UNIVERSIDAD PANAMERICANA
FACULTAD DE INGENIERÍA

Marzo – Junio 2020

Maestría en Ciencia de datos

estadística

31 DE MARZO DE 2020
AGENDA

- LECTURA ASIGNADA DE TAREA, COMENTARIOS
- EJERCICIOS DE APLICACIÓN DE D.P.BINOMIAL EN R
- DISTRIBUCIONES DE PROBABILIDAD
  - RESOLUCIÓN DE EJERCICIOS EN R

EL IMPACTO DEL TURISMO EN EL CASCO VIEJO DE BILBAO MEDIANTE LOS MODELOS ECONOMÍA COLABORATIVA: UNA APROXIMACIÓN A TRAVÉS DE UNA DISTRIBUCIÓN BINOMIAL NEGATIVA

1. ¿Qué se hizo en este estudio?
2. ¿Cómo se hizo?
3. ¿Para qué se hizo?

(**Teorema de Chebyshev**) La probabilidad de que cualquier variable aleatoria X tome un valor dentro de k desviaciones estándar de la media es de al menos $1 - 1/k^2$. Es decir,

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

k= número de desviaciones estándar

Para $k = 2$ el teorema establece que la variable aleatoria X tiene una probabilidad de al menos $1 - 1/2^2 = 3/4$ de caer dentro de dos desviaciones estándar a partir de la media; es decir, que tres cuartas partes o más de las observaciones de cualquier distribución se localizan en el intervalo $\mu \pm 2\sigma$. De manera similar, el teorema afirma que al menos ocho novenos de las observaciones de cualquier distribución caen en el intervalo $\mu \pm 3\sigma$.

1. DISTRIBUCIÓN DE POISSON
2. DISTRIBUCIÓN HIPERGEOMÉTRICA
3. DISTRIBUCIÓN HIPERGEOMÉTRICA MULTIVARIADA
4. DISTRIBUCIÓN BINOMIAL NEGATIVA Y GEOMÉTRICA
5. DISTRIBUCIÓN NORMAL

DISTRIBUCIÓN DE POISSON

Propiedades de un proceso de Poisson:

1. El número de resultados que ocurren en un intervalo o región específica es independiente del número que ocurre en cualquier otro intervalo de tiempo o región del espacio disjunto. De esta forma vemos que el proceso de Poisson no tiene memoria.
2. La probabilidad de que ocurra un solo resultado durante un intervalo de tiempo muy corto o en una región pequeña es proporcional a la longitud del intervalo o al tamaño de la región, y no depende del número de resultados que ocurren fuera de este intervalo de tiempo o región.
3. La probabilidad de que ocurra más de un resultado en tal intervalo de tiempo corto o que caiga en tal región pequeña es insignificante.

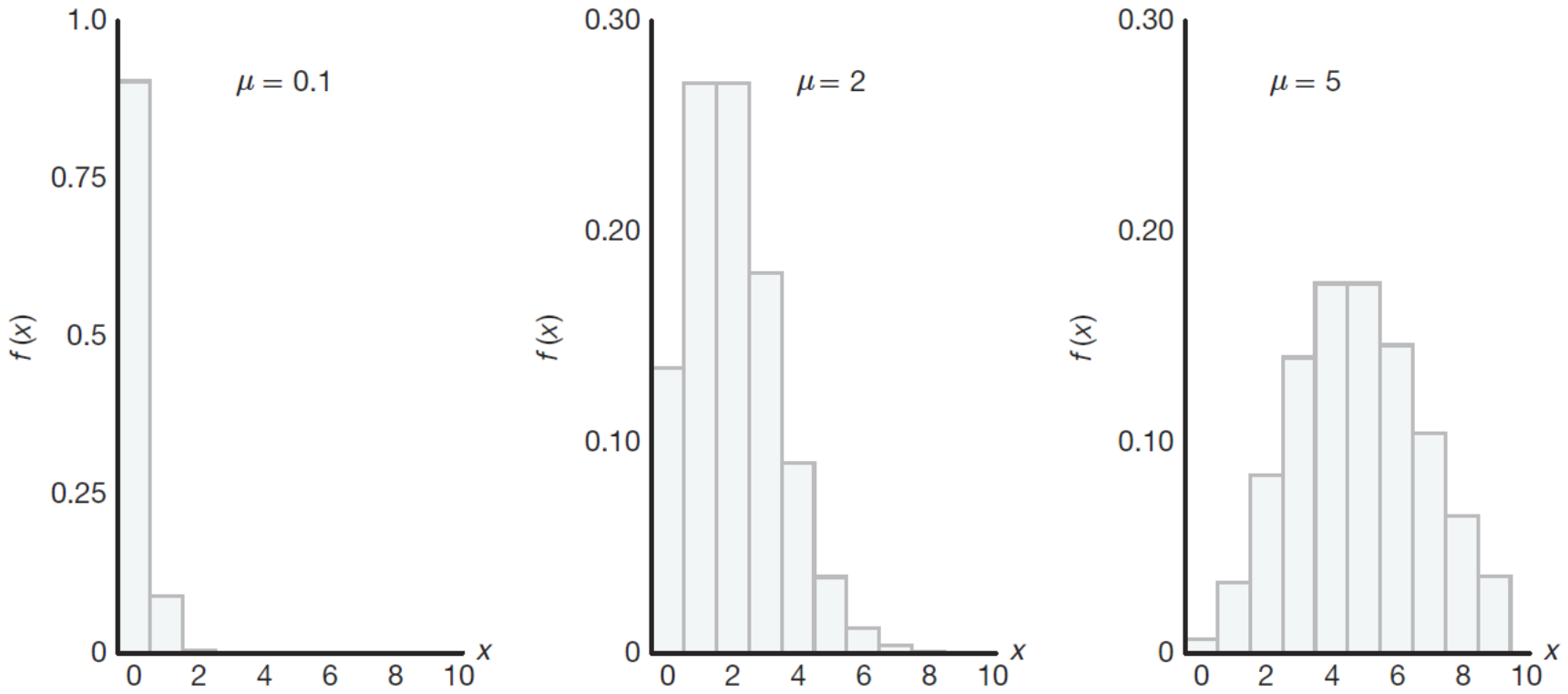
Sea X una variable aleatoria binomial con distribución de probabilidad $b(x; n, p)$. Cuando $n \to \infty$, $p \to 0$, y $np \overset{n\to\infty}{\longrightarrow} \mu$ permanece constante,

$$b(x; n, p) \overset{n\to\infty}{\longrightarrow} p(x; \mu).$$

La distribución de probabilidad de la variable aleatoria de Poisson X, la cual representa el número de resultados que ocurren en un intervalo de tiempo dado o región específicos y se denota con t, es

$$p(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

donde λ es el número promedio de resultados por unidad de tiempo, distancia, área o volumen y $e = 2.71828\ldots$



La curva será más simétrica en la medida en que la media es mayor

$$p(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

Durante un experimento de laboratorio el número promedio de partículas radiactivas que pasan a través de un contador en un milisegundo es 4. ¿Cuál es la probabilidad de que entren 6 partículas al contador en un milisegundo dado?

$$p(6;4) = \frac{e^{-4}4^6}{6!} = \sum_{x=0}^{6} p(x;4) - \sum_{x=0}^{5} p(x;4) = 0.8893 - 0.7851 = 0.1042$$

El numero promedio de camiones-tanque que llega cada día a cierta ciudad portuaria es 10. Las instalaciones en el puerto pueden alojar a lo sumo 15 camiones-tanque por día. ¿Cual es la probabilidad de que en un día determinado lleguen mas de 15 camiones y se tenga que rechazar algunos?

$$P(X > 15) = 1 - P(X \leq 15) = 1 - \sum_{x=0}^{15} p(x;10) = 1 - 0.9513 = 0.0487$$

Si hay doce coches cruzando un puente por minuto en promedio, encuentre la probabilidad de tener diecisiete o más coches cruzando el puente en un minuto en particular.

DISTRIBUCIÓN HIPERGEOMÉTRICA

La distribución hiper geométrica no requiere independencia y se basa en el muestreo que se realiza **sin reemplazo**

Experimento hiper geométrico:

1. De un lote de N artículos se selecciona una muestra aleatoria de tamaño n sin reemplazo.
2. k de los N artículos se pueden clasificar como éxitos y $N - k$ se clasifican como fracasos.

$$h(x;N,\ n,\ k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad \text{máx}\{0, n-(N-k)\} \leq x \leq \text{mín}\{n, k\}$$

$$\binom{k}{n} = \frac{k!}{n!\,(k-n)!}$$

Número de combinaciones de n ítems que se
Pueden muestrear en una población de 20 ítems.

Lotes con 40 componentes cada uno que contengan 3 o más defectuosos se consideran inaceptables. El procedimiento para obtener muestras del lote consiste en seleccionar 5 componentes al azar y rechazar el lote si se encuentra un componente defectuoso. ¿Cuál es la probabilidad de, que en la muestra, se encuentre exactamente un componente defectuoso, si en todo el lote hay 3 defectuosos?

$n = 5, N = 40, k = 3$ y $x = 1$

$$h(1; 40, 5, 3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011$$

Sólo 30% de las veces se detecta un lote con 3 defectuosos

Interpretación de los resultados de media y varianza, de acuerdo con el teorema de Chebyshev

$$\mu = \frac{(5)(3)}{40} = \frac{3}{8} = 0.375,$$

$$\sigma^2 = \left(\frac{40-5}{39}\right)(5)\left(\frac{3}{40}\right)\left(1-\frac{3}{40}\right) = 0.3113.$$

La media y la varianza de la distribución hipergeométrica $h(x; N, n, k)$ son

$$\mu = \frac{nk}{N} \text{ y } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1-\frac{k}{N}\right).$$

Si calculamos la raíz cuadrada de 0.3113, encontramos que $\sigma = 0.558$. Por lo tanto, el intervalo que se requiere es $0.375 \pm (2)(0.558)$, o de -0.741 a 1.491. El teorema de Chebyshev establece que el número de componentes defectuosos que se obtienen cuando, de un lote de 40 componentes, se seleccionan 5 al azar, de los cuales 3 están defectuosos, tiene una probabilidad de al menos 3/4 de caer entre -0.741 y 1.491. Esto es, al menos tres cuartas partes de las veces los 5 componentes incluirán menos de 2 defectuosos. ⌟

Cuando n es pequeña en comparacion con N, se puede utilizar una distribución binomial para aproximar la distribución hiper geométrica.

Por regla general la aproximación es buena cuando $n/N \leq 0.05$

Distribución hipergeométrica multivariada

Si N artículos se pueden dividir en las k celdas $A_1, A_2, \ldots, A_k$ con $a_1, a_2, \ldots, a_k$ elementos, respectivamente, entonces la distribución de probabilidad de las variables aleatorias $X_1, X_2, \ldots, X_k$, que representan el número de elementos que se seleccionan de $A_1, A_2, \ldots, A_k$ en una muestra aleatoria de tamaño n, es

$$f(x_1, x_2, \ldots, x_k; a_1, a_2, \ldots, a_k, N, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}},$$

con $\sum_{i=1}^{k} x_i = n$ y $\sum_{i=1}^{k} a_i = N$.

Se usa un grupo de 10 individuos para un estudio de caso biológico. El grupo contiene 3 personas con sangre tipo O, 4 con sangre tipo A y 3 con tipo B. ¿Cuál es la probabilidad de que una muestra aleatoria de 5 contenga 1 persona con sangre tipo O, 2 personas con tipo A y 2 personas con tipo B?

Si se utiliza la extensión de la distribución hipergeométrica con $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $a_1 = 3$, $a_2 = 4$, $a_3 = 3$, $N = 10$ y $n = 5$, vemos que la probabilidad que se desea es

$$f(1, 2, 2; 3, 4, 3, 10, 5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14}.$$

Distribución binomial negativa

Si ensayos independientes repetidos pueden dar como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número del ensayo en el que ocurre el k-ésimo éxito, es

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \ldots$$

En la serie de campeonato de la NBA (National Basketball Association), el equipo que gane 4 de 7 juegos será el ganador. Suponga que los equipos A y B se enfrentan en los juegos de campeonato y que el equipo A tiene una probabilidad de 0.55 de ganarle al equipo B.

a) ¿Cuál es la probabilidad de que el equipo A gane la serie en 6 juegos?

b) ¿Cuál es la probabilidad de que el equipo A gane la serie?

c) Si ambos equipos se enfrentaran en la eliminatoria de una serie regional y el triunfador fuera el que ganara 3 de 5 juegos, ¿cuál es la probabilidad de que el equipo A gane la serie?

$$b^*(x;k,p) = \binom{x-1}{k-1} p^k q^{x-k}$$

a) $b^*(6;4,0.55) = \binom{5}{3}0.55^4(1-0.55)^{6-4} = 0.1853.$

b) P(el equipo A gana la serie de campeonato) es

$$b^*(4;4,0.55) + b^*(5;4,0.55) + b^*(6;4,0.55) + b^*(7;4,0.55)$$
$$= 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083.$$

c) P(el equipo A gana la eliminatoria) es

$$b^*(3;3,0.55) + b^*(4;3,0.55) + b^*(5;3,0.55)$$
$$= 0.1664 + 0.2246 + 0.2021 = 0.5931.$$

Distribución binomial geométrica

Si pruebas independientes repetidas pueden tener como resultado un éxito con probabilidad p y un fracaso con probabilidad $q = 1 - p$, entonces la distribución de probabilidad de la variable aleatoria X, el número de la prueba en el que ocurre el primer éxito, es

$$g(x; p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots$$

Se sabe que en cierto proceso de fabricación uno de cada 100 artículos, en promedio, resulta defectuoso. ¿Cuál es la probabilidad de que el quinto artículo que se inspecciona, en un grupo de 100, sea el primer defectuoso que se encuentra?

Si utilizamos la distribución geométrica con $x = 5$ y $p = 0.01$, tenemos

$$g(5; 0.01) = (0.01)(0.99)^4 = 0.0096.$$

En “momentos ajetreados” un conmutador telefónico está muy cerca de su límite de capacidad, por lo que los usuarios tienen dificultad para hacer sus llamadas. Sería interesante saber cuántos intentos serían necesarios para conseguir un enlace telefónico. Suponga que la probabilidad de conseguir un enlace durante un momento ajetreado es $p = 0.05$. Nos interesa conocer la probabilidad de que se necesiten 5 intentos para enlazar con éxito una llamada.

$$P(X = x) = g(5; 0.05) = (0.05)(0.95)^4 = 0.041$$

La media y la varianza de una variable aleatoria que sigue la distribución geométrica son

$$\mu = \frac{1}{p} \text{ y } \sigma^2 = \frac{1-p}{p^2}.$$

Distribuciones continuas de probabilidad
Distribución uniforme

La función de densidad de la variable aleatoria uniforme continua X en el intervalo $[A, B]$ es

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B, \\ 0, & \text{en otro caso.} \end{cases}$$

Suponga que el tiempo máximo que se puede reservar una sala de conferencias grande de cierta empresa son cuatro horas. Con mucha frecuencia tienen conferencias extensas y breves. De hecho, se puede suponer que la duración X de una conferencia tiene una distribución uniforme en el intervalo $[0, 4]$.

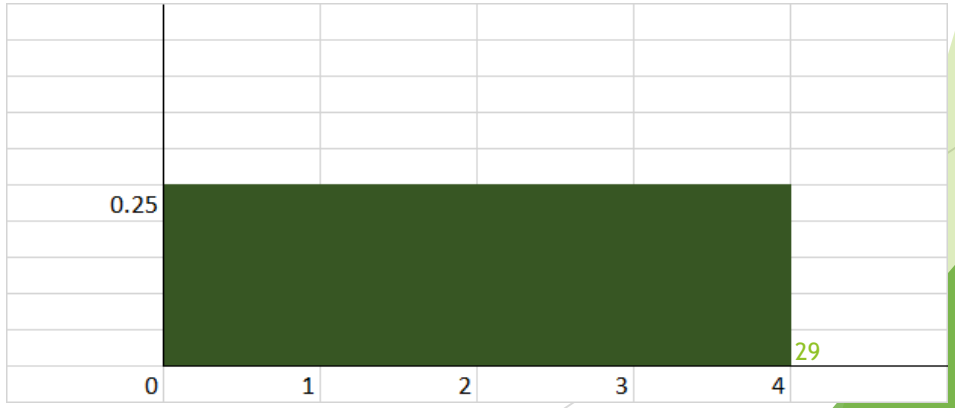
a) ¿Cuál es la función de densidad de probabilidad?

b) ¿Cuál es la probabilidad de que cualquier conferencia determinada dure al menos 3 horas?

a) La función de densidad apropiada para la variable aleatoria X distribuida uniformemente en esta situación es

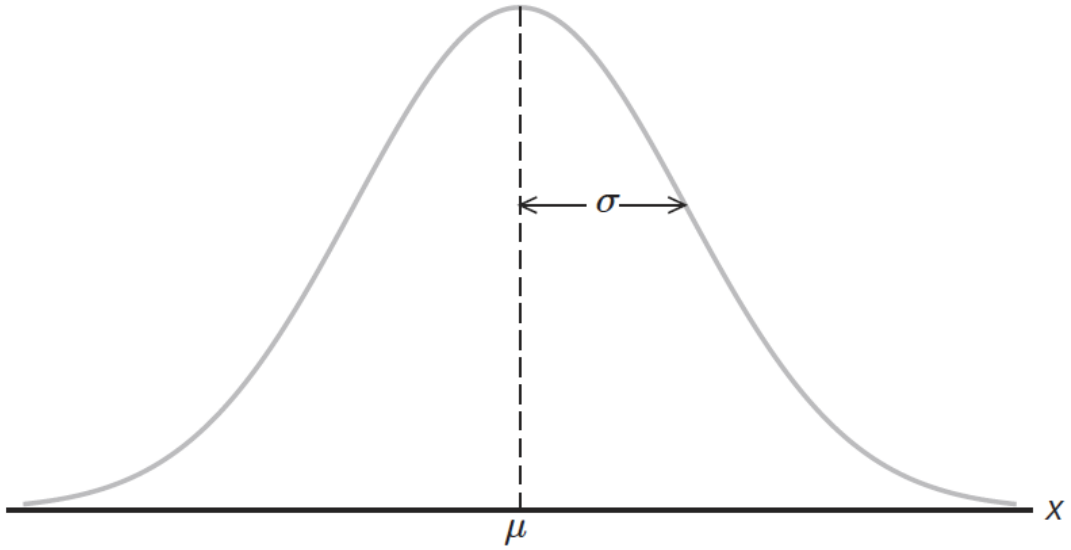
$$f(x) = \begin{cases} \frac{1}{4}, & 0 \le x \le 4, \\ 0, & \text{en otro caso.} \end{cases}$$

b) $P[X \ge 3] = \int_3^4 \frac{1}{4}\, dx = \frac{1}{4}$



Distribución normal o Gaussiana

Una variable aleatoria continua *X* que tiene la distribución en forma de campana de siguiente figura, se denomina **variable aleatoria normal**.



La densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es

$$n(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad -\infty<x<\infty,$$

Área bajo la curva normal

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)\,dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx,$$

Con el fin de hacer más simples el cálculo del área bajo la curva:

podemos transformar todas las observaciones de cualquier variable aleatoria normal *X* en un nuevo conjunto de observaciones de una variable aleatoria normal *Z* con media 0 y varianza 1. Esto se puede realizar mediante la transformación:

$$Z = \frac{X - \mu}{\sigma}.$$

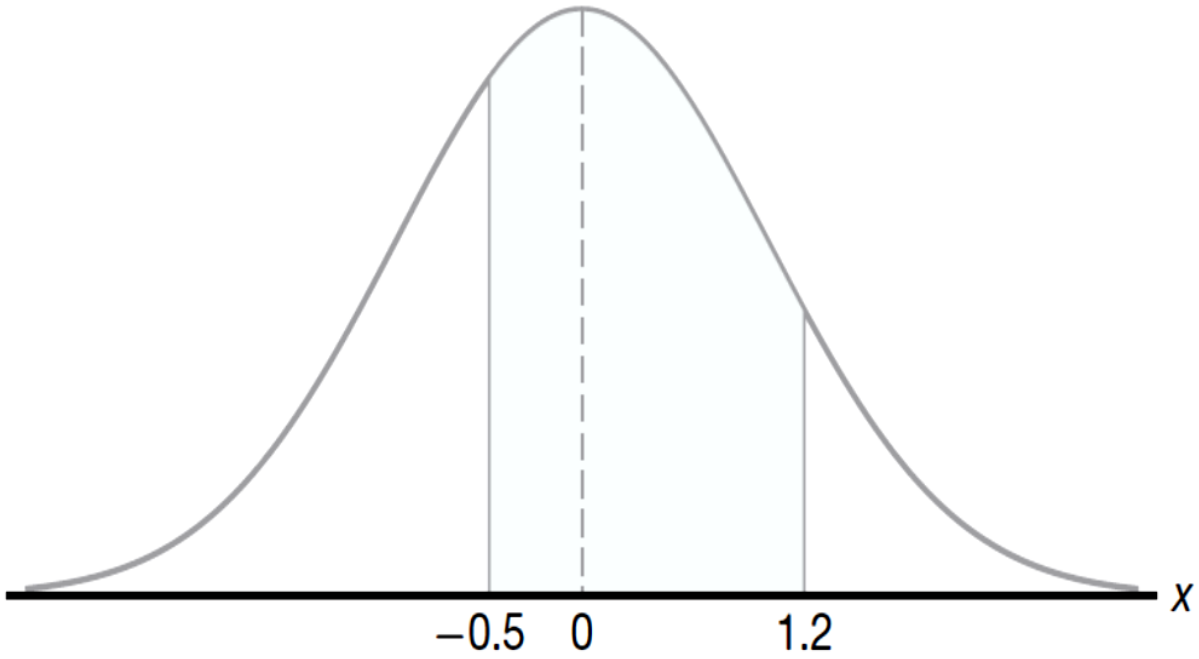
$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= \int_{z_1}^{z_2} n(z; 0, 1)\, dz = P(z_1 < Z < z_2),$$

donde Z se considera una variable aleatoria normal con media 0 y varianza 1.

La distribución de una variable aleatoria normal con media 0 y varianza 1 se llama **distribución normal estándar**.

Dada una variable aleatoria X que tiene una distribución normal con $\mu = 50$ y $\sigma = 10$, calcule la probabilidad de que X tome un valor entre 45 y 62.



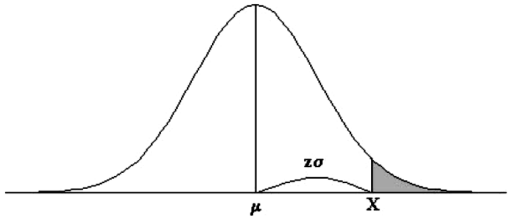
Los valores z que corresponden a $x_1 = 45$ y $x_2 = 62$ son

$$z_1 = \frac{45 - 50}{10} = -0.5 \text{ y } z_2 = \frac{62 - 50}{10} = 1.2$$

Esta área se puede calcular restando el área a la izquierda de la ordenada $z = -0.5$ de toda el área a la izquierda de $z = 1.2$. Si usamos la tabla A.3, tenemos

$$P(45 < X < 62) = P(-0.5 < Z < 1.2) = P(Z < 1.2) - P(Z < -0.5)$$
$$= 0.8849 - 0.3085 = 0.5764.$$

Áreas bajo la curva normal



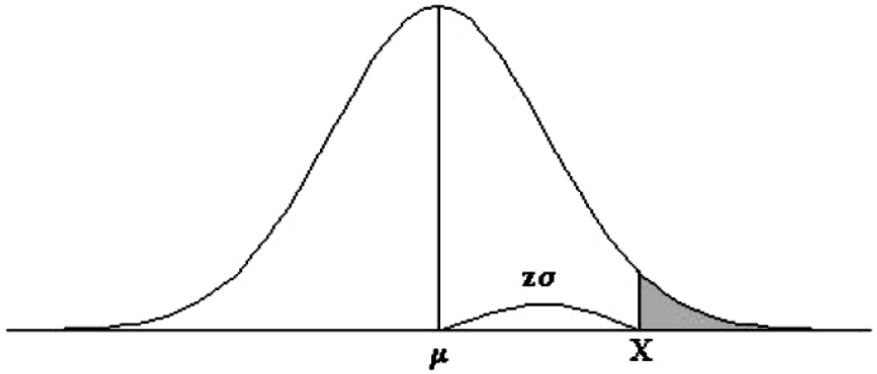
Ejemplo:

$$Z = \frac{X - \mu}{\sigma}$$

P [Z > 1] = 0.1587
P [Z > 1.96] = 0.0250

Desv. normal x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

Áreas bajo la curva normal



Ejemplo:

$$Z = \frac{X - \mu}{\sigma}$$

P [Z > 1] = 0.1587
P [Z > 1.96] = 0.0250

Desv. normal x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233