área total bajo la curva es 1. La distribución ji-cuadrado correspondientes a varios valores ν se muestran en la Figura 11.2. El máximo de Y ocurre en $\chi^2 = \nu - 2$ para $\nu \geq 2$.

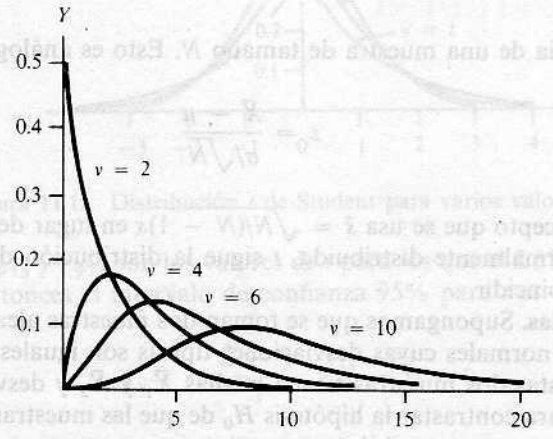


Figura 11.2. Distribuciones ji-cuadrado para varios valores de ν.

INTERVALOS DE CONFIANZA PARA LA DISTRIBUCION JI-CUADRADO

Como se hizo con la distribución normal y con la distribución de Student, podemos definir los intervalos y límites de confianza 95%, 99%, u otros, usando la tabla de la distribución ji-cuadrado

en el Apéndice IV. De este modo podemos estimar, dentro de límites especificados, la desviación típica de la población en términos de una desviación típica muestral s.

Por ejemplo, si $\chi^2_{.025}$ y $\chi^2_{.975}$ son los valores de χ^2 (llamados *valores críticos*) para los que el 2.5% del área está en cada cola de la distribución, entonces el intervalo de confianza 95% es

$$\chi^2_{.025} < \frac{Ns^2}{\sigma^2} < \chi^2_{.975} \tag{10}$$

del cual vemos que σ se estima que estará en el intervalo

$$\frac{s\sqrt{N}}{\chi_{.975}} < \sigma < \frac{s\sqrt{N}}{\chi_{.025}} \tag{11}$$

con el 95% de confianza. Otros intervalos de confianza se hallan de forma parecida. Los valores de $\chi_{.025}$ y $\chi_{.975}$ representan, respectivamente, los valores 2.5 y 97.5 percentil.

El Apéndice IV da los valores percentiles correspondientes al número de grados de libertad ν. Para grandes $\nu(\nu \geq 30)$, podemos utilizar el hecho de que $(\sqrt{2\chi^2} - \sqrt{2\nu - 1})$ está casi normalmente distribuida con media 0 y desviación típica 1; luego se pueden usar tablas de la distribución normal si $\nu \geq 30$. Entonces, si χ^2_p y z_p son los p-ésimos percentiles de la distribución ji-cuadrado y de la distribución normal, respectivamente, tenemos

$$\chi^2_p = \tfrac{1}{2}(z_p + \sqrt{2\nu - 1})^2 \tag{12}$$

En esos casos, hay muy buen acuerdo con los resultados obtenidos en los Capítulos 8 y 9

Para otras aplicaciones de la distribución ji-cuadrado, véase el Capítulo 12.

GRADOS DE LIBERTAD

Para el cálculo de un estadístico tal como (1) u (8), es necesario emplear tanto observaciones de muestras como propiedades de ciertos parámetros de la población. Si estos parámetros son desconocidos, hay que estimarlos a partir de la muestra.

El *número de grados de libertad* de un estadístico, generalmente denotado por ν, se define como el número N de observaciones independientes en la muestra (o sea, el tamaño de la muestra) menos el número k de parámetros de la población, que debe ser estimado a partir de observaciones muestrales. En símbolos, $\nu = N - k$.

En el caso del estadístico (1), el número de observaciones independientes en la muestra es N, de donde podemos calcular $\bar{X}$ y s. Sin embargo, como debemos estimar μ, $k = 1$ y $\nu = N - 1$).

En el caso del estadístico (8), el número de observaciones independientes en la muestra es N, de donde podemos calcular s. Sin embargo, como debemos estimar σ, $k = 1$ y $\nu = N - 1$.

LA DISTRIBUCION F

Como hemos visto, es importante en algunas aplicaciones conocer la distribución de muestreo de la diferencia en medias $(\bar{X}_1 - \bar{X}_2)$ de dos muestras. De la misma manera, podemos necesitar la

distribución de muestreo de la diferencia en varianzas $(S_1^2 - S_2^2)$. Resulta, sin embargo, que esta distribución es complicada, por lo que en lugar de eso, consideramos el estadístico S_1^2/S_2^2, ya que un cociente grande o pequeño indicará una gran diferencia, mientras un cociente cercano a 1 indica una pequeña diferencia. Su distribución de muestreo se llama *distribución F*, en honor de R. A. Fisher.

Más concretamente, sean dos muestras, 1 y 2, de tamaños N_1 y N_2, respectivamente, tomadas de dos poblaciones normales (o casi) con varianzas σ_1^2 y σ_2^2. Definamos el estadístico

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} \tag{13}$$

donde

$$\hat{S}_1^2 = \frac{N_1 S_1^2}{N_1 - 1} \qquad \hat{S}_2^2 = \frac{N_2 S_2^2}{N_2 - 1} \tag{14}$$

(véase pág. 208). Entonces la distribución de muestreo de F se llama distribución F de Fisher, o en breve, distribución F, con $\nu_1 = N_1 - 1$ y $\nu_2 = N_2 - 1$ grados de libertad. Esta distribución viene dada por

$$Y = \frac{CF^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}} \tag{15}$$

donde C es una constante que depende de ν_1 y ν_2 tal que el área total bajo la curva es 1. La curva tiene una forma del tipo que indica la Figura 11.3, aunque esa forma puede variar considerablemente según los valores de ν_1 y ν_2.

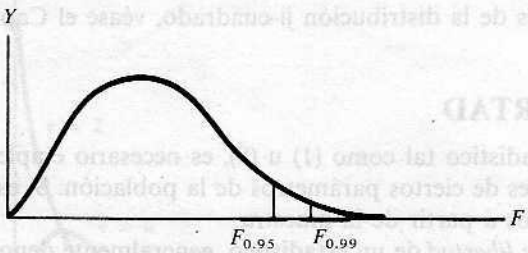


Figura 11.3.

Los Apéndices V y VI dan valores percentiles de F para los que las áreas en la cola de la derecha son 0.05 y 0.01, denotadas $F_{.95}$ y $F_{.99}$, respectivamente. Representando los niveles de significación 5% y 1%, éstos se pueden usar para determinar si la varianza S_1^2 es significativamente mayor que S_2^2, o no. En la práctica, la muestra con mayor varianza se elige como muestra 1.

PROBLEMAS RESUELTOS

DISTRIBUCION t DE STUDENT

11.1. La Figura 11.4 recoge el gráfico de la distribución de Student con 9 grados de libertad. Hallar el valor de t_1 para el que (*a*) el área sombreada de la derecha es 0.05, (*b*) el área total sombreada es 0.05, (*c*) el

área total sin sombrear es 0.99, (*d*) el área en sombra de la izquierda es 0.01 y (*e*) el área a la izquierda de t_1 es 0.90.

Solución

(*a*) Si el área sombreada de la derecha es 0.05, el área a la izquierda de t_1 es $(1 - 0.05) = 0.95$ y t_1 es el 95 percentil, $t_{.95}$. En el Apéndice III, buscamos el 9 en la columna encabezada con v, y después nos desplazamos a la derecha hasta la columna $t_{.95}$; el resultado, 1.83, es el valor pedido de t.

(*b*) Si el área total sombreada es 0.05, la de la derecha es 0.025 por simetría. Luego el área a la izquierda de t_1 es $(1 - 0.025) = 0.975$ y t_1 representa el 97.5 percentil, $t_{.975}$. En el Apéndice III encontramos que el valor requerido de t es 2.26.

(*c*) Si el área total sin sombrear es 0.99, el área en sombra es $(1 - 0.99) = 0.01$, y su mitad derecha es 0.005. En el Apéndice III vemos que $t_{.995} = 3.25$.

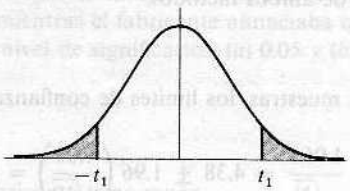


Figura 11.4.

(*d*) Si el área sombreada de la izquierda es 0.01, por simetría la de la derecha es igual. El Apéndice III da $t_{.99} = 2.82$. Luego el valor crítico de t para el cual el área sombreada de la izquierda es 0.01 es -2.82.

(*e*) Si el área a la izquierda de t_1 es 0.90, el t_1 corresponde al 90 percentil $t_{.90}$, que según el Apéndice III es igual a 1.38.

11.2. Hallar los valores críticos de t para los que el área de la cola derecha de la distribución t es 0.05 si el número de grados de libertad, v, es (*a*) 16, (*b*) 27 y (*c*) 200.

Solución

En el Apéndice III, columna $t_{.95}$, hallamos los valores (*a*) 1.75 para $v = 16$; (*b*) 1.70 para $v = 27$; y (*c*) 1.645 para $v = 200$. (Este último es el valor que se obtendría de la curva normal; en el Apéndice III corresponde a la entrada marcada ∞ en la última fila).

11.3. Los coeficientes de confianza 95% (con dos colas) para la distribución normal vienen dados por ± 1.96. ¿Cuáles son los correspondientes coeficientes para la distribución t si (*a*) $v = 9$, (*b*) $v = 20$, (*c*) $v = 30$ y (*d*) $v = 60$?

Solución

Para los coeficientes de confianza 95% (con dos colas), el área total sombreada en la Figura 11.4 ha de ser 0.05. Así que el área de la cola derecha es 0.025 y el correspondiente valor crítico de t es $t_{.975}$. Entonces los coeficientes de confianza pedidos son $\pm t_{.975}$; para los valores dados de v, son (*a*) ± 2.26, (*b*) ± 2.09, (*c*) ± 2.04 y (*d*) ± 2.00.

11.4. Una muestra de 10 medidas del diámetro de una esfera dan una media $\bar{X} = 4.38$ cm y una desviación típica $s = 0.06$ cm. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para el diámetro verdadero.

Solución

(*a*) Los límites de confianza 95% vienen dados por $\bar{X} \pm t_{.975}(s/\sqrt{N-1})$.

Como $\nu = N - 1 = 10 - 1 = 9$, encontramos $t_{.975} = 2.26$ [véase también el Problema 11.3(*a*)]. Entonces, usando $\bar{X} = 4.38$ y $s = 0.06$, los requeridos límites de confianza 95% son $4.38 \pm 2.26(0.06/\sqrt{10-1}) = 4.38 \pm 0.0452$ cm. Luego podemos tener 95% de confianza de que la verdadera media está entre $(4.38 - 0.045) = 4.335$ cm y $(4.38 + 0.045) = 4.425$ cm.

(*b*) Los límites de confianza 99% están dados por $\bar{X} \pm t_{.995}(s/\sqrt{N-1})$.

Para $\nu = 9$, $t_{.995} = 3.25$. Entonces los límites de confianza 99% son $4.38 \pm 3.25(0.06/\sqrt{10-1}) = 4.38 \pm 0.0650$ cm, y el intervalo de confianza 99% es 4.315 a 4.445 cm.

11.5. (*a*) Repetir el Problema 11.4 suponiendo que son válidos los métodos de la teoría de grandes muestras.

(*b*) Comparar los resultados de ambos métodos.

Solución

(*a*) En el método de grandes muestras, los límites de confianza 95% son

$$\bar{X} \pm \frac{1.96\sigma}{\sqrt{N}} = 4.38 \pm 1.96\left(\frac{0.06}{\sqrt{10}}\right) = 4.38 \pm 0.037 \text{ cm}$$

donde se ha usado la desviación típica muestral 0.06 como estimación de σ. Análogamente, los límites de confianza 99% son

$$\bar{X} \pm \frac{2.58\sigma}{\sqrt{N}} = 4.38 \pm 2.58\left(\frac{0.06}{\sqrt{10}}\right) = 4.38 \pm 0.049 \text{ cm}$$

(*b*) En cada caso, los límites de confianza obtenidos usando la teoría exacta (pequeñas muestras) son mayores que los obtenidos por métodos de grandes muestras. Era de esperar, porque la precisión disponible con pequeñas muestras es menor que con muestras grandes.

11.6. Hace tiempo, una máquina producía arandelas de 0.05 pulgadas(in) de espesor. Para determinar si sigue en buen estado, se toma una muestra de 10 arandelas, que dan un espesor medio de 0.053 in con desviación típica de 0.003 in. Contrastar la hipótesis de que la máquina sigue funcionando bien, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Queremos decidir entre las hipótesis:

H_0: $\mu = 0.050$, y la máquina sigue en buen estado.

H_1: $\mu \neq 0.050$, y la máquina está deteriorada.

Por tanto, se precisa un contraste de dos colas. Bajo la hipótesis H_0, tenemos

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{0.053 - 0.050}{0.003}\sqrt{10-1} = 3.00$$

(*a*) Para un test de dos colas al nivel de significación 0.05, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t está en el intervalo $-t_{.975}$ a $t_{.975}$, que para $10 - 1 = 9$ grados de libertad es desde -2.26 a 2.26.

Rechazarla en caso contrario.

Como $t = 3.00$, rechazamos H_0 al nivel 0.05.

(*b*) Para un test de dos colas al nivel de significación 0.01, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t está en el intervalo $-t_{.995}$ a $t_{.995}$, que para $10 - 1 = 9$ grados de libertad es desde -3.25 a 3.25.

Rechazarla en caso contrario.

Como $t = 3.00$, aceptamos H_0 al nivel 0.01.

Como podemos rechazar H_0 al nivel 0.05 pero no al 0.01, decimos que el resultado de la muestra es *probablemente significativo* (véase final del Problema 10.5). Sería recomendable revisar la máquina o al menos tomar otra muestra.

11.7. Una prueba con 6 sogas de un cierto fabricante dio una tensión media de ruptura de 7750 lb y una desviación típica de 145 lb, mientras el fabricante anunciaba que era de 8000 lb. ¿Puede sostenerse la afirmación del fabricante al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Hemos de decidir entre:

H_0: $\mu = 8000$ lb, y el fabricante tiene razón.

H_1: $\mu < 8000$ lb, y el fabricante no tiene razón.

Hay que aplicar un contraste de una cola. Bajo la hipótesis H_0, tenemos

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N - 1} = \frac{7750 - 8000}{145}\sqrt{6 - 1} = -3.86$$

(*a*) Para un contraste de una cola al nivel de significación 0.05, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t es mayor que $-t_{.95}$, que para $6 - 1 = 5$ grados de libertad quiere decir $t > -2.01$.

Rechazar H_0 en caso contrario.

Como $t = -3.86$, rechazamos H_0.

(*b*) Para un contraste de una cola al nivel de significación 0.01, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t es mayor que $-t_{.99}$, que para 5 grados de libertad quiere decir $t > -3.36$.

Rechazar H_0 en caso contrario.

Como $t = -3.86$, rechazamos H_0.

Deducimos que es muy improbable que el fabricante tuviese razón.

11.8. Los cocientes de inteligencia(IQ) de 16 estudiantes de un barrio dieron una media de 107 con desviación típica 10, y 14 estudiantes de otro barrio dieron media 112 con desviación típica 8. ¿Hay diferencia significativa entre los IQ de los dos grupos al nivel de significación (*a*) 0.01 y (*b*) 0.05?

Solución

Si μ_1 y μ_2 denotan los IQ medios de la población de ambos barrios, respectivamente, tenemos que decidir entre:

H_0: $\mu_1 = \mu_2$, y no hay diferencia esencial entre los dos barrios.

H_1: $\mu_1 \neq \mu_2$, y hay diferencia significativa entre ellos.

Bajo la hipótesis H_0,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde } \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Luego

$$\sigma = \sqrt{\frac{16(10)^2 + 14(8)^2}{16 + 14 - 2}} = 9.44 \qquad \text{y} \qquad t = \frac{112 - 107}{9.44\sqrt{1/16 + 1/14}} = 1.45$$

(*a*) Con un contraste bilateral al nivel de significación 0.01, rechazaríamos H_0 si t estuviera fuera del rango $-t_{.995}$ a $t_{.995}$, que para $(N_1 + N_2 - 2) = (16 + 14 - 2) = 28$ grados de libertad es el rango -2.76 a 2.76. Así pues, no podemos rechazar H_0 al nivel de significación 0.01.

(*b*) Con un contraste bilateral al nivel de significación 0.05, rechazaríamos H_0 si t estuviera fuera del rango $-t_{.975}$ a $t_{.975}$, que para 28 grados de libertad es el rango -2.05 a 2.05. Así pues, no podemos rechazar H_0 al nivel de significación 0.01.

Concluimos que no hay diferencia significativa entre los dos grupos.

11.9. Con el fin de probar un fertilizante, se tomaron 24 parcelas de la misma área, de las que la mitad se trataron con ese fertilizante y las otras no (el grupo de control); por lo demás, las condiciones fueron idénticas para todas ellas. La producción media de trigo en las parcelas sin tratar fue de 4.8 bushels(bu) con desviación típica de 0.40 bu, y en las tratadas fue 5.1 bu con desviación típica de 0.36 bu. ¿Podemos concluir que se produjo mejora a causa del fertilizante de significación (*a*) 1% y (*b*) 5 %?

Solución

Si μ_1 y μ_2 denotan las producciones medias de trigo de las poblaciones tratada y sin tratar, respectivamente, hemos de decidir entre:

H_0: $\mu_1 = \mu_2$, y la diferencia es fortuita.

H_1: $\mu_1 > \mu_2$, y el fertilizante mejora la cosecha.

Bajo la hipótesis H_0,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Así pues $\sigma = \sqrt{\dfrac{12(0.40)^2 + 12(0.36)^2}{12 + 12 - 2}} = 0.397$ y $t = \dfrac{5.1 - 4.8}{0.397\sqrt{1/12 + 1/12}} = 1.85$

(*a*) Con un contraste de una cola al nivel de significación 0.01, rechazaremos H_0 si t es mayor que $t_{.99}$, que para $(N_1 + N_2 - 2) = (12 + 12 - 2) = 22$ grados de libertad es 2.51. Luego no podemos rechazar H_0 al nivel de significación 0.01.

(*b*) Con un contraste de una cola al nivel de significación 0.05, rechazaremos H_0 si t es mayor que $t_{.95}$, que para 22 grados de libertad es 1.72. Luego podemos rechazar H_0 al nivel de significación 0.05.

Concluimos que la mejora causada por el fertilizante es *probablemente significativa*. No obstante, antes de sacar conclusiones definitivas sería deseable una evidencia más nítida.

DISTRIBUCION JI-CUADRADO

11.10. El gráfico de la distribución ji-cuadrado con 5 grados de libertad se muestra en la Figura 11.5. Hallar los valores críticos de χ^2 para los que (*a*) el área sombreada a la derecha es 0.05, (*b*) el área total en sombra es 0.05, (*c*) el área sombreada de la izquierda es 0.10 y (*d*) el área sombreada a la derecha es 0.01.

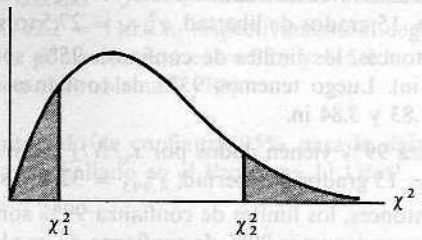


Figura 11.5.

Solución

(*a*) Si el área sombreada de la derecha es 0.05, el área a la izquierda de χ_2^2 es $(1 - 0.05) = 0.95$ y χ_2^2 representa el 95 percentil, χ_{95}^2. Buscando en el Apéndice IV el 5 bajo la columna v, y entonces desplazándonos a la derecha hasta la columna $\chi_{.95}^2$, resulta 11.1, que es el requerido valor crítico de χ^2.

(*b*) Como la distribución no es simétrica, hay muchos valores críticos para los que el área total sombreada es 0.05. Por ejemplo, la de la derecha podría ser 0.04 y la de la izquierda 0.01. Es costumbre, sin embargo, salvo que se especifique lo contrario, escoger ambas iguales. En este caso, cada área será de 0.025.

Si el área sombreada a la derecha es 0.025, el área a la izquierda de χ_2^2 es $1-0.025 = 0.975$ y χ_2^2 representa el 95 percentil, $\chi_{.975}^2$, que por el Apéndice IV es 12.8. Análogamente, si el área sombreada de la izquierda es 0.025, el área a la izquierda de χ_1^2 es 0.025 y χ_1^2 representa el 2.5 percentil, $\chi_{.025}^2$, que es 0.831. Luego los valores críticos son 0.831 y 12.8.

(*c*) Si el área sombreada de la derecha es 0.10, χ_1^2 representa el 10° percentil, $\chi_{.10}^2$, que es 1.61.

(*d*) Si el área sombreada de la derecha es 0.01, el área a la izquierda de χ_2^2 es 0.99 y χ_2^2 representa el 99 percentil, $\chi_{.99}^2$, que es 15.1.

11.11. Hallar los valores críticos de χ^2 para los cuales el área de la cola derecha de la distribución ji-cuadrado es 0.05, siendo el número de grados de libertad, v, igual a (*a*) 15, (*b*) 21 y (*c*) 50.

Solución

Usando el Apéndice IV, se ven en la columna encabezada por $\chi_{.95}^2$ los valores (*a*) 25.0 para $v = 15$, (*b*) 32.7 para $v = 21$ y (*c*) 67.5 para $v = 50$.

11.12. Hallar la mediana de χ^2 correspondiente a (*a*) 9, (*b*) 28 y (*c*) 40 grados de libertad.

Solución

Usando el Apéndice IV, vemos en la columna encabezada por $\chi_{.50}^2$ (ya que la mediana es el 50 percentil) el valor (*a*) 8.34 para $v = 9$; (*b*) 27.3 para $v = 28$; y (*c*) 39.3 para $v = 40$.

Conviene fijarse en que las medianas son casi iguales al número de grados de libertad. De hecho, para $v > 10$, los valores de la mediana son $(v - 0.7)$, como se ve en la tabla.

11.13. La desviación típica de las alturas de 16 estudiantes varones tomados al azar en un colegio de 1000 alumnos es 2.40 in. Hallar los límites de confianza (*a*) 95% y (*b*) 99% de la desviación típica para todos los estudiantes de ese colegio.

Solución

(*a*) Los límites de confianza 95% vienen dados por $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{.975} = 27.5$ (o sea $\chi_{.975} = 5.24$) y $\chi^2_{.025} = 6.26$ (o sea $\chi_{.005} = 2.50$). Entonces, los límites de confianza 95% son $2.40\sqrt{16}/5.24$ y $2.40\sqrt{16}/2.50$ (es decir, 1.83 y 3.84 in). Luego tenemos 95% de confianza de que la desviación típica de la población está entre 1.83 y 3.84 in.

(*b*) Los límites de confianza 99% vienen dados por $s\sqrt{N}/\chi_{.995}$ y $s\sqrt{N}/\chi_{.005}$

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{.995} = 32.8$ (o sea $\chi_{.995} = 5.73$) y $\chi^2_{.005} = 4.60$, es decir $\chi_{.025} = 2.14$). Entonces, los límites de confianza 99% son $2.40\sqrt{16}/5.73$ y $2.40\sqrt{16}/2.14$ (es decir, 1.68 y 4.49 in). Luego tenemos 99% de confianza de que la desviación típica de la población está entre 1.68 y 4.49 in.

11.14. Hallar $\chi^2_{.95}$ para (*a*) $\nu = 50$ y (*b*) $\nu = 100$ grados de libertad.

Solución

Para $\nu > 30$ podemos usar el que $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ está casi normalmente distribuida con media 0 y desviación típica 1. Así que si z_p es el valor z percentil de la distribución normal canónica, podemos escribir, con muy buena aproximación,

$$\sqrt{2\chi_p^2} - \sqrt{2\nu - 1} = z_p \quad \text{o sea} \quad \sqrt{2\chi_p^2} = z_p + \sqrt{2\nu - 1}$$

de donde $\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2\nu - 1})^2$.

(*a*) Si $\nu = 50$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(50) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{99})^2 = 67.2$, que está en buen acuerdo con el valor 67.5 dado en el Apéndice IV.

(*b*) Si $\nu = 100$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(100) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{199})^2 = 124.0$ (valor real = 124.3).

11.15. La desviación típica de las vidas medias de una muestra de 200 lámparas es 100 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de todas las lámparas de ese tipo.

Solución

(*a*) Los límites de confianza 95% están dados por $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 200 - 1 = 199$ grados de libertad, encontramos (como en el Problema 11.14)

$$\chi^2_{.975} = \tfrac{1}{2}(z_{.975} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(1.96 + 19.92)^2 = 239$$
$$\chi^2_{.025} = \tfrac{1}{2}(z_{.025} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-1.96 + 19.92)^2 = 161$$

de donde $\chi_{.975} = 15.5$ y $\chi_{.025} = 12.7$. Entonces los límites de confianza 95% son $100\sqrt{200}/15.5 = 91.2$ h y $100\sqrt{200}/12.7 = 111.3$ h, respectivamente. Luego estamos 95% confiados de que la desviación típica de la población está entre 91.2 y 111.3 h.

Comparar esto con el Problema 9.17(*a*).

(*b*) Los límites de confianza 99% están dados por $s\sqrt{N}/\chi_{.995}$ y $s\sqrt{N}/\chi_{.005}$.

Para $\nu = 200 - 1 = 199$ grados de libertad,

$$\chi^2_{.995} = \tfrac{1}{2}(z_{.995} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(2.58 + 19.92)^2 = 253$$
$$\chi^2_{.005} = \tfrac{1}{2}(z_{.005} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-2.58 + 19.92)^2 = 150$$

de donde $\chi_{.995} = 15.9$ y $\chi_{.005} = 12.2$. Entonces los límites de confianza 99% son $100\sqrt{200}/15.9 = 88.9$ h y $100\sqrt{200}/12.2 = 115.9$ h, respectivamente. Luego estamos 99% confiados de que la desviación típica de la población está entre 88.9 y 115.9 h.

Comparar esto con el Problema 9.17(*b*).

11.16. ¿Es posible obtener un intervalo de confianza 95% para la desviación típica de la población cuya anchura sea menor que la del hallado en el Problema 11.15(*a*)?

Solución

Los límites de confianza para la desviación típica de la población hallados en el Problema 11.15(*a*) se obtuvieron escogiendo valores críticos de χ^2 tales que el área en cada cola era 2.5%. Es posible hallar otros límites de confianza eligiendo valores críticos de χ^2 para los que la suma de las áreas en las dos colas sea 5%, pero con áreas desiguales en las colas.

En la Tabla 11.1 se han recogido varios de tales valores críticos (obtenidos por los métodos del Problema 11.14), y los correspondientes intervalos de confianza 95%. De ahí vemos que un intervalo 95% con anchura de sólo 19.8 es el que va desde 91.0 a 110.8. Se puede lograr otro con menor anchura todavía continuando de esa forma, usando valores críticos como $\chi_{.031}$ y $\chi_{.981}$, $\chi_{.032}$ y $\chi_{.982}$, etc. En general, sin embargo, el decrecimiento que se consigue en el intervalo es despreciable y no merece la pena el trabajo exigido.

Tabla 11.1

Valores críticos	Intervalo de confianza del 95%	Anchura
$\chi_{.01} = 12.44$, $\chi_{.96} = 15.32$	92.3 a 113.7	21.4
$\chi_{.02} = 12.64$, $\chi_{.97} = 15.42$	91.7 a 111.9	20.2
$\chi_{.03} = 12.76$, $\chi_{.98} = 15.54$	91.0 a 110.8	19.8
$\chi_{.04} = 12.85$, $\chi_{.99} = 15.73$	89.9 a 110.0	20.1

11.17. Tiempo atrás, la desviación típica de los pesos de ciertos envases llenados por una máquina era 0.25 onzas(oz). Una muestra aleatoria de 20 envases ha dado una desviación típica de 0.32 oz. ¿Es significativo el aparente aumento en la variabilidad al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Hemos de decidir entre las hipótesis:

H_0: $\sigma = 0.25$ oz, y el resultado observado es fortuito.

H_1: $\sigma > 0.25$ oz, y la variabilidad ha aumentado realmente.

El valor de χ^2 para la muestra es

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{20(0.32)^2}{(0.25)^2} = 32.8$$

(*a*) Usando un contraste unilateral, rechazaríamos H_0 al nivel de significación 0.05 si el valor de χ^2 para que la muestra fuese mayor que $\chi^2_{.95}$, que es igualar a 30.1 para $\nu = 20 - 1 = 19$ grados de libertad. Así pues, rechazaríamos H_0 al nivel de significación 0.05.

(*b*) Usando un contraste unilateral, rechazaríamos H_0 al nivel de significación 0.01 si el valor de χ^2 para la muestra fuese mayor que $\chi^2_{.99}$, que es igual a 36.2 para 19 grados de libertad. Así pues, no rechazaríamos H_0 al nivel de significación 0.01.

Concluimos que la variabilidad ha crecido probablemente. Debiera hacerse una revisión de esa máquina.

DISTRIBUCION F

11.18. Dos muestras de tamaños 9 y 12 se han tomado en dos poblaciones normalmente distribuidas con varianzas respectivas 16 y 25. Si las varianzas muestrales son 20 y 8, determinar si la primera muestra tiene una varianza significativamente mayor que la segunda al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Para las dos muestras, 1 y 2, tenemos $N_1 = 9$, $N_2 = 12$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $S_1^2 = 20$ y $S_2^2 = 8$. Luego

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} = \frac{(9)(20)/(9 - 1)(16)}{(12)(8)/(12 - 1)(25)} = 4.03$$

(*a*) Los grados de libertad para el numerador y el denominador de F son $\nu_1 = N_1 - 1 = 9 - 1 = 8$ y $\nu_2 = N_2 - 1 = 12 - 1 = 11$. Entonces del Apéndice V vemos que $F_{.95} = 2.95$. Como la $F = 4.03$ calculada es mayor que 2.95, concluimos que la varianza de la muestra 1 es significativamente mayor que la de la muestra 2 al nivel de significación 0.05.

(*b*) Para $\nu_1 = 8$ y $\nu_2 = 11$, hallamos en el Apéndice VI que $F_{.01} = 4.74$. Luego no podemos concluir que la muestra 1 tenga varianza mayor que la muestra 2 al nivel de significación 0.01.

11.19. Se toman dos muestras de tamaños 8 y 10 de dos poblaciones normalmente distribuidas con varianzas respectivas 20 y 36. Hallar la probabilidad de que la varianza de la primera sea doble que la de la segunda.

Solución

Tenemos $N_1 = 8$, $N_2 = 10$, $\sigma_1^2 = 20$ y $\sigma_2^2 = 36$. Por tanto,

$$F = \frac{8S_1^2/(7)(20)}{10S_2^2/(9)(36)} = 1.85\,\frac{S_1^2}{S_2^2}$$

El número de grados de libertad para el numerador y el denominador son $\nu_1 = N_1 - 1 = 8 - 1 = 7$ y $\nu_2 = N_2 - 1 = 10 - 1 = 9$. Ahora bien, si S_1^2 es más del doble que S_2^2, entonces

$$F = 1.85\,\frac{S_1^2}{S_2^2} > (1.85)(2) = 3.70$$

Buscando 3.70 en los Apéndices V y VI, hallamos que la probabilidad es menor que 0.05 pero mayor que 0.01. Valores más precisos requieren una tabulación más exhaustiva de la distribución F.

PROBLEMAS SUPLEMENTARIOS

DISTRIBUCION t DE STUDENT

11.20. Para una distribución de Student con 15 grados de libertad, hallar el valor de t_1 tal que (*a*) el área a su derecha sea 0.01, (*b*) el área a su izquierda sea 0.95, (*c*) el área a su derecha sea 0.10, (*d*) la suma de áreas a la derecha de t_1 y a la izquierda de $-t_1$ sea 0.01 y (*e*) el área entre $-t_1$ y t_1 sea 0.95.

11.21. Hallar los valores críticos de t para los que el área de la cola derecha de la distribución t es 0.01 si el número de grados de libertad ν, es igual a (*a*) 4, (*b*) 12, (*c*) 25, (*d*) 60 y (*e*) 150.

11.22. Hallar los valores de t_1 para la distribución de Student que satisfacen cada una de las condiciones siguientes:

(*a*) El área entre $-t_1$ y t_1 es 0.90 y $\nu = 25$.
(*b*) El área a la izquierda de $-t_1$ es 0.025 y $\nu = 20$.
(*c*) La suma de áreas a la derecha de t_1 y a la izquierda de $-t_1$ es 0.01 y $\nu = 5$.
(d) El área a la derecha de t_1 es 0.55 y $\nu = 16$.

11.23. Si una variable U tiene una distribución de Student con $\nu = 10$, hallar la constante C tal que (*a*) $\Pr\{U > C\} = 0.05$, (*b*) $\Pr\{-C \leq U \leq C\} = 0.98$, (*c*) $\Pr\{U \leq C\} = 0.20$ y (*d*) $\Pr\{U \geq C\} = 0.90$.

11.24. Los coeficientes de confianza 99% (con dos colas) para la distribución normal vienen dados por $\pm$ 2.58. ¿Cuáles son los correspondientes coeficientes para la distribución t de Student si (*a*) $\nu = 4$, (*b*) $\nu = 12$, (*c*) $\nu = 25$, (*d*) $\nu = 30$ y (*e*) $\nu = 40$?

11.25. Una muestra de 12 medidas de la tensión de ruptura de hilos de algodón da una media de 7.38 gramos (g) y una desviación típica de 1.24 g. Hallar los limites de confianza (*a*) 95% y (*b*) 99% para la verdadera tensión de ruptura.

11.26. Repetir el Problema 11.25 en el supuesto de que los métodos de grandes muestras fuesen aplicables, y comparar los resultados obtenidos.

11.27. Cinco medidas del tiempo de reacción de un individuo ante cierto estímulo se han registrado como 0.28, 0.30, 0.27, 0.33 y 0.31 segundos. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para el tiempo real de reacción.

11.28. La vida media de las lámparas producidas por una empresa era, en tiempos, de 1120 h con desviación típica de 125 h. Una muestra reciente de 8 lámparas da una vida media de 1070 h. Contrastar la hipótesis de que la vida media de esas lámparas no ha cambiado, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.29. En el Problema 11.28, contrastar la hipótesis $\mu = 1120$ h frente a la hipótesis alternativa $\mu < 1120$ h, usando nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.30. Las especificaciones para la fabricación de cierta aleación exigen un 23.2% de cobre. Una muestra de 10 análisis del producto ha revelado un contenido medio de cobre del 23.5% con desviación típica de 0.24%. ¿Podemos concluir que el producto cumple las especificaciones al nivel de significación (*a*) 0.01 y (*b*) 0.05?

11.31. En el Problema 11.30, contrastar la hipótesis de que el contenido medio de cobre es mayor de lo especificado, usando nivel de significación (*a*) 0.01 y (*b*) 0.05.

11.32. Un técnico sostiene que introduciendo un nuevo tipo de maquinaria en un proceso de producción se puede disminuir sustancialmente el tiempo requerido en la producción. A causa del alto costo de mantenimiento, el empresario piensa que salvo que se reduzca ese tiempo en al menos un 8%, no vale la pena tal inversión. Seis experiencias arrojan una disminución media del

tiempo de producción del 8.4% con desviación típica de 0.32%. Con nivel de significación (*a*) 0.01 y (*b*) 0.05, contrastar la hipótesis de que el proceso merece ser renovado.

11.33. Con gasolina de la marca *A*, el número medio de millas por galón que recorren 5 automóviles similares en igualdad de condiciones es 22.6 con desviación típica 0.48. Con gasolina de otra marca *B*, el resultado es 21.4 con desviación típica 0.54. Usando un nivel de significación 0.05, investigar si la marca *A* es de mejor calidad que la *B*.

11.34. Dos tipos de soluciones químicas, *A* y *B*, han sido probadas para ver su pH (grado de acidez de la solución). El análisis de 6 muestras de *A* arroja un pH medio 7.52 con desviación típica 0.024, mientras que 5 muestras de *B* dan un pH medio 7.49 con desviación típica 0.032. Usando el nivel de significación 0.05, determinar si los dos tipos de soluciones tienen distinto pH.

11.35. En un examen de psicología, 12 estudiantes de una clase obtuvieron media de 78 con desviación típica 6, y 15 de otra clase consiguieron media de 74 con desviación típica 8. Mediante un nivel de significación 0.05, determinar si el primer grupo es superior al segundo.

DISTRIBUCION JI-CUADRADO

11.36. Para una distribución ji-cuadrado con 12 grados de libertad, hallar el valor de χ_c^2 tal que (*a*) el área a la derecha de χ_c^2 es 0.05, (*b*) el área a la izquierda de χ_c^2 es 0.99 y (*c*) el área a la derecha de χ_c^2 es 0.025.

11.37. Hallar los valores críticos de χ^2 para los cuales el área de la cola derecha de la distribución ji-cuadrado es 0.05 si el número de grados de libertad, ν, es igual (*a*) 8, (*b*) 19, (*c*) 28 y (*d*) 40.

11.38. Repetir el Problema 11.37 si el área de la cola de la derecha es 0.01.

11.39. (*a*) Hallar χ_1^2 y χ_2^2 tales que el área bajo la distribución ji-cuadrado correspondiente a $\nu = 20$ entre χ_1^2 y χ_2^2 es 0.95, suponiendo áreas iguales a la derecha de χ_2^2 y a la izquierda de χ_1^2.

(*b*) Probar que si la suposición de áreas iguales en (*a*) se omite, los valores χ_1^2 y χ_2^2 no son únicos.

11.40. Si la variable U tiene una distribución ji-cuadrado con $\nu = 7$, hallar χ_1^2 y χ_2^2 tales que (*a*) $\Pr\{U > \chi_2^2\} = 0.025$, (*b*) $\Pr\{U < \chi_1^2\} = 0.50$, (*c*) $\Pr\{\chi_1^2 \leq U \leq \chi_2^2\} = 0.90$.

11.41. La desviación típica de las vidas medias de 10 bombillas es 120 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de las bombillas de esa clase.

11.42. Rehacer el Problema 11.41 si 25 bombillas diesen esa misma desviación típica de 120 h.

11.43. Hallar (*a*) $\chi_{.05}^2$ y (*b*) $\chi_{.95}^2$ para $\nu = 150$.

11.44. Hallar (*a*) $\chi_{.025}^2$ y (*b*) $\chi_{.975}^2$ para $\nu = 250$.

11.45. Probar que para grandes valores de ν, una buena aproximación de χ^2 viene dada por $(\nu + z_p\sqrt{2\nu})$, donde z_p es el p-ésimo percentil de la distribución normal canónica.

11.46. Resolver el Problema 11.39 usando la distribución ji-cuadrado si una muestra de 100 bombillas da la misma desviación típica de 120 h. Comparar los resultados con los obtenidos por los métodos del Capítulo 9.

11.47. ¿Cuál es el intervalo de confianza 95% del Problema 11.44 que tiene anchura mínima?

11.48. La desviación típica de las tensiones de ruptura de ciertos cables producidos por una empresa es 240 lb. Tras un cambio en el proceso de producción, una muestra de 8 cables dio una desviación típica de 300 lb. Investigar si es significativo ese crecimiento en variabilidad, usando nivel de significación (*a*) 0.05, y (*b*) 0.01.

11.49. La desviación típica de las temperaturas anuales en una ciudad a lo largo de 100 años es 16 °F. Usando la temperatura media del día 15 de cada mes durante los últimos 15 años, ha resultado una desviación

típica de 10 °F. Contrastar la hipótesis de que las temperaturas en esa ciudad son menos variables que en el pasado, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

DISTRIBUCION F

11.50. Hallar los valores de F en cada caso:

(*a*) $F_{.95}$ con $\nu_1 = 8$ y $\nu_2 = 10$
(*b*) $F_{.99}$ con $\nu_1 = 24$ y $\nu_2 = 11$
(*c*) $F_{.95}$ con $N_1 = 16$ y $N_2 = 25$
(*d*) $F_{.99}$ con $N_1 = 21$ y $N_2 = 23$

11.51. Calcular $F_{.95}$ con $\nu_1 = 22$ y $\nu_2 = 27$.

11.52. En dos poblaciones normalmente distribuidas con varianzas 40 y 60, se toman muestras respectivas de tamaño 10 y 15. Si las varianzas muestrales son 90 y 50, determinar si la muestra 1 tiene varianza significativamente mayor que la muestra 2, al nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.53. Dos empresas *A* y *B* producen lámparas eléctricas, cuyas vidas medias están muy normalmente distribuidas, con desviaciones típicas de 20 y 27 h, respectivamente. Si seleccionamos 16 lámparas de *A* y 20 de *B* y las desviaciones típicas de sus vidas medias resultan ser 15 y 40 h respectivamente, ¿podemos concluir a los niveles de significación (*a*) 0.05 y (*b*) 0.01 que la variabilidad de las de *A* es significativamente menor que la de las de *B*?

CAPITULO 12

Test ji-cuadrado

FRECUENCIAS OBSERVADAS Y TEORICAS

Como ya hemos visto repetidamente, los resultados obtenidos por muestreo no siempre coinciden exactamente con los esperados teóricamente de acuerdo con las leyes de las probabilidades. Por ejemplo, aunque consideraciones teóricas conducen a esperar 50 caras y 50 cruces en 100 tiradas de una moneda (buena), es raro que ocurra eso exactamente.

Supongamos que en una muestra particular un conjunto de sucesos posibles $E_1, E_2, E_3, \ldots, E_k$ (véase Tabla 12.1) se observa que ocurren con frecuencias $o_1, o_2, o_3, \ldots, o_k$, llamadas *frecuencias observadas*, y que según las leyes de las probabilidades, se espera que sucedan con frecuencias $e_1, e_2, e_3, \ldots, e_k$, llamadas *frecuencias esperadas* o *teóricas*.

Tabla 12.1

Suceso	E_1	E_2	E_3	$\cdots$	E_k
Frecuencia observada	o_1	o_2	o_3	$\cdots$	o_k
Frecuencia esperada	e_1	e_2	e_3	$\cdots$	e_k

A menudo deseamos saber si las frecuencias observadas difieren significativamente de las esperadas. Para el caso en que sólo son posibles dos sucesos E_1 y E_2 (llamado a veces una *dicotomía* o *clasificación dicotómica*), como es el caso de cara o cruz, piezas defectuosas o no, etc., el problema se resuelve satisfactoriamente por los métodos de los anteriores capítulos. En este capítulo consideramos el problema general.

DEFINICION DE χ^2

Una medida de la discrepancia existente entre las frecuencias observadas y esperadas viene proporcionada por el estadístico χ^2 (léase ji-cuadrado) dado por

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \qquad (1)$$

donde si la frecuencia total es N,

$$\sum o_j = \sum e_j = N \tag{2}$$

Una expresión equivalente a la fórmula (1) es (véase Prob. 12.11)

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{3}$$

Si $\chi^2 = 0$, las frecuencias observadas y teóricas coinciden completamente; mientras que si $\chi^2 > 0$, no coinciden exactamente. A valores más grandes de χ^2, mayor discrepancia entre las frecuencias observadas y esperadas.

La distribución muestral de χ^2 se aproxima muy bien por la distribución ji-cuadrado

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \tag{4}$$

(ya considerada en el Capítulo 11) si las frecuencias esperadas son al menos iguales a 5, y mejora para valores más grandes.

El número de grados de libertad, ν, viene dado por

1. $\nu = k - 1$ si las frecuencias esperadas se pueden calcular sin tener que estimar los parámetros de la población a partir de estadísticos muestrales. Nótese que hemos restado 1 de k a causa de la ligadura (2), que establece que si conocemos $k - 1$ de las frecuencias esperadas, la restante puede determinarse ya.
2. $\nu = k - 1 - m$ si las frecuencias esperadas se pueden calcular sólo estimando m parámetros de la población a partir de estadísticos de la muestra.

CONTRASTES DE SIGNIFICACION

En la práctica, las frecuencias esperadas se calculan sobre la base de una hipótesis H_0. Si bajo tal hipótesis el valor calculado para χ^2 dado por (1) o (3) es mayor que algún valor crítico (tal como $\chi^2_{.95}$ o $\chi^2_{.99}$, que son los valores críticos de los niveles de significación 0.05 y 0.01 respectivamente), debemos concluir que las frecuencias observadas difieren *significativamente* de las frecuencias esperadas y rechazaremos H_0 al correspondiente nivel de significación; en caso contrario, la aceptaremos (o al menos no la rechazaremos). Este procedimiento se llama el *test o contraste ji-cuadrado* de hipótesis o significación.

Hay que hacer constar que debe mirarse con suspicacia en circunstancias en las que χ^2 sea *demasiado próximo a cero*, pues es raro que las frecuencias observadas coincidan *demasiado bien* con las frecuencias esperadas. Para examinar tales situaciones, podemos determinar si el valor calculado de χ^2 es menor que $\chi^2_{.05}$ o $\chi^2_{.01}$, en cuyo caso hablaremos de decidir que el acuerdo es *demasiado bueno* al nivel de significación 0.05 ó 0.01, respectivamente.

EL TEST JI-CUADRADO PARA LA BONDAD DE AJUSTE

El test ji-cuadrado puede utilizarse para determinar la calidad del ajuste mediante distribuciones teóricas (como la distribución normal o la distribución binomial) de distribuciones empiricas (o sea, las obtenidas de los datos de la muestra). Véanse Problemas 12.12 y 12.13.

TABLAS DE CONTINGENCIA

La Tabla 12.1, en la que las frecuencias observadas ocupan una sola fila, se llama una *tabla de clasificación de entrada única*. Como el número de columnas es k, también se le llama una tabla $1 \times k$ (leído «1 por k»). Extendiendo estas ideas, podemos llegar a *tablas de doble entrada*, o *tablas* $h \times k$, en las que las frecuencias observadas ocupan h filas y k columnas. Tales tablas se suelen llamar *tablas de contingencia*.

Correspondiendo a cada frecuencia observada en una tabla de contingencia $h \times k$, hay una *frecuencia esperada* (o *teórica*) que se calcula sujeta a ciertas hipótesis de acuerdo con las leyes de las probabilidades. Estas frecuencias, que ocupan las celdas de una tabla de contingencia, se llaman *frecuencias de celda*. La frecuencia total en cada fila o en cada columna se llama la *frecuencia marginal*.

Para investigar el acuerdo entre las frecuencias observadas y las frecuencias esperadas, calculamos el estadístico

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \tag{5}$$

donde la suma se toma sobre todas las celdas de una tabla de contingencia y donde los símbolos o_j y e_j representan, respectivamente, las frecuencias observadas y frecuencias esperadas de la j-ésima celda. Esta suma, análoga a la ecuación (1), contiene hk términos. La suma de todas las frecuencias observadas se denota por N y es igual a la suma de todas las frecuencias esperadas [comparar con la ecuación (2)].

Como antes, el estadístico (5) tiene una distribución muestral dada muy aproximadamente por (4), supuesto que las frecuencias esperadas no sean demasiado pequeñas. El número de grados de libertad, ν, de esta distribución ji-cuadrado viene dado por $h > 1$ y $k > 1$ por:

1. $\nu = (h - 1)(k - 1)$ si las frecuencias esperadas se pueden calcular sin recurrir a estimaciones muestrales de los parámetros de la población. Para una demostración de esto, véase el Problema 12.18.
2. $\nu = (h - 1)(k - 1) - m$ si las frecuencias esperadas sólo se pueden calcular mediante estimación de m parámetros de la población a partir de estadísticos de la muestra.

Los contrastes de significación para las tablas $h \times k$ son similares a los de las tablas $1 \times k$. Las frecuencias esperadas se hallan sujetas a una hipótesis particular H_0. Una hipótesis común es suponer que las dos clasificaciones son mutuamente independientes.

Las tablas de contingencia se pueden generalizar a más dimensiones. Así, por ejemplo, podemos tener tablas $h \times k \times l$, donde están presentes tres clasificaciones.

CORRECCION DE YATES A LA CONTINUIDAD

Cuando se aplican resultados de distribuciones continuas a datos discretos, pueden hacerse ciertas correcciones a la continuidad, como se ha visto en capítulos precedentes. Una corrección similar existe cuando se usa la distribución ji-cuadrado. La corrección consiste en reformular la ecuación (1) como

$$\chi^2 \text{ (corregido)} = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} + \cdots + \frac{(|o_k - e_k| - 0.5)^2}{e_k} \tag{6}$$

y se llama *corrección de Yates*. Una modificación análoga existe para (5).

En general, la corrección se hace sólo cuando el número de grados de libertad es $\nu = 1$. Para grandes muestras, esto da prácticamente los mismos resultados que el χ^2 sin corregir, pero pueden surgir dificultades cerca de los valores críticos (véase Prob. 12.8). Para pequeñas muestras donde cada frecuencia esperada está entre 5 y 10, es quizás mejor comparar ambos valores de χ^2, corregido y sin corregir. Si ambos llevan a la misma conclusión acerca de la hipótesis, tal como el rechazo al nivel de significación 0.05, rara vez surgen dificultades. Si conducen a diferente conclusión, uno debe pensar en aumentar el tamaño de la muestra o, si ello no es factible, en emplear métodos de probabilidad que involucren la *distribución multinomial* del Capítulo 6.

FORMULAS SIMPLES PARA CALCULAR

Existen fórmulas sencillas para calcular χ^2 que implican tan sólo las frecuencias observadas. Lo que sigue da los resultados para tablas de contingencia 2×2 y 2×3 (véanse Tablas 12.2 y 12.3, respectivamente).

Tablas 2×2

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1N_2N_AN_B} \tag{7}$$

Tabla 12.2

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.3

	I	II	III	Total
A	a_1	a_2	a_3	N_A
B	b_1	b_2	b_3	N_B
Total	N_1	N_2	N_3	N

donde $\Delta = a_1b_2 - a_2b_1$, $N = a_1 + a_2 + b_1 + b_2$, $N_1 = a_1 + b_1$, $N_2 = a_2 + a_2$, $N_A = a_1 + a_2$, y $N_B = b_1 + b_2$ (véase Prob. 12.19). Con corrección de Yates esto se convierte en

$$\chi^2 \text{ (corregido)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N(|\Delta| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} \tag{8}$$

Tablas 2×3

$$\chi^2 = \frac{N}{N_A}\left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3}\right] + \frac{N}{N_B}\left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3}\right] - N \tag{9}$$

donde hemos usado el resultado general válido para todas las tablas de contingencia (véase Problema 12.43):

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{10}$$

El resultado (9) para tablas de contingencia $2 \times k$, con $k > 3$, admite generalización (véase Problema 12.46).

COEFICIENTE DE CONTINGENCIA

Una medida del grado de interrelación, asociación o dependencia de las clasificaciones en una tabla de contingencia viene dada por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{11}$$

que se llama el *coeficiente de contingencia*. Cuanto mayor es C, mayor es el grado de asociación. El número de filas y de columnas en la tabla de contingencia determina el máximo valor de C, que nunca es mayor que 1. Si el número de filas y columnas de una tabla de contingencia es igual a k, el máximo valor de C está dado por $\sqrt{(k-1)/k}$ (véanse Problemas 12.22, 12.52 y 12.53).

CORRELACION DE ATRIBUTOS

Ya que las clasificaciones en una tabla de contingencia describen a menudo características de individuos u objetos, se les conoce como *atributos*, y el grado de dependencia, asociación o interrelación se llama la *correlación de atributos*. Para tablas $k \times k$, definimos

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{12}$$

como el coeficiente de contingencia entre atributos (o clasificaciones). Este coeficiente está entre 0 y 1 (véase Prob. 12.24). Para tablas 2×2 en las que $k = 2$, la *correlación se llama tetracórica*.

El problema general de correlación de variables numéricas se considera en el Capítulo 14.

PROPIEDAD ADITIVA DE χ^2

Supongamos que los resultados de experimentos repetidos dan valores muestrales de χ^2 dados por $\chi_1^2, \chi_2^2, \chi_3^2, \ldots$ con $\nu_1, \nu_2, \nu_3, \ldots$ grados de libertad, respectivamente. Entonces el resultado de todos esos experimentos puede considerarse equivalente a un valor de χ^2 dado por $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots$ con $\nu_1 + \nu_2 + \nu_3 + \cdots$ grados de libertad (véase Prob. 12.25).

PROBLEMAS RESUELTOS

EL TEST JI-CUADRADO

12.1. En 200 tiradas de una moneda, han salido 115 caras y 85 cruces. Contrastar la hipótesis de que la moneda es buena, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Las frecuencias observadas de caras y cruces son $o_1 = 115$ y $o_2 = 85$, respectivamente, y las frecuencias esperadas (si la moneda es buena) son $e_1 = 100$ y $e_2 = 100$, respectivamente. Entonces

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Como el número de categorías, o clases (caras, cruces) es $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

(*a*) El valor crítico $\chi^2_{.95}$ para 1 grado de libertad es 3.84. Así pues, como $4.50 > 3.84$, rechazamos la hipótesis de que la moneda es buena al nivel de significación 0.05.

(*b*) El valor crítico $\chi^2_{.99}$ para 1 grado de libertad es 6.63. Así pues, como $4.50 < 6.63$, no podemos rechazar la hipótesis de que la moneda es buena al nivel de significación 0.01.

Concluimos que los resultados observados son *probablemente significativos* y que la moneda es *probablemente falsa*. Para comparar este método con los usados previamente, véase el Problema 12.3.

12.2. Rehacer el Problema 12.1 usando la corrección de Yates.

Solución

$$\chi^2(\text{corregido}) = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} = \frac{(|105 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100}$$

$$= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

Como $4.205 > 3.84$ y $4.205 < 6.63$, las conclusiones alcanzadas en el Problema 12.1 son válidas. Para comparar con métodos previos, ver el Problema 12.3.

12.3. Resolver el Problema 12.1 usando la aproximación normal a la distribución binomial.

Solución

Bajo la hipótesis de que la moneda es buena, la media y la desviación típica del número de caras esperadas en 200 tiradas son $\mu = Np = (200)(0.5) = 100$ y $\sigma = \sqrt{Npq} = \sqrt{(200)(0.5)(0.5)} = 7.07$, respectivamente.

Primer método

$$115 \text{ caras en unidades estándar} = \frac{115 - 100}{7.07} = 2.12$$

Usando el nivel de significación 0.05 y un contraste de dos colas, rechazaríamos la hipótesis de que la moneda es buena si z cae fuera de intervalo -1.96 a 1.96. Con nivel de significación 0.01, el

intervalo correspondiente sería de -2.58 a 2.58. Se sigue que (como en el Problema 12.1) podemos rechazarla al nivel 0.05 pero no al 0.01.

Nótese que el cuadrado del recuento estándar anterior $(2.12)^2 = 4.50$, es lo mismo que el valor de χ^2 obtenido en el Problema 12.1. Este es siempre el caso para un test ji-cuadrado que involucre dos categorías (véase Prob. 12.10).

Segundo método

Usando corrección de continuidad, 115 o más caras es equivalente a 114.5 o más caras. Pero 114.5 en unidades estándar $= (114.5 - 100)/7.07 = 2.05$. Eso lleva a las mismas conclusiones que el primer método.

Nótese que el cuadrado de ese valor estándar es $(2.05)^2 = 4.20$, que coincide con el valor de χ^2 corregido por continuidad con la corrección de Yates del Problema 12.2. Esto sucede siempre para un test ji-cuadrado que implique a dos categorías a las que se ha aplicado la corrección de Yates.

12.4. La Tabla 12.4 muestra las frecuencias observadas y las frecuencias esperadas al lanzar un dado 120 veces. Contrastar la hipótesis de que el dado es bueno, con un nivel de significación de 0.05.

Tabla 12.4

Cara del dado	1	2	3	4	5	6
Frecuencia observada	25	17	15	23	24	16
Frecuencia esperada	20	20	20	20	20	20

Solución

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6}$$

$$= \frac{25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00$$

Como el número de categorías, o clases (caras 1, 2, 3, 4, 5 y 6), es $k = 6$, $\nu = k - 1 = 6 - 1 = 5$. El valor crítico $\chi^2_{.95}$ para 5 grados de libertad es 11.1. Así que $5.00 < 11.1$ y no podemos rechazar la hipótesis de que el dado es bueno.

Para 5 grados de libertad, $\chi^2_{.05} = 1.15$, así que $\chi^2 = 5.00 > 1.15$. Se deduce que el acuerdo no es excepcionalmente bueno, y debemos mirarlo con recelo.

12.5. La Tabla 12.5 recoge la distribución de los dígitos 0, 1, 2, ..., 9 en una tabla de números aleatorios de 250 dígitos. ¿Difiere la distribución observada de la esperada de forma significativa?

Tabla 12.5

Dígito	0	1	2	3	4	5	6	7	8	9
Frecuencia observada	17	31	29	18	14	20	35	30	20	36
Frecuencia esperada	25	25	25	25	25	25	25	25	25	25

Solución

$$\chi^2 = \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \cdots + \frac{(36-25)^2}{25} = 23.3$$

El valor $\chi^2_{.99}$ para $v = k - 1 = 9$ grados de libertad es 21.7 y $23.3 > 21.7$. Por tanto, concluimos que la distribución observada difiere significativamente de la esperada al nivel de significación 0.01. Luego dicha tabla de números aleatorios merece cierto recelo.

12.6. En su experimento con guisantes. Gregor Mendel observó que 315 eran redondos y amarillos, 108 redondos y verdes, 101 rugosos y amarillos y 32 rugosos y verdes. De acuerdo con su teoría de la herencia, esos números debían estar en la proporción 9:3:3:1. ¿Hay alguna evidencia para dudar de su teoría al nivel de significación (*a*) 0.01 y (*b*) 0.05?

Solución

El número total de guisantes es $315 + 108 + 101 + 32 = 556$. Como los números esperados están en la proporción 9:3:3:1 (y $9 + 3 + 3 + 1 = 16$), esperaríamos

$\frac{9}{16}(556) = 312.75$ lisos y amarillos $\quad \frac{3}{16}(556) = 104.25$ rugosos y amarillos

$\frac{3}{16}(556) = 104.25$ lisos y verdes $\quad \frac{1}{16}(556) = 34.75$ rugosos y verdes

$$\text{Luego } \chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470$$

Como hay 4 categorías, $k = 4$ y el número de grados de libertad es $v = 4 - 1 = 3$.

(*a*) Para $v = 3$, $\chi^2_{.99} = 11.3$, y, por tanto, no podemos rechazar la teoría al nivel 0.01.
(*b*) Para $v = 3$, $\chi^2_{.95} = 7.81$, y, por tanto, no podemos rechazar al nivel 0.05.

Concluimos que teoría y experimentos están en buen acuerdo.

Nótese que para 3 grados de libertad, $\chi^2_{.05} = 0.352$ y $\chi^2 = 0.470 > 0.352$. Así pues, aunque el acuerdo es bueno, los resultados obtenidos están sujetos a un error de muestreo razonable.

12.7. Una urna contiene un gran número de fichas de 4 colores diferentes: rojo, naranja, amarillo y verde. Una muestra de 12 fichas ha dado 2 rojas, 5 naranjas, 4 amarillas y 1 verde. Contrastar la hipótesis de que la urna contiene iguales proporciones de los cuatro colores.

Solución

Bajo la hipótesis de proporciones idénticas, se esperarían 3 fichas de cada color. Como estos números esperados son menores que 5, la aproximación ji-cuadrado será errónea. Para evitar eso, combinamos categorías de modo que el número esperado en cada una sea al menos 5.

Si deseamos rechazar la hipótesis, debemos combinarlas de manera tal que la evidencia en contra de la hipótesis sea más nítida. Ello se logra en nuestro caso considerando las categorías «rojo o verde» y «naranja o amarillo», para las cuales la muestra daba 3 y 9 fichas, respectivamente. Como el número esperado en cada categoría bajo la hipótesis de proporciones iguales es 6, tenemos

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

Para $v = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Luego no podemos rechazarla al nivel de significación 0.05 (aunque sí al 0.01). Cabe concebir que los resultados observados pudieran ser fruto del azar, aunque haya igual proporción presente de cada color.

Otro método

Con corrección de Yates se obtiene

$$\chi^2 = \frac{(|3 - 6| - 0.5)^2}{6} + \frac{(|9 - 6| - 0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

que conduce a las mismas conclusiones que antes. Era de esperar, claro está, pues la corrección de Yates siempre *reduce* el valor de χ^2.

Hay que hacer notar que si se hubiera usado la aproximación χ^2 a pesar de que las frecuencias son demasiado pequeñas, se hubiera obtenido

$$\chi^2 = \frac{(2 - 3)^2}{3} + \frac{(5 - 3)^2}{3} + \frac{(4 - 3)^2}{3} + \frac{(1 - 3)^2}{3} = 3.33$$

Como para $\nu = 4 - 1 = 3$, $\chi^2_{.95} = 7{,}81$, llegaríamos a la misma conclusión de antes. Desgraciadamente, la aproximación χ^2 para pequeñas frecuencias es pobre; por tanto, cuando no sea aconsejable combinar frecuencias, debemos recurrir a los métodos exactos de probabilidad del Capítulo 6.

12.8. En 360 tiradas de un par de dados, han salido 74 sietes y 24 onces. Con nivel de significación 0.05, contrastar la hipótesis de que los dados son buenos.

Solución

Un par de dados puede caer de 36 formas. Un 7 ocurre de 6 formas y un 11 en 2 formas. Luego $\Pr\{\text{siete}\} = \frac{6}{36} = \frac{1}{6}$ y $\Pr\{\text{once}\} = \frac{2}{36} = \frac{1}{18}$. Por tanto, en 360 tiradas esperaríamos $360/6 = 60$ sietes y $360/18 = 20$ onces, de modo que

$$\chi^2 = \frac{(74 - 60)^2}{60} + \frac{(24 - 20)^2}{20} = 4.07$$

Para $\nu = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Luego, como $4.07 > 3.84$, estaríamos inclinados a rechazar la hipótesis de que los dados son buenos. Usando la corrección de Yates, sin embargo, encontramos

$$\chi^2(\text{corregido}) = \frac{(|74 - 60| - 0.5)^2}{60} + \frac{(|24 - 20| - 0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

Así que sobre la base del χ^2 corregido no podemos rechazarla al nivel de significación 0.05.

En general, para grandes muestras como las de este ejemplo, los resultados usando la corrección de Yates son más fiables. No obstante, como incluso los valores corregidos de χ^2 están tan cerca del valor crítico, dudamos en tomar decisiones en un sentido u otro. En tales casos es quizás mejor aumentar el tamaño de la muestra si estamos interesados especialmente en el nivel de significación 0.05 por alguna razón; de otro modo, podríamos rechazar la hipótesis a algún otro nivel (tal como 0.01) si ello es satisfactorio.

12.9. Un estudio sobre 320 familias con 5 hijos reveló la distribución de la Tabla 12.6. ¿Es consistente el resultado con la hipótesis de que los nacimientos de chicos y chicas son igualmente probables?

Tabla 12.6

Número de chicos y chicas	5 chicos 0 chicas	4 chicos 1 chica	3 chicos 2 chicas	2 chicos 3 chicas	1 chico 4 chicas	0 chicos 5 chicas	Total
Número de familias	18	56	110	88	40	8	320

Solución

Sea p = probabilidad de que nazca un chico y $q = 1 - p$ la de una chica. Entonces, las probabilidades de (5 chicos), (4 chicos y 1 chica), ... (5 chicas) vienen dadas por los términos del desarrollo del binomio

$$(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

Si $p = q = \frac{1}{2}$, tenemos

$$\Pr\{5 \text{ chicos y } 0 \text{ chicas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32} \qquad \Pr\{2 \text{ chicos y } 3 \text{ chicas}\} = 10(\tfrac{1}{2})^2(\tfrac{1}{2})^3 = \tfrac{10}{32}$$
$$\Pr\{4 \text{ chicos y } 1 \text{ chica}\} = 5(\tfrac{1}{2})^4(\tfrac{1}{2}) = \tfrac{5}{32} \qquad \Pr\{1 \text{ chico y } 4 \text{ chicas}\} = 5(\tfrac{1}{2})(\tfrac{1}{2})^4 = \tfrac{5}{32}$$
$$\Pr\{3 \text{ chicos y } 2 \text{ chicas}\} = 10(\tfrac{1}{2})^3(\tfrac{1}{2})^2 = \tfrac{10}{32} \qquad \Pr\{0 \text{ chicos y } 5 \text{ chicas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32}$$

Así que el número esperado de familias con 5, 4, 3, 2, 1 y 0 chicos se obtiene multiplicando las probabilidades anteriores por 320, y los resultados son 10, 50, 100, 100, 50 y 10, respectivamente. Por tanto,

$$\chi^2 = \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} = 12.0$$

Como $\chi^2_{.95} = 11.1$ y $\chi^2_{.99} = 15.1$ para $\nu = 6 - 1 = 5$ grados de libertad, podemos rechazar la hipótesis al nivel de significación 0.05 pero no al 0.01. Así pues, concluimos que los resultados son probablemente significativos, y los nacimientos de chicos y chicas no son equiprobables.

12.10. Probar que un test ji-cuadrado con sólo dos categorías es equivalente al contraste de significación para proporciones (o sea, el test 2) de la página 226.

Solución

Si P es la proporción muestral para la categoría I, p la proporción de la población y N la frecuencia total, podemos describir la situación por medio de la Tabla 12.7. Entonces, por definición,

$$\chi^2 = \frac{(NP - Np)^2}{Np} + \frac{[N(1-P) - N(1-p)]^2}{Nq} = \frac{N^2(P-p)^2}{Np} + \frac{N^2(P-p)^2}{Nq}$$
$$= N(P-p)^2\left(\frac{1}{p} + \frac{1}{q}\right) = \frac{N(P-p)^2}{pq} = \frac{(P-p)^2}{pq/N}$$

que es el cuadrado del estadístico z de la página 226.

Tabla 12.7

	I	II	Total
Frecuencia observada	NP	$N(1-P)$	N
Frecuencia esperada	Np	$N(1-p) = Nq$	N

12.11. (*a*) Probar que la fórmula (1) de este capítulo se puede escribir

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N$$

(*b*) Usar el resultado de la parte (*a*) para verificar el valor de χ^2 calculado en el Problema 12.6.

Solución

(*a*) Por definición,

$$\chi^2 = \sum \frac{(o_j - e_j)^2}{e_j} = \sum \left(\frac{o_j^2 - 2o_j e_j + e_j^2}{e_j} \right)$$

$$= \sum \frac{o_j^2}{e_j} - 2\sum o_j + \sum e_j = \sum \frac{o_j^2}{e_j} - 2N + N = \sum \frac{o_j^2}{e_j} - N$$

donde se ha usado la fórmula (2) de este capítulo.

(*b*) $$\chi^2 = \sum \frac{o_j^2}{e_j} - N = \frac{(315)^2}{312.75} + \frac{(108)^2}{104.25} + \frac{(101)^2}{104.25} + \frac{(32)^2}{34.75} - 556 = 0.470$$

BONDAD DEL AJUSTE

12.12. Usar el test ji-cuadrado para determinar la bondad del ajuste de los datos de la Tabla 7.4 del Problema 7.31.

Solución

$$\chi^2 = \frac{(38-33.2)^2}{33.2} + \frac{(144-161.9)^2}{161.9} + \frac{(342-316.2)^2}{316.2} + \frac{(287-308.7)^2}{308.7} + \frac{(164-150.7)^2}{150.7} + \frac{(25-29.4)^2}{29.4}$$
$$= 7.54$$

Como el número de parámetros utilizados en la estimación de las frecuencias esperadas es $m = 1$ (a saber, el parámetro p de la distribución binomial), $\nu = k - 1 - m = 6 - 1 - 1 = 4$.

Para $\nu = 4$, $\chi^2_{.95} = 9.49$. Así que el ajuste de los datos es bueno.

Para $\nu = 4$, $\chi^2_{.05} = 0.711$. Así pues, como $\chi^2 = 7.54 > 0.711$. El acuerdo no es tan extremadamente bueno como para ser increíble.

12.13. Determinar la bondad del ajuste de los datos en la Tabla 7.6 del Problema 7.33.

Solución

$$\chi^2 = \frac{(5 - 4.13)^2}{4.13} + \frac{(18 - 20.68)^2}{20.68} + \frac{(42 - 38.92)^2}{38.92} + \frac{(27 - 27.71)^2}{27.71} + \frac{(8 - 7.43)^2}{7.43} = 0.959$$

Como el número de parámetros utilizados en la estimación de las frecuencias esperadas es $m = 2$ (a saber la media μ y la desviación σ de la distribución normal), $\nu = k - 1 - m = 5 - 1 - 2 = 2$.

Para $\nu = 2$, $\chi^2_{.95} = 5.99$. Luego concluimos que el ajuste es muy bueno.

Para $\nu = 2$, $\chi^2_{.05} = 0.103$. Así pues, como $\chi^2 = 0.959 > 0.103$, el ajuste no es «demasiado bueno».

TABLA DE CONTINGENCIA

12.14. Resolver el Problema 10.20 usando el test ji-cuadrado.

Solución

Las condiciones del Problema se presentan en la Tabla 12.8(*a*). Bajo la hipótesis nula H_0 de que el suero no tiene efecto, esperaríamos 70 personas curadas en cada grupo, como indica la Tabla 12.8(*b*). Nótese que H_0 equivale a decir que la recuperación es independiente del uso del suero (o sea, las clasificaciones son independientes).

Tabla 12.8(*a*). Frecuencias observadas

	Curados	No curados	Total
Grupo A (usando suero)	75	25	100
Grupo B (sin suero)	65	35	100
Total	140	60	200

Tabla 12.8(*b*). Frecuencias esperadas bajo H_0

	Curados	No curados	Total
Grupo A (usando suero)	70	30	100
Grupo B (sin suero)	70	30	100
Total	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

Para determinar el número de grados de libertad, consideremos la Tabla 12.9, que es la misma que la 12.8 excepto que sólo muestra los totales. Es claro que somos libres de colocar sólo un número en cualquiera de las 4 celdas vacías, ya que una vez hecho eso los números en las restantes celdas vacías quedan fijados por los totales indicados. Luego hay 1 grado de libertad.

Tabla 12.9

	Curados	No curados	Total
Grupo A			100
Grupo B			100
Total	140	60	200

Otro método

Por la fórmula (véase Problema 12.18), $\nu = (h-1)(k-1) = (2-1)(2-1) = 1$. Como $\chi^2_{.95} = 3.84$ para 1 grado de libertad y como $\chi^2 = 2.38 < 3.84$, concluimos que los resultados *no son significativos* al nivel 0.05. Somos incapaces, en consecuencia, de rechazar H_0 a este nivel, y o bien concluimos que el suero no es efectivo o aplazamos la decisión, a la espera de más observaciones.

Nótese que $\chi^2 = 2.38$ es el cuadrado del z, $z = 1.54$, obtenido en el Problema 10.20. En general, el test ji-cuadrado que involucra proporciones muestrales en una tabla de contingencia 2×2 es equivalente a un contraste de significación de diferencias en proporciones usando la aproximación normal, como en la página 228. (Véase Prob. 12.20).

Hacemos notar también que un contraste de una cola usando χ^2 es equivalente a uno de dos colas usando χ ya que, por ejemplo, $\chi^2 > \chi^2_{.95}$ corresponde a $\chi > \chi_{.95}$ o $\chi < -\chi_{.95}$. Como para tablas de contingencia 2×2, χ^2 es el cuadrado de z, se sigue que χ es lo mismo que z para este caso. Así pues, un rechazo de la hipótesis al nivel 0.05 usando χ^2 equivale a un rechazo en un contraste de dos colas al nivel 0.10 usando z.

12.15. Repetir el Problema 12.14 haciendo la corrección de Yates.

Solución

$$\chi^2(\text{corregido}) = \frac{(|75-70|-0.5)^2}{70} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|35-30|-0.5)^2}{30} = 1.93$$

Luego las conclusiones del Problema 12.14 son válidas. Lo cual se podía haber visto de golpe recordando que la corrección de Yates siempre decrece el valor de χ^2.

12.16. La Tabla 12.10 muestra los números de estudiantes aprobados y suspendidos por tres profesores: Mr. X, Mr. Y y Mr. Z. Contrastar la hipótesis de que las proporciones de suspendidos por los tres profesores son iguales.

Tabla 12.10. Frecuencias observadas

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados	50	47	56	153
Suspensos	5	14	8	27
Total	55	61	64	180

Solución

Bajo la hipótesis H_0 de que las proporciones de estudiantes suspendidos por los tres profesores son iguales, hubieran suspendido 27/180 = 15% de los estudiantes y aprobado el 85%. En ese caso Mr. X, por ejemplo, hubiera suspendido al 15% de 55 estudiantes y hubiera aprobado al 85% de esos 55. Las frecuencias esperadas bajo H_0 se recogen en la Tabla 12.11. Tenemos pues

$$\chi^2 = \frac{(50-46.75)^2}{46.75} + \frac{(47-51.85)^2}{51.85} + \frac{(56-54.40)^2}{54.40} + \frac{(5-8.25)^2}{8.25} + \frac{(14-9.15)^2}{9.15} + \frac{(8-9.60)^2}{9.60} = 4.84$$

Para determinar el número de grados de libertad, consideremos la Tabla 12.12, que es la misma que las Tablas 12.10 y 12.11 excepto que sólo muestra los totales. Es claro que tenemos la libertad de

sólo un número en una celda vacía de la primera columna y uno en una celda vacía de la segunda o tercera columna, tras lo cual todos los demás números de las otras casillas quedan fijados unívocamente por los totales indicados. Luego hay 2 grados de libertad en este caso.

Tabla 12.11. Frecuencias esperadas bajo H_0

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados	88% de 55 = 46.75	85% de 61 = 51.85	85% de 64 = 54.40	153
Suspensos	15% de 55 = 8.25	15% de 61 = 9.15	15% de 64 = 9.60	27
Total	55	61	64	180

Tabla 12.12

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados				153
Suspensos				27
Total	55	61	64	180

Otro método

Por la fórmula, $\nu = (h - 1)(k - 1) = (2 - 1)(3 - 1) = 2$. Como $\chi^2_{.95} = 5.99$, no podemos rechazar H_0 al nivel 0.05. Nótese, no obstante, que como $\chi^2_{.90} = 4.61$, podemos rechazar H_0 al nivel 0.10 si estamos dispuestos a correr el riesgo de uno entre 10 de equivocarnos.

12.17. Usar la fórmula (9) de este capítulo para calcular el valor de χ^2 para el Problema 12.16.

Solución

Tenemos $a_1 = 50$, $a_2 = 47$, $a_3 = 56$, $b_1 = 5$, $b_2 = 14$, $b_3 = 8$, $N_A = a_1 + a_2 + a_3 = 153$, $N_B = b_1 + b_2 + b_3 = 27$, $N_1 = a_1 + b_1 = 55$, $N_2 = a_2 + b_2 = 61$, $N_3 = a_3 + b_3 = 64$ y $N = N_A + N_B = N_1 + N_2 + N_3 = 180$. Luego

$$\chi^2 = \frac{N}{N_A}\left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3}\right] + \frac{N}{N_B}\left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3}\right] - N$$

$$= \frac{180}{153}\left[\frac{(50)^2}{55} + \frac{(47)^2}{61} + \frac{(56)^2}{64}\right] + \frac{180}{27}\left[\frac{(5)^2}{55} + \frac{(14)^2}{61} + \frac{(8)^2}{64}\right] - 180 = 4.84$$

12.18. Probar que para una tabla de contingencia $h \times k$ el número de grados de libertad es $(h - 1) \times (k - 1)$, donde $h > 1$ y $k > 1$.

Solución

En una tabla con h filas y k columnas, podemos dejar de lado un número en cada columna, porque tales números se pueden recuperar por el conocimiento de los totales de filas y columnas. Se sigue que tenemos la libertad de colocar sólo $(h-1)(k-1)$ números en la tabla, ya que los demás se determinan univocamente. Luego el número de grados de libertad es $(h-1)(k-1)$. Este resultado vale si se conocen los parámetros de la población necesarios para obtener las frecuencias esperadas.

12.19. (*a*) Probar que para la tabla de contingencia recogida en la Tabla 12.13(*a*).

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

(*b*) Ilustrar el resultado de la parte (*a*) con los datos del Problema 12.14.

Tabla 12.13(*a*). Resultados observados

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.13(*b*). Resultados esperados

	I	II	Total
A	N_1N_A/N	N_2N_A/N	N_A
B	N_1N_B/N	N_2N_B/N	N_B
Total	N_1	N_2	N

Solución

(*a*) Como en el Problema 12.14, los resultados esperados bajo una hipótesis nula se muestran en la Tabla 12.13(*b*). Entonces

$$\chi^2 = \frac{(a_1 - N_1N_A/N)^2}{N_1N_A/N} + \frac{(a_2 - N_2N_A/N)^2}{N_2N_A/N} + \frac{(b_1 - N_1N_B/N)^2}{N_1N_B/N} + \frac{(b_2 - N_2N_B/N)^2}{N_2N_B/N}$$

Pero $$a_1 - \frac{N_1N_A}{N} = a_1 - \frac{(a_1+b_1)(a_1+a_2)}{a_1+b_1+a_2+b_2} = \frac{a_1b_2 - a_2b_1}{N}$$

Análogamente, $$a_2 - \frac{N_2N_A}{N} \quad \text{y} \quad b_1 - \frac{N_1N_B}{N} \quad \text{y} \quad b_2 - \frac{N_2N_B}{N}$$

son también iguales a $$\frac{a_1b_2 - a_2b_1}{N}$$

Así que podemos escribir

$$\chi^2 = \frac{N}{N_1N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_1N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2$$

que se simplifica a $$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

(*b*) En el Problema 12.14, $a_1 = 75$, $a_2 = 25$, $b_1 = 65$, $b_2 = 35$, $N_1 = 140$, $N_2 = 60$, $N_A = 100$, $N_B = 100$ y $N = 200$; entonces, como se ha obtenido antes.

$$\chi^2 = \frac{200[(75)(35) - (25)(65)]^2}{(140)(60)(100)(100)} = 2.38$$

Usando la corrección de Yates, el resultado es el mismo que en el Problema 12.15:

$$\chi^2 \text{ (corregido)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} = \frac{200[|(75)(35) - (25)(65)| - 100]^2}{(140)(60)(100)(100)} = 1.93$$

12.20. Probar que un test ji-cuadrado que implique a dos proporciones muestrales es equivalente a un contraste de significación de diferencias en proporciones mediante la aproximación normal (véase página 228).

Solución

Sean P_1 y P_2 dos proporciones muestrales, y sea p la proporción de la población. Con referencia al Problema 12.19, se tiene

$$P_1 = \frac{a_1}{N_1} \quad P_2 = \frac{a_2}{N_2} \quad 1 - P_1 = \frac{b_2}{N_1} \quad 1 - P_2 = \frac{b_2}{N_2} \tag{13}$$

y

$$p = \frac{N_A}{N} \quad 1 - p = q = \frac{N_B}{N} \tag{14}$$

Por tanto,

$$a_1 = N_1P_1 \quad a_2 = N_2P_2 \quad b_1 = N_1(1 - P_1) \quad b_2 = N_2(1 - P_2) \tag{15}$$

y

$$N_A = Np \quad N_B = N_q \tag{16}$$

Usando las ecuaciones (15) y (16), del Problema 12.19 deducimos

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B} = \frac{N[N_1P_1N_2(1 - P_2) - N_2P_2N_1(1 - P_1)]^2}{N_1N_2NpNq}$$

$$= \frac{N_1N_2(P_1 - P_2)^2}{Npq} = \frac{(P_1 - P_2)^2}{pq(1/N_1 + 1/N_2)} \quad (\text{porque } N = N_1 + N_2)$$

que es el cuadrado del estadístico z dado en la página 228.

TABLA DE CONTINGENCIA

12.21. Hallar el coeficiente de contingencia para los datos de la tabla de contingencia del Problema 12.14.

Solución

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

12.22. Hallar el máximo valor de C para la tabla 2 × 2 del Problema 12.14.

Solución

El máximo de C ocurre cuando las dos clasificaciones son perfectamente dependientes o asociadas. En tal caso, todos los que toman el suero se recuperan y todos los que no lo toman siguen enfermos. La tabla de contingencia aparece en la Tabla 12.14.

Tabla 12.14

	Curados	No curados	Total
Grupo A (usando suero)	100	0	100
Grupo B (sin suero)	0	100	100
Total	100	100	200

Como las frecuencias esperadas de celda, supuesta completa independencia, son todas 50,

$$\chi^2 = \frac{(100-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(100-50)^2}{50} = 200$$

Así que el máximo de C es $\sqrt{\chi^2/(\chi^2+N)} = \sqrt{200/(200+200)} = 0.7071$.

En general, para dependencia perfecta en una tabla de contingencia donde los números de filas y columnas son ambos k, las únicas frecuencias de celda no nulas se producen en la diagonal desde la esquina superior izquierda hasta la inferior derecha. Para tales casos, $C_{\text{máx}} = \sqrt{(k-1)/k}$. (Véase Problemas 12.52 y 12.53.)

CORRELACION DE ATRIBUTOS

12.23. Para la Tabla 12.8 del Problema 12.14, hallar el coeficiente de contingencia (*a*) sin y (*b*) con la corrección de Yates.

Solución

(*a*) Como $\chi^2 = 2.38$, $N = 200$, y $k = 2$, se tiene

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{2.38}{200}} = 0.1091$$

lo que indica poca correlación entre recuperación y uso del suero.

(*b*) Por el Problema 12.15, r (corregido) $= \sqrt{1.93/200} = 0.0982$.

12.24. Probar que el coeficiente de contingencia para tablas de contingencia, como se definió en la ecuación (12) de este capítulo, está entre 0 y 1.

Solución

Por el problema 12.53, el máximo valor de $\sqrt{\chi^2/(\chi^2+N)}$ es $\sqrt{(k-1/k}$. Luego

$$\frac{\chi^2}{\chi^2+N} \leqslant \frac{k-1}{k} \quad k\chi^2 \leqslant (k-1)(\chi^2+N) \quad k\chi^2 \leqslant k\chi^2 - \chi^2 + kN - N$$

$$\chi^2 \leqslant (k-1)N \quad \frac{\chi^2}{N(k-1)} \leqslant 1 \quad \text{y} \quad r = \sqrt{\frac{\chi^2}{N(k-1)}} \leqslant 1$$

Puesto que $\chi^2 \geqslant 0$, $r \geqslant 0$. Así que, $0 \leqslant r \leqslant 1$, como deseábamos probar.

PROPIEDAD ADITIVA DE χ^2

12.25. Para contrastar una hipótesis se ha realizado tres veces un experimento. Los valores resultantes de χ^2 son 2.37, 2.86 y 3.54, cada uno de los cuales corresponde a un grado de libertad. Probar que mientras H_0 no se puede rechazar al nivel 0.05 sobre la base de uno sólo de esos experimentos, sea cual sea, sí se puede rechazar cuando se combinan los tres.

Solución

Los valores de χ^2 obtenidos al combinar los tres experimentos es, de acuerdo con la *propiedad aditiva*, $\chi^2 = 2.37 + 2.86 + 3.54 = 8.77$ con $1 + 1 + 1 = 3$ grados de libertad. Como $\chi^2_{.95}$ para 3 grados de libertad es 7.81, podemos rechazar H_0 al nivel de significación 0.05. Pero como $\chi^2_{.95} = 3.84$ para 1 grado de libertad, no se puede rechazarla sobre la base de un solo experimento.

Al combinar experimentos en los que se obtienen valores de χ^2 correspondientes a 1 grado de libertad, la corrección de Yates se omite debido a que tiene tendencia a corregir en exceso.

PROBLEMAS SUPLEMENTARIOS

EL TEST JI-CUADRADO

12.26. En 60 lanzamientos de una moneda han salido 37 caras y 23 cruces. Usando nivel de significación (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que la moneda es buena.

12.27. Repetir el Problema 12.26 usando la corrección de Yates.

12.28. En un largo período de tiempo, los grados dados por un grupo de profesores en un curso particular han dado como promedio 12% Aes, 18% Bes, 40% Ces, 18% Des y 12% Efes. Un nuevo profesor da 22 Aes, 34 Bes, 66 Ces, 16 Des y 12 Efes en dos semestres. Determinar al nivel de significación 0.05 si el profesor nuevo sigue la norma de grados de los otros.

12.29. Se lanzan tres monedas 240 veces con el número de caras que recoge, junto con los resultados esperados bajo la hipótesis de que las monedas son buenas, la Tabla 12.15. Contrastar la hipótesis al nivel de significación.

Tabla 12.15

	Fr. observada	F. esperada
Caras 0	24	30
Caras 1	108	90
Caras 2	95	90
Caras 3	23	30

12.30. La Tabla 12.16 indica el número de libros prestados en una biblioteca pública durante una semana concreta. Contrastar la hipótesis de que el número de libros prestados no depende del día de la semana, usando nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 12.16

	N.º de libros prestados
Lunes	135
Martes	108
Miércoles	120
Jueves	114
Viernes	146

12.31. Una urna contiene 6 fichas rojas y 3 blancas. Se sacan dos al azar, se anotan sus colores y se devuelven a la urna. Este proceso se realiza 120 veces, y los resultados los presenta la Tabla 12.17.

(*a*) Calcular las frecuencias esperadas.
(*b*) Determinar al nivel de significación 0.05 si los resultados obtenidos son consistentes con los esperados.

Tabla 12.17

	Número de extracciones
0 Rojas 2 Blancas	6
1 Roja 1 Blanca	53
2 Rojas 0 Blancas	61

12.32. Se toman al azar 200 tuercas de las producidas por cada una de 4 máquinas. Las defectuosas encontradas fueron 2, 9, 10 y 3. Determinar si hay una diferencia significativa entre las máquinas, usando nivel de significación 0.05.

BONDAD DEL AJUSTE

12.33. (*a*) Usando el test ji-cuadrado, determinar la bondad del ajuste de los datos de la Tabla 7.9 del Problema 7.75. (*b*) ¿Es «demasiado bueno» el ajuste? Trabajar al nivel de significación 0.05.

12.34. Usar el test ji-cuadrado para juzgar la bondad del ajuste de los datos en (*a*) la Tabla 3.8 del Problema 3.59 y (*b*) la Tabla 3.10 del Problema 3.61. Usar un nivel de significación de 0.05 y determinar en cada caso si el ajuste es «demasiado bueno».

12.35. Usar el test ji-cuadrado para determinar la bondad del ajuste de los datos en (*a*) la Tabla 7.9 del Problema 7.79 y (*b*) la Tabla 7.10 del Problema 7.80. ¿Es consistente el resultado de la parte (*a*) con el del Problema 12.33?

TABLA DE CONTINGENCIA

12.36. La Tabla 12.18 recoge el resultado de un experimento para investigar el efecto de la vacunación de animales de laboratorio contra una cierta enfermedad. Con nivel de significación (*a*) 0.01 y (*b*) 0.05 contrastar la hipótesis de que no hay diferencia entre los grupos con y sin vacuna (o sea, que vacuna y enfermedad son independientes).

Tabla 12.18

	Enfermaron	No enfermaron
Vacunados	9	42
No vacunados	17	28

Tabla 12.19

	Aprobados	Suspensos
Clase *A*	72	17
Clase *B*	64	23

12.37. Rehacer el Problema 12.36 usando la corrección de Yates.

12.38. La Tabla 12.19 muestra el número de estudiantes en las clases *A* y *B* que aprobaron y suspendieron un examen propuesto a ambos grupos. Al nivel de significación (*a*) 0.05 y (*b*) 0.01 contrastar la hipótesis de que no hay diferencia entre las dos clases. Resolver el problema con y sin corrección de Yates.

12.39. A una parte de los pacientes con insomnio se les administró un tipo de píldoras inductoras del sueño y a los demás píldoras de azúcar (aunque ellos *creían* tomar un somnífero). Se les preguntó más tarde si las píldoras hacían efecto, con las respuestas que contiene la Tabla 12.20. Supuesto que los pacientes contestaron con sinceridad, contrastar la hipótesis de que no hay diferencia entre ambos tipos de píldoras al nivel de significación 0.05.

Tabla 12.20

	Durmieron bien	No durmieron bien
Tomaron píldoras somníferas	44	10
Tomaron píldoras inocuas	81	35

12.40. Ante una propuesta de política exterior, demócratas y republicanos adjudicaron sus votos como muestra la Tabla 12.21. Al nivel de significación (*a*) 0.01 y (*b*) 0.05, contrastar la hipótesis de que no hay diferencia entre los dos partidos en lo que a dicha propuesta se refiere.

Tabla 12.21

	Demócratas	Republicanos
A favor	85	118
En contra	78	61
Indecisos	37	25

12.41. La Tabla 12.22 presenta la relación entre las notas de estudiantes en matemáticas y física. Contrastar la hipótesis de que ambas son independientes, usando nivel de significación (*a*) 0.05 y (*b*) 0.01

Tabla 12.22

Física	Matemáticas		
	Calific. altas	Calific. bajas	Calific. medias
Calific. altas	56	71	12
Calific. medias	47	163	38
Calific. bajas	14	42	85

12.42. La Tabla 12.23 recoge los resultados de un estudio sobre si la edad de los conductores, de 21 años o más, afecta al número de accidentes que sufren (incluidos pequeños percances). Al nivel de significación (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que el número de accidentes es independiente de la edad del conductor. ¿Qué posibles dificultades en las técnicas de muestreo, o qué otras consideraciones, podrían afectar a las conclusiones?

Tabla 12.23

Edad del conductor	Número de accidentes			
	0	1	2	>2
21-30	748	74	31	9
31-40	821	60	25	10
41-50	786	51	22	6
51-60	720	66	16	5
61-70	672	50	15	7

12.43. (*a*) Probar que $\chi^2 = \sum(o_j^2/e_j) - N$ para todas las tablas de contingencia, donde N es la frecuencia total de todas las celdas.

(*b*) Usando el resultado de la parte (*a*), resolver el Problema 12.41.

12.44. Si N_i y N_j denotan, respectivamente, la suma de frecuencias de la i-ésima fila y de la j-ésima columna de una tabla de contingencia (las *frecuencias marginales*), probar que la frecuencia esperada para la celda que está en la i-ésima fila y en la j-ésima columna es N_iN_j/N, donde N es la frecuencia total de todas las celdas.

12.45. Demostrar la fórmula (9) de este capítulo. (*Ayuda:* Usar los Problemas 12.43 y 12.44.)

12.46. Extender el resultado de la fórmula (9) a las tablas de contingencia $2 \times k$, con $k > 3$.

12.47. Probar la fórmula (8) de este capítulo.

12.48. Por analogía con las ideas desarrolladas para tablas de contingencia $h \times k$, discutir las tablas de contingencia $h \times k \times l$ citando sus posibles aplicaciones.

COEFICIENTE DE CONTINGENCIA

12.49. La Tabla 12.24 presenta la relación entre el color del pelo y el de los ojos en una muestra de 200 estudiantes.

(*a*) Hallar el coeficiente de contingencia sin y con corrección de Yates.

(*b*) Comparar el resultado de (*a*) con el coeficiente de contingencia máximo.

Tabla 12.24

Color de los ojos	Color del cabello	
	Rubio	No rubio
Azul	49	25
No azul	30	96

12.50. Hallar el coeficiente de contingencia para los datos de (*a*) el Problema 12.36 y (*b*) el Problema 12.38, sin y con corrección de Yates.

12.51. Hallar el coeficiente de contingencia para los datos del Problema 12.41.

12.52. Probar que el coeficiente de contingencia máximo para una tabla de contingencia 3×3 es $\sqrt{\frac{2}{3}} = 0.8165$ aproximadamente.

12.53. Probar que el coeficiente de contingencia máximo de una tabla de contingencia $k \times k$ es $\sqrt{(k-1)/k}$.

CORRELACION DE ATRIBUTOS

12.54. Hallar el coeficiente de correlación para los datos de la Tabla 12.24.

12.55. Hallar el coeficiente de correlación para los datos de la (*a*) Tabla 12.18 y (*b*) Tabla 12.19 sin y con corrección de Yates.

12.56. Hallar el coeficiente de correlación entre las notas de matemáticas y física de la Tabla 12.22.

12.57. Si C es el coeficiente de contingencia para una tabla de contingencia $k \times k$ y r es el correspondiente coeficiente de correlación, probar que $r = C/\sqrt{(1 - C^2)(k - 1)}$.

PROPIEDAD ADITIVA DE χ^2

12.58. Para contrastar una hipótesis, se ha realizado cinco veces un experimento. Los valores resultantes de χ^2, cada uno correspondiendo a 4 grados de libertad, son 8.3, 9.1, 8.9, 7.8 y 8.6, respectivamente. Probar que mientras H_0 no puede ser rechazada al nivel 0.05 sobre la base de cada experimento por separado, puede rechazarse al nivel 0.005 atendiendo al resultado combinado de los cuatro experimentos.

CAPITULO 13

Ajuste de curvas y el método de mínimos cuadrados

RELACIONES ENTRE VARIABLES

En la práctica encontramos a menudo que existen relaciones entre dos (o más) variables. Por ejemplo, los pesos de las personas dependen en cierta medida de sus alturas, las circunferencias de los círculos dependen de los radios, y la presión de una masa de gas dada depende de su volumen y de su temperatura.

Suele ser deseable expresar tales relaciones en forma matemática determinando una ecuación que conecte a las variables.

AJUSTE DE CURVAS

Para hallar una ecuación que relacione las variables, el primer paso es recoger datos que muestren valores correspondientes de las variables bajo consideración. Así por ejemplo, supongamos que X e Y denotan, respectivamente, la altura y el peso de personas adultas; entonces una muestra de N individuos revelaría las alturas $X_1, X_2, ..., X_N$ y los pesos correspondientes $Y_1, Y_2, ..., Y_N$.

El próximo paso es marcar los puntos (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) sobre un sistema de coordenadas rectangulares. El conjunto de puntos resultante se llama a veces un *diagrama de dispersión*.

A partir del diagrama de dispersión es posible, con frecuencia visualizar una curva suave que aproxima los datos. Tal curva se llama una *curva aproximante*. En la Figura 13.1, por ejemplo, los datos parecen aproximarse bien a una línea recta, y decimos que hay una *relación lineal* entre las variables. En la Figura 13.2, sin embargo, aunque existe una relación entre las variables, no es lineal, y se dice que es una *relación no lineal*.

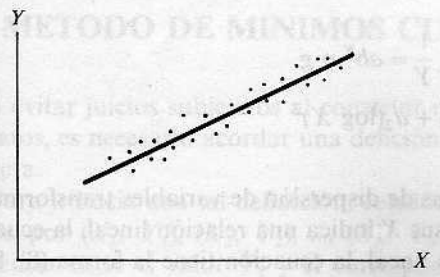


Figura 13.1.

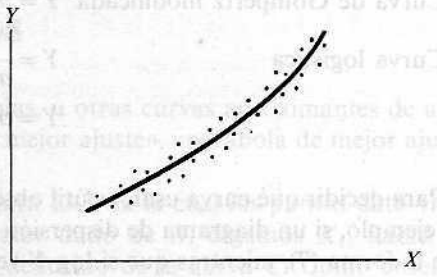


Figura 13.2.

El problema general de hallar ecuaciones de curvas aproximantes que se ajusten a un conjunto de datos se llama *ajuste de curvas*.

ECUACIONES DE CURVAS APROXIMANTES

Varios tipos comunes de curvas aproximantes y sus ecuaciones se presentan en la lista adjunta para facilitar posteriores referencias. Todas las letras excepto X e Y representan constantes. Las variables X e Y se llaman *variable independiente y dependiente*, respectivamente, aunque estos papeles se pueden intercambiar.

Línea recta	$Y = a_0 + a_1X$	(1)
Parábola, o curva cuadrática	$Y = a_0 + a_1X + a_2X^2$	(2)
Curva cúbica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3$	(3)
Curva cuártica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4$	(4)
Curva de grado n	$Y = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$	(5)

Los lados derechos de las ecuaciones anteriores se llaman *polinomios* de grado uno, dos, tres, cuatro y n, respectivamente. Las funciones definidas por las cuatro primeras ecuaciones se llaman a veces funciones *lineal*, *cuadrática*, *cúbica* y *cuártica*, respectivamente.

He aquí algunas otras de las muchas ecuaciones que se utilizan frecuentemente en la práctica:

Hipérbola	$Y = \dfrac{1}{a_0 + a_1X}$ ó $\dfrac{1}{Y} = a_0 + a_1X$	(6)
Curva exponencial	$Y = ab^X$ ó $\log Y = \log a + (\log b)X = a_0 + a_1X$	(7)
Curva geométrica	$Y = aX^b$ ó $\log Y = \log a + b(\log X)$	(8)
Curva exponencial modificada	$Y = ab^X + g$	(9)
Curva geométrica modificada	$Y = aX^b + g$	(10)
Curva de Gompertz	$Y = pq^{b^X}$ ó $\log Y = \log p + b^X(\log q) = ab^X + g$	(11)
Curva de Gompertz modificada	$Y = pq^{b^X} + h$	(12)
Curva logística	$Y = \dfrac{1}{ab^X + g}$ ó $\dfrac{1}{Y} = ab^X + g$	(13)
	$Y = a_0 + a_1(\log X) + a_2(\log X)^2$	(14)

Para decidir qué curva usar, es útil obtener diagramas de dispersión de variables transformadas. Por ejemplo, si un diagrama de dispersión de $\log Y$ versus X indica una relación lineal, la ecuación tiene la forma (7), mientras que si $\log Y$ versus $\log X$ es lineal, la ecuación tiene la forma (8). Suele usarse papel gráfico especial para facilitar la decisión sobre qué curva usar. El *papel gráfico* que tiene sólo una escala calibrada logarítmicamente se llama *semilogarítmico* (o *semilog*), y el que tiene las dos escalas logarítmicas se llama *papel log-log*.

AJUSTE DE CURVAS A MANO

A menudo puede recurrirse a la intuición personal a la hora de dibujar una curva que ajuste un conjunto de datos. Esto se conoce como *método de ajuste de curvas a mano*. Si el tipo de ecuación de esa curva es conocido, es posible obtener las constantes de la ecuación eligiendo tantos puntos de la curva como constantes haya en la ecuación. Por ejemplo, si la curva es una recta, son necesarios dos puntos; si es una parábola, son precisos tres puntos. El método tiene la desventaja de que diferentes observadores obtendrán distintas curvas y ecuaciones.

LA RECTA

El tipo más sencillo de curva aproximante es una línea recta, cuya ecuación puede escribirse

$$Y = a_0 + a_1 X \tag{15}$$

Dados cualesquiera dos puntos (X_1, Y_1) y (X_2, Y_2) sobre la recta, se pueden determinar las contantes a_0 y a_1. La ecuación así obtenida se puede expresar

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \quad \text{o sea} \quad Y - Y_1 = m(X - X_1) \tag{16}$$

donde

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

se llama la *pendiente* de la recta y representa el cambio en Y dividido por el correspondiente cambio en X.

Cuando la ecuación se escribe en la forma (15), la constante a_1 es la pendiente m. La constante a_0, que es el valor de Y cuando $X = 0$, se llama la *Y-intersección*.

EL METODO DE MINIMOS CUADRADOS

Para evitar juicios subjetivos al construir rectas, parábolas, u otras curvas aproximantes de ajuste de datos, es necesario acordar una defición de «recta de mejor ajuste», «parábola de mejor ajuste», etcétera.

Para ir hacia una tal definición, consideremos la Figura 13.3, en la cual los puntos dato vienen dados por (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N). Para un valor dado de X, digamos X_1, habrá una diferencia entre el valor Y_1 y el correspondiente valor deducido de la curva C. Como enseña la figura, denotamos esta diferencia por D_1, que se llama a veces *desviación*, *error* o *residual*, y puede ser positiva, negativa o nula. Análogamente, asociadas a los datos X_2, ..., X_N se obtienen desviaciones D_2, ..., D_N.

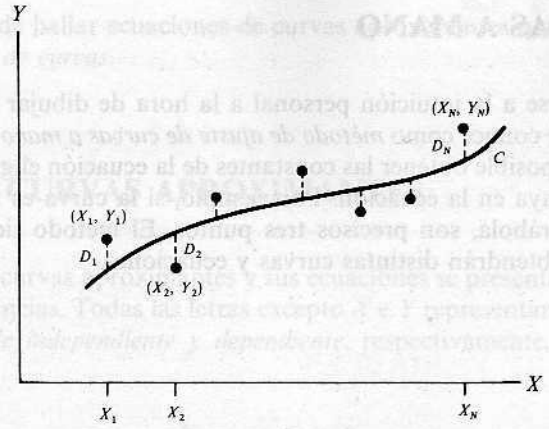


Figura 13.3.

Una medida de la «bondad del ajuste» de la curva C a los datos dados viene proporcionada por la cantidad $D_1^2 + D_2^2 + \cdots + D_N^2$. Si es pequeña, el ajuste es bueno; si es grande, el ajuste es malo. Hacemos, por tanto, la siguiente

Definición. De todas las curvas que aproximan un conjunto dado de datos, la que tiene la propiedad de que $D_1^2 + D_2^2 + \cdots + D_N^2$ es mínimo se llama una *curva de ajuste óptimo*.

Una tal curva se dice que ajusta los datos en el sentido de *mínimos cuadrados* y se llama una *curva de mínimos cuadrados*. Así pues, una recta con esa propiedad se llama *recta de mínimos cuadrados*, una parábola con esa propiedad se llama *parábola de mínimos cuadrados*, etc.

Es habitual emplear la definición precedente cuando X es la variable independiente e Y la dependiente. Si la variable dependiente es X, la definición se modifica considerando desviaciones horizontales en lugar de verticales, lo que viene a ser como intercambiar los ejes X e Y. Estas dos definiciones conducen, en general, a curvas distintas de mínimos cuadrados. Salvo que se especifique lo contrario, consideraremos a Y como la variable dependiente y a X como la independiente.

Es posible definir otras curvas de mínimos cuadrados considerando distancias perpendiculares desde cada uno de los puntos a la curva, en vez de distancias verticales u horizontales, pero no son de uso común.

LA RECTA DE MINIMOS CUADRADOS

La recta de mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ tiene por ecuación

$$Y = a_0 + a_1 X \tag{17}$$

donde las constantes a_0 y a_1 quedan fijadas al resolver simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{18}$$

que se llaman las *ecuaciones normales para la recta de mínimos cuadrados* (17). Las constantes a_0 y a_1 de las ecuaciones (18) se pueden hallar, si se desea, de las fórmulas

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \tag{19}$$

Las ecuaciones normales (18) son fáciles de recordar sin más que observar que la primera se obtiene formalmente sumando en ambos lados de (17) [o sea, $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$], mientras la segunda se obtiene formalmente multiplicando primero ambos lados de (17) por X y sumando después [o sea, $\sum XY = \sum X(a_0 + a_1 X) = a_0 \sum X + a_1 \sum X^2$]. Nótese que esto no es una deducción de las ecuaciones normales, sino sólo una forma de recordarlas. Nótese además que en las ecuaciones (18) y (19) hemos usado la notación abreviada $\sum X$, $\sum XY$, etc., en lugar de $\sum_{j=1}^{N} X_j$, $\sum_{j=1}^{N} X_j Y_j$, etc.

El trabajo requerido para hallar una recta de mínimos cuadrados se puede aliviar en ocasiones transformando los datos de manera que $x = X - \bar{X}$ e $y = Y - \bar{Y}$. La ecuación de la recta de mínimos cuadrados se puede escribir entonces (véase Prob. 13.15).

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x \tag{20}$$

En particular, si X es tal que $\sum X = 0$ (es decir, $\bar{X} = 0$), esto se convierte en

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X \tag{21}$$

La ecuación (20) implica que $y = 0$ cuando $x = 0$; así que la recta de mínimos cuadrados pasa por el punto $(\bar{X}, \bar{Y})$, llamado *centroide* o *centro de gravedad*, de los datos.

Si se toma X como variable dependiente, escribimos (17) como $X = b_0 + b_1 Y$. Entonces los resultados anteriores son válidos si se intercambian X e Y, y se sustituyen a_0 y a_1 por b_0 y b_1, respectivamente. La recta de mínimos cuadrados resultante, sin embargo, no es generalmente la misma que la obtenida antes [véanse Probs. 13.11 y 13.15(*d*)].

RELACIONES NO LINEALES

Las relaciones no lineales pueden reducirse en ocasiones a relaciones lineales por un apropiado cambio de variables (véase Prob. 13.21).

LA PARABOLA DE MINIMOS CUADRADOS

La parábola de mínimos cuadrados que aproxima el conjunto de puntos (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) tiene ecuación dada por

$$Y = a_0 + a_1 X + a_2 X^2 \tag{22}$$

donde las constantes a_0, a_1 y a_2 se determinan al resolver simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X_2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{23}$$

llamadas *ecuaciones normales de la parábola de mínimos cuadrados* (22).

Las ecuaciones (23) se recuerdan fácilmente observando que se pueden obtener formalmente multiplicando (22) por 1, X y X^2, respectivamente, y sumando en ambos lados de las ecuaciones resultantes. Esta técnica puede extenderse para obtener ecuaciones normales para curvas cúbicas de mínimos cuadrados curvas cuárticas de mínimos cuadrados, y en general cualquiera de las curvas de mínimos cuadrados correspondientes a la ecuación (5).

Como en el caso de la recta de mínimos cuadra dos, las ecuaciones (23) se simplifican si se elige X de modo $\sum X = 0$. También se produce simplificación tomando como nuevas variables $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

REGRESION

A menudo deseamos estimar, basados en datos de una muestra, el valor de una variable Y correspondiente a un valor dado de la variable X. Ello se puede hacer estimando el valor de Y mediante una curva de mínimos cuadrados que ajuste los datos. La curva resultante se llama una *curva de regresión de Y sobre X*, ya que Y se estima a partir de X.

Si queremos estimar el valor de X a paritr de un valor dado de Y, hemos de usar una *curva de regresión de X sobre Y*, que viene a ser un intercambio de las variables en el diagrama de dispersión de modo que X sea la variable dependiente e Y la independiente. Eso equivale a sustituir las desviaciones verticales en la definición de la curva de mínimos cuadrados en la página 291 por desviaciones horizontales.

En general, la recta o curva de regresión de Y sobre X no es la misma que la de X sobre Y.

APLICACIONES A SERIES EN EL TIEMPO

Si la variable independiente X es el tiempo, los datos muestran los valores de Y en varios instantes. Datos ordenados en el tiempo se llaman *series en el tiempo*. La recta o curva de regresión de Y sobre X en este caso se suele llamar una *recta o curva de tendencia*, y se utilizan en estimación y predicción.

PROBLEMAS EN MAS DE DOS VARIABLES

Los problemas que involucran a más de dos variables pueden tratarse de manera análoga a los de dos variables. Por ejemplo, puede haber una relación entre tres variables X, Y y Z descrita por la ecuación

$$Z = a_0 + a_1 X + a_2 Y \tag{24}$$

que se llama una *ecuación lineal en las variables X, Y y Z.*

En un sistema de coordenadas rectangulares tridimensional esa ecuación representa un plano, y los puntos (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ..., (X_N, Y_N, Z_N) de la muestra pueden «dispersarse» no lejos de ese plano, que se llama un *plano aproximante*.

Por extensión del método de mínimos cuadrados, podemos hablar de un *plano de mínimos cuadrados* que aproxima los datos. Si estamos estimando Z a partir de valores de X e Y, se le llama un *plano de regresión de Z sobre X e Y*. Las ecuaciones normales correspondientes al plano de mínimos cuadrados (24) vienen dadas por

$$\begin{aligned} \sum Z &= a_0 N + a_1 \sum X + a_2 \sum Y \\ \sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\ \sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2 \end{aligned} \tag{25}$$

y se pueden memorizar como obtenidas de (24) multiplicándola por 1, X, Y sucesivamente, y sumando después.

Cabe considerar también ecuaciones más complicadas que (24), que representan *superficies de regresión*. Si el número de variables es mayor que tres, se pierde la intuición geométrica ya que se requieren espacios de 4, 5, ... dimensiones.

Los problemas de estimación de una variable a partir de dos o más variables se llaman problemas de *regresión múltiple* y se considerarán con más detalle en el Capítulo 15.

PROBLEMAS RESUELTOS

RECTAS

13.1. (*a*) Construir una recta que aproxime los datos de la Tabla 13.1.
(*b*) Hallar una ecuación para esa recta.

Tabla 13.1

X	2	3	5	7	9	10
Y	1	3	7	11	15	17

Solución

(*a*) Marcar los puntos (2, 1), (3, 3), (5, 7), (7, 11), (9, 15) y (10, 17) en un sistema rectangular de coordenadas, como indica la Figura 13.4. Es claro de esa figura que todos los puntos están en una recta (dibujada a trazos); así que una recta ajusta esos datos *exactamente*.

(*b*) Para hallar la ecuación de la recta dada por

$$Y = a_0 + a_1 X \tag{26}$$

sólo se necesitan dos puntos. Escogemos los puntos (2, 1) y (3, 3), por ejemplo. Para el punto (2, 1), $X = 2$ y $Y = 1$; sustituyendo esos valores en (26) se ve que

$$1 = a_0 + 2a_1 \tag{27}$$

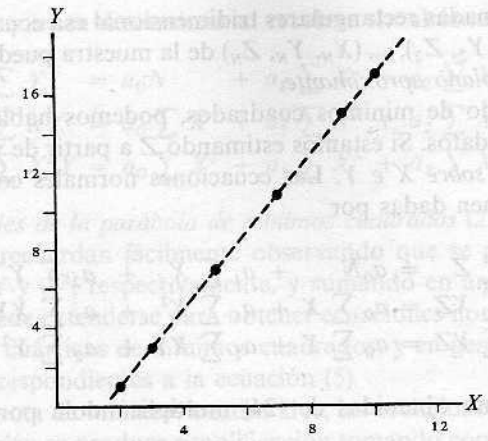


Figura 13.4.

Análogamente, para los puntos (3, 3), $X = 3$ e $Y = 3$; sustituyendo esos valores en (26) se obtiene

$$3 = a_0 + 3a_1 \tag{28}$$

Resolviendo (27) y (28) simultáneamente, $a_0 = -3$ y $a_1 = 2$, y la requerida ecuación es

$$Y = -3 + 2X \quad \text{o sea} \quad Y = 2X - 3$$

Como comprobación, véase que los puntos (5, 7), (7, 11), (9, 15) y (10, 17) están también sobre esa recta.

13.2. En el Problema 13.1 hallar (*a*) Y cuando $X = 4$, (*b*) Y cuando $X = 15$, (*c*) Y cuando $X = 0$, (*d*) X cuando $Y = 7.5$, (*e*) X cuando $Y = 0$ y (*f*) el crecimiento en Y correspondiente a un crecimiento unidad en X.

Solución

Suponemos que para otros valores de X e Y distintos de los especificados en la Tabla 13.1 es válida la misma relación $Y = 2X - 3$.

(*a*) Si $X = 4$, $Y = 2(4) - 3 = 8 - 3 = 5$. Como estamos hallando el valor de Y correspondiente a un valor de X incluido *entre* dos valores dados de X, este proceso se llama *interpolación lineal.*

(*b*) Si $X = 15$, $Y = 2(15) - 3 = 30 - 3 = 27$. Como estamos hallando el valor de Y correspondiente a un valor de X *exterior* a los valores dados de X, este proceso se llama *extrapolación lineal.*

(*c*) Si $X = 0$, $Y = 2(0) - 3 = 0 - 3 = -3$. El valor de Y cuando $X = 0$ se llama *Y-intersección.* Es el valor de Y en el punto donde la recta (extendida si es preciso) corta al eje Y.

(*d*) Si $Y = 7.5$, $7.5 = 2X - 3$; entonces $2X = 7.5 + 3 = 10.5$ y $X = 10.5/2 = 5.25$.

(*e*) Si $Y = 0$, $0 = 2X - 3$; entonces $2X = 3$ y $X = 1{,}5$. El valor de X cuando $Y = 0$ se llama X-intersección. Es el valor de X en el punto donde la recta (extendida si es preciso) corta al eje X.

(*f*) Si X crece una unidad de 2 a 3, Y crece de 1 a 3, un cambio de dos unidades. Si X crece de 2 a 10, o sea $(10 - 2) = 8$ unidades, Y crece de 1 a 17, un cambio de $(17 - 1) = 16$ unidades; esto es, Y crece 2 unidades por cada unidad que crece X.

En general, si ΔY denota el cambio en Y debido a un cambio en X de ΔX entonces el cambio en Y por unidad de cambio en X viene dado por $\Delta Y/\Delta X = 2$. Esto se llama la pendiente de la

recta y es siempre igual a a_1 en la ecuación $Y = a_0 + a_1X$. La constante a_0 es la *Y-intersección* de la recta [véase parte (*c*)].

Las cuestiones anteriores se pueden contestar también directamente del gráfico, Figura 13.4.

13.3. (*a*) Probar que la ecuación de una recta que pasa por los puntos (X_1, Y_1) y (X_2, Y_2) viene dada por

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

(*b*) Hallar la ecuación de una recta que pasa por los puntos $(2, -3)$ y $(4, 5)$.

Solución

(*a*) La ecuación de la recta es

$$Y = a_0 + a_1X \tag{29}$$

Como (X_1, Y_1) está en la recta,

$$Y_1 = a_0 + a_1X_1 \tag{30}$$

Como (X_2, Y_2) está en la recta,

$$Y_2 = a_0 + a_1X_2 \tag{31}$$

Restando la ecuación (30) de (29),

$$Y - Y_1 = a_1(X - X_1) \tag{32}$$

Restando la ecuación (30) de (31),

$$Y_2 - Y_1 = a_1(X_2 - X_1) \qquad \text{o sea} \qquad a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

Sustituyendo este valor de a_1 en la ecuación (32), obtenemos

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

como se pedía. La cantidad

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

abreviada usualmente como m, representa el cambio en Y dividido por el correspondiente cambio en X y es la pendiente de la recta. La ecuación pedida puede escribirse $Y - Y_1 = m(X - X_1)$.

(*b*) *Primer método* [usando el resultado de la parte (*a*)]

Correspondiendo al primer punto $(2, -3)$, tenemos $X_1 = 2$ e $Y_1 = -3$; para el segundo, $(4, 5)$, tenemos $X_2 = 4$ e $Y_2 = 5$. Luego la pendiente es

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{5 - (-3)}{4 - 2} = \frac{8}{2} = 4$$

y la ecuación pedida es

$$Y - Y_1 = m(X - X_1) \qquad \text{o sea} \qquad Y - (-3) = 4(X - 2)$$

que se puede expresar $Y + 3 = 4(X - 2)$, o sea $Y = 4X - 11$.

Segundo método [usando el método del Problema 13.1(*b*)]

La ecuación de una recta es $Y = a_0 + a_1X$. Como el punto $(2, -3)$ está en la recta, $-3 = a_0 + 2a_1$, y como el punto $(4, 5)$ está en la recta, $5 = a_0 + 4a_1$; resolviendo esas dos ecuaciones simultáneamente, obtenemos $a_1 = 4$ y $a_0 = -11$. Luego la ecuación pedida es

$$Y = -11 + 4X \qquad \text{o sea} \qquad Y = 4X - 11$$

13.4. Dar una interpretación gráfica de la parte (*a*) del Problema 13.3.

Solución

La Figura 13.5 muestra la recta que pasa por los puntos P y Q, de coordenadas (X_1, Y_1) y (X_2, Y_2), respectivamente. El punto R, con coordenadas (X, Y), representa cualquier otro punto sobre esa recta.

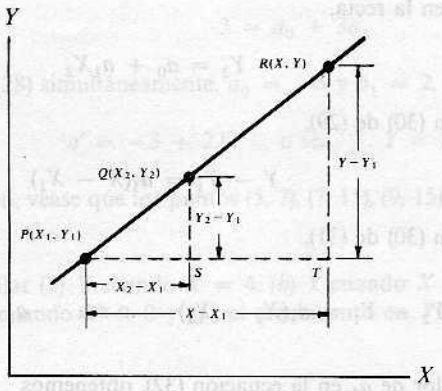


Figura 13.5.

Por semejanza de los triángulos PRT y PQS

$$\frac{RT}{TP} = \frac{QS}{SP} \qquad \text{o sea} \qquad \frac{Y - Y_1}{X - X_1} = \frac{Y_2 - Y_1}{X_2 - X_1} \tag{33}$$

Entonces, multiplicando ambos lados por $X - X_1$,

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

que es la ecuación solicitada para la recta.

Nótese que cada uno de los cocientes en la ecuación (33) es la pendiente m; eso puede escribirse $Y - Y_1 = m(X - X_1)$.

13.5. Hallar (*a*) la pendiente, (*b*) la ecuación, (*c*) la Y-intersección, y (*d*) la X-intersección de la recta que pasa por los puntos $(1, 5)$ y $(4, -1)$.

Solución

(*a*) $(X_1 = 1, Y_1 = 5)$ y $(X_2 = 4, Y_2 = -1)$. Luego

$$m = \text{pendiente} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{-1 - 5}{4 - 1} = \frac{-6}{3} = -2$$

El signo negativo de la pendiente indica que cuando X crece, Y decrece, tal como se ve en la Figura 13.6.

(*b*) La ecuación de la recta es

$$Y - Y_1 = m(X - X_1) \quad \text{o sea} \quad Y - 5 = -2(X - 1)$$

Es decir, $$Y - 5 = -2X + 2 \quad \text{o sea} \quad Y - 7 - 2X$$

Esto puede obtenerse también por el segundo método del Problema 13.3(*b*).

(*c*) La Y-intersección, que es el valor de Y cuando $X = 0$, viene dada por $Y = 7 - 2(0) = 7$. Eso puede verse directamente en la Figura 13.6.

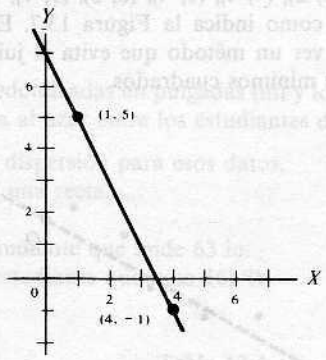


Figura 13.6.

(*d*) La X-intersección es el valor de X cuando $Y = 0$. Sustituyendo $Y = 0$ en la ecuación $Y = 7 - -2X$, o sea $0 = 7 - 2X$, o sea $2X = 7$ y $X = 3.5$. Eso puede verse directamente en la Figura 13.6.

13.6. Hallar las ecuaciones de una recta que pase por el punto (4, 2) y sea paralela a la recta $2X + 3Y = 6$.

Solución

Si dos rectas son paralelas, sus pendientes son iguales. De $2X + 3Y = 6$ tenemos $3Y = 6 - 2X$, o sea $Y = 2 - \frac{2}{3}X$, así que la pendiente de la recta es $m = -\frac{2}{3}$. Luego la ecuación de la recta pedida es

$$Y - Y_1 = m(X - X_1) \quad \text{o sea} \quad Y - 2 = -\tfrac{2}{3}(X - 4)$$

que también se puede escribir $2X + 3Y = 14$.

Otro método

Cualquier recta paralela a $2X + 3Y = 6$ tiene ecuación $2X + 3Y = c$. Para hallar c, hacemos $X = 4$ e $Y = 2$. Entonces $2(4) + 3(2) = c$, o sea $c = 14$, y la ecuación buscada es $2X + 3Y = 14$.

13.7. Hallar la ecuación de una recta cuya pendiente es -4 y cuya Y-intersección es 16.

Solución

En la ecuación $Y = a_0 + a_1X$, $a_0 = 16$ es la Y-intersección y $a_1 = -4$ es la pendiente. Así pues, la ecuación buscada es $Y = 16 - 4X$.

13.8. (*a*) Construir una recta que aproxime los datos de la Tabla 13.2.
(*b*) Hallar la ecuación de esa recta.

Tabla 13.2

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Solución

(*a*) Marcar los puntos (1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8) y (14, 9) sobre un sistema de coordenadas rectangulares, como indica la Figura 13.7. En la figura se ha trazado una recta aproximante *a mano*. Para ver un método que evita el juicio subjetivo, consultar el Problema 13.11, que usa el método de mínimos cuadrados.

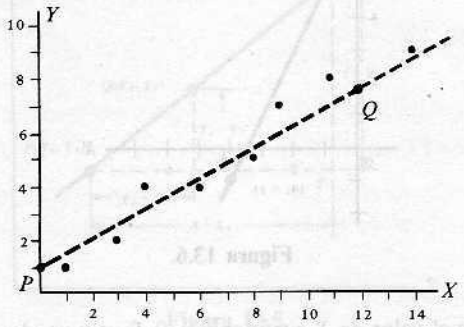


Figura 13.7.

(*b*) Para obtener una ecuación de esa recta, escojamos dos puntos en ella, tales como P y Q; las coordenadas de P y Q, según el gráfico, son aproximadamente (0, 1) y (12, 7.5). La ecuación de la recta es $Y = a_0 + a_1X$. Luego para que (0, 1) esté en ella, ha de ser $1 = a_0 + a_1(0)$, y para que esté el punto (12, 7.5), ha de ser $7.5 = a_0 + 12a_1$; como la primera de estas ecuaciones da $a_0 = 1$, la segunda nos dice que $a_1 = 6.5/12 = 0.542$. Por tanto, la requerida ecuación es $Y = 1 + 0.542X$.

Otro método

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \quad \text{e} \quad Y - 1 = \frac{7.5 - 1}{12 - 0}(X - 0) = 0.542X$$

Así pues $Y = 1 + 0.542X$.

13.9. (*a*) Comparar los valores de Y obtenidos de la recta aproximante con los de la Tabla 13.2.
(*b*) Estimar el valor de Y cuando $X = 10$.

Solución

(*a*) Para $X = 1$, $Y = 1 + 0.542(1) = 1.542$, o sea 1.5. Para $X = 3$, $Y = 1 + 0.542(3) = 2.626$, o 2,6. Los valores de Y correspondientes a otros valores de X se obtienen de la misma manera. Los valores de Y estimados por la ecuación $Y = 1 + 0.542X$ se denotan por Y_{est}. Estos valores estimados, junto con los verdaderos datos de la Tabla 13.2, se recogen en la Tabla 13.3.
(*b*) El valor estimado de Y cuando $X = 10$ es $Y = 1 + 0.542(10) = 6.42$ o sea 6.4.

Tabla 13.3

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9
Y_{est}	1.5	2.6	3.2	4.3	5.3	5.9	7.0	8.6

13.10. La Tabla 13.4 da las alturas redondeadas en pulgadas (in) y los pesos en libras (lb) de una muestra de 12 estudiantes varones tomada al azar entre los estudiantes de primer año del State College.

(*a*) Obtener un diagrama de dispersión para esos datos.
(*b*) Aproximar los datos con una recta.
(*c*) Hallar su ecuación.
(*d*) Estimar el peso de un estudiante que mide 63 in.
(*e*) Estimar la altura de un estudiante que pesa 168 lb.

Tabla 13.4

Altura X (in)	70	63	72	60	66	70	74	65	62	67	65	68
Peso Y (lb)	155	150	180	135	156	168	178	160	132	145	139	152

Solución

(*a*) El diagrama de dispersión, véase Figura 13.8, se obtiene marcando los puntos (70, 155), (63, 150), (72, 180), ..., (68, 152).

(*b*) Una recta que aproxima a los datos se ve en trazos en la Figura 13.8. No es sino una de las muchas posibles rectas que se podían haber construido.

(*c*) Escoger un par de puntos arbitrarios P y Q de esa recta. Sus coordenadas según el gráfico vienen a ser (60, 130) y (72, 170). Por tanto

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad Y - 130 = \frac{170 - 130}{72 - 60}(X - 60) \qquad Y = \frac{10}{3}X - 70$$

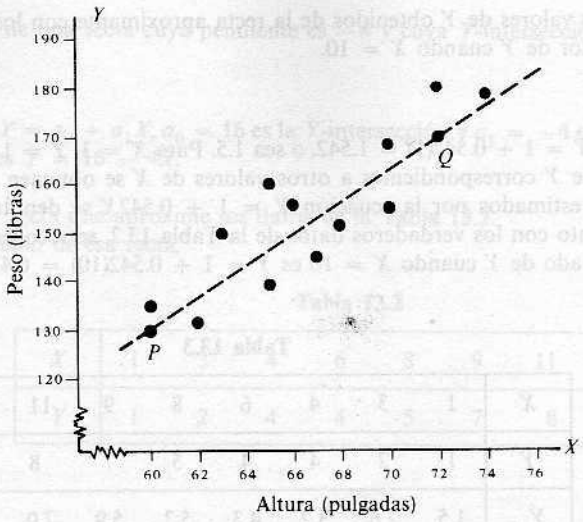


Figura 13.8.

(*d*) Si $X = 63$, entonces $Y = \frac{10}{3}(63) - 70 = 140$ lb.

(*e*) Si $Y = 168$, entonces $168 = \frac{10}{3}X - 70$, $\frac{10}{3}X = 238$ y $X = 71.4$, o sea 71 in.

LA RECTA DE MINIMOS CUADRADOS

13.11. Ajustar una recta de mínimos cuadrados a los datos del Problema 13.8 usando (*a*) X como variable independiente y (*b*) X como variable dependiente.

Solución

(*a*) La ecuación de la recta es $Y = a_0 + a_1 X$. Las ecuaciones normales son

$$\sum Y = a_0 N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

El trabajo exigido para calcular las sumas se puede ordenar como en la Tabla 13.5. Si bien la columna de la derecha no es necesaria para esta parte del problema, la usaremos en (*b*).

Tabla 13.5

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

Puesto que hay ocho pares de valores de X e Y, $N = 8$ y las ecuaciones normales se convierten en

$$\begin{aligned} 8a_0 + 56a_1 &= 40 \\ 56a_0 + 524a_1 &= 364 \end{aligned}$$

Resolviendo simultáneamente, $a_0 = \frac{6}{11}$, o sea 0.545; $a_1 = \frac{7}{11}$, o sea 0.636; y la recta de mínimos cuadrados pedida es $Y = \frac{6}{11} + \frac{7}{11}X$, o sea $Y = 0.545 + 0.636X$.

Otro método

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \quad \text{o sea} \quad 0.545$$

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \quad \text{o sea} \quad 0.636$$

Luego $Y = a_0 + a_1X$, o sea $Y = 0.545 + 0.636X$, como antes.

(*b*) Si se considera X como variable dependiente e Y como independiente, la ecuación de la recta de mínimos cuadrados es $X = b_0 + b_1Y$ y las ecuaciones normales son

$$\begin{aligned} \sum X &= b_0N + b_1\sum Y \\ \sum XY &= b_0\sum Y + b_1\sum Y^2 \end{aligned}$$

Por la Tabla 13.5 las ecuaciones normales se convierten en

$$\begin{aligned} 8b_0 + 40b_1 &= 56 \\ 40b_0 + 256b_1 &= 364 \end{aligned}$$

de donde $b_0 = -\frac{1}{2}$, o sea $-0{,}50$ y $b_1 = \frac{3}{2}$, o sea 1.50. Estos valores pueden deducirse también de

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Luego la ecuación solicitada de la recta de mínimos cuadrados es $X = b_0 + b_1Y$, o sea $X = -0.50 + 1.50Y$.

Nótese que resolviendo esa ecuación obtenemos $Y = \frac{1}{3} + \frac{2}{3}X$, o sea $Y = 0{,}333 + 0{,}667X$, que es distinta de la recta a la que llegamos en la parte (*a*).

13.12. Dibujar las dos rectas del Problema 13.11.

Solución

Los gráficos de las rectas $Y = 0.545 + 0.636X$ y $X = -0.500 + 1.50Y$, se muestran en la Figura 13.9. Hagamos notar que en este caso son casi coincidentes, lo cual indica que los datos están muy bien descritos por una relación lineal.

La recta de la parte (*a*) del Problema 13.11 se suele llamar la *recta de regresión de Y sobre X*, y se

usa para estimar Y en valores dados de X. La recta de la parte (b) del Problema 13.11 se suele llamar la *recta de regresión de X sobre Y*, y se usa para estimar X en valores dados de Y.

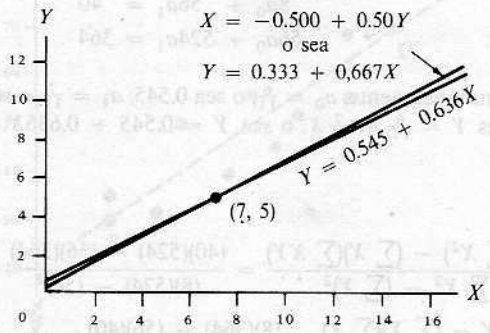


Figura 13.9.

13.13. (a) Probar que las rectas de mínimos cuadrados obtenidas en el Problema 13.11 se cortan en el punto $(\bar{X}, \bar{Y})$.
(b) Estimar el valor de Y cuando $X = 12$.
(c) Estimar el valor de X cuando $Y = 3$.

Solución

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7 \qquad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

Luego el punto $(\bar{X}, \bar{Y})$, llamado el *centroide*, es (7, 5).

(a) El punto (7,5) está en la recta $Y = 0{,}545 + 0{,}636X$; o, más exactamente, $Y = \frac{6}{11} + \frac{7}{11}X$, pues $5 = \frac{6}{11} + \frac{7}{11}(7)$. El punto (7, 5) está en la recta $X = -\frac{1}{2} + \frac{3}{2}Y$, ya que $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Otro método

Las ecuaciones de las dos rectas son $Y = \frac{6}{11} + \frac{7}{11}X$ y $X = -\frac{1}{2} + \frac{3}{2}Y$. Resolviendo simultáneamente se encuentra $X = 7$ e $Y = 5$. Luego las rectas se cortan en el punto (7, 5).

(b) Haciendo $X = 12$ en la recta de regresión de Y(Problema 13.11), $Y = 0.545 + 0636(12) = 8.2$.
(c) Haciendo $Y = 3$ en la recta de regresión de X(Problema 13.11), $X = -0.50 + 1.50(3) = 4.0$.

13.14. Probar que una recta de mínimos cuadrados siempre pasa por el punto $(\bar{X}, \bar{Y})$.

Solución

Caso 1: (X es la variable independiente)

La ecuación de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \tag{34}$$

Una ecuación normal para la recta de mínimos cuadrados es

$$\sum Y = a_0 N + a_1 \sum X \tag{35}$$

Dividiendo la ecuación (35) a ambos lados por N tenemos

$$\bar{Y} = a_0 + a_1\bar{X} \tag{36}$$

Restando (36) de (34), la recta de mínimos cuadrados se puede expresar

$$Y - \bar{Y} = a_1(X - \bar{X}) \tag{37}$$

que demuestra que la recta pasa por el punto $(\bar{X}, \bar{Y})$.

Caso 2: (X es la variable dependiente)

Procediendo como en el caso 1, pero intercambiando X e Y y sustituyendo las constantes a_0 y a_1 por b_0 y b_1, respectivamente, vemos que la recta de mínimos cuadrados se puede escribir

$$X - \bar{X} = b_1(Y - \bar{Y}) \tag{38}$$

lo cual indica que la recta pasa por el punto $(\bar{X}, \bar{Y})$.

Nótese que las rectas (37) y (38) no son coincidentes; se cortan en $(\bar{X}, \bar{Y})$.

13.15. (*a*) Considerando X como variable independiente, probar que la ecuación de la recta de mínimos cuadrados se puede escribir como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{es decir} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

donde $x = X - \bar{X}$ donde $y = Y - \bar{Y}$.

(*b*) Si $\bar{X} = 0$, demostrar que la recta de mínimos cuadrados de la parte (*a*) se escribe

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) Escribir la ecuación de la recta de mínimos cuadrados correspondiente a la de la parte (*a*) si Y es la variable independiente.

(*d*) Verificar que las rectas en las partes (*a*) y (*c*) no son necesariamente la misma.

Solución

(*a*) La ecuación (37) se puede escribir $y = a_1x$, donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. Además, de la solución simultánea de las ecuaciones normales (18) tenemos

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum (x + \bar{X})(y + \bar{Y}) - [\Sigma (x + \bar{X})][\sum (y + \bar{Y})]}{N\sum (x + \bar{X})^2 - [\sum (x + \bar{X})]^2}$$

$$= \frac{N\sum (xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum (x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + N\bar{X})^2}$$

$$= \frac{N\sum xy + N\bar{Y}\sum x + N\bar{X}\sum y + N^2\bar{X}\bar{Y} - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum x^2 + 2N\bar{X}\sum x + N^2\bar{X}^2 - (\sum x + N\bar{X})^2}$$

Pero $\sum x = \sum (X - \bar{X}) = 0$ y $\sum y = \sum (Y - \bar{Y}) = 0$; por tanto, lo anterior se reduce a

$$a_1 = \frac{N\sum xy + N^2\bar{X}\bar{Y} - N^2\bar{X}\bar{Y}}{N\sum x^2 + N^2\bar{X}^2 - N^2\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

Esto puede escribirse como

$$a_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY - \bar{Y}\sum x}{\sum x^2} = \frac{\sum xY}{\sum x^2}$$

Asi que la recta de mínimos cuadrados es $y = a_1 x$; es decir,

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{o sea} \quad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

(*b*) Si $\bar{X} = 0$, $x = X - \bar{X} = X$. Entonces de

$$y = \left(\frac{\sum xY}{\sum x^2}\right)$$

se tiene $$y = \left(\frac{\sum XY}{\sum X^2}\right)X \quad \text{o sea} \quad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

Otro método

Las ecuaciones normales de la recta de mínimos cuadrados $Y = a_0 + a_1 X$ son

$$\sum Y = a_0 N + a_1 \sum X \quad \text{y} \quad \sum XY = a_0 \sum X + a_1 \sum X^2$$

Si $\bar{X} = (\sum X)/N = 0$, entonces $\sum X = 0$ y las ecuaciones normales pasan a ser

$$\sum Y = a_0 N \quad \text{y} \quad \sum XY = a_1 \sum X^2$$

de donde $$a_0 = \frac{\sum Y}{N} = \bar{Y} \quad \text{y} \quad a_1 = \frac{\sum XY}{\sum X^2}$$

Luego la ecuación pedida de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \quad \text{o sea} \quad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) Intercambiando X e Y, o sea x e y, podemos ver como en (*a*) que

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

(*d*) Por la parte (*a*), la recta de mínimos cuadrados es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \tag{39}$$

Por la parte (*c*), la recta de mínimos cuadrados es

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

o sea

$$y = \left(\frac{\sum y^2}{\sum xy}\right) x \tag{40}$$

Como en general $$\frac{\sum xy}{\sum x^2} \neq \frac{\sum y^2}{\sum xy}$$

la recta de mínimos cuadrados (39) y (40) son diferentes en general. Sin embargo, intersectan en $x = 0$ e $y = 0$ [o sea, en el punto $(\bar{X}, \bar{Y})$].

13.16. Si $X' = X + A$ e $Y' = Y + B$, donde A y B son constantes, probar que

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = a_1'$$

Solución

$$x' = X' - \bar{X}' = (X + A) - (\bar{X} + A) = X - \bar{X} = x$$
$$y' = Y' - \bar{Y}' = (Y + B) - (\bar{Y} + B) = Y - \bar{Y} = y$$

Entonces $$\frac{\sum xy}{\sum x^2} = \frac{\sum x'y'}{\sum x'^2}$$

y el resultado se sigue del Problema 13.15. Un resultado similar se aplica a b_1.

Este resultado es útil porque nos capacita para simplificar cálculos al obtener la recta de regresión restando constantes adecuadas de las variables X e Y (véase el segundo método del Problema 13.17).

Nota: El resultado no es válido si $X' = c_1X + A$ e $Y' = c_2Y + B$ a menos que $c_1 = c_2$.

13.17. Ajustar una recta de mínimos cuadrados a los datos del Problema 13.10 usando (*a*) X como variable independiente y (*b*) X como variable dependiente.

Solución

Primer método

(*a*) Del Problema 13.15(*a*) sabemos que la recta requerida es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. El trabajo de calcular las sumas se puede organizar como sugiere la Tabla 13.6. De sus dos primeras columnas hallamos $\bar{X} = 802/12 = 66.8$ e $\bar{Y} = 1850/12 = 154.2$. La última columna se utilizará en la parte (*b*).

La recta de mínimos cuadrados pedida es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x = \frac{616.32}{191.68} x = 3.32x$$

o sea $Y - 154.2 = 3.22(X - 66.8)$, que se puede escribir $Y = 3.22X - 60.9$ Esta ecuación se llama la *recta de regresión de Y sobre X*, y se usa para estimar Y para valores dados de X.

(*b*) Si X es la variable dependiente, la recta en cuestión es

$$x = \left(\frac{\sum xy}{\sum y^2}\right) y = \frac{616.32}{2659.68} y = 0.232y$$

Tabla 13.6

Altura X	Peso Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	−3.8	−4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	−6.8	−19.2	130.56	46.24	368.64
66	156	−0,8	1.8	−1.44	0,64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	−1.8	5.8	−10.44	3.24	33.64
62	132	−4.8	−22.2	106.52	23.04	492.84
67	145	0.2	−9.2	− 1.84	0.04	84.64
65	139	−1.8	−15.2	27.36	3.24	231.04
68	152	1.2	−2.2	−2.64	1.44	4.84
$\sum X = 802$	$\sum Y = 1850$			$\sum xy = 616.32$	$\sum x^2 = 191.68$	$\sum y^2 = 2659.68$
$\bar{X} = 66.8$	$\bar{Y} = 154.2$					

que se puede expresar como $X - 66.8 = 0.232(Y - 154.2)$, o sea $X = 31.0 + 0.232\ Y$. Esta es la *recta de regresión de X sobre Y*, usada para estimar X para valores de Y dados.

Nótese que el método del Problema 13.11 es también aplicable si se desea.

Segundo método

Usando el resultado del Problema 13.16, podemos sustraer constantes adecuadas de X e Y. Escogemos sustraer 65 de X y 150 de Y. Con ello los resultados se muestran en la Tabla 13.7.

Tabla 13.7

X'	Y'	X'^2	$X'Y'$	Y'^2
5	5	25	25	25
−2	0	4	0	0
7	30	49	210	900
−5	−15	25	75	225
1	6	1	6	36
5	18	25	90	324
9	28	81	252	784
0	10	0	0	100
−3	−18	9	54	324
2	−5	4	−10	25
0	−11	0	0	121
3	2	9	6	4
$\sum X' = 22$	$\sum Y' = 50$	$\sum X'^2 = 232$	$\sum X'Y' = 708$	$\sum Y'^2 = 2868$

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = \frac{(12)(708) - (22)(50)}{(12)(232) - (22)^2} = 3.22$$

$$b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = \frac{(12)(708) - (50)(22)}{(12)(2868) - (50)^2} = 0{,}232$$

Como $\bar{X} = 65 + 22/12 = 66.8$ e $\bar{Y} = 150 + 50/12 = 154.2$, las ecuaciones de regresión son $Y - 154.2 = 3.2(X - 66.8)$ y $X - 66.8 = 0.232\,(Y - 154.2)$; esto es, $Y = 3.22\,X - 60.9$ y $X = 0.232\,Y + 31.0$, de acuerdo con el primer método.

13.18. (*a*) Dibujar, en un mismo par de ejes, los gráficos de las dos rectas del Problema 13.17.
(*b*) Estimar el peso de un estudiante que mide 63 in.
(*c*) Estimar la altura de un estudiante que pesa 168 lb.

Solución

(*a*) Las dos rectas se muestran en la Figura 13.10, junto a los puntos dato originales. Obsérvese que se cortan en $(\bar{X}, \bar{Y})$, o sea (66.8, 154.2).
(*b*) Para estimar Y a partir de X usaremos la recta de regresión de Y sobre X, dada en el Problema 13.17 por $Y = 3.22X - 60.9$. Entonces, si $X = 63$, $Y = 3.22(63) - 60.9 = 142$ lb.
(*c*) Para estimar X a partir de Y usaremos la recta de regresión de X sobre Y, dada en el Problema 13.17 por $X = 31.0 + 0.232Y$. Entonces, si $Y = 168$, $X = 31.0 + 0.232(168) = 70.0$ in.

Los resultados de las partes (*b*) y (*c*) deben compararse con los del Problema 13.10, partes (*d*) y (*c*).

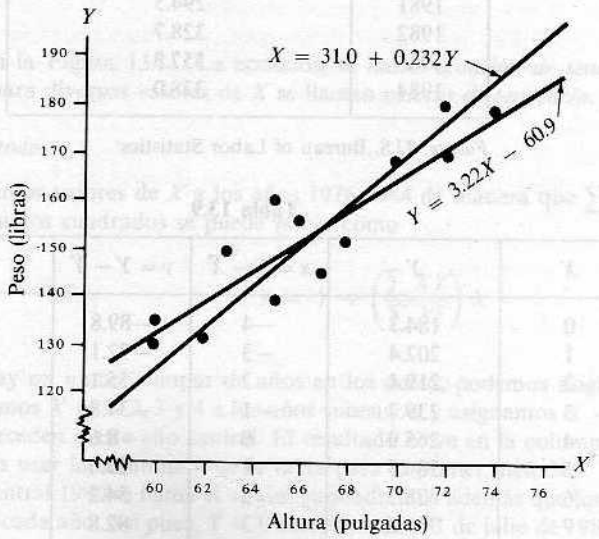


Figura 13.10.

APLICACIONES A SERIES EN EL TIEMPO

13.19. El índice de costes sanitarios en EE.UU. para los años 1976-1984, tomando como 100 el del año 1967, se da en la Tabla 13.8.

(*a*) Representar los datos gráficamente.
(*b*) Hallar la ecuación de una recta de mínimos cuadrados que ajuste esos datos.

(*c*) Estimar el índice para el año 1985 y comparar con el valor real, 396.1.
(*d*) Estimar el índice para 1975 y comparar con el valor verdadero, 168.6.

Solución

(*a*) Véase Figura 13.11.
(*b*) *Primer método*

Usar las ecuaciones $y = (\sum xy/\sum x^2)x$, donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. La Tabla 13.9 resume la tarea. La requerida ecuación es $y = (1511.3/60)x$, o sea $y = 25.19x$, que se puede escribir

$$Y - 274.5 = 25.19(X - 4) \quad \text{o sea} \quad Y = 173.7 + 25.19X$$

Tabla 13.8

Año	Indice de costes sanitarios en EE. UU. (1967 = 100)
1976	184.7
1977	202.4
1978	219.4
1979	239.7
1980	265.9
1981	294.5
1982	328.7
1983	357.3
1984	378.0

Fuente: U.S. Bureau of Labor Statistics.

Tabla 13.9

Año	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
1976	0	184.7	−4	−89.8	16	359.2
1977	1	202.4	−3	−72.1	9	216.3
1978	2	219.4	−2	−55.1	4	110.2
1979	3	239.7	−1	−34.8	1	34.8
1980	4	265.9	0	−8.6	0	0.0
1981	5	284.5	1	20.0	1	20.0
1982	6	328.7	2	54.2	4	108.4
1983	7	357.3	3	82.8	9	248.4
1984	8	378.0	4	103.5	16	414.0
	$\sum X = 36$	$\sum Y = 2470.6$			$\sum x^2 = 60$	$\sum xy = 1511.3$
	$\bar{X} = 4$	$\bar{Y} = 274.5$				

donde el origen $X = 0$ es el año 1976 (se suele tomar la mitad del año, el 1 de julio de 1976) y la unidad de X es 1 año. El gráfico de esta recta, llamada a veces una *recta de tendencia*, se muestra

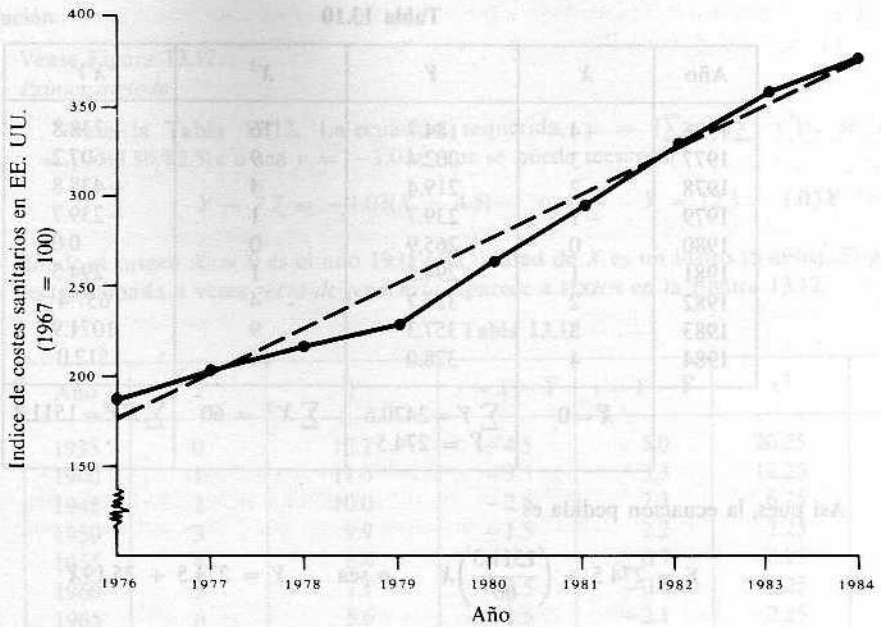


Figura 13.11.

en trazos en la Figura 13.11. La ecuación se llama *ecuación de tendencia*, y los valores de Y calculados para diversos valores de X se llaman *valores de tendencia*.

Segundo método

Si asignamos valores de X a los años 1976-1984 de manera que $\sum X = 0$, la ecuación de la recta de mínimos cuadrados se puede poner como

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right) X$$

Como hay un número impar de años en los datos, podemos asignar $X = 0$ al año central, 1980; asignamos $X = 1, 2, 3$ y 4 a los años sucesivos; y asignamos $X = -1, -2, -3$ y -4 a los años que preceden a este año central. El resultado se ve en la columna 2 de la Tabla 13.10 y es equivalente a usar la columna 4 de la tabla para el primer método.

El año central 1980 se llama el *origen*; supondremos además que los valores de Y se refieren al 1 de julio de cada año. Así pues, $X = 0$ corresponde al 1 de julio de 1980; $X = -1$ al 1 de julio de 1979, etc. Los cálculos se resumen en la Tabla 13.10.

Tabla 13.10

Año	X	Y	X^2	XY
1976	−4	184.7	16	−738.8
1977	−3	202.4	9	−607.2
1978	−2	219.4	4	−438.8
1979	−1	239.7	1	−239.7
1980	0	265.9	0	0.0
1981	1	294.5	1	294.5
1982	2	328.7	4	657.4
1983	3	357.3	9	1071.9
1984	4	378.0	16	1512.0
	$\bar{X}=0$	$\sum Y=2470.6$ $\bar{Y}=274.5$	$\sum X^2=60$	$\sum XY=1511.3$

Así pues, la ecuación pedida es

$$Y = 274.5 + \left(\frac{1511.3}{60}\right)X \quad \text{o sea} \quad Y = 274.5 + 25.19X$$

donde el origen $X = 0$ es el año 1980 y la unidad de X es 1 año. Para desplazar el origen a 1976, 4 años antes, sustituimos X por $X - 4$, con lo que se llega a la ecuación $Y = 274.5 + 25.19(X - 4)$, o $Y = 173.7 + 25.19X$, como en el primer método.

El segundo método es mejor que el primero porque reduce el trabajo de cálculo. Sin embargo, mientras el primer método es aplicable en todos los casos, el segundo exige modificaciones en el caso de un número de años par en los datos. Para tal modificación, ver el segundo método del Problema 13.20(*b*).

(*c*) Usar la ecuación de tendencia $Y = 173.7 + 25.19X$, donde $X = 0$ corresponde al año 1976. Entonces el año 1985 corresponde a $X = 9$, luego el valor de Y para 1985 es $Y = 173.7 + 25.19(9) = 400.4$.

El mismo resultado se puede obtener de la ecuación de tendencia $Y = 274.5 + 25.19X$, donde el origen $X = 0$ corresponde al año 1980, haciendo $X = 5$.

(*d*) Usando la ecuación de tendencia $Y = 173.7 + 25.19X$, con $X = -1$, encontramos el valor $Y = 173.7 + 25.19(-1) = 148.5$.

13.20. La Tabla 13.11 indica el censo de trabajadores agrícolas en EE. UU. los años 1935, 1940, 1945, ..., 1980, en millones.

(*a*) Representar los datos gráficamente.

(*b*) Hallar una ecuación para la recta de mínimos cuadrados que ajuste esos datos.

(*c*) Predecir el censo de trabajadores agrícolas en los años 1990 y 2000, suponiendo que la tendencia se mantenga.

Tabla 13.11

Año	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Trabajadores agrícolas en EE. UU. (millones)	12.7	11.0	10.0	9.9	8.4	7.1	5.6	4.5	4.3	3.7

Fuente: U.S. Department of Agriculture.

Solución

(*a*) Véase Figura 13.12.

(*b*) *Primer método*

Véase la Tabla 13.12. La ecuación requerida, $y = (\sum xy/\sum x^2)x$, se convierte en $y = (-84.80/82.5)x$ o sea $y = -1{,}03x$, que se puede reescribir

$$Y - 7.7 = -1.03(X - 4.5) \qquad \text{o sea} \qquad Y = 12.3 - 1.03X$$

donde el origen $X = 0$ es el año 1935 y la unidad de X es un lustro (5 años). El gráfico de esta recta, llamada a veces *recta de tendencia*, aparece a trazos en la Figura 13.12.

Tabla 13.12

Año	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
1935	0	12.7	−4.5	5.0	20.25	−22.50
1940	1	11.0	−3.5	3.3	12.25	−11.55
1945	2	10.0	−2.5	2.3	6.25	−5.75
1950	3	9.9	−1.5	2.2	2.25	−3.30
1955	4	8.4	−0.5	0.7	0.25	−0.35
1960	5	7.1	0.5	−0.6	0.25	−0.30
1965	6	5.6	1.5	−2.1	2.25	−3.15
1970	7	4.5	2.5	−3.2	6.25	−8.00
1975	8	4.3	3.5	−3.4	12.25	−11.90
1980	9	3.7	4.5	−4.0	20.25	−18.00
	$\sum X = 45$	$\sum Y = 77.2$			$\sum x^2 = 82.5$	$\sum xy = -84.80$
	$\bar{X} = 4.5$	$\bar{Y} = 7.7$				

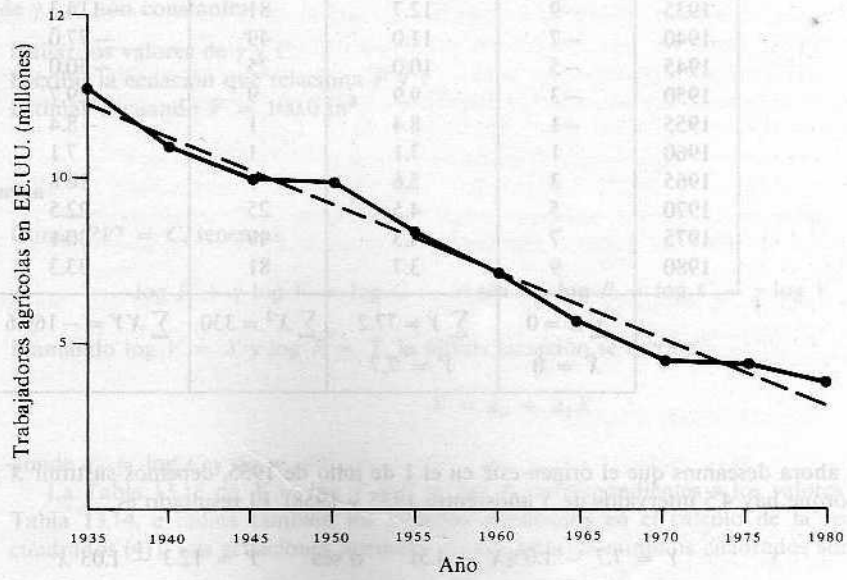


Figura 13.12.

Segundo método

En este método queremos asignar valores de X a los años de modo que $\sum X = 0$. Como hay un número par de años, no hay año central y no se puede usar el segundo método del Problema 13.19(*b*). No obstante, podemos asociar los números -0.5 y 0.5 a los años centrales, 1955 y 1960, de manera que 1965, 1970, 1975 y 1980 están representados por 1.5, 2.5, 3.5 y 4.5 y 1950, 1945, 1940 y 1935 lo están por -1.5, -2.5, -3.5 y -4.5. Esto viene a ser esencialmente la columna 4 de la Tabla 13.12.

Además, para evitar fracciones, doblamos esos valores, obteniendo la columna 2 de la Tabla 13.13. Nótese que con estos valores de X el origen $X = 0$ está a medio camino entre el 1 de julio de 1955 y el 1 de julio de 1960, que es el 1 de enero de 1958 o el 31 de diciembre de 1957. La unidad de X es medio lustro, o sea 2.5 años. Como $\bar{X} = 0$, la ecuación pedida tiene la forma $Y = \bar{Y} + (\sum XY/\sum X^2)X$, que da (véase Tabla 13.13)

$$Y = 7.7 + \left(\frac{-169.6}{330}\right)X \qquad \text{o sea} \qquad Y = 7.7 - 0.514X$$

donde el origen $X = 0$ corresponde al 1 de enero de 1958, y X se mide en unidades de 2.5 años.

Si queremos medir X en intervalos de 5 años en vez de 2.5 años, debemos reemplazar X por $2X$, con lo que la ecuación es

$$Y = 7.7 - 1.028X \qquad \text{o sea} \qquad Y = 7.7 - 1.03X$$

donde el origen es el 1 de enero de 1958, y X se mide en unidades de 5 años.

Tabla 13.13

Año	X	Y	X^2	XY
1935	-9	12.7	81	-114.3
1940	-7	11.0	49	-77.0
1945	-5	10.0	25	-50.0
1950	-3	9.9	9	-29.7
1955	-1	8.4	1	-8.4
1960	1	7.1	1	7.1
1965	3	5.6	9	16.8
1970	5	4.5	25	22.5
1975	7	4.3	49	30.1
1980	9	3.7	81	33.3
	$\sum X = 0$ $\bar{X} = 0$	$\sum Y = 77.2$ $\bar{Y} = 7.7$	$\sum X^2 = 330$	$\sum XY = -169.6$

Si ahora deseamos que el origen esté en el 1 de julio de 1935, debemos sustituir X por $X - 4.5$ (porque hay 4.5 intervalos de 5 años entre 1935 y 1958). El resultado es

$$Y = 7.7 - 1.03(X - 4.5) \qquad \text{o sea} \qquad Y = 12.3 - 1.03X$$

Esto coincide con la ecuación obtenida en el primer método.

(*c*) Usando el primer método en la parte (*b*), los años 1990 y 2000 corresponden a $X = 11$ y $X = 13$, respectivamente. Entonces

$$Y = 12.3 - 1.03X = 12.3 - 1.03(11) = 0.97 \text{ millones en } 1990$$
$$Y = 12.3 - 1.03X = 12.3 - 1.03(13) = -1.09 \text{ millones en } 2000$$

Mientras el primer resultado de un millón de trabajadores agrícolas en 1990 es posible, especialmente a la vista de las nuevas tecnologías y de las importaciones agrícolas, el segundo resultado es claramente imposible. Hemos de concluir que la tendencia que muestra la Tabla 13.13 no se mantendrá por mucho tiempo.

ECUACIONES NO LINEALES REDUCIBLES A FORMA LINEAL

13.21. La Tabla 13.14 presenta valores experimentales de la presión P de una masa dada de gas correspondiente a varios valores del volumen V.

Tabla 13.14

Volumen V en pulgadas cúbicas (in^3)	54.3	61.8	72.4	88.7	118.6	194.0
Presión P en libras por pulgada cuadrada (lb/in^2)	61.2	49.5	37.6	28.4	19.2	10.1

De acuerdo con la Termodinámica, existe una relación del tipo $PV^\gamma = C$ entre las variables P y V, donde γ y C son constantes.

(*a*) Hallar los valores de γ y C.
(*b*) Escribir la ecuación que relaciona P y V.
(*c*) Estimar P cuando $V = 100.0$ in^3.

Solución

Como $PV^\gamma = C$, tenemos

$$\log P + \gamma \log V = \log C \quad \text{o sea} \quad \log P = \log C - \gamma \log V$$

Llamando $\log V = X$ y $\log P = Y$, la última ecuación se escribe

$$Y = a_0 + a_1 X \tag{41}$$

donde $a_0 = \log C$ y $a_1 = -\gamma$.

La Tabla 13.15 da $X = \log V$ e $Y = \log P$, correspondientes a los valores de V y P de la Tabla 13.14, e indica también los cálculos implicados en el cálculo de la recta de mínimos cuadrados (41). Las ecuaciones normales de esa recta de mínimos cuadrados son

$$\sum Y = a_0 N + a_1 \sum X \quad \text{y} \quad \sum XY = a_0 \sum X + a_1 \sum X^2$$

Tabla 13.15

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

de donde

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = 4.20 \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = -1.40$$

Luego $Y = 4.20 - 1.40\,X$.

(*a*) Como $a_0 = 4.20 = \log C$ y $a_1 = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ y $\gamma = 1.40$.
(*b*) La ecuación requerida en términos de P y V puede escribirse $PV^{1.40} = 16{,}000$.
(*c*) Cuando $V = 100$, $X = \log V = 2$ e $Y = \log P = 4.20 - 1.40(2) = 1.40$. Entonces $P =$ antilog $1.40 = 25.1$ lb/in^2.

13.22. Resolver el Problema 13.21 representando los datos en papel log-log.

Solución

Obtenemos primero un punto para cada par de valores de la presión P y del volumen V en la Tabla 13.14, y marcamos esos puntos en papel log-log, como indica la Figura 13.13. Entonces trazamos una recta que aproxime esos puntos (la recta de la figura esté trazada «a mano»). El gráfico resultante muestra que hay una relación lineal entre $\log P$ y $\log V$ representable por la ecuación

$$\log P = a_0 + a_1 \log V \qquad \text{o sea} \qquad Y = a_0 + a_1 X$$

La pendiente a_1, que es negativa en este caso, viene dada numéricamente por el cociente de longitudes de AB y AC (usando una unidad de longitud apropiada). La medida en este caso da $a_1 = -1.4$.

Para hallar a_0, se necesita un punto sobre la recta. Por ejemplo, cuando $V = 100$, $P = 25$ en el gráfico; por tanto, $a_0 = \log P - a_1 \log V = \log 25 + 1.4 \log 100 = 1.4 + (1.4)(2) = 4.2$, y en consecuencia tenemos $\log P + 1.4 \log V = 4.2$ $\log PV^{1.4} = 4.2$ y $PV^{1.4} = 16{,}000$.

LA PARABOLA DE MINIMOS CUADRADOS

13.23. La Tabla 13.16 da la población de EE. UU. en los años 1880-1980 en intervalos de 10 años.

(*a*) Hallar la ecuación de una parábola de mínimos cuadrados que ajuste los datos.
(*b*) Calcular los valores de tendencia para los años dados en la Tabla 13.16, y compararlos con los verdaderos.
(*c*) Estimar la población en 1990 y 2000.
(*d*) Estimar la población en 1870 y 1860, y comparar con los valores reales (véase página 18).

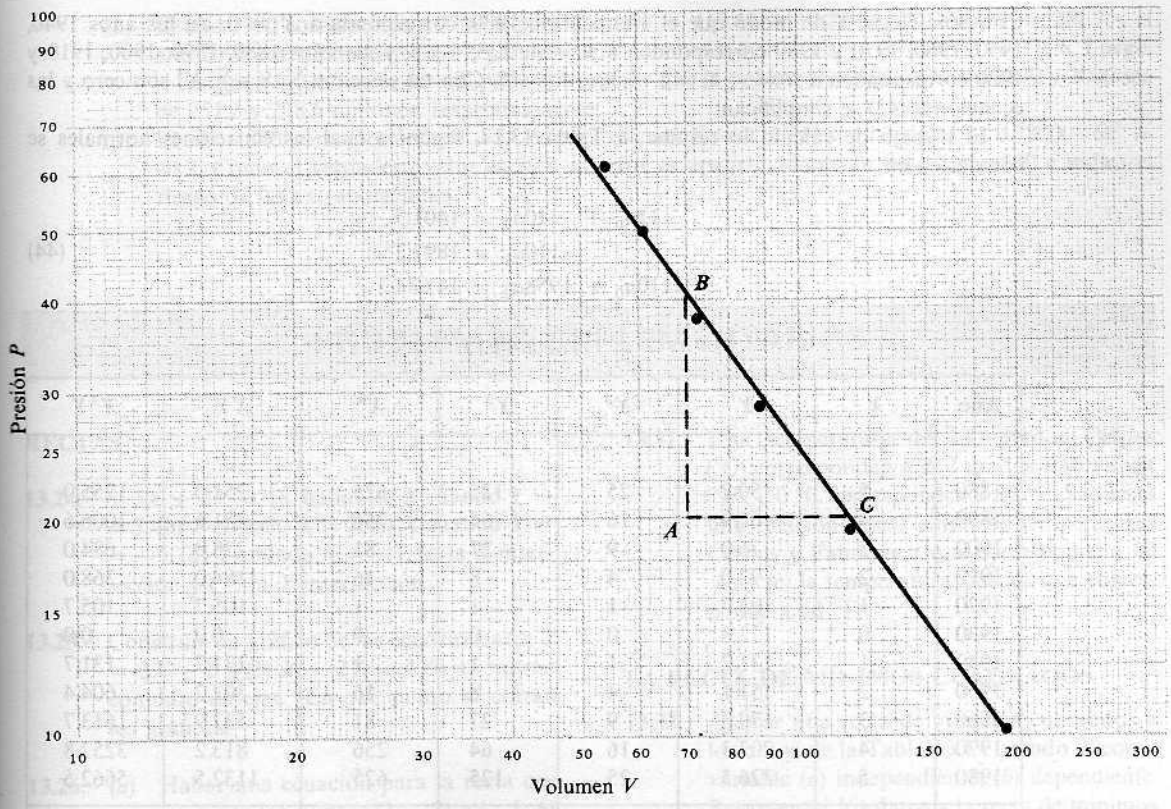


Figura 13.13.

Tabla 13.16

Año	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
Población de EE. UU. (millones)	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1	179.3	203.3	226.5

Fuente: U.S. Bureau of the Census.

Solución

(*a*) Sean X, Y, respectivamente, el año y la población en ese año. La ecuación de una parábola de mínimos cuadrados que ajuste los datos es

$$Y = a_0 + a_1 X + a_2 X^2 \tag{42}$$

donde a_0, a_1 y a_2 se deducen de las ecuaciones normales

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X_2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{43}$$

Conviene elegir X de modo que el año central, 1930, corresponda a $X = 0$; así los años 1940, 1950, 1960, 1970 y 1980 corresponden a $X = 1, 2, 3, 4$ y 5; y los años 1880, 1890, 1900, 1910 y 1920 corresponden a $X = -1, -2, -3, -4$ y -5. Con tal elección, $\sum X$ y $\sum X^3$ son cero y las ecuaciones (43) se simplifican.

El trabajo de cálculo lo resume la Tabla 13.17, según la cual las ecuaciones normales se convierten en

$$\begin{aligned} 11a_0 + 110a_2 &= 1401.5 \\ 110a_1 &= 1897.2 \\ 110a_0 + 1958a_2 &= 14{,}684.2 \end{aligned} \tag{44}$$

Tabla 13.17

Año	X	Y	X^2	X^3	X^4	XY	X^2Y
1880	−5	50.2	25	−125	625	−251.0	1255.0
1890	−4	62.9	16	−64	256	−251.6	1006.6
1900	−3	76.0	9	−27	81	−228.0	684.0
1910	−2	92.0	4	−8	16	−184.0	368.0
1920	−1	105.7	1	−1	1	−105.7	105.7
1930	0	122.8	0	0	0	0.0	0.0
1940	1	131.7	1	1	1	131.7	131.7
1950	2	151.1	4	8	16	302.2	604.4
1960	3	179.3	9	27	81	537.9	1613.7
1970	4	203.3	16	64	256	813.2	3252.8
1980	5	226.5	25	125	625	1132.5	5662.5
	$\sum X = 0$	$\sum Y = 1401.5$	$\sum X^2 = 110$	$\sum X^3 = 0$	$\sum X^4 = 1958$	$\sum XY = 1897.2$	$\sum X^2Y = 14{,}684.2$

De la segunda ecuación en (44), $a_1 = 17.25$; de la primera, $a_0 = 119.61$; y de la tercera, $a_2 = 0.7800$. Luego la ecuación buscada es

$$Y = 119.61 + 17.25X + 0.7800X^2 \tag{45}$$

donde el origen $X = 0$ es el 1 de julio de 1930 y la unidad de X son 10 años.

(*b*) Los valores de tendencia se obtienen haciendo $X = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4$ y 5 en la ecuación (45). Estos valores de tendencia, junto con los valores reales, se recogen en la Tabla 13.18. Vemos que el acuerdo es bueno.

Tabla 13.18

Año	$X=-5$ 1880	$X=-4$ 1890	$X=-3$ 1900	$X=-2$ 1910	$X=-1$ 1920	$X=0$ 1930	$X=1$ 1940	$X=2$ 1950	$X=3$ 1960	$X=4$ 1970	$X=5$ 1980
Valor de tendencia	52.9	63.1	74.9	88.2	103.1	119.6	137.6	157.2	178.4	201.1	225.4
Valor real	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1	179.3	203.3	226.5

(*c*) El año 1990 corresponde a $X = 6$, para el que $Y = 119{,}61 + 17.25(6) + 0.7800(6)^2 = 251.2$, y el año 2000 corresponde a $X = 7$, para el que $Y = 119.61 + 17.25(7) + 0.7800(7)^2 = 278.6$. Luego, si continúa la tendencia actual, podemos esperar que la población de EE. UU. en 1990 y 2000 sea de 251.2 y 278.6 millones, respectivamente.

(*d*) El año 1870 corresponde a $X = -6$, para el cual $Y = 119.61 + 17.25(-6) + 0.7800(-6)^2 = 44.2$. Como el verdadero valor es 39.8, el error es aproximadamente del 10 por 100 e indica el riesgo de las extrapolaciones.

PROBLEMAS SUPLEMENTARIOS

RECTAS

13.24. Si $3X + 2Y = 18$, hallar (*a*) X cuando $Y = 3$, (*b*) Y cuando $X = 2$, (*c*) X cuando $Y = -5$, (*d*) Y cuando $X = -1$, (*e*) la X-intersección y (*f*) la Y-intersección.

13.25. Construir un gráfico de las ecuaciones (*a*) $Y = 3X - 5$ y (*b*) $X + 2Y = 4$ en un mismo conjunto de ejes. ¿En qué punto se cortan los gráficos?

13.26. (*a*) Hallar una ecuación para la recta que pasa por los puntos $(3, -2)$ y $(-1, 6)$.
(*b*) Determinar sus intersecciones con los ejes.
(*c*) Hallar el valor de Y correspondiente a $X = 3$ y $X = 5$.
(*d*) Verificar directamente sobre el gráfico las respuestas de (*a*), (*b*) y (*c*).

13.27. Hallar una ecuación para la recta de pendiente $2/3$ y cuya Y-intersección es -3.

13.28. (*a*) Hallar la pendiente y la Y-intersección de la recta $3X - 5Y = 20$.
(*b*) ¿Cuál es la ecuación de una recta paralela a la de la parte (*a*) y que pasa por el punto $(2, -1)$?

13.29. Hallar (*a*) la pendiente, (*b*) la Y-intersección y (*c*) la ecuación de la recta que pasa por los puntos (5, 4) y (2, 8).

13.30. Hallar la ecuación de una recta cuyas intersecciones X e Y son 3 y -5, respectivamente.

13.31. Una temperatura de 100 grados Celsius (°C) corresponden a 212 grados Fahrenheit (°F), y 0 °C corresponden a 32 °F. Supuesta una relación lineal entre las temperaturas Celsius y Fahrenheit que corresponde a 80 C, y (*c*) la temperatura Celsius que corresponde a 68 °F.

LA RECTA DE MINIMOS CUADRADOS

13.32. Ajustar una recta de mínimos cuadrados a los datos de la Tabla 13.19 usando X como variable (*a*) independiente, (*b*) dependiente. Representar los datos y la recta de mínimos cuadrados sobre unos mismos ejes de coordenadas.

Tabla 13.19

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

13.33. Para los datos del Problema 13.32, hallar (*a*) los valores de Y cuando $X = 5$ y $X = 12$ y (*b*) el valor de X cuando $Y = 7$.

13.34. (*a*) Obtener una ecuación, por el método «a mano», para una recta que ajuste los datos del Problema 13.32.
(*b*) Usando el resultado de (*a*), resolver el Problema 13.33.

13.35. La Tabla 13.20. presenta las notas en Algebra y Física de 10 estudiantes elegidos al azar entre un grupo muy numeroso.

Tabla 13.20

Algebra (X)	Física (Y)
75	82
80	78
93	86
65	72
87	91
71	80
98	95
68	72
84	89
77	74

(a) Representar los datos.
(b) Hallar una recta de mínimos cuadrados que ajuste los datos, usando X como variable independiente.
(c) Hallar una recta de mínimos cuadrados que ajuste los datos, usando Y como variable independiente.
(d) Si un estudiante tiene 75 en Algebra, ¿cuál es su nota esperada en Física?
(e) Si un estudiante tiene 95 en Física, ¿cuál es su nota esperada en Algebra?

13.36. La Tabla 13.21 muestra la tasa de natalidad en EE. UU. durante 1920-1980, en intervalos de 10 años.
(a) Representar los datos.
(b) Hallar una recta de mínimos cuadrados que ajuste esos datos.
(c) Calcular los valores de tendencia y compararlos con los verdaderos.

Tabla 13.21

Año	Tasa de natalidad por cada 1000 habitantes
1920	27.7
1930	21.3
1940	19.4
1950	24.1
1960	23.7
1970	18.4
1980	15.9

Fuente: National Center for Health Statistics.

(d) Predecir la tasa de natalidad en los años 1990 y 2000, y discutir las posibles causas de error en tal predicción.

13.37. La Tabla 13.22 recoge los porcentajes de la población de EE. UU. de 65 años o más, para los años 1890-1980.
(a) Representar los datos.
(b) Ajustar los datos con una recta de mínimos cuadrados.
(c) Calcular los valores de tendencia y compararlos con los verdaderos.
(d) Predecir el porcentaje de esas edades para los años 1990 y 2000, y discutir las posibles causas de error en esa predicción.
(e) ¿Cuándo se esperaría que el porcentaje alcance 25, 35 y 50 % y qué hipótesis hay que hacer para responder?

Tabla 13.22

Año	Porcentaje
1890	3.84
1900	4.05
1910	4.29
1920	4.67
1930	5.40
1940	6.85
1950	8.12
1960	9.30
1970	9.89
1980	11.35

Fuente: U.S. Bureau of the Census.

CURVAS DE MINIMOS CUADRADOS

13.38. Ajustar una parábola de mínimos cuadrados, $Y = a_0 + a_1X + a_2X^2$, a los datos de la Tabla 13.23.

Tabla 13.23

X	Y
0	2.4
1	2.1
2	3.2
3	5.6
4	9.3
5	14.6
6	21.9

13.39. El tiempo necesario para detener un coche tras percibir un peligro es el tiempo de reacción (el tiempo entre la percepción del peligro y la aplicación de los frenos) más el tiempo de frenada (lo que tarda en detenerse bajo la acción de los frenos). La Tabla 13.24. da la distancia D (en pies) que recorre antes de pararse un coche que circula a V millas por hora, a partir del instante en que se ha percibido el peligro.

(*a*) Representar los datos.
(*b*) Ajustar una parábola de mínimos cuadrados de la forma $D = a_0 + a_1 V + a_2 V^2$ a los datos.
(*c*) Estimar D cuando $V = 45$ mi/h y 80 mi/h.

Tabla 13.24

Velocidad V (mi/h)	Distancia de frenado D (pies)
20	54
30	90
40	138
50	206
60	292
70	396

13.40. La Tabla 13.25 presenta las poblaciones masculina y femenina de EE. UU. durante 1920-1980.

(*a*) Representar las diferencias entre esas dos poblaciones.
(*b*) Usando una ecuación apropiada, hallar una curva de mínimos cuadrados que ajuste esos datos.
(*c*) Estimar la diferencia para los años 1990 y 2000.
(*d*) Determinar en qué año habrá una proporción 2:1 de mujeres a hombres. Al determinar esto, ¿qué hipótesis hay que hacer?

Tabla 13.25

Año	Población masculina	Población femenina
1920	53.90	51.81
1930	62.14	60.64
1940	66.06	65.61
1950	75.19	76.14
1960	88.33	90.99
1970	98.93	104.31
1980	110.05	116.49

Fuente: U.S. Bureau of the Census.

13.41. Resolver el Problema 13.40 usando el cociente en vez de la diferencia entre poblaciones.

13.42. Resolver el Problema 13.37 con una parábola de mínimos cuadrados y comparar los resultados.

13.43. El número de bacterias por unidad de volumen en un cultivo tras X horas viene dado en la Tabla 13.26.

(*a*) Representar los datos en papel semilog, usando escala logarítmica para Y y escala aritmética para X.
(*b*) Ajustar una curva de mínimos cuadrados de la forma $Y = ab^X$ a los datos y explicar por qué esa ecuación particular debe dar buenos resultados.
(*c*) Comparar los valores de Y obtenidos de esa ecuación con los valores reales.
(*d*) Estimar el valor de Y cuando $X = 7$.

Tabla 13.26

Número de horas (X)	Numero de bacterias por unidad de volumen (Y)
0	32
1	47
2	65
3	92
4	132
5	190
6	275

13.44. En el Problema 13.43, mostrar cómo un gráfico en papel semilog puede ser utilizado para la obtención de la ecuación requerida sin recurrir al método de mínimos cuadrados.

CAPITULO 14

Teoría de la correlación

CORRELACION Y REGRESION

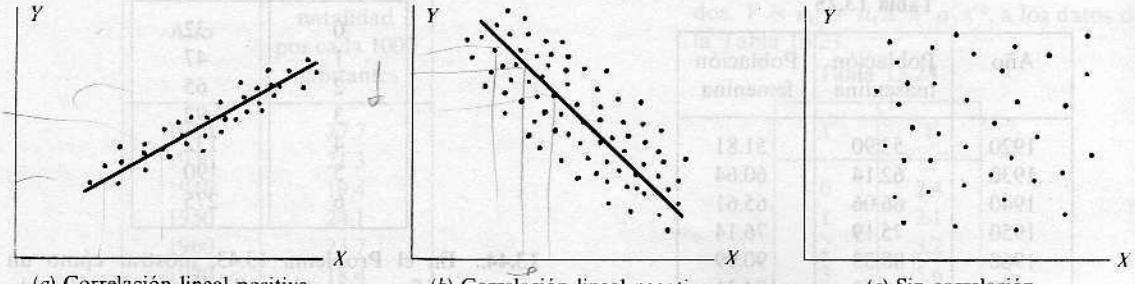
En el último capítulo hemos considerado el problema de la *regresión* o *estimación* de una variable (la variable dependiente) de una o más variables relacionadas (las variables independientes). En este capítulo tratamos el problema cercano de la *correlación*, o grado de interconexión entre variables, que intenta determinar *con qué precisión* describe o explica la relación entre variables una ecuación lineal o de cualquier otro tipo.

Si todos los valores de las variables satisfacen una ecuación exactamente, decimos que las variables están *perfectamente correlacionadas* o que hay *correlación perfecta* entre ellas. Así, las circunferencias C y los radios r de todos los círculos están perfectamente correlacionados porque $C = 2\pi r$. Si se lanzan dos dados 100 veces, no hay relación entre las puntuaciones de ambos dados (a menos que estén trucados) es decir, *no están en correlación*. Variables tales como el peso y la altura de las personas tienen una cierta correlación.

Cuando sólo están en juego dos variables, hablamos de *correlación simple* y *regresión simple*. En otro caso, se habla de *correlación múltiple* y *regresión múltiple*. Este capítulo considera sólo correlación simple. La correlación y regresión múltiples se analizarán en el Capítulo 15.

CORRELACION LINEAL

Si X e Y son las dos variables en cuestión, un *diagrama de dispersión* muestra la localización de los puntos (X, Y) sobre un sistema rectangular de coordenadas. Si todos los puntos del diagrama de dispersión parecen estar en una recta, como en las Figuras 14.1(*a*) y 14.1(*b*), la correlación se llama *lineal*. En tales casos, como ya hemos visto en el Capítulo 13, una ecuación lineal es adecuada a efectos de regresión (o estimación).



(*a*) Correlación lineal positiva (*b*) Correlación lineal negativa (*c*) Sin correlación

Figura 14.1.

Si Y tiende a crecer cuando X crece, como en la Figura 14.1(a), la *correlación* se dice *positiva*, o *directa*. Si Y tiende a decrecer cuando X crece, como en la Figura 14.1(b), la *correlación* se dice *negativa*, o *inversa*.

Si todos los puntos parecen estar sobre una cierta curva, la correlación se llama *no lineal*, y una ecuación no lineal será apropiada para la regresión, como hemos visto en el Capítulo 13. Es claro que la correlación no lineal puede ser positiva o negativa.

Si no hay relación entre las variables, como en la Figura 14.1(c), decimos que *no hay correlación* entre ellas.

MEDIDAS DE CORRELACION

Podemos determinar de forma *cualitativa* con qué precisión describe una curva dada la relación entre variables por observación directa del propio diagrama de dispersión. Por ejemplo, se ve que una recta es mucho más conveniente para describir la relación entre X e Y para los datos de la Figura 14.1(a) que para los de la Figura 14.1(b), porque hay menos dispersión relativa a la recta en la Figura 14.1(a).

Si hemos de enfrentarnos al problema de la dispersión de datos muestrales respecto de rectas o curvas de modo *cuantitativo*, será necesario definir *medidas de correlación*.

LA RECTA DE REGRESION DE MINIMOS CUADRADOS

Consideremos primero el problema de ver con qué calidad explica una recta la relación entre dos variables. Para ello, necesitaremos las ecuaciones de la recta de regresión de mínimos cuadrados obtenidas en el Capítulo 13. Tal como vimos, la recta de regresión de mínimos cuadrados de Y sobre X es

$$Y = a_0 + a_1 X \tag{1}$$

donde a_0 y a_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{2}$$

de las que se deduce

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \end{aligned} \tag{3}$$

Análogamente, la recta de regresión de X sobre Y es

$$X = b_0 + b_1 Y \tag{4}$$

donde b_0 y b_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum X &= b_0 N + b_1 \sum Y \\ \sum XY &= b_0 \sum X + b_1 \sum Y^2 \end{aligned} \tag{5}$$

obteniéndose

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2}$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} \tag{6}$$

Las ecuaciones (1) y (4) se pueden escribir, respectivamente, como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{y} \quad x = \left(\frac{\sum xy}{\sum y^2}\right)y \tag{7}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

Las ecuaciones de regresión son idénticas si y sólo si todos los puntos del diagrama de dispersión están en una recta. En tal caso hay una *correlación lineal perfecta* entre X e Y.

ERROR TIPICO DE ESTIMACION

Si denotamos por Y_{est} el valor de Y para valores dados de X, tal como se estima a partir de la ecuación (1), una medida de la dispersión respecto de la recta de regresión de Y sobre X viene proporcionada por la cantidad

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}} \tag{8}$$

que se llama el *error típico de estimación* de Y sobre X.

Si se usa la recta de regresión (4), un error típico de estimación análogo de la estimación de X sobre Y se define como

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{est})^2}{N}} \tag{9}$$

En general, $s_{Y.X} \neq s_{X.Y}$.

La ecuación (8) se puede formular

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \tag{10}$$

que puede ser más conveniente para el cálculo (véase Prob. 14.3). Existe una expresión similar para (9).

El error típico de estimación tiene propiedades análogas a las de la desviación típica. Por ejemplo, si construimos rectas paralelas a la de regresión de Y sobre X a distancias verticales respectivas $s_{Y.X}$, $2s_{Y.X}$, y $3s_{Y.X}$ de ella, encontraremos, si N es lo bastante grande, que estarían incluidos entre esas rectas aproximadamente el 68%, 95% y 99.7% de los puntos muestrales.

Igual que la desviación típica modificada

$$\hat{s} = \sqrt{\frac{N}{N-1}}\, s$$

era útil para pequeñas muestras, será útil un error típico de estimación modificado dado por

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}}\, s_{Y.X}$$

Por esta razón, algunos estadísticos prefieren definir (8) ó (9) con $N - 2$ en lugar de N en el denominador.

VARIACION EXPLICADA Y VARIACION INEXPLICADA

La *variación total* de Y se define como $\sum(Y - \bar{Y})^2$: esto es, la suma de los cuadrados de las desviaciones de los valores de Y respecto de la media Y. Como se ve en el Problema 14.7, eso se puede escribir

$$\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2 \tag{11}$$

El primer término de la derecha en la ecuación (11) se llama la *variación explicada*, mientras que el segundo se llama la *variación inexplicada* (porque las desviaciones $Y_{est} - \bar{Y}$ tienen un esquema definido mientras las desviaciones $Y - Y_{est}$ se comportan de modo caótico, impredictible). Resultados similares son válidos para la variable X.

COEFICIENTE DE CORRELACION

El coeficiente entre la variación explicada y la variación total se llama *coeficiente de determinación*. Si la variación explicada es cero (o sea, toda la variación es variación inexplicada), ese cociente es 0. Si la variación inexplicada es cero (o sea, toda la variación es explicada), el cociente es 1. En los demás casos, está entre 0 y 1. Como nunca es negativo, denotaremos ese cociente por r. La cantidad r, llamada *coeficiente de correlación*, viene dada por

$$r = \pm\sqrt{\frac{\text{variación explicada}}{\text{variación total}}} = \pm\sqrt{\frac{\sum(Y_{est} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}} \tag{12}$$

y varía entre -1 y $+1$. Se usan los signos $+$ y $-$ para las correlaciones positivas y negativas, respectivamente. Nótese que r es una cantidad adimensional, es decir, no depende de las unidades empleadas.

Usando las ecuaciones (8) y (11) y el hecho de que la desviación típica de Y es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \tag{13}$$

encontramos que la ecuación (12) se puede escribir, independientemente del signo, como

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \quad \text{o sea} \quad s_{Y.X} = s_Y\sqrt{1 - r^2} \tag{14}$$

Ecuaciones similares existen cuando se intercambian X e Y.

Para el caso de correlación lineal, la cantidad r es la misma tanto si es X como Y la variable independiente. Así pues, r es una buena medida de la correlación lineal entre dos variables.

OBSERVACIONES SOBRE EL COEFICIENTE DE CORRELACION

Las definiciones del coeficiente de correlación en (12) y (14) son completamente generales y se pueden usar tanto para relaciones lineales como no lineales, con la única diferencia de que Y se calcula de una ecuación de regresión no lineal en lugar de una lineal, y que se omiten los signos $+$ y $-$. En tal caso, la ecuación (8), que define el error típico de estimación, es perfectamente general. La (10), sin embargo, que sólo se aplica a regresión lineal, debe ser modificada. Si, por ejemplo, la ecuación de estimación es

$$Y = a_0 + a_1X + a_2X^2 + \cdots + a_{n-1}X^{n-1} \tag{15}$$

la ecuación (10) queda sustituida por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY - \cdots - a_{n-1}\sum X^{n-1}Y}{N} \tag{16}$$

En tal caso el *error típico de estimación modificado* (discutido previamente en este capítulo) es

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-n}}\, s_{Y.X}$$

donde la cantidad $N - n$ se llama el número de *grados de libertad*.

Hay que insistir en que en todo caso el valor calculado de r mide el grado de relación con referencia al tipo de ecuación que se adopta. Así pues, si se supone una ecuación lineal y (12) o (14) dan un valor de r próximo a cero, eso significa que no hay apenas correlación lineal entre las variables. No obstante, no quiere decir que no haya correlación en absoluto, pues puede haber una fuerte *correlación no lineal* entre ellas. En otras palabras, el coeficiente de correlación mide la

bondad del ajuste entre: (1) la ecuación adoptada y (2) los datos. A menos que se especifique lo contrario, el término *coeficiente de correlación* se usará para el *coeficiente de correlación lineal.*

Hemos de hacer constar que un coeficiente de correlación alto (o sea, cercano a 1 ó -1) no indica necesariamente una dependencia directa de las variables. Puede haber una alta correlación entre el número de libros publicados cada año y el número de tormentas cada año. Tales ejemplos constituyen lo que se llama *correlaciones sin sentido*, o *espúreas.*

FORMULAS MOMENTO-PRODUCTO PARA EL COEFICIENTE DE CORRELACION LINEAL

Si se supone una relación lineal entre dos variables, la ecuación (12) se convierte en

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \tag{17}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$ (véase Prob. 14.10). Esta fórmula, que da automáticamente el signo apropiado de r, se llama la *fórmula momento-producto* y muestra claramente la simetría entre X e Y.

Si escribimos

$$s_{XY} = \frac{\sum xy}{N} \qquad s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \tag{18}$$

entonces s_X y s_Y se reconocen como la desviación típica de las variables X e Y, mientras que s_X^2 y s_Y^2 son sus varianzas. La nueva cantidad s se llama la *covarianza* de X e Y. En términos de símbolos de (18), la fórmula (17) se reescribe

$$r = \frac{s_{XY}}{s_X s_Y} \tag{19}$$

Nótese que r no es sólo independiente de la elección de unidades de X e Y, sino también de la elección del origen.

FORMULAS CORTAS DE CALCULO

La fórmula (17) se puede escribir en la forma equivalente

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \tag{20}$$

que se usa con frecuencia al calcular r (véanse Probs. 14.15 y 14.16).

Para datos agrupados como en una *tabla de frecuencias de dos variables*, o en una *distribución de frecuencias de dos variables* (véase Prob. 14.17), conviene usar un *método de compilación* como en los capítulos previos. En tal caso, la fórmula (20) se escribe

$$r = \frac{N\sum f_X u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

(véase Prob. 14.18). Por conveniencia en los cálculos cuando se recurre a esa fórmula, se usa una *tabla de correlación* (véase Prob. 14.19).

Para datos agrupados, las (18) se expresan

$$s_{XY} = c_X c_Y \left[\frac{\sum f_X u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right) \right] \tag{22}$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} \tag{23}$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} \tag{24}$$

donde c_X y c_Y son las anchuras de intervalos de clase (supuestas constantes) de las variables X e Y. Nótese que (23) y (24) son equivalentes a la fórmula (11) del Capítulo 4.

La fórmula (19) es equivalente a (21), como se ve sin más que usar (22) a (24).

RECTAS DE REGRESION Y EL COEFICIENTE DE CORRELACION LINEAL

La ecuación de la recta de mínimos cuadrados $Y = a_0 + a_1 X$, la recta de regresión de Y sobre X, se puede escribir

$$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad \text{o sea} \qquad y = \frac{rs_Y}{s_X} x \tag{25}$$

Análogamente, la recta de regresión de X sobre Y, $X = b_0 + b_1 Y$, puede expresarse como

$$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad \text{o sea} \qquad x = \frac{rs_X}{s_Y} y \tag{26}$$

Las pendientes de las rectas en las ecuaciones (25) y (26) son iguales si y sólo si $r = \pm 1$. En tal caso las dos rectas son idénticas y hay correlación lineal perfecta entre X e Y. Si $r = 0$, las rectas son perpendiculares y no hay correlación lineal entre X e Y. Así pues, el coeficiente de correlación lineal mide la separación de ambas rectas de regresión.

Obsérvese que si (25) y (26) se escriben como $Y = a_0 + a_1 X$ y $X = b_0 + b_1 Y$, respectivamente, entonces $a_1 b_1 = r^2$ (véase Prob. 14.22).

CORRELACION DE SERIES EN EL TIEMPO

Si las variables X e Y dependen del tiempo, es posible que pueda existir una relación entre X e Y aun cuando no sea una dependencia directa y pueda producir «correlación espúrea». El coeficiente

de correlación se obtiene simplemente considerando los pares de valores (X, Y) correspondientes a varios tiempos y procediendo como de costumbre, haciendo uso de las fórmulas anteriores (véase Problema 14.28).

Es posible intentar correlacionar valores de una variable X en ciertos tiempos con valores correspondientes de X en tiempos anteriores. Tales correlaciones se llaman *autocorrelaciones*.

CORRELACION DE ATRIBUTOS

Los métodos descritos en este capítulo no nos capacitan para considerar la correlación de variables que sean de naturaleza no numérica, tales como los *atributos* de individuos (color del pelo, de los ojos, etc.). Para una discusión de la correlación de atributos, véase el Capítulo 12

TEORIA MUESTRAL DE LA CORRELACION

Los N pares de valores (X, Y) de dos variables pueden verse como muestras de una población de todos los pares posibles. Como están en juego dos variables, se llama una *población de dos variables*, que supondremos tiene una *distribución normal de dos variables*.

Podemos pensar en un coeficiente de correlación de población teórico, denotado por ρ, que se estima por el coeficiente de correlación r de la muestra. Contrastes de hipótesis o significación relativos a varios valores de ρ exigen conocer la distribución muestral de r. Para $\rho = 0$ esta distribución es simétrica, y se puede usar un estadístico con distribución de Student. Para $\rho \neq 0$, la distribución es sesgada y en tal caso una transformación debida a Fisher produce un estadístico que es aproximadamente normal. Los siguientes contrastes resumen los procedimientos implicados:

1. **Contraste de hipótesis** $\rho = 0$. Aquí usamos el hecho de que el estadístico

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{27}$$

tiene una distribución de Student con $\nu = N - 2$ grados de libertad (véanse Probs. 14.31 y 14.32).

2. **Contraste de hipótesis** $\rho = \rho_0 \neq 0$. Aquí usamos el hecho de que el estadístico

$$Z = \tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.1513\,\log_{10}\left(\frac{1+r}{1-r}\right) \tag{28}$$

donde $e = 2.71828...$, está casi normalmente distribuido con media y desviación típica dadas por

$$\mu_Z = \tfrac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513\,\log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right) \quad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{29}$$

Las ecuaciones (28) y (29) se pueden utilizar también para hallar límites de confianza para el coeficiente de correlación (véanse Probs. 14.33 y 14.34). La ecuación (28) se llama *transformación Z de Fisher*.

3. **Significación de una diferencia entre coeficiente de correlación.** Para determinar si dos coeficientes de correlación, r_1 y r_2, sacados de muestras de tamaños N_1 y N_2, respectivamente,

difieren significativamente uno de otro, calculamos Z_1 y Z_2 correspondientes a r_1 y r_2 usando (28). Y utilizamos entonces el hecho de que el estadístico de contraste

$$z = \frac{Z_1 - Z_2 - \mu_{Z1-Z2}}{\sigma_{Z1-Z2}} \tag{30}$$

donde

$$\mu_{Z1-Z2} = \mu_{Z1} - \mu_{Z2}$$

y

$$\sigma_{Z1-Z2} = \sqrt{\sigma_{Z1}^2 + \sigma_{Z2}^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

está normalmente distribuido (véase Prob. 14.35).

TEORIA MUESTRAL DE LA REGRESION

La ecuación de regresión $Y = a_0 + a_1 X$ se obtiene a partir de los datos de la muestra. A menudo estamos interesados en la correspondiente ecuación de regresión para la población de la que procede el muestreo. He aquí tres contrastes relativos a dicha población:

1. **Contraste de hipótesis** $a_1 = A_1$. Para contrastar la hipótesis de que el coeficiente de regresión a_1 es igual a cierto valor A_1 especificado, usamos el hecho de que el estadístico

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N - 2} \tag{31}$$

tiene distribución de Student con $N - 2$ grados de libertad. Esto se puede también utilizar para hallar intervalos de confianza para los coeficientes de regresión de la población a partir de los valores de la muestra (véanse Probs. 14.36 y 14.37).

2. **Contraste de hipótesis para valores de predicción.** Sea Y_0 la predicción para el valor de Y correspondiente a $X = X_0$ tal como se estima a partir de la ecuación de regresión muestral (o sea $Y_0 = a_0 + a_1 X_0$). Sea Y_p la predicción del valor de Y correspondiente a $X = X_0$ para la población. Entonces el estadístico

$$t = \frac{Y_0 - Y_p}{s_{Y.X}\sqrt{N + 1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N - 2} = \frac{Y_0 - Y_p}{\hat{s}_{Y.X}\sqrt{1 + 1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{32}$$

tiene distribución de Student con $N - 2$ grados de libertad. De donde pueden hallarse límites de confianza para las predicciones de los valores poblacionales (véase Prob. 14.38).

3. **Contraste de hipótesis para predicciones de valores medios.** Sea Y_0 el valor de predicción de Y correspondiente a $X = X_0$ estimado a partir de la ecuación de regresión muestral (o sea, $Y_0 = a_0 + a_1 X_0$). Denotemos por $\bar{Y}_p$ la predicción del *valor medio* de Y correspondiente a $X = X_0$ para la población. Entonces el estadístico

$$t = \frac{Y_0 - \bar{Y}_p}{s_{Y.X}\sqrt{1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N - 2} = \frac{Y_0 - \bar{Y}_p}{\hat{s}_{Y.X}\sqrt{1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{33}$$

tiene distribución de Student con $N - 2$ grados de libertad. De ahí se pueden reducir límites de confianza para las predicciones de los valores medios de la población (véase Prob. 14.39).

PROBLEMAS RESUELTOS

DIAGRAMA DE DISPERSION Y RECTAS DE REGRESION

14.1. La Tabla 14.1 da en pulgadas las respectivas alturas X e Y de una muestra de 12 padres y sus hijos mayores.

(*a*) Construir un diagrama de dispersión.
(*b*) Hallar la recta de regresión de mínimos cuadrados de Y sobre X.
(*c*) Hallar la recta de regresión de mínimos cuadrados de X sobre Y.

Tabla 14.1

Altura X del padre	65	63	67	64	68	62	70	66	68	67	69	71
Altura Y del hijo	68	66	68	65	69	66	68	65	71	67	68	70

Solución

(*a*) El diagrama de dispersión se obtiene marcando los puntos (X, Y) en un sistema rectangular de coordenadas, como ilustra la Figura 14.2.

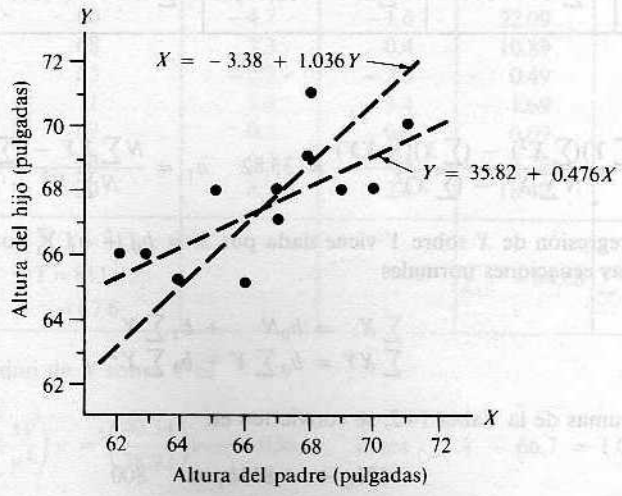


Figura 14.2.

(*b*) La recta de regresión de Y sobre X viene dada por $Y = a_0 + a_1X$, donde a_0 y a_1 se obtienen resolviendo las ecuaciones normales

$$\sum Y = a_0 N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

Las sumas se indican en la Tabla 14.2, de la que las ecuaciones normales pasan a ser

$$12a_0 + 800a_1 = 811$$
$$800a_0 + 53{,}418a_1 = 54{,}107$$

y de aquí concluimos que $a_0 = 35.82$ y $a_1 = 0.476$, y por tanto $Y = 35.82 + 0.476X$. El gráfico de esta ecuación aparece en la Figura 14.2.

Tabla 14.2

X	Y	X^2	XY	Y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53{,}418$	$\sum XY = 54{,}107$	$\sum Y^2 = 54{,}849$

Otro método

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = 35.82 \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = 0.476$$

(*c*) La recta de regresión de X sobre Y viene dada por $X = b_0 + b_1 Y$, donde b_0 y b_1 se obtienen resolviendo las ecuaciones normales

$$\sum X = b_0 N + b_1 \sum Y$$
$$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

Usando las sumas de la Tabla 14.2, se convierten en

$$12b_0 + 811b_1 = 800$$
$$811b_0 + 54{,}849b_1 = 54{,}107$$