

Función de Distribución Normal

La distribución normal es, con mucho, la más utilizada de todas las distribuciones de probabilidad. Es una distribución de variable continua, con recorrido $(-\infty ; +\infty)$. Fue descubierta por Carl Gauss, matemático, físico y astrónomo alemán de principios del siglo XIX, al estudiar la distribución de los errores en las observaciones astronómicas.

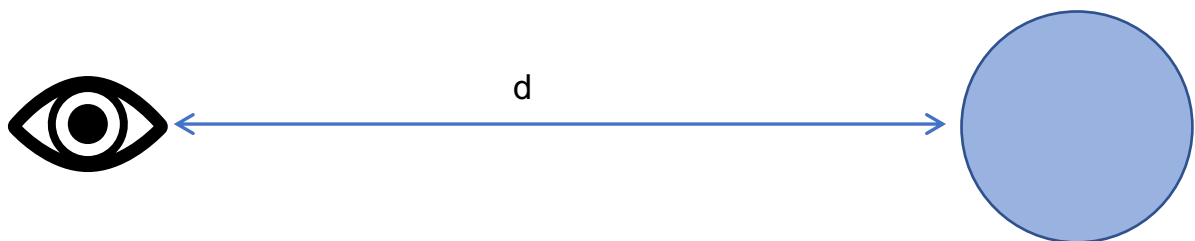
Debe su importancia a tres razones fundamentales:

Por un lado, un gran número de fenómenos reales se pueden modelizar con esta distribución (tales el caso de las características cuantitativas de casi todas las grandes poblaciones).

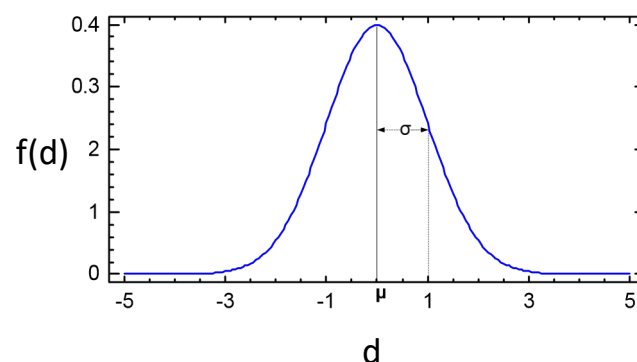
Por otro lado, muchas de las distribuciones de uso frecuente tienden a aproximarse a la distribución normal, bajo ciertas condiciones; y,

Por último, en virtud del *Teorema Central del límite*, todas aquellas variables que puedan considerarse causadas por un gran número de pequeños efectos (como pueden ser los errores de observación de una medición) tienden a distribuirse con una distribución normal.

Para comprender mejor la distribución normal imaginemos su descubrimiento, en un pequeño gráfico. Gauss definió la Distribución Normal al realizar mediciones acerca de la posición y distancia hacia una estrella en el cielo, con las herramientas de su época. Se encontró pues midiendo la distancia a ese astro (d): evidentemente cada distancia observada no era igual a las anteriormente tomadas, pues se producían errores de medición, o bien por error en la lectura, o bien por lo instrumento que usaba.



Teóricamente un error podía ser tan grande que la distancia observada fuera muy grande (asociando ese valor a más infinito) y, también teóricamente, a menos infinito. Plasmando sobre un gráfico xy las distancias medidas en cada observación realizada, cuantas más observaciones realizaba, acabó trazando una curva continua similar a la siguiente figura donde representaba sobre el eje x las distancias d medidas.



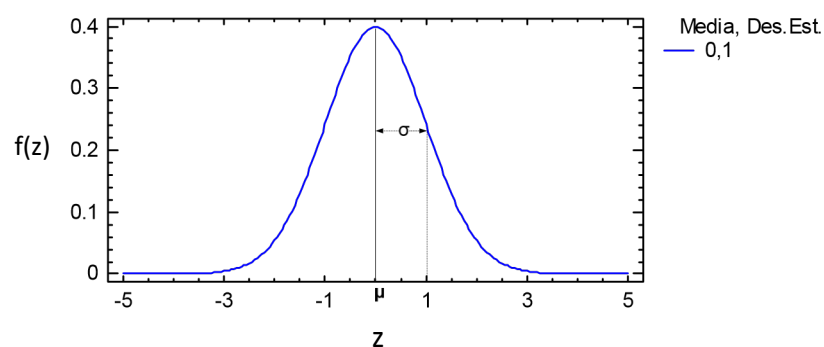
Si sobre el eje y se representaba la frecuencia de cada observación, acabó surgiendo una curva en forma de campana, en donde el valor μ (valor medio) dividía al área bajo la curva en dos partes iguales, con lo que esa media era también la **Mediana**. Otra consideración al visualizar la curva era que el valor medio tenía la mayor frecuencia dentro de los valores medidos, con lo que era también la **Moda**.

En conjunto, las mediciones acabaron tomando una determinada forma, hoy conocida como **“Campana de Gauss”**.

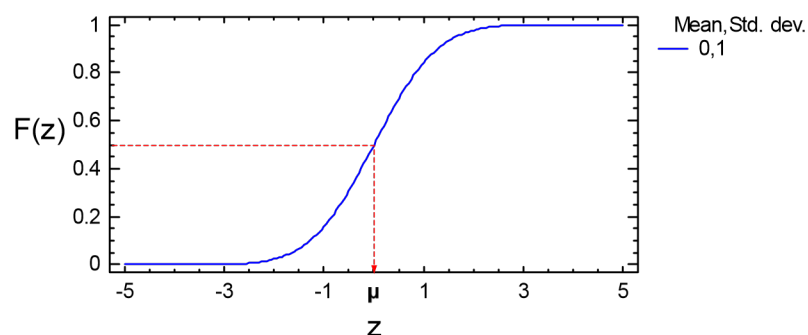
Recordando que la función de probabilidad de las variables continuas se llama **función densidad de probabilidad**, la curva anterior es una *distribución de probabilidad de una variable continua*. Si nos planteamos calcular la expresión analítica de dicha forma llegaremos a la función de la curva, o *función de densidad*, que por sus características es la de la **Distribución Normal**. Esta función depende, como se observa, de dos parámetros: uno, el valor central de la curva, y otro, la distancia de dicho valor a los puntos de inflexión (es posible comprobar que dichos parámetros son la **media** y la **desviación estándar** de esa función).

La distribución normal es entonces una distribución de variable continua que queda especificada por dos parámetros, de los que depende su función de densidad, y que resultan ser la **media** y la **desviación estándar** de la distribución. Su estudio teórico, por ser una variable continua, se realiza a partir de su función de densidad.

La **Distribución Normal** es el modelo de distribución más utilizado en la práctica, ya que multitud de fenómenos se comportan según esa distribución.



La Distribución Normal tipificada



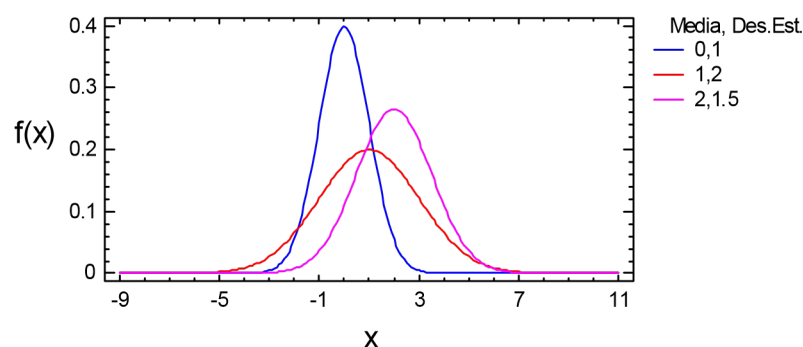
Función de Distribución de una Variable Normal tipificada

Cualquier variable continua que tenga una Distribución Normal cumple con las características ya mencionadas (Media = Mediana = Moda). Por lo tanto, el 50% de los valores están a la derecha del valor central y otro 50% a la izquierda, y se la define por sus **dos parámetros**:

$$X \rightarrow N(\mu, \sigma^2)$$

μ : es el valor medio de la distribución y es precisamente donde se sitúa el centro de la curva que habíamos denominado "*Campana de Gauss*".

σ^2 : es la varianza. Indica si los valores están más o menos alejados del valor central: si la varianza es baja, los valores están próximos a la media; si es alta, entonces los valores están muy dispersos.



Tres variables aleatorias continuas de Distribución Normal

La función densidad de probabilidad de una variable continua Normal está dada por la siguiente ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cuando la media de una Distribución Normal es 0 y la varianza es 1, se la denomina "**normal tipificada**", y su ventaja reside en que, aplicando la ecuación anterior, se ha podido tabular la probabilidad acumulada para cada punto de la curva de esta distribución. Con esas tablas es factible resolver la gran mayoría de los problemas prácticos asociados a este tipo de distribución de probabilidad.

Además, puede comprobarse que **toda distribución normal se puede transformar en una normal tipificada**:

Ejemplo: Dada una variable aleatoria que sigue el modelo de una distribución normal con media 10 y varianza 4, transformarla en una normal tipificada.

$$X \rightarrow N(10, 4)$$

Para transformarla en una normal tipificada **se crea una nueva variable (Z)** que será **igual a la anterior (X) menos su media y dividida por su desviación estándar** (que es la raíz cuadrada de la varianza)

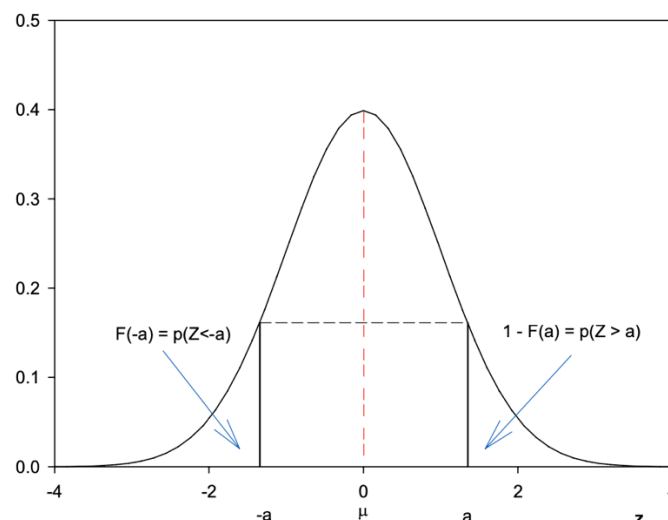
$$Z = \frac{X - \mu}{\sigma}$$

En este caso, los valores de la nueva variable estarán dados por: $Z = \frac{X-10}{2}$

La nueva variable tiene una distribución normal tipificada:

$$Z \rightarrow N(0,1),$$

y es posible calcular la probabilidad acumulada para cada valor de esta nueva variable.



Otra característica interesante deriva del carácter simétrico de esta distribución.

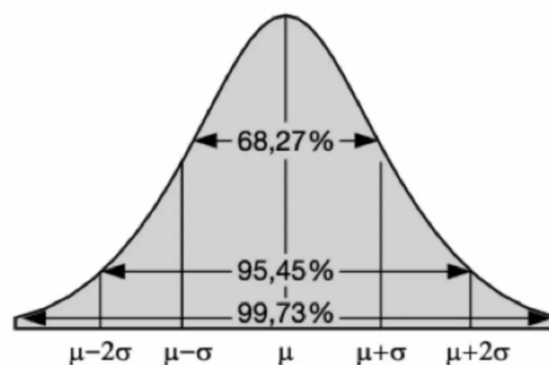
Si tomamos dos valores de la variable, $z_1 = -a$ y $z_2 = a$, luego tal como surge de la gráfica encima,

$$F(-a) = P(Z < -a) = P(Z > a) = 1 - F(a) \quad \therefore \quad \boxed{F(a) = 1 - F(-a)} \quad \text{y} \quad \boxed{F(-a) = 1 - F(a)}$$

Esta característica es de suma utilidad para resolver algunos problemas asociados con esta distribución de probabilidad.

Una última característica importante de las distribuciones normales de gran utilidad en temas prácticos tales como el Control de Calidad es que, si tomamos distintos rangos de media y desvío estándar, podemos identificar rápidamente el porcentaje de casos que quedan englobados en cada uno. En efecto, podemos calcular los datos para el siguiente cuadro:

- Valores entre $(\mu-\sigma)$ y $(\mu+\sigma)$ = 68,27 %
- Valores entre $(\mu-2\sigma)$ y $(\mu+2\sigma)$ = 95,45 %
- Valores entre $(\mu-3\sigma)$ y $(\mu+3\sigma)$ = 99,73 %



O sea, conociendo los dos parámetros para una determinada distribución normal, podemos por ejemplo asegurar que el 99,73% de los casos quedan englobados en el intervalo $(\mu - 3\sigma, \mu + 3\sigma)$, o lo que es equivalente, que en 6σ se concentra ese 99,73% de una población que se comporta según una distribución normal. Podemos inferir que en 8σ encontraremos aproximadamente el 100% de la población.

¿Cómo se lee una Tabla de Distribución de Probabilidad Normal?

Veamos una *Tabla de Distribución Norma Tipificada*. La columna de la izquierda nos indica el valor de la variable cuya probabilidad acumulada queremos conocer. La primera fila nos indica el segundo decimal del valor de la variable que estamos consultando. En la intersección de la fila y columna correspondiente encontraremos la probabilidad acumulada buscada.

Ejemplo: Queremos conocer la probabilidad acumulada para $z = 1,75$. Entonces buscamos en la columna de la izquierda el valor 1,7 y en la primera fila el valor 0,05. La casilla en la que se interceptan la fila y columna correspondientes es la probabilidad acumulada ($F(1,75) = 0,9599 = 95.99\%$).

Debemos tener en cuenta que la Tabla nos da la **probabilidad acumulada**, es decir, la que va desde el inicio del Recorrido de la función normal tipificada hasta dicho valor. No nos da la probabilidad concreta en ese punto. En una distribución continua en el que la variable puede tomar infinitos valores, la probabilidad en un punto concreto es prácticamente despreciable. Por eso consideramos que la probabilidad de que tome un valor específico es cero.

Ejemplo: Imaginemos que una variable continua puede tomar valores entre 0 y 5. La probabilidad de que tome exactamente el valor 2 es despreciable, ya que podría tomar infinitos valores: por ejemplo: 1,99, 1,994, 1,9906, 1,9998, 2,0025, etc.

Veamos **otros ejemplos:**

Probabilidad acumulada en el valor 0,67: la respuesta es 0,7486

Probabilidad acumulada en el valor 1,30: la respuesta es 0,9032

Probabilidad acumulada en el valor 2,19: la respuesta es 0,9857

Un comentario final respecto del uso de la Tabla.

Supongamos que, en lugar de buscar la probabilidad acumulada asociada a un valor de Z, queremos buscar el valor de Z asociado a una determinada probabilidad.

Por ejemplo, queremos encontrar el valor de Z asociado a $P(Z) = 0,89$.

En ese caso, para sistematizar la búsqueda, nos conviene centrarnos en la segunda columna de la Tabla (aquella que corresponde al valor 0,00 para el segundo dígito decimal de Z). Si descendemos por esa columna, llegaremos al valor 0,9032, que es el primero que supera 0,89.

Si retrocedemos a la fila inmediata anterior (identificada por $Z = 1,2$), y nos desplazamos por esa fila hacia la derecha, encontraremos como valor más cercano al valor buscado a $Z = 1,23$.

De donde $P(Z = 1,23) = 0,8907$ y habremos llegado al valor de nuestra variable buscado.

A continuación, listamos una serie de ejercicios cortos para completar la familiarización con el uso de la Tabla de Distribución Normal.

- a) $p(Z \leq 1) = F(1) = 0.8413$
- b) $p(Z \leq 1.95) = F(1.95) = 0.974$
- c) $p(Z \leq -1) = 1 - F(1) = 1 - 0.8413 = 0.1587$
- d) $p(Z \leq -1.95) = 1 - p(Z \leq 1.95) = 0.026$
- e) $p(1 < Z \leq 1.95) = F(1.95) - F(1) = 0.974 - 0.8413 = 0.1327$
- f) $p(-1.95 < Z \leq -1) = F(-1) - F(-1.95) = 0.1587 - 0.026 = 0.1327$
- g) $p(-2 \leq Z < 0) = F(0) - F(-2) = 0.5 - [1 - F(2)] = -0.5 + 0.9773 = 0.4773$

Veamos ahora, una serie de ejercicios que requieren de esa misma Tabla para resolver distintas distribuciones normales.

Ejemplo 1: El salario medio de los empleados de una empresa se distribuye según una distribución normal, con media 50 mil pesos y desvío estándar de 4 mil pesos. Calcular el porcentaje de empleados con un sueldo inferior a 45 mil pesos.

Lo primero que haremos es transformar esa distribución en una normal tipificada. Para ello, crearemos una nueva variable (Z) que será igual a la anterior (X) menos su media y dividida por su desvío estándar.

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

En el ejemplo, la nueva variable sería:

$$Z = \frac{X - 50}{4}$$

Esta nueva variable se distribuye como una normal tipificada. Necesitamos calcular $P(X < 45)$ para ello debemos encontrar que valor le corresponde a la variable Z para el valor de la variable $X = 45$:

$$Z = \frac{X - 50}{4} = \frac{45 - 50}{4} = -1,25$$

Ahora tenemos que calcular $P(Z < -1,25) = F(-1,25)$. Dado que tenemos una Tabla sólo con valores positivos de Z y que toda distribución normal es simétrica, podremos calcular ese valor a partir de su complemento.

Recordemos que $F(-1,25) = 1 - F(1,25)$ por lo que $P(Z < -1,25) = 1 - P(Z < 1,25)$

De la Tabla podemos identificar que $F(1,25) = 0,8944$ y por lo tanto,

$$P(-1,25) = 1 - 0,8944 = 0,1056$$

Luego, el porcentaje de empleados con sueldos inferiores a \$45.000 será 10,56%.

Ejercicio 2: Según una encuesta, la renta media de los egresados de una Facultad después de 2 años de su egreso es de 290 mil pesos/año, con una varianza de 784. Se supone que esa renta se distribuye según una distribución normal. Calcular:

- a) Porcentaje de esa población con una renta inferior a 236 mil pesos anuales.
- b) Renta a partir de la cual se sitúa el 10% de la población con mayores ingresos.
- c) Ingresos mínimo y máximo que engloba al 60% de la población con renta media.

Solución:

- a) Comencemos por el porcentaje de esa población con una renta inferior a 236 mil pesos anuales. Lo primero que debemos hacer es calcular la normal tipificada (Z):

$$Z = \frac{X-290}{28} \quad (*)$$

(*) Recordemos que el denominador es la desviación estándar (o sea, la raíz cuadrada de la varianza)

Reemplazando X en la ecuación anterior, surge que el valor de Z equivalente a 236 mil pesos anuales es -1,93

Entonces, $P(X \leq 236) = P(Z \leq -1,93)$

Ahora tenemos que ver cuál es la probabilidad acumulada hasta ese valor. Como en el ejemplo inicial nuestra Tabla sólo contiene valores positivos, por lo que debemos calcular esa probabilidad a partir del complemento.

Por lo tanto: $P(Z \leq -1,93) = P(Z \geq 1,93)$

Y de la Tabla, $P(Z > 1,93) = 1 - P(Z \leq 1,93) = 1 - 0,9732 \cong 0,0278$

Luego, de darse la distribución estimada en este ejercicio, sólo el 2,78% de la población tendría una renta inferior a 236 mil pesos anuales.

- b) Veamos ahora cuál es el nivel de ingresos a partir del cual se sitúa el 10% de la población con renta más elevada.

Vemos en la tabla el valor de la variable tipificada cuya probabilidad acumulada es el 0,9 (90%), lo que quiere decir que por encima se sitúa el 10% superior.

Ese valor, según lo visto anteriormente corresponde a $Z \cong 1,28$ (aproximadamente).

Ahora calculamos la variable normal X equivalente a ese valor de la normal tipificada:

$$1,28 = \frac{X-290}{28}$$

Despejando X, $X = (28 * 1,28) + 290 \cong 325,8$

O sea, llegamos al valor es 325,8 miles de pesos anuales. Por lo tanto, aquellas personas con ingresos superiores a esa cifra constituyen el 10% de la población con renta más elevada.

- c) Por último, veamos el nivel de ingresos mínimo y máximo que engloba al 60% de la población con renta media

Recordemos que la distribución de probabilidad normal es simétrica. Por lo tanto, si queremos identificar el 60 % de los casos con rentas medias, eso implica que deberemos buscar el recorrido de x entre $\mu - 30\%$ y $\mu + 30\%$, o lo que es equivalente buscar en Tabla los valores de la variable normalizada Z que se correspondan con las probabilidades acumuladas 0,20 ($50\%-30\% = 20\%$) y 0,80 ($50\%+30\% = 80\%$).

O sea que el segmento (Z_1, Z_2) englobará al 60% de la población con ingresos medios, y podremos determinar Z_1 y Z_2 recordando que $P(Z_1) = 0,20$ y $P(Z_2) = 0,80$.

Tal como habíamos explicado anteriormente, bajando en la Tabla de Distribución Normal Tipificada por la 2° columna, encontramos que el primer valor mayor a 0,80 es 0,8159. Por lo tanto, tomando la fila anterior ($Z=0,8$) y desplazándonos hacia la derecha vemos que para $Z=0,84$ el valor de probabilidad tabulado es 0,7995, o sea aproximadamente 0,80.

Como habíamos comprobado que $P(-Z) = 1 - P(Z)$, y como en nuestro caso $Z_1 = -Z_2$, entonces aplicando aquella igualdad a nuestro caso $P(-Z_2) = 1 - P(Z_2)$,

y luego, $P(-0,84) = 1 - P(0,84) = 1 - 0,80 = 0,20$

Reemplazando los valores Z_1 y Z_2 de nuestra variable tipificada, podemos calcular los valores de X_1 y X_2 asociados. En efecto,

$$X_1 = (28 * (-0,84)) + 290 \cong 266,5 \quad \text{y} \quad X_2 = (28 * (+0,84)) + 290 \cong 313,5$$

Por lo tanto, las personas con ingresos superiores a 266,5 miles de pesos anuales e inferiores a 313,5 miles de pesos anuales. constituyen el 60% de la población con un nivel medio de renta.

Ejercicio 3: La vida media de los habitantes de un país es de 68 años, con una varianza de 25. Se hace un estudio en una pequeña ciudad de 10.000 habitantes:

- a) ¿Cuántas personas es razonable pensar que superarán los 75 años?
- b) ¿Cuántos vivirán menos de 60 años?

Solución:

a) Personas que vivirán (previsiblemente) más de 75 años.

Calculamos el valor de la normal tipificada equivalente a 75 años,

$$z = \frac{75-68}{5} = 1,40$$

Por lo tanto de la Tabla,

$$P(X > 75) = P(Z > 1,40) = 1 - P(Z \leq 1,40) = 1 - 0,9192 = 0,0808$$

Luego, el 8,08% de la población (unos 808 habitantes) es razonable estimar que vivirán más de 75 años.

b) Personas que vivirán (previsiblemente) menos de 60 años.

Calculamos el valor de la normal tipificada equivalente a 60 años,

$$z = \frac{60-68}{5} = -1,60$$

Por lo tanto también de la Tabla,

$$P(X \leq 60) = P(Z \leq -1,60) = P(Z > 1,60) = 1 - P(Z \leq 1,60) = 1 - 0,9452 = 0,0548$$

Luego, el 5,48% de la población (548 habitantes) no llegarán probablemente a esta edad.

Ejercicio 4: El consumo medio anual de cerveza de los habitantes de un país es de 59 litros, con una varianza de 36. Se supone que se distribuye según una distribución normal.

- a) Si usted presume de buen bebedor, ¿por lo menos cuántos litros de cerveza tendría que beber al año para pertenecer al 5% de la población que más bebe?
- b) Si usted bebe 50 litros de cerveza al año y su mujer lo califica de borracho, ¿qué podría argumentar en su defensa?

Solución:

- a) Analicemos al 5% de la población que más bebe.

Vemos en la Tabla el valor de la variable normal tipificada cuya probabilidad acumulada es el 0,95 (95%), por lo que por arriba estaría el 5% restante.

Ese valor corresponde a $Z \cong 1,645$ (aprox.). Ahora calculamos la variable normal X equivalente a ese valor de la normal tipificada:

$$1,645 = \frac{X-59}{6} \quad \text{y luego,} \quad X = (1,645 * 6) + 59 \cong 67,9$$

Del despeje anterior surge el valor 67,9 y por lo tanto, tendría usted que beber más de 67,9 litros al año para pertenecer a ese "selecto" club de grandes bebedores de cerveza.

- b) Usted bebe 50 litros de cerveza al año. ¿Es usted un borracho?

Vamos a ver en que nivel de la población se situaría usted en función de su consumo de cerveza. Calculamos el valor de la normal tipificada correspondiente a 50 litros:

$$Z = \frac{50-59}{6} = -1,50$$

$$\text{Luego, } P(X < 50) = P(Z < -1,50) = P(Z > 1,50) = 1 - P(Z \leq 1,50) = 1 - 0,9332 = 0,0668$$

Luego, tan sólo un 6,68% de la población bebe menos que usted. Parece un argumento de suficiente peso para que dejen de catalogarlo de "enamorado de la bebida".

Ejercicio 5: Al examen de un concurso profesional por oposición se han presentado 2.000 aspirantes. La nota media ha sido un 6,5, con una varianza de 1,5.

- a) Tan sólo hay 250 plazas. Usted ha obtenido un 7,7. ¿Sería oportuno ir organizando una fiesta para celebrar su éxito?
- b) Va a haber una 2ª oportunidad para el 20% de notas más altas que no se hayan clasificado en el primer intento. ¿A partir de que nota se podrá participar en esta "repecha"?

Solución:

- a) Ud. ha obtenido usted un 7,7 en su examen y duda acerca de ir organizando un festejo por su aprobación.

Vamos a ver con ese 7,7 en qué nivel porcentual ha quedado situado. Para ello vamos a comenzar por calcular el valor de la normal tipificada equivalente.

$$Z = \frac{7,7-6,5}{1,049} = 1,144$$

A este valor de Z le corresponde una probabilidad acumulada aproximada (ver Tabla) de 0,8739 ($\cong 87,4\%$), lo que quiere decir que por encima de usted se encuentra un 12,6% de los aspirantes.

Si se han presentado 2.000 aspirantes, ese 12,6% equivale a unos 252 aspirantes. Por lo que, si hay 250 plazas disponibles, usted está muy cerca del punto de corte y es probable le convenga aguardar un poco para ir organizando los festejos.

- b) Puntaje para participar de una 2ª oportunidad.

Ahora veamos el punto de corte para la "repesca" que incluirá al 20% de los candidatos examinados con mejores notas, pero que no han aprobado.

Vemos en la tabla el valor de la normal tipificada que acumula el 80% de la probabilidad, ya que por arriba sólo quedaría el 20% restante.

Este valor de Z corresponde a 0,84 (aproximadamente). Ahora calculamos el valor de la variable continua normal X equivalente:

$$0,84 = \frac{X-6,5}{1,049} \text{ y luego, } X = (0,84 * 1,049) + 6,5 \cong 7,4$$

Al despejar X cuyo valor aproximado es 7,4 podemos determinar la nota a partir de la cual se convocará a la "repesca".