Manual Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Sep 2017

Departamento de Matemática Aplicada y Estadística
CEU San Pablo

CEU
Universidad
San Pablo

Términos de la licencia

Contenidos

Variables aleatorias continuas

Variables aleatorias continuas

Las variables aleatorias continuas, a diferencia de las discretas, se caracterizan porque pueden tomar cualquier valor en un intervalo real. Es decir, el conjunto de valores que pueden tomar no sólo es infinito, sino que además es no numerable.

Tal densidad de valores hace imposible el cálculo de las probabilidades de cada uno de ellos, y por tanto no podemos definir los modelos de distribución de probabilidad por medio de una función de probabilidad como en el caso discreto.

Por otro lado, la medida de una variable aleatoria continua suele estar limitada por la precisión del proceso o instrumento de medida. Por ejemplo, cuando se dice que una estatura es 1.68 m, no se está diciendo que es exactamente 1.68 m, sino que la estatura está entre 1.675 y 1.685 m, ya que el instrumento de medida sólo es capaz de precisar hasta cm.

Así pues, en el caso de variables continuas, *no tiene sentido medir probabilidades de valores aislados, sino que se medirán probabilidades de intervalos.*

Función de densidad de probabilidad

Para conocer cómo se distribuye la probabilidad entre los valores de una variable aleatoria continua se utiliza la función de densidad.

Definición (Función de densidad de probabilidad)

La *función de densidad de probabilidad* de una variable aleatoria continua X es una función $f(x)$ que cumple las siguientes propiedades:

- Es no negativa: $f(x) \geq 0 \ \forall x \in \mathbb{R}$,
- El área acumulada entre la función y el eje de abscisas es 1, es decir,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

- Las probabilidad de que X tome un valor entre a y b es igual al área que queda por debajo de la función de densidad y el eje de abcisas limitada por a y b, es decir,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

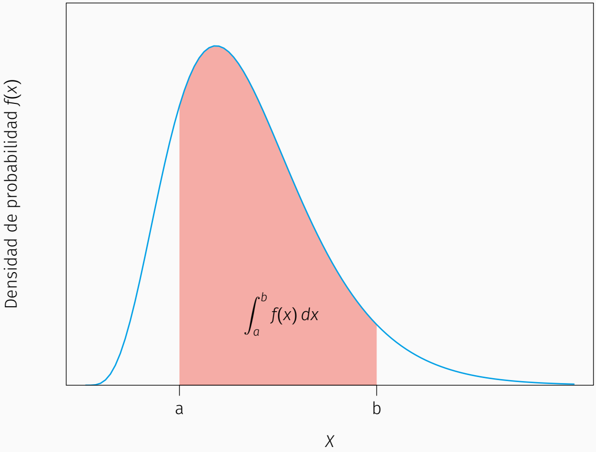
Al igual que para las variables discretas, también tiene sentido medir probabilidades acumuladas por debajo de un determinado valor.

Definición (Función de distribución)

La *función de distribución* de una variable aleatoria continua X es una función $F(x)$ que asocia a cada valor a la probabilidad de que la variable tome un valor menor o igual que a, es decir,

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)\, dx.$$

Cálculo de probabilidades como áreas



$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Dada la siguiente función:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0, \end{cases}$$

veamos que se trata de una función de densidad.

Como es evidente que la función es no negativa, solo que da por comprobar que el área por debajo de ella es 1.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} e^{-x}\,dx =$$
$$= \left[-e^{-x}\right]_0^{\infty} = -e^{-\infty} + e^0 = 1.$$

Calculemos la probabilidad de que la variable tome un valor entre 0 y 2.

$$P(0 \leq X \leq 2) = \int_0^2 f(x)\,dx = \int_0^2 e^{-x}\,dx = \left[-e^{-x}\right]_0^2 = -e^{-2} + e^0 = 0.8646.$$

Estadísticos poblacionales

El cálculo de los estadísticos poblacionales es similar al caso discreto, pero utilizando la función de densidad en lugar de la función de probabilidad, y extendiendo la suma discreta a la integral en todo el rango de la variable.

Los más importantes son:

- Media o esperanza matemática:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

- Varianza:

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

- Desviación típica:

$$\sigma = +\sqrt{\sigma^2}$$

Sea X una variable aleatoria con la siguiente función de densidad:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0 \end{cases}$$

Su media es

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{0} xf(x)\,dx + \int_{0}^{\infty} xf(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} xe^{-x}\,dx =$$
$$= \left[-e^{-x}(1+x)\right]_0^{\infty} = 1.$$

y su varianza vale

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_{-\infty}^{0} x^2 f(x)\,dx + \int_{0}^{\infty} x^2 f(x)\,dx - \mu^2 =$$
$$= \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} x^2 e^{-x}\,dx - \mu^2 = \left[-e^{-x}(x^2+2x+2)\right]_0^{\infty} - 1^2 = 2e^0 - 1 = 1.$$

Modelos de distribución continuos

Existen varios modelos de distribución de probabilidad que aparecen bastante a menudo en la naturaleza y también como consecuencia de los procesos de muestreo aleatorio simple. Los más importantes son:

- Distribución Uniforme continua.
- Distribución Normal.
- Distribución T de Student.
- Distribución Chi-cuadrado.
- Distribución F de Fisher-Snedecor.

Distribución Uniforme continua $U(a, b)$

Cuando todos los valores de una variable continua son equiprobables, se dice que la variable sigue un *modelo de distribución uniforme continuo*.

Definición (Distribución Uniforme continua $U(a, b)$)

Una variable aleatoria continua X, sigue un modelo de distribución de probabilidad *uniforme*, y se nota $X \sim U(a, b)$, si su rango es $\operatorname{Ran}(X) = [a, b]$ y su función de densidad es

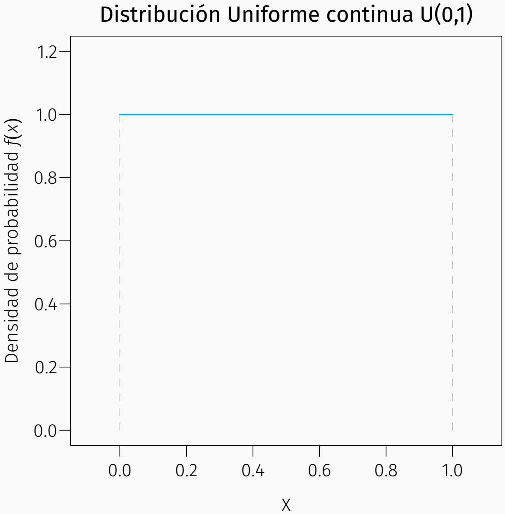
$$f(x) = \frac{1}{b-a} \quad \forall x \in [a, b]$$

Obsérvese que a y b son el mínimo y el máximo del rango respectivamente, y que la función de densidad es constante.

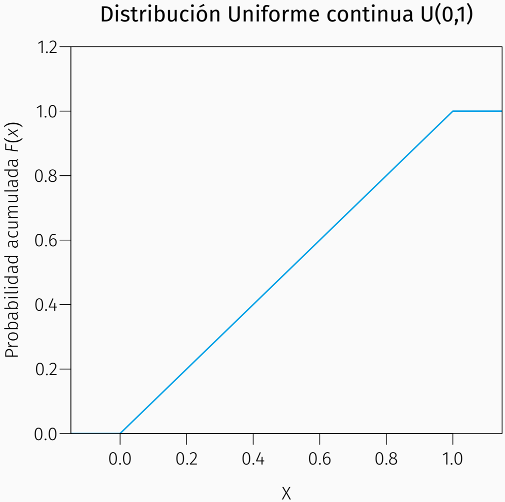
Su media y varianza valen

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}.$$

La generación aleatoria de un número real entre 0 y 1 sigue un modelo de distribución uniforme continuo $U(0, 1)$.



Como la función de densidad es constante, la función de distribución presenta un crecimiento lineal.

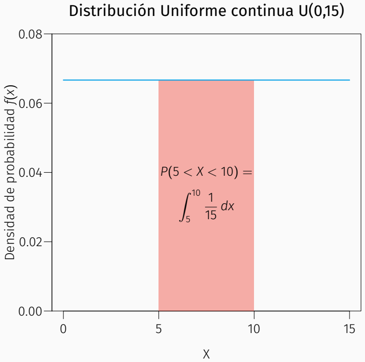


Cálculo de probabilidades con una Uniforme continua
Ejemplo de espera de un autobús

Un autobús pasa por una parada cada 15 minutos. Si una persona puede llegar a la parada en cualquier instante, *¿cuál es la probabilidad de que espere entre 5 y 10 minutos?*

En este caso, la variable X que mide el tiempo de espera sigue un modelo de distribución uniforme continua $U(0,15)$ ya que cualquier valor entre los 0 y los 15 minutos es equipobrable. Así pues, la probabilidad de esperar entre 5 y 10 minutos es

$$P(5 \leq X \leq 10) = \int_5^{10} \frac{1}{15}\, dx = \left[\frac{X}{15}\right]_5^{10} = \\ = \frac{10}{15} - \frac{5}{15} = \frac{1}{3}.$$



Además, el tiempo medio de espera será $\mu = \frac{0+15}{2} = 7.5$ minutos.

Distribución de probabilidad Normal $N(\mu, \sigma)$

El modelo de distribución normal es, sin duda, el modelo de distribución continuo más importante, ya que es el que más a menudo se presenta en la naturaleza.

Definición (Distribución de probabilidad Normal $N(\mu, \sigma)$)

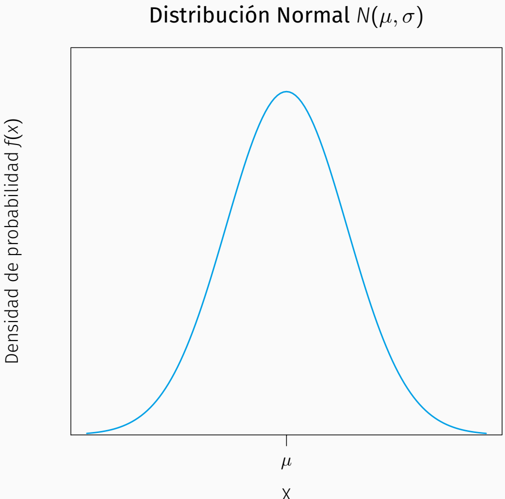
Una variable aleatoria continua X sigue un modelo de distribución *normal* de parámetros μ y σ, y se nota $X \sim N(\mu, \sigma)$, si su rango es $\mathbb{R}$ y su función de densidad vale

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

Los dos parámetros μ y σ son la media y la desviación típica de la población respectivamente.

Función de densidad de la Normal

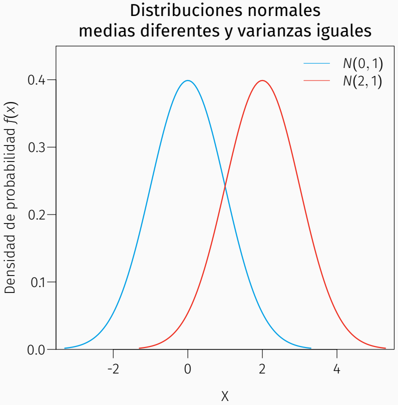
La gráfica de la función de densidad de la distribución normal tiene forma de una especie de campana, conocida como *campana de Gauss* (en honor a su descubridor).

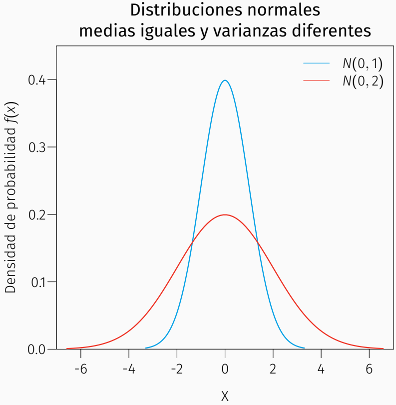


Función de densidad de la Normal

La forma de la campana de Gauss depende de sus dos parámetros:

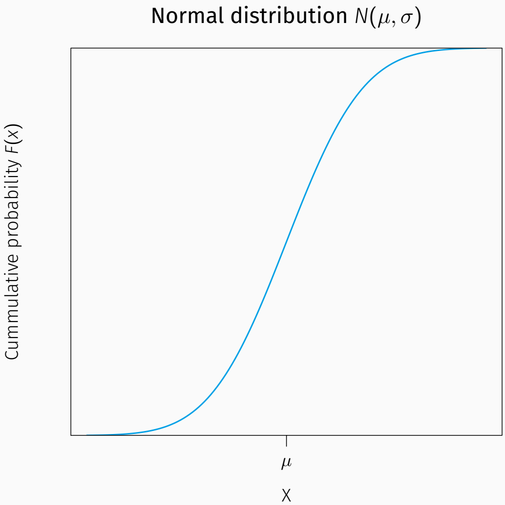
- La media μ determina dónde está centrada la campana.
- La desviación típica σ determina la anchura de la campana.





Función de distribución de la Normal

La gráfica de la función de distribución tiene forma de S.



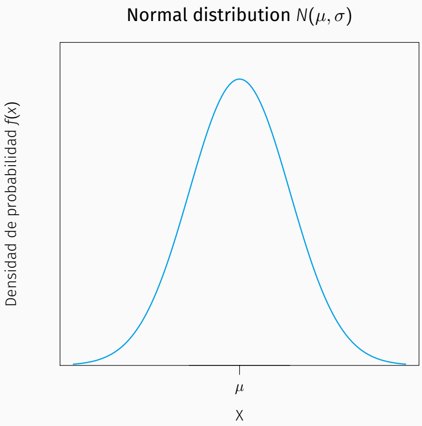
Propiedades de la distribución Normal

- La función de densidad es simétrica respecto a la media y por tanto, su coeficiente de asimetría es $g_1 = 0$.
- También es mesocúrtica, ya que la función de densidad tiene forma de campana de Gauss, y por tanto, su coeficiente de apuntamiento vale $g_2 = 0$.
- La media, la mediana y la moda coinciden

$$\mu = Me = Mo.$$

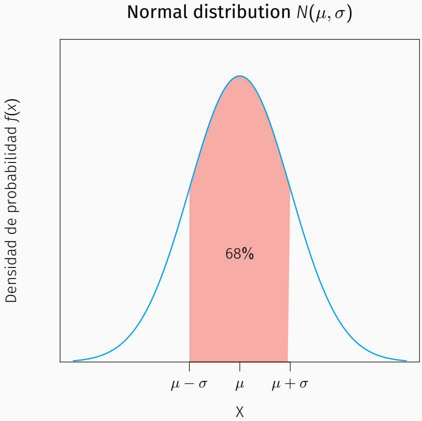
- Tiende asintóticamente a 0 cuando x tiende a $\pm\infty$.

Propiedades de la distribución Normal



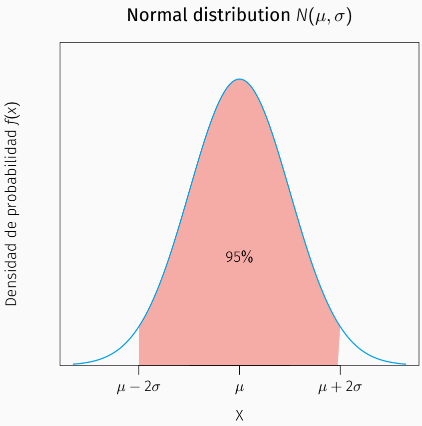
Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$



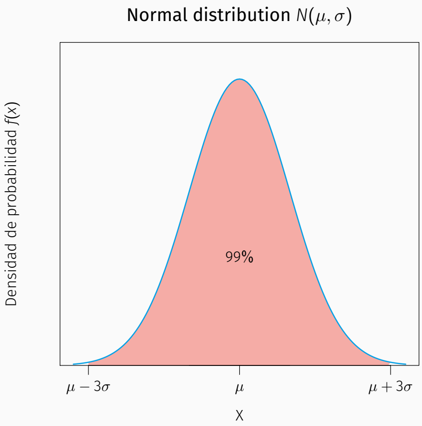
Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95,$$



Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95,$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.99.$$



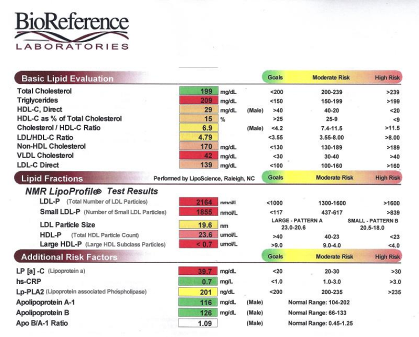
es decir, casi la totalidad de los individuos de la población presentarán

En un estudio se ha comprobado que el nivel de colesterol total en mujeres sanas de entre 40 y 50 años sigue una distribución normal de media de 210 mg/dl y desviación típica 20 mg/dl.

Atendiendo a las propiedades de la campana de Gauss, se tiene que

- El 68% de las mujeres sanas tendrán el colesterol entre 210 ± 20 mg/dl, es decir, entre 190 y 230 mg/dl.
- El 95% de las mujeres sanas tendrán el colesterol entre $210 \pm 2 \cdot 20$ mg/dl, es decir, entre 170 y 250 mg/dl.
- El 99% de las mujeres sanas tendrán el colesterol entre $210 \pm 3 \cdot 20$ mg/dl, es decir, entre 150 y 270 mg/dl.

En la analítica sanguínea suele utilizarse el intervalo $\mu \pm 2\sigma$ para detectar posibles patologías. En el caso del colesterol, dicho intervalo es $[170 \text{ mg/dl}, 250 \text{ mg/dl}]$. Cuando una persona tiene el colesterol fuera de estos límites, se tiende a pensar que tiene alguna patología, aunque ciertamente podría estar sana,pero la probabilidad de que eso ocurra es sólo de un 5%.


BioReference LABORATORIES

Basic Lipid Evaluation				Goals	Moderate Risk	High Risk
Total Cholesterol	199	mg/dL		<200	200-239	>239
Triglycerides	209	mg/dL		<150	150-199	>199
HDL-C, Direct	29	mg/dL	(Male)	>40	40-20	<20
HDL-C as % of Total Cholesterol	15	%		>25	25-9	<9
Cholesterol / HDL-C Ratio	6.9		(Male)	<4.2	7.4-11.5	>11.5
LDL/HDL-C Ratio	4.79			<3.55	3.55-8.00	>8.00
Non-HDL Cholesterol	170	mg/dL		<130	130-189	>189
VLDL Cholesterol	42	mg/dL		<30	30-40	>40
LDL-C Direct	139	mg/dL		<100	100-160	>160
Lipid Fractions Performed by LipoScience, Raleigh, NC				Goals	Moderate Risk	High Risk
NMR LipoProfile® Test Results						
LDL-P (Total Number of LDL Particles)	2164	nmol/l		<1000	1300-1600	>1600
Small LDL-P (Number of Small LDL Particles)	1855	nmol/L		<117	437-617	>839
LDL Particle Size	19.6	nm		LARGE - PATTERN A 23.0-20.6		SMALL - PATTERN B 20.5-18.0
HDL-P (Total HDL Particle Count)	23.6	umol/L		>40	40-23	<23
Large HDL-P (Large HDL Subclass Particles)	< 0.7	umol/L		>9.0	9.0-4.0	<4.0
Additional Risk Factors				Goals	Moderate Risk	High Risk
LP [a] -C (Lipoprotein a)	39.7	mg/dL		<20	20-30	>30
hs-CRP	0.7	mg/L		<1.0	1.0-3.0	>3.0
Lp-PLA2 (Lipoprotein associated Phospholipase)	201	ng/dL		<200	200-235	>235
Apolipoprotein A-1	116	mg/dL	(Male)	Normal Range: 104-202		
Apolipoprotein B	126	mg/dL	(Male)	Normal Range: 66-133		
Apo B/A-1 Ratio	1.09		(Male)	Normal Range: 0.45-1.25		

El teorema central del límite

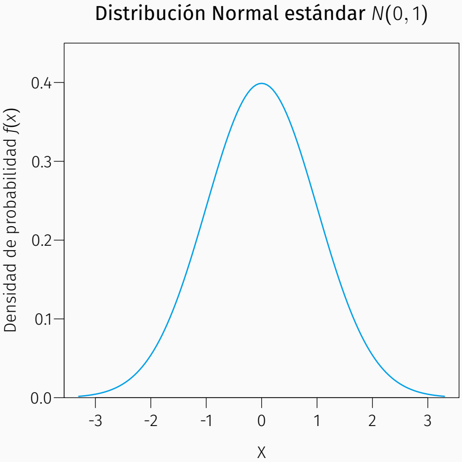
El comportamiento anterior lo presentan muchas variables continuas físicas y biológicas.

Si se piensa por ejemplo en la distribución de las estaturas, se verá que la mayor parte de los individuos presentan estaturas en torno a la media, tanto por arriba, como por debajo, pero que a medida que van alejándose de la media, cada vez hay menos individuos con dichas estaturas.

La justificación de que la distribución normal aparezca de manera tan frecuente en la naturaleza la encontramos en el **teorema central del límite**, que veremos más adelante, y que establece que si una variable aleatoria continua proviene de un experimento aleatorio cuyos resultados son debidos a un conjunto muy grande de factores independientes que actúan sumando sus efectos, entonces sigue una distribución aproximadamente normal.

La distribución Normal estándar $N(0,1)$

De todas las distribuciones normales, la más importante es la que tiene media $\mu = 0$ y desviación típica $\sigma = 1$, que se conoce como **normal estándar** y se designa por $Z \sim N(0,1)$.



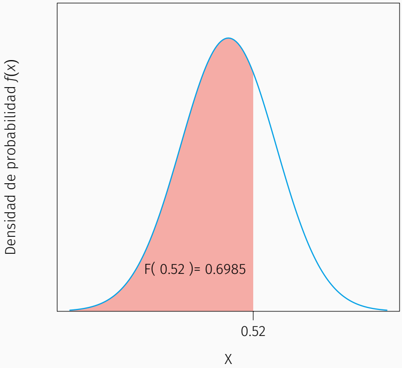
Para evitar tener que calcular probabilidades integrando la función de densidad de la normal estándar es habitual utilizar su función de distribución, que se da en forma de tabla como la siguiente.

Por ejemplo, para calcular $P(Z \leq 0.52)$

z	0.00	0.01	0.02	$\cdots$
0.0	0.5000	0.5040	0.5080	$\cdots$
0.1	0.5398	0.5438	0.5478	$\cdots$
0.2	0.5793	0.5832	0.5871	$\cdots$
0.3	0.6179	0.6217	0.6255	$\cdots$
0.4	0.6554	0.6591	0.6628	$\cdots$
0.5	0.6915	0.6950	0.6985	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

$0.52 \rightarrow$ fila 0.5 + columna 0.02

Distribución Normal estándar $N(0, 1)$

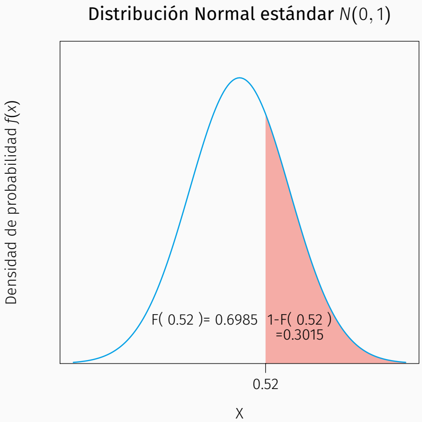


Cálculo de probabilidades con la Normal estándar

Probabilidades acumuladas por encima de un valor

Cuando tengamos que calcular probabilidades acumuladas por encima de un determinado valor (cola de acumulación a la derecha), podemos hacerlo por medio de la probabilidad del suceso contrario. Por ejemplo,

$$P(Z > 0.52) = 1 - P(Z \leq 0.52) = 1 - F(0.52) = 1 - 0.6985 = 0.3015.$$



Tipificación

Ya se ha visto cómo calcular probabilidades con una distribución normal estándar, pero *¿qué hacer cuando la distribución normal no es la estándar?* En tal caso hay que transformar la variable normal en una normal estándar.

Teorema (Tipificación)

Si X es una variable que sigue una distribución de probabilidad Normal de media μ y desviación típica σ, $X \sim N(\mu, \sigma)$, entonces la variable resultante de restarle a X su media μ y dividir por su desviación típica σ, sigue un modelo de distribución de probabilidad Normal estándar,

$$X \sim N(\mu, \sigma) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Esta transformación lineal se conoce como *transformación de tipificación* y la variable resultante Z se conoce como *normal tipificada*.

Así pues, para calcular probabilidades de una variable Normal que no sea la Normal estándar, se aplica primero la transformación de tipificación y después se puede utilizar la función de distribución de la Normal estándar.

Supóngase que la nota de un examen sigue un modelo de distribución de probabilidad normal $N(\mu = 6, \sigma = 1.5)$. ¿Qué porcentaje de suspensos habrá en la población?

Como X no es la normal estándar, primero se le aplica la transformación de tipificación $Z = \dfrac{X-\mu}{\sigma} = \dfrac{X-6}{1.5}$,

$$P(X < 5) = P\left(\frac{X-6}{1.5} < \frac{5-6}{1.5}\right) = P(Z < -0.67).$$

Después se puede usar la tabla de la función de distribución de la Normal estándar:

$$P(Z < -0.67) = F(-0.67) = 0.2514.$$

Así pues, habrán suspendido el 25.14% de los alumnos.

Distribución de probabilidad chi-cuadrado $\chi^2(n)$

Definición (Distribución de probabilidad Chi-cuadrado $\chi^2(n)$)

Dadas n variables aleatorias independientes $Z_1, \ldots, Z_n$, todas ellas siguiendo un modelo de distribución Normal estándar, entonces la variable

$$\chi^2(n) = Z_1^2 + \cdots + Z_n^2.$$

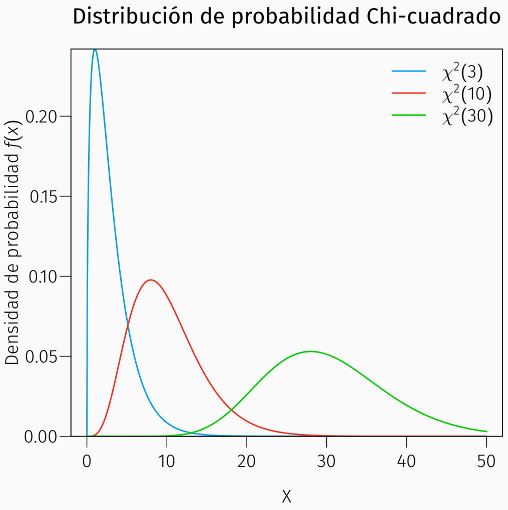
sigue un modelo de distribución de probabilidad *chi-cuadrado de n grados de libertad*.

Su rango es $\mathbb{R}^+$ y su media y varianza valen

$$\mu = n, \qquad \sigma^2 = 2n.$$

Como se verá más adelante, la distribución chi-cuadrado juega un papel importante en la estimación de la varianza poblacional y en el estudio de la relación entre variables cualitativas.

Función de densidad de la distribución chi-cuadrado



Propiedades de la distribución chi-cuadrado $\chi^2(n)$

- No toma valores negativos.
- Si $X \sim \chi^2(n)$ e $Y \sim \chi^2(m)$, entonces

$$X + Y \sim \chi^2(n+m).$$

- Al aumentar el número de grados de libertad, se aproxima asintóticamente a una normal.

Definición (Distribución de probabilidad T de Student $T(n)$)

Dada una variable Z que sigue un modelo de distribución de probabilidad Normal estándar, $Z \sim N(0,1)$, y una variable X que sigue un modelo de distribución de probabilidad Chi-cuadrado con n grados de libertad, $X \sim \chi^2(n)$, independientes, la variable

$$T = \frac{Z}{\sqrt{X/n}},$$

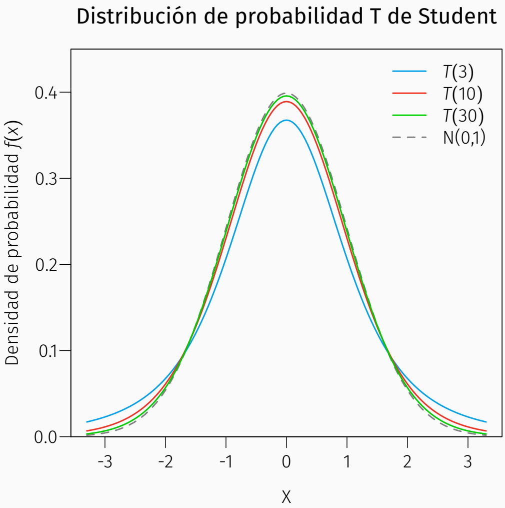
sigue un modelo de distribución de probabilidad *T de Student de n grados de libertad*.

Su rango es $\mathbb{R}$ y su media y varianza valen

$$\mu = 0, \qquad \sigma^2 = \frac{n}{n-2} \text{ si } n > 2.$$

Como se verá más adelante, la distribución T de Student juega un papel importante en la estimación la media poblacional.

Función de densidad de la distribución T de Student



Propiedades de la distribución T de Student $T(n)$

- La media, la mediana y la moda coinciden, $\mu = Me = Mo$.
- Es simétrica, $g_1 = 0$.
- Tiende asintóticamente a la distribución Normal estándar a medida que aumentan los grados de libertad. En la práctica, para $n \geq 30$ ambas distribuciones son aproximadamente iguales.

$$T(n) \overset{n\to\infty}{\approx} N(0,1).$$

Distribución de probabilidad F de Fisher-Snedecor $F(m, n)$

Definición (Distribución de probabilidad F de Fisher-Snedecor $F(m, n)$)

Dadas dos variables independientes X e Y, siguiendo modelos de distribución Chi-cuadrado con m y n grados de libertad respectivamente, $X \sim \chi^2(m)$ e $Y \sim \chi^2(n)$, entonces la variable

$$F = \frac{X/m}{Y/n},$$

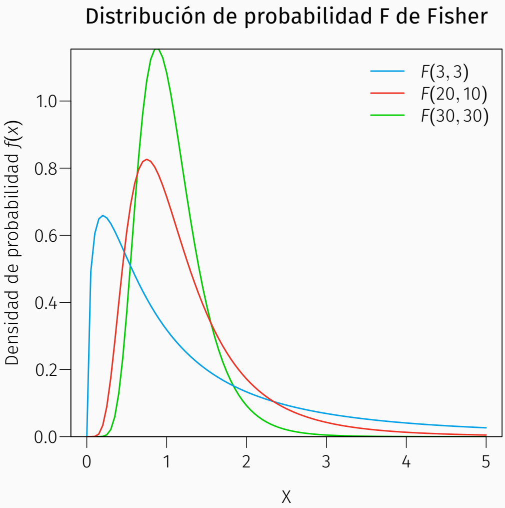
sigue un modelo de distribución de probabilidad *F de Fisher-Snedecor de m y n grados de libertad*.

Su rango es $\mathbb{R}^+$ y su media y varianza valen

$$\mu = \frac{n}{n-2}, \qquad \sigma^2 = \frac{2n^2(m+n2)}{m(n-2)^2(n-4)} \text{ si } n > 4.$$

Como se verá más adelante, la distribución F de Fisher-Snedecor juega un papel importante en la comparación de varianzas poblacionales y en el análisis de la varianza.

Función de densidad de la distribución F de Fisher-Snedecor $F(m, n)$



Propiedades de la distribución F de Fisher-Snedecor $F(m,n)$

- No está definida para valores negativos.
- De la definición se deduce que

$$F(m,n) = \frac{1}{F(n,m)}$$

de manera que si llamamos $f(m,n)_p$ al valor que cumple que $P(F(m,n)f(m,n)_p) = p$, entonces se cumple

$$f(m,n)_p = \frac{1}{f(n,m)_{1p}}$$

Esto resulta muy útil para utilizar las tablas de su función de distribución.