AWAKELAB

BASECAMP

Ciencia de Datos

Inferencia Estadística

Objetivo de la jornada

- Realizar cálculos de probabilidad seleccionando la distribución de probabilidad requerida para resolver un problema.

Distribución de probabilidades.

Una distribución de probabilidad es aquella que permite establecer toda la gama de resultados probables de ocurrir en un experimento determinado. Es decir, describe la probabilidad de que un evento se realice en el futuro.

Variables Aleatorias

Una función que asocia un número real, perfectamente definido, a cada punto muestral.

A veces las variables aleatorias (v.a.) están ya implícitas en los puntos muestrales.

Ejemplos:

- Experiencia consistente en medir la presión sistólica de 100 individuos. Un punto muestral (resultado de un experimento) es ya un número (presión sistólica). La v.a. está implícita.
- Una mujer portadora de hemofilia tiene 3 hijos ¿Cuál es el espacio muestral apropiado para estudiar la posible hemofilia de estos?

  W = {sss, ssn, sns, snn, nss, nsn, nns, nnn}

  Se podría definir una variable que asignará a cada punto muestral el número de orden en el espacio muestral.

X: sss $\rightarrow$1; ssn $\rightarrow$2; sns $\rightarrow$3;...

Pero otra posible v.a.: a cada punto muestral el número de s. X: sss $\rightarrow$ 3; ssn $\rightarrow$2; …

Los conjuntos pueden ser:

Discretos: número finito o infinito numerable de elementos.

Continuos: número infinito no numerable de elementos.

Las v.a. definidas sobre espacios muestrales discretos se llaman v.a. discretas y las definidas sobre espacios muestrales continuos se llaman continuas.

Una v.a. puede ser continua, aunque nosotros sólo podamos acceder a un subconjunto finito de valores. P.e. la presión arterial es una v.a. continua pero sólo podemos acceder a un conjunto finito de valores por la limitación de los aparatos de medida.

En general, las medidas dan lugar a v.a. continuas y los conteos a v.a. discretas.

Inducción de la probabilidad a variable aleatoria

Las v.a permiten definir la probabilidad como una función numérica (de variable real) en lugar de como una función de conjunto.

Ejemplo: Tiramos una moneda 3 veces. Representamos cara por c y cruz por z.

Ω = {ccc, ccz, czc, zcc, czz, zcz, zzc, zzz}

La probabilidad de cada suceso elemental es 1/8. Por ejemplo p(ccc)=1/8, ya que la probabilidad de sacar cara en una tirada es 1/2 según la definición clásica y las tiradas son independientes.

Definimos la v.a. X: número de caras, que puede tomar los valores {0, 1, 2, 3}. Se buscan todos los puntos muestrales que dan lugar a cada valor de la variable y a ese valor se le asigna la probabilidad del suceso correspondiente.

x	Sucesos	p_x
0	{zzz}	1/8
1	{czz, zcz, zzc}	3/8
2	{ccz, czc, zcc}	3/8
3	{ccc}	1/8

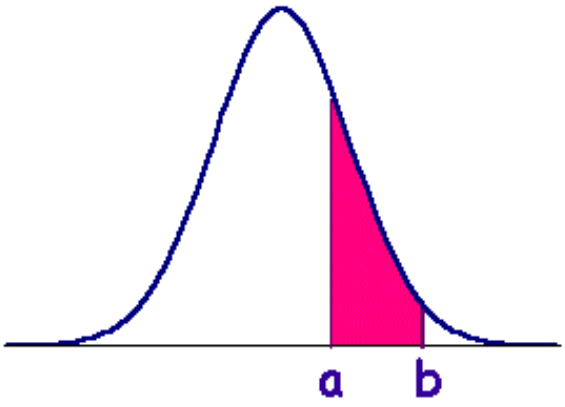
A esta función se le denomina función densidad de probabilidad (fdp), que desgraciadamente "funciona" de distinta manera en las variables discreta que en las continuas. En el caso de las variables discretas, como en el ejemplo, es una función que para cada valor de la variable da su probabilidad.

Ejemplo 4: Supongamos la variable tipo histológico de un tumor, con los valores 1, 2, 3, 4. Si la fdp fuera

x	f(x)
1	0,22
2	0,27
3	0,30
4	0,21

significa que la probabilidad del tipo 2 es 0,27, etc.

Para variables continuas la probabilidad de que una variable tome cualquier valor concreto es 0, por lo tanto la fdp sólo permite calcular la probabilidad para un intervalo del tipo (a<X<b), mediante el área bajo la curva de la fdp.



Para las variables aleatorias de interés hay tablas, y programas de ordenador, donde buscar esos valores.

Distribución acumulativa o función de distribución

Sea X una variable aleatoria discreta cuyos valores suponemos ordenados de menor a mayor. Llamaremos **función de distribución de la variable** X, y escribiremos $F(x)$ a la función:

$$F(x) = P(X \leq x)$$

La función de distribución asocia a cada valor de la variable aleatoria la probabilidad acumulada hasta ese valor.

Para el ejemplo de tirar la moneda 3 veces, Representamos cara por c y cruz por z.

x	f(x)	F(x)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8

Media de una variable aleatoria

Una **variable aleatoria discreta** es aquella que toma únicamente valores discretos o aislados, mientras que una **variable aleatoria continua** es aquella que toma valores de un intervalo.

Ejemplo:

La variable que asigna el número de autos vendidos en una agencia automotriz es una **variable discreta** que puede tomar los valores $0, 1, 2, 3, \ldots$.

Ejemplo:

La variable que asigna la temperatura en una ciudad es una **variable continua** que puede tomar los valores $[a, b]$, donde a y b representan la temperatura mínima y máxima respectivamente que se alcanza en un día.

Esperanza matemática o media

Es una medida de tendencia central que se emplea para designar mediante un solo valor a una colección de elementos y se representa por μ

$$\mu = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n = \sum_{i=1}^{n} x_i \cdot p_i$$

Varianza de una variable aleatoria

Es una medida de dispersión que se emplea para indicar que tan cercanos de la media se encuentran los elementos de la colección y se representa por σ^2. Si la varianza es cero, entonces los elementos coinciden con la media; mientras mayor sea la varianza, mayor dispersión.

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot p_i - \mu^2$$

Distribución Normal

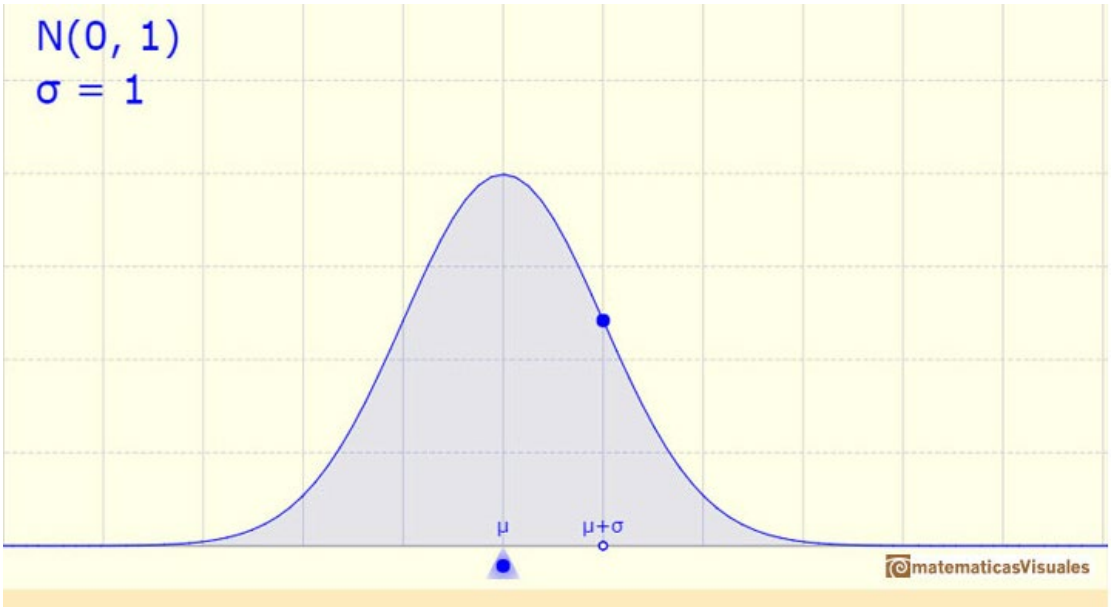
La **distribución normal** es la más notable en la Estadística y el Cálculo de Probabilidad. A ella se ajustan infinidad de variables aleatorias continuas presentes en la naturaleza, las ciencias sociales, las finanzas o la biomedicina. Llamamos a esa característica: *X*.

Queda definida por dos parámetros: la **media** y la **desviación estándar** (o desviación típica). Se denota por **N(*μ, σ*)**.

La distribución normal es un ejemplo importante referido a una variable aleatoria continua (la variable puede tomar cualquier valor real)

Podemos usar la distribución normal como una herramienta para calcular probabilidades. Por ejemplo, puede usarse para aproximar la distribución binomial (calcular probabilidades de la distribución binomial con números 'grandes' no ha sido tarea sencilla). Esta propiedad está en el origen de la curva normal.

La función de densidad de una distribución normal tiene forma de campana. Es simétrica en torno a la media. El área total bajo la curva es 1 (como corresponde a una función de densidad).



.

La densidad está concentrada en torno a la media y se hace muy pequeña conforme nos alejamos del centro por la derecha o la izquierda (las 'colas' de la distribución). Cuanto más alejado es el valor del centro de la función de densidad menos probable es observar ese valor.

Dos parámetros determinan una distribución normal: la media y la desviación estándar. Por tanto, puede ser adecuado hablar de las distribuciones normales, en plural, y decir que son una familia biparamétrica de distribuciones. Luego veremos que la más simple de ellas juega un papel destacado. Se denota por **N(μ, σ).**

μ	media
σ	desviación estándar
σ^2	varianza

Si una variable aleatoria sigue una distribución normal podemos escribirlo con esta notación:

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

La media de la distribución determina el centro de la gráfica de la función de densidad.

Si cambiamos la media la forma de la gráfica no cambia, simplemente se traslada a derecha o izquierda.

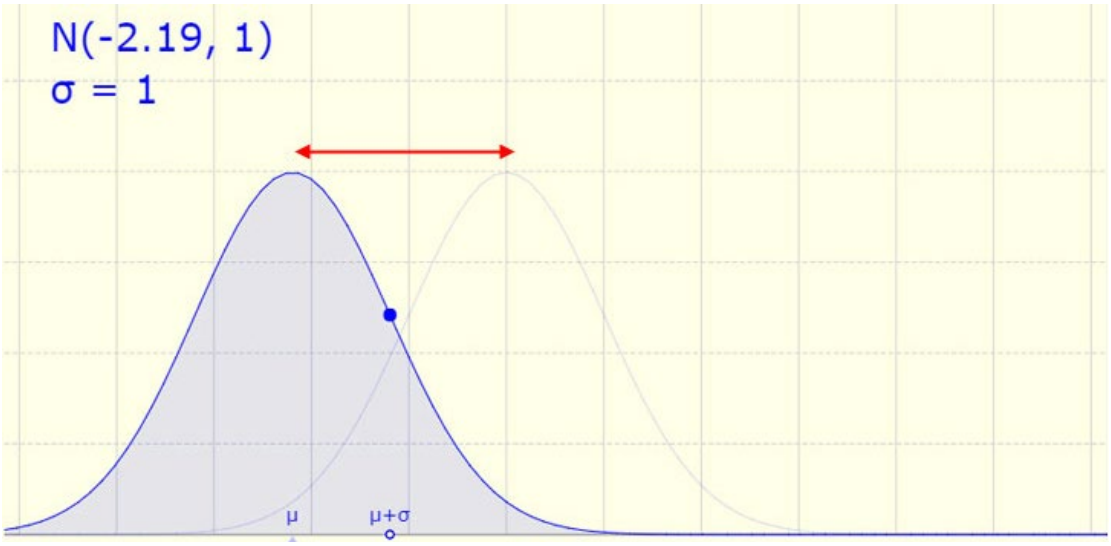


imagen: www.matematicasvisuales.com

La función de densidad tiene dos puntos de inflexión que están localizados a una distancia de la media de una desviación típica (más y menos).

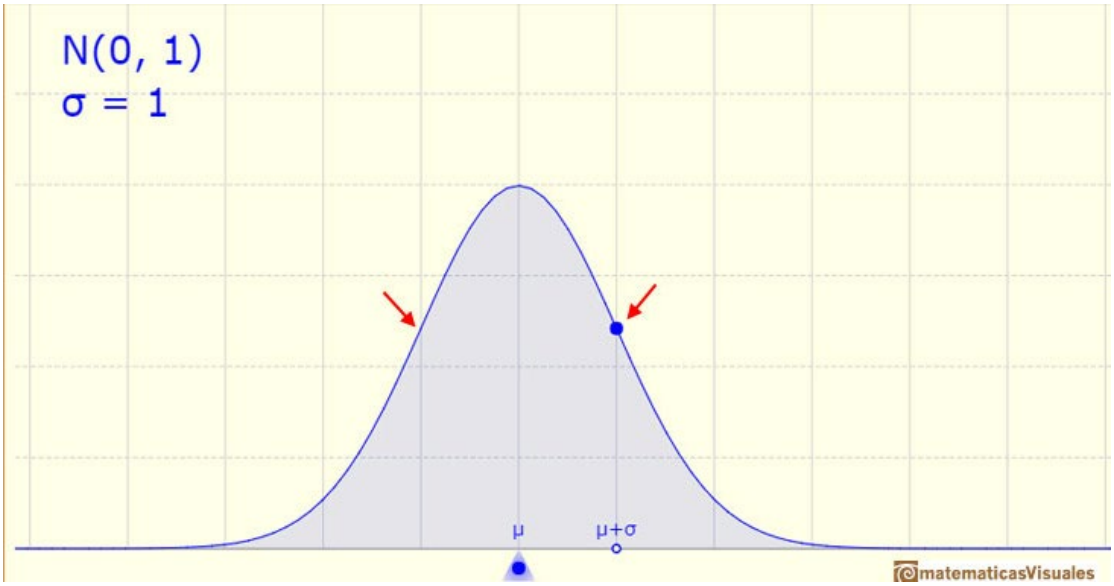


imagen: www.matematicasvisuales.com

Aumentando la desviación estándar (si no modificamos la media, el centro de la gráfica no cambia) la forma de la curva cambia. La curva se hace más ancha y menos alta, es decir, la dispersión aumenta. Cuanto mayor es la desviación estándar mayor es la dispersión de la variable.

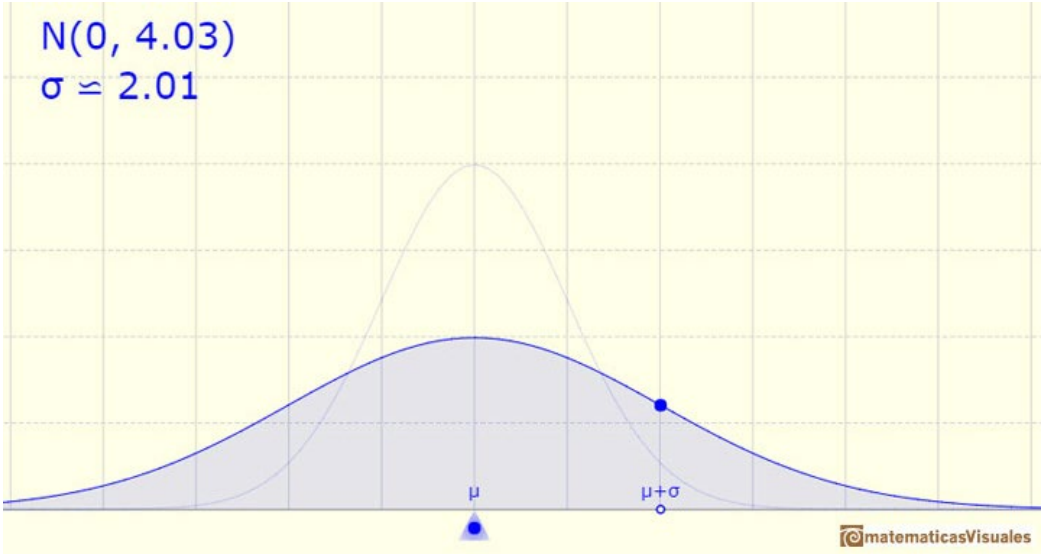


imagen: www.matematicasvisuales.com

Si la desviación estándar es pequeña la curva es más alta y estrecha. La dispersión de la variable es menor.

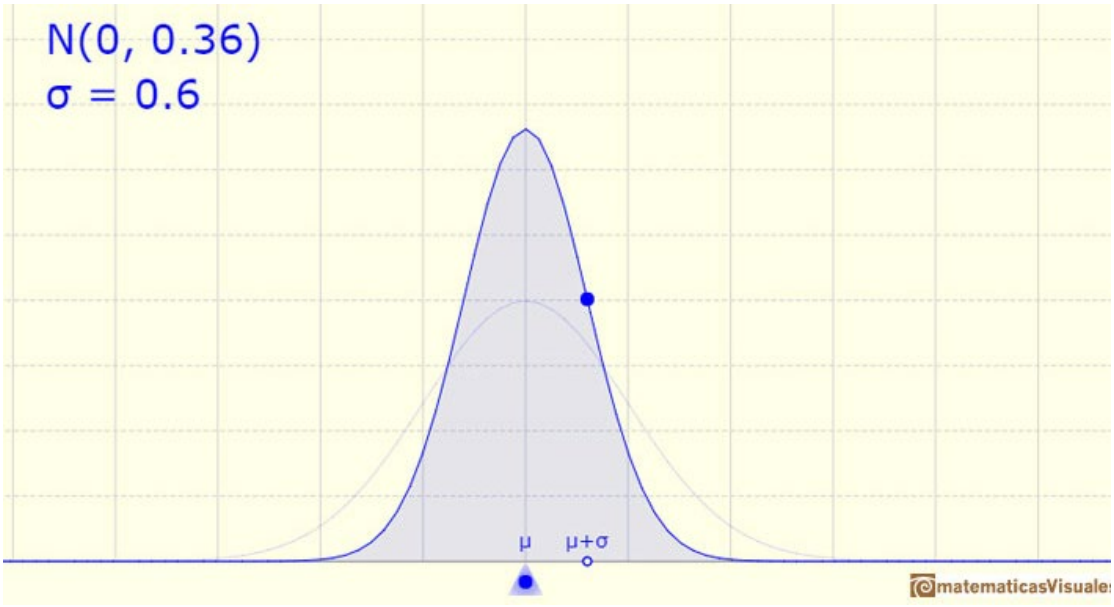


imagen: www.matematicasvisuales.com

El ejemplo más sencillo es la llamada distribución normal estándar. Es el caso especial cuando la media es igual a 0 y la varianza es 1.

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

Juega un papel importante en los cálculos a través de un proceso que llamamos estandarización o tipificación de la variable.

La función de distribución (a veces se añade la palabra 'acumulada') tiene forma de S. A cada valor de x le corresponde la probabilidad de que la variable aleatoria X tome valores menores o iguales a x. A partir de la función de densidad (en nuestro caso, la 'campana'), para calcular los valores de la función de distribución se calcula el área bajo la curva desde menos infinito hasta x. Se trata de una integral que, en el caso de la distribución normal, sólo puede hacerse numéricamente.

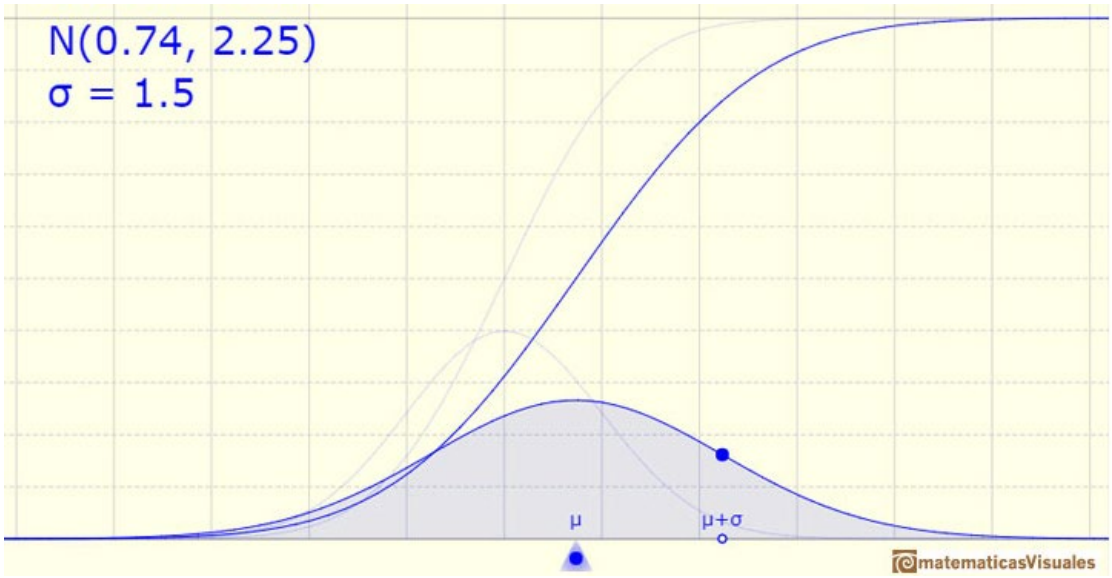


imagen: www.matematicasvisuales.com

El caso particularmente importante es el de la distribución normal estándar. Usamos tablas y ordenadores para hacer estos cálculos. Una notación habitual para este caso de la función de distribución es:

$$\Phi(x) = P[Z \leq x]$$

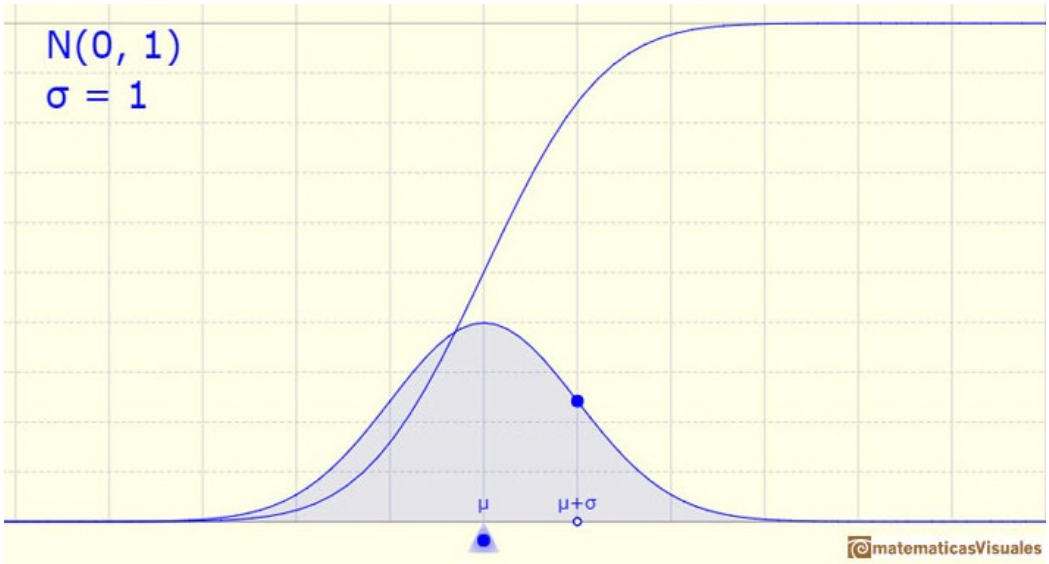


imagen: www.matematicasvisuales.com

La media está representada por un triángulo y se puede interpretar como un punto de equilibrio. Al arrastrarlo se modifica también la media.

Arrastrando el punto sobre la curva (que es uno de los dos puntos de inflexión de la curva) se modifica la desviación típica.

Podemos ver la función de distribución acumulada y cómo cambia al modificar la media (simple traslación) y la desviación típica (reflejando la mayor o menor dispersión de la variable).

Los puntos rojos controlan la escala vertical y horizontal de la gráfica.

"La distribución normal fue estudiada por Gauss (1809) en relación con la distribución de los errores en medidas astronómicas. Por este motivo se usa a veces el termino 'campana de Gauss' para referirnos a la función de densidad. Dos antecedentes importantes son de Moivre(1738) y Laplace (1774)."

Características de la distribución normal

Las propiedades de la **distribución normal** son un conjunto de características que describen la distribución normal.

En otras palabras, las propiedades de la distribución normal son el motivo por el cual esta distribución es tan versátil y utilizada en gran magnitud.

La distribución normal es un modelo teórico capaz de aproximar satisfactoriamente un valor de una variable aleatoria a un valor real. En otras palabras, la distribución normal adapta una variable aleatoria a una función que depende de la **media** y la **desviación típica.** Es decir, la **función** y la **variable aleatoria** tendrán la misma representación pero con ligeras diferencias.

La distribución normal es muy conocida y se emplea en la mayoría de los casos porque gran parte de las asunciones y teoría **estadística** se basa en la distribución normal. A destacar, la distribución normal es simétrica, sólo depende de dos parámetros y tiene una única moda (unimodal).

Características

1. Simétrica respecto a su media. En otras palabras, la media actúa como espejo en la distribución y hace que ambas colas sean idénticas y, por tanto, simétricas.
2. Media = Moda = Mediana. Las medidas de centralización son iguales porque la distribución es simétrica.
3. La distribución cambia de curvatura o tiene puntos de inflexión en los puntos del eje horizontal:

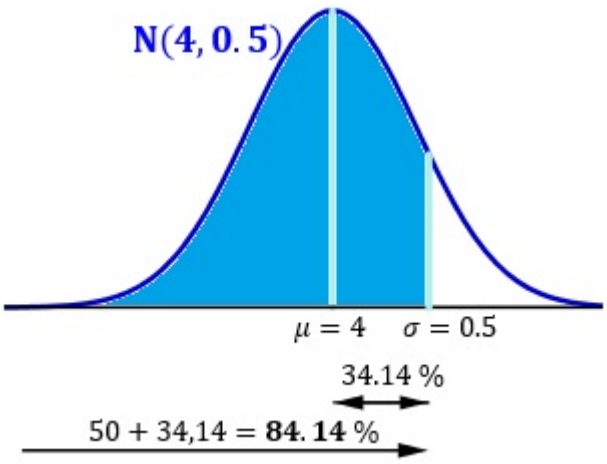
Cálculo de probabilidad con la distribución normal

El tiempo medio de duración de una batería de la flota de vehículos de una gran empresa es de cuatro años, con una desviación típica de medio año. ¿Qué probabilidad hay de que la batería de un vehículo escogido al azar haya durado igual o menos de cuatro años y medio?

Este caso no requiere recurrir a estandarizar y uso de tablas, ya que se sabe que en una distribución normal, la probabilidad de encontrar un suceso comprendido entre la media y la desviación típica es la mitad del 68,26 %, o sea del 34,14 %. Por lo tanto, la probabilidad buscada será:

$$P(X < 4{,}5) = 50\,\% + 34{,}13\% = 84{,}16\,\%$$

Gráficamente:



Habrá un 84,14 % de probabilidades de elegir al azar una batería que haya durado igual o menos de cuatro años y medio.

Distribución normal estándar

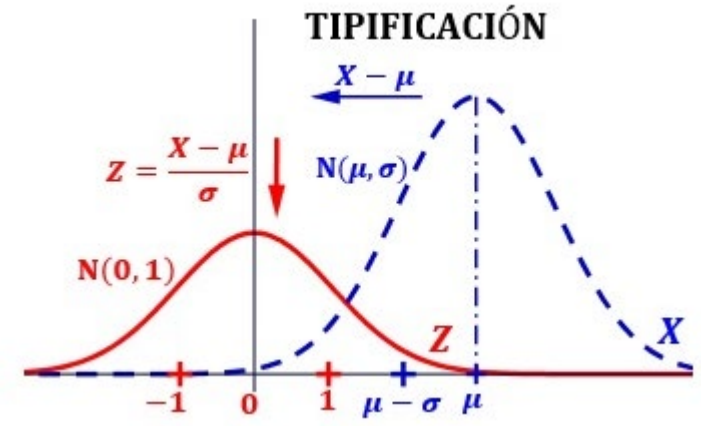
La **distribución normal estándar** o **distribución normal tipificada** es una distribución normal singular cuya denominación es **N(0, 1)**. Su variable, **Z** es el producto de una transformación o cambio de variable de la variable **X**. Esta transformación se llama **tipificación**:

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

La función de densidad de probabilidad de la distribución normal estándar o tipificada es:

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

Con la tipificación, respecto a las gráficas, se produce un desplazamiento horizontal hacia el centro de coordenadas (0, 0) y un desplazamiento en la forma vertical (hacia arriba o hacia abajo):



Es decir, de la distribución normal con variable *X*:

$$X \rightarrow \mathrm{N}(\mu, \sigma)$$

Después de la tipificación se llega a la **distribución normal estándar**:

$$Z \rightarrow \mathrm{N}(0, 1)$$

Al tipificar una variable *X* y llegar a una distribución normal estandarizada, se consigue la comparación entre distribuciones diferentes, al tiempo que se facilita el cálculo de la probabilidad mediante el uso de una **tabla normal estándar**.

La tabla que se ofrece a continuación proporciona directamente la probabilidad acumulada de que un suceso sea igual o menor que un valor positivo de *Z*:

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Distribución Binomial

Una distribución binomial es una distribución de probabilidad discreta que describe el número de éxitos al realizar n experimentos independientes entre sí, acerca de una variable aleatoria.

Existen una gran diversidad de experimentos o sucesos que pueden ser caracterizados bajo esta distribución de probabilidad. Imaginemos el lanzamiento de una moneda en el que definimos el suceso “sacar cara” como el éxito. Si lanzamos 5 veces la moneda y contamos los éxitos (sacar cara) que obtenemos, nuestra distribución de probabilidades se ajustaría a una distribución binomial.

Por lo tanto, la distribución binomial se entiende como una serie de pruebas o ensayos en la que solo podemos tener 2 resultados (éxito o fracaso), siendo el éxito nuestra variable aleatoria.

$$p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

Siendo:

n es el número de pruebas.

k es el número de éxitos.

p es la probabilidad de éxito.

q= **1-p** es la probabilidad de fracaso.

Características de la distribución binomial

Para que una variable aleatoria se considere que sigue una distribución binomial, tiene que cumplir las siguientes propiedades:

- En cada ensayo, experimento o prueba solo son posibles dos resultados (éxito o fracaso).
- La probabilidad del éxito ha de ser constante. Esta se representa mediante la letra p. La probabilidad de que salga cara al lanzar una moneda es 0,5 y esta es constante dado que la moneda no cambia en cada experimento y las probabilidades de sacar cara son constantes.
- La probabilidad de fracaso ha de ser también constante. Esta se representa mediante la letra q = 1-p. Es importante fijarse que mediante esa ecuación, sabiendo p o sabiendo q, podemos obtener la que nos falte.
- El resultado obtenido en cada experimento es independiente del anterior. Por lo tanto, lo que ocurra en cada experimento no afecta a los siguientes.
- Los sucesos son mutuamente excluyentes, es decir, no pueden ocurrir los 2 al mismo tiempo. No se puede ser hombre y mujer al mismo tiempo o que al lanzar una moneda salga cara y cruz al mismo tiempo.
- Los sucesos son colectivamente exhaustivos, es decir, al menos uno de los 2 ha de ocurrir. Si no se es hombre, se es mujer y, si se lanza una moneda, si no sale cara ha de salir cruz.
- La variable aleatoria que sigue una distribución binomial se suele representar como X~(n,p), donde n representa el número de ensayos o experimentos y p la probabilidad de éxito.

Distribución de probabilidad acumulada

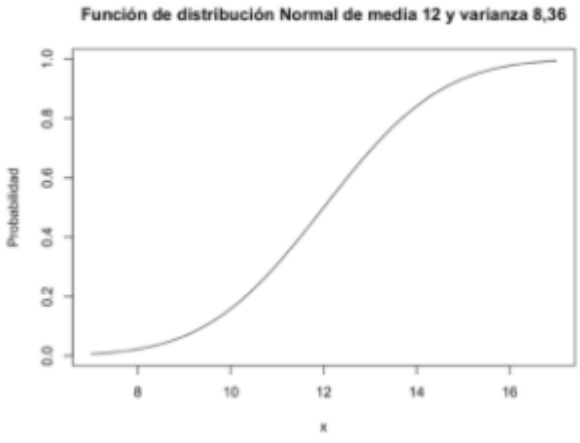
La distribución de probabilidad acumulada (FDA) es una **función matemática** que depende de una variable aleatoria real y de una distribución de probabilidad determinada que devuelve la probabilidad de que la variable sea igual o menor que un valor concreto.

En otras palabras, la distribución de probabilidad acumulada es una función matemática que se emplea para saber la probabilidad de que una variable aleatoria tome valores más pequeños o iguales que un número en concreto, sea cual sea su distribución.

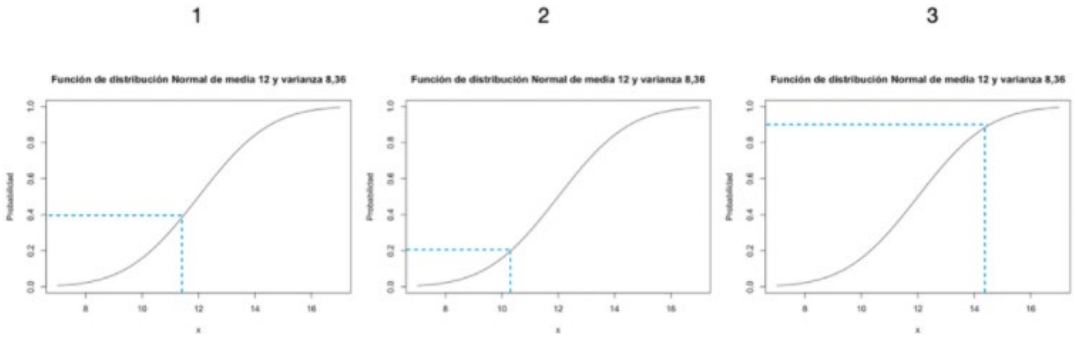
La distribución de probabilidad acumulada también recibe el nombre de **función de distribución** (FD) y acostumbra a denotarse como F(x) para diferenciarla de la función de densidad f(x).

Ejemplo:

Sitúa en la siguiente gráfica las siguientes probabilidades:



1. 40%
2. 20%
3. 90%



A diferencia de la función de densidad de probabilidad, en la función de distribución las probabilidades son puntos de la curva y no áreas. Este ejercicio se podría plantear también sabiendo la observación (eje horizontal) y buscar la probabilidad asociada.

Cálculo de probabilidades con la distribución binominal

Ejemplo:

Un jugador encesta con probabilidad 0.55. Calcula la probabilidad de que al tirar 6 veces enceste:

a) 4 veces. b) todas las veces c) ninguna vez

Desarrollo:

- Se define la Distribución binomial:

  B(n,p)

  n= número de intentos, p= probabilidad de éxitos. q=1-0.55 = 0.45

  Entonces,

  B(n,p) = B(6, 0.55)

a. P(x=4) = 6 ncr 4 x 0.55^4 x 0.45^2 = 15 * 0.55^4 * 0.45^2 = 0.2779

b. P(x=6) = 6 ncr 6 x 0.55^6 x 0.45^0 = 1 * 0.55^6 * 0.45^0 = 0.0277

c. P (x=0) = 6 ncr 0 x 0.55^0 x 0.45^6 = 1 * 0.55^0 * 0.45^6 = 0.0083

Referencias

[1] Variables aleatorias

http://www.hrc.es/bioest/estadis_21.html

[2] Distribución Normal

https://economipedia.com/definiciones/propiedades-de-la-distribucion-normal.html

[2] Distribución Binomial

https://www.youtube.com/watch?v=ckTxTTxkZpg

https://www.profesorenlinea.cl/matematica/Distribucion_binomial.html