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,*

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

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

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

where $\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ 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

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

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?

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 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

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

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

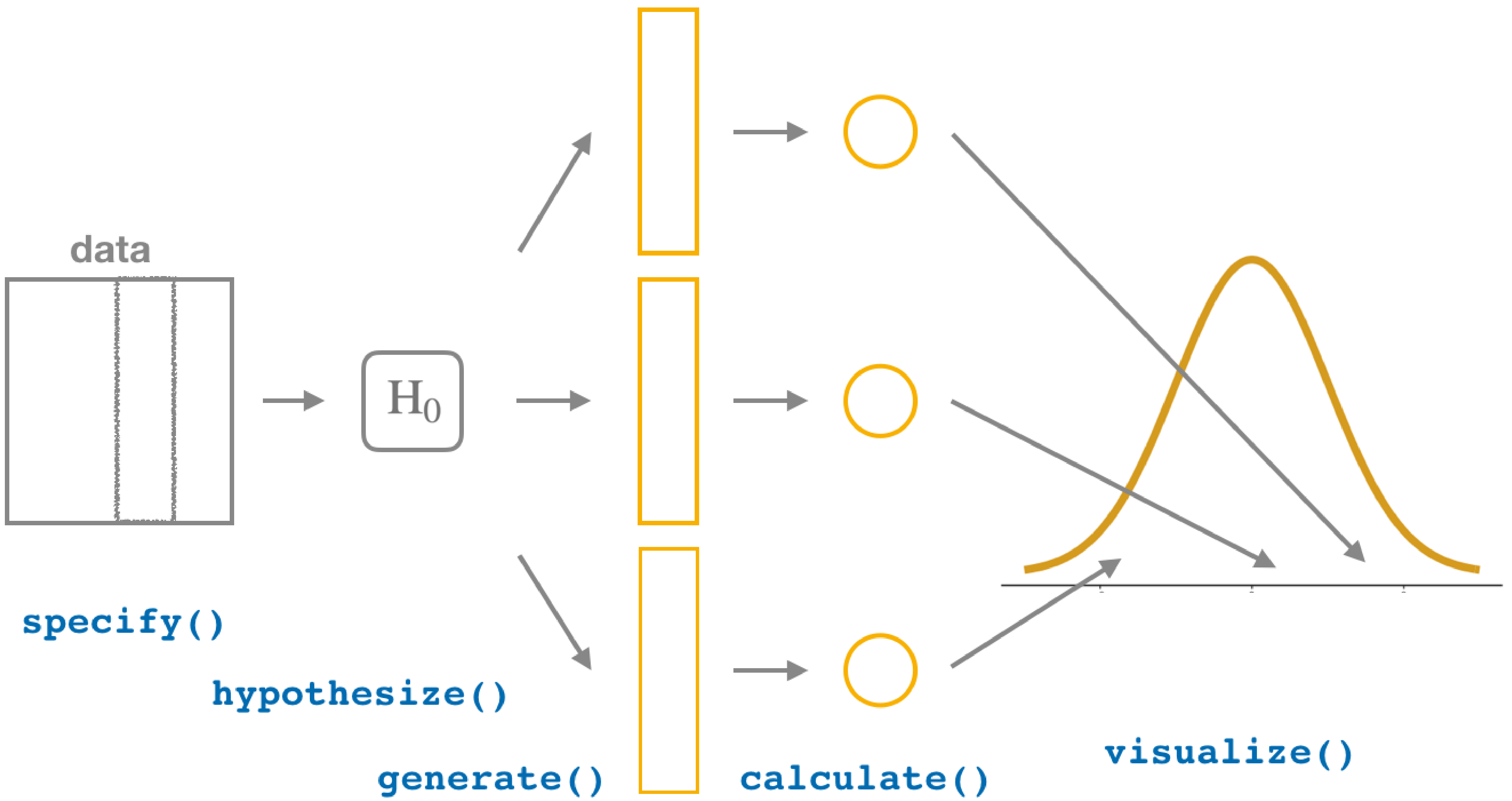
Halle un intervalo de 98% confianza para la diferencia entre las medias de los tiempos de secado bajo las dos condiciones ambientales. Suponga que las muestras son independientes entre si y provienen de poblaciones normales.

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.

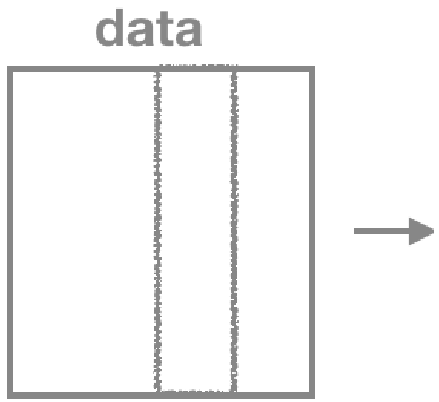
INTERVALOS DE CONFIANZA
DATA SCIENCE

PAQUETE infer – Flujo de trabajo



PAQUETE infer

Extraer variables para la inferencia

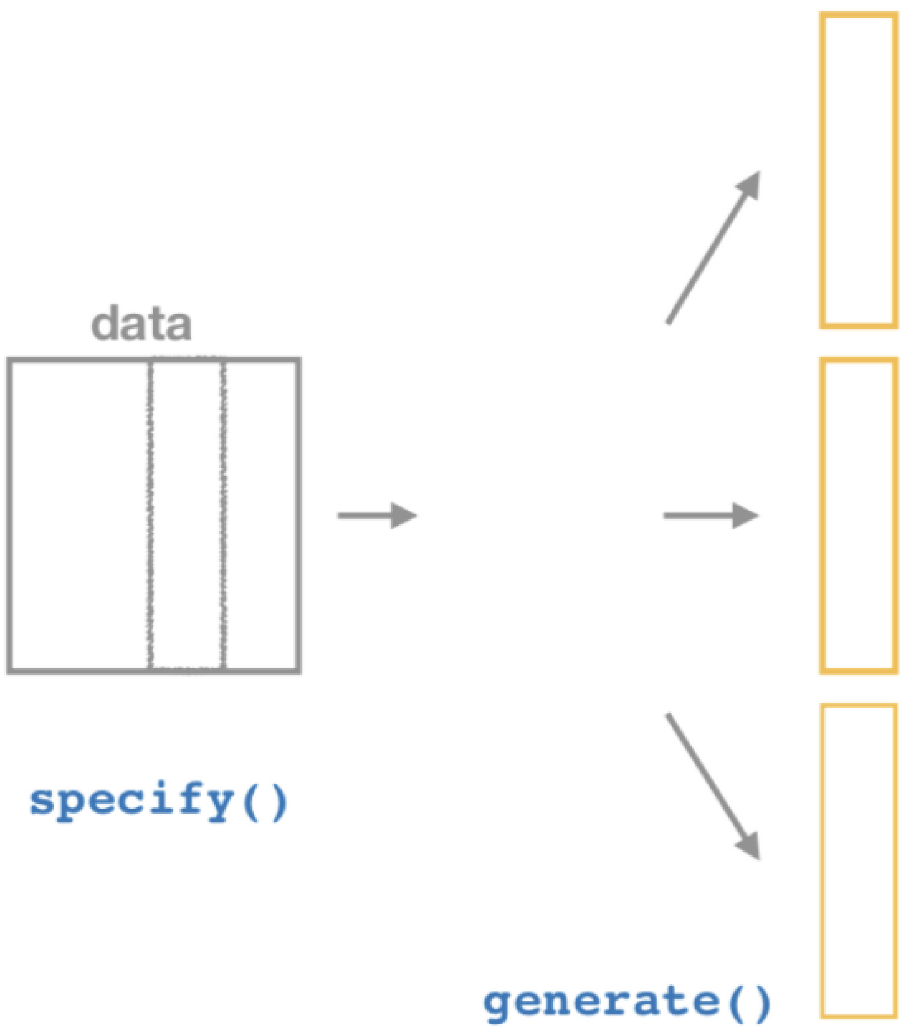


`specify()`

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

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

PAQUETE infer

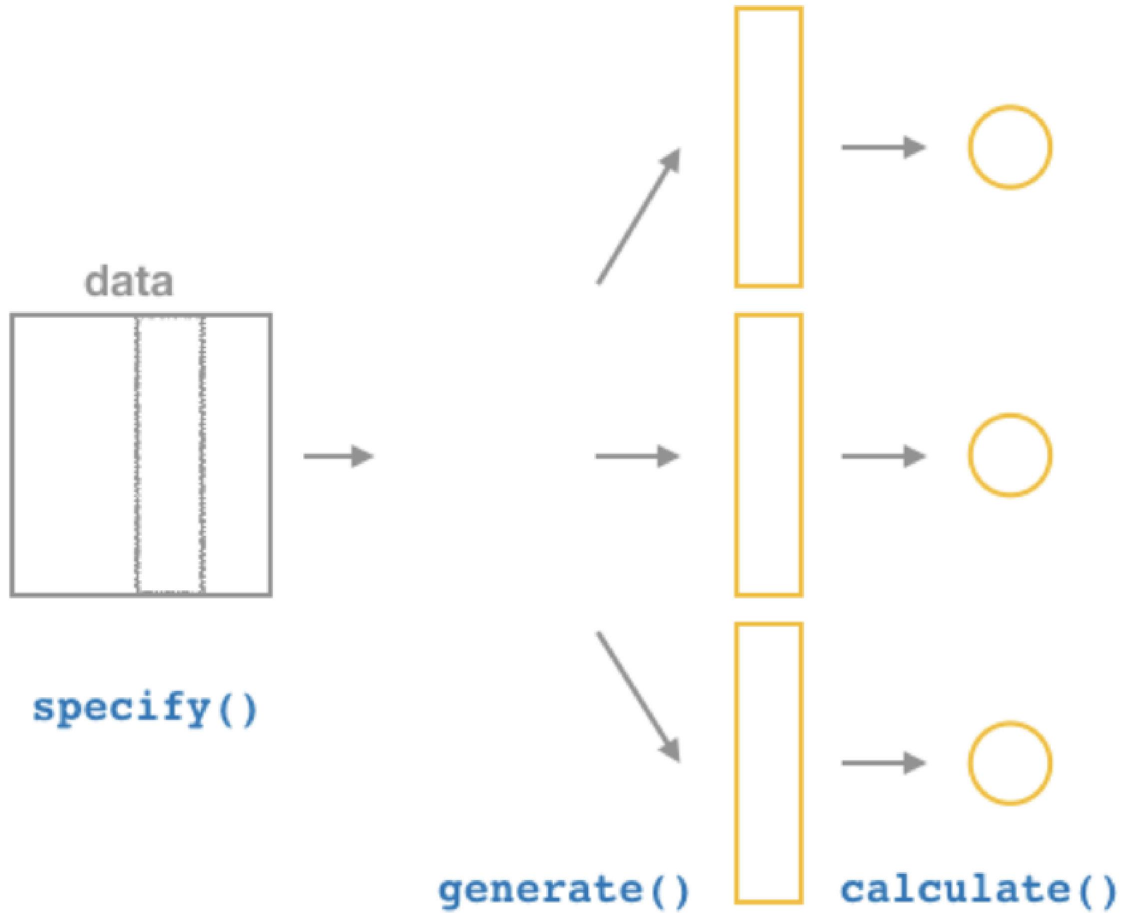


Generar réplicas

```
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

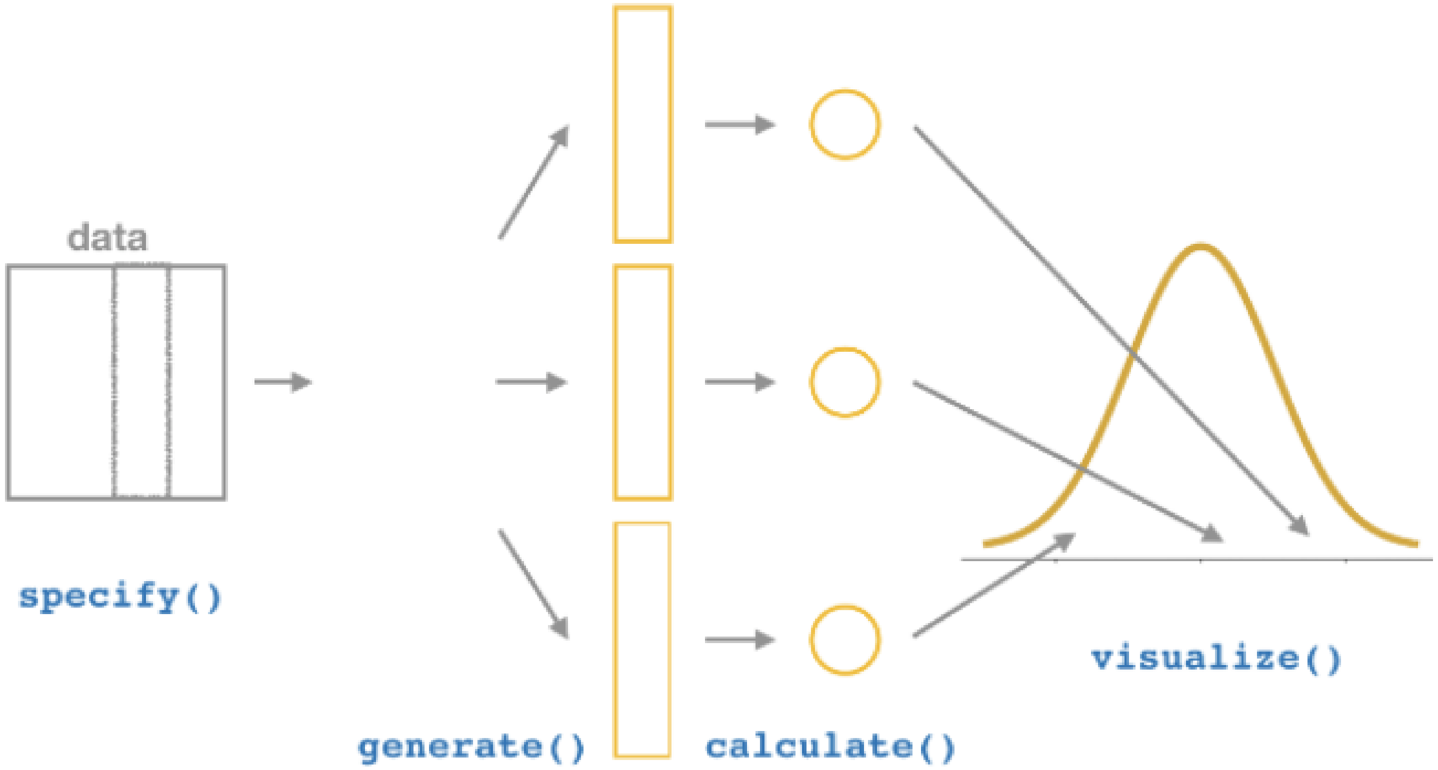


Cálculo de la estadística en cada réplica

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

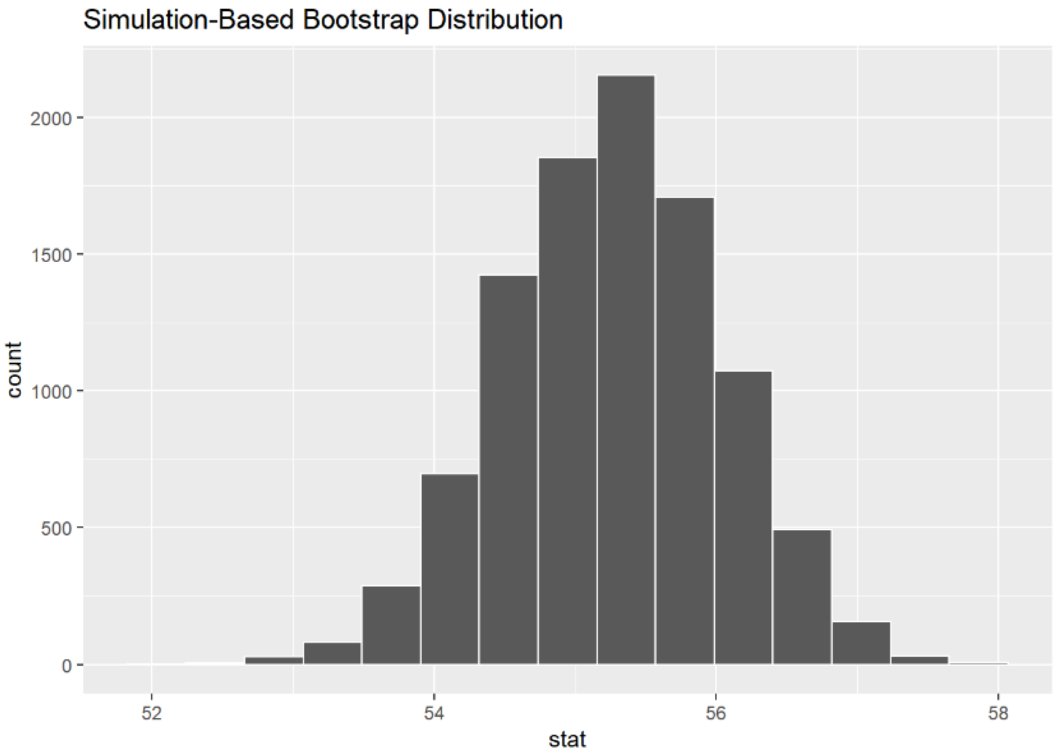
```
## Rows: 10,000
## Columns: 2
## $ replicate <int> 1, 2, 3, 4, 5, 6, 7, 8, 9,
## $ stat      <dbl> 54.7625, 55.8175, 55.1050,
```

PAQUETE infer

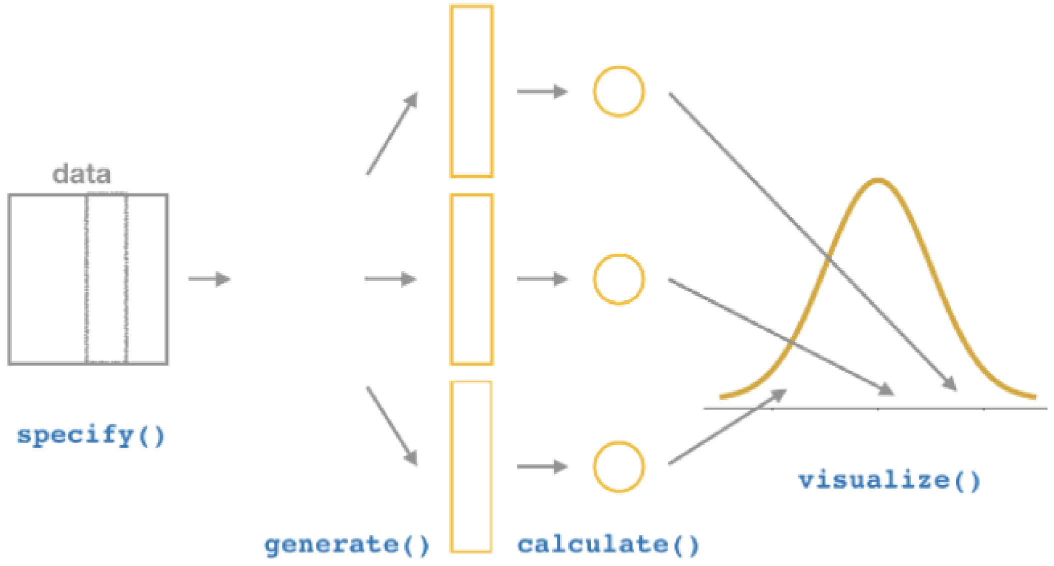


Visualización

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

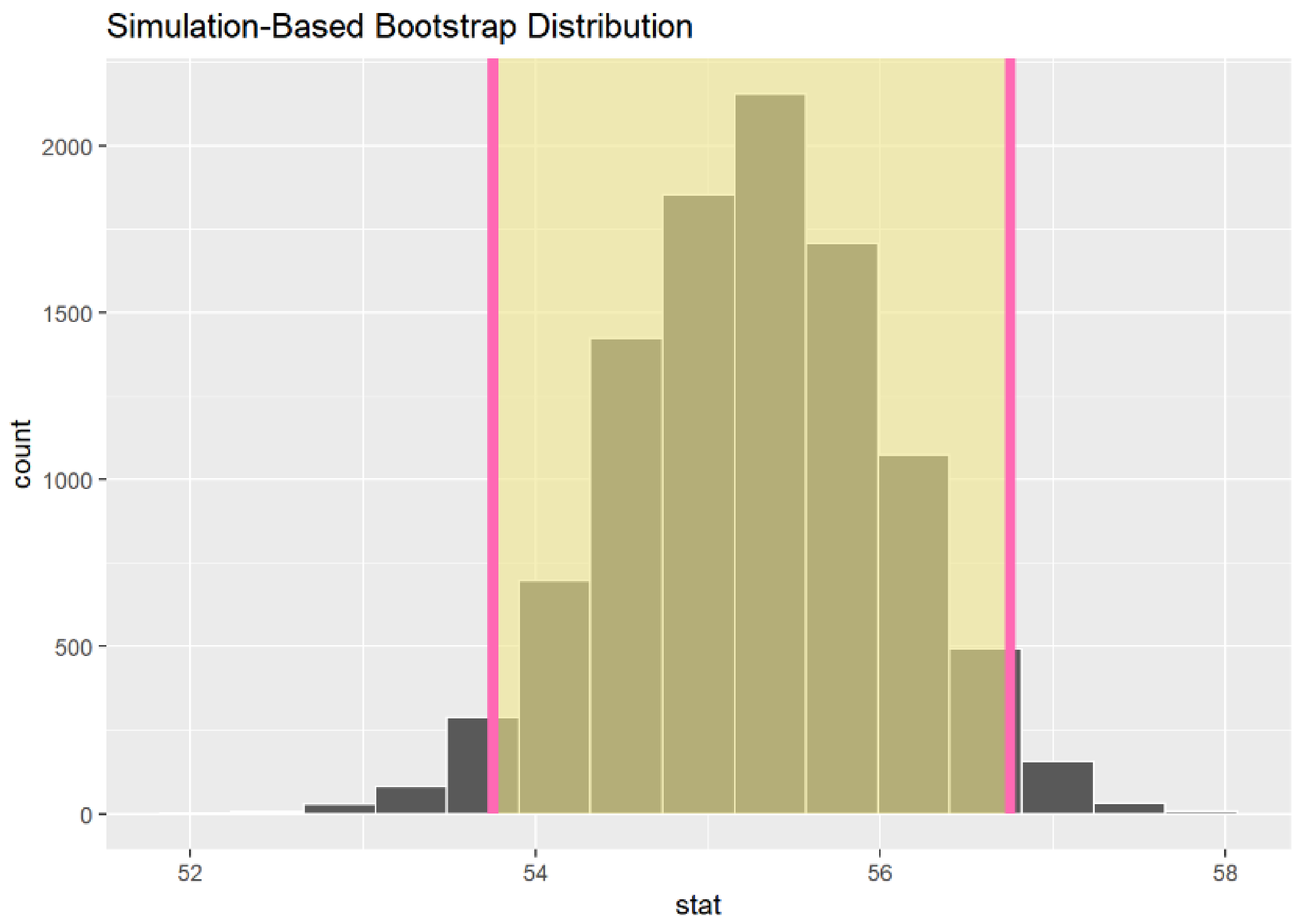


PAQUETE infer



Visualización

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



COSIDERACIONES

La función **specify()** permite especificar la variable de resultado y las variables explicativas o escribirlo como una formula.

Usage

```
specify(x, formula, response = NULL, explanatory = NULL, success = NULL)
```

En el caso de variables categóricas establecidas como *response* se debe especificar el suceso que representa el éxito en el argumento *success.*

PRUEBAS DE HIPÓTESIS

HIPÓTESIS DE INVESTIGACIÓN

- Los establecimientos que tienen nevera venden más producto
- La intención de compra es mayor en Pasto que en Bogotá
- El nivel de satisfacción es mayor en lo Urbana que en lo rural
- Entre los 5 productos, el producto 1 agrada más que todos

SISTEMA DE HIPÓTESIS

More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

REGIÓN DE RECHAZO

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$
\begin{aligned}
X \in R &\implies \text{reject } H_0 \\
X \notin R &\implies \text{retain (do not reject) } H_0
\end{aligned}
$$

Usually, the rejection region R is of the form

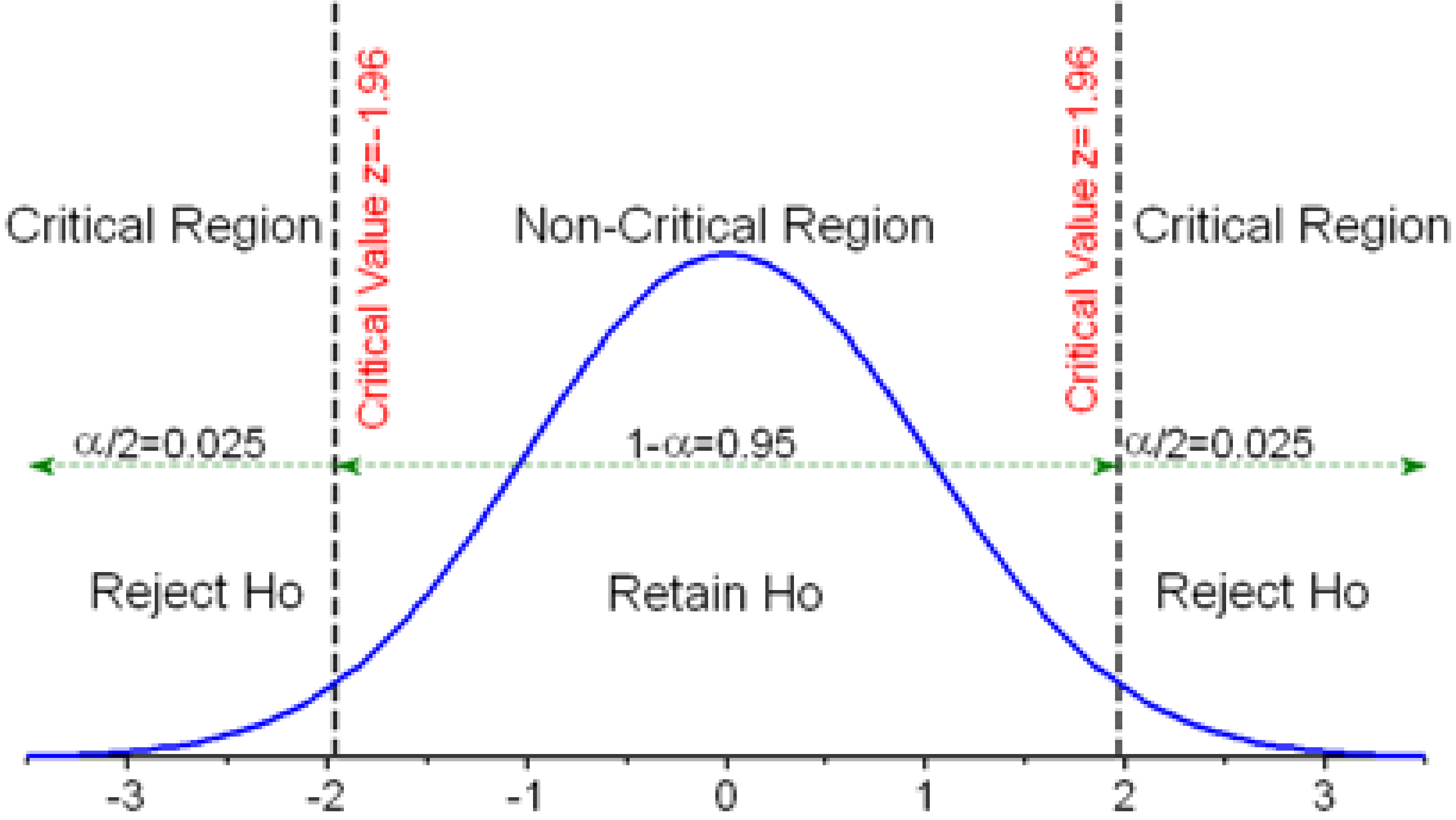
$$R = \left\{ x : \ T(x) > c \right\} \tag{10.2}$$

where T is a **test statistic** and c is a **critical value**. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

PRUEBA DE HIPÓTESIS

Ho: Hipótesis Nula

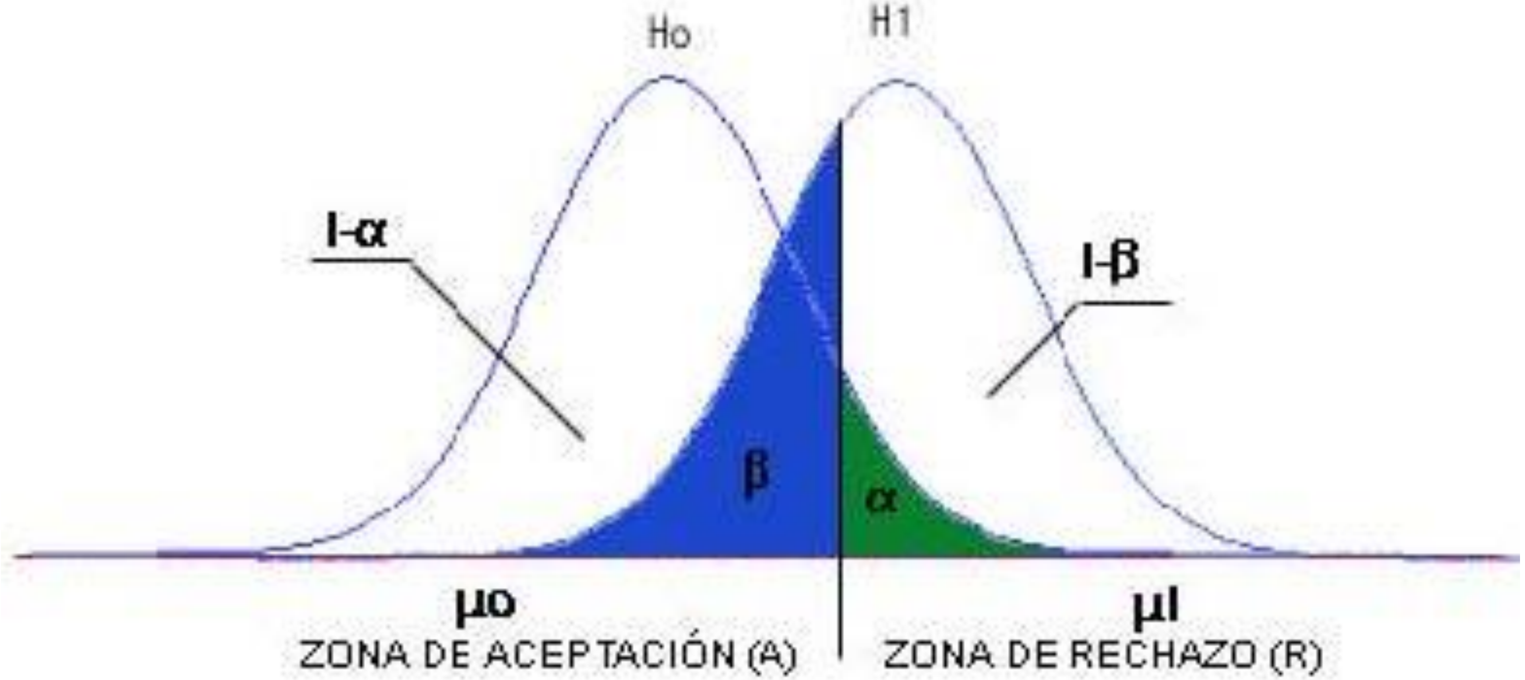
K1: Hipótesis Alterna



TIPOS DE ERROR

Contraste de hipótesis		Resultado real	
		Ho	H1
Resultado encontrado	Ho	Acierto	Error tipo II
	H1	Error tipo I	Acierto

TIPOS DE ERROR



POTENCIA DE UNA PRUEBA ESTADÍSTICA

Definition. *The* **power function** *of a test with rejection region R is defined by*

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

The **size** *of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

A test is said to have **level** α *if its size is less than or equal to* α*.*

TIPOS DE HIPÓTESIS

A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**. A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called a **two-sided test**. A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a **one-sided test**. The most common tests are two-sided.

EJEMPLO

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Consider the test:

$$\text{reject } H_0 \text{ if } T > c$$

where $T = \overline{X}$. The rejection region is

$$R = \Big\{(x_1, \ldots, x_n) : \; T(x_1, \ldots, x_n) > c\Big\}.$$

Let Z denote a standard Normal random variable. The power function is

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}_\mu\left(\overline{X} > c\right) \\
&= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).
\end{aligned}$$

¿Cuál es el tipo de hipótesis?

Dibuje la función de potencia en **R**

NOTA

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma\, \Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

We reject when $\overline{X} > \sigma\, \Phi^{-1}(1-\alpha)/\sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n}\,(\overline{X} - 0)}{\sigma} > z_\alpha.$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. ■

PRUEBA DE WALD

Denominado así en honor a Abraham Wald (1902-1950), quién murió en un accidente aéreo en India en 1950.

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ *is asymptotically Normal:*

$$\frac{(\widehat{\theta} - \theta_0)}{\widehat{\text{se}}} \rightsquigarrow N(0, 1).$$

The size α **Wald test** *is: reject* H_0 *when* $|W| > z_{\alpha/2}$ *where*

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\text{se}}}.$$

PRUEBA DE WALD

Theorem. *Asymptotically, the Wald test has size α, that is,*

$$\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) \to \alpha$$

as $n \to \infty$.

Proof. Under $\theta = \theta_0$, $(\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$. Hence, the probability of rejecting when the null $\theta = \theta_0$ is true is

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}\left(|Z| > z_{\alpha/2}\right) \\ &= \alpha \end{aligned}$$

where $Z \sim N(0,1)$. ∎

PRUEBA DE WALD

Theorem. *Suppose the true value of* θ *is* $\theta_\star \neq \theta_0$. *The power* $\beta(\theta_\star)$ — *the probability of correctly rejecting the null hypothesis* — *is given (approximately) by*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right).$$

Example (Comparing Two Prediction Algorithms). We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$ write

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

CRITERIO DEL P-VALOR

Reporting "reject H_0" or "retain H_0" is not very informative. Instead, we could ask, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest α at which the test rejects and we call this number the p-value.

Definition. *Suppose that for every* $\alpha \in (0,1)$ *we have a size* α *test with rejection region* R_α. *Then,*

$$\text{p-value} = \inf\Big\{\alpha :\ T(X^n) \in R_\alpha\Big\}.$$

That is, the p-value is the smallest level at which we can reject H_0.

CRITERIO DEL P-VALOR

Informally, the p-value is a measure of the evidence against H_0: the smaller the p-value, the stronger the evidence against H_0. Typically, researchers use the following evidence scale:

p-value	evidence
$< .01$	very strong evidence against H_0
$.01 - .05$	strong evidence against H_0
$.05 - .10$	weak evidence against H_0
$> .1$	little or no evidence against H_0

Warning! A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.