Clase 10: La distribución de Student

Matías Carrasco

5 de octubre de 2019

Índice

1. $\overline{X}_n$ y S_n son independientes

En lo que resta del curso vamos a enfocarnos en estadística de poblaciones normales. Es decir, supondremos que los datos se ajustan a la distribución normal, y veremos que los estimadores de μ y σ^2 tienen distribuciones que podemos calcular.

Vamos a probar primero una propiedad sorprendente que tienen las muestras normales, y es que la media muestral $\overline{X}_n$ y el desvío muestral S_n son independientes. Una demostración relativamente sencilla se puede hacer aplicando la fórmula de cambio de variable, pero es casi inmediata usando la simetría rotacional de la distribución normal.

Supongamos por el momento que $X_1,\dots,X_n$ son i.i.d. con distribución normal estándar. La densidad conjunta (en dimensión n) es

$$p(x)=\prod_{i=1}^{n}\varphi(x_i)=\frac{1}{(2\pi)^{n/2}}e^{-\|x\|^2/2},$$

en donde hemos usado la notación

1. $x=(x_1,\dots,x_n)$ es un vector en $\mathbb{R}^n$;
2. $\|x\|^2=\sum_{i=1}^{n}x_i^2$ es la norma al cuadrado de x, que representa la distancia de x al origen elevada al cuadrado.

Al igual que en dimensión 2, vemos que la densidad $p(x)$ depende solamente de la distancia al origen, y por lo tanto presenta la misma simetría rotacional.

Comencemos por una observación trivial. Si proyectamos ortogonalmente el vector

$$(X_1,\dots,X_n)$$

sobre el plano generado por las primeras $n-1$ coordenadas, obtenemos el vector

$$(X_1,\ldots,X_{n-1},0),$$

cuya distancia al origen esta dada por

$$R^2 = X_1^2+\cdots+X_{n-1}^2.$$

Notar que como X_n es independiente de $X_1,\ldots,X_{n-1}$, entonces X_n y R^2 son independientes.

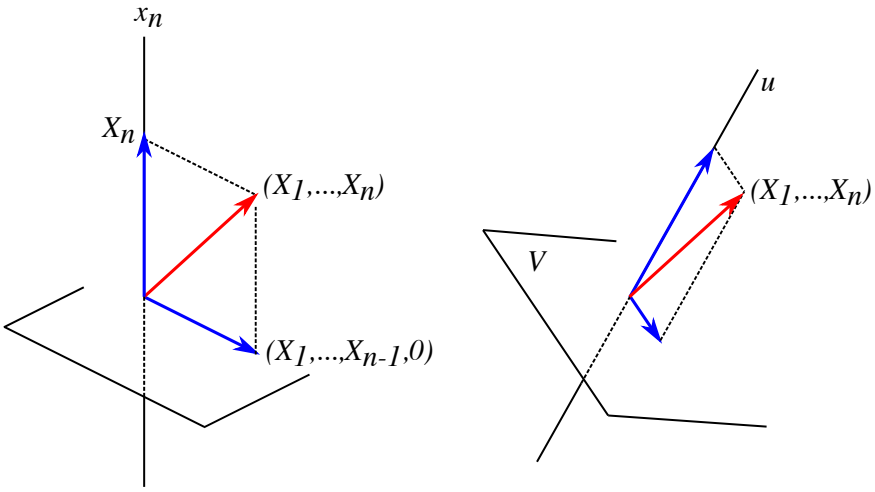


Figura 1: Proyección de $(X_1,\ldots,X_n)$. A la izquierda sobre el plano de las primeras $n-1$ coordenadas, a la derecha sobre una dirección dada u y su complemento ortogonal V.

Lo mismo ocurre en cualquier otra dirección. Si elegimos una dirección representada por un vector unitario u, la proyección del vector $X=(X_1,\ldots,X_n)$ sobre u tiene coeficiente $X\cdot u$ en donde el punto indica el producto escalar de vectores. La proyección sobre el complemento ortogonal V de u es entonces

$$X_V = X-(X\cdot u)u.$$

Por la simetría rotacional, las variables $X\cdot u$ y $R_V^2 = \|X_V\|^2$ son independientes. Más aún, $X\cdot V$ tiene distribución $N(0,1)$ y R_V^2 tiene la misma distribución que $R^2 = X_1^2+\cdots+X_{n-1}^2$.

Llamemos $\{e_1,\ldots,e_n\}$ a la base canónica de $\mathbb{R}^n$. Si tomamos

$$u = \frac{1}{\sqrt{n}}(1,\ldots,1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i,$$

entonces la proyección de X es

$$X\cdot u = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \sqrt{n}\overline{X}_n.$$

Además, el espacio V ortogonal a u es

$$V = \{x\in\mathbb{R}^n : x\cdot u = 0\} = \left\{x\in\mathbb{R}^n : \sum_{i=1}^n x_i = 0\right\},$$

y la proyección de X sobre V es

$$X_V = X-(\sqrt{n}\overline{X}_n)u = \sum_{i=1}^n (X_i-\overline{X}_n)e_i$$

cuya norma al cuadrado es

$$R_V^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = (n-1)S_n^2.$$

Hemos probado entonces que $\overline{X}_n$ y S_n son independientes. Más aún, podemos decir que $(n-1)S_n^2$ tiene la misma distribución que R^2.

Si $X_1, \ldots, X_n$ son i.i.d. con distribución $N(\mu, \sigma^2)$ en lugar de estándar, basta observar que

$$X_i = \sigma Z_i + \mu, \quad i = 1, \ldots, n$$

de modo que

$$\overline{X}_n = \sigma \overline{Z}_n + \mu, \quad S_n^2(X) = \sigma^2 S_n^2(Z).$$

De lo anterior sabemos que $\overline{Z}_n$ y $S_n^2(Z)$ son independientes, por lo que también lo son $\overline{X}_n$ y $S_n^2(X)$.

Sean $X_1, \ldots, X_n$ i.i.d. con distribución $N(\mu, \sigma^2)$. Entonces $\overline{X}_n$ y S_n^2 son independientes. Más aún:

1. $\overline{X}_n$ tiene distribución $N(\mu, \sigma^2/n)$;
2. $(n-1)S_n^2/\sigma^2$ tiene la misma distribución que la suma de $n-1$ cuadrados de normales independientes.

2. Las distribución *t* de Student

Recordar que si $Z_1, \ldots, Z_k$ son variables aleatorias independientes con distribución $\mathrm{N}(0,1)$, entonces la suma de sus cuadrados,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

tiene (por definición) distribución *chi-cuadrado con k grados de libertad*.

En la sección anterior probamos que si $X_1, \ldots, X_n$ son i.i.d. con distribución $N(\mu, \sigma^2)$, entonces $(n-1)S_n^2/\sigma^2$ tiene la misma distribución que la suma de los cuadrados de $n-1$ normales estándar independientes. Pero esta es precisamente la distribución $\chi^2(n-1)$.

Distribución de S_n^2

La distribución de $(n-1)S_n^2/\sigma^2$ es χ^2 con $n-1$ grados de libertad.

La *distribución t de Student* con k grados de libertad es la distribución de una variable T que se puede escribir como

$$T = \frac{Z}{\sqrt{V/k}} = Z\sqrt{\frac{k}{V}},$$

en donde

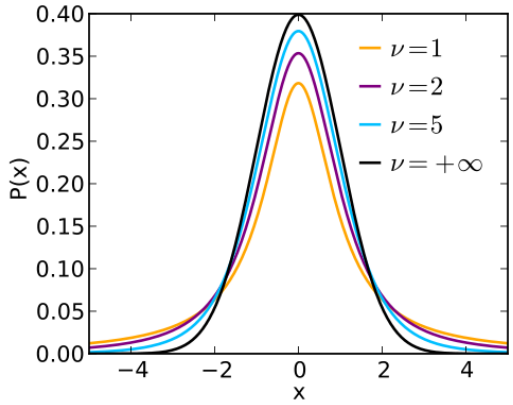


Figura 2: Se muestra las densidades t_ν para varios valores de ν. En el gráfico también se muestra la densidad normal estándar que corresponde a t_ν con $\nu = +\infty$.

- Z es una variable $\mathrm{N}(0,1)$;
- V tiene distribución chi-cuadrado con k grados de libertad;
- Z y V son independientes.

La distribución de Student tiene función de densidad dada por

$$p_T(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}},$$

en donde k es el número de grados de libertad y Γ es la función Gamma. Al igual que con la χ^2, la fórmula de la densidad no la usaremos nunca. Calcularemos probabilidades usando las tablas de la t.

Notar la semejanza con el gráfico de la densidad normal. A pesar de la apariencia, las colas de la distribución t son un poco más grandes que las de la normal estándar. A medida que los grados de libertad k crecen, la $t(k)$ se aproxima más y más a la normal estándar.

Consideremos los estimadores muestrales

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \;\text{ y }\; S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Podemos resumir entonces lo hecho en esta clase del siguiente modo:

- $\overline{X}_n$ y S_n^2 son independientes;
- $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$;
- y $\frac{\sqrt{n}(\overline{X}_n-\mu)}{S_n}$ tiene distribución de Student con $n-1$ grados de libertad.

La tabla al final de las notas muestra algunos valores críticos de la t de Student. Se debe leer de la misma forma que la tabla de la χ^2. Por ejemplo, el valor crítico asociado a la probabilidad 0.05 y 10 grados de libertad es 1.812. Esto lo escribimos $t_{10}(0.05) = 1.812$. En general, $t_k(\alpha)$ es el valor que verifica $\mathbf{P}(t_k \geq t_k(\alpha)) = \alpha$.

3. El test t de Student

¿Qué podemos hacer si n es relativamente pequeño y la varianza σ^2 es desconocida? En el caso de datos normales, podemos utilizar un test t (o test de Student).

Un test t es aquel en el cual el estadístico tiene distribución t de Student bajo la hipótesis nula. Como hemos visto en las clases anteriores, si $X_1, \ldots, X_n$ son i.i.d. con distribución normal $N(\mu, \sigma^2)$, entonces

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$$

tiene distribución t de Student con $n-1$ grados de libertad. Aquí S_n denota el desvío estándar muestral

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

que se calcula a partir de los datos observados (dividimos entre $n-1$ y no entre n).

Ejemplo 1

Una compañía fabrica barras de jabón que se supone pesan en promedio 280 gramos. Se asume que el peso de una barra de jabón es una variable aleatoria con distribución normal. Se toma una muestra de tamaño $n = 20$ para el control de calidad y se obtiene

$$\overline{X}_n = 289 \text{ gr}, \quad S_n = 22 \text{ gr}.$$

¿Es el peso de las barras compatible con el peso de la etiqueta?

Como en los ejemplos anteriores, diseñamos un TdH para responder a esta pregunta.

1. *Elegir la hipótesis nula H_0*. El fabricante afirma que las barras pesan en promedio 280 gr. Si denotamos por μ la esperanza del peso de una barra de jabón, podemos escribir la hipótesis nula como $H_0 : \mu = 280$.

2. *Decidir si H_A es a una o a dos colas*. Nos interesa que el jabón pese lo que indica la etiqueta. Si el peso es menor o mayor significa que algo está fallando en el proceso de fabricación. Así que $H_A : \mu \neq 280$ es a dos colas.

3. *Elegir un estadístico*. Hasta aquí el diseño del TdH viene siendo igual al del test z. La diferencia ahora es que no conocemos la varianza σ^2 de la distribución de los pesos de las barras de jabón. Esto nos impide tomar como estadístico

   $$Z = \frac{\sqrt{n}(\overline{X}_n - 280)}{\sigma}$$

   pues no sabríamos calcularlo a partir de los datos observados (¿qué σ ponemos en la fórmula?). El truco consiste en cambiar σ por S_n.

   Usaremos el estadístico

   $$T = \frac{\sqrt{n}(\overline{X}_n - 280)}{S_n} = \frac{\sqrt{20}(\overline{X}_{20} - 280)}{22}.$$

El valor observado de T es en nuestro caso

$$T_{\text{obs}} = \frac{\sqrt{20}(289-280)}{22} = 1.83$$

4. *Elegir un nivel de significación y determinar la región de rechazo.* Para variar un poco, tomemos $\alpha = 0.1$. Lo natural es rechazar H_0 cuando el promedio $\overline{X}_{20}$ se aleja bastante de 280. Esto es, rechazar cuando

$$\left|\overline{X}_{20} - 280\right| \geq k.$$

Podemos escribir una desigualdad similar usando el estadístico T

$$|T| = \left|\frac{\sqrt{20}(\overline{X}_{20} - 280)}{S_{20}}\right| \geq c.$$

que es una región del tipo $I = (-\infty, -c] \cup [c, +\infty)$.

Para calcular c debemos usar el nivel de significación elegido. Debemos resolver

$$\alpha = \mathbf{P}(T \in I | H_0) = \mathbf{P}(|T| \geq c | \mu = 280) = F_{t,n-1}(-c) + 1 - F_{t,n-1}(c),$$

en donde hemos denotado por $F_{t,n-1}$ la f.d.a. de la distribución t de Student con $n-1$ grados de libertad.

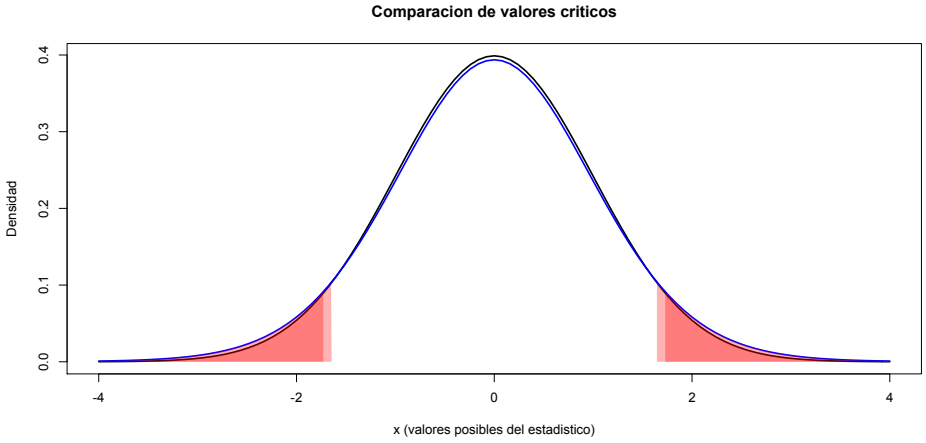
La distribución t también es simétrica en torno al cero, por lo que

$$F_{t,n-1}(-c) = 1 - F_{t,n-1}(c).$$

Vemos entonces que debemos resolver

$$F_{t,n-1}(c) = 1 - \alpha/2.$$

Es decir, c es el valor crítico $t_{n-1}(\alpha/2) = t_{19}(0.05) = 1.73$ (lo obtenemos de la tabla).



En la figura de arriba mostramos la comparación entre el valores crítico 1.645 de la normal (el test z) con el de Student (el test t) de 19 grados de libertad. Aunque la diferencia es chica, lo correcto es usar el t en este caso. Si hubiéramos usado el valor crítico de la normal, el nivel de significación sería 0.12 y no 0.1 como queremos.

5. *Determinar la(s) potencia(s)*. He aquí la mala noticia. El aparentemente insignificante cambio de σ por S_n que hicimos para poder calcular el estadístico hace que no podamos calcular de forma sencilla las potencias de un test t. Así que seguiremos adelante sin conocer la potencia.

Como $T_{\text{obs}} = 1.83 \geq 1.73$, rechazamos H_0. Esto lo podemos hacer calculando el p-valor a dos colas:

$$\text{pval}\,(T_{\text{obs}}) = \mathbf{P}\,(|T| \geq |T_{\text{obs}}||H_0) = 2(1 - F_{t,n-1}(1.83)) = 0.083,$$

que es menor que $\alpha = 0.1$. ■

4. El estadístico t a partir de la razón de verosimilitud

Apliquemos el método de la razón de verosimilitud a la situación modelada en el test t. En este caso $X_1, \ldots, X_n$ son i.i.d. normales $N(\mu, \sigma^2)$, con ambos μ y σ^2 desconocidos, pero estamos interesados en hacer un TdH para μ. Es decir, σ^2 es un parámetro molesto.

Supongamos que queremos hacer un test a dos colas

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_A : \mu \neq \mu_0 \end{cases}$$

Si bien H_0 parece simple pues consiste de un solo valor de μ_0, no lo es. Aunque asumamos H_0 es cierta, no podemos determinar la distribución de las $X_i's$, justamente por el parámetro molesto σ^2.

En este caso

$$P_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}.$$

Recordar que la densidad normal tiene la fórmula

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

La función de verosimilitud es

$$L\left(\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right),$$

por lo que al tomar logaritmo y poner $\mu = \mu_0$, obtenemos

$$\ell\left(\mu_0, \sigma^2\right) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

En este caso, μ_0 está fijo y debemos hallar el máximo variando σ^2. Así que

$$\frac{d\ell}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu_0)^2$$

que es cero cuando

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

De aquí concluimos que

$$\sup_{\sigma^2}\left\{L\left(\mu_0, \sigma^2\right)\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)^{-n/2} e^{-n/2}.$$

Para el denominador, ya sabemos de clases anteriores que los estimadores de máxima verosimilitud de μ y σ^2 son

$$\overline{X}_n \quad \text{y} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

respectivamente. Substituyéndolos en la definición de $L(\mu, \sigma^2)$ obtenemos

$$\sup\left\{L\left(\mu, \sigma^2\right)\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right)^{-n/2} e^{-n/2}.$$

Es decir, la razón de verosimilitud es

$$q_L = \left(\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}\right)^{-n/2}.$$

Esto puede ser escrito en una forma más conveniente. Notar que

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu_0)^2 &= \sum_{i=1}^{n}\left(\left(X_i - \overline{X}_n\right) + \left(\overline{X}_n - \mu_0\right)\right)^2 \\ &= \sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 + n\left(\overline{X}_n - \mu_0\right)^2\end{aligned}$$

por lo que

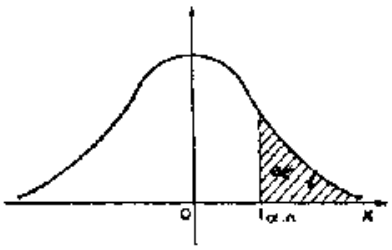
$$q_L = \left(1 + \frac{n\left(\overline{X}_n - \mu_0\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}\right)^{-n/2} = \left(1 + \frac{1}{n-1}T^2\right)^{-n/2}.$$

Ahora, la región de rechazo $\{q_L \leq k\}$ se puede escribir como $\{|T| \geq c\}$ para una constante c que depende de n y k.

Podemos calcular el valor de c para que el nivel de significación sea α. Como T, bajo la hipótesis nula, tiene distribución de Student con $n-1$ grados de libertad, vemos que la ecuación

$$\mathbf{P}(|T| \geq c | H_0) = \alpha$$

equivale a tomar $c = t_{n-1}(\alpha/2)$. Luego, rechazamos H_0 si $|T_{\text{obs}}| \geq t_{n-1}(\alpha/2)$.



$\alpha/2$ / gl	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,192	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,648	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,256	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291