ESTADÍSTICA CON R

Ernesto Correa Velandia

Máster Data Science
Edición 20

Febrero 21, 2020

Contenido

Estadística

Ciencia que utiliza un conjunto de datos medidos sobre una población, para obtener, a partir de ellos, inferencias basadas en el cálculo de probabilidades.

Población

Conjunto total de personas, objetos, ideas sometidos a observaciones estadísticas.

Muestra

Subconjunto de la población.

Población vs Muestra

El Método Estadístico

1. Planteamiento del problema (definimos el objeto de investigación)
2. Planificación del trabajo de campo (¿Cómo y qué vamos a medir?)
3. Recopilación o medición de la información (Recogida y depuración de la información)
4. Análisis descriptivo de los datos (Organización y presentanción de la información)
5. Inferencia estadística (generalizar de la muestra a la población)
6. Validación del modelo (pruebas de diagnóstico del modelo)
7. Interpretación

¿Qué tipo de información medimos?

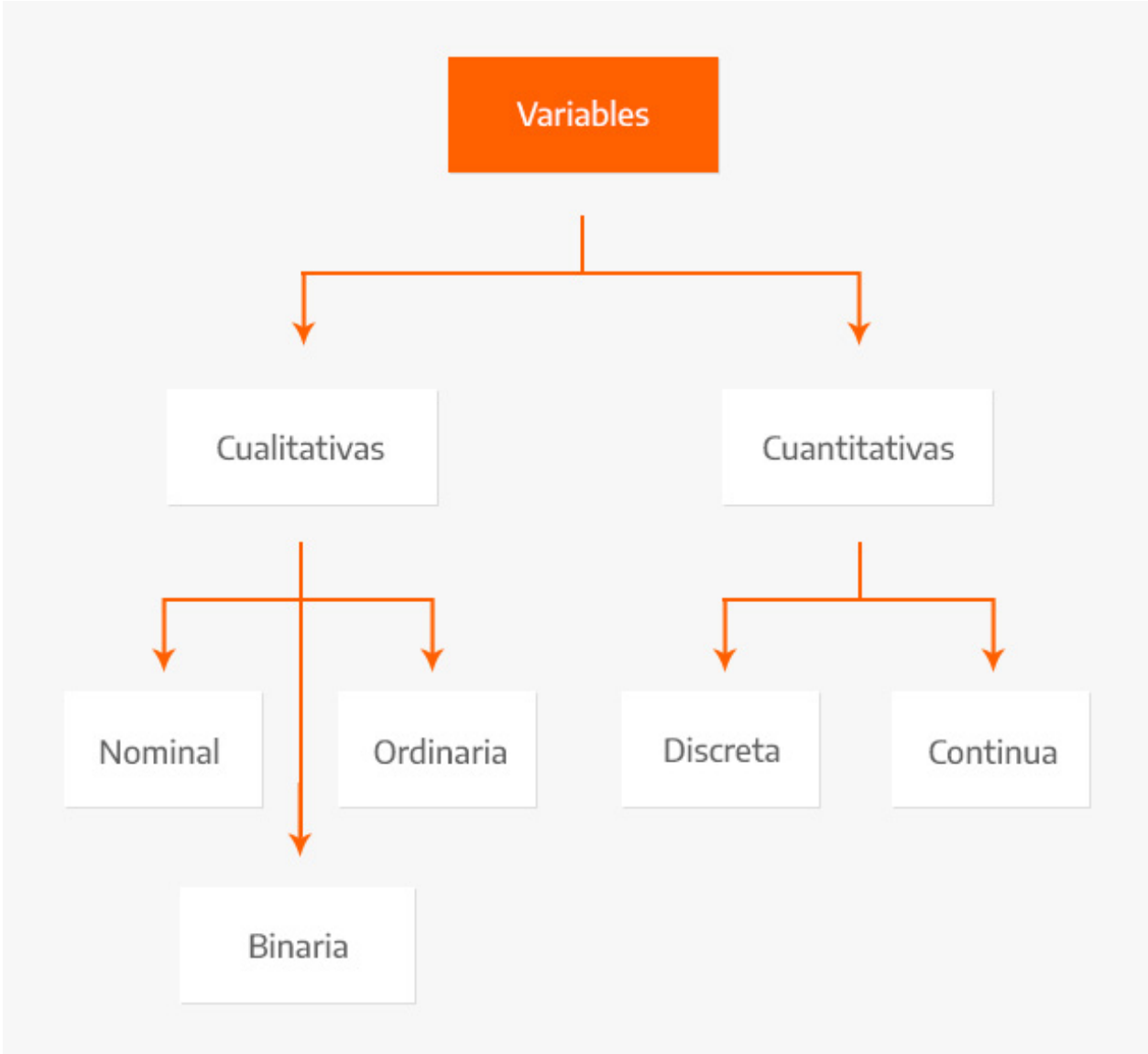
Variables aleatorias

Cada una de las propiedades, rasgos o cualidades que poseen los elementos de una población y que son objeto de estudio. Deben poseer la propiedad de ser medibles. Se conocen los posibles resultados pero es imposible predecir el resultado de antemano.

Pueden ser:

- Variables cualitativas o categóricas
- Variables cuantitativas o numéricas

Variables aleatorias



Variables aleatorias cualitativas

Los valores que toman no se pueden cuantificar. Cada uno de estos valores se denomina categoría, clase o modalidad. Pueden ser:

- Variables nominales establecen categorías, por ejemplo sexo
- Variables ordinales establecen un orden, por ejemplo grado de satisfacción de un cliente
- Variables binarias o dicotómicas solo pueden tomar 2 valores, si y no;

Variables aleatorias cuantitativas

Los valores que toman se pueden cuantificar. Dependiendo del conjunto de posibles resultados Pueden ser:

- Variables Discretas usualmente toman valores en los números enteros, por ejemplo número de hijos
- Variables Continuas usualmente toman valores en los números reales, por ejemplo precio de un producto.

Juegan un papel muy importante, ya que toda la teoría matemática se desarrolla a partir de ellas.

Estadística Descriptiva

Los principales objetivos de esta parte son:

1. Obtener tablas de frecuencias de un conjunto de datos
2. Obtener medidas de posición, dispersión y forma de un conjunto de datos
3. Obtener representaciones gráficas que resuman un conjunto de datos
4. Detectar valores atípicos o fuera de rango en un conjunto de datos

Tablas de frecuencias

Son tablas que ordenan los datos teniendo en cuenta la frecuencia con la que aparecen.

Frecuencia Absoluta

$$n_i = \text{número de veces que ocurre } x_i$$

Frecuencia Relativa

$$f_i = \frac{n_i}{n}, \quad n = \sum n_i$$

Frecuencia Acumulada

$$N_k = n_1 + n_2 + \ldots + n_k$$

Frecuencia Relativa Acumulada

$$N_i = \frac{N_i}{n}$$

Tablas de Frecuencias

Ejemplo de Tabla de Frecuencias				
Xi	Frecuencia Absoluta (ni)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa (fi= ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)
1	10	10	0,09	0,09
2	15	15	0,13	0,22
3	17	17	0,15	0,37
4	20	20	0,18	0,54
5	7	7	0,06	0,61
6	12	12	0,11	0,71
7	15	15	0,13	0,84
8	18	18	0,16	1,00
Total	114	114	1	1
Fuente: Propia				

Intervalos	Frecuencia absoluta f_i	Frecuencia absoluta acumulada F_i	Frecuencia relativa n_i	Frecuencia relativa acumulada N_i
[5 - 5,5)	1	1	0,04	0,04
[5,5 - 6)	2	3	0,08	0,13
[6 - 6,5)	3	6	0,13	0,25
[6,5 - 7)	4	10	0,17	0,42
[7 - 7,5)	8	18	0,33	0,75
[7,5 - 8)	1	19	0,04	0,79
[8 - 8,5)	5	24	0,21	1,00
Total	**24**		**1,00**	

Medidas de posición

Son magnitudes calculadas a partir de los datos que dan información acerca de dónde están ubicados los datos

Media aritmética

$$\bar{x} = \frac{\sum x_i}{n} = \sum f_i x_i$$

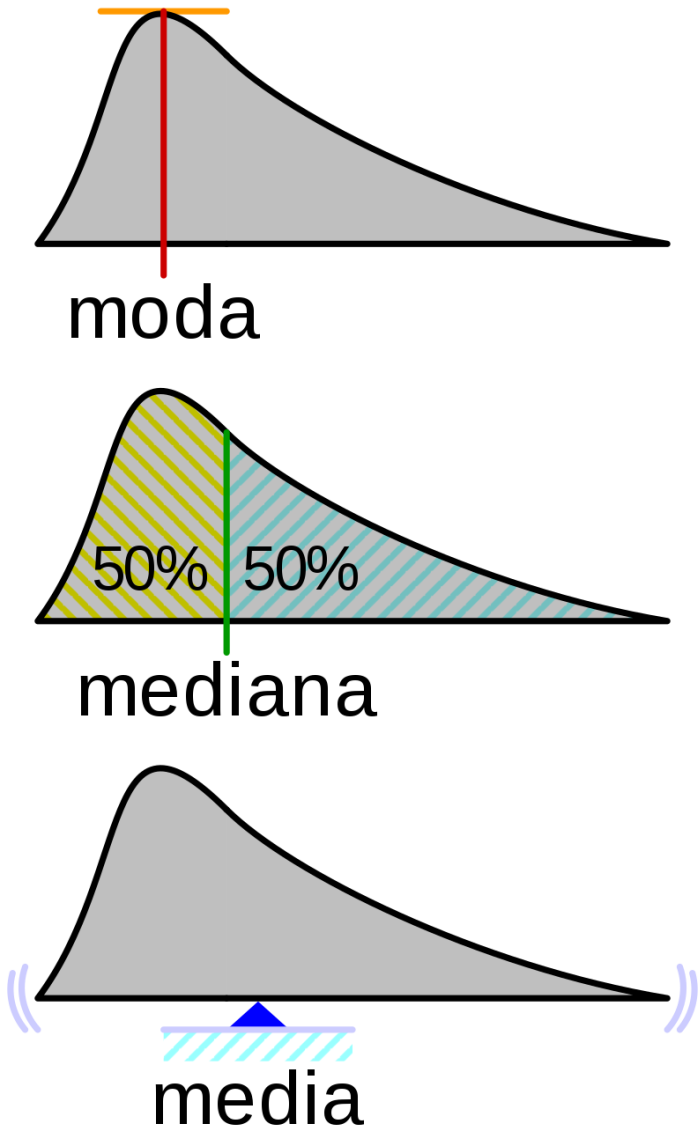
Mediana

Ordenados los datos de menor a mayor, es el valor que deja el $50\,\%$ a la izquierda y el $50\,\%$ a la derecha. Si el número de datos es par, se tomará el punto medio entre los dos valores centrales. Se nota como Me

Moda

Es el valor que más veces se repite. Puede ser que ser que el conjunto de datos sea multimodal, es decir, que haya más de una moda.

Medidas de posición



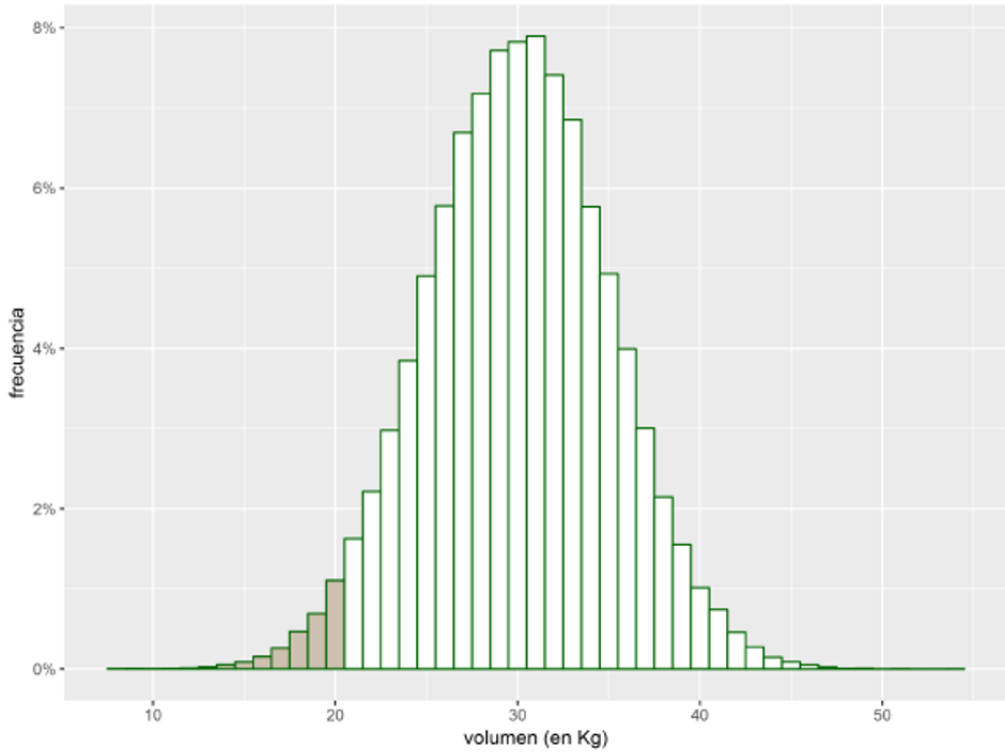
Percentiles Una vez ordenados los datos menor a mayor, los percentiles indican el valor de la variable por debajo del cual se encuentran un porcentaje dado en las observaciones de la muestra. Se notan P_r, por ejemplo, P_{20} representa el valor de la varible bajo el cual se encuentra el $20\,\%$ de las observaciones.

$$P_{50} = Me$$

Los cuartiles $Q_1, Q_2, Q_3,$ y Q_4 dividen la muestra de datos en cuatro partes iguales, entonces tenemos que $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$ y $Q_4 = P_{100} = \text{máx}\{x_i\}$ El rango intercuartílico RI se define

$$RI = Q_3 - Q_1$$

Percentil 20



Medidas de dispersión

Parámetros estadísticos que indican como se alejan los datos respecto de la media aritmética. Sirven como indicador de la variabilidad de los datos.

Rango Indica la dispersión entre los valores extremos de una variable. se calcula como la diferencia entre el mayor y el menor valor de la variable. Se denota como R.

$$R = x_{(n)} - x_{(1)}$$

Varianza Es la media aritmética del cuadrado de las desviaciones respecto a la media

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Propiedades de la varianza

1. La varianza será siempre un valor positivo o cero, en el caso de que las mediciones sean todas iguales.
2. Si a todos los valores de la variable se les suma un número la varianza no varía
3. Si todos los valores de la variable se multiplican por un número la varianza queda multiplicada por el cuadrado de dicho número.
4. La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.
5. $$\sigma^2 = \sum_{i=1}^{n} \frac{x_i^2}{n} - \overline{x}^2$$

Desviación Típica Es la raiz cuadrada de la Varianza. Está representada por σ

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Algunas Propiedades son:

1. σ es siempre un valor positivo o cero, en el caso de que lasmediciones sean iguales.
2. Cuanta más pequeña sea σ mayor será la concentración de datos alrededor de la media

Coeficiente de Variación Permite comparar las dispersiones de dos variables distintas, siempre que sus medias sean positivas. Se calcula para cada una de las variables y los valores que se obtienen se comparan entre sí.

$$C.V = \frac{\sigma}{\overline{x}}$$

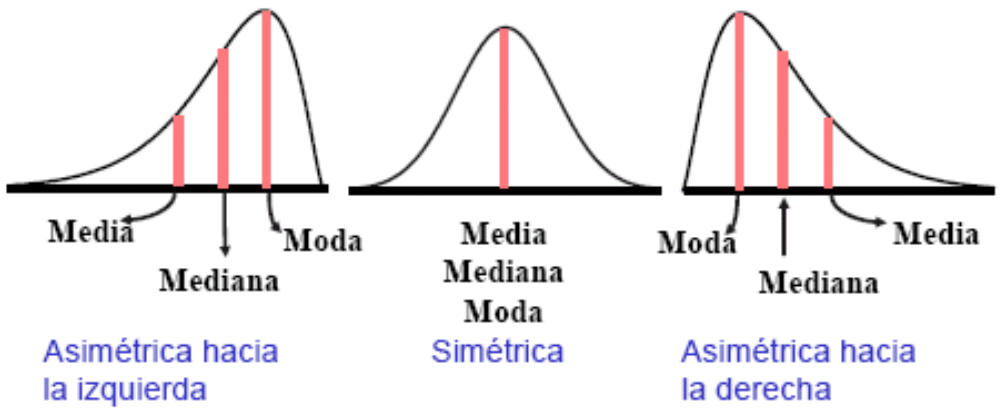
- La mayor dispersión corresponderá al valor del coeficiente de variación mayor.

Ejemplo: Se tomaron dos muestras de la misma población, la primera tiene $\overline{x} = 140$, $\sigma_x = 28{,}28$ y la segunda $\overline{w} = 150$, $\sigma_w = 24$
¿Cúal de las dos muestras tiene una mejor dispersión de los datos?

Medidas de Forma

Coeficiente de Asimetría Este coeficiente mide la simetría de la distribución de la variable aleatoria respecto a la media aritmética.

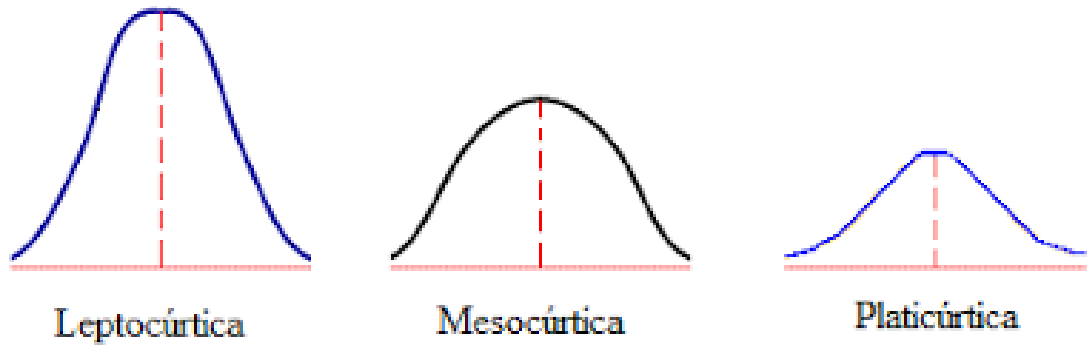
$$g_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{n\sigma^3}$$



- $g_3 < 0$ Hay menos datos a la izquierda de la media
- $g_3 > 0$ Hay menos datos a la derecha de la media
- $g_3 \approx 0$ Hay simetría respecto a la media

Coeficiente de Kurtosis Este coeficiente mide el apuntamiento de la distribución de la variable aleatoria.

$$g_4 = \frac{\sum (x_i - \bar{x})^4 n_i}{n\sigma^4} - 3$$



- $g_4 < 0$ Planicúrtica
- $g_4 > 0$ Leptocúrtica
- $g_4 \approx 0$ mesocúrtica.

¿Podemos saber si existe relación entre dos variables?

Covarianza poblacional:

$$\sigma(x, y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

Covarianza muestral:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})$$

Correlación entre dos variables:

$$corr(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

R logo CC-BY-SA 4.0 Hadley Wickham and others at RStudio

Función de probabilidad f

La Función de probabilidad o distribución de probabilidad de una variable aleatoria, es una función que asigna a cada suceso la probabilidad de que dicho suceso ocurra.

Por ejemplo, si definimos la variable aleatoria X como el resultado al lanzar un dado normal, esta variable puede tomar los valores $X = 1, 2, 3, 4, 5, 6$. Por lo tanto es una varible aleatoria discreta.

x	1	2	3	4	5	6
$P[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Función de distribución F

La Función de distribución de una variable aleatoria, es una función definida sobre los números reales $\mathbb{R}$ cuyo valor en cada $x \in \mathbb{R}$ es la probabilidad de que la variable aleatoria sea menor o igual que x. Depende del tipo de variable aleatoria, discreta o continua.

$$F(x) = \begin{cases} \sum_{x_i \leq x} f(x_i), & \textit{si es una v. a. discreta} \\ \int_{-\infty}^{x} f(t)dt, & \textit{si es una v. a. continua} \end{cases}$$

Algunas Distribuciones Importantes

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Discrete uniform	$p(x) = \frac{1}{n}$ $x = 1, 2, \ldots, n$		$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Hyper-geometric	$\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$ $\text{Max}[0, n-(N-N_1)]$ $\leq x \leq \text{Min}(n, N_1)$		$\mu = n\theta$ $\theta = \frac{N_1}{N}$	$\sigma^2 = \frac{N-n}{N-1} n\theta(1-\theta)$ $\theta = \frac{N_1}{N}$
Bernoulli	$\theta^x (1-\theta)^{1-x}$ $x = 0, 1 \quad 0 \leq \theta \leq 1$	$\theta e^t + (1-\theta)$	θ	$\theta(1-\theta)$

Algunas Distribuciones Importantes

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Binomial	$\binom{n}{x}\theta^x 1-\theta^{n-x}$ $x=0,1,\ldots,n;\ 0\le\theta\le 1$	$(\theta e^t+(1-\theta))^n$	$n\theta$	$n\theta(1-\theta)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$ $x=0,1,\ldots;\ \lambda>0$	$e^{\lambda(e^t-1)}$	λ	λ
Uniform	$f(x)=\frac{1}{b-a}$ $a\le x\le b$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$ $-\infty<x<\infty,-\infty<\mu<\infty,$ $\sigma>0$	$e^{\mu t+(\sigma^2t^2/2)}$	μ	σ^2
Chi-square	$\frac{1}{2^{n/2}\Gamma(n/2)}w^{n/2-1}e^{-w/2}$ $w\ge 0, n>0$	$(1-2t)^{-n/2}$	n	$2n$
Student-t	$f(t)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}$ $\left(1+\frac{t^2}{n}\right)^{(n+1)/2}$ $-\infty<t<\infty$		0	$\frac{n}{n-2}$

Resumen de las funciones asociadas a las distribuciones

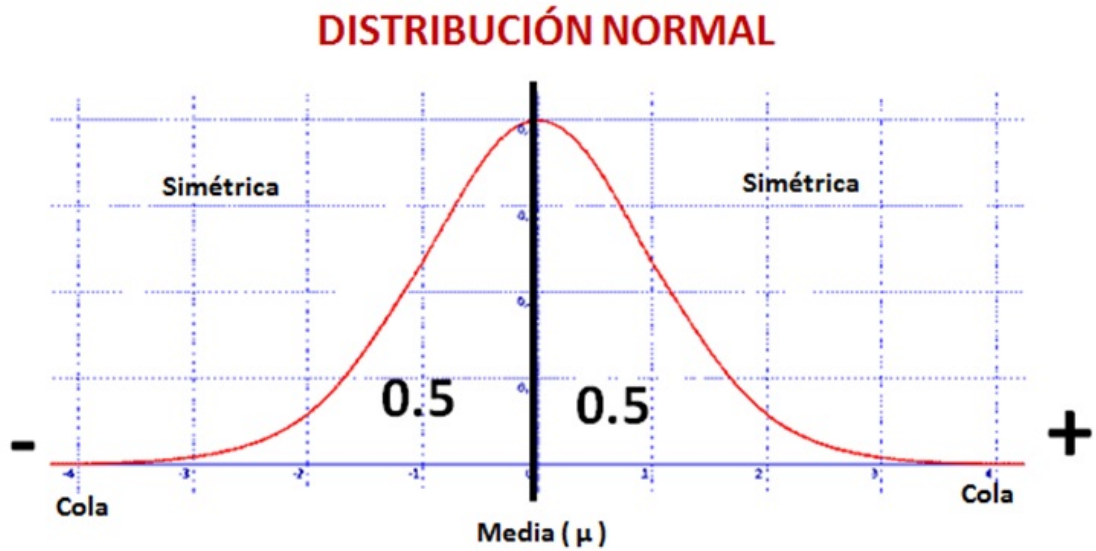
	Densidad o masa	Función de distribución	Función cuantil	Muestras aleatorias
Binomial(n,p)	`dbinom`(x,n,p)	`pbinom`(x,n,p)	`qbinom`$(prob,n,p)$	`rbinom`$(muestras,n,p)$
Poisson(λ)	`dpois`(x,λ)	`ppois`(x,λ)	`qpois`$(prob,\lambda)$	`rpois`$(muestras,\lambda)$
Geométrica(p)	`dgeom`(x,p)	`pgeom`(x,p)	`qgeom`$(prob,p)$	`rpois`$(muestras,p)$
Binomial negativa(a,p)	`dnbinom`(x,a,p)	`pnbinom`(x,a,p)	`qnbinom`$(prob,a,p)$	`rnbinom`$(muestras,a,p)$
Exponencial(λ)	`dexp`(x,λ)	`pexp`(x,λ)	`qexp`$(prob,\lambda)$	`rexp`$(muestras,\lambda)$
Gamma(a,λ)	`dgamma`(x,a,λ)	`pgamma`(x,a,λ)	`qgamma`$(prob,a,\lambda)$	`rgamma`$(muestras,a,\lambda)$
Normal(μ,σ)	`dnorm`(x,μ,σ)	`pnorm`(x,μ,σ)	`qnorm`$(prob,\mu,\sigma)$	`rnorm`$(muestras,\mu,\sigma)$

Distribución Normal

- Se llama distribución normal o distribución de Gauss
- Representa a variables aleatorias continuas
- Permite modelar numerosos fenómenos naturales, sociales, psicológicos, etc...
- Teorema del Límite Central. Bajo ciertas hipótesis adecuadas, la distribución normal es el límite (muestras grandes) de la suma de variables aleatroias independientres.
- Es muy cómoda y fácil de manipular matemáticamente
- Satisface la propiedad de reproductividad
- Depende de dos parámetros, la media y la varianza.
- No debemos abusar de sus beneficios

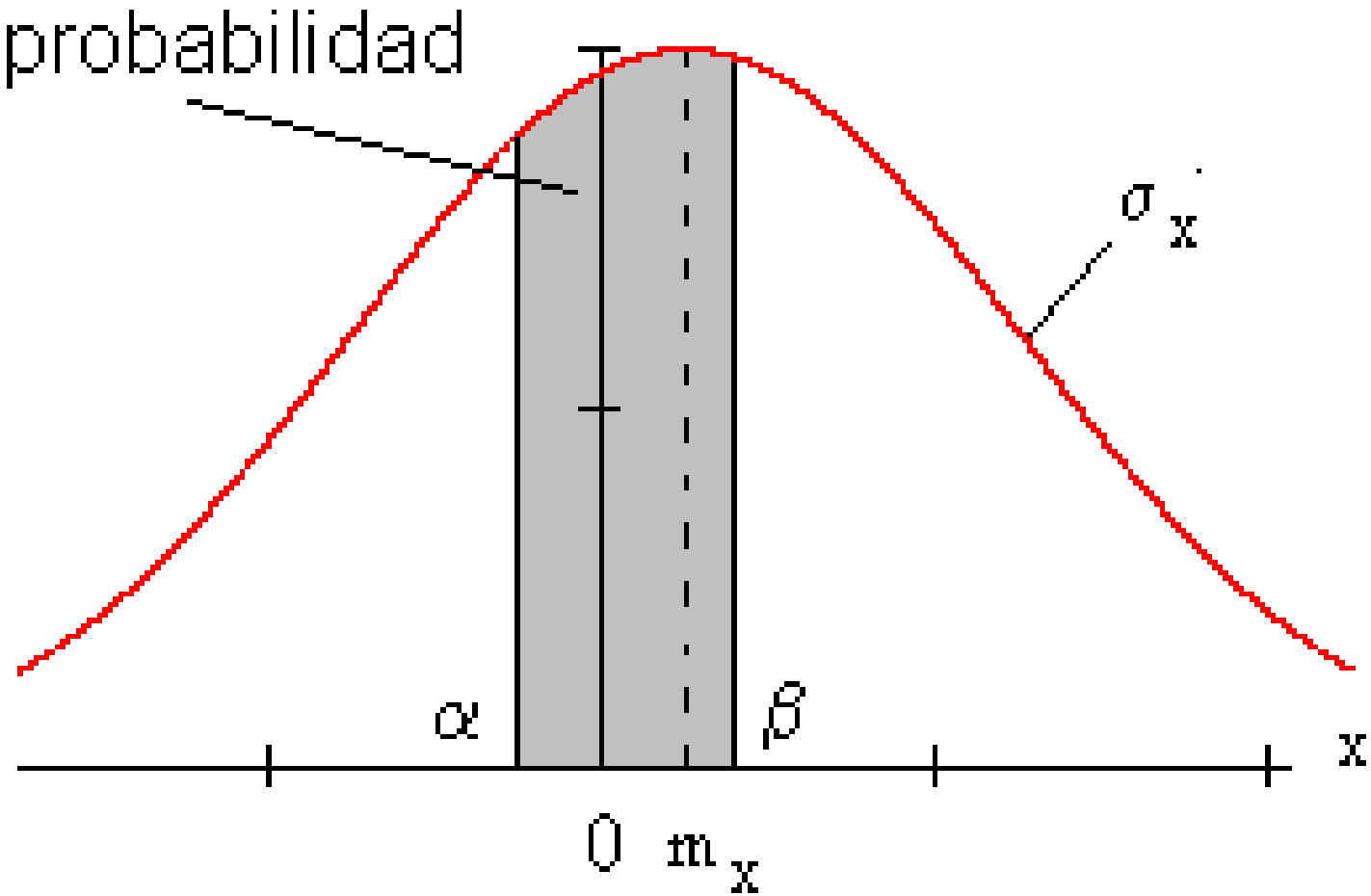
Distribución Normal

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



Distribución Normal

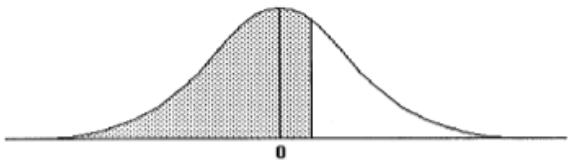
$$P(\alpha \leq x \leq \beta) = \int_{\alpha}^{\beta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx.$$



Distribución Normal Estándar

TABLA-T3: DISTRIBUCIÓN NORMAL ESTANDAR

$$Z \approx N(\mu = 0; \sigma^2 = 1)$$



$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t/2} \, dt$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327

$$P(Z \leq 0{,}34) = 0{,}63307$$

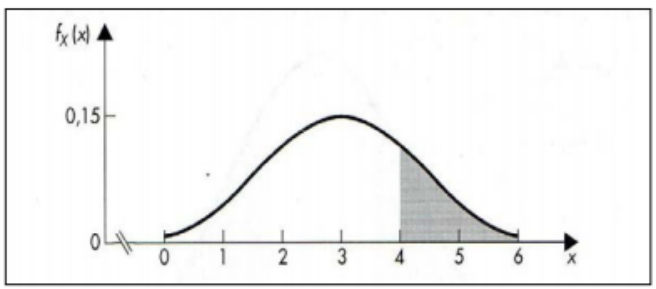
$$P(Z \leq 1{,}36) = 0{,}91308$$

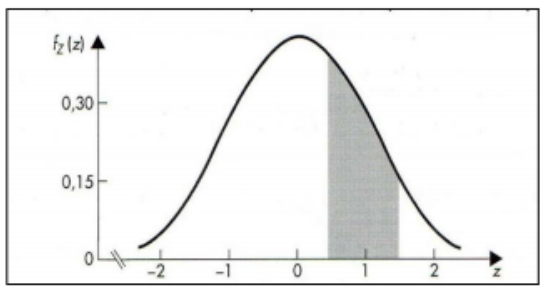
Distribución Normal Estándar

Probabilidades de intervalos

Si X es una variable Normal de media 3 y desviación 2. Calcular la probabilidad de que tome un valor entre 4 y 6.

$$X \to N(3, 2) \qquad \Rightarrow \qquad Z \to N(0, 1)$$





$$P(4 < X < 6) = P\left(\frac{4-3}{2} < \frac{X-\mu}{\sigma} < \frac{6-3}{2}\right) = P(0{,}5 < Z < 1{,}5) =$$

$$= P(Z < 1{,}5) - P(Z < 0{,}5) = 0{,}9332 - 0{,}6915 = 0{,}2417$$

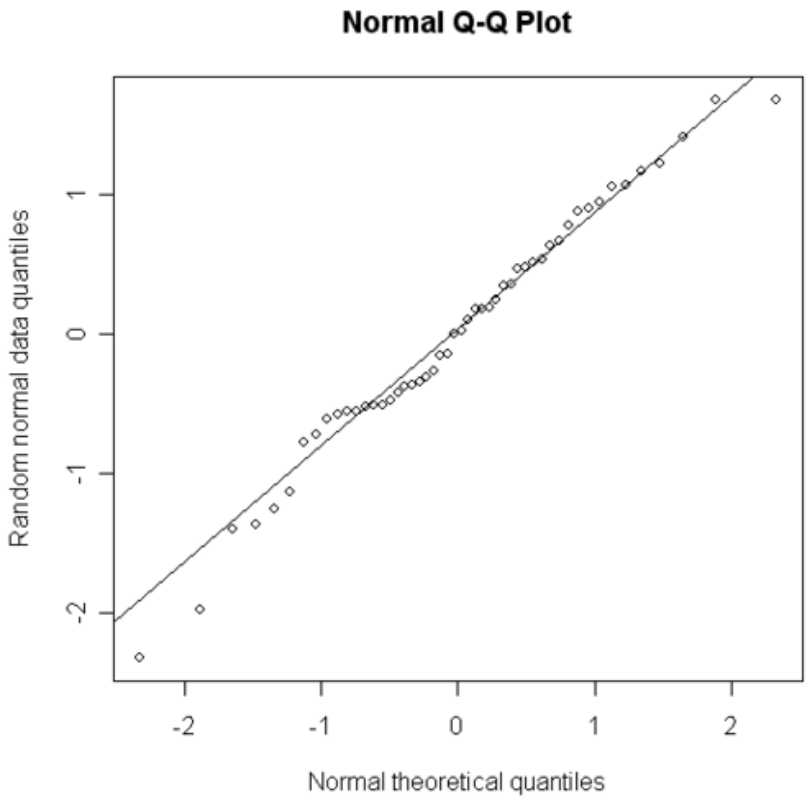
Pruebas de Normalidad

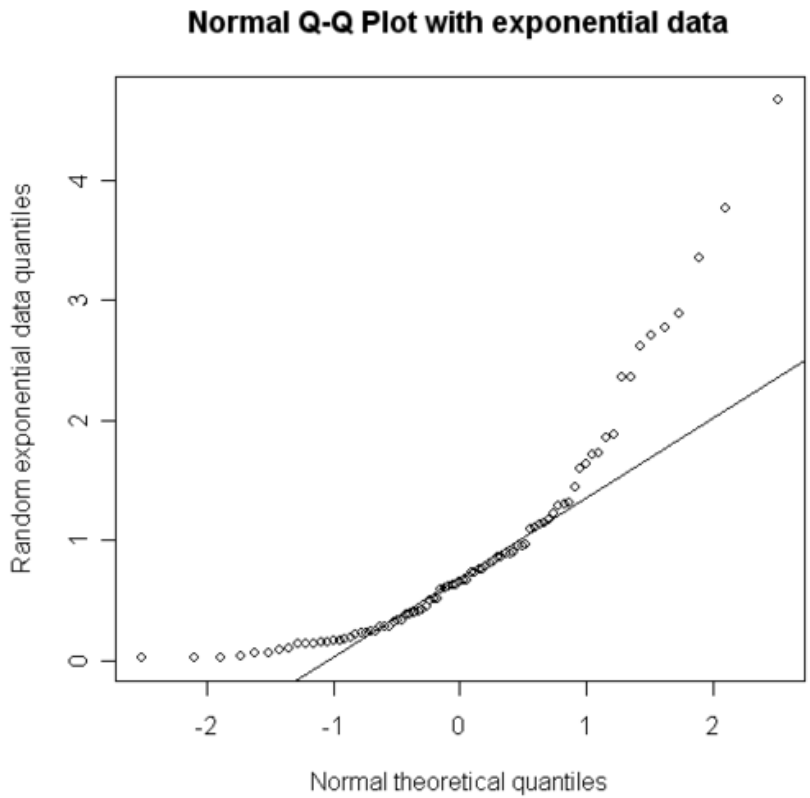
- Analizando el histograma de frecuencias y las medidas de forma (asimetría y kurtosis)
- Analizando el gráfico box-plot.
- Mediante el análisis de gráficas **Q**uantile-**Q**uantile-**P**lot
- Contrastes de normalidad (**KOLMOGOROV-SMIRNOV; SHAPIRO-WILK**)

Quantile-Quantile-Plot

- El quantile quantile plot es una técnica gráfica para determinar si dos conjuntos de datos provienen de poblaciones con una distribución común
- Es una gráfica de los cuantiles del primer conjunto de datos contra los cuantiles del segundo conjunto de datos.
- Se traza una línea de referencia de 45 grados. Si los dos conjuntos provienen de una población con la misma distribución, los puntos deberían caer aproximadamente a lo largo de esta línea de referencia.
- Muchos aspectos distributivos pueden ser probados simultáneamente. Ubicación, escala, simetría y la presencia de valores atípicos se pueden detectar desde esta gráfica.

Quantile-Quantile-Plot





R logo CC-BY-SA 4.0 Hadley Wickham and others at RStudio

Estimación puntual y por Intervalos de confianza

- Estimar los parámetros mediante el método de máxima verosimilitud. Este método se basa en la optimización de la función de verosimilitud (congruencia entre parámetros y datos).
- Obtener intervalos de confianza para los parámetros de una distribución

Vamos a suponer que tenemos una muestra aleatoria de las mediciones de una variable. Con estos datos vamos a estimar los parámetros de la distribución que suponemos siguen (generalmente Normal).

Estimadores para la distribución Normal

Cómo vimos, la distribución Normal depende de dos paraámetros, la media μ y la desviación estándar σ. En este caso los estimadores por el método de máxima verosimilitud son

$$\mu = \bar{x} = \frac{\sum x_i}{n}, \quad \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n\text{-}1}$$

Para cada tipo de distribución existen diferentes estimadores,

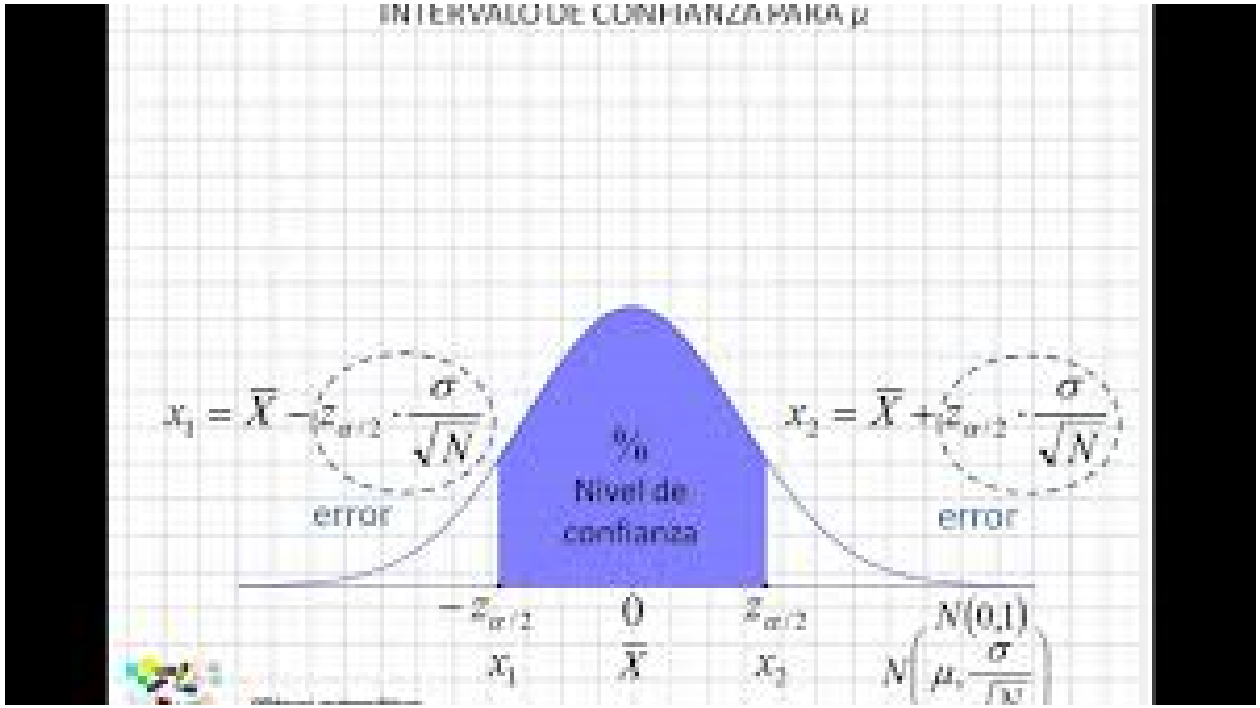
ajuste.nombre-distribución en R.

Estimación por Intervalos de confianza

- Un intervalo de confianza es un intervalo de números que contiene los valores más plausibles para nuestro parámetro de población.
- Proporcionan el valor de un estadístico mediante un intervalo, bajo una confianza.

1-α	α/2	$Z\alpha_{/2}$	Intervalo de confianza
0,90	0,05	1,645	$(\bar{X} - 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 1{,}645 \cdot \frac{\sigma}{\sqrt{n}})$
0,95	0,025	1,96	$(\bar{X} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}})$
0,99	0,005	2,575	$(\bar{X} - 2{,}572 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 2{,}575 \cdot \frac{\sigma}{\sqrt{n}})$

Estimación por Intervalos de confianza



Estimación mediante pruebas de Hipótesis

- Otra manera de hacer inferencia es haciendo una afirmación acerca del valor que puede tomar el parámetro de la población bajo estudio.
- Esta afirmación puede estar basada en alguna creencia o experiencia pasada que será contrastada con la evidencia que nosotros obtengamos a través de la información contenida en la muestra. Esto es a lo que llamamos Prueba de Hipótesis

Elementos de una Prueba de Hipótesis

- Hipótesis nula H_0, debe ser la que suponemos cierta de partida.
- Hipótesis alternativa H_a
- Estadístico de prueba
- Región de rechazo.
- El objetivo no es determinar cual de las dos Hipótesis es correcta, por el contrario, debemos determinar si aceptamos o no la Hipótesis nula

Tipos de Errores

	H_0 Verdadera	H_0 Falsa
Rechazamos H_0	Error Tipo I P(error Tipo I) = α	Decisión Correcta
No Rechazamos H_0	Decisión Correcta	Error Tipo II P(error Tipo II) = β

p-valor

- El valor p es el tamaño más pequeño a para el que se rechaza H0. Es decir,

$$p - valor = P[Rechazar\ el\ estadístico\ muestral\ |H_0\ es\ cierta]$$

Si $p - valor > \alpha\ (0.05)$ se acepta la hipótesis nula H_0

- El valor p expresa evidencia contra H0: cuanto más pequeño es el valor p, más fuerte es la evidencia contra H0.
- Generalmente, el valor de p se considera pequeño cuando p <0.01 y grande cuando p> 0.1.
- El valor p no es la probabilidad de que el H0 sea verdadero.

Contraste bilateral para la media poblacional

Supongamos que conocemos la varianza de la población.

$$H_0 : \mu = \mu_0; \quad H_a : \mu \neq \mu_0.$$

En ese caso

$$p - valor = 2 * P(Z > z_0), \quad z_0 = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Supongamos que NO conocemos la varianza de la población. En ese caso

$$p - valor = 2 * P(t_{n-1} > t_0), \quad t_0 = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

R logo CC-BY-SA 4.0 Hadley Wickham and others at RStudio

¡Gracias por su atención!