

Técnicas de análise cuantitativas y cualitativas

Sesión 1

Eduardo Corbelle Rico

Máster Universitario en Xestión Sustentable da Terra e o Territorio
Universidade de Santiago de Compostela

Curso 2015–2016

Objetivos de la sesión

- Estadística descriptiva
 - Estadísticos de posición y dispersión
 - Principales tipos de gráficos
- Estadística inferencial
 - Conceptos básicos de muestreo
 - Intervalos de confianza
 - Contrastes de hipótesis
- Introducción a R

Contidos

- 1 Presentación de la materia
- 2 Estadística descriptiva
- 3 Estadística inferencial: error de muestreo, contraste hipótesis

Descripción de la materia

Título Técnicas de análisis cuantitativas y cualitativas

Carácter Obligatoria

Duración 3 ECTS (75 h)

- 21 h presenciales
- 3 h tutorías
- 51 h trabajo individual

Calendario

Octubre 2015

l	m	m	x	v	s	d
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Profesor

Eduardo Corbelle Rico

eduardo.corbelle@usc.es

Escola Politécnica Superior, Lugo
Pav. II, planta 2^a inferior (despacho 11)

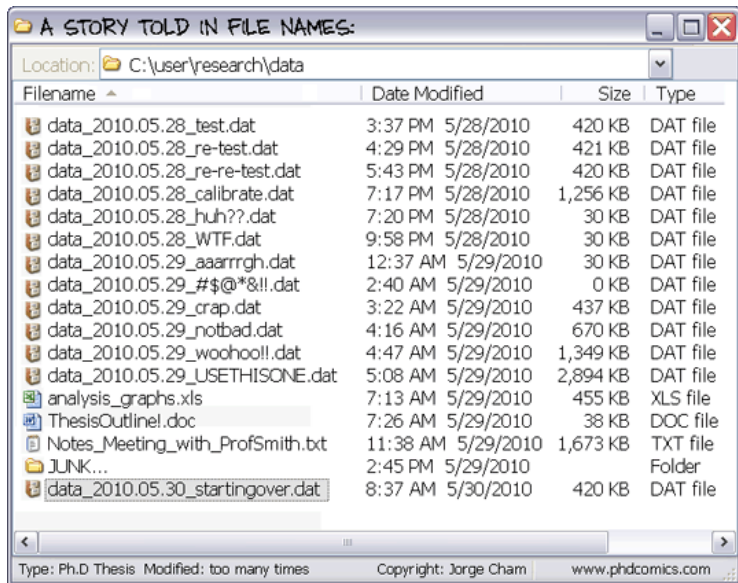
Tutorías: lunes y viernes, 9–12h

Crterios de evaluacón

- Resolución de 4 casos prácticos
- Entrega inicial hasta 29 de noviembre de 2015
 - Recomendaciones de mejora: 20 de diciembre de 2015
- Fecha límite de entrega: 17 de enero de 2016



- Lenguaje y entorno de programación para análisis estadístico y gráfico
- Software libre (GPL)
- Multiplataforma
- www.r-project.org



“Piled Higher and Deeper”, Jorge Cham, www.phdcomics.com.

Replicable / reproducible

Replicability ←————→ Reproducibility

Reproduction of the original results using the same tools

by the original author on the same machine

by someone in the same lab/using a different machine

by someone in a different lab

Reproduction using different software, but with access to the original code

Completely independent reproduction based only on text description, without access to the original code

Davison, 2014

http://rrcns.readthedocs.org/en/latest/reproducible_research.html



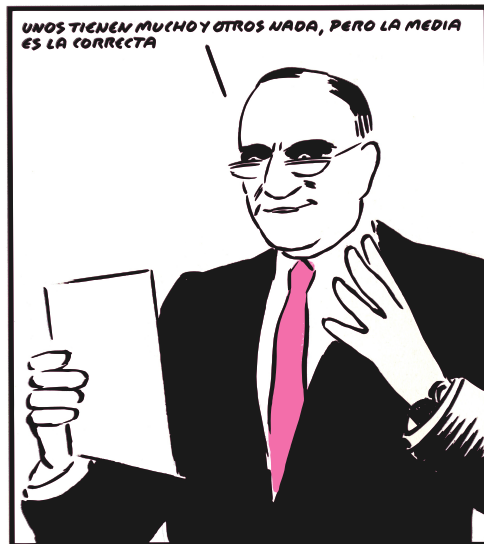
- Entorno de desarrollo para R
- Software libre (AGPL)
- Multiplataforma
- www.rstudio.com/

- 1 Presentación de la materia
- 2 Estadística descriptiva
- 3 Estadística inferencial: error de muestreo, contraste hipótesis

Salario bruto medio en España, 2010

22 790 euros/año

Instituto Nacional de Estadística. Encuesta de estructura salarial 2010 (resultados provisionales), <http://www.ine.es/prensa/np720.pdf>



“Unos tienen mucho y otros nada, pero la media es la correcta”

El Roto, *Viñetas para una crisis*, Mondadori, 2011.

Estadísticos descriptivos: medidas de posición

- Media aritmética (*arithmetic mean*)
- Mediana (*median*)
- Moda (*mode*)

Estadísticos descriptivos: medidas de posición

- Media aritmética (*arithmetic mean*)
- Mediana (*median*)
- Moda (*mode*)

Salarios brutos, 2010

Salario medio	22 790 euros
Salario mediano	19 017 euros
Salario modal	16 489 euros

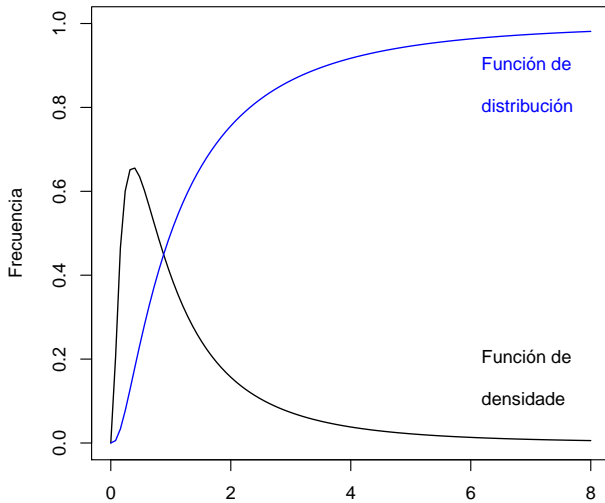
Instituto Nacional de Estadística. Encuesta de estructura salarial 2010 (resultados provisionales). URL: <http://www.ine.es/prensa/np720.pdf>

Estadísticos descriptivos: medidas de dispersión

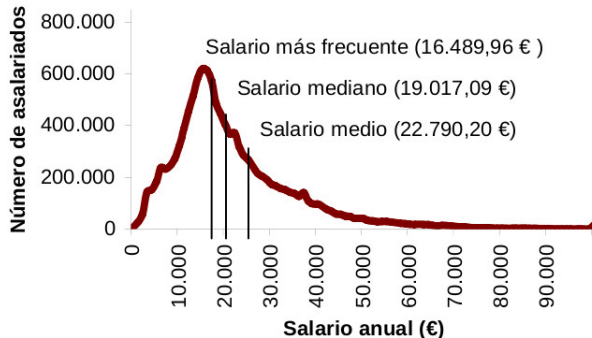
Desv. típica (<i>std. deviation</i>)	$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{n}}$
Varianza (<i>variance</i>)	σ^2
Coeficiente de variación	$cv = \frac{\sigma}{\mu}$
Desv. absoluta mediana	$mad = M(X - M(X))$
Recorrido (<i>range</i>)	$max(X) - min(X)$

Representación gráfica

Función de densidad e distribución

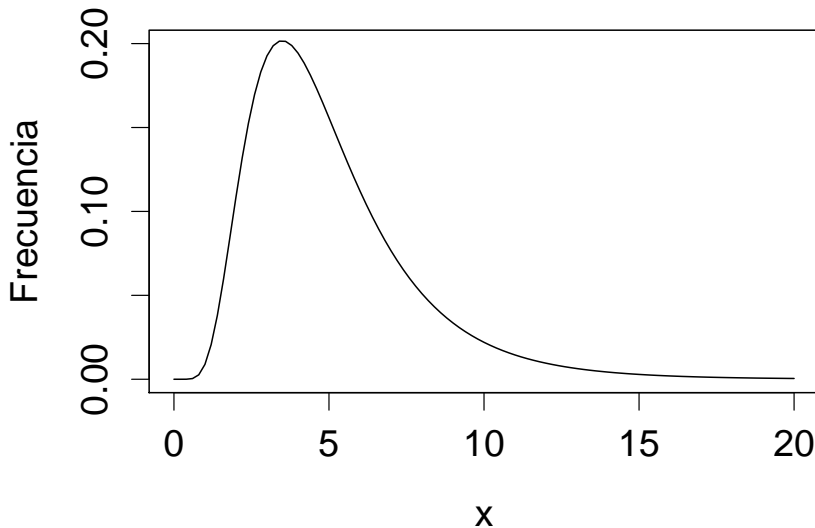


Distribución del salario bruto anual

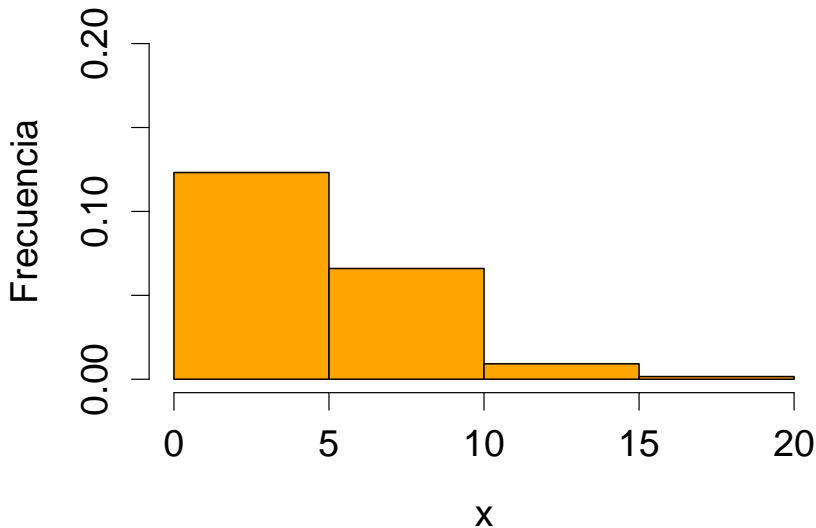


Instituto Nacional de Estadística. Encuesta de estructura salarial 2010 (resultados provisionales), <http://www.ine.es/prensa/np720.pdf>

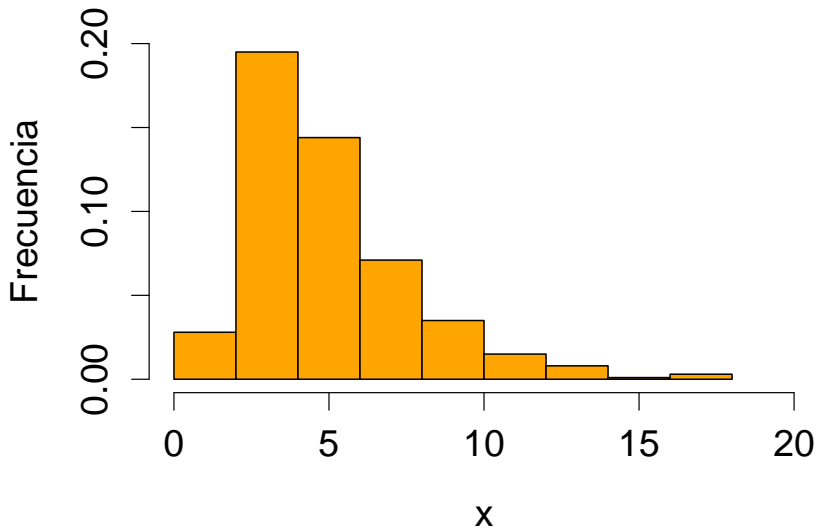
Histograma



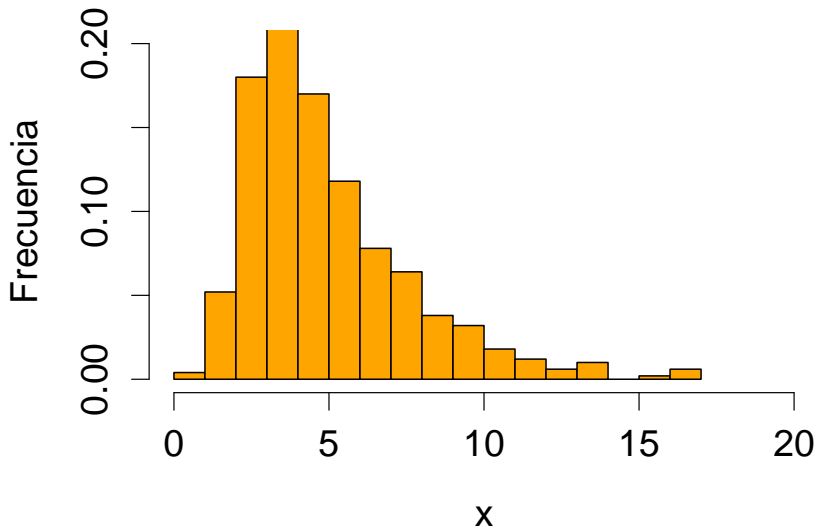
Histograma



Histograma



Histograma



Histograma

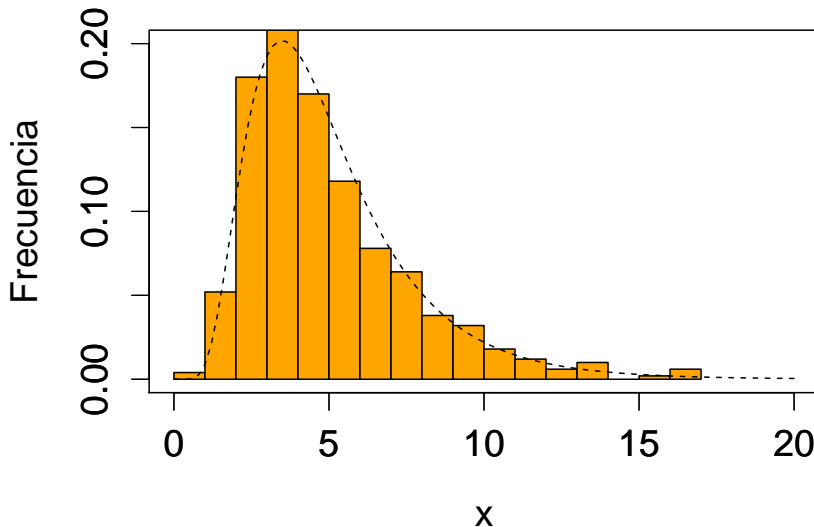


Diagrama de barras

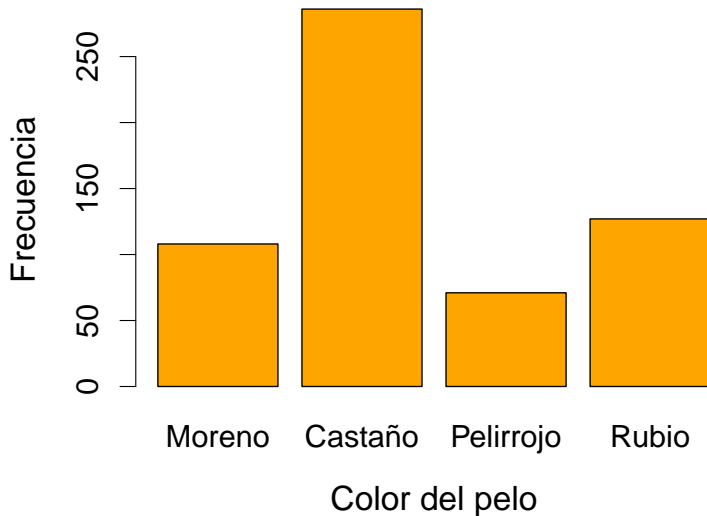


Diagrama de caja

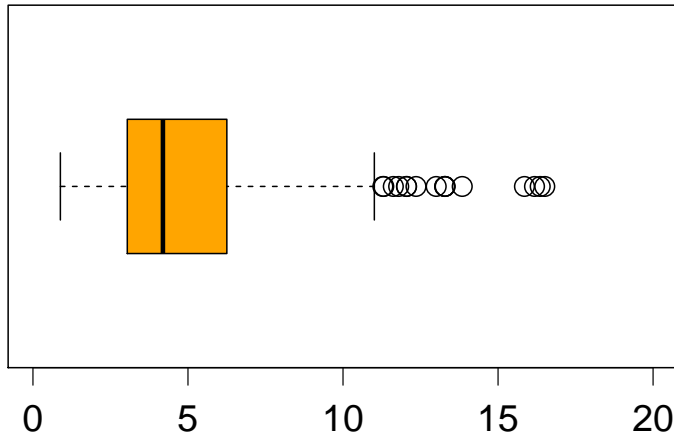
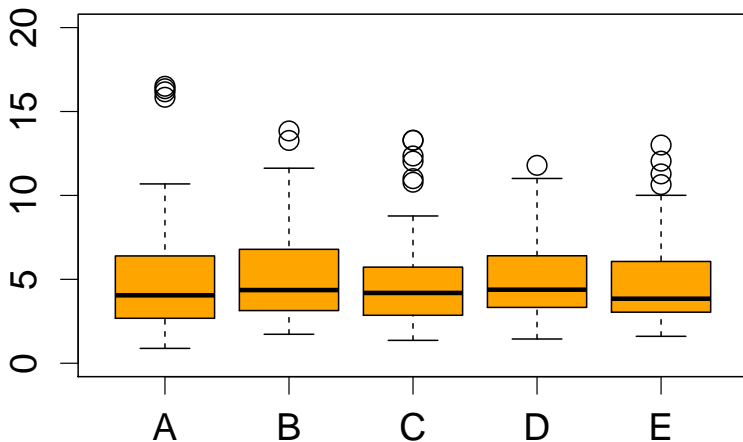


Diagrama de caja



Práctica 1

- Repaso de estadística descriptiva
- Introducción a R y RStudio



1 Presentación de la materia

2 Estadística descriptiva

3 Estadística inferencial: error de muestreo, contraste hipótesis



Estadística inferencial

Inferir

(Del lat. inferre, llevar a).

1. tr. Sacar una consecuencia o deducir algo de otra cosa.

Censo y muestra



Foto: US Census Bureau. *A census taker at work*, ca. 1940.

Utilidad de una muestra

¿Hasta qué punto...

- la muestra **representa** a la población?
- podemos **confiar** en lo que inferimos a partir de ella?

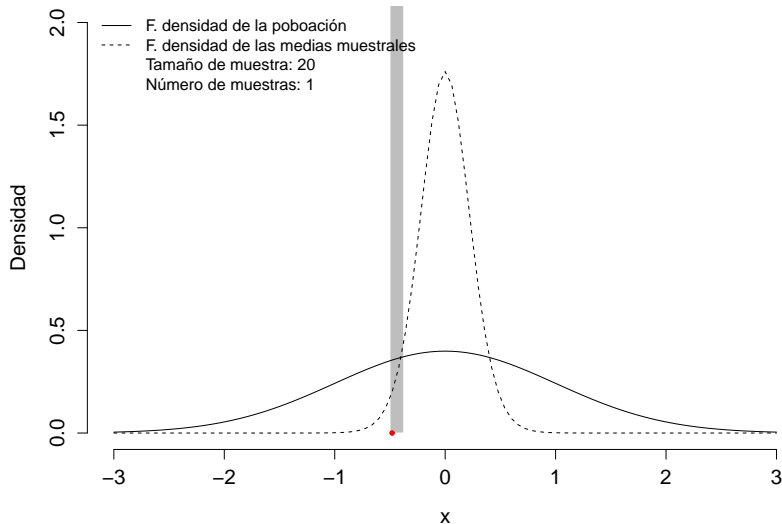
Representatividad

Muestra representativa

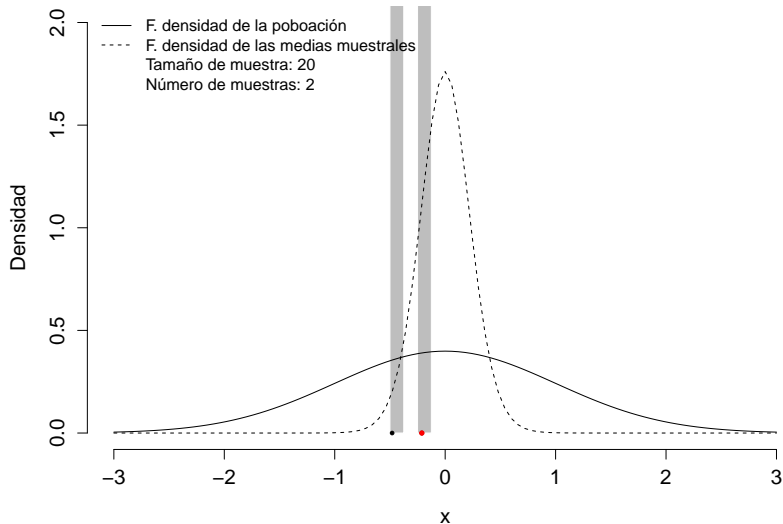
Una muestra es **representativa** de la población si todos los individuos tuvieron la misma probabilidad de haber sido escogidos

Muestra aleatoria simple (MAS) y variantes

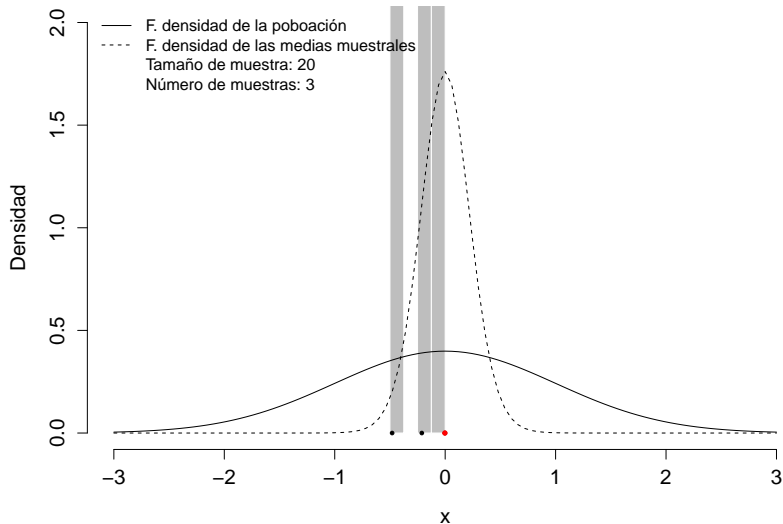
Error de muestreo



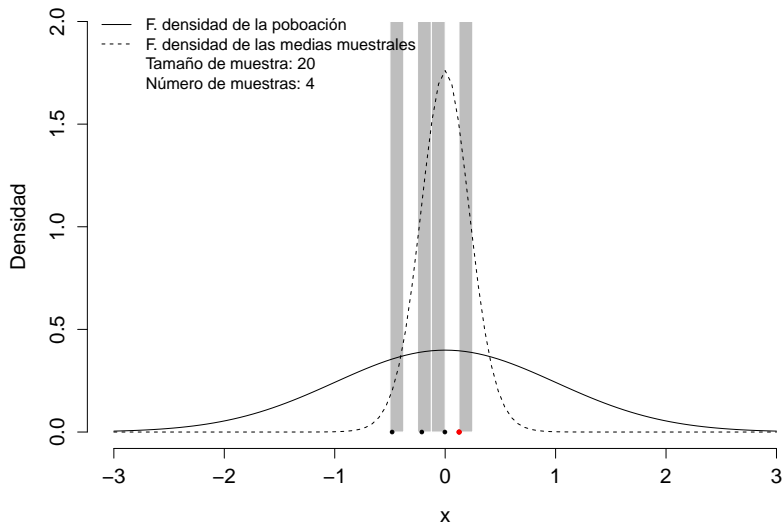
Error de muestreo



Error de muestreo

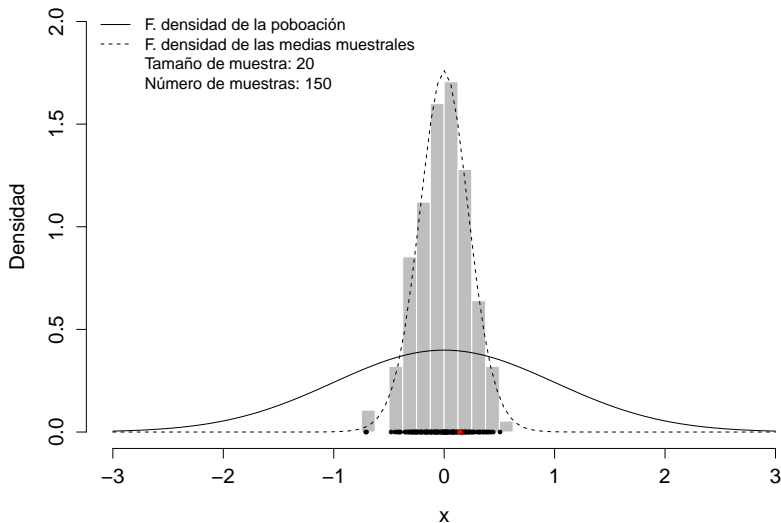


Error de muestreo



Error de muestreo

Error de muestreo



Error de muestreo

Desviación típica del estimador (p.ej. media muestral).

Su valor depende...

- De la heterogeneidad (dispersión) **de la población**
- Del tamaño **de la muestra**

Estimación por intervalo

Estimación de la media de la población (μ)

Media muestral $\bar{x} = \frac{\sum x_i}{n}$

Estimación por intervalo

Estimación de la media de la población (μ)

Media muestral $\bar{x} = \frac{\sum x_i}{n}$

Error de muestreo $S_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

Estimación por intervalo

Estimación de la media de la población (μ)

Media muestral $\bar{x} = \frac{\sum x_i}{n}$

Error de muestreo $S_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

Distribución de \bar{x}

- Normal, si
 - $X \sim N(\mu, \sigma)$; σ conocida
 - σ conocida; $n > 30$
 - $n > 100$
- t de Student con $n - 1$ grados de libertad

Estimación por intervalo

Estimación de la proporción de la población (P)

Proporción muestral p

Estimación por intervalo

Estimación de la proporción de la población (P)

Proporción muestral p

Error de muestreo $S_p \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Estimación por intervalo

Estimación de la proporción de la población (P)

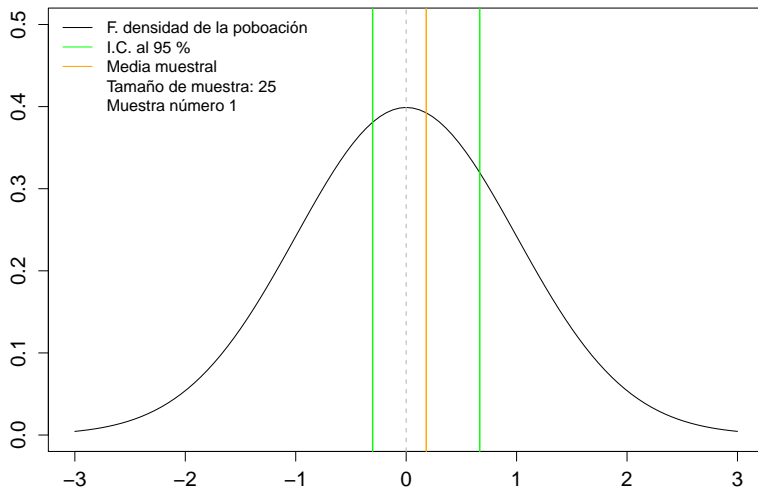
Proporción muestral p

Error de muestreo $S_p \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}$

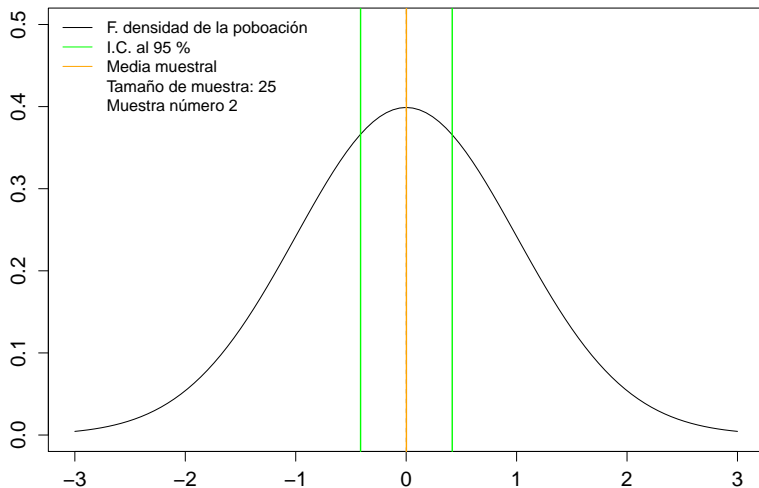
Distribución de \hat{p}

- Normal, si
 - $n \geq 30$, $n\hat{p} > 5$, y $n\hat{q} > 5$

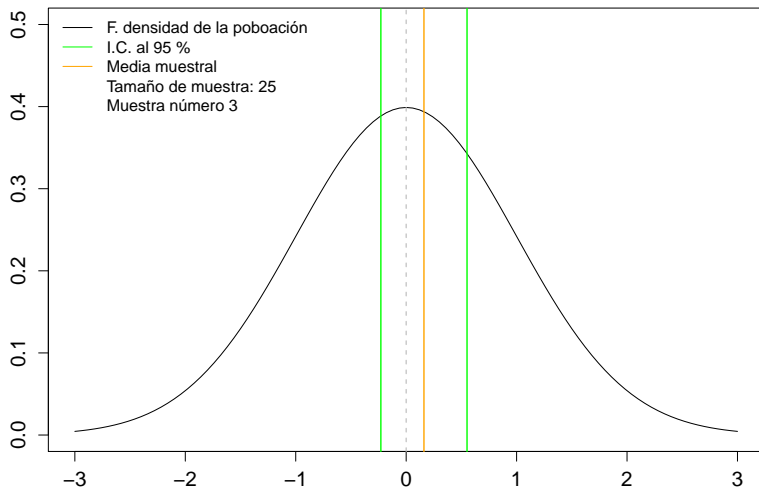
Estimación por intervalo



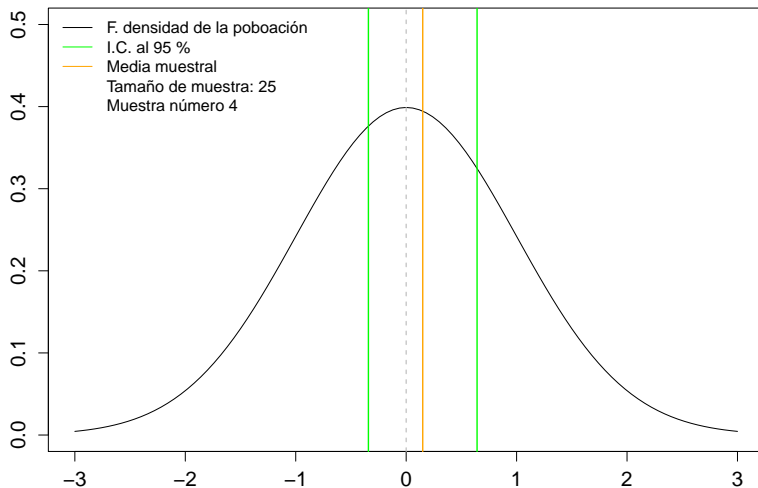
Estimación por intervalo



Estimación por intervalo

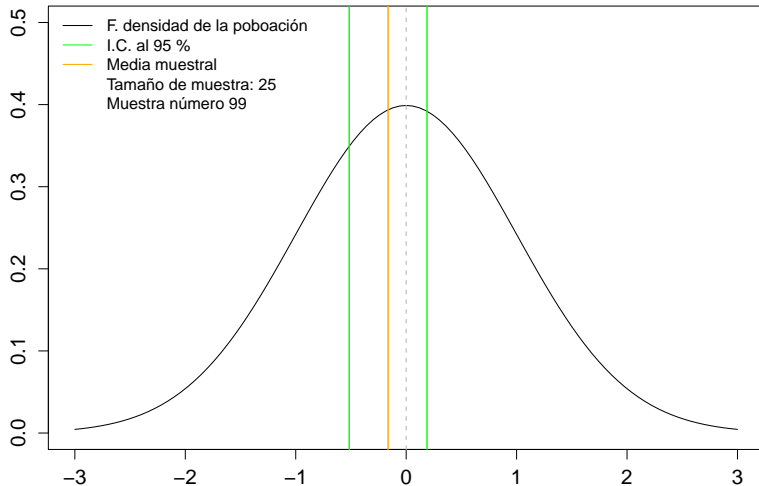


Estimación por intervalo



Estimación por intervalo

Estimación por intervalo



Contraste de hipótesis

Ejemplo: sacos de café

Peso medio según envasado: 50 kg

Contraste de hipótesis

Ejemplo: sacos de café

Peso medio según envasado: 50 kg

10 sacos al azar

48.8 47.7 46.8 47.2 51.6 49.0 50.4 50.9 48.0 48.7

media muestral: 48.9 kg

Presentación de la materia
oooooooo

Estadística descriptiva
oooooooooooo

Estadística inferencial: error de muestreo, contraste hipótesis
oooooooooooooooo●oooooooooooo



Contraste de hipótesis

Dos hipótesis en estudio

H_0 Hipótesis nula

H_1 Hipótesis alternativa

Contraste de hipótesis

Dos hipótesis en estudio

H_0 Hipótesis nula

H_1 Hipótesis alternativa

$$H_0 \quad \bar{x} = 50 \text{ kg}$$

$$H_1 \quad \bar{x} \neq 50 \text{ kg}$$

¿Existe evidencia suficiente para descartar H_0 ...
y aceptar, por lo tanto, H_1 ?

Contraste de hipótesis

«Si fuera cierta la hipótesis nula H_0 , la muestra, o mejor T , debería de comportarse de una determinada manera [...]. Si extraída una muestra al azar, acontece un suceso para T que tenía poca probabilidad de ocurrir si fuera cierta H_0 —es decir, bajo H_0 — o bien es que hemos tenido tan mala suerte de haber elegido una muestra “muy rara”, o, lo que es más probable, la hipótesis nula era falsa».

García Pérez, A (2003), *Estadística aplicada: Conceptos básicos*,
Universidad Nacional de Educación a Distancia, Madrid.

Errores posibles

Error de tipo I

Rechazar H_0 siendo, en realidad, cierta

Error de tipo II

No rechazar H_0 siendo falsa

Errores posibles

Error de tipo I

Rechazar H_0 siendo, en realidad, cierta

Error de tipo II

No rechazar H_0 siendo falsa

Elección	H_0 es en realidad...	
	Cierta	Falsa
Rechazar	Error Tipo I	Éxito
No rechazar	Éxito	Error Tipo II

Control sobre los errores

Podemos ejercer cierto control sobre los errores...

De tipo I: al escoger el nivel de significación (p-valor, α)

$$P(\text{Rechazar } H_0 | H_0 \text{ cierta}) = P(\text{Error de tipo 1}) = \alpha$$

De tipo II: mediante la *potencia* del test ($1 - \beta$)

$$P(\text{Rechazar } H_0 | H_1 \text{ cierta}) = 1 - P(\text{Error de tipo 2}) = 1 - \beta$$

Crterios de decisión

Simples y más habituales

$\alpha < 0'05$ Rechazar H_0

$\alpha > 0'05$ Aceptar H_0

Crterios de decisión

Simples y más habituales

$\alpha < 0'05$ Rechazar H_0

$\alpha > 0'05$ Aceptar H_0

Alternativa más conservadora

$\alpha < 0'01$ Rechazar H_0

$\alpha > 0'20$ Aceptar H_0

Resto de casos Aumentar la potencia (aumentar n)

Volviendo al café

Contraste bilateral

$$H_0 \quad \bar{x} = 50 \text{ kg}$$

$$H_1 \quad \bar{x} \neq 50 \text{ kg}$$

Volviendo al café

Contraste bilateral

$$H_0 \quad \bar{x} = 50 \text{ kg}$$

$$H_1 \quad \bar{x} \neq 50 \text{ kg}$$

```
> t.test(x, mu=50, alternative="two.sided")
```

One Sample t-test

data: x

t = -2.1492, df = 9, p-value = 0.06011

alternative hypothesis: true mean is not equal to 50

95 percent confidence interval:

47.76272 50.05728

sample estimates:

mean of x

48.91

Volviendo al café

Contraste unilateral

$$H_0 \quad \bar{x} \geq 50 \text{ kg}$$

$$H_1 \quad \bar{x} < 50 \text{ kg}$$

Volviendo al café

Contraste unilateral

$$H_0 \bar{x} \geq 50 \text{ kg}$$

$$H_1 \bar{x} < 50 \text{ kg}$$

```
> t.test(x, mu=50, alternative="less")
```

One Sample t-test

data: x

t = -2.1492, df = 9, p-value = 0.03006

alternative hypothesis: true mean is less than 50

95 percent confidence interval:

-Inf 49.83968

sample estimates:

mean of x

48.91

Práctica 2

Contrastes de hipótesis con R y RStudio

Ejercicio de evaluación 1

Errores de muestreo e intervalos de confianza

Fecha de entrega recomendada: 25 de octubre.