REGRESIÓN CON DOS VARIABLES: ESTIMACIÓN DE INTERVALOS Y PRUEBA DE HIPÓTESIS

Teoría de la estimación:

- Estimación puntual
- Estimación por intervalos

ESTIMACIÓN DE INTERVALOS: IDEAS BÁSICAS[1]

Lo que se busca es averiguar qué tan cerca está el valor estimado del valor verdadero. Para ello se buscan dos números positivos δ y α, este último situado entre 0 y 1 tal que la probabilidad de que el intervalo aleatorio $(\hat{\beta}_i - \delta, \hat{\beta}_i + \delta)$ contenga el verdadero β_2 sea 1- α:

[1] Una pregunta que se deriva al estimar los parámetros es ***¿qué tan confiables son?*** Recuerde que en estadística la confiabilidad de un estimador se mide por su error estándar.

Intervalo de confianza

$$\Pr\,(\underbrace{\hat{\beta}_i - \delta \ \leq \beta_i \ \leq \ \hat{\beta}_i + \delta}_{\text{Límites de confianza}}) = 1-\alpha$$

1- α: coeficiente de confianza.

α : Nivel de significancia o probabilidad de cometer error tipo I

ERROR TIPO I: CONSISTE EN RECHAZAR UNA HIPÓTESIS VERDADERA.
ERROR TIPO II: CONSISTE EN NO RECHAZAR UNA HIPÓTESIS FALSA.

No olvide: minimizar un error, incrementa el otro. Depende de usted la decisión que tome.

PARA TENER PRESENTE:

- El intervalo no dice que la probabilidad que β_i se encuentre entre los límites de confianza sea 1- α.

- Se interpreta como la probabilidad que el intervalo aleatorio ***CONTENGA*** el verdadero valor de β_i es 1- α.

- Es un intervalo aleatorio, esto significa que varía de una muestra a otra debido a que se basa en $\hat{\beta}_i$, quien por naturaleza es aleatorio.

- Los enunciados derivados deben entenderse en un sentido de largo plazo o de muestreo repetido. Al construir intervalos de confianza en el 1- α de los casos el intervalo contendrá el verdadero valor del parámetro.

- El intervalo se mantiene aleatorio mientras $\hat{\beta}_i$ sea desconocido. Para una muestra específica al obtenerse el valor numérico de $\hat{\beta}_i$, el intervalo deja de ser aleatorio y queda fijo. Por ende la probabilidad de contener el verdadero β_i será 1 ó 0.

NUNCA LEEA EL INTERVALO COMO LA PROBABILIDAD DE QUE EL VERDADERO VALOR DE β_i SE ENCUENTRE O CAIGA EN EL INTERVALO ES 1- α

INTERVALO DE CONFIANZA PARA LOS COEFICIENTES DE REGRESIÓN β_1 Y β_2

Del capítulo 4 recuerde que:

$$U_i \sim N(0, \sigma^2) \qquad \hat{u}_i \sim N(0, \sigma^2)$$

Si se conoce la verdadera varianza poblacional (σ^2) se utiliza la distribución normal:

$$Z = \frac{(\hat{\beta}_2 - \beta_2)}{ee(\hat{\beta}_2)} \qquad ee(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

Pero esto rara vez ocurre, por tanto está determinada por el estimador insesgado $\hat{\sigma}^2$ y es utiliza una distribución t:

$$t = \frac{(\hat{\beta}_2 - \beta_2)}{ee(\hat{\beta}_2)} \qquad ee(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

$\boldsymbol{Para\ \hat{\beta}_2}$**:**

$$\Pr(-t_{\alpha/2} \leq t \ \leq t_{\alpha/2}) = 1 - \alpha$$

$$\Pr\left[-t_{\alpha/2} \leq \frac{(\hat{\beta}_2 - \beta_2)}{ee(\hat{\beta}_2)} \ \leq t_{\alpha/2}\right] = 1 - \alpha$$

$$\Pr[\hat{\beta}_2 - t_{\alpha/2}ee(\hat{\beta}_2) \leq \beta_2 \ \leq \hat{\beta}_2 + t_{\alpha/2}ee(\hat{\beta}_2)] = 1 - \alpha$$

$$\hat{\beta}_2 \pm t_{\alpha/2}ee(\hat{\beta}_2)$$

Para $\boldsymbol{\widehat{\beta}_1}$**:**

$$\Pr\left[\hat{\beta}_1 - t_{\alpha/2} ee(\hat{\beta}_1) \leq \beta_1 \ \leq \hat{\beta}_1 + t_{\alpha/2} ee(\hat{\beta}_1)\right] = 1 - \alpha$$

$\hat{\beta}_1 \pm t_{\alpha/2} ee(\hat{\beta}_1)$ con probabilidad $1 - \alpha$

PRUEBA DE HIPÓTESIS

Una prueba de hipótesis no es otra cosa que un procedimiento mediante el cual se rechaza o no la hipótesis nula. Lo que se busca es saber si es compatible o incompatible una observación con alguna o algunas hipótesis planteadas.

¿QUÉ ES UNA HIPÓTESIS?

R: ES UN ESTAMENTO O DECLARACIÓN QUE SE REALIZA ACERCA DE UN PARÁMETRO QUE SE DESEA PROBAR

En estadística la hipótesis planteada se denomina **HIPÓTESIS NULA** ($\boldsymbol{H_0}$) y frente a ésta se plantea la **HIPÓTESIS ALTERNATIVA o MANTENIDA** ($\boldsymbol{H_1}$). La prueba de hipótesis puede ser simple o compuesta. La teoría estadística en materia de hipótesis se encarga de diseñar reglas que permitan decidir si rechazar o no una hipótesis nula, para ello existen dos métodos mutuamente complementarios: el intervalo de confianza y la prueba de significancia.

$\boldsymbol{H_0}: \beta_i = 1$

$\boldsymbol{H_1}: \beta_i > (<)\ 1 \longrightarrow$ una cola

$\beta_i \neq 1 \longrightarrow$ dos colas

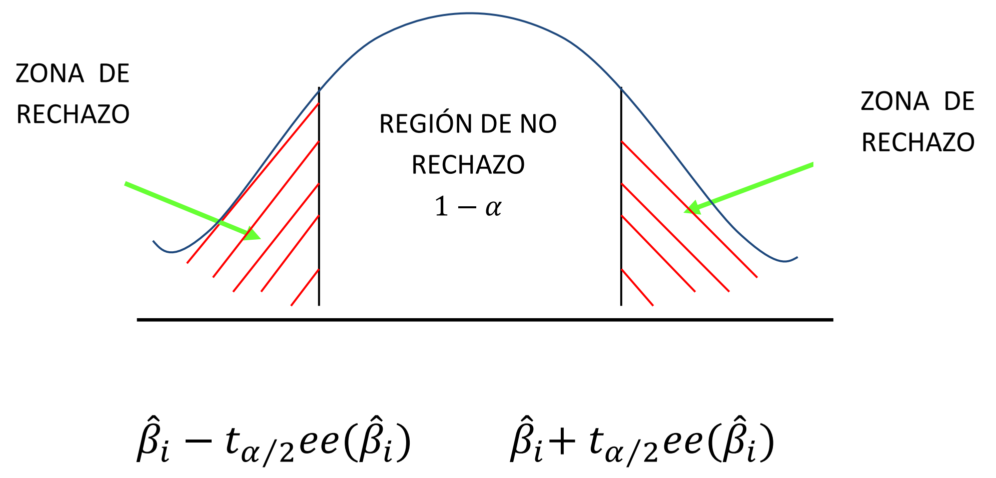
Para una comprensión completa de las pruebas de hipótesis, hay que recordar que los parámetros β_i son características desconocidas de la población, que nunca se conocerán con certeza

1. MÉTODO DEL INTERVALO DE CONFIANZA: conocida también como la prueba de dos colas, utiliza la distribución t de Student. El procedimiento es el siguiente:

a) Planteamiento de la hipótesis:

$\boldsymbol{H_0}: \beta_i = \beta$
$\boldsymbol{H_1}: \beta_i \neq \beta$

NOTA: recuerde que su intervalo de confianza será tan bueno como los supuestos subyacentes empleados para construirlo.



b) Estimación del intervalo de confianza:

$$\Pr\left[\hat{\beta}_i - t_{\alpha/2} ee(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2} ee(\hat{\beta}_i)\right] = 1-\alpha$$

c) Regla de decisión: rechace H_0 si β se contiene en una zona de rechazo. Esto significa que el hallazgo es estadísticamente significativo.

CUANDO NO SE RECHAZA $H_0 H_0$ SIGNIFICA QUE EL HALLAZGO NO ES ESTADÍSTICAMENTE SIGNIFICATIVO.

Ejemplo 1: Pruebe mediante un intervalo de confianza del 95% que β_2 es estadísticamente cero.

Impacto de los gastos en publicidad. Wall Street Journal 1984.

Dependent Variable: IMPRESIONES
Method: Least Squares
Sample: 1 21
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.16269	7.089479	3.126139	0.0056
GASTO	0.363174	0.097120	3.739425	0.0014

R-squared	0.423951	Mean dependent var	40.46667
Adjusted R-squared	0.393633	S.D. dependent var	30.18061
S.E. of regression	23.50152	Akaike info criterion	9.242400
Sum squared resid	10494.11	Schwarz criterion	9.341878

Log likelihood	-95.04520	Hannan-Quinn criter.	9.263989
F-statistic	13.98330	Durbin-Watson stat	2.371839
Prob(F-statistic)	0.001389		

Procedimiento:

a) Planteamiento de la hipótesis:

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

b) Estimación del intervalo de confianza:

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2} ee(\hat{\beta}_2) \leq \boldsymbol{\beta_2} \leq \hat{\beta}_2 + t_{\alpha/2} ee(\hat{\beta}_2)\right] = 1 - \alpha$$

$$[0.363174 - 2.093 * 0.097120 \leq \boldsymbol{\beta_2} \leq 0.363174 + 2.093 * 0.097120]$$

$$[0.363174 - 0.203272 \leq \boldsymbol{\beta_2} \leq 0.363174 + 0.203272]$$

$[0.159905 \leq \boldsymbol{\beta_2} \leq 0.566446]$

[0.159905; 0.566446]

c) Regla de decisión: Se rechaza la hipótesis nula, es decir, β_2 es estadísticamente distinto de cero, por tanto el hallazgo es estadísticamente significativo.

Salida del Eviews 7.0:

Coefficient Confidence Intervals
Sample: 1 21
Included observations: 21

		95% CI	
Variable	Coefficient	Low	High
C	22.16269	7.324244	37.00114
GASTO	0.363174	0.159899	0.566449

2. MÉTODO DE LA PRUEBA DE SIGNIFICANCIA[2] (TEST DE SIGNIFICANCIA):

Una prueba de significancia es un procedimiento mediante el cual se utilizan los resultados muestrales para verificar la verdad o falsedad de una hipótesis nula. Lo fundamental detrás del test de significancia es el ***estadístico de prueba.*** La metodología es la siguiente:

a) Planteamiento de la hipótesis:

$H_0: \beta_i = \beta$
$H_1: \beta_i \neq \beta$

[2] Desarrollado independientemente por R. Fisher y conjuntamente por Neyman y Pearson

b) Selección del estadístico de prueba:

$$t_c = \frac{(\hat{\beta}_i - \beta_i)}{ee(\hat{\beta}_i)} \sim t_{n-k}$$

n= tamaño de la muestra.   k= número de parámetros a estimar

c) Aplicar la regla de decisión:

$t_c > t_\alpha$ RECHAZAR H_0

$t_c < t_\alpha$ NO RECHAZAR H_0

Si se tiene a mano el P-value:

$P - value > \alpha$ NO RECHAZAR H_0

$P - value < \alpha$ RECHAZAR H_0

d) Establecer la conclusión.

Ejemplo 2: Pruebe mediante un test de significancia que $\beta_2 = 0$

Relación tasa de interés e inflación en Venezuela 1982-2005

Dependent Variable: TI
Method: Least Squares
Sample: 1982 2005
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.93345	3.566080	3.907217	0.0008
INF	0.219672	0.089295	2.460058	0.0222

R-squared	0.215739	Mean dependent var	21.03333
Adjusted R-squared	0.180091	S.D. dependent var	11.33273
S.E. of regression	10.26165	Akaike info criterion	7.574361
Sum squared resid	2316.634	Schwarz criterion	7.672532
Log likelihood	-88.89233	Hannan-Quinn criter.	7.600405
F-statistic	6.051885	Durbin-Watson stat	0.970288
Prob(F-statistic)	0.022218		

Procedimiento:

a) Planteamiento de la hipótesis:

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

b) Selección del estadístico de prueba:

$$t_c = \frac{(\hat{\beta}_i - \beta_i)}{ee(\hat{\beta}_i)} \sim t_{n-k}$$

$t_c = 2.460058$

$t_\alpha = 2.074$

c) Aplicar la regla de decisión:

$t_c > t_\alpha$ RECHAZAR H_0

P-value = 0.0222
α = 0.05

$P - value < \alpha$ RECHAZAR H_0

d) Conclusión: se rechaza la hipótesis nula lo que significa que β_2 es estadísticamente distinto de cero, es decir, β_2 es estadísticamente significativo.

CARACTERÍSTICAS QUE HACEN AL tc UN ESTADÍSTICO RAZONABLE:

1)Como el error estándar de $\hat{\beta}_i$ es siempre positivo, t_c tiene el mismo signo que $\hat{\beta}_i$

2) Para un valor dado del error estándar de $\hat{\beta}_i$, un valor mayor de $\hat{\beta}_i$ lleva a un valor mayor de t_c .Si $\hat{\beta}_i$ se vuelve más negativo, ocurre lo mismo con t_c .

TOME EN CUENTA:

- Por naturaleza las hipótesis no se aceptan. Esto se debe a que no existe evidencia estadística suficiente que permita rechazarla. Consecuentemente la aceptación de una hipótesis nula siempre debe tener presente que puede existir otra hipótesis nula compatible con los datos.

"… de la misma manera que una corte pronuncia un veredicto de "no culpable" en lugar de decir "inocente", así la conclusión de un estadístico de prueba es la de "no rechazar "en lugar de aceptar". Kmenta (1971)

- Una hipótesis nula de uso frecuente es $H_0: \beta_i = 0$. Prueba si Y está relacionada con X, mediante un intervalo de confianza o un test de significación. Sin embargo, estas pruebas se pueden obviar al adopta la regla práctica "2-t":

SI EL NÚMERO DE GRADOS DE LIBERTAD ES NO MENOR DE 20 Y EL NIVEL DE SIGNIFICANCIA SE FIJA EN 5%, ENTONCES LA $\boldsymbol{H_0: \beta_i = 0}$ SE PUEDE RECHAZAR SI EL VALOR DEL t CALCULADO ES SUPERIOR A DOS (tc > 2) EN VALOR ABSOLUTO.

- Es muy importante que el investigador plantee sus hipótesis antes de llevar a cabo la investigación.

- Rechazar o no una hipótesis nula depende en gran medida del α seleccionado, el nivel de significancia o probabilidad de cometer error tipo I.

LA SELECCIÓN DEL NIVEL DE SIGNIFICANCIA NO ES UN HECHO SAGRADO DEPENDERÁ DEL TIPO DE EXPERIMENTO Y LAS IMPLICACIONES QUE CONLLEVE. RECUERDESE QUE REDUCIR LA PROBABILIDAD DE COMETER ERROR TIPO I AUMENTA LA PROBABILIDAD DE COMETER ERROR TIPO II.

- EL P-VALUE NO ES UN ESTADÍSTICO DE PRUEBA, ES UNA PROBABILIDAD ASOCIADA AL ESTADÍSTICO DE PRUEBA. También se conoce como el nivel observado o exacto de significancia, en realidad es la verdadera probabilidad de cometer error tipo I.

TÉCNICAMENTE EL P-VALUE ES EL NIVEL MÁS BAJO AL CUAL SE PUEDE RECHAZAR LA HIPÓTESIS NULA.

- No debe confundirse la significancia estadística con la significancia económica o práctica.

- Usando el intervalo de confianza, el test de significancia o el P-value la conclusión a la que se llegue debe ser la misma.

Ejercicio 1: La siguiente regresión muestra la relación entre salario por hora en función de la educación, medida en años de estudio. Demuestre mediante un intervalo de confianza y un test de significancia individual la significancia estadística de la educación.

$$\widehat{Wage} = 0.284 + 0.92educ$$
$$(.104) \quad (.007)$$

N=526. Coeficiente de determinación= 0.316

Ejercicio 2: El desempeño de los estudiantes en determinada escuela se mide por el porcentaje de estudiantes que obtienen una puntuación aprobatoria en la prueba estandarizada de matemática del décimo grado (math10) en el programa de evaluación de Michigan (MEAP: Michigan Educational Assessment Program). El tamaño de le escuela se mide por el número de estudiantes inscritos (enroll). Pruebe la significancia estadística de la variable dependiente. ¿Cuál es su conclusión?

$$\widehat{math}10 = 2.274 - 0.0002enroll$$
$$(6.113) \quad (0.00010)$$

N=408. Coeficiente de determinación= 0.0541

Salida del R: regresión que muestra la relación existente entre el PIB real per cápita y el agregado monetario M1.

```
Call:
lm(formula = PIBrpc ~ M, data = Informal)
Residuals:
Min       1Q   Median       3Q      Max
-0.20895 -0.09612 -0.03246  0.03688  0.43700

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0729     0.1564  13.256 5.74e-12 ***
M            -1.1063     0.3470  -3.188  0.00425 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1615 on 22 degrees of freedom
(1 observation deleted due to missingness)  Multiple R-squared: 0.316,	Adjusted R-squared: 0.2849
F-statistic: 10.17 on 1 and 22 DF,  p-value: 0.004247
```

PREDICCIÓN: el fin último de todo modelo econométrico es hacer predicción, esta puede ser:

a) Predicción media: la predicción del valor de la media condicional de Y correspondiente a un valor escogido de X.

Ejemplo 3: Predecir el gasto en consumo promedio para un PIB de 6062 mil millones de dólares.

Gasto en consumo personal en función del PIB real de USA. 1982-1996

Dependent Variable: CONSUMO
Method: Least Squares
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-184.0780	46.26198	-3.979034	0.0016
PIB	0.706408	0.007827	90.24707	0.0000
R-squared	0.998406	Mean dependent var		3964.087
Adjusted R-squared	0.998284	S.D. dependent var		489.6614

$$\widehat{Consumo} = -184.07 + 0.70PIB$$

$$\widehat{Consumo} = -184.07 + 0.70(6062)$$

$$\widehat{Consumo} = 4059.33$$

b) Usar predicción para evaluar el modelo: consiste en regresar el modelo nuevamente dejando por fuera las últimas observaciones.

Ejemplo 4: Productividad y datos relacionados, sector negocios 1959-1997. Tabla 3.6 D. Gujarati

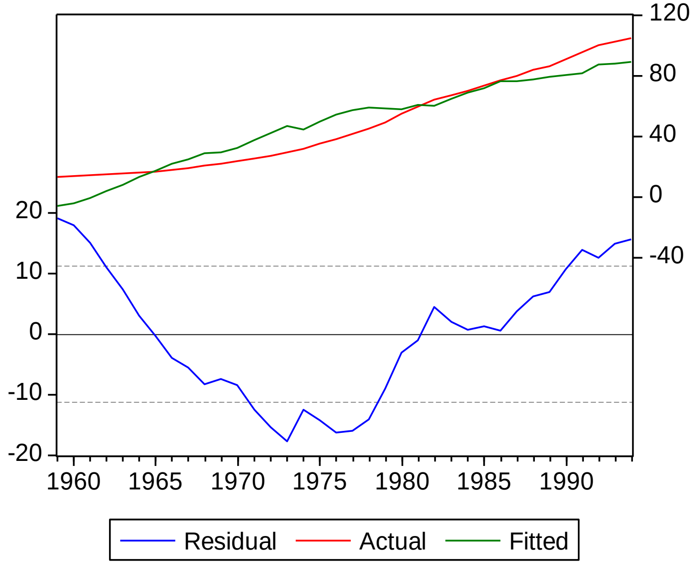
Dependent Variable: CSN
Method: Least Squares
Sample(adjusted): 1959 1994
Included observations: 36 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-101.1226	10.08997	-10.02209	0.0000
PSN	1.885811	0.125233	15.05841	0.0000

R-squared	0.869610	Mean dependent var	48.19167
Adjusted R-squared	0.865775	S.D. dependent var	30.58127
S.E. of regression	11.20399	Akaike info criterion	7.724369

$$CSN = -101.12 + 1.88PSN$$

Año	CSN observado	CSN estimado
1995	106.8	88.76
1996	110.7	93.83
1997	114.9	97.03



Lecturas recomendadas:

- ✓ Gujarati, D. (2003). Econometría. 4ta. Edición McGraw Hill. Capítulo 5.
- ✓ Novales, A. (1993). Econometría. 2da. Edición. McGraw Hill Interamericana. Capítulo 4.
- ✓ Wooldridge, J. (2010). Introducción a la econometría un enfoque moderno. Capítulos 4 y 5.

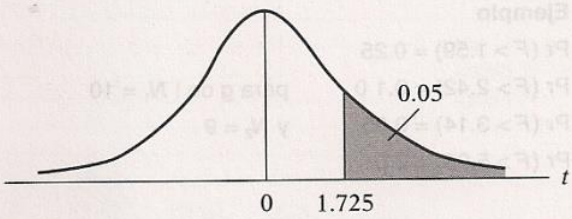
TABLA D.2 PUNTOS PORCENTUALES DE LA DISTRIBUCIÓN *t*

Ejemplo

Pr ($t > 2.086$) = 0.025

Pr($t > 1.725$) = 0.05 para g de l = 20

Pr ($|t| > 1.725$) = 0.10



g de l \ Pr	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Nota: La probabilidad más baja que aparece en el encabezamiento de cada columna es el área en una cola; la probabilidad más alta es el área en ambas colas.

Fuente: De E. S. Pearson y H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3a. ed., tabla 12, Cambridge University Press, Nueva York, 1966. Reproducido con permiso de los editores y de los depositarios de *Biometrika*.