INFERENCIA ESTADÍSTICA

Maestría en estadística aplicada

Universidad de Nariño

Material preparado por:

Giovany Babativa

CONVERGENCE

Theorem (The Central Limit Theorem (CLT)). *Let* $X_1, \ldots, X_n$ *be* IID *with mean* μ *and variance* σ^2. *Let* $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. *Then*

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0,1)$. *In other words,*

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

DISTRIBUCIÓN MUESTRAL DE S^2

Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces el estadístico

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $v = n - 1$ grados de libertad.

BERRY-ESSÈEN THEOREM

Theorem. *Assume the same conditions as the CLT. Then,*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

Para n pequeño se aproxima por una distribución t-student con n-1 gl.

DISTRIBUCIÓN t (VARIANZA DESCONOCIDA)

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}}$$

Para n mayor que 30, se aproxima por una distribución t-student con n-1 gl.

Use el argumento **fun=dt** en la geometría ***geom_fuction*** y use varias capas para mostrar el resultado con v = c(4, 10, 20, 50, 100, 200) y agregue una distribución normal.

DISTRIBUCIÓN F

$$F = \frac{U/v_1}{V/v_2},$$

Es el cociente entre dos v.a. con distribución chi-cuadrado

Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente, entonces,

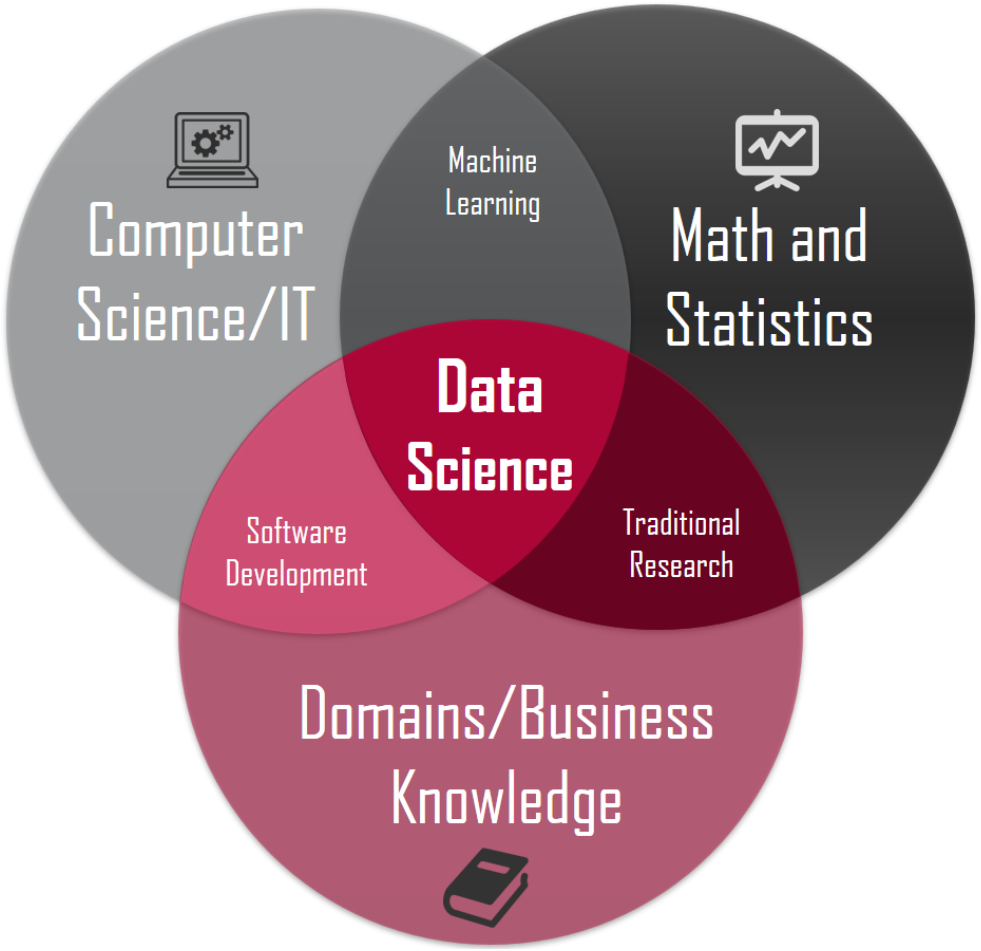
$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

ESTADÍSTICA INFERENCIAL

STATISTICAL LEARNING

DATA SCIENCE



INTRODUCCIÓN

Statistical inference, or "learning" as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is:

Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?

In some cases, we may want to infer only some feature of F such as its mean.

SOPORTE TEÓRICO

A **statistical model** $\mathfrak{F}$ is a set of distributions (or densities or regression functions). A **parametric model** is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R},\ \sigma > 0 \right\}.$$

This is a two-parameter model. We have written the density as $f(x;\mu,\sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, a parametric model takes the form

$$\mathfrak{F} = \left\{ f(x;\,\theta) :\ \theta \in \Theta \right\}$$

SOPORTE TEÓRICO

Métodos paramétricos: Se denominan así porque suponen que los datos provienen de alguna distribución F indexada por un parámetro. El supuesto que se impone sobre la distribución de X, en pruebas de hipótesis sobre medias, es que ésta es de familia normal. Cuando se aplican estos métodos sin que se cumplan los supuestos requeridos las pruebas tienden a no conservar el error de tipo I, luego las conclusiones extraídas de la misma pueden ser erradas.

Métodos no paramétricos: Se denominan así porque no se impone ningún supuesto sobre la distribución de X (métodos libres de la distribución) o los supuestos que se imponen son solo condiciones de regularidad (continuidad, simetría), pero estos datos aún se encuentran dados por una familia de distribuciones *F* indexada por infinitos parámetros.

ESTIMACIÓN PUNTUAL

Point estimation refers to providing a single "best guess" of some quantity of interest. The quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable.

> By convention, we denote a point estimate of θ by $\widehat{\theta}$ or $\widehat{\theta}_n$. Remember that θ is a fixed, unknown quantity. The estimate $\widehat{\theta}$ depends on the data so $\widehat{\theta}$ is a random variable.

More formally, let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n).$$

SESGO – ESTIMADOR INSESGADO

$$\text{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta.$$

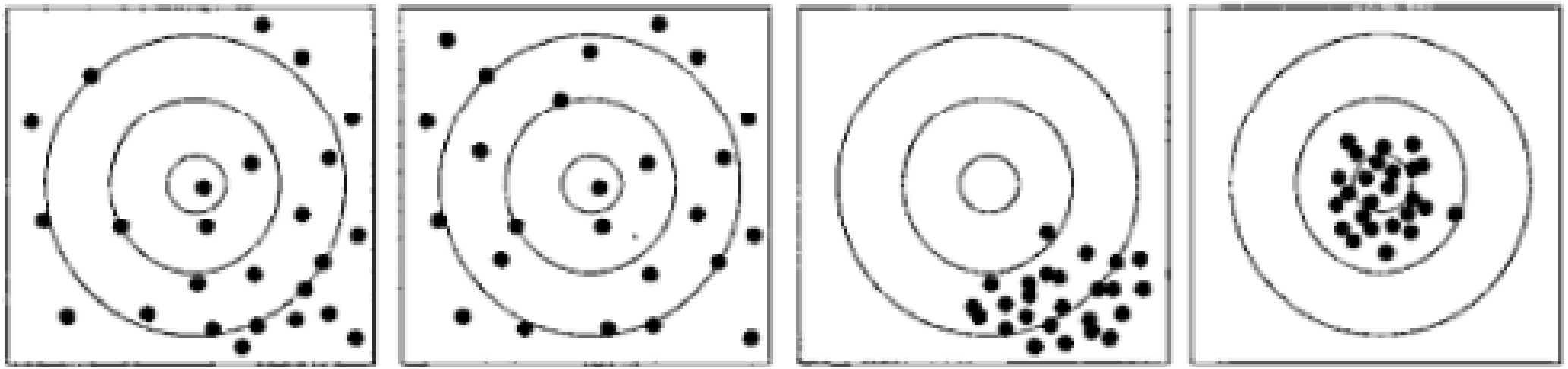
$\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}(\widehat{\theta}_n) = \theta$.

Se dice que un estadístico $\hat{\Theta}$ es un **estimador insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Definition. *A point estimator* $\widehat{\theta}_n$ *of a parameter* θ *is* **consistent** *if* $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.

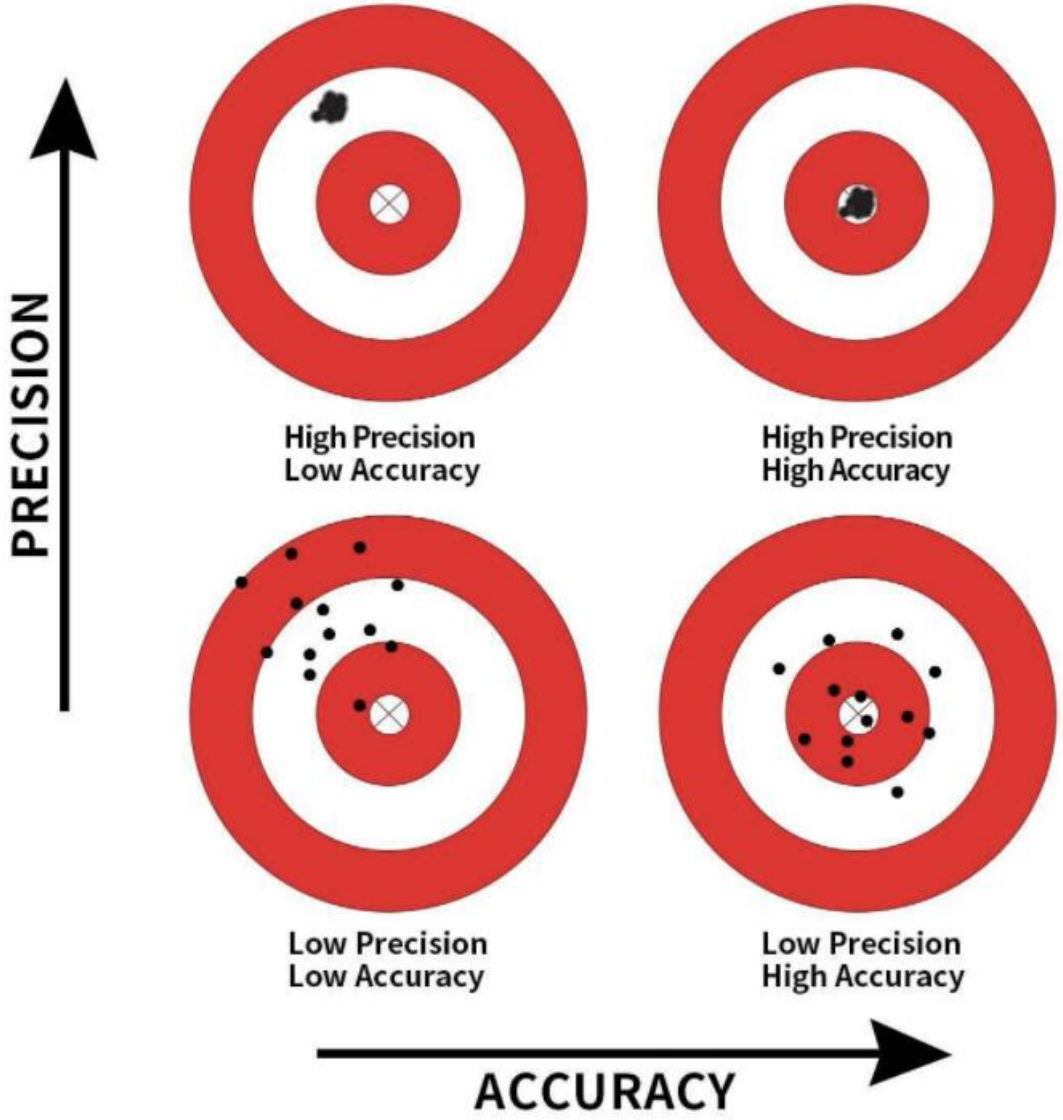
PROPIEDADES DE UN ESTIMADOR



(a) Biased, imprecise (b) Unbiased, imprecise (c) Biased, precise (d) Unbiased, precise

PROPIEDADES DE UN ESTIMADOR

Sesgo (Exactitud) vs Precisión



¿El tamaño de la muestra resuelve todos los problemas?

PROPIEDADES DE UN ESTIMADOR


DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

PARÁMETROS DE INTERÉS

Often, we are only interested in some function $T(\theta)$. For example, if $X \sim N(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ then $\mu = T(\theta)$ is called the **parameter of interest** and σ is called a **nuisance parameter**. The parameter of interest might be a complicated function of θ as in the following example.

MÉTODO DE LOS MOMENTOS

Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_k)$ has k components. For $1 \leq j \leq k$, define the j^{th} **moment**

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x)$$

and the j^{th} **sample moment**

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

MÉTODO DE LOS MOMENTOS

Definition. *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned}$$

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu'_1 = \mathrm{E}X^1, \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu'_2 = \mathrm{E}X^2, \\ &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu'_k = \mathrm{E}X^k. \end{aligned}$$

EJEMPLO

Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\widehat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$. By equating these we get the estimator

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad \blacksquare$$

EJEMPLO

Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations[1]

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\begin{aligned} \widehat{\mu} &= \overline{X}_n \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \quad \blacksquare \end{aligned}$$

$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$

MÉTODO DE LA MÁXIMA VEROSIMILITUD

Definition. *The* **likelihood function** *is defined by*

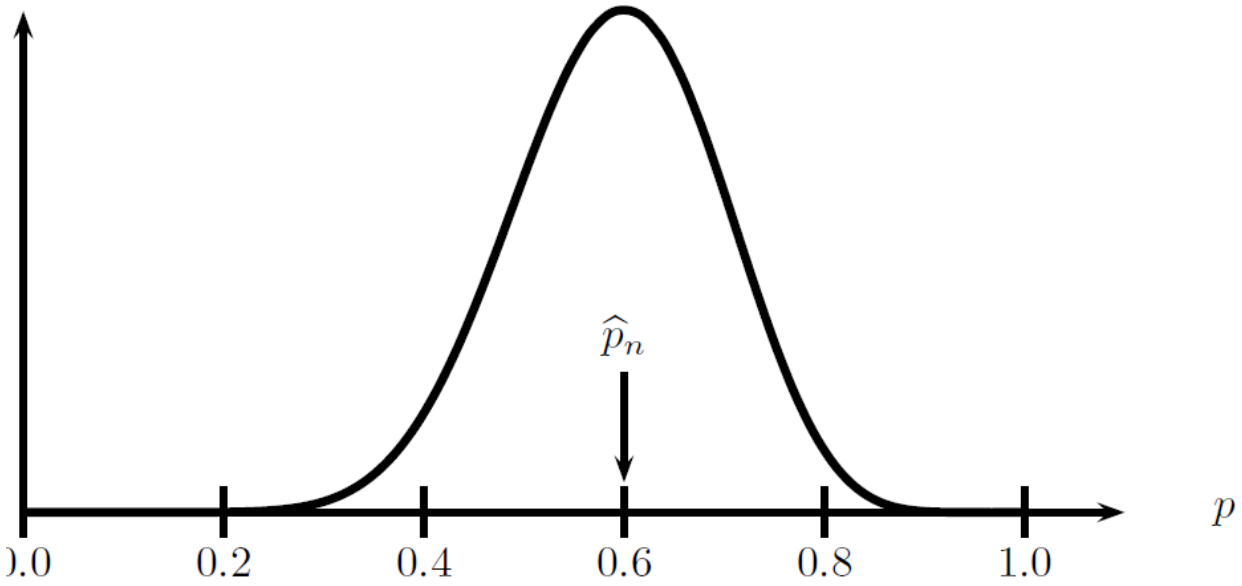
$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

The **log-likelihood function** *is defined by* $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we **treat it is a function of the parameter** θ**.** Thus, $\mathcal{L}_n : \Theta \to [0, \infty)$. The likelihood function is not a density function: in general, it is **not** true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

MÉTODO DE LA MÁXIMA VEROSIMILITUD

Definition. *The* **maximum likelihood estimator** MLE, *denoted by* $\widehat{\theta}_n$, *is the value of* θ *that maximizes* $\mathcal{L}_n(\theta)$.



EJEMPLO

Example. Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p. Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(p) = S \log p + (n - S) \log(1 - p).$$

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\widehat{p}_n = S/n$.

EJEMPLO

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned} \mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \end{aligned}$$

EJEMPLO

$$= \quad \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where $\overline{X} = n^{-1}\sum_i X_i$ is the sample mean and $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2$ and then expanding the square. The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood. ■

ESTIMADORES COMUNES

Scenario	Population parameter	Notation	Point estimate	Symbol(s)
1	Population proportion	p	Sample proportion	$\hat{p}$
2	Population mean	μ	Sample mean	$\overline{x}$ or $\widehat{\mu}$
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proportions	$\hat{p}_1 - \hat{p}_2$
4	Difference in population means	$\mu_1 - \mu_2$	Difference in sample means	$\overline{x}_1 - \overline{x}_2$
5	Population regression slope	β_1	Fitted regression slope	b_1 or $\widehat{\beta}_1$

A VECES NO ES POSIBLE USAR MÉTODOS EXACTOS

Open Access Article

Logistic Biplot by Conjugate Gradient Algorithms and Iterated SVD

by Jose Giovany Babativa-Márquez [1,2,*] and José Luis Vicente-Villardón [1]

[1] Department of Statistics, University of Salamanca, 37008 Salamanca, Spain

[2] Facultad de Ciencias de la Salud y del Deporte, Fundación Universitaria del Área Andina, Bogotá 1321, Colombia

* Author to whom correspondence should be addressed.

Academic Editor: Liangxiao Jiang

Mathematics **2021**, *9*(16), 2015; **https://doi.org/10.3390/math9162015**

Received: 24 June 2021 / Revised: 19 August 2021 / Accepted: 19 August 2021 / Published: 23 August 2021

(This article belongs to the Special Issue **Multivariate Statistics: Theory and Its Applications**)

https://www.mdpi.com/2227-7390/9/16/2015/htm

CONSISTENCIA

Definition. *A point estimator* $\widehat{\theta}_n$ *of a parameter* θ *is* **consistent** *if* $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

ERROR ESTÁNDAR

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}.$$

ERROR CUADRÁTICO MEDIO

$$\mathrm{MSE} = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2.$$

The MSE *can be written as*

$$\mathrm{MSE} = \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}_\theta(\widehat{\theta}_n).$$

EJEMPLO

Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\widehat{p}_n = n^{-1} \sum_i X_i$. Then $\mathbb{E}(\widehat{p}_n) = n^{-1} \sum_i \mathbb{E}(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\mathsf{se} = \sqrt{\mathbb{V}(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. The estimated standard error is $\widehat{\mathsf{se}} = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

■

EJEMPLO

Example. Returning to the coin flipping example, we have that $\mathbb{E}_p(\widehat{p}_n) = p$ so the bias $= p - p = 0$ and $\mathsf{se} = \sqrt{p(1-p)/n} \to 0$. Hence, $\widehat{p}_n \xrightarrow{\mathrm{P}} p$, that is, $\widehat{p}_n$ is a consistent estimator. ■

DISTRIBUCIÓN ASINTÓTICA

Definition. *An estimator is* **asymptotically Normal** *if*

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1).$$

EJERCICIOS

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\widehat{\lambda} = n^{-1} \sum_{i=1}^{n} X_i$. Find the bias, se, and MSE of this estimator.

Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = 2\overline{X}_n$. Find the bias, se, and MSE of this estimator.

ESTIMACIÓN DE FUNCIONES ESTADÍSTICAS

Función de distribución empírica

Let $X_1, \ldots, X_n \sim F$ be an IID sample where F is a distribution function on the real line. We will estimate F with the empirical distribution function.

Definition. *The* **empirical distribution function** $\widehat{F}_n$ *is the* CDF *that puts mass* $1/n$ *at each data point* X_i. *Formally,*

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$

PROPIEDADES

Theorem. *At any fixed value of* x,

$$\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \\
\text{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{\text{P}} F(x).
\end{aligned}$$

Theorem (The Glivenko-Cantelli Theorem). *Let* $X_1, \ldots, X_n \sim F$. *Then*

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{P}} 0.$$

BOOTSTRAP

The **bootstrap** is a method for estimating standard errors and computing confidence intervals. Let $T_n = g(X_1, \ldots, X_n)$ be a **statistic**, that is, T_n is any function of the data. Suppose we want to know $\mathbb{V}_F(T_n)$, the variance of T_n. We have written $\mathbb{V}_F$ to emphasize that the variance usually depends on the unknown distribution function F. For example, if $T_n = \overline{X}_n$ then $\mathbb{V}_F(T_n) = \sigma^2/n$ where $\sigma^2 = \int (x-\mu)^2 dF(x)$ and $\mu = \int x dF(x)$. Thus the variance of T_n is a function of F. The bootstrap idea has two steps:

Step 1: Estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}_n}(T_n)$.

Step 2: Approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ using simulation.

BOOTSTRAP

Suppose we draw an IID sample $Y_1, \ldots, Y_B$ from a distribution G. By the law of large numbers,

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{\text{P}} \int y\, dG(y) = \mathbb{E}(Y)$$

as $B \to \infty$. So if we draw a large sample from G, we can use the sample mean $\overline{Y}_n$ to approximate $\mathbb{E}(Y)$. In a simulation, we can make B as large as we like, in which case, the difference between $\overline{Y}_n$ and $\mathbb{E}(Y)$ is negligible. More generally, if h is any function with finite mean then

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow{\text{P}} \int h(y) dG(y) = \mathbb{E}(h(Y))$$

as $B \to \infty$.

BOOTSTRAP

as $B \to \infty$. In particular,

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y})^2 = \frac{1}{B}\sum_{j=1}^{B}Y_j^2 - \left(\frac{1}{B}\sum_{j=1}^{B}Y_j\right)^2$$

$$\xrightarrow{\text{P}} \int y^2 dF(y) - \left(\int y dF(y)\right)^2 = \mathbb{V}(Y).$$

Hence, we can use the sample variance of the simulated values to approximate $\mathbb{V}(Y)$.

ESTIMACIÓN DE LA VARIANZA BOOTSTRAP

According to what we just learned, we can approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ by simulation. Now $\mathbb{V}_{\widehat{F}_n}(T_n)$ means "the variance of T_n if the distribution of the data is $\widehat{F}_n$." How can we simulate from the distribution of T_n when the data are assumed to have distribution $\widehat{F}_n$? The answer is to simulate $X_1^*, \ldots, X_n^*$ from $\widehat{F}_n$ and then compute $T_n^* = g(X_1^*, \ldots, X_n^*)$. This constitutes one draw from the distribution of T_n. The idea is illustrated in the following diagram:

$$\begin{array}{rccccc} \text{Real world} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\ \text{Bootstrap world} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

How do we simulate $X_1^*, \ldots, X_n^*$ from $\widehat{F}_n$?

ESTIMACIÓN DE LA VARIANZA BOOTSTRAP

drawing an observation from $\widehat{F}_n$ is equivalent to drawing one point at random from the original data set.

Thus, to simulate $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ it suffices to draw n observations with replacement from $X_1, \ldots, X_n$.

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)^2 .$$

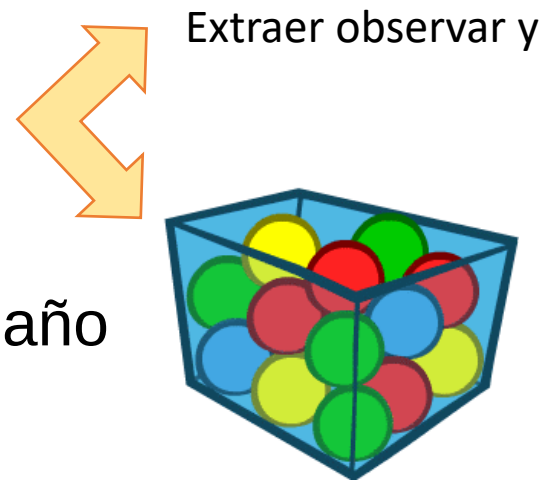
EJERCICIO

Diseñe un instrumento y registre:

- El nombre del estudiante.
- Sexo.
- Gasto de la semana.
- Marca de tecnología preferida.

Suponga que esta es una muestra aleatoria de todos los estudiantes de la maestría.

EJERCICIO



Usando el conjunto de datos, importe a R y

1. Seleccione una muestra con reemplazo del mismo tamaño

¿por qué reemplazamos? Porque de lo contrario, ¡terminaríamos con la misma muestra original! reemplazar induce la *variación de muestreo* .

2. Compare los histogramas de la muestra original y la remuestra 1.
3. Realice un procedimiento de remuestreo 40 veces y calcule la distribución Bootstrap del estimador
4. Realice el procedimiento de remuestreo 1000 veces y calcule la distribución Bootstrap.

ESTIMACIÓN POR INTERVALO

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

In words, (a, b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

REMARKS

Warning! C_n is random and θ is fixed.

Commonly, people use 95 percent confidence intervals, which corresponds to choosing $\alpha = 0.05$. If θ is a vector then we use a **confidence set** (such as a sphere or an ellipse) instead of an interval.

Warning! There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ since θ is a fixed quantity, not a random variable.

EL CONCEPTO DE CONFIABILIDAD

Una vez hecha la estimación la validez está dada por la estimación por intervalo

Definición

Si $\widehat{\theta}$ es una función basada en una suma de variables aleatorias independientes el teorema central de límite permite encontrar una expresión para la estimación por intervalo bajo ciertas condiciones de regularidad. En caso de que $\mathbb{E}(\widehat{\theta}) = \theta$ se espera con una confiabilidad del $(1-\alpha)100\,\%$ que:

$$\theta \in \left(\widehat{\theta} - z_{1-\alpha/2}\sqrt{V(\widehat{\theta})}, \widehat{\theta} + z_{1-\alpha/2}\sqrt{V(\widehat{\theta})}\right),$$

donde $z_{1-\alpha/2}$ es el percentil correspondiente en una distribución normal estándar.

¿QUE SIGNIFICA QUE HAYA UN 95% DE CONFIABILIDAD?

CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA

Theorem (Normal-based Confidence Interval). *Suppose that* $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. *Let* Φ *be the* CDF *of a standard Normal and let* $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, *that is,* $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ *and* $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ *where* $Z \sim N(0,1)$. *Let*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}).$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha.$$

PROOF. Let $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\mathsf{se}}$. By assumption $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$. Hence,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} < z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1-\alpha. \quad \blacksquare
\end{aligned}$$

CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA

BOOTSTRAP: Método del error estándar

The Normal Interval. The simplest method is the Normal interval

$$T_n \pm z_{\alpha/2}\ \widehat{\mathsf{se}}_{\mathrm{boot}}$$

where $\widehat{\mathsf{se}}_{\mathrm{boot}} = \sqrt{v_{\mathrm{boot}}}$ is the bootstrap estimate of the standard error. This interval is not accurate unless the distribution of T_n is close to Normal.

CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA

BOOTSTRAP: Método del percentil

Percentile Intervals. The **bootstrap percentile interval** is defined by

$$C_n = \left(\theta^*_{\alpha/2},\ \theta^*_{1-\alpha/2}\right).$$

EJERCICIO

Considere los resultados del gasto

1. Genere el espacio muestral de las muestras de tamaño n = 3.
2. Considere los valores del gasto en cada elemento seleccionado para cada muestra.
3. Estime la media y desviación estándar a partir de cada muestra.
4. Construya un intervalo usando una probabilidad del 95%.
5. Determine el porcentaje de intervalos que cubren al valor verdadero.

Use un enfoque Bootstrap para este ejercicio. Es posible que lo hagamos con una muestra más grande? .

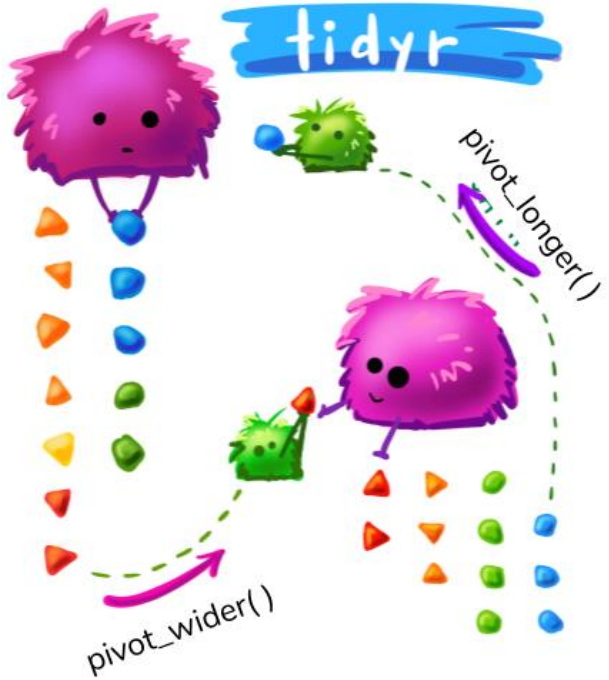
VERBOS -- TIDYVERSE

Base larga a base ancha **spread()**, **pivot_wider()**

Base ancha a base larga **gather()**, **pivot_longer()**

Separar cadena en columnas **separate()**

Unir columnas **unite()**

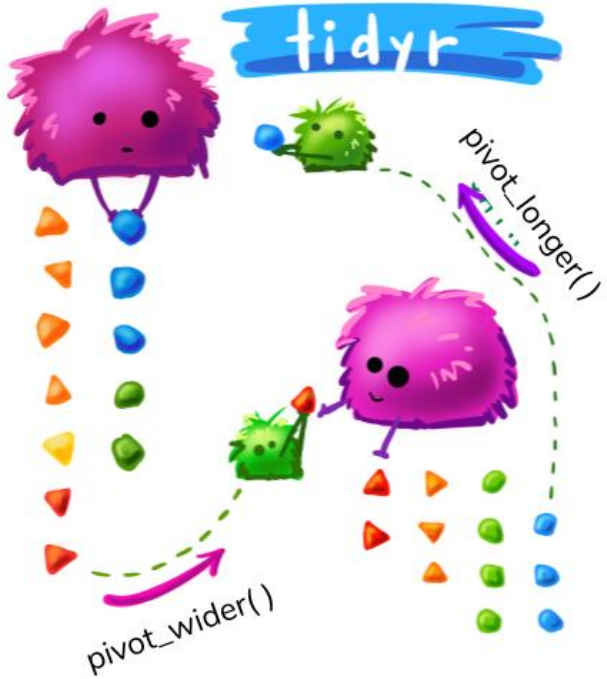


PIVOT

```
cases <- tribble(
  ~Country, ~"2011", ~"2012", ~"2013",
       "FR",    7000,    6900,    7000,
       "DE",    5800,    6000,    6200,
       "US",   15000,   14000,   13000
)
```

country	year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

```
cases %>% pivot_longer(-country, names_to = "year", values_to = "n")
```



EJEMPLO

Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Then $\mathbb{V}(\widehat{p}_n) = n^{-2}\sum_{i=1}^n \mathbb{V}(X_i) = n^{-2}\sum_{i=1}^n p(1-p) = n^{-2}np(1-p) = p(1-p)/n$. Hence, $\mathsf{se} = \sqrt{p(1-p)/n}$ and $\widehat{\mathsf{se}} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$. By the Central Limit Theorem, $\widehat{p}_n \approx N(p, \widehat{\mathsf{se}}^2)$. Therefore, an approximate $1-\alpha$ confidence interval is

$$\widehat{p}_n \pm z_{\alpha/2}\widehat{\mathsf{se}} = \widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

INTERVALO DE CONFIANZA: MEDIA

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población de la que se conoce su varianza σ^2, lo que da un intervalo de confianza de $100(1-\alpha)\%$ para μ es

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

INTERVALO DE CONFIANZA: MEDIA

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$ tiene una distribución t de Student con $n - 1$ grados de libertad.

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

¿Cuál es la restricción paramétrica?

SUPUESTOS

If we assume the data come from a parametric model, then it is a good idea to check that assumption. One possibility is to check the assumptions informally by inspecting plots of the data. For example, if a histogram of the data looks very bimodal, then the assumption of Normality might be questionable. A formal way to test a parametric model is to use a **goodness-of-fit test.**

INTERVALOS DE CONFIANZA

Diferencia de medias con varianzas conocidas

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Diferencia de medias con varianzas desconocidas pero iguales

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t \text{ con } v = n_1 + n_2 - 2$$

Diferencia de medias con varianzas desconocidas y diferentes

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

MUESTRAS PAREADAS

$$D_i = X_{1i} - X_{2i}.$$

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad

INTERVALO DE CONFIANZA PARA LA PROPORCIÓN

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado del $100(1 - \alpha)\%$ para el parámetro binomial p se obtiene por medio de

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

DIFERENCIA DE PROPORCIONES

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatorias de tamaños n_1 y n_2, respectivamente, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, un intervalo de confianza aproximado del $100(1 - \alpha)\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$ es dado por

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

INTERVALO DE CONFIANZA PARA LA VARIANZA

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza del $100(1-\alpha)\%$ para σ^2 es

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son valores χ^2 con $v = n-1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente, a la derecha.

COCIENTE DE LAS VARIANZAS

Si s_1^2 y s_2^2 son las varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, tomadas de poblaciones normales, entonces un intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2 es

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1),$$

donde $f_{\alpha/2}(v_1, v_2)$ es un valor f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha, y $f_{\alpha/2}(v_2, v_1)$ es un valor f similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad.

EJEMPLOS

El contenido de ácido sulfúrico de 7 contenedores similares es de 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, y 9.6 litros. Calcule un intervalo de confianza del 95% para el contenido promedio de todos los contenedores suponiendo una distribución aproximadamente normal.

$$\bar{x} = 10.0 \quad \text{y} \quad s = 0.283.$$

```
> qt(p = 0.025, df=6)
[1] -2.446912
> qt(p = 0.975, df=6)
[1] 2.446912
```

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right)$$

$$9.74 < \mu < 10.26.$$

EJEMPLOS

Los siguientes son los tiempos de secado (minutos) de hojas cubiertas de poliuretano bajo dos condiciones ambientales diferentes:

Condición 1	55.6	56.1	61.8	55.9	51.4	59.9	54.3	62.8	58.5	55.8
	58.3	60.2	54.2	50.1	57.1	57.5	63.6	59.3	60.9	61.8
Condición 2	55.1	43.5	51.2	46.2	56.7	52.5	53.5	60.5	52.1	47.0
	53.0	53.8	51.6	53.6	42.9	52.0	55.1	57.1	62.8	54.8

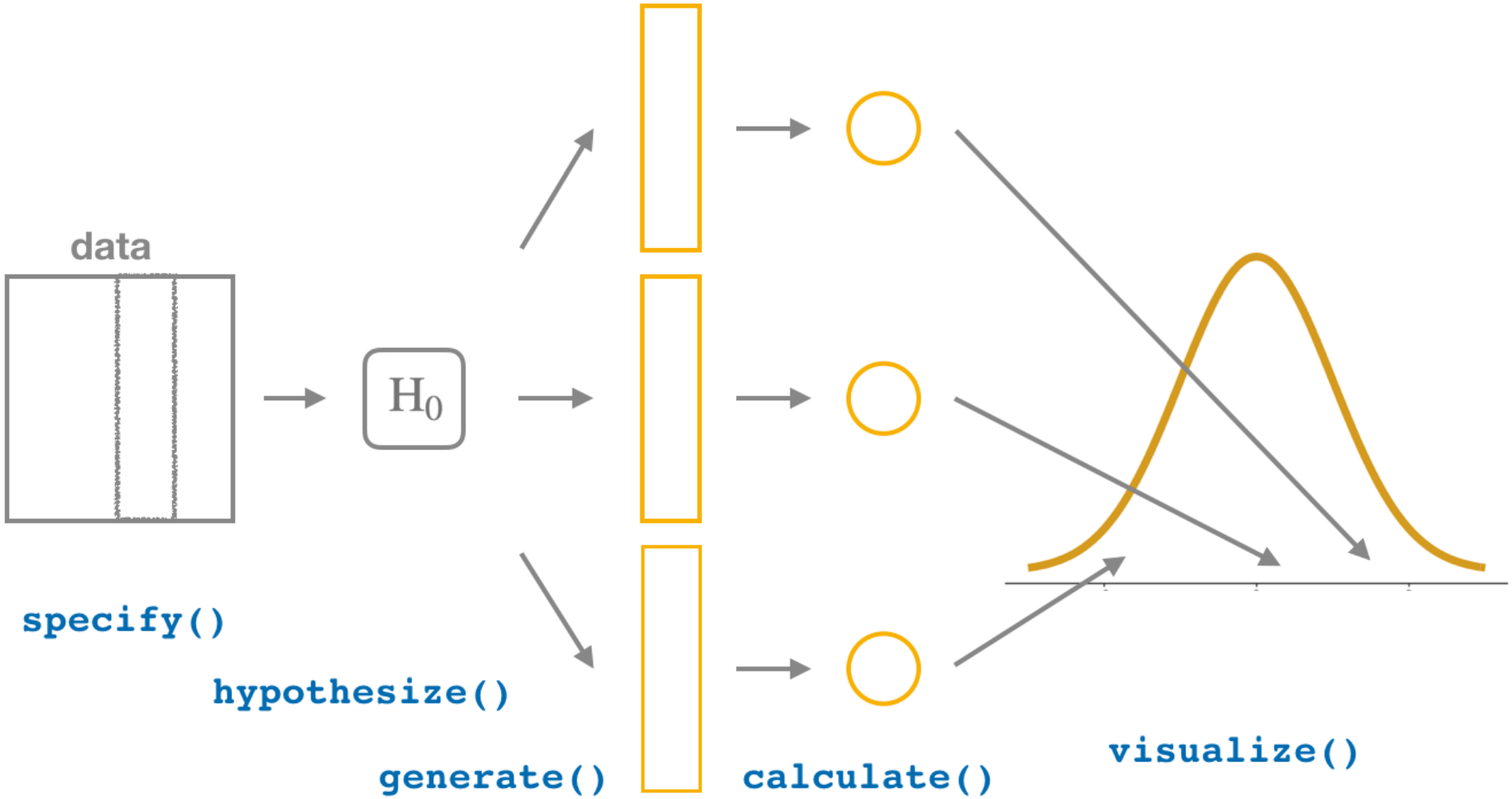
Halle un intervalo de 98% confianza para la diferencia entre las medias de los tiempos de secado bajo las dos condiciones ambientales. Suponga que las muestras son independientes entre si y provienen de poblaciones normales.

Halle un intervalo de 95% de confianza para la proporción de hojas cubiertas de poliuretano con tiempos de secado mayores que 60. No discrimine por condición ambiental.

Calcule el intervalo de confianza del 95% para la diferencia de proporciones entre las condiciones ambientales.

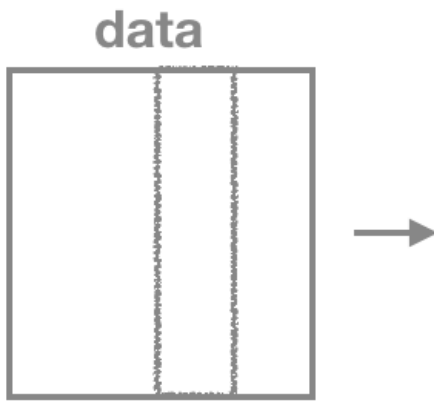
INTERVALOS DE CONFIANZA
DATA SCIENCE

PAQUETE infer – Flujo de trabajo



PAQUETE infer

Extraer variables para la inferencia

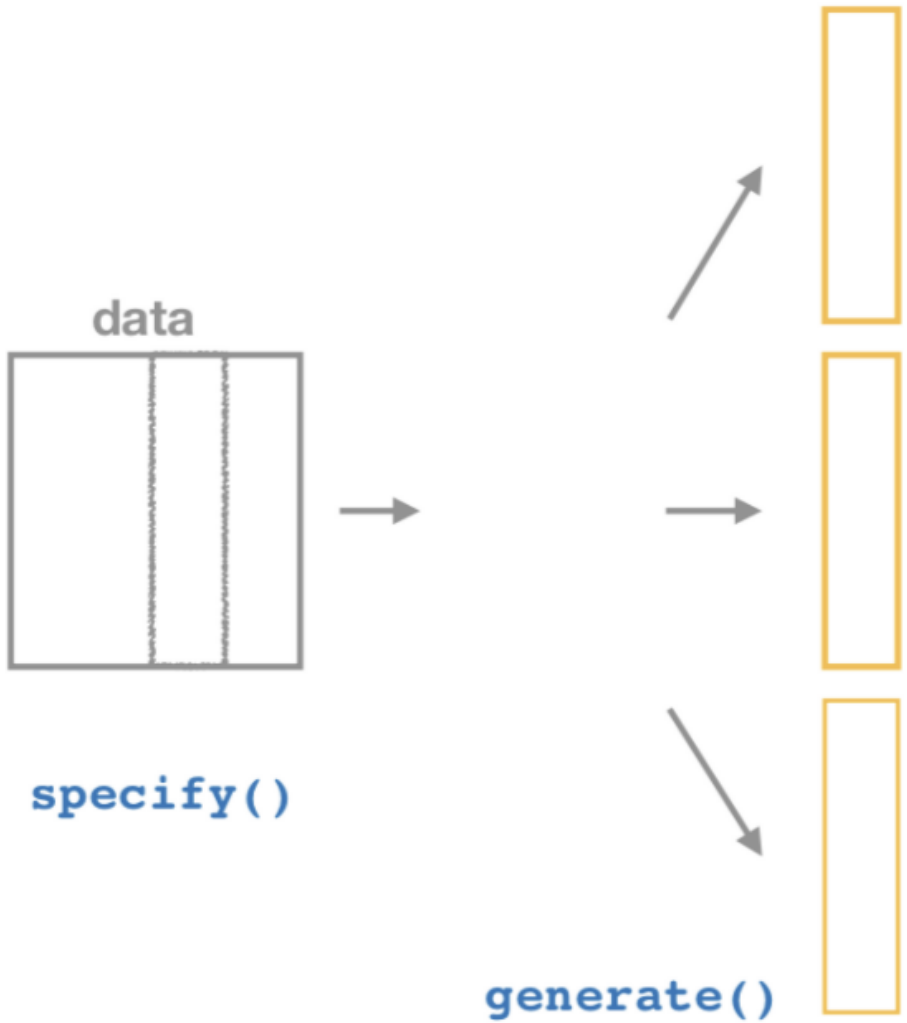


specify()

```
datap %>%
  specify(formula = tiempo ~ NULL) %>%
  glimpse()
```

```
## Rows: 40
## Columns: 1
## $ tiempo <dbl> 55.6, 55.1, 56.1, 43.5,
```

PAQUETE infer

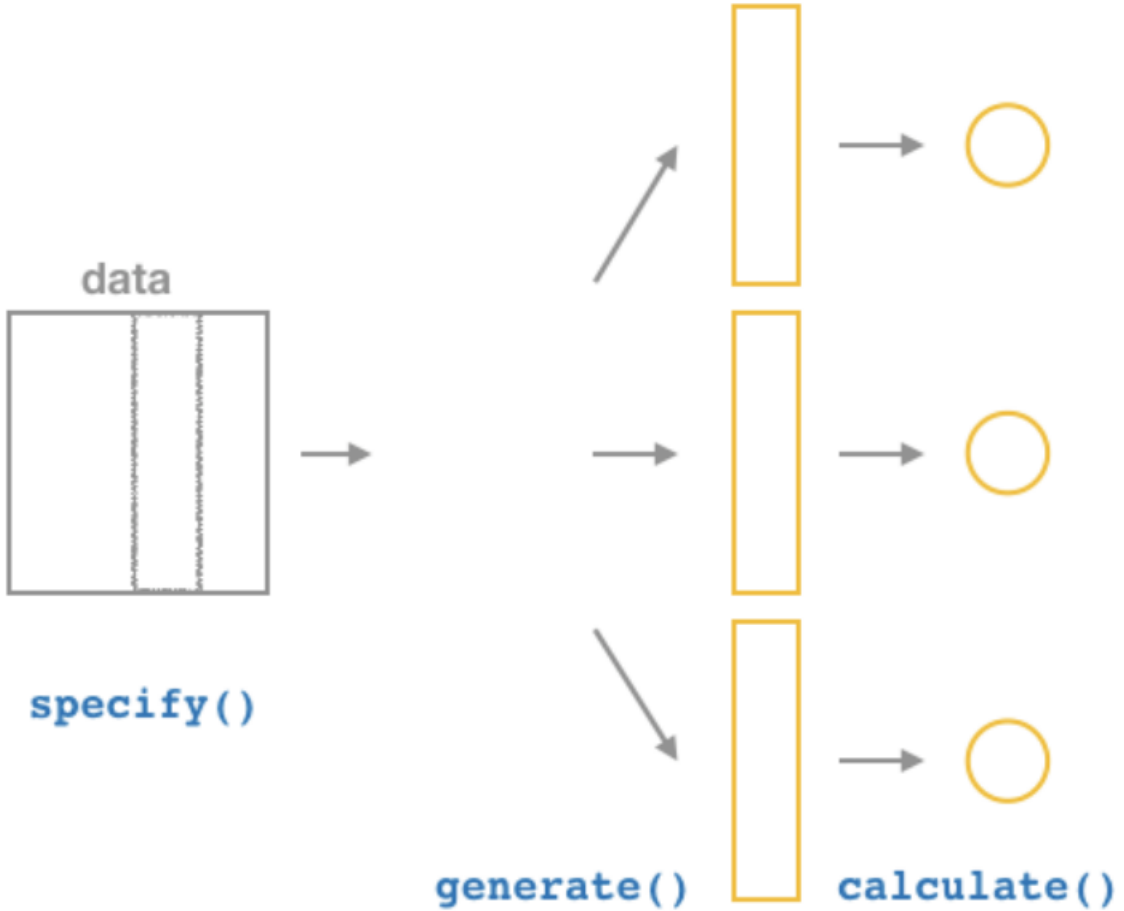


Generar réplicas

```
datap %>%
  specify(formula = tiempo ~ NULL) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  glimpse()
```

```
## Rows: 400,000
## Columns: 2
## Groups: replicate [10,000]
## $ replicate <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
## $ tiempo    <dbl> 55.1, 43.5, 54.8, 61.8, 42.9,
```

PAQUETE infer

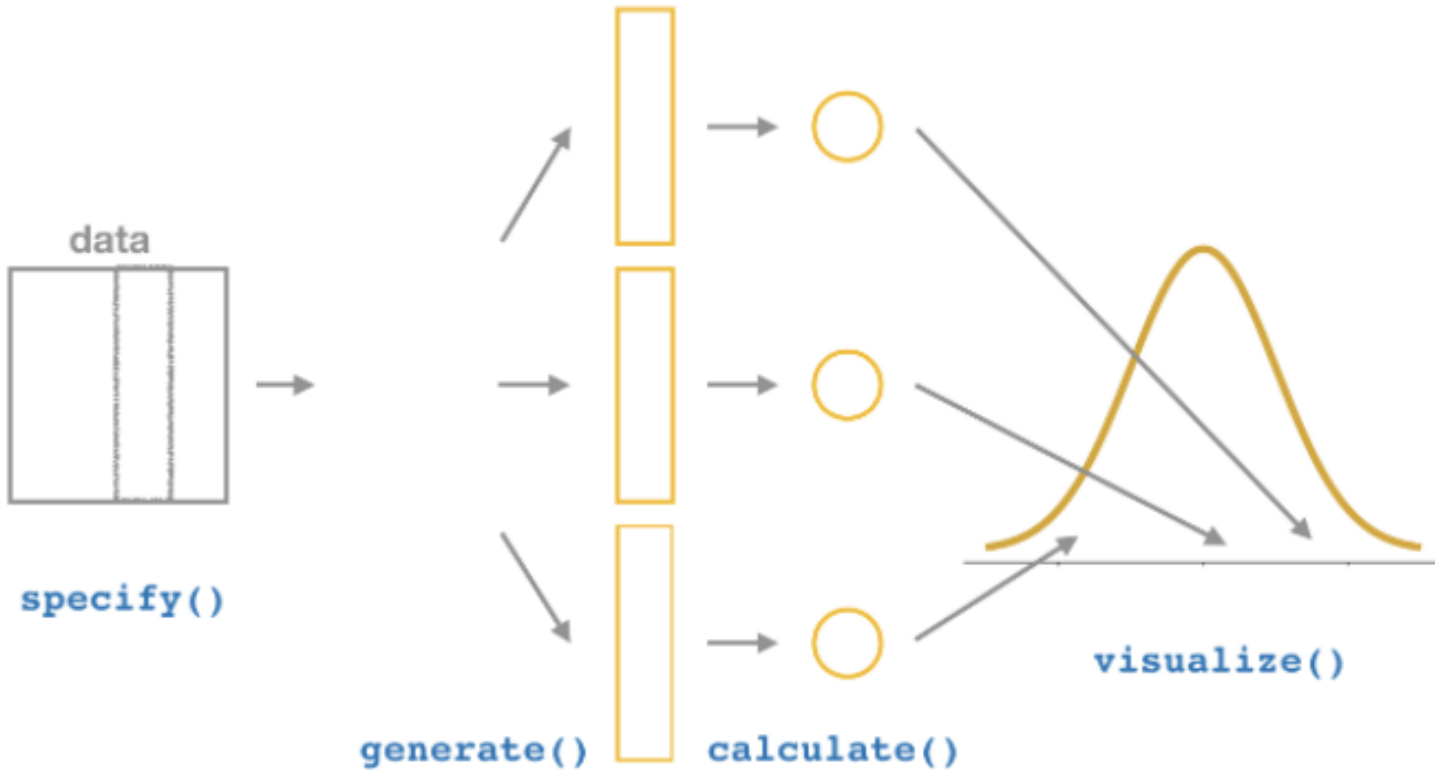


Cálculo de la estadística en cada réplica

```
dist.boot <- datap %>%
  specify(formula = tiempo ~ NULL) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  glimpse()
```

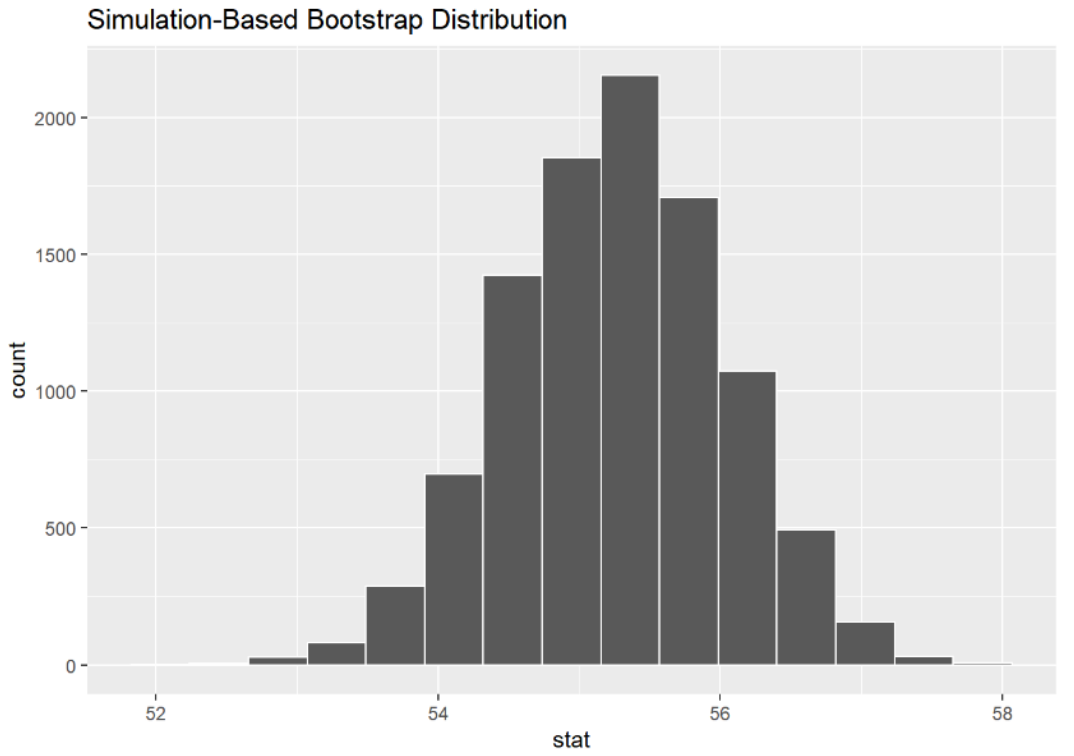
```
## Rows: 10,000
## Columns: 2
## $ replicate <int> 1, 2, 3, 4, 5, 6, 7, 8, 9,
## $ stat      <dbl> 54.7625, 55.8175, 55.1050,
```

PAQUETE infer

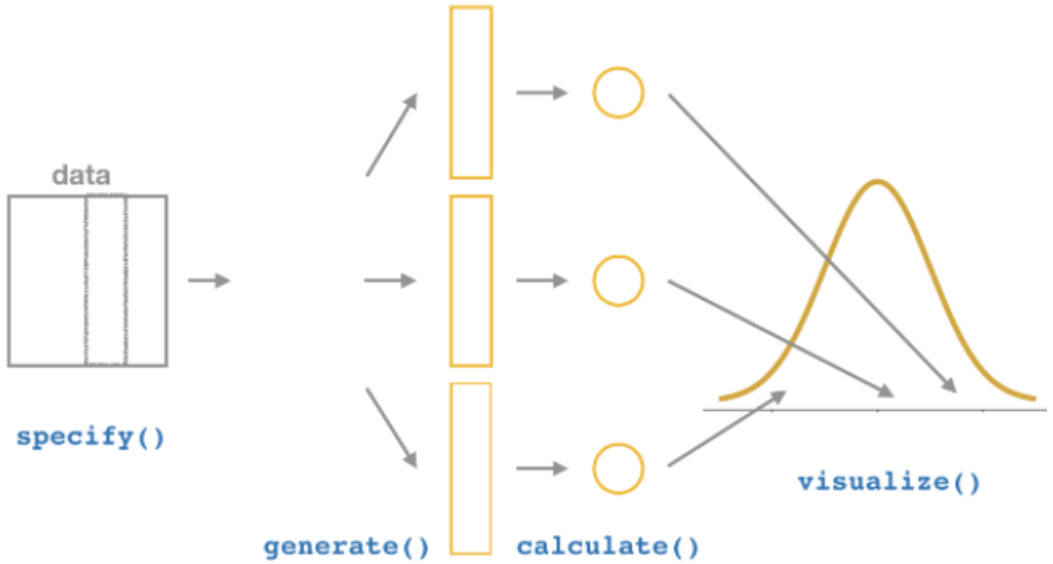


Visualización

```
visualise(dist.boot)
```



PAQUETE infer



Visualización

```
visualize(dist.boot) +
  shade_ci(endpoints = ic_perc, color = "hotpink", fill = "khaki")
```