Variable aleatoria continua y distribución normal

CURSO	**TEMA**	**WWW.DANIPARTAL.NET**
1ºBach	Estadística 03	Colegio Marista “La Inmaculada” de Granada

INFORMACIÓN GENERAL

Variables aleatoria continua. Función de densidad y función de distribución normal o gaussiana. Tipificar la variable aleatoria. Ejercicios para operar con la tabla de probabilidad acumulada. Aproximación de la binomial a la normal (Teorema central del límite).

Vídeo asociado:

https://youtu.be/Ao7hjpPWj9E

VARIABLE CONTINUA

Sabemos que, si la variable toma un conjunto de valores continuo, el número de casos totales se hace tan grande que al aplicar la regla de Laplace la probabilidad de un caso favorable tiende a cero. Por ejemplo: probabilidad de que un alumno obtenga una nota de 6,325 en un examen de matemáticas realizado a 100 alumnos de una asignatura.

Que la probabilidad según la regla de Laplace tienda a 0 no significa que el suceso sea imposible. Significa que el total de casos posibles es muy, muy, muy grande en comparación con un único caso favorable.

Al igual que hicimos con variable discreta, podemos definir una función de probabilidad, una función de distribución y una serie de parámetros estadísticos

DISTRIBUCIÓN DE PROBABILIDAD PARA VARIABLES ALEATORIAS CONTINUAS

Como la probabilidad de un valor concreto x_i en una variable aleatoria tiende a 0, no tiene sentido hablar de función de probabilidad P(X) para una valor concreto X=x_i sino de una función densidad f(X) que al ser integrada en un intervalo $[x_i, x_j]$ nos devuelve la probabilidad de que X tome un valor dentro de ese intervalo:

$$\boldsymbol{P(x_i \leq X \leq x_j) = \int_{x_i}^{x_j} f(x)dx}$$

Fíjate que si $x_i = x_j \rightarrow P(x_i \leq X \leq x_i) = \int_{x_i}^{x_i} f(x)dx = 0$

Lo cual es coherente con nuestra hipótesis de partida: la probabilidad de alcanzar un valor concreto en una variable continua es 0.

Acumulando la probabilidad para valores menores o iguales a x_i tendremos la función de distribución:

$$\boldsymbol{F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x_i} f(x)dx}$$

Donde el límite inferior “menos infinito” significa que consideramos todos los valores de la variable X inferiores a x_i. Es obvio que la probabilidad acumulada en el intervalo $[x_i, x_j]$ se obtiene a partir de la definición primera de la función de densidad.

$$\boldsymbol{F(x_j) - F(x_i) = P(x_i \leq X \leq x_j) = \int_{x_i}^{x_j} f(x)dx}$$

Si consideramos todos los valores posibles que pueda tomar la variable aleatoria X, su probabilidad acumulada total debe ser igual a 1. Por eso nuestra integral cumple:

$$\boldsymbol{\int_{-\infty}^{+\infty} f(x)dx = 1}$$

Donde "menos y más infinito" indican que se acumulan todos los valores posible de la variable de la función que estamos integrando.

PARÁMETROS DE UNA VARIABLE ALEATORIA CONTINUA

Si X es una variable aleatoria continua que toma valores en el intervalo $[a, b]$ y f(x) es su función de densidad que cumple las condiciones descritas en el apartado anterior, podemos calcular sus correspondientes parámetros estadísticos.

Media o esperanza matemática:

$$\mu = \int_a^b x \cdot f(x)dx$$

Varianza:

$$\sigma^2 = \int_a^b (x-\mu)^2 \cdot f(x)dx$$

Desviación típica: la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\int_a^b (x-\mu)^2 \cdot f(x)dx}$$

Recuerda que la media es un dato que representa el promedio del conjunto de todos los valores. El grado de acierto o no de que ese valor medio sea un buen representante de la distribución lo da, como siempre, la varianza y la desviación típica. A menor desviación típica mejor será la media como representante de la variable.

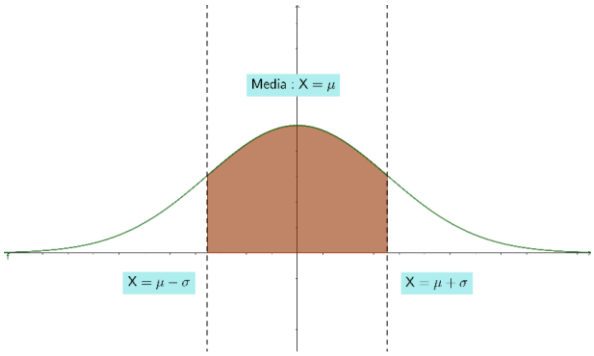
LA DISTRIBUCIÓN NORMAL O GAUSSIANA

Una variable X se dice que sigue una distribución normal N(μ, σ) de media μ y desviación típica σ si los valores posibles de la variable ocupan toda la recta real (desde $-\infty$ hasta $+\infty$) y si su función de densidad puede expresarse como:

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

Decir que X sea una variable continua siempre será un modelo ideal, ya que existen infinitos decimales en los números reales. Pero es una muy buena aproximación para multitud de problemas de estadística.

La gráfica de la distribución normal toma la forma de la conocida campana de Gauss. Es simétrica respecto al valor medio μ (la media coincide con la mediana) y muchas variables aleatorias se pueden aproximar a esta forma (peso de una población de una determinada edad, altura de las personas de un curso, etc.).



La distribución gaussiana aglomera:

- el 68,26% de los valores de la variable X en el intervalo $[\mu-\sigma, \mu+\sigma]$.
- el 95,44% de los valores de la variable X en el intervalo $[\mu-2\sigma, \mu+2\sigma]$.
- el 99,73% de los valores de la variable X en el intervalo $[\mu-3\sigma, \mu+3\sigma]$.

TIPIFICAR LA VARIABLE PARA OBTENER LA DISTRIBUCIÓN NORMAL ESTÁNDAR

¿Tenemos que estudiar integrales para calcular la probabilidad acumulada por la distribución normal en un intervalo?

Afortunadamente no. Esta distribución ha sido tan estudiada a lo largo de los años que existen tablas con los valores acumulados para diferentes intervalos. Para usar estas tablas necesitamos convertir la distribución N(μ, σ) en una distribución normal de media 0 y desviación típica 1: N(0,1).

Este proceso se conoce como tipificar la variable. Quedando la función de densidad de la variable normal tipificada:

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

¿Cómo pasar de N(μ, σ) a N(0,1)?

Definiendo una nueva variable aleatoria Z que será:

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

De esta forma la función de distribución cumple (probabilidad acumulada desde menos infinito hasta el valor concreto X=x):

$$\boldsymbol{F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = P(Z \leq \frac{x-\mu}{\sigma})}$$

Pudiendo obtener el valor $\boldsymbol{P(Z \leq \frac{x-\mu}{\sigma})}$ de la tabla normal estándar o tipificada, ya que Z es una variable que sigue la distribución N(0,1). En la siguiente hoja encontrarás una imagen de estos valores tabulados. Esta tabla te la darán el día del examen de Selectividad.

Un ejemplo:

Sea la distribución normal N(60,5). Es decir, la media vale 60 y la desviación típica 5. Calcular la probabilidad de que la variable aleatoria X sea igual o inferior a 62.

Es decir, preguntan por la probabilidad acumulada (función de distribución) siguiente:

$$P(X \leq 62)$$

Tipificamos la variable, de tal forma que:

$$P(X \leq 62) = P(Z \leq \frac{62-\mu}{\sigma}) = P(Z \leq \frac{62-60}{5})$$

$$P(X \leq 62) = P(Z \leq \frac{2}{5}) = P(Z \leq 0{,}40)$$

La probabilidad acumulada para $X \leq 62$ coincide con la probabilidad acumulada para la variable tipificada $Z \leq 0{,}40$. Mirando la tabla de la distribución normal inferior (probabilidad acumulada inferior) comprobamos:

$$P(Z \leq 0{,}40) = 0{,}6554$$

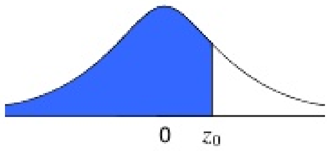
Podemos afirmar que existe una probabilidad del 65,54% de que al escoger un valor de la muestra el valor de la variable X sea igual o inferior a 62.

Tabla de la distribución normal N(0,1) para probabilidad acumulada inferior

μ = Media

σ = Desviación típica

$$P(z \le z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{\frac{-z^2}{2}} dz$$



Tipificación: $z_0 = \dfrac{x-\mu}{\sigma}$

z_0	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	z_0
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359	**0,0**
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753	**0,1**
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141	**0,2**
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517	**0,3**
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879	**0,4**
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224	**0,5**
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549	**0,6**
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852	**0,7**
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133	**0,8**
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389	**0,9**
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621	**1,0**
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830	**1,1**
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015	**1,2**
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177	**1,3**
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319	**1,4**
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441	**1,5**
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545	**1,6**
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633	**1,7**
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706	**1,8**
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767	**1,9**
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817	**2,0**
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857	**2,1**
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890	**2,2**
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916	**2,3**
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936	**2,4**
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952	**2,5**
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964	**2,6**
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974	**2,7**
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981	**2,8**
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986	**2,9**
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900	**3,0**
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929	**3,1**
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950	**3,2**
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965	**3,3**
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976	**3,4**
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983	**3,5**
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989	**3,6**
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992	**3,7**
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995	**3,8**
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997	**3,9**

1−α	90%	92%	94%	95%	96%	97%	98%	99%
α	10%	8%	6%	5%	4%	3%	2%	1%
$z_{\alpha/2}$	1,645	1,751	1,881	1,960	2,054	2,170	2,326	2,576
z_{α}	1,282	1,405	1,555	1,645	1,751	1,881	2,054	2,326

Siendo:

1 − α = Nivel de confianza

α = Nivel de significación

www.vaxasoftware.com/indexes.html

Siguiendo con el ejemplo anterior de la distribución normal N(60,5).

Pueden preguntarnos por la probabilidad de que la variable aleatoria esté por encima de un valor concreto. Por ejemplo:

$$P(X \geq 62)$$

Si la tabla tipificada nos da los valores acumulados inferiormente, ¿cómo obtener la probabilidad acumulada superior (por encima de x=62)?

Sabemos que la probabilidad total es 1 y que la distribución normal es simétrica respecto la media, por lo que se cumple:

$$\boldsymbol{P(Z \geq \frac{x-\mu}{\sigma}) = 1 - P(Z \leq \frac{x-\mu}{\sigma})}$$

Anteriormente ya habíamos tipificado la variable y obtenido el valor de:

$$P(X \leq 62) = P(Z \leq \frac{62-60}{5}) = P(Z \leq 0{,}40) = 0{,}6554$$

Por lo tanto:

$$P(X \geq 62) = P(Z \geq \frac{62-60}{5}) = P(Z \geq 0{,}40) = 1 - P(Z \leq 0{,}40)$$

$$P(Z \geq 0{,}40) = 1 - P(Z \leq 0{,}40) = 1 - 0{,}6554 = 0{,}3446$$

Resultando una probabilidad del 34,56%.

¿Qué ocurre si nos preguntan por un valor tipificado Z negativo (por debajo del valor medio z=0)?

La tabla tipificada ofrece valor a partir de z=0. Pero al ser simétrica respecto a la media, podemos razonar de la siguiente forma **(ojo con el sentido de las desigualdades)**:

$$\boldsymbol{P(Z \leq -k) = P(Z \geq k) = 1 - P(Z \leq k)}$$

Siendo "k" un valor real positivo.

Por ejemplo:

$$P(Z \leq -0{,}40) = P(Z \geq 0{,}40) = 1 - P(Z \leq 0{,}40) = 1 - 0{,}6554 = 0{,}3446$$

La probabilidad sería del 34,56%.

¿Cómo calcular la probabilidad acumulada superior de un valor tipificado Z negativo?

Nuevamente usamos la simetría alrededor de la media de la distribución normal.

$$\boldsymbol{P(Z \geq -k) = P(Z \leq k)}$$

Por ejemplo:

$$P(Z \geq -0{,}40) = P(Z \leq 0{,}40) = 0{,}6554$$

La probabilidad sería del 65,54%.

¿Y si buscamos la probabilidad acumulada en un intervalo $[\boldsymbol{a}, \boldsymbol{b}]$?

Muy sencillo: a la probabilidad acumulada para el extremo superior $x = b$ le restamos la probabilidad acumulada para el extremo inferior $x = a$.

$$\boldsymbol{P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)}$$

Por ejemplo:

$$P(0{,}25 \leq Z \leq 0{,}40) = P(Z \leq 0{,}40) - P(Z \leq 0{,}25) = 0{,}6554 - 0{,}5987 = 0{,}0567$$

La probabilidad acumulada en el intervalo sería del 5,67%.

OBTENER VALOR DE LA VARIABLE TIPIFICADA Z QUE ACUMULA UNA DETERMINADA PROBABILIDAD CONOCIDA

En los ejemplos resueltos anteriormente nos daban un valor de Z y nos pedían obtener la probabilidad acumulada por debajo o por encima de ese valor de Z. Ahora nos plantean el problema inversa: nos dan una probabilidad acumulada "p" y debemos obtener el valor de Z que acumula dicha probabilidad "p".

Ejemplo:

Dada una distribución normal N(0,1) obtener el valor de la variable tipificada Z cuya función de distribución acumula el 75% de las observaciones.

Miramos en la tabla y localizamos el valor más cercano a 0,75 y que lo supere. Y leemos:

- Para z = 0,67 la probabilidad acumulada inferior es 0,7486.
- Para z = 0,68 la probabilidad acumulada inferior es 0,7517

Por lo tanto, aproximando según los valores de la tabla, el valor que acumula el 75% de las observaciones sería z = 0,68.

Si quisiéramos obtener un valor más preciso, podríamos plantear una interpolación lineal entre los valores 0,67 y 0,68. Pero, de cara a Selectividad, es suficiente razonar con los datos de la tabla tal y como hemos descrito.

OBTENER VALOR DE LA MEDIA O DE LA DESVIACIÓN TÍPICA A PARTIR DE UN DATO DE LA DISTRIBUCIÓN NORMAL TIPIFICADA

Si en un ejercicio no conocemos la media o la desviación típica de la distribución normal, pero sí conocemos un valor de probabilidad acumulada para un dato concreto de la variable, podemos deducir el dato desconocido a partir de la expresión que tipifica a la variable.

Ejemplo:

Supongamos una distribución normal de media μ desconocida y de desviación típica igual a 17. Sabemos que el valor x=200 de la variable aleatoria acumula el 40% de las observaciones. ¿Cuánto vale la media de la distribución?

Debemos tipificar la variable X a la variable Z con media 0 y desviación 1. Para un valor concreto "x" podemos normalizar con la expresión:

$$z = \frac{x - \mu}{\sigma}$$

Para x=200 y $\sigma = 17$ tendremos:

$$z = \frac{200 - \mu}{17}$$

El valor "z" podemos obtenerlo del dato de que, en la curva tipificada, debe acumular el 0,40. Como la tabla de probabilidad acumulada con la que solemos trabajar da los resultados a partir de 0,50 tendremos que razonar así: el valor "z" que buscamos será simétrico respecto a la media z=0 de aquel valor que deje por encima de él un 40%. Y el valor que deja por encima de él un 40% es el que acumula un 60%. Por lo tanto:

$$P(Z = z) \leq 0{,}60$$

Mirando la tabla, observamos:

$$P(Z = 0{,}25) \leq 0{,}5987$$

$$P(Z = 0{,}26) \leq 0{,}6026$$

Aproximamos al valor z=0,26. Por lo que el valor simétrico respecto a la media z=0 será z=-0,26. Si llevamos este valor a la expresión tipificada podremos despejar el valor de la media.

$$-0{,}26 = \frac{200 - \mu}{17}$$

$$-4{,}42 = 200 - \mu$$

$$\mu = 204{,}42$$

¿CUÁNDO APROXIMAR LA BINOMIAL A LA NORMAL? TEOREMA CENTRAL DEL LÍMITE

La distribución normal es tan importante porque muchas distribuciones, bajo determinadas condiciones, pueden aproximarse a la normal. Incluso distribuciones de variable discreta, como la binomial, pueden aproximarse a la distribución normal que es de variable continua.

Esto es una consecuencia del Teorema central del límite. De este teorema vamos a tomar únicamente la siguiente aplicación:

Si el tamaño de la muestra "n" es muy grande (superior a 30) y la probabilidad "p" no se acerca ni a 0 ni a 1, podemos aproximar la variable aleatoria X que sigue una distribución normal B(n,p) a una variable continua X que sigue una distribución normal $N(n \cdot p, \sqrt{n \cdot p \cdot (1-p)})$.

Esta aproximación es válida si:

$n \cdot p \geq 5$

$\mathrm{n} \cdot (1-\mathrm{p}) \geq 5$

Esta aproximación es muy útil cuando debamos calcular la probabilidad de un intervalo de éxitos tras varios ensayos, lo cual llevaría con las ecuaciones binomiales a tener que repetir varias veces las operaciones para cada uno de los valores contenidos dentro de dicho intervalo. Con la aproximación a la normal se ahorra mucho tiempo de operaciones.

Ejemplo:

El 5% de los teléfonos móviles fabricados por una empresa no satisface las especificaciones técnicas.

Se extrae al azar una muestra de 2.000 móviles y queremos hallar la probabilidad de que la muestra contenga más de 120 teléfonos defectuosos.

La variable X sigue una distribución binomial B(2.000,0,05) ya que en cada ensayo solo hay dos posibilidades: funcionamiento correcto o incorrecto.

Como n=2.000 y p=0,05 se cumple:

$2.000 \cdot 0{,}05 = 100 \geq 5$

$2.000 \cdot (1 - 0{,}05) = 1.900 \geq 5$

Por lo que podemos aproximar la binomial B(2.000,0,05) a la normal $N(100, \sqrt{95})$.

De esta forma:

$$P(X > 120) = P(Z > \frac{120 - 100}{9{,}7}) = P(Z > 2{,}06) = 1 - P(Z \leq 2{,}06) = 1 - 0{,}9803 = 0{,}0197$$

La probabilidad solución es del 1,97%.

Un detalle: hemos considerado $P(X > 120)$ sin el signo igual porque, al ser una aproximación de binomial a normal, no debemos olvidar que en una variable discreta la probabilidad de un valor concreto no tiende a 0 (como sí pasa en la variable continua). Por eso el enunciado indica que buscamos la probabilidad de superar los 120 teléfonos defectuosos, y no dice que deba ser mayor o igual.

Si quisiéramos en esta aproximación obtener el valor concreto de P(X=120) podríamos razonar de la siguiente manera: tomar en la aproximación normal un intervalo de amplitud 1 centrado en el valor X=120:

$P(X = 120) \cong P(120 - 0{,}5 \leq X \leq 120 + 0{,}5)$