



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Herramientas estadísticas para Big Data

Introducción a la Inferencia
Estadística,
Muestreo y Preproceso de
datos

Máster **Big Data** Analytics

Departamento de
Estadística e Investigación
Operativa Aplicadas y
Calidad

Valencia, Octubre 2017

Elena Vázquez

Contenidos

1. Conceptos básicos
2. Probabilidad
3. Variables aleatorias y distribuciones
4. Inferencia en muestras grandes
5. Técnicas de muestreo
6. Preprocesamiento de datos

Glosario

Enlaces de interés

Bibliografía



4 Inferencia en muestras grandes

1. Introducción
 2. Muestreo y distribuciones en el muestreo
 3. Estimación puntual
 4. Intervalos de confianza
 5. Test o contrastes de hipótesis
- Resumen y consideraciones

4 Inferencia en muestras grandes

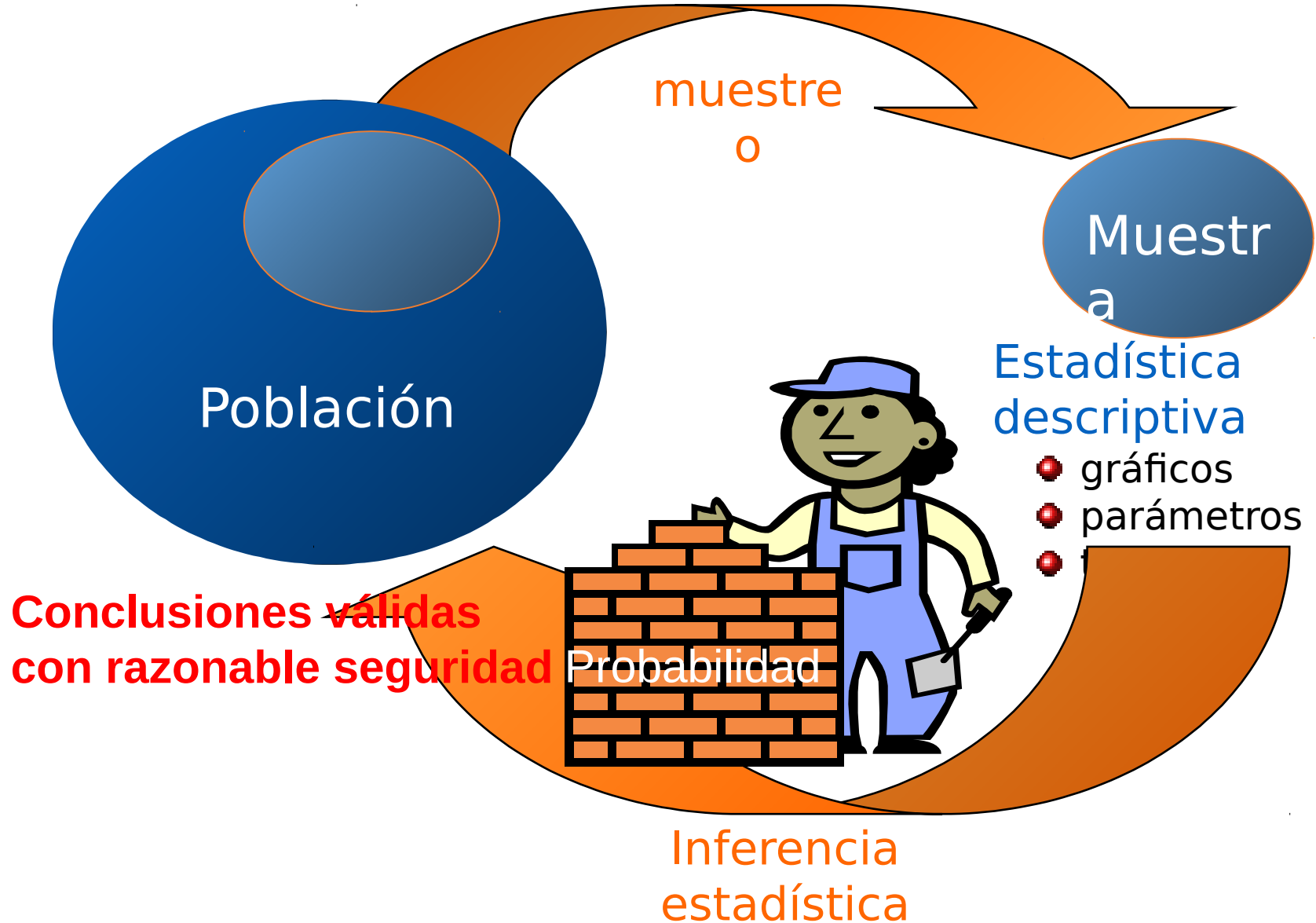
- No vamos a ver todos los tipos de contrastes e intervalos de confianza con el tiempo disponible.
- El objetivo es que conozcáis, entendáis los conceptos básicos de inferencia que permitan aplicar la lógica de los contrastes e intervalos de confianza y la terminología asociada para que podáis seguir aprendiendo por vosotros mismos.

Introducción

- Los analistas necesitamos responder a preguntas que se plantean sobre fenómenos del mundo real.
- Para ello:
 - Planteamos **hipótesis**
 - Recogemos **datos** el mundo real
 - **Verificamos** las hipótesis
- Verificar dichas hipótesis implica construir **modelos estadísticos** del fenómeno estudiado

La **inferencia estadística** permite extrapolar las conclusiones obtenidas con los datos observados sobre los fenómenos al mundo real mediante modelos

Introducción



Introducción

- Ya hemos visto algunos “modelos” como son la **función de densidad** o la **función de probabilidad** que permiten conocer lo probables o improbables que son cada uno de los **valores** que puede tomar una variable
 - Normal
 - Uniforme
 - Binomial
 - Poisson
 - ...
- Además, necesitamos **otros modelos** para conocer lo probables o improbables que son los **parámetros** (media, varianza, ...) que caracterizan una variable

Introducción

Muestra: Lo que tenemos

Frecuencia relativa (%)

% de veces que sale un 5 al lanzar el dado

Parámetros muestrales

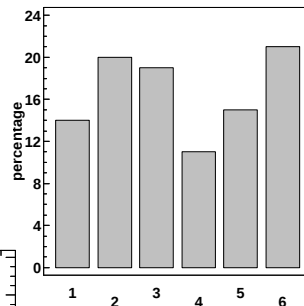
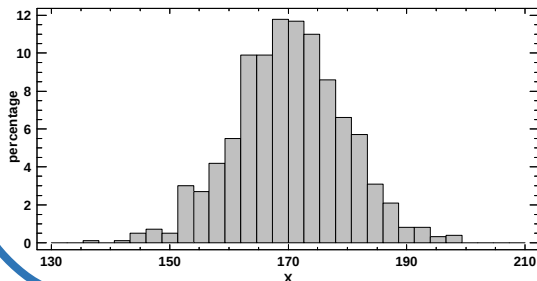
\bar{X}

S^2

S

Cuartiles, ...

Distribución frecuencias



Población: Lo teórico, lo ideal

Probabilidad (tanto por uno)

$$P(A) = P(X=5)$$

Parámetros poblacionales

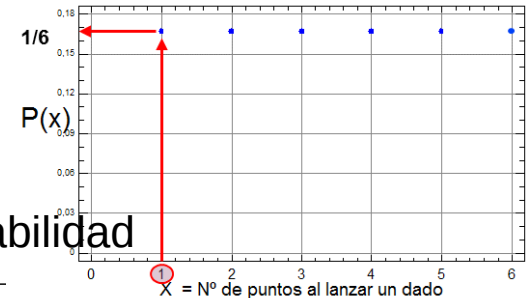
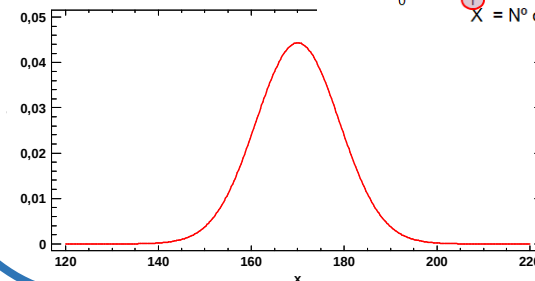
μ o $E(X)$

σ^2

σ

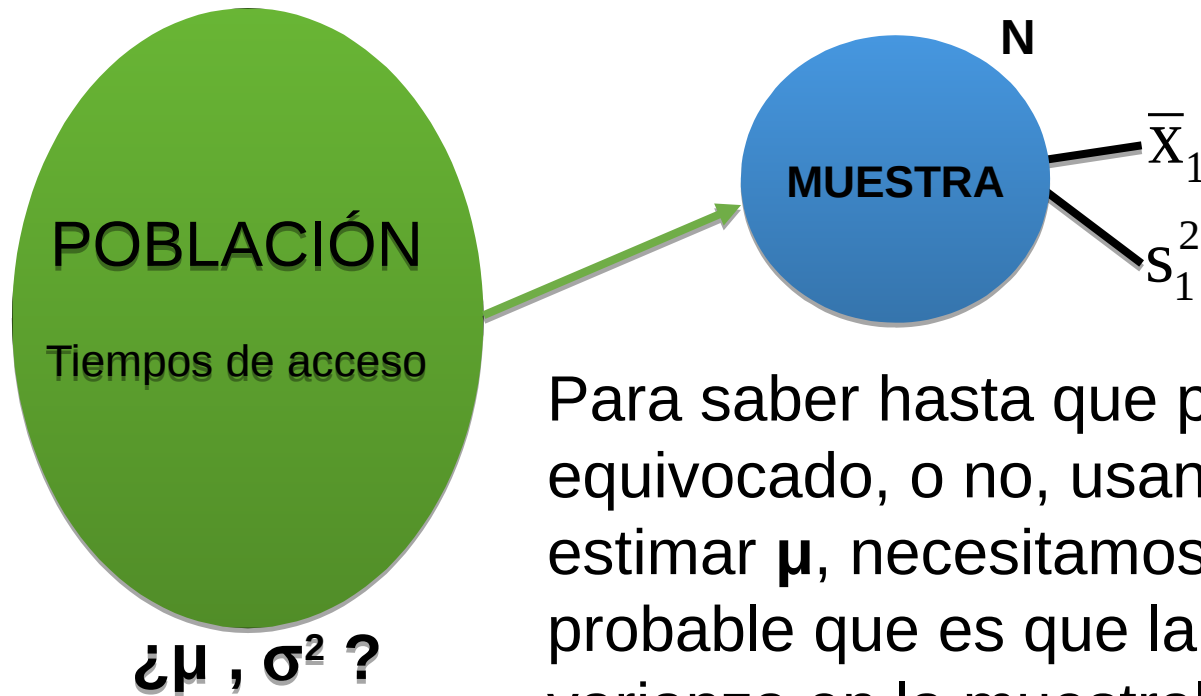
Cuartiles, ...

Distribución probabilidad



Muestreo y distribuciones en el muestreo

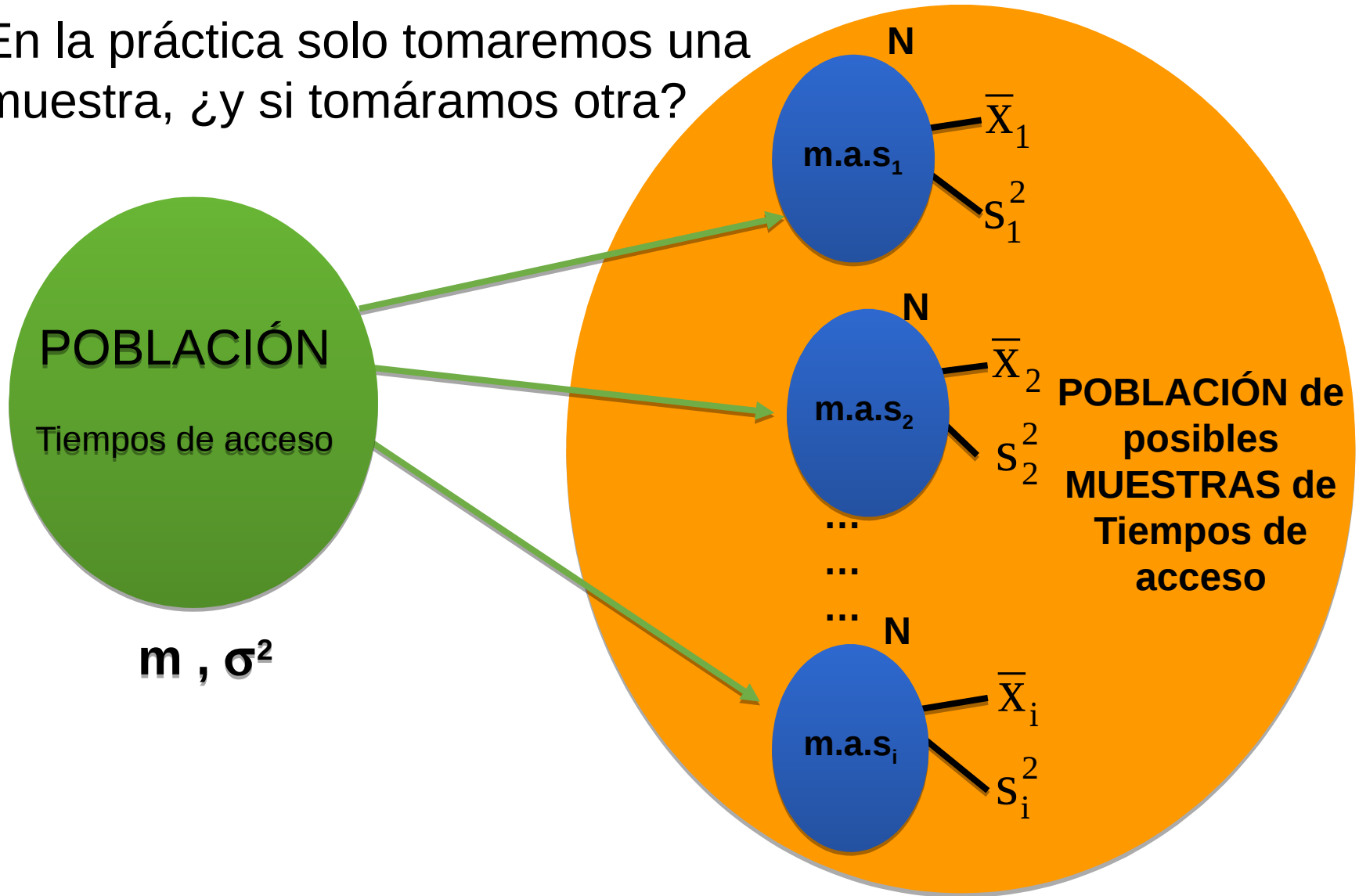
Para estimar la media y la varianza de la población, por ejemplo, obtenemos una muestra y calculamos los respectivos parámetros muestrales.



Para saber hasta que punto nos hemos equivocado, o no, usando \bar{X} para estimar μ , necesitamos conocer lo probable que es que la media o la varianza en la muestral hayan sido las que son si suponemos que la media en la población es μ

Variabilidad muestral

En la práctica solo tomaremos una muestra, ¿y si tomáramos otra?

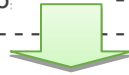


Los parámetros muestrales son variables aleatorias

v.a $X = \{\text{Peso bolsas de naranjas (gr)}\}$ **Población**

Muestras

	PES01	PES02	PES03	PES04	PES05	PES06
Count	15	15	15	15	15	15
Average	1993,6	2002,94	2000,01	1993,29	2001,57	2001,8
Median	1992,0	2002,41	1996,78	2001,74	2008,43	1999,07
Variance	391,971	665,367	353,233	449,589	366,657	267,42
Standard deviation	19,7983	25,7947	18,7945	21,2035	19,1483	16,353
Lower quartile	1980,0	1984,67	1986,51	1972,14	1987,88	1989,29
Upper quartile	2013,0	2016,16	2014,87	2009,29	2015,57	2010,83
Std. skewness	-0,405564	0,143095	0,0315826	-0,063688	-1,34455	0,909537
Std. kurtosis	-0,593681	-0,162126	-0,734175	-1,27034	-0,34034	0,229404



\bar{X} es la MEDIA MUESTRAL y es una VARIABLE ALEATORIA:

$\bar{X} = \{1993,6, 2002,94, 2000,01, 1993,29, 2001,57, 2001,8\}$

S^2 es la VARIANZA MUESTRAL y es una VAR. ALEATORIA:

$S^2 = \{391,971, 665,367, 353,233, 449,589, 366,657, 267,42\}$

Idem para el resto de parámetros muestrales (simetría, r, parámetros de un modelo de regresión, etc)

Distribución de los parámetros muestrales

•Cualquier **parámetro muestral** (\bar{X}, S^2, r, b_i , etc) es una **variable aleatoria**:

- Tendrá sus **parámetros** (posición, dispersión,...)
- Seguirá una distribución (**distribución muestral**), que depende de:
 - Distribución población original
 - Tamaño de la muestra (N)

Por razones de tiempo nos centraremos en la media y varianza, pero lo fundamental será los mismo para otros parámetros y el razonamiento análogo.

Estadístico

Un **estadístico** es cualquier función de los parámetros muestrales. Un parámetro muestral es también un estadístico:

$$\bar{X}, \quad S^2, \quad 3\bar{X}, \quad \bar{X} - 4, \quad \frac{S}{10}, \quad \frac{\bar{X} - 1993'6}{19'8 / \sqrt{15}} \quad \dots$$

- Establece las **relaciones que ligan los parámetros muestrales y poblacionales**, proporcionando las evidencias respecto de la población contenida en la muestra que necesitamos en el proceso de inferencia.
- **Algunos** estadísticos importantes **reciben el nombre de la distribución de probabilidad que los caracteriza** (t, F, χ^2 , etc)

Distribución de la media muestral

La **media de la media muestral** es la media poblacional:

$$m_{\bar{X}} = m$$

La **varianza de la media muestral** es la varianza de la población dividida por el tamaño N de la muestra:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N} \quad \sigma_{\bar{X}}^2 \xrightarrow{N \rightarrow \infty} 0$$

\bar{X} es suma de v.a. independientes con la misma distribución, por lo que la **distribución muestral de la media** es:

$$\bar{X} \underset[N \rightarrow \infty]{(TCL)} \approx N(m_{\bar{X}} = m, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{N})$$

Error típico o estándar $ET = \sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$

TCL es Teorema Central del Límite

Distribución de la varianza muestral

La media de la varianza muestral es la varianza poblacional:

$$m_{S^2} = \sigma^2$$

La varianza de la varianza muestral tiene una expresión compleja.

$$\sigma_{S^2}^2 \xrightarrow{N \rightarrow \infty} 0$$

La distribución muestral de la varianza es muy asimétrica pero, por el TCL y si N es muy grande:

$$S^2 \underset{\substack{N \rightarrow \infty \\ (TCL)}}{\approx} N(m_{S^2} = \sigma^2, \sigma_{S^2}^2 \approx 0)$$

TCL es Teorema Central del Límite

Muestreo en poblaciones normales

- Los resultados expuestos en los apartados anteriores son completamente generales → son válidos sea cual sea la distribución de la población muestreada (Binomial, Normal...)
- Cuando dicha población es Normal es posible establecer ciertos resultados adicionales de gran importancia dentro de la metodología de la Inferencia Estadística:
 - $\bar{X} \sim \text{Normal}$
 - Los estadísticos que aparecen siguen otras distribuciones características como:
 - χ^2
 - t de Student
 - F de Snedecor

Estas distribuciones no modelizan la pauta de variabilidad de ninguna variable real.

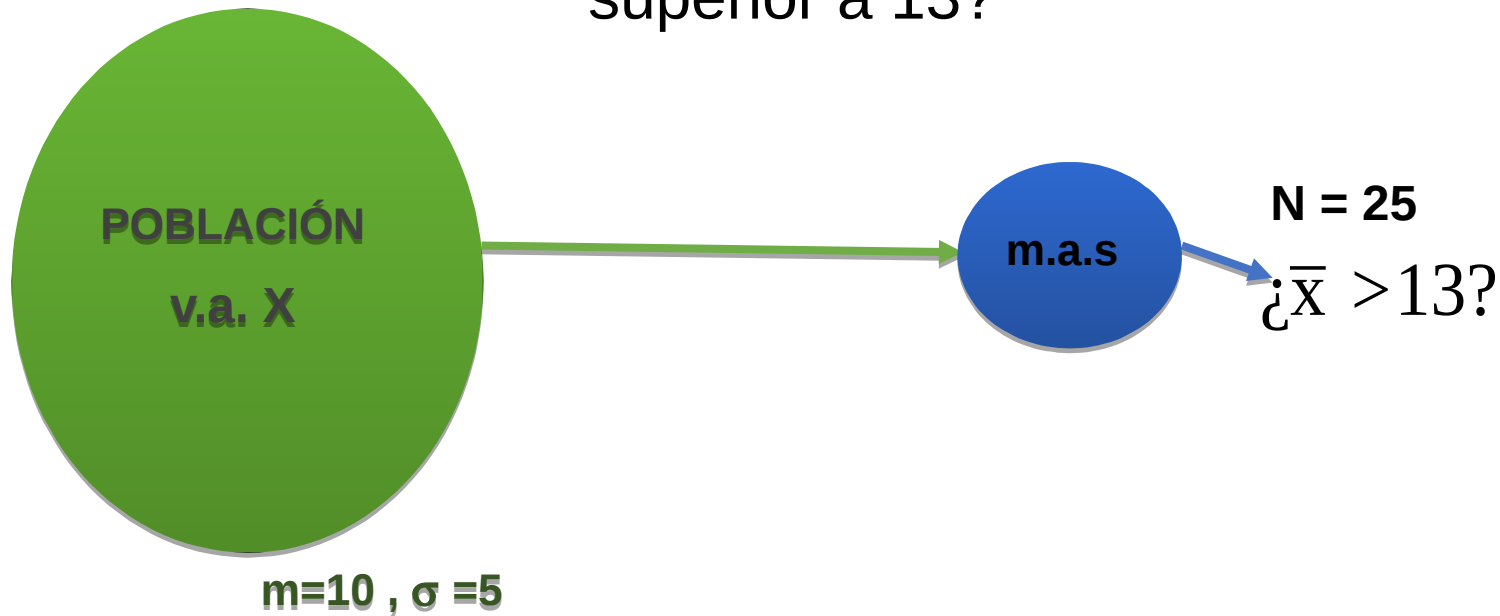
Son distribuciones que surgen en el proceso de inferencia estadística.

Media muestral de poblaciones normales



$$\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \approx \text{Normal}(0,1)$$

¿Es muy probable que una muestra, de tamaño 25, extraída de una población Normal con $m=10$ y $\sigma=5$ proporcione una media muestral superior a 13?

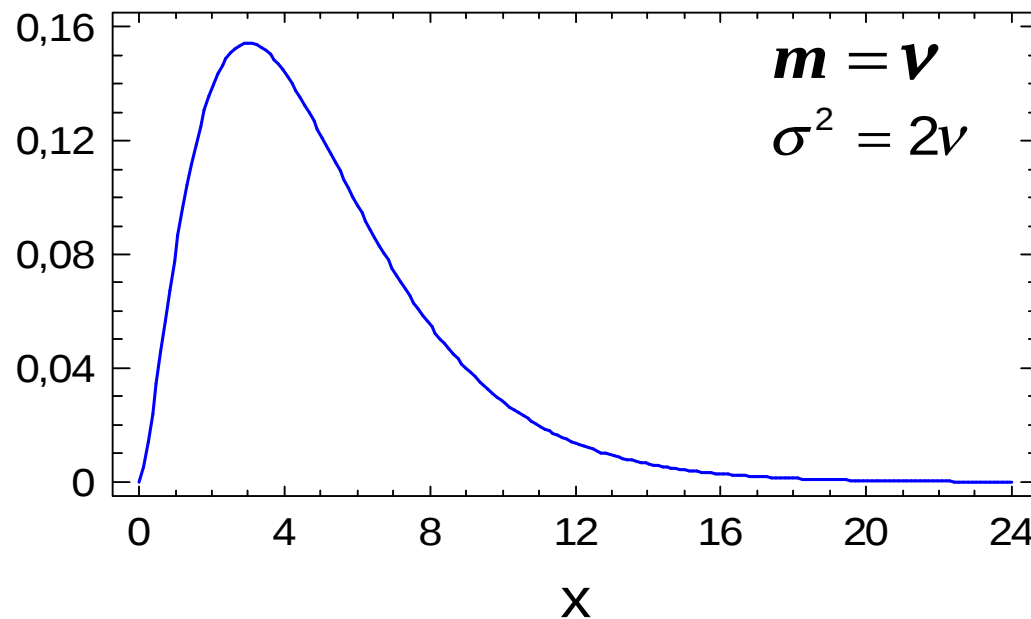
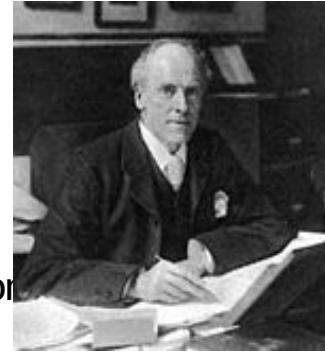


La distribución χ^2



- Importante en el estudio de la S^2 de una muestra de una población normal (*K. Pearson*)

- Se denota: χ^2_v **Grados de libertad (g.l.)**



Deg. of freedom
— 5

Sólo toma valores positivos

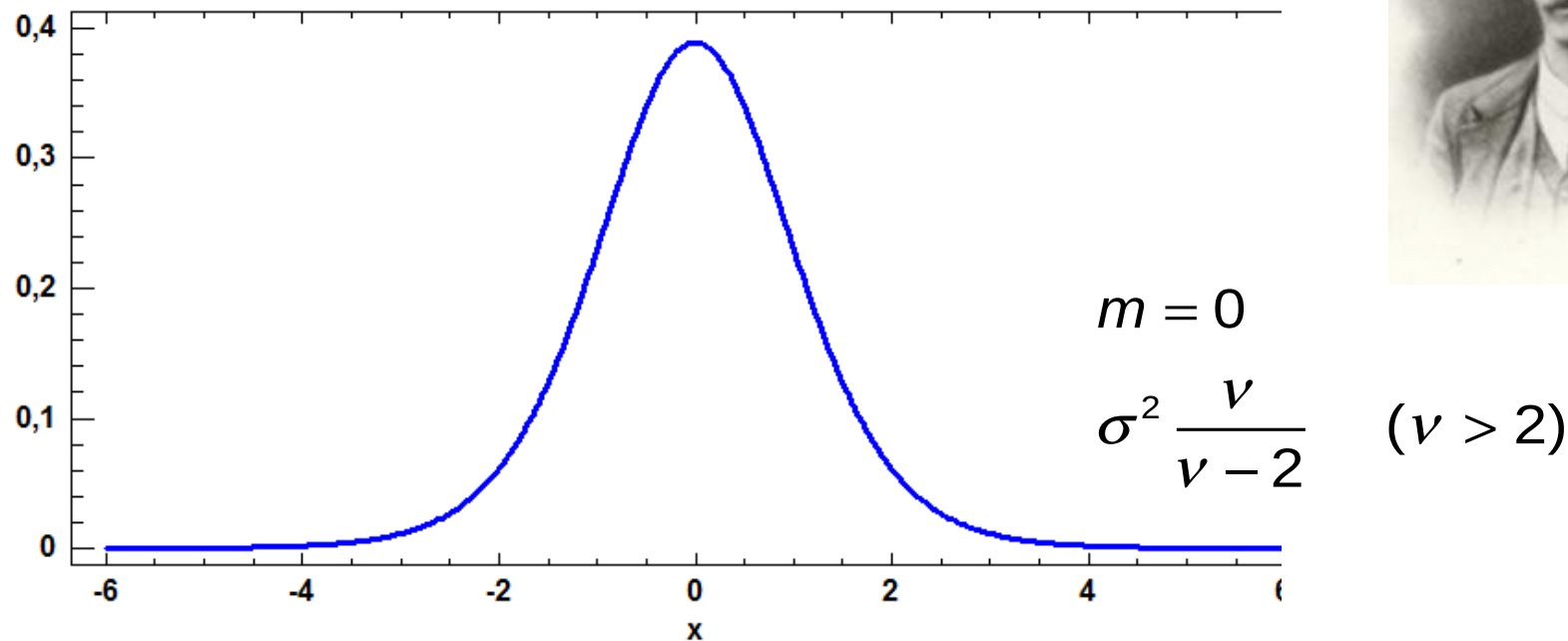
Si $X \sim N(m, \sigma^2)$ y S^2 es la varianza en una muestra de tamaño N

$$(N-1) \frac{S^2}{\sigma^2} \sim \chi^2_{N-1}$$

La distribución t de Student



- Importante en la inferencia respecto de la media de una población normal (*W. S. Gosset*)



La forma de la $f(x)$ se parece mucho a una $N(0,1)$

Si $X \sim N(m, \sigma^2)$ y \bar{X} y S^2 son la media y la varianza en una muestra de tamaño N

$$t = \frac{\bar{X} - m}{s/\sqrt{N}} \sim t_{N-1}$$

La distribución t de Student y la Normal(0,1)

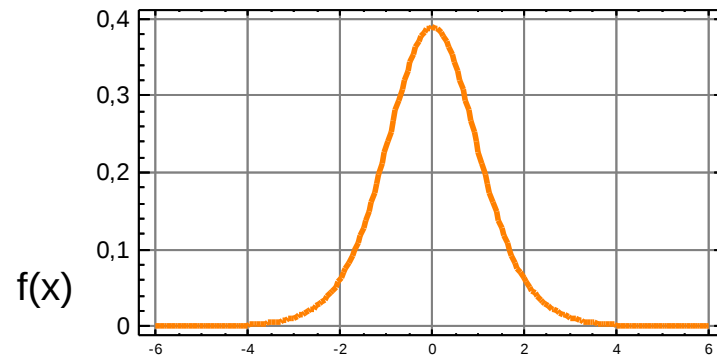
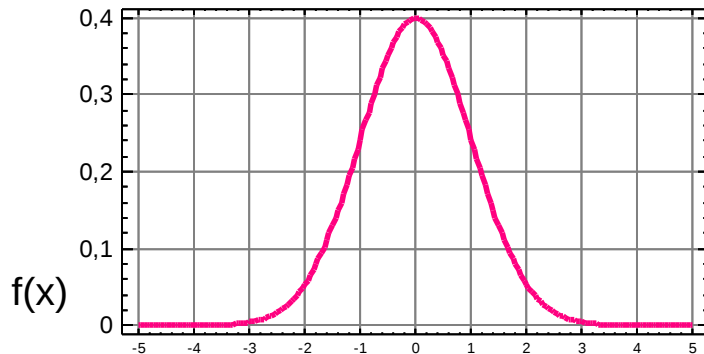


Apréciese la analogía entre:

$$\frac{\bar{X} - m}{\sigma / \sqrt{N}} \sim N(0,1)$$

$$\frac{\bar{X} - m}{s / \sqrt{N}} \sim t_{N-1}$$

Y



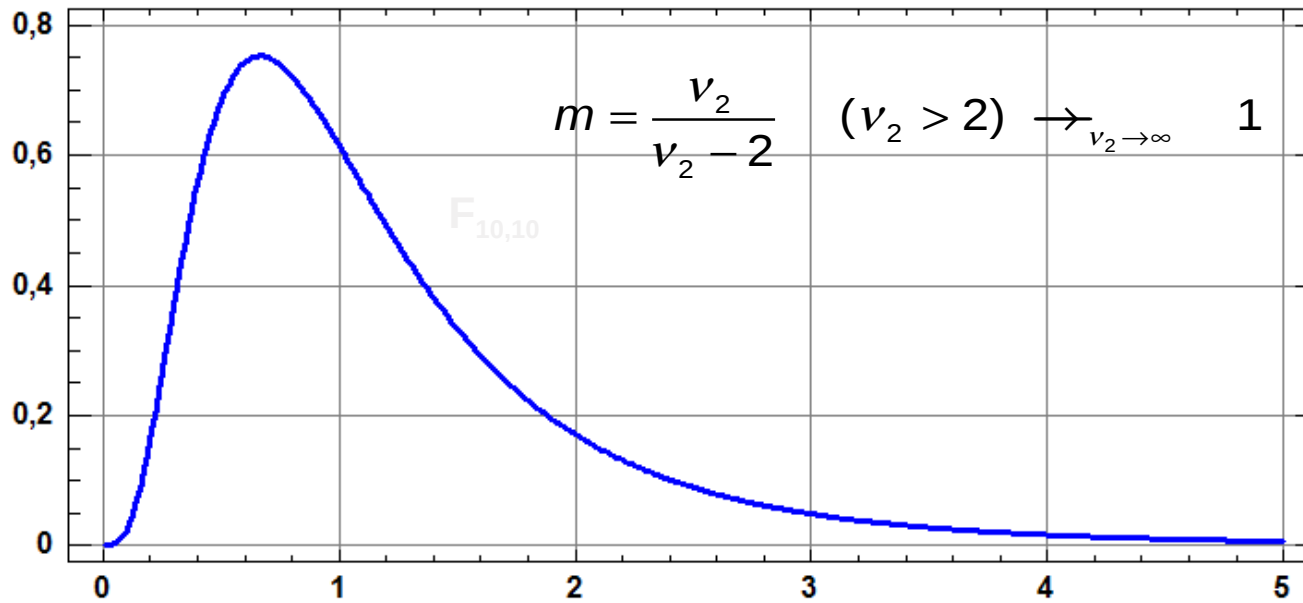
σ = desv. Típica de la población (teórica)

s = desv. Típica de la muestra (estimación de la desv. Típica de la población)

La distribución F de Snedecor



- En el estudio de los modelos de Regresión Lineal y de Análisis de la Varianza desempeña un papel fundamental la distribución F, denominada así por Snedecor.



Si $X_1 \sim N(m_1, \sigma_1^2)$, $X_2 \sim N(m_2, \sigma_2^2)$ son independientes y S_1^2 y S_2^2 son las varianzas muestrales de X_1 y X_2 (tamaños N_1 y N_2)

Si $\sigma_1^2 = \sigma_2^2 \Rightarrow \frac{S_1^2}{S_2^2} \sim F_{N_1-1, N_2-1}$

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{N_1-1, N_2-1}$$

Las distribuciones t, F y χ^2



Distribución t

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp
```

Distribución χ^2

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

Distribución F

```
df(x, df, ncp = 0, log = FALSE)
pf(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qf(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rf(n, df, ncp = 0)
```



Ejemplo gráfico de $f(t)$ t



```
> ##### GRÁFICO DE FUNCIÓN DE DENSIDAD #####  
>  
#####  
  
> ## Ejemplo: t CON 9 G.L.  
  
> x<-seq(-4,4,0.05)  
  
> y<-dt(x,9)  
  
> plot(x, y, type = "l", col="red", xlab="t 9 g.l.",  
ylab = "f(x)", main = "Función de densidad")  
  
> grid(col = "lightgray", lty = "dotted")
```

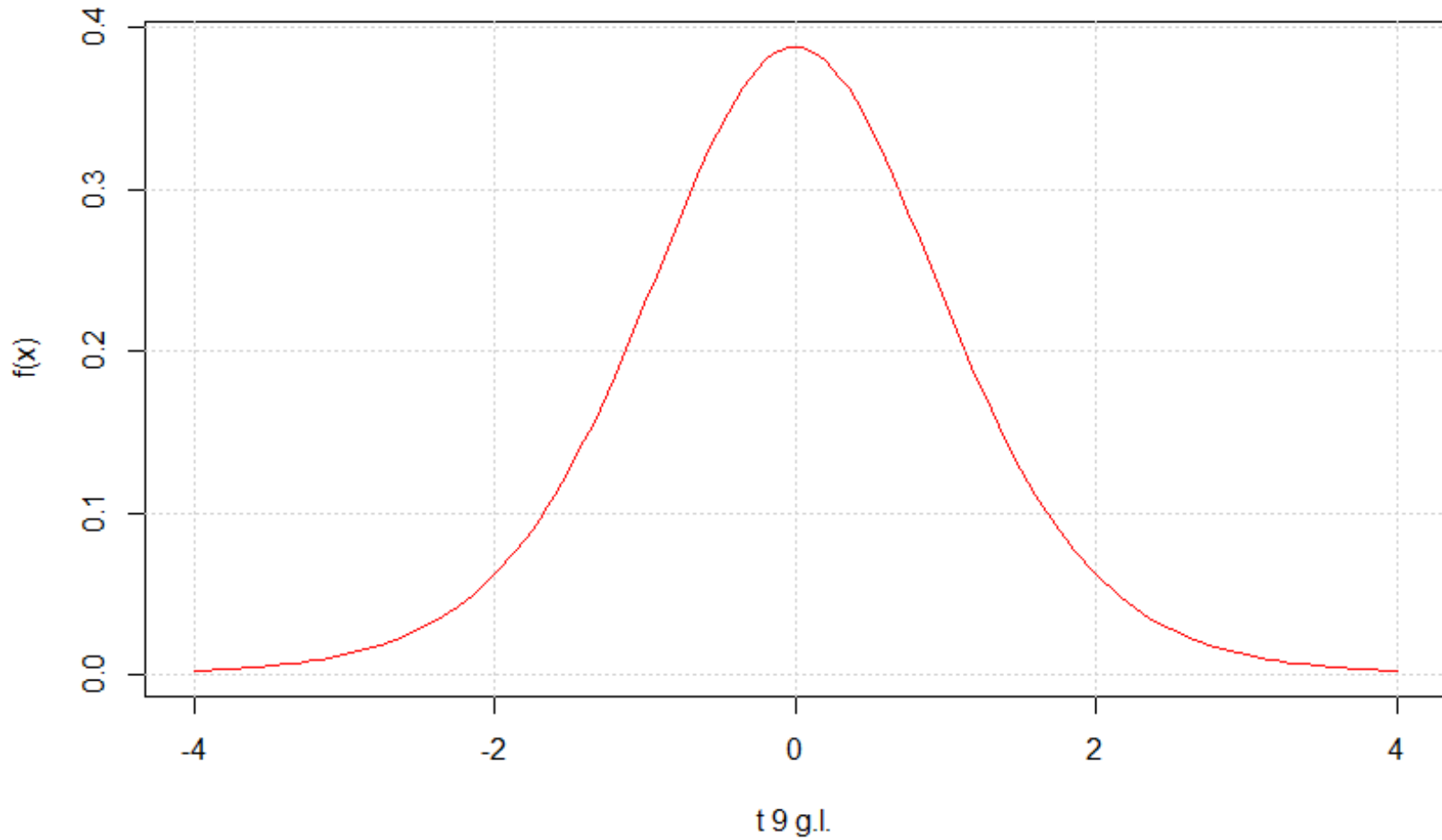
Repetir lo mismo con $x \leftarrow \text{seq}(-10, 10, 0.05)$



Ejemplo gráfico de $f(t)$ t



Función de densidad



Error estándar de la media ET o SE

- La desviación típica muestral (S) nos da una idea sobre el grado en el que la media (\bar{X}) representa adecuadamente a los datos observados.
- Análogamente, la desviación típica de la media muestral ($\sigma_{\bar{X}}$) nos da una idea sobre el grado en el que la media de la media muestral ($m_{\bar{X}}$) representa adecuadamente a la población.

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$ y un estimador de σ es S :

$$SE = ET = \hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{N}}$$

El **error estándar** nos proporciona una medida de lo probable que es que una muestra sea representativa de la población.

Error estándar de la media

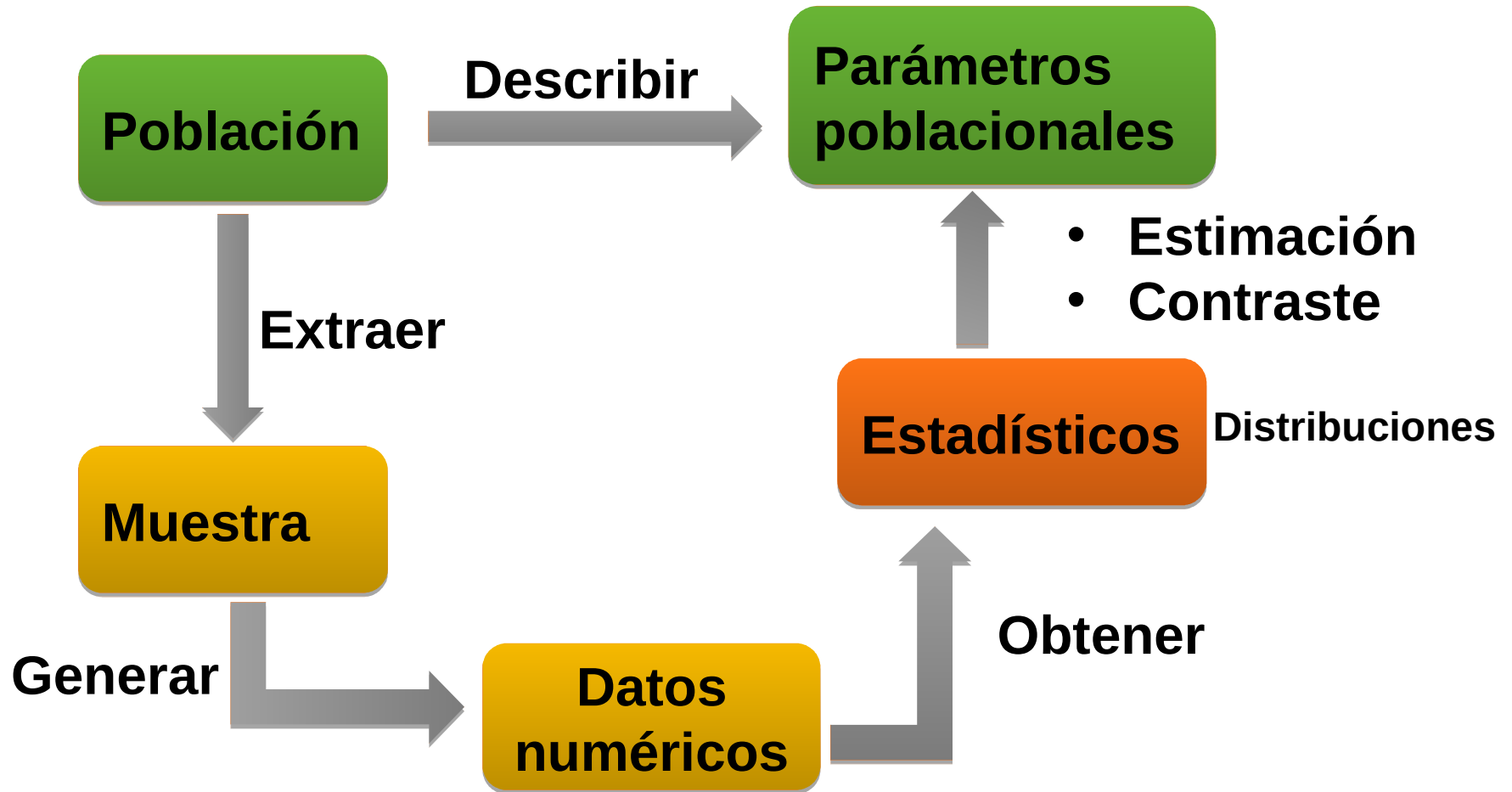
- **E.T. grande indica** gran variabilidad (con respecto a la media de la población) entre las medias muestrales, por lo que tomada una muestra (lo que haremos en realidad) podría ser poco probable que su media se pareciera a la de la población (**muestra poco representativa**)
- **E.T. pequeña indica** poca variabilidad (con respecto a la media de la población) entre las medias muestrales, por lo que tomada una muestra (lo que haremos en realidad) puede ser muy probable que su media se pareciera a la de la población (**muestra representativa**)

- . ¿Cómo de probable es que la muestra tenga la media o la desviación típica o r o b , etc (**parámetro muestral**) ... que tiene?
- . ¿Cómo de probable es la media poblacional (**parámetro poblacional**) tome un determinado valor?
- . ¿Cómo de probable es que la media de una muestra (u otro parámetro) sea diferente de la de otra...?

Necesitamos saber qué modelo de probabilidad (**distribución**) tiene la variable aleatoria estudiada y sus parámetros muestrales o su diferencia o su cociente, etc (**estadísticos**)

La **base Inferencia Estadística** reside en el conocimiento de

- Las relaciones entre los parámetros muestrales y poblacionales.
- La distribución de estas relaciones o estadísticos, que a su vez depende de la distribución de los parámetros muestrales.
- La distribución de la población de las variables estudiadas.



• Parámetros o características muestrales:

- Se utiliza el alfabeto latino
- S , r , \bar{X} , ...

Parámetros o características poblacionales:

- Se utiliza el alfabeto griego
- σ , ρ , μ , etc

Estimaciones de parámetros o características poblacionales :

- Se utiliza el símbolo “ \wedge ” encima de la letra correspondiente
- Estimación de σ es $\hat{\sigma}$

- Cuando el número de individuos de la muestra o **tamaño de la muestra** es muy grande ($N > 30$), por el T.C.L:

- La **media y varianza muestrales siguen siempre una distribución normal**:

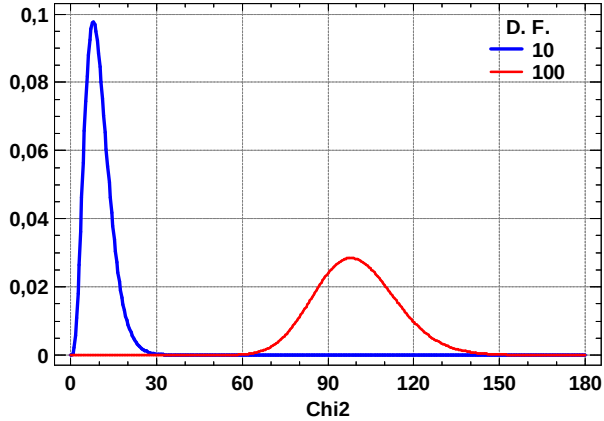
$$S^2_{T.C.L} \approx N(m_{S^2} = \sigma^2, \sigma_{S^2}^2 \approx 0) \quad \bar{X}_{T.C.L} \approx N(m_{\bar{X}} = m, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{N} \approx 0)$$

- Las distribuciones que aparecen en el muestreo de poblaciones normales χ^2 , t de Student y F de Snedecor también **tienden a distribuirse como normales**.
- Las **distribuciones** de muchas de las **variables estudiadas**, como la Binomial, por ejemplo, también se **aproximan a la distribución normal**.
- El **error estándar** tiende a ser pequeño (**no desaparece**).
- El número de **grados de libertad** es muy grande.

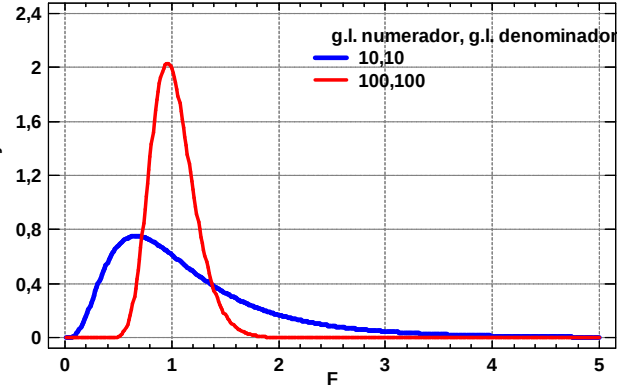
Muestreo en Big Data

Resumiendo

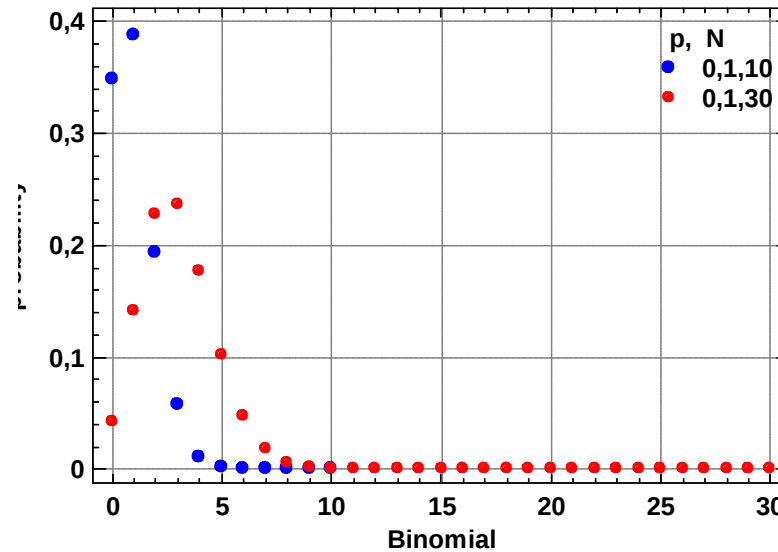
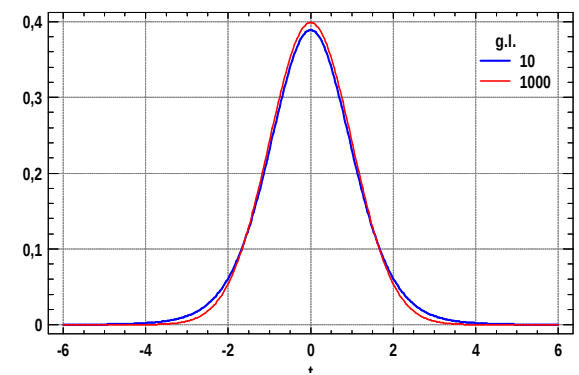
Chi-Square Distribution



F (variance ratio) Distribution



Student's t Distribution



Estimación de los parámetros

- Habitualmente NO conocemos el valor real de los **parámetros** en la **población** (β_j) ✉ Necesitamos estimarlos (b_j) a partir de algo conocido .
- En algunos casos podemos usar los **parámetros** de la **muestra: estimación puntual**

Parámetro poblacional	Descripción	Estimador															
μ o m	Media	<table> <tr> <th>Parámetro poblacional</th><th>Descripción</th><th>Estimador</th></tr> <tr> <td>μ o m</td><td>Media</td><td>\bar{X}</td></tr> <tr> <td>σ^2</td><td>Varianza</td><td>S^2</td></tr> <tr> <td>p</td><td>Proporción</td><td>p</td></tr> <tr> <td>ρ</td><td>Coefficiente de correlación</td><td>r</td></tr> </table>	Parámetro poblacional	Descripción	Estimador	μ o m	Media	\bar{X}	σ^2	Varianza	S^2	p	Proporción	p	ρ	Coefficiente de correlación	r
Parámetro poblacional	Descripción	Estimador															
μ o m	Media	\bar{X}															
σ^2	Varianza	S^2															
p	Proporción	p															
ρ	Coefficiente de correlación	r															
σ^2	Varianza	S^2															
P	Proporción	p															
ρ	Coefficiente de correlación	r															

Estimación de relaciones entre parámetros

- O sus relaciones:

Parámetro poblacional	Descripción	Estimador																		
$\mu_1 - \mu_2$	Diferencia de medias	<table> <tr> <th>Parámetro poblacional</th><th>Descripción</th><th>Estimador</th></tr> <tr> <td>$\mu_1 - \mu_2$</td><td>Diferencia de medias</td><td>$\bar{X}_1 - \bar{X}_2$</td></tr> <tr> <td>$\sigma_1^2 - \sigma_2^2$</td><td>Diferencia de varianzas</td><td>$S_1^2 - S_2^2$</td></tr> <tr> <td>$P_1 - P_2$</td><td>Diferencia de proporciones</td><td>$p_1 - p_2$</td></tr> <tr> <td>σ_1^2 / σ_2^2</td><td>Ratio de varianzas</td><td>S_1^2 / S_2^2</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> </table>	Parámetro poblacional	Descripción	Estimador	$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$	$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$	$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$	σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2
Parámetro poblacional	Descripción	Estimador																		
$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$																		
$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$																		
$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$																		
σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2																		
...																		
	Diferencia de varianzas	<table> <tr> <th>Parámetro poblacional</th><th>Descripción</th><th>Estimador</th></tr> <tr> <td>$\mu_1 - \mu_2$</td><td>Diferencia de medias</td><td>$\bar{X}_1 - \bar{X}_2$</td></tr> <tr> <td>$\sigma_1^2 - \sigma_2^2$</td><td>Diferencia de varianzas</td><td>$S_1^2 - S_2^2$</td></tr> <tr> <td>$P_1 - P_2$</td><td>Diferencia de proporciones</td><td>$p_1 - p_2$</td></tr> <tr> <td>σ_1^2 / σ_2^2</td><td>Ratio de varianzas</td><td>S_1^2 / S_2^2</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> </table>	Parámetro poblacional	Descripción	Estimador	$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$	$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$	$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$	σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2
Parámetro poblacional	Descripción	Estimador																		
$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$																		
$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$																		
$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$																		
σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2																		
...																		
$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$																		
	Ratio de varianzas	<table> <tr> <th>Parámetro poblacional</th><th>Descripción</th><th>Estimador</th></tr> <tr> <td>$\mu_1 - \mu_2$</td><td>Diferencia de medias</td><td>$\bar{X}_1 - \bar{X}_2$</td></tr> <tr> <td>$\sigma_1^2 - \sigma_2^2$</td><td>Diferencia de varianzas</td><td>$S_1^2 - S_2^2$</td></tr> <tr> <td>$P_1 - P_2$</td><td>Diferencia de proporciones</td><td>$p_1 - p_2$</td></tr> <tr> <td>σ_1^2 / σ_2^2</td><td>Ratio de varianzas</td><td>S_1^2 / S_2^2</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> </table>	Parámetro poblacional	Descripción	Estimador	$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$	$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$	$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$	σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2
Parámetro poblacional	Descripción	Estimador																		
$\mu_1 - \mu_2$	Diferencia de medias	$\bar{X}_1 - \bar{X}_2$																		
$\sigma_1^2 - \sigma_2^2$	Diferencia de varianzas	$S_1^2 - S_2^2$																		
$P_1 - P_2$	Diferencia de proporciones	$p_1 - p_2$																		
σ_1^2 / σ_2^2	Ratio de varianzas	S_1^2 / S_2^2																		
...																		
...																		

Estimación de los parámetros

- Otras veces necesitamos algún procedimiento, hay muchos pero, el denominador común a ellos reside en obtener los b_j de forma que el ajuste entre el modelo y los datos observados sea el mejor posible.

- Método **de los Mínimos Cuadrados** (*Least Squares*)

$$\text{Minimizar } SC = \sum_{i=1}^N (\text{observado}_i - \text{modelo}_i)^2$$

- Método **de la Máxima Verosimilitud** (*Maximun Likelihood*)
- **Bootstrapping**
- Métodos **no paramétricos**
- etc

Estimación puntual

```
## Estimación maximo verosímil de  
parámetros mediante fitdistr()
```



fitdistr(*muestra*, “*modelo*”)

- ***muestra***: contiene los valores de la variable aleatoria a ajustar (sin valores perdidos)
- ***modelo***: la distribución de la variable aleatoria
- El **resultado** de la función es:
 - Una estimación del parámetro/s (**estimate**)¹
 - Error estándar (ET) de la estimación o estimaciones (**sd**)

¹Este valor también lo proporciona el resultado de las funciones t.test y prop.test que veremos en el apartado de contrastes de hipótesis.

Estimación de μ y σ de una Normal



```
> ajuste.Normal<-fitdistr(gasto, "normal")
>
> # El resultado del ajuste es una estimación para los parámetros  $\mu$ 
y  $\sigma$ 
> ajuste.Normal$estimate
      mean      sd
10.042287  1.882111

> # el error estándar para la estimación de los parámetros  $\mu$  y  $\sigma$ 
> ajuste.Normal$sd
      mean      sd
0.1657106  0.1171751
```

Estimaciones $\hat{m} = 10.042287$ y $\hat{\sigma} = 1.882111$

Error de estimación de $m = 0.1657106$

Error de estimación de $\sigma = 0.1171751$

Test o contrastes de hipótesis

Permiten **plantear** una **hipótesis** sobre la **población**¹ y **decidir**, a partir de la información contenida en la **muestra**, si dicha muestra **confirma** o **desmiente** dicha hipótesis, con una determinada probabilidad de equivocarnos pequeña y conocida de antemano (α)

- Es el procedimiento más ampliamente utilizado para responder a las preguntas que se plantean sobre fenómenos del mundo real en el proceso de investigación.
 - Ronald Fisher tuvo la idea de calcular probabilidades para evaluar evidencias
 - Jerzy Neyman and Egon Pearson usaron dichas probabilidades en la verificación de hipótesis

¹ Generalmente sobre un parámetro o parámetros de la población

Método o procedimiento para el contraste

1. Enunciar la **hipótesis**
2. Elegir el **estadístico** (según tipo de prueba)
3. Elegir un **nivel de confianza** y construir la **zona de aceptación para el valor del estadístico si H_0 fuera cierta**,
Fuera de este intervalo sólo se encuentran el 100% de los casos más raros. La **zona de rechazo** se denomina **región crítica**, y su probabilidad (área¹) es el **nivel de significación (α)**. El nivel de significación lo fija el investigador.
4. **Verificar la hipótesis** extrayendo una muestra adecuada y obtener de ella el correspondiente **estadístico**.
5. **Decidir**. Si el valor del estadístico calculado en la muestra cae dentro de la zona de aceptación se acepta la hipótesis y si no se rechaza. Para ello se compara:
 - **valor crítico y estadístico** o
 - **P-valor y α**

¹ Área bajo la curva de la función de densidad de la distribución del estadístico

Enunciado de hipótesis para el parámetro Θ

“Hipótesis Nula” H_0 : es la hipótesis de partida a contrastar y refleja el conocimiento previo de la situación.

Como si de un juicio penal se tratara, mientras no se demuestre lo contrario... “nada cambia”, “no hay efecto”, “nada influye”,... $\Theta = \Theta_0$

“Hipótesis Alternativa” H_1 : es la hipótesis con la que se contrasta la H nula. En el caso de hipótesis sobre un parámetro poblacional, se puede enunciar de varias formas:

Contraste bilateral

$$H_0 : \Theta = \Theta_0$$

$$H_1 : \Theta \neq \Theta_0$$

Contraste unilateral

$$H_0 : \Theta \geq \Theta_0$$

$$H_1 : \Theta < \Theta_0$$

Contraste unilateral

$$H_0 : \Theta \leq \Theta_0$$

$$H_1 : \Theta > \Theta_0$$

Lógica de los contrastes

1. Se asume que la H_0 es cierta
2. Elegimos un modelo estadístico (**estadístico**) que sabemos cómo se comporta si H_0 es cierta (**la distribución del estadístico**).
3. Ajustamos el modelo a la muestra (**calculamos ese estadístico a partir de la muestra**)
4. Evaluamos el ajuste del modelo, **calculando la probabilidad** de que el modelo se ajuste bien a los datos si la H_0 fuera cierta (**p-valor**)
5. Si el
 - p-valor es grande, entonces no tenemos suficiente evidencia como para rechazar H_0 (aceptaríamos H_0)
 - Cuanto más pequeño es el p-valor más confianza tenemos para acepta la H_1 (o rechazar H_0)

Test de hipótesis para la μ bilateral

- Se dispone de los datos referentes al gasto efectuado por 129 clientes en una tienda on line.
- Tras realizar un análisis descriptivo se sabe que

$$\bar{X} = 10,05452 \quad S = 1,908163 \quad N = 129$$

- El analista necesita confirmar si se puede asumir que el gasto medio de todos los potenciales clientes podría considerarse, en promedio, de 10 euros.

Contraste bilateral

$H_0 : \mu = 10$ euros

$H_1 : \mu \neq 10$ euros

θ es μ , en este caso

Test de hipótesis para la m

1. Se asume que la **H0** es **cierta**
2. Elegimos un **estadístico** que sabemos cómo se comporta (distribuye) si H0 es cierta.

$$\frac{\bar{X} - m}{S/\sqrt{N}} \sim t_{N-1} = t_{128} \rightarrow N(0,1)$$

3. Ajustamos el estadístico a la muestra

$$t = z = \frac{10,05452 - 10}{1,9081/\sqrt{129}} = 0,325$$

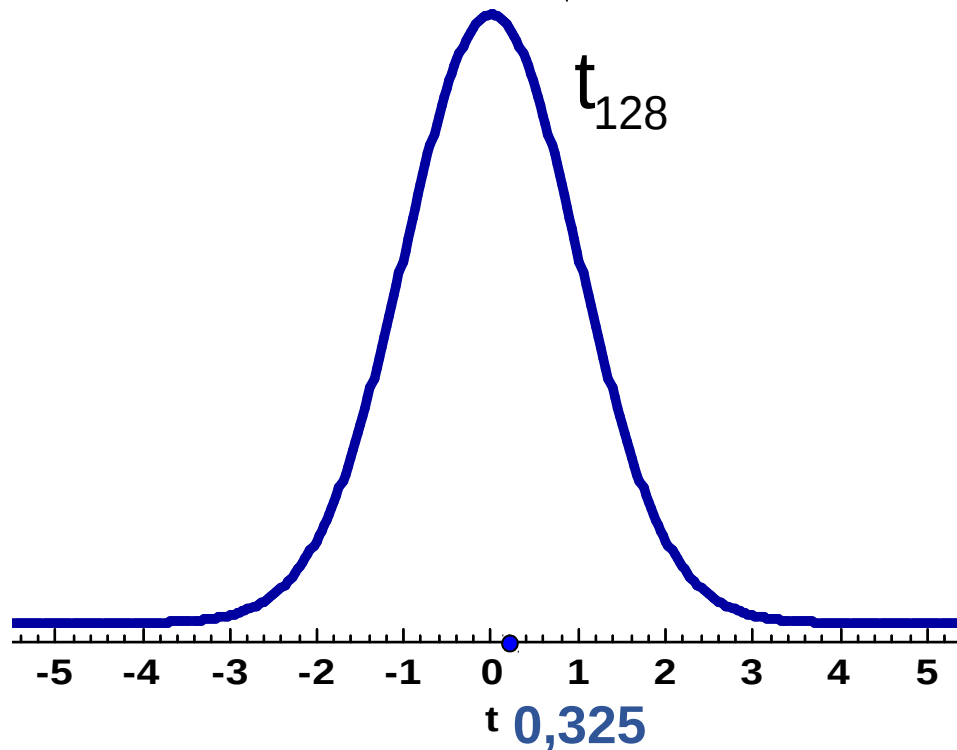
4. Evaluamos el ajuste del modelo.

¿Cuánto de probable es que el valor 0,325 sea un valor de t_{128} ?

Evaluación del ajuste del estadístico

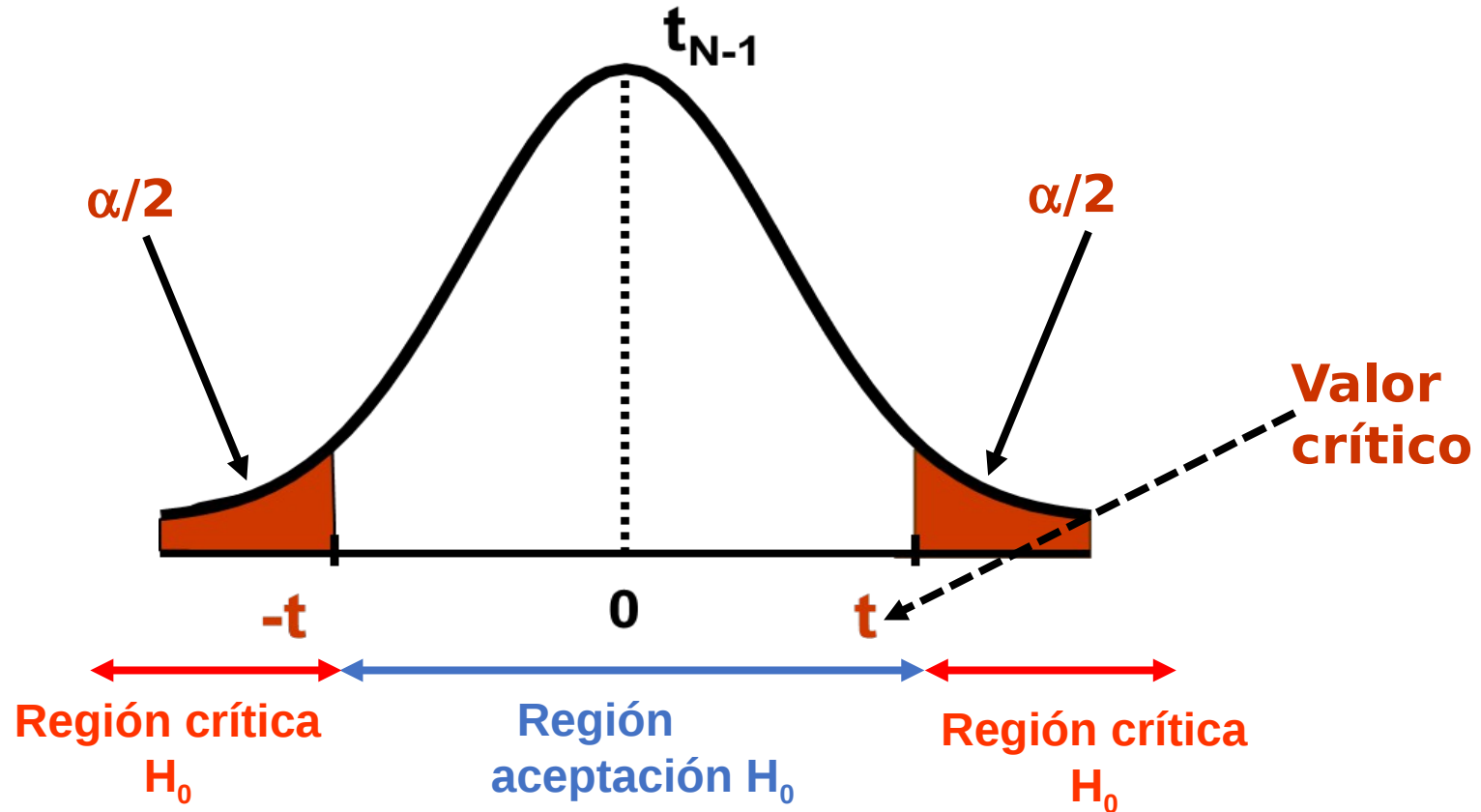
¿Es probable es que el valor $t=0,325$ sea un valor de t_{128} ?

Si la **H0** es cierta, $\frac{\bar{X} - m}{S/\sqrt{N}} \sim t_{128}$



Región crítica y región de aceptación

¿A partir de que valor se asume que el estadístico en nuestra muestra **NO** sigue una distribución t_{N-1} ?

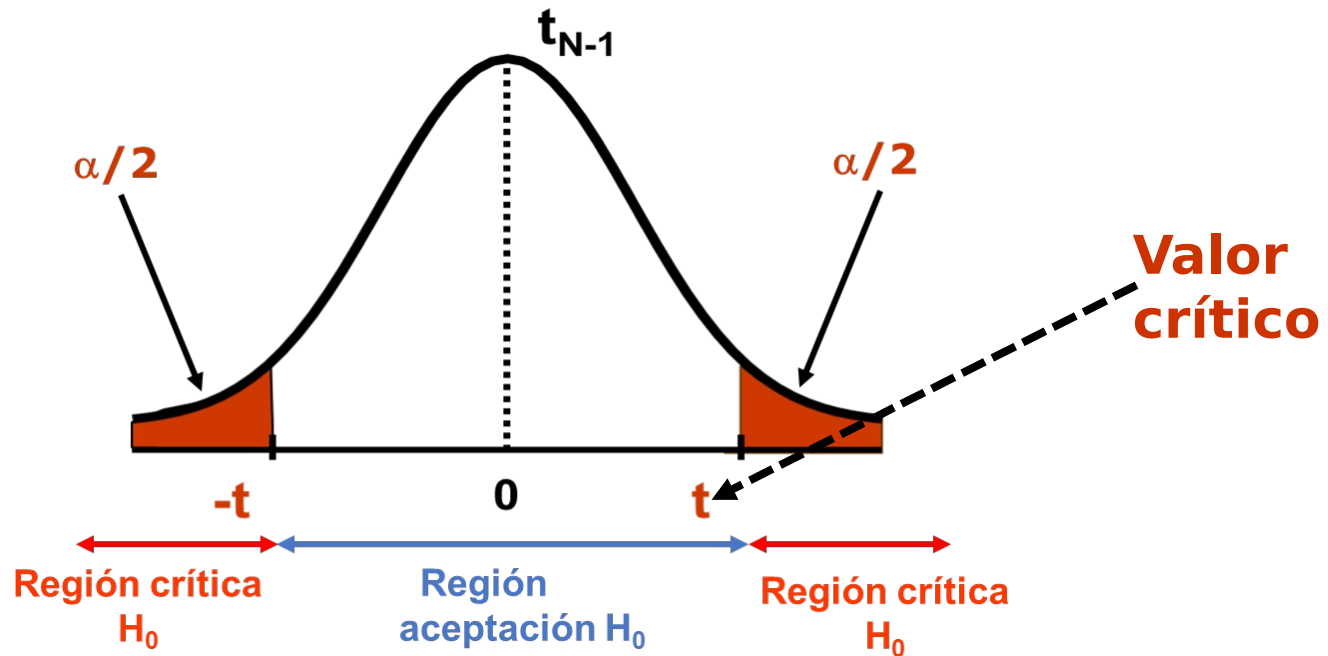


Donde α es una probabilidad “pequeña” que fija el investigador
Habitualmente α es 0,1 o 0,05 o 0,01 (**Nivel de Significación**)

Región crítica y región de aceptación

La **región crítica** es el conjunto de valores del estadístico de contraste que nos induce a rechazar la hipótesis nula.

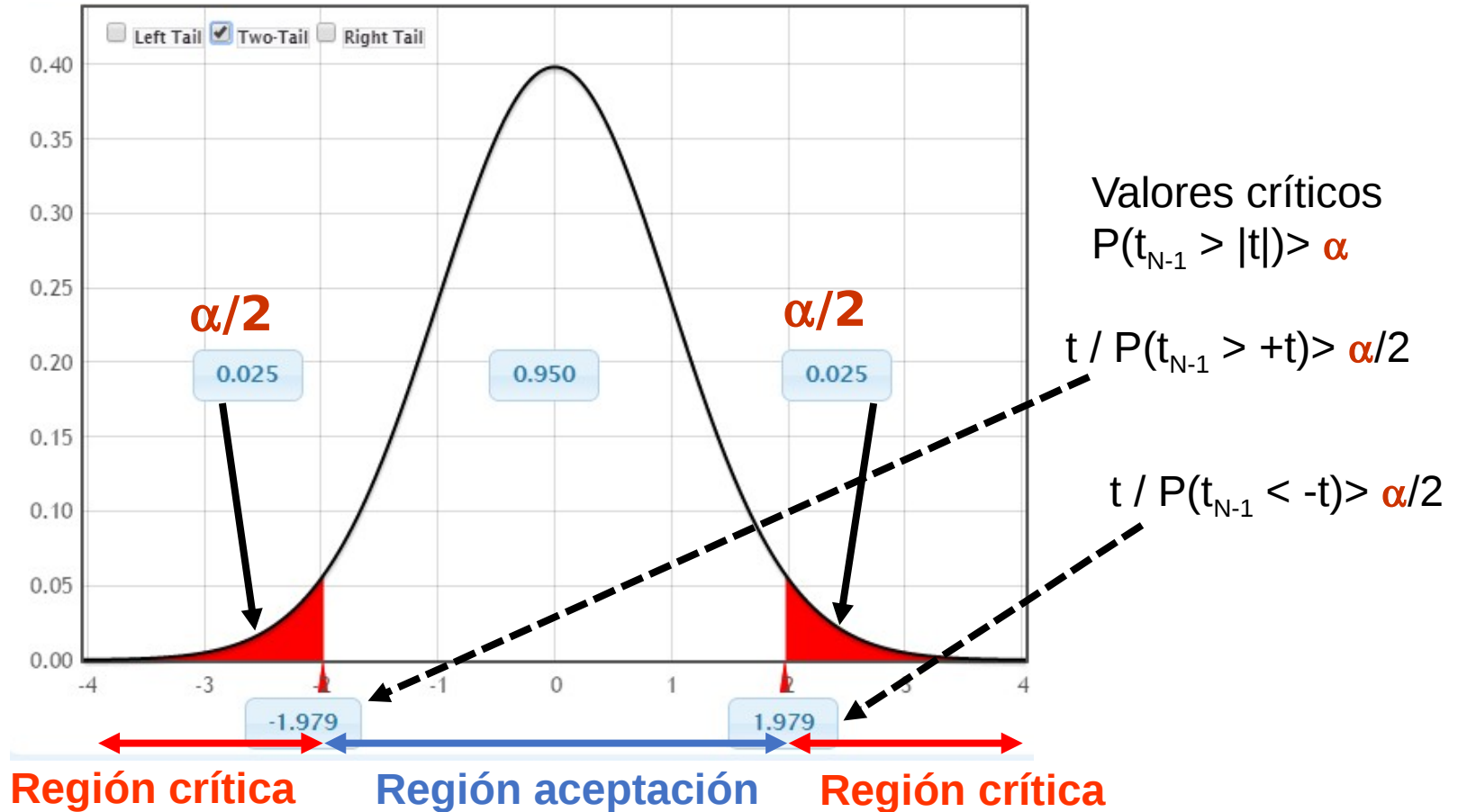
La **región de aceptación** es el conjunto de los valores del estadístico que nos induce a aceptar la hipótesis nula



Evaluación del ajuste del estadístico

¿Es probable es que el valor $t=0,325$ sea un valor de t_{128} ?

$$\alpha = 0,05$$



Valor crítico



Ejemplo IC del gasto total por cliente en una tienda on line (X). $X \sim N(m, \sigma)$

Variable "GASTO". Data set "datos1". Archivo "datos.RData"

```
## Obtención parámetros muestrales media y S
```

```
media.gasto<-mean(gasto, na.rm=T)
```

```
S.gasto<-sd(gasto, na.rm=T)
```

```
N.gasto<-length(gasto)
```

```
> N.compra1
```

```
[1] 129
```

```
> media.gasto
```

```
[1] 10.05452
```

```
> S.gasto
```

```
[1] 1.908163
```

```
# Desviación típica de la media muestral (standar error)
```

```
SE.gasto<-S.gasto/sqrt(N.gasto)
```

```
> SE.gasto
```

```
[1] 0.1680044
```



Valor crítico



Ejemplo IC del gasto total por cliente en una tienda on line (X). $X \sim N(m, \sigma)$

Variable "GASTO". Data set "datos1". Archivo "datos.RData"

```
## Obtención del valor crítico para el contraste sobre m
```

```
### Valores críticos
```

```
# Valor de una t con n-1 g.l. que deja a su derecha la mitad de alfa
```

```
# Valor de una t con n-1 g.l. con una probabilidad de ser superado
```

```
# igual a la mitad de alfa
```

```
t.negativo <- qt(0.05/2, 129-1)
```

$$t / P(t_{N-1} < -t) > \alpha/2$$

```
# Valor de una t con n-1 g.l. que deja a su izquierda la mitad de alfa
```

```
# Valor de una t con n-1 g.l. con una probabilidad de NO ser superado
```

```
# igual a la mitad de alfa
```

```
t.positivo <- qt(0.05/2, 129-1, lower.tail = F)
```

$$t / P(t_{N-1} > +t) > \alpha/2$$

```
t.negativo
```

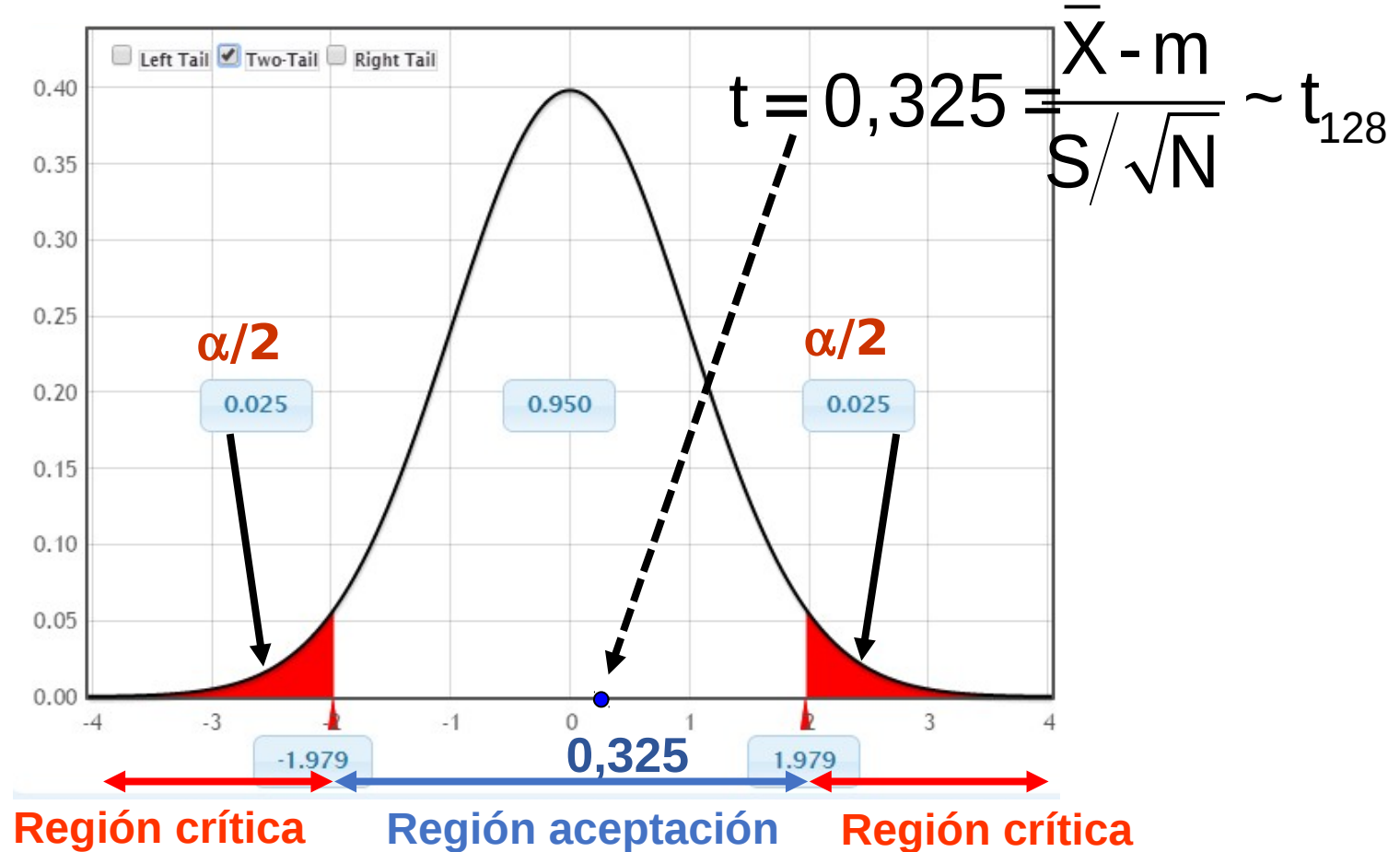
```
t.positivo
```

```
[1] -1.97671
```

```
[1] 1.978671
```

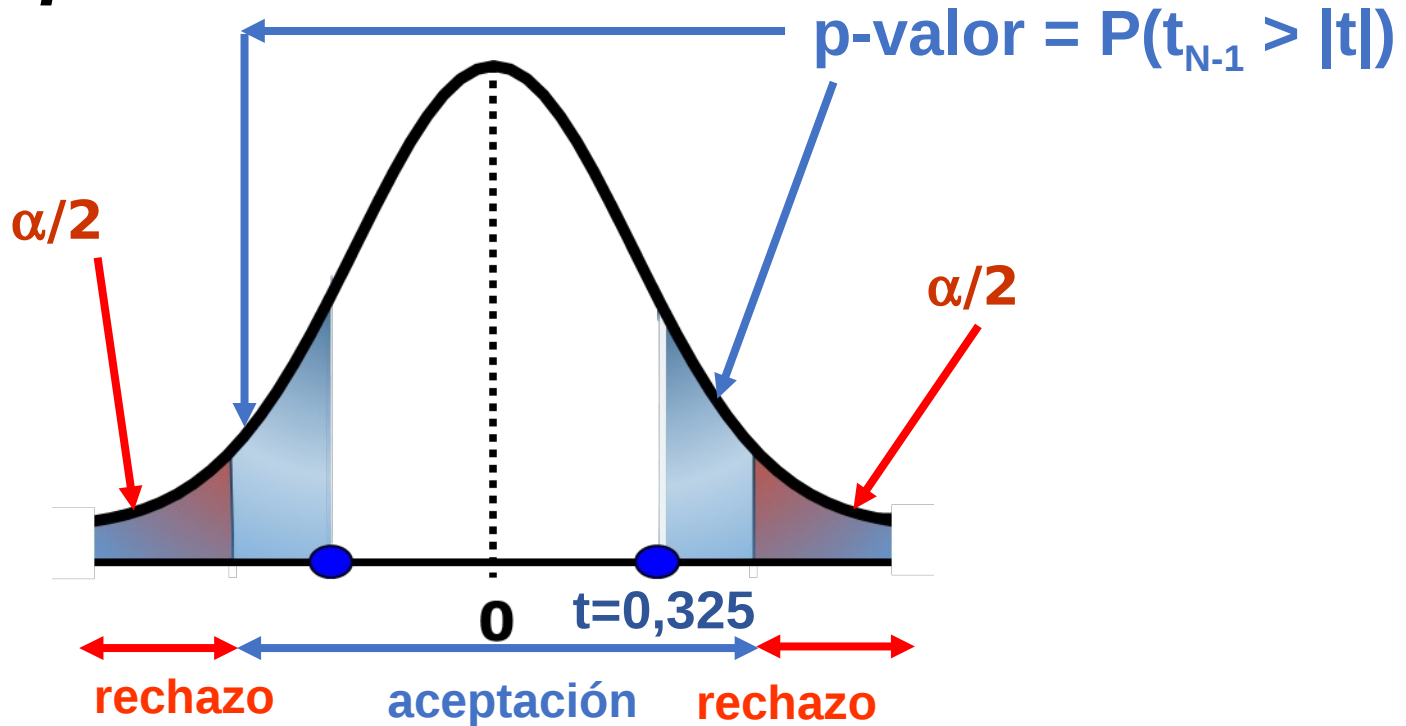
Evaluación del ajuste del estadístico

$$\alpha = 0,05$$



No hay suficiente evidencia para rechazar H_0 , por lo que es razonable concluir que el que el gasto medio de todos los potenciales clientes podría considerarse, en promedio, de 10 euros ($\alpha = 0,05$)

p-valor o *p-value*



Si **p-valor** $> \alpha$ $\Rightarrow P(t_{N-1} > |t|) > \alpha$ \Rightarrow es muy probable que el estadístico obtenido para nuestra muestra siga la distribución prevista si H_0 es cierta \Rightarrow **Aceptar H_0**

Si **p-valor** $\leq \alpha$ $\Rightarrow P(t_{N-1} > |t|) \leq \alpha$ \Rightarrow es muy poco probable que el estadístico obtenido para nuestra muestra siga la distribución prevista si H_0 es cierta \Rightarrow **Rechazar H_0**

P-valor



Ejemplo IC del gasto total por cliente en una tienda on line (X). $X \sim N(m, \sigma)$

Variable "GASTO". Data set "datos1". Archivo "datos.RData"

```
## Obtención del p-valor para el contraste sobre m
```

```
### p-valor
```

```
# Probabilidad de que una t con n-1 g.l.
```

```
# supere el valor del estadístico
```


```
pvalor <- pt(0.3245404, 129-1)
```

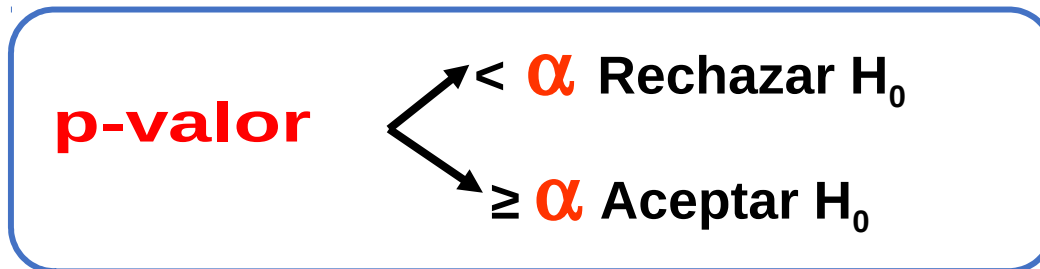
$P(t_{N-1} < |\text{estadístico } t|)$

```
pvalor
```

```
[1] 0.6269706
```

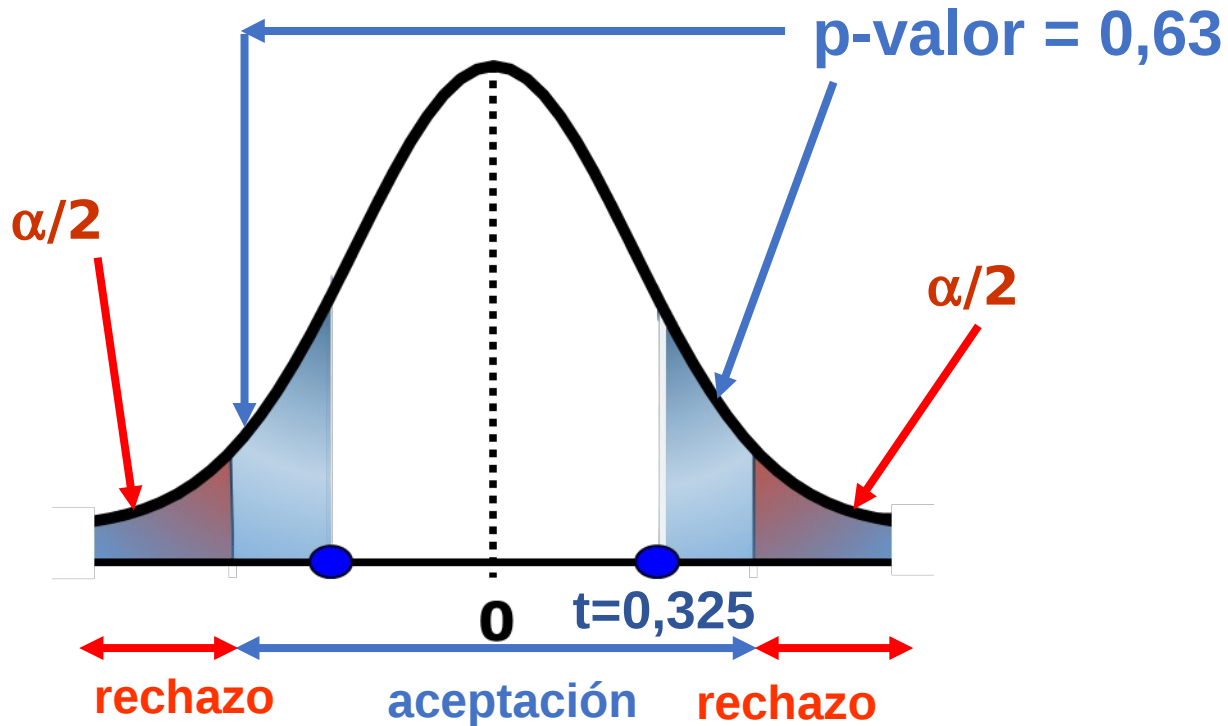
Decisión del contraste y **p-valor**

- Habitualmente se utiliza el **p-valor** (*p-value*) para decidir sobre el contraste.
- Intuitivamente, el **p-valor** nos informa de lo probable que es que lo observado en la muestra haya sido lo que ha sido, si suponemos cierta la hipótesis nula.
- Para determinar la magnitud del **p-valor** hay que compararlo con algo  se compara con **α** o **Nivel de Significación**, que es el **Riesgo de Primera Especie**



Decisión del contraste

$$\alpha = 0,05$$



Como **p-valor (0,63) > α (0,05)** ✉ Aceptar H_0


No hay suficiente evidencia para rechazar H_0 , por lo que es razonable concluir que el que el gasto medio de todos los potenciales clientes podría considerarse, en promedio, de 10 euros ($\alpha = 0,05$)

Tipos de error y compromiso

		Lo que pasa en realidad, ... nunca lo sabremos	
		H_0 es verdadera	H_0 es falsa
Decisión que tomamos	Aceptar H_0	Decisión correcta	Decisión incorrecta Error de tipo II 2ª especie
	Rechazar H_0	Decisión incorrecta Error de tipo I 1ª especie	Decisión correcta

Los valores más utilizados son $\alpha = 0,05$, $\alpha = 0,05$, $\alpha = 0,01$

Tipos de error y compromiso

- **Riesgo** de **1ª especie** (α): probabilidad de cometer un **error de 1ª especie** (**Rechazar H_0 cuando es verdadera**)
- **Riesgo** de **2ª especie** (β): probabilidad de cometer un **error de 2ª especie** (**Aceptar H_0 cuando es falsa**)
- Si disminuyo α  β aumenta y viceversa (para un N dado)
- Sólo se puede bajar α y β a la vez si aumentamos el tamaño de la muestra (N).



Compromiso

Riesgo de **1ª especie**: $\alpha = 0,05$ ó $\alpha = 0,01$

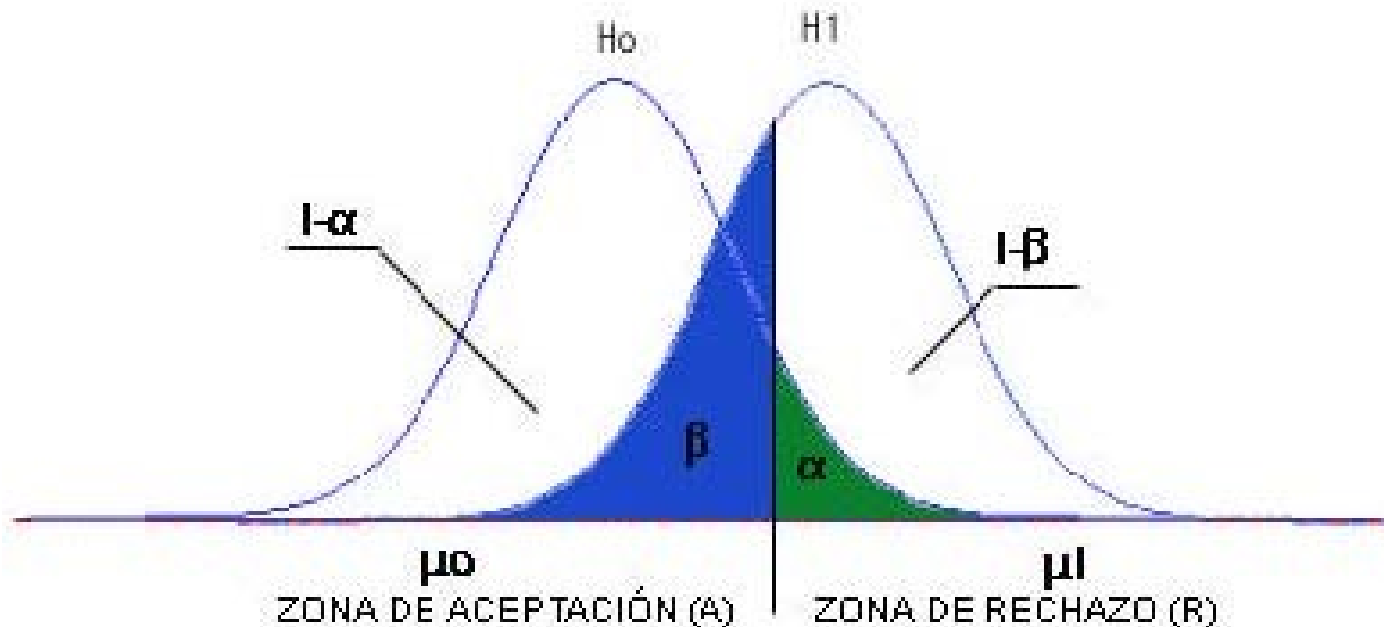
Potencia estadística: $(1-\beta) = 0,8$ ó más

La elección del nivel de significación depende del tipo de estudio y de la literatura especializada

Tipos de error y compromiso

La probabilidad de cometer un error de tipo II, β , es un valor desconocido que depende de tres factores:

- La hipótesis H_1 que consideremos verdadera.
- El valor de α .
- El tamaño del error típico (desviación típica) de la distribución muestral utilizada para efectuar el contraste.



Nivel de significación y Nivel de confianza

Tradicionalmente se utilizan los términos:

Nivel de Significación ✉ **Riesgo de 1ª especie (α)**

Nivel de Confianza $(1 - \alpha)\% = 95\% \text{ ó } 99\%$

• Ejemplo:

- ***Nivel de Significación*** $\alpha = 0,05 \leftrightarrow$ ***Nivel de Confianza*** 95%
- ***Nivel de Significación*** $\alpha = 0,01 \leftrightarrow$ ***Nivel de Confianza*** 99%

Decisión del test y valores habituales **p-valor** (p)

P-valor	Significado
$p \geq 0,1$	No hay evidencia contra la H_0 .
$0,05 \leq p < 0,1$	Evidencia débil contra la H_0
$0,01 \leq p < 0,05$	Evidencia moderada contra H_0
$0,001 \leq p < 0,01$	Fuerte evidencia contra H_0
$p < 0,001$	Evidencia muy fuerte contra H_0

El α fijado por el analista o investigador siempre debe acompañar a la decisión del contraste o test.

P-VALOR NO ES UN NÚMERO MÁGICO

Test de hipótesis para la m unilateral

- Se dispone de los datos referentes al gasto efectuado por 500 clientes en una tienda on line.
- Tras realizar un análisis descriptivo se sabe que

$$\bar{X} = 10,05452 \quad S = 1,908163 \quad N = 129$$

- El analista necesita confirmar si se puede asumir que el gasto medio de todos los potenciales clientes podría considerarse, en promedio, superior a 10 euros.

Contraste bilateral

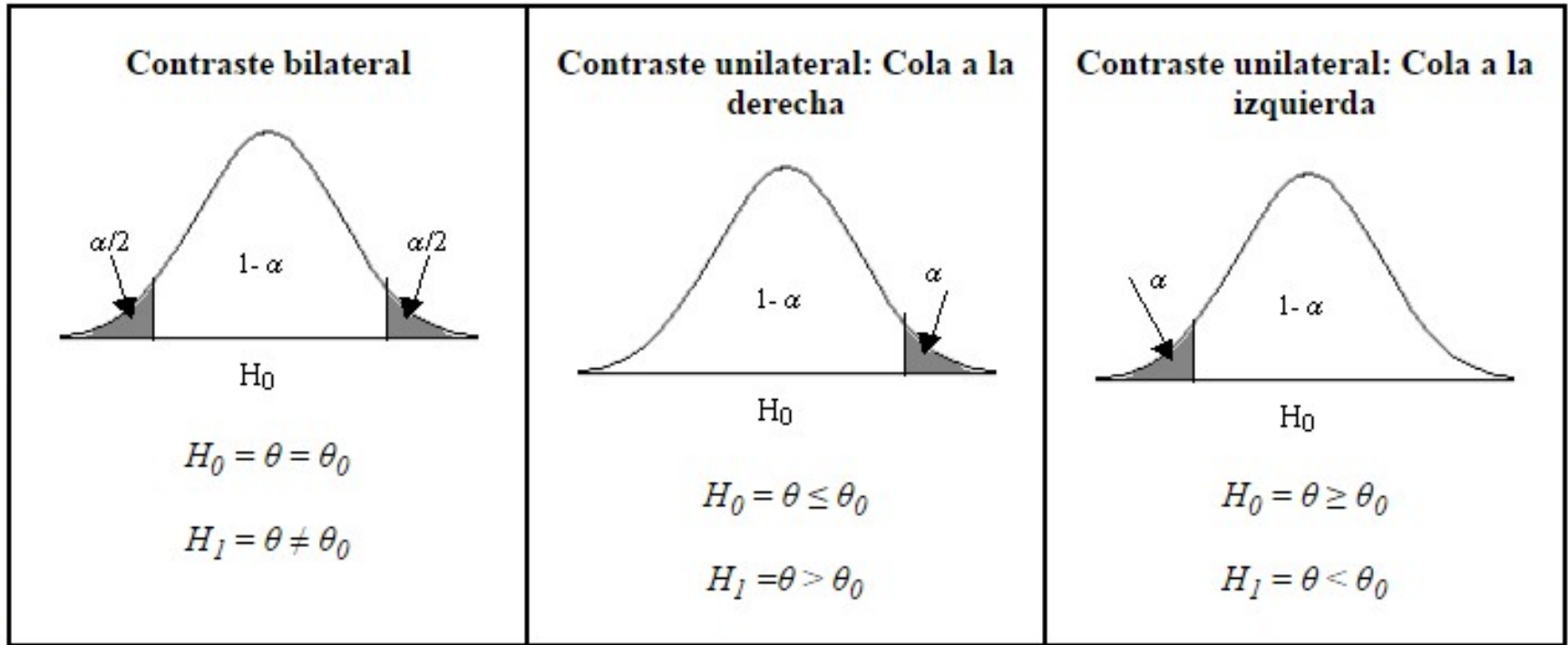
$$H_0 : m \geq 10 \text{ euros}$$

$$H_1 : m < 10 \text{ euros}$$

El bilateral es que habitualmente se utiliza. Salvo que la dirección del efecto (positivo o negativo) tenga importancia sobre la investigación, los unidireccionales no se suelen usar.

Test de hipótesis para la μ unilateral

- Lo único que cambia en el test, es que la región de rechazo queda sólo en una cola de la distribución del estadístico de contraste, t , en este test.
- Esto es válido para cualquier parámetro



Test de hipótesis para la P

Contraste bilateral

$H_0 : P = p_0$ Θ es P, en

$H_1 : P \neq p_0$ este caso

Estadístico de contraste que sabemos cómo se comporta (distribuye) si H_0 es cierta.

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)}} \sim N(0, 1)$$

Test para la comparación de medias en poblaciones independientes

Contraste bilateral

$$H_0 : m_1 = m_2 \quad \text{vs} \quad m_1 - m_2 = 0$$

$$H_1 : m_1 \neq m_2 \quad \text{vs} \quad m_1 \neq m_2 = 0$$

- Con varianzas conocidas (iguales o no)

$$\frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim N(0,1)$$

- Con varianzas desconocidas, pero iguales

$$\frac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{S_{\bar{X}_1 - \bar{X}_2}} \sim t_{(N_1 - 1) + (N_2 - 1)} \rightarrow N(0,1)$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S^2}{N_1} + \frac{S^2}{N_2}} \quad S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{(N_1 - 1) + (N_2 - 1)}$$

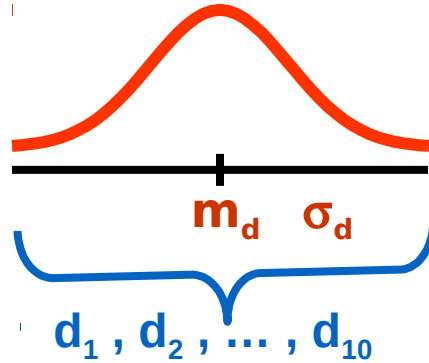
Sobra la igualdad de varianzas

- Si $\sigma^2_1 \neq \sigma^2_2$ ✉ el test de comparación de medias visto, y la correspondiente fórmula para el intervalo de confianza, tienen sólo carácter aproximado.
- Sin embargo, dicho test es bastante “robusto” frente al incumplimiento de esta hipótesis de homocedasticidad (igualdad de varianzas), especialmente si el número de observaciones en ambas muestras es parecido.
- En la práctica es razonable seguir utilizando, con carácter aproximado los procedimientos de comparación de medias expuestos en este capítulo, aunque existan diferencias entre las varianzas poblacionales, especialmente si N es grande (**Big Data**)

Test para la comparación de medias en poblaciones apareadas

Población

Muestreo



$$m_d = m_1 - m_2$$

$$d_i = X_{i1} - X_{i2}$$

Contraste bilateral

$$H_0 : m_d = 0$$

$$H_1 : m_d \neq 0$$

$$\frac{\bar{d}}{s_d / \sqrt{N}} \sim N(0,1)$$

- \bar{d} media de las diferencias entre las muestras
- S_d desviación típica de las diferencias entre las muestras

Test para la comparación de proporciones

Contraste bilateral

$$H_0 : p_1 = p_2 \quad \text{vs} \quad p_1 - p_2 = 0$$

$$H_1 : p_1 \neq p_2 \quad \text{vs} \quad p_1 \neq p_2 \neq 0$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{N_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{N_2}}} \sim N(0, 1)$$

Test de hipótesis para la σ^2

Contraste bilateral

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

$$(N-1) \frac{S^2}{\sigma^2} \sim \chi^2_{N-1}$$

Test para la comparación de varianzas

Contraste bilateral

$$H_0 : \sigma^2_1 = \sigma^2_2 \quad \text{vs} \quad \sigma^2_1 / \sigma^2_2 = 1$$

$$H_1 : \sigma^2_1 \neq \sigma^2_2 \quad \text{vs} \quad \sigma^2_1 / \sigma^2_2 \neq 1$$

$$\frac{s_1^2 / \frac{1}{2}}{s_2^2 / \frac{2}{2}} \sim F_{(N_1-1), (N_2-1)}$$

Existen varios métodos para hacer inferencia sobre desviaciones típicas de poblaciones normales. El más común de estos métodos, la prueba F para comparar la dispersión de dos poblaciones normales.

A diferencia de los procedimientos t para medias, la prueba F y otros procedimientos para hacer inferencia sobre desviaciones típicas son extremadamente sensibles a la falta de normalidad de las distribuciones. Esta falta de robustez no mejora con muestras grandes.

En la práctica es difícil decir si un valor significativo de F es una buena evidencia a favor de que las dispersiones poblacionales son distintas o simplemente es un signo de que las poblaciones no son normales

No es recomendable intentar hacer inferencia sobre desviaciones típicas de poblaciones en un curso básico de estadística aplicada, es mejor evaluar la normalidad de las distribuciones gráficamente, poniendo especial atención en la posible existencia de observaciones atípicas y en la falta de simetría de la distribución, y utilizar la versión del estadístico t de dos muestras que asume heterocedasticidad

Test de hipótesis para el coeficiente de correlación r

Contraste bilateral

$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

$$\frac{r - 0}{\sqrt{\frac{1 - r^2}{N - 2}}} \sim t_{N-1}$$

r es el coeficiente de correlación de Pearson o Kendall o Spearman entre x e y

Intervalos de confianza IC

Es un **intervalo** calculado de forma que se tenga una **probabilidad “alta”** de que el verdadero valor del parámetro poblacional (Θ) que tratamos de estimar esté dentro de dicho intervalo:

$$IC_{\theta}^{nc\%} = [v1, v2] / P(v1 \leq \theta \leq v2) = p \text{ "alta"}$$

- Θ es una constante (no una v.a) y desconocida generalmente que puede hacer referencia a la media, la varianza, el cociente de varianzas, una diferencia de medias,...de una variable (o un estadístico, en general)
- **nc%** se denomina **nivel de confianza** y es una probabilidad alta en %, generalmente:
 - 90%
 - 95%
 - 99%

Intervalo de confianza para la media

- Es un intervalo calculado de forma que se tenga una probabilidad “alta” de que el verdadero valor de la media poblacional (**m**) que tratamos de estimar esté dentro de dicho intervalo:

$$IC_m^{nc\%} = [v1, v2] / P(v1 \leq m \leq v2) = p \text{ "alta"}$$

- Vamos a calcular un $IC_m^{95\%}$, el proceso será idéntico para cualquier otro nivel de confianza y la lógica será la misma para cualquier otro parámetro.

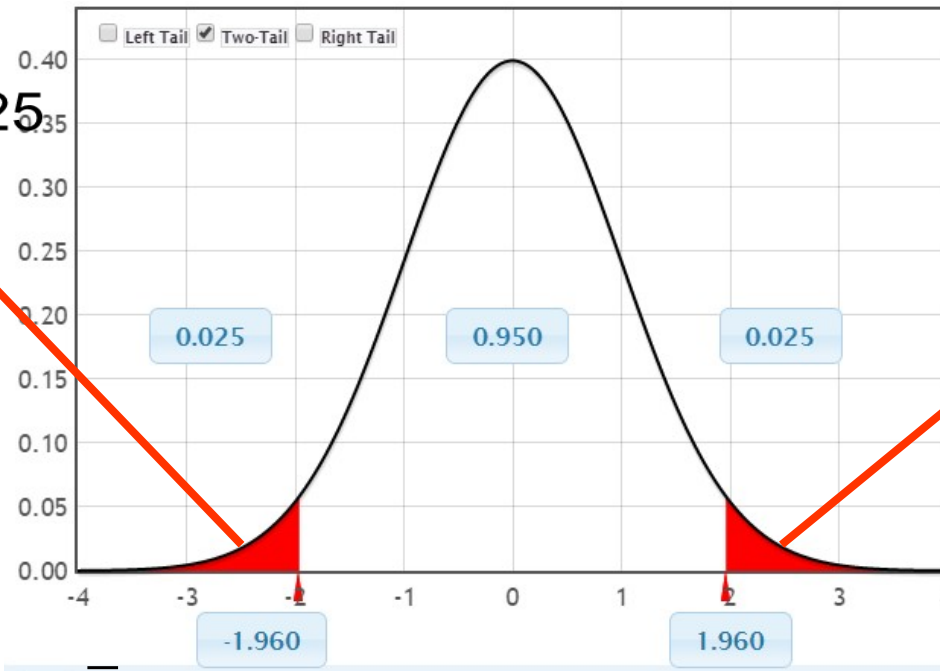
$$\bar{X} \underset{\substack{N \rightarrow \infty \\ (TCL)}}{\approx} N(m_{\bar{X}} = m, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}) \quad Z = \frac{\bar{X} - m}{\sigma / \sqrt{n}} \approx N(0,1)$$

Intervalo de confianza para la media

Los valores z_1 y z_2 tales que $P(Z \leq z_1) = 0,025$ y $P(Z \geq z_2) = 0,025$ son:

$$P(Z \leq z_1) = 0,025 = \alpha/2$$

$$P(Z \geq z_2) = 0,025 = \alpha/2$$



$$Z_1 = -1,96 = \frac{\bar{X} - m}{\sigma / \sqrt{n}}$$

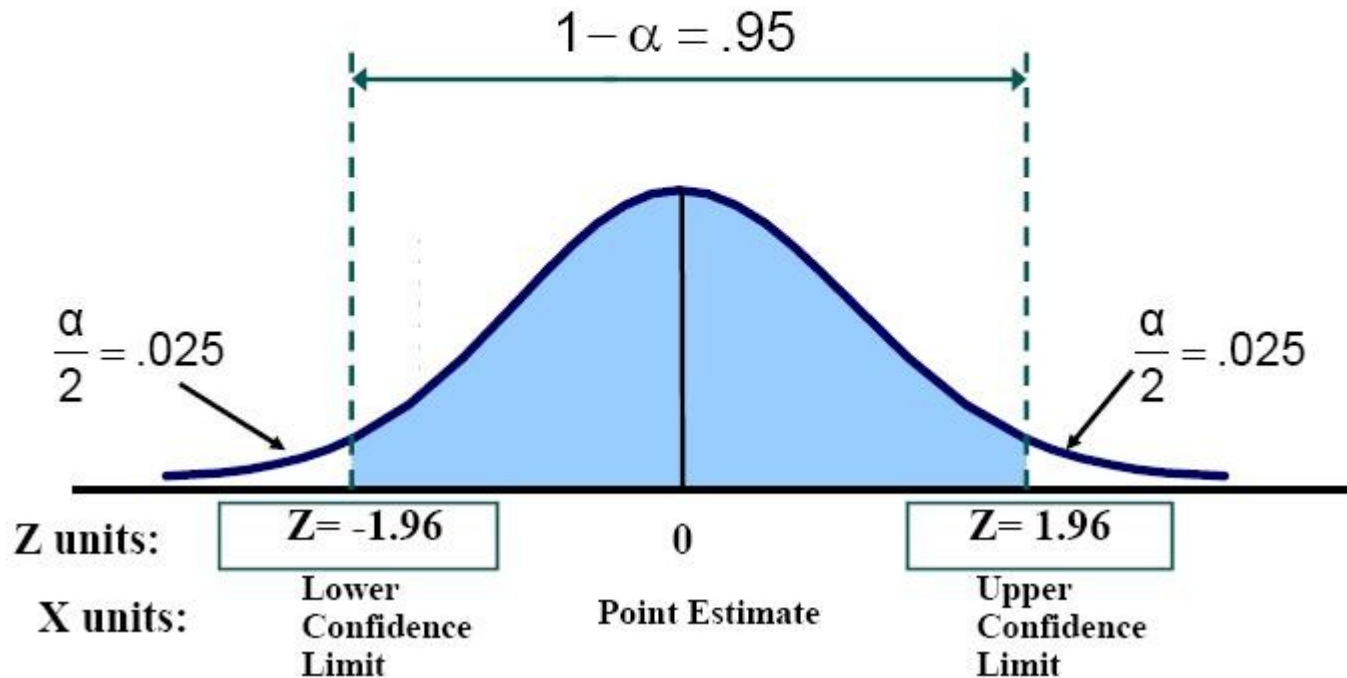
$$Z_2 = +1,96 = \frac{\bar{X} - m}{\sigma / \sqrt{n}}$$

$$IC_m^{95\%} = \left[\bar{X} - 1,96 \frac{\sigma}{\sqrt{N}}, \bar{X} + 1,96 \frac{\sigma}{\sqrt{N}} \right]$$

Nivel de significación y Nivel de confianza

Finding the Critical Value, Z

Consider a 95% confidence interval:



Intervalo de confianza para la media

$$IC_m^{nc\%} = \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{N}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{N}} \right]$$

$z_{\alpha/2}$ es el valor de una variable Normal(0, 1) tal que:
 $P(Z \geq z) = \alpha/2$ y $P(Z \leq -z) = \alpha/2$

Si no conocemos la varianza $\hat{\sigma}_{\bar{X}} = SE = \frac{S}{\sqrt{N}}$

$$t = \frac{\bar{X} - m}{\frac{S}{\sqrt{n}}} \approx t_{n-1}$$

Como en **Big Data** $N \rightarrow \infty$, este IC sirve:

- Sea cual sea la distribución de X
- Conozcamos la σ de X o su estimación S

Intervalo de confianza para la media

El caso general para el cálculo del IC para μ es:

$$IC_m^{nc\%} = \left[\bar{X} - t_{n-1}^{\alpha/2} \frac{S}{\sqrt{N}}, \bar{X} + t_{n-1}^{\alpha/2} \frac{S}{\sqrt{N}} \right]$$

$t_{n-1}^{\alpha/2}$ es el valor de una variable t con n-1 gl tal que:

$$P(t_{n-1}^{\alpha/2} \geq +t) = \alpha/2 \quad \text{y} \quad P(t_{n-1}^{\alpha/2} \leq -t) = \alpha/2$$

nc% es el **nivel de confianza**

α es el nivel de significación

$$(1 - \alpha)\% = \text{nc}\%$$

Si n es grande, prácticamente no hay diferencia entre la t y la $N(0,1)$

Ejemplo de IC para m



Ejemplo IC del gasto total por cliente en una tienda on line (X). $X \sim N(m, \sigma)$

IC para la media del gasto de los clientes en la web de dicha tienda

Variable "GASTO". Data set "datos1". Archivo "datos.RData"

```
## Obtención parámetros muestrales media y S
```

```
media.gasto<-mean(gasto, na.rm=T)
```

```
S.gasto<-sd(gasto, na.rm=T)
```

```
N.gasto<-length(gasto)
```

```
> N.gasto
```

```
[1] 131
```

```
> media.gasto
```

```
[1] 10.04229
```

```
> S.gasto
```

```
[1] 1.889449
```

```
# Desviación típica de la media muestral (standar error)
```

```
SE.gasto<-S.gasto/sqrt(N.gasto)
```

```
> SE.gasto
```

```
[1] 0.1663567
```

Ejemplo de IC para m



```
# valores críticos nivel de confianza 90%. Alfa=0.1. N(0,1)
alfa=0.1
z0.05<-qnorm(alfa/2, lower.tail=F)
z0.05
li<-media.gasto-z0.05*SE.gasto
ls<-media.gasto+z0.05*SE.gasto

# IC para la media del GASTO. Nivel de confianza 90%
c(li,ls)
> c(li,ls)
9.768654 10.315919
```

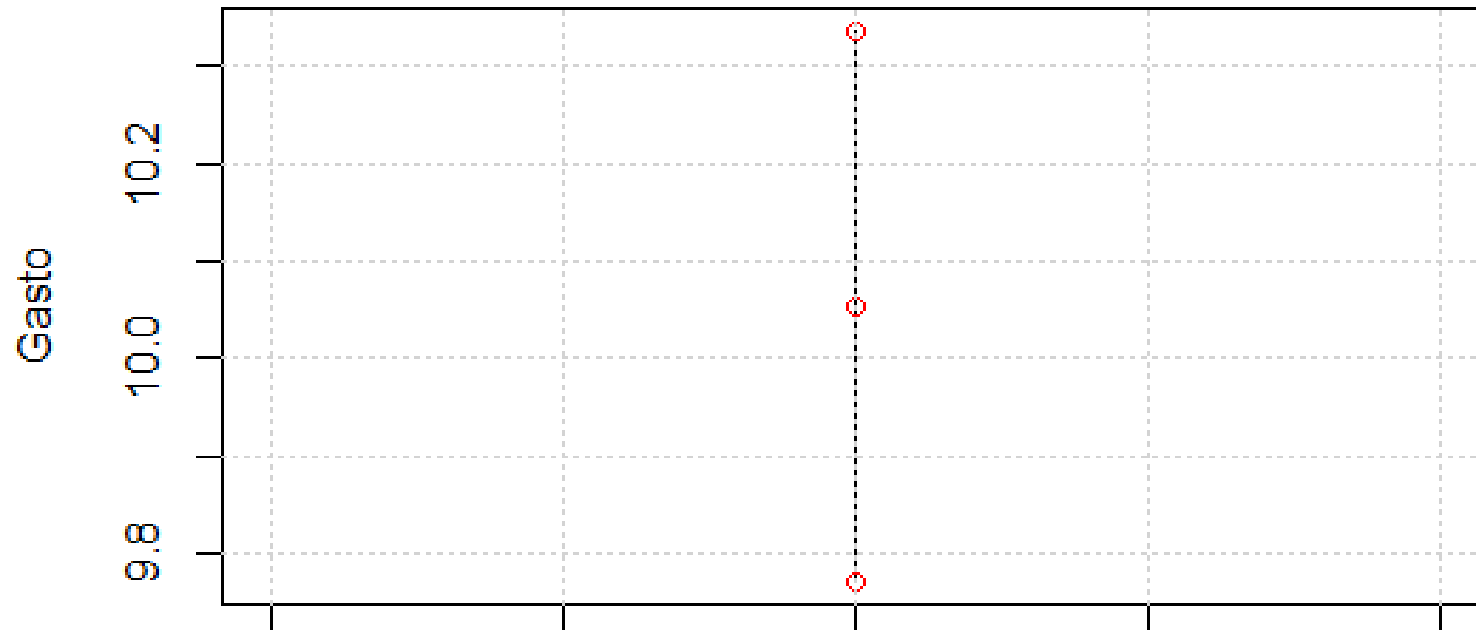
IC al para la media del gasto de los clientes en la web de dicha tienda
con nivel de confianza 90%

$$IC_m^{90\%} = [9,768 , 10,316]$$




Ejemplo de IC para m

IC para la media del GASTO al 90 %



Contrastes, estimación puntual e intervalos de confianza



- Un intervalo de confianza es un intervalo calculado a partir de los datos y que tiene una probabilidad elevada (**1- α**) de contener el valor desconocido del parámetro (θ) 
- Es el conjunto de todas las hipótesis compatibles o aceptables con un **α** dado.
- Es más informativo que dar sólo el resultado del test
- En general

Si $\theta_0 \in IC$  **Aceptar H_0**



Si $\theta_0 \notin IC$  **Rechazar H_0**

Contrastes, estimación puntual e intervalos de confianza



- Comparación de medias $m_1 = m_2$ ✉ $m_1 - m_2 = 0$

Si $0 \in IC$  Aceptar H_0
Si $0 \notin IC$  Rechazar H_0

- Comparación de proporciones $p_1 = p_2$ ✉ $p_1 - p_2 = 0$

Si $0 \in IC$  Aceptar H_0
Si $0 \notin IC$  Rechazar H_0

- Comparación de varianzas $\sigma^2_1 = \sigma^2_2$ ✉ $\sigma^2_1 / \sigma^2_2 = 1$

Si $1 \in IC$  Aceptar H_0
Si $1 \notin IC$  Rechazar H_0

Test de hipótesis para la m con IC

Θ es m , en este caso

Contraste bilateral

$H_0 : m = 10$ euros

$H_1 : m \neq 10$ euros

IC al para la media del gasto de los clientes en la web de dicha tienda con nivel de confianza 95%

$$IC_m^{95\%} = [9,716, 10,368]$$

Si $m_0 = 10 \in IC \longrightarrow$ **Aceptar H_0**

Si $m_0 = 10 \notin IC \longrightarrow$ **Rechazar H_0**

$m_0 = 10 \in IC_m^{95\%} = [9,716, 10,368] \longrightarrow$ **Aceptar H_0**

¡ Conclusión: la Hipótesis $m=10$ euros es aceptable !

Test de hipótesis para la m con IC

Contraste bilateral

H0 : m = 9 euros

H1 : m ≠ 9 euros

$$IC_m^{95\%} = [9,716, 10,368]$$

$$\vartheta \notin IC_m^{95\%} = [9,716, 10,368]$$

La hipótesis m=9 euros NO es aceptable

Contraste bilateral

H0 : m = 11 euros

H1 : m ≠ 11 euros

$$IC_m^{90\%} = [9,769, 10,316]$$

$$\vartheta \notin IC_m^{90\%} = [9,769, 10,316]$$

La hipótesis m=11 euros NO es aceptable

Intervalo de confianza para una proporción

- En vez de estudiar una v.a. que sigue una distribución Normal como el GASTO, o cualquier otra distribución si N es grande, estudiamos una v.a. que sigue una **distribución Binomial**
- Se sabe que si $X \sim B(n, p)$
$$X = \sum_{i=1}^n w_i / w_i \sim B(1, p) \text{ Bernoulli e independientes}$$
- Así, en vez de construir un IC para la media de la distribución ($m=np$), puede resultar más útil construir un IC para la proporción p
- Existen **dos alternativas** a la hora de construir un intervalo de confianza para p :
 - Considerar la **aproximación asintótica** de la distribución Binomial en la distribución Normal.
 - Utilizar un **método exacto**.

Intervalo de confianza para una proporción

Como en **Big Data** $N \rightarrow \infty$, usaremos la **aproximación asintótica**

$$X \sim B(n, p) \xrightarrow{\text{Aproximación}} X \sim N\left(m = np, \sigma = \sqrt{np(1-p)}\right)$$

Si usamos la proporción (**P**) en vez del número de veces que se verifica el suceso (**X**)¹:

$$\boxed{P = \frac{X}{n}} \xrightarrow{\text{Aproximación}} P \sim N\left(m = p, \sigma = \sqrt{p(1-p)}\right)$$

$$P \sim N\left(m = p, \sigma = \sqrt{p(1-p)}\right) \xrightarrow{\text{Tipificando}} Z = \frac{P - p}{\sqrt{p(1-p)}}$$
$$Z \sim N(0,1)$$

Por definición, si $X \sim B(n, p)$ $\Rightarrow X = \text{"Nº de veces que se da el suceso"}$

Intervalo de confianza para una proporción

El IC para una proporción se obtiene, siguiendo la forma general del IC para la m :

$$IC_m^{nc\%} = \left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{N}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{N}} \right]$$

- Sustituyendo \bar{X} por la proporción observada asociada al suceso: \hat{p}
- Sustituyendo S por la desviación típica observada de la proporción: $\sqrt{\hat{p}(1 - \hat{p})}$

$z_{\alpha/2}$ es el valor de una variable Normal(0, 1) tal que:
 $P(Z \geq z) = \alpha/2$ y $P(Z \leq -z) = \alpha/2$

Intervalo de confianza para una proporción

El caso general para el cálculo del IC es:

$$IC_P^{nc\%} = \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

nc% es el **nivel de confianza**
 α es el nivel de significación
 $(1 - \alpha)\%$ = nc%

El IC para P con la corrección de continuidad es: $IC_P^{nc\%} = \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} + \frac{1}{2n}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} + \frac{1}{2n} \right]$

Como en **Big Data** $N \rightarrow \infty$, esta corrección sería prácticamente 0

Ejemplo de IC para P



Sea X la variable definida como “nº de clientes que compran en la primera visita a la web” de una determinada tienda on line que no tiene nada que ver con la otra tienda. $X \sim \text{Binomial}(n, p)$

Vamos a construir un **IC** para la **proporción (P)** de clientes que compran en la **primera visita** a la web de dicha tienda

```
## Obtención parámetros muestrales p y S squared
frec.compra1<-table(compra1)
prop.compra1<-prop.table(frec.compra1)
p.compra1<-prop.compra1["si"] # o bien p.compra1<-prop.compra1[2]
N.compra1<-length(compra1)
S2.compra1<-p.compra1*(1-p.compra1)
> p.compra1
      0.09923664
> S2.compra1
      0.08938873
> N.compra1
[1] 131
```



Ejemplo de IC para P



```
# Desviación típica de la proporción muestral (standar error)
SE.compra1<-sqrt(S2.compra1/N.compra1)

# valores críticos nivel de confianza 90%. Alfa=0.1. N(0,1)
alfa=0.1
z0.05<-qnorm(alfa/2, lower.tail=F) # o bien z0.95<-qnorm(alfa/2,
lower.tail=T)
z0.05
li<-p.compra1-z0.05*SE.compra1
ls<-p.compra1+z0.05*SE.compra1

# IC para la proporción de compras a la primera visita. Nivel de
confianza 90%
c(li*100,ls*100)
  si      si
5.626984 14.220344
```

IC al para la proporción de los clientes que compran en la primera visita a la web de una tienda on line con nivel de confianza 90%

$$IC_P^{90\%} = [5,627\% , 14,22\%]$$

Intervalo de confianza para la varianza

- Es un intervalo calculado de forma que se tenga una probabilidad “alta” de que el verdadero valor de la varianza poblacional (σ^2) que tratamos de estimar esté dentro de dicho intervalo:

$$IC_{\sigma^2}^{nc\%} = [v1, v2] / P(v1 \leq \sigma^2 \leq v2) = p \text{ "alta"}$$

- Vamos a calcular un $IC_{\sigma^2}^{95\%}$, el proceso será idéntico para cualquier otro nivel de confianza.

$$\bar{X} \underset{\substack{N \rightarrow \infty \\ (TCL)}}{\approx} N(m_{\bar{X}} = m, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}})$$

$$\frac{(N-1)S^2}{\sigma^2} \approx \chi_{N-1}^2$$

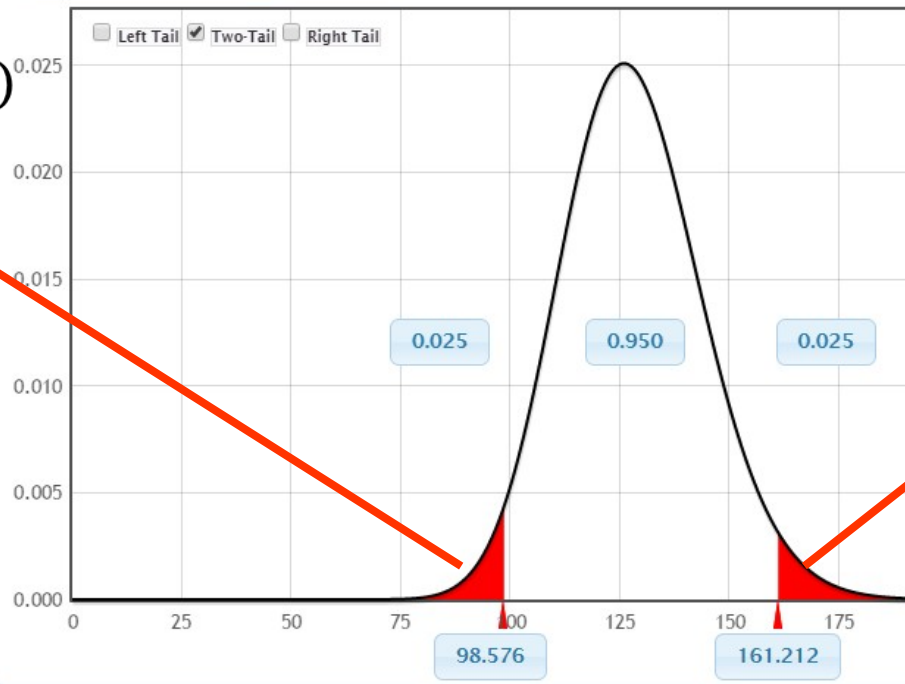
Intervalo de confianza para la varianza

Los valores g_1 y g_2 / $P\left(g_1 \leq \frac{(N-1)S^2}{\sigma^2} \leq g_2\right) = 0,95$ son:

$$\left. \begin{array}{l} \sigma^2 \leq \frac{(N-1)S^2}{g_1} \\ \sigma^2 \geq \frac{(N-1)S^2}{g_2} \end{array} \right\} IC_{\sigma^2}^{95\%} = \left[\frac{(N-1)S^2}{g_2}, \frac{(N-1)S^2}{g_1} \right]$$

$$P(g_1 \leq \chi_{N-1}^2) = 0,025 = \alpha/2$$

$$g_1 = 98,576$$



$$P(g_2 \geq \chi_{N-1}^2) = 0,025 = \alpha/2$$

$$g_2 = 161,212$$

Intervalo de confianza para la varianza

El caso general para el cálculo del IC es:

$$IC_{\sigma^2}^{nc\%} = \left[\frac{(N-1)S^2}{g_2}, \frac{(N-1)S^2}{g_1} \right]$$

g_1 es el valor de una variable χ^2_{N-1} tal que: $P(g_1 \leq \chi^2_{N-1}) = \alpha/2$

g_2 es el valor de una variable χ^2_{N-1} tal que: $P(g_2 \geq \chi^2_{N-1}) = \alpha/2$

El IC para la desviación típica sería:

$$IC_{\sigma}^{nc\%} = \left[\sqrt{\frac{(N-1)S^2}{g_2}}, \sqrt{\frac{(N-1)S^2}{g_1}} \right]$$

nc% es el **nivel de confianza**

α es el nivel de significación

$(1 - \alpha)\% = \text{nc}\%$

Como en **Big Data** $N \rightarrow \infty$,
este IC sirve sea cual sea la
distribución de X

Ejemplo de IC para σ^2



Ejemplo IC del gasto total por cliente en una tienda on line (X). $X \sim N(\mu, \sigma)$

IC para la varianza del gasto de los clientes en la web de dicha tienda

Variable "GASTO". Data set "datos1". Archivo "datos.RData"

```
## Obtención parámetros muestrales media y varianza
```

```
> media.gasto<-mean(gasto, na.rm=T)
```

```
> S2.gasto<-var(gasto, na.rm=T)
```

```
> N.gasto<-length(gasto)
```

```
> media.gasto
```

```
[1] 10.04229
```

```
> S2.gasto
```

```
[1] 3.570017
```

```
> N.gasto
```

```
[1] 129
```



Ejemplo de IC para σ^2



```
# valores críticos nivel de confianza 95%. Alfa=0.05. chi 2 con N-1
g.l.
> alfa=0.05
> nc $\leftarrow$ (1-alfa)*100
> gl.gasto $\leftarrow$ N.gasto-1
> g1<-qchisq(alfa/2, gl.gasto, lower.tail=T)
> g2<-qchisq(alfa/2, gl.gasto, lower.tail=F)
> g1
[1] 98.5756
> g2
[1] 161.2087
> li $\leftarrow$ (N.gasto-1)*S2.gasto/g2
> ls $\leftarrow$ (N.gasto-1)*S2.gasto/g1
```



Ejemplo de IC para σ^2



```
> # IC para la varianza del GASTO. Nivel de confianza 95%
> nc
[1] 95
> c(li,ls)
[1] 2.834599 4.635652
>
> # IC para la desviación típica del GASTO. Nivel de confianza 95%
> nc
[1] 95
> c(sqrt(li),sqrt(ls))
[1] 1.683627 2.153056
```

IC al para la varianza y desviación típica del gasto de los clientes en la web de dicha tienda con nivel de confianza 90%

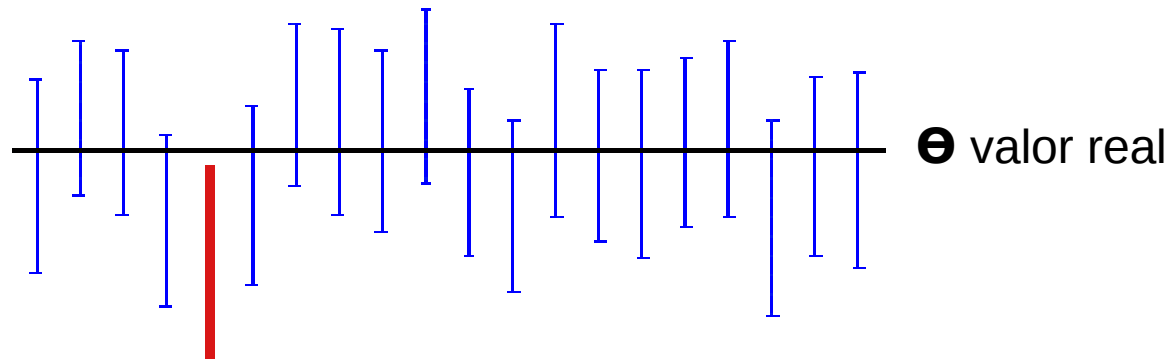
$$IC_{\sigma^2}^{95\%} = [2,835, 4,636] \quad IC_{\sigma}^{95\%} = [1,684, 2,153]$$



Intervalos de Confianza

- Poblaciones normales o no si $N \rightarrow \infty$ (Big Data)
- Para un parámetro poblacional de la distribución: μ , σ , σ^2
 - Para la comparación de dos parámetros poblacionales: μ , σ
 - Comparación de medias
 - Muestras independientes
 - » Varianzas conocidas o desconocidas
 - » Iguales
 - » Diferentes
 - Muestras dependientes (datos apareados)
 - Comparación de varianzas
- Poblaciones binomiales $B(n,p)$
 - Para un parámetro poblacional de la distribución: p
 - Para la comparación de dos parámetros poblacionales
- etc

- Es una forma más informativa de dar una estimación sobre un parámetro.
- También puede usarse en la decisión de tests o contrastes de hipótesis.
- **Ejemplo:** Si tomamos 20 muestras de una misma población y construimos un IC a partir de cada una de esas 20 muestras, en el 95% de éstas (19 barras azules), el verdadero valor del parámetro poblacional estará en el IC construido



Paso 1

- Conocimiento del problema y objetivos

Paso 2

- Plantear H_0 y H_1

Paso 3

- Preparar el experimento y tomar la muestra

Paso 4

- Analizar la muestra: descriptivo (normalidad, datos anómalos,...)

Paso 5

- Identificar el tipo de prueba (estadísticos)

Paso 6

- Determinar nivel de significación α

Paso 7

- Realizar la prueba (t, F, ...) y/o obtener IC

Paso 8

- Decidir: Aceptar o Rechazar hipótesis

Paso 9

- Interpretar resultados y redactar conclusiones

Procedimiento para el contraste

- 1º Determinar, claramente, la hipótesis nula H_0 y la hipótesis alternativa H_1 .
- 2º Elegir el nivel de significación.
- 3º Seleccionar un estadístico cuya distribución muestral sea conocida en el caso de que la hipótesis nula sea cierta.
- 4º Determinar la región crítica.
- 5º Calcular el valor del estadístico de contraste para la muestra elegida.
- 6º Sacar las conclusiones estadísticas del contraste (aceptar o rechazar H_0).
- 7º Sacar las conclusiones no estadísticas (biológicas, médicas, económicas, etc.) a que nos llevan los resultados estadísticos.

Contrastes e IC con R



- Hemos visto como calcular algún IC
- Los IC también aparecen como resultado de los contrastes al utilizar las funciones correspondientes.
- Conociendo la lógica y estadísticos de los contrastes se pueden elaborar scripts para cualquiera de ellos.
- R dispone de algunas funciones que realizan la mayoría de los tests vistos.
- Se usan básicamente las funciones:
 - **t.test()**
 - **prop.test()**
 - **var.test()**
 - **cor.test()**

NOTA: El t.test usa un estadístico t y no la distribución normal, pero los resultados en **Big Data** son prácticamente idénticos.



t.test



```
t.test(x, y = NULL, alternative = c("two.sided", "less",  
"greater"), mu = 0, paired = FALSE, var.equal = FALSE,  
conf.level = 0.95)
```

- **x** es un vector de datos correspondiente a una de las muestras o a la única muestra del problema. Si estamos haciendo un test sobre la media de una población, x contendrá la única muestra; si estamos realizando un test de comparación de medias, x será la primera de las dos muestras.
- **y** correspondería con la segunda muestra en un test de comparación de medias. Si no es el caso y estamos en un test sobre una sola población, simplemente no se incluye.
- **alternative** especifica la dirección de la hipótesis alternativa. Como puede verse, tiene 3 posibles valores, "**two.sided**" (bilateral), "**less**" (unilateral a la izquierda) y "**greater**" (unilateral a la derecha).
- **mu** es el valor hipotético con el que se compara la media o la diferencia de medias en el contraste (hipótesis nula)
- **paired** especifica si las dos muestras x e y, en caso de que aparezcan, son apareadas o no. En el caso en el que aparecen dos muestras, **var.equal** especifica si podemos suponer varianzas iguales o no.
- **conf.level** es el nivel de confianza de los intervalos que se mostrarán asociados al test (en tanto por 1).



```
prop.test(x, n, p = NULL, alternative = c("two.sided",  
"less", "greater"), conf.level = 0.95, correct = TRUE)
```

- **x** puede especificar dos cosas. O bien simplemente el número de éxitos, o bien, mediante una matriz de dos columnas, el número de éxitos y de fracasos en cada muestra.
- **n** especifica el número de datos de la muestra en el caso en que x sea el número de éxitos, y es ignorado en el caso en que x proporcione también el número de fracasos.
- **p** es el vector de probabilidades de éxito bajo la hipótesis nula. Debe ser un vector de la misma dimensión que el número de elementos especificado en x.
- **alternative** especifica la dirección de la hipótesis alternativa, tomando los valores "two.sided", "greater" o "less".
- **conf.level** es el nivel de confianza de los intervalos que se muestran entre los resultados.
- **correct** especifica si se usa la corrección por continuidad de Yates. Obsérvese que la opción por defecto es que sí se use esta corrección.

cor.test()



```
cor.test(x, y, alternative = c("two.sided", "less",  
"greater"), method = c("pearson", "kendall", "spearman"),  
exact = NULL, conf.level = 0.95, continuity =  
FALSE, ...)
```



var.test()



```
var.test(x, y, ratio = 1, alternative = c("two.sided",  
"less", "greater"), conf.level = 0.95, ...)
```

```
var.test(formula, data, subset, na.action, ...)
```



Ejemplo de test para m



```
#### Ejemplo IC del gasto total por cliente en una tienda on
line
####  $X \sim N(m, \sigma)$ 
#### Vamos a realizar un test sobre la media del gasto de los
clientes en la web de dicha tienda
#### Variable "GASTO". Data set "datos1". Archivo "datos.RData"

## Eliminamos los valores perdidos de la variable.
gasto<-na.omit(datos1$GASTO)

## Parámetros muestrales
media.gasto<-mean(gasto, na.rm=T)
S.gasto<-sd(gasto, na.rm=T)
N.gasto<-length(gasto)
media.gasto
S.gasto
N.gasto
```

Contraste bilateral

$H_0 : m = 9$ euros

$H_1 : m \neq 9$ euros



Ejemplo de test para m



```
> #####  
> ## Contraste BILATERAL. Nivel de confianza 95%  
> ## Usando la función t.test()  
> #####  
> ## Recordad su sintaxis  
> ?t.test()  
>  
> t.test(gasto, alternative = "two.sided", mu=9, conf.level = 0.95)  
One Sample t-test  
data:  gasto  
t = 6.2654, df = 128, p-value = 5.213e-09  
alternative hypothesis: true mean is not equal to 9  
95 percent confidence interval:  
  9.713122 10.371452  
sample estimates:  
mean of x  
 10.04229
```



Ejemplo de test para m

Resumiendo



One Sample t-test **Contraste sobre una muestra**

data: gasto (**muestra**)

t = 6.2654 (**estadístico¹**) df = 128 (**grados de libertad**) p-value =
5.213e-09

P-valor

alternative hypothesis: true mean is not equal to 9
(**H1 : m \neq 9 euros**)

95 percent confidence interval:

9.713122 10.371452 **Intervalo de confianza de la m al 95% nc**

sample estimates: **Algunas estimaciones de la muestra, como la media**
mean of x
10.04229

¹Estadístico t, pero si la muestra es suficientemente grande (mayor que 30) como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra.

Ejemplo de test para m

Resumiendo



Contraste bilateral

H0 : m = 9 euros

H1 : m ≠ 9 euros

- Resolución del contraste mediante IC al 95%:

$$IC_m^{95\%} = [9.713122, 10.371452]$$

$9 \notin IC_m^{95\%}$ ✉ La hipótesis m=9 euros NO es aceptable

- Resolución mediante el p-valor:

(p-value = 5.213e-09) << 0,001 ✉ Evidencia muy fuerte contra H0

¡ Conclusión: la Hipótesis m=9 euros NO es aceptable !

Ejemplo de test para m



```
> #####  
> ## Contraste UNILATERAL. Nivel de confianza 95%  
> ## Usando la función t.test()  
> #####
```

```
> t.test(gasto, y=NULL, alternative = "greater", mu=9,  
conf.level = 0.95)
```

One Sample t-test

data: gasto

t = 6.2654, df = 128, p-value = 2.607e-09

alternative hypothesis: true mean is greater than 9

95 percent confidence interval:

9.76666 Inf

sample estimates:

mean of x

10.04229

Contraste unilateral

H0 : $m \leq 9$ euros

H1 : $m > 9$ euros

Ejemplo de test para m

Resumiendo



One Sample t-test

data: gasto (**muestra**)

t = 6.2654 (**estadístico¹**) df = 128 (**grados de libertad**) p-value =
2.607e-09

P-valor

alternative hypothesis: true mean is greater than to 9
(**H1 : m > 9 euros**)

95 percent confidence interval:

9.713122 Inf **Intervalo de confianza de la m al 95% nc**
Observad que el límite superior es infinito

sample estimates: **Algunas estimaciones de la muestra, como la media**
mean of x
10.04229

¹Estadístico t, pero si la muestra es suficientemente grande (mayor que 30) como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra.



Ejemplo de test para m

Resumiendo



Contraste unilateral

$H_0 : m \leq 9$ euros

$H_1 : m > 9$ euros

- Resolución del contraste mediante IC al 95%:

$$IC_m^{95\%} = [9.713122, \infty]$$

$9 \notin IC_m^{95\%}$ ✉ La hipótesis $m \leq 9$ euros NO es aceptable

- Resolución mediante el p-valor:

(p-value = 2.607e-09) $\ll 0,001$ ✉ Evidencia muy fuerte contra H_0

¡ Conclusión: la Hipótesis $m \leq 9$ euros NO es aceptable !

Test para la comparación de medias en poblaciones independientes



```
### Ejemplo de comparación del gasto total por cliente en una tienda on line (X) entre hombres y mujeres
```

```
### Vamos a realizar un test de comparación de medias de poblaciones independientes.
```

```
## Selección de la muestra de gasto para mujeres ("mujer")  
gasto.mujer<-datos1[which(datos1$SEX0=="mujer"), "GASTO"]
```

```
## Eliminamos los valores perdidos de la muestra.  
gasto.mujer<-na.omit(gasto.mujer)
```

```
## Selección de la muestra de gasto para hombres ("varon")  
gasto.varon<-datos1[which(datos1$SEX0=="varon"), "GASTO"]
```

```
## Eliminamos los valores perdidos de la muestra.  
gasto.varon<-na.omit(gasto.varon)
```



Test para la comparación de medias en poblaciones independientes



```
## Test t bilateral de comparación de medias de 2 muestras
## independientes asumiendo igualdad de varianzas

t.test(gasto.mujer, gasto.varon, alternative="two.sided", mu=0,
var.equal=T, conflevel=0.95)
```

Two Sample t-test

```
data: gasto.mujer and gasto.varon
t = -0.47231, df = 127, p-value = 0.6375
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8728396  0.5364620
sample estimates:
mean of x mean of y
 9.928857 10.097046
```

Contraste bilateral

$H_0 : m_1 = m_2 \quad \Rightarrow \quad m_1 - m_2 = 0$

$H_1 : m_1 \neq m_2 \quad \Rightarrow \quad m_1 - m_2 \neq 0$

Ejemplo de test para comparación de m



Resumiendo

Two Sample t-test

data: gasto.mujer and gasto.varon (**muestra1 y muestra2**)

t = -0.47231 (**estadístico¹**) df = 127 (**grados de libertad**) p-value = 0.6375

P-valor

alternative hypothesis: true difference in means is not equal to 0
(**H1 : m1 - m2 \neq 0 euros**)

95 percent confidence interval:

-0.8728396 0.5364620 **Intervalo de confianza de la diferencia de medias**

sample estimates: **Algunas estimaciones de las muestras, como sus medias**

mean of x mean of y

9.928857 10.097046

¹Estadístico t, pero si la muestra es suficientemente grande (mayor que 30) como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra.



Ejemplo de test para comparación de m



Resumiendo

Contraste bilateral

$$H_0 : m_1 = m_2 \quad \text{✉} \quad m_1 - m_2 = 0$$

$$H_1 : m_1 \neq m_2 \quad \text{✉} \quad m_1 \neq m_2 \neq 0$$

- Resolución del contraste mediante IC al 95%:

$$IC_{m_1 - m_2}^{95\%} = [-0.8728396, 0.5364620]$$

$0 \in IC_{m_1 - m_2}^{95\%}$ ✉ La hipótesis de que las medias son iguales es aceptable

- Resolución mediante el p-valor:

$(p\text{-value} = 0.6375) \gg 0,1$ ✉ No hay evidencia contra la H_0

¡ Conclusión: la Hipótesis de que el gasto de hombre y mujeres es el mismo es aceptable!

Test para la comparación de medias en poblaciones apareadas



```
### Ejemplo de comparación del gasto total por cliente en una tienda on line (X) entre el gasto en la primera visita a la web o primera compra (gasto1) y el gasto en la segunda (gasto2).
```

```
### Comparación de medias de poblaciones apareadas: se comparan dos características para un mismo individuo de la población.
```

```
### Variables "GASTO" y "GASTO2". Data set "datos1". Archivo "datos.RData"
```

```
> ## Contraste BILATERAL. Nivel de confianza 99%
```

```
> ## Asumiendo homocedasticidad
```

Contraste bilateral

$H_0 : m_d = 0$

$H_1 : m_d \neq 0$

```
> t.test(gasto1, gasto2, alternative="two.sided",  
mu=0, paired=T, var.equal = T, conf.level=0.99)
```

Ejemplo de test para comparación de m apareadas

Resumiendo



Paired t-test

data: gasto1 and gasto2 (**muestra1 y muestra2**)

t = -0.055973(**estadístico¹**) df = 128(**grados de libertad**) p-value = 0.9555

P-valor

alternative hypothesis: true difference in means is not equal to 0
(**H1 : md \neq 0 euros**)

99 percent confidence interval:

-0.6842853 0.6556032 **Intervalo de confianza de la diferencia de medias**
99%

sample estimates: **Algunas estimaciones de las muestras, como la media de las diferencias entre el gasto de la primera y la segunda compra**

mean of the differences

-0.01434109

¹Estadístico t, pero si la muestra es suficientemente grande (mayor que 30) como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra.

Ejemplo de test para comparación de m apareadas

Resumiendo



Contraste bilateral

$$H_0 : m_1 = m_2 \quad \text{✉} \quad m_1 - m_2 = 0$$

$$H_1 : m_1 \neq m_2 \quad \text{✉} \quad m_1 \neq m_2 \neq 0$$

- Resolución del contraste mediante IC al 95%:

$$IC_{m_1 - m_2}^{95\%} = [-0.8728396, 0.5364620]$$

$\varpi \in IC_{m_1 - m_2}^{95\%}$ ✉ La hipótesis de que las medias son iguales es aceptable

- Resolución mediante el p-valor:

(p-value = 0.6375) >> 0,1 ✉ No hay evidencia contra la H_0

¡ Conclusión: la Hipótesis de que el gasto en la primera visita a la web (gasto1) y el gasto en la segunda (gasto2) es el mismo es aceptable!

Ejemplo de test para la varianza



R no da con un menú específico el intervalo de confianza para la varianza.

Admitiendo la hipótesis de normalidad en la población de partida, se puede calcular el intervalo de confianza para la varianza programando el código según hemos visto.

Es posible que la varianza teórica del gasto de los clientes sea de 2 euros²?

Contraste bilateral

$H_0 : \sigma^2 = 2$

$H_1 : \sigma^2 \neq 2$

```
> ## IC PARA LA VARIANZA DE UNA DISTRIBUCIÓN EN MUESTRAS  
GRANDES
```

```
> ### Vamos a construir un IC para la varianza del gasto de los  
> ### clientes en la web de dicha tienda
```

```
> ### Variable "GASTO". Data set "datos1". Archivo
```



Ejemplo de test para la varianza



```
> gasto<-na.omit(datos1$GASTO)
> media.gasto<-mean(gasto, na.rm=T)
> S2.gasto<-var(gasto, na.rm=T)
> N.gasto<-length(gasto)
> # valores críticos nivel de confianza 90%. Alfa=0.1. chi 2
con N-1 g.l.
> alfa<-0.1
> nc<-(1-alfa)*100
> gl.gasto<-N.gasto-1
> g1<-qchisq(alfa/2, gl.gasto, lower.tail=T)
> g2<-qchisq(alfa/2, gl.gasto, lower.tail=F)
> li<-(N.gasto-1)*S2.gasto/g2
> ls<-(N.gasto-1)*S2.gasto/g1
> # IC para la varianza del GASTO. Nivel de confianza 90%
> c(li,ls)
[1] 2.940465 4.442245
```



Contraste bilateral

$$H_0 : \sigma^2 = 2$$

$$H_1 : \sigma^2 \neq 2$$

- Resolución del contraste mediante IC al 90%:

$$IC_{\sigma^2}^{95\%} = [2.940465, 4.442245]$$

$$2 \notin IC_{\sigma^2}^{95\%}$$

- ✉ La hipótesis de que la varianza es 2 euros² NO es aceptable

¡ Conclusión: la Hipótesis de que la varianza (variabilidad) del gasto es 2 euros² NO es aceptable!



Test para la comparación de varianzas



```
> ### Ejemplo de comparación de la variabilidad del volumen de  
gasto total por cliente en una tienda on line (X) entre hombres y  
mujeres  
  
> ### Vamos a realizar un test de comparación de varianzas de  
poblaciones independientes.  
  
> ### Variables "SEX0" y "GAST0". Data set "datos1". Archivo  
"datos.RData"  
  
>  
>  
  
> ## Test de F para comparación de varianzas  
  
> var.test(gasto.mujer, gasto.varon, ratio=1,  
alternative="two.sided", conf.level=0.95)
```

Contraste bilateral

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$



Test para la comparación de varianzas



Resumiendo

F test to compare two variances

data: gasto.mujer and gasto.varon (**muestra1 y muestra2**)

F = 0.79919 (**estadístico¹**) num df = 38 (**grados de libertad**) p-value = 0.4305

denom df = 84

P-valor

alternative hypothesis: true ratio of variances is not equal to 1
(**H1 : $\sigma_1^2 / \sigma_2^2 \neq 1$ euros²**)

95 percent confidence interval:

0.4814966 1.3953079 **Intervalo de confianza para el ratio de varianzas al 95%**

¹Estadístico F, pero si la muestra es suficientemente grande como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra.



Test para la comparación de varianzas

Resumiendo



Contraste bilateral

$$H_0 : \sigma^2_1 = \sigma^2_2 \quad \Rightarrow \quad \sigma^2_1 / \sigma^2_2 = 1$$

$$H_1 : \sigma^2_1 \neq \sigma^2_2 \quad \Rightarrow \quad \sigma^2_1 / \sigma^2_2 \neq 1$$

- Resolución del contraste mediante IC al 95%:

$$IC_{\sigma^2_1/\sigma^2_2}^{95\%} = [0.4814966 \ 1.3953079]$$

1 $\in IC_{\sigma^2_1/\sigma^2_2}^{95\%}$ \Rightarrow La hipótesis de que las varianzas son iguales es aceptable

- Resolución mediante el p-valor:

(p-value = 0.4305) \gg 0,1 \Rightarrow No hay evidencia contra la H_0

¡ Conclusión: la Hipótesis de que la variabilidad en el gasto es la misma entre hombres y mujeres es aceptable!

Ejemplo de test para p



```
#### Sea X la variable definida como "nº de clientes  
que compren en la primera visita a la web" de una  
determinada tienda on line.
```

```
####  $X \sim \text{Binomial}(n, p)$ 
```

```
#### Vamos a realizar un test sobre la proporción (P)  
de clientes que compren en la primera visita a la web  
de dicha tienda
```

```
#### Variable "COMPRA1". Data set "datos1". Archivo  
"datos.RData"
```

**¿Es admisible que la proporción de clientes que
compran en la primera visita sea del 5%?**



Ejemplo de test para p



```
# TEST de hipótesis, H0: p=0.05. H1: p<>0.05. Nivel de  
confianza 90%
```

```
prop.test(frec.compra1["si"], N.compra1, p=0.05,  
alternative="two.sided", conf.level=0.9, correct=F)
```

1-sample proportions test without continuity correction

data: frec.compra1["si"] out of N.compra1, null probability 0.05

X-squared = 6.6858, df = 1, p-value = 0.009718

alternative hypothesis: true p is not equal to 0.05

90 percent confidence interval:

0.06405002 0.15064223

sample estimates:

p

0.09923664

Contraste bilateral

H0 : P = 0,05

H1 : P ≠ 0,05

Ejemplo de test para p

Resumiendo



1-sample proportions test without continuity correction

data: freq.compra1["si"] out of N.compra1, null probability 0.05

X-squared¹ = 6.6858, df = 1, p-value = 0.009718

P-valor

Valor de p a contrastar (5%)

alternative hypothesis: true p is not equal to 0.05

90 percent confidence interval:

0.06405002 0.15064223 Intervalo de confianza para el p al 90%

sample estimates: **Proporción observada de clientes que compran en la 1ª visita**

p

0.09923664

¹Estadístico χ^2 , (no paramétrico), también podría usarse el **binom.test**, pero si la muestra es suficientemente grande (mayor que 30) como en Big Data, su valor será muy parecido al estadístico z (Normal), tanto más parecido cuanto más grande sea la muestra. Podríamos usar el **t.test**. Sin embargo, si N es grande, las diferencias entre los tests es pequeña, así usamos **prop.test** (χ^2) que nos sirve para la comparación también.



Ejemplo de test para p

Resumiendo



Contraste bilateral

$H_0 : P = 0,05$

$H_1 : P \neq 0,05$

- Resolución del contraste mediante IC al 95%:

$$IC_P^{90\%} = [0.06405002, 0.15064223]$$

0,05 $\notin IC_P^{90\%}$



La hipótesis de que la P es 0,05 NO es aceptable

- Resolución mediante el p-valor:

$0,001 < (\text{p-value} = \mathbf{0.009718}) < 0,01$



Fuerte evidencia contra la H_0

¡ Conclusión: la Hipótesis de que la proporción clientes que compran en la 1ª visita es del 5% NO es aceptable!

Ejemplo de test para r



```
# TEST de hipótesis, H0: r=0 H1: p<>0 Nivel de  
confianza 95% para ESTATURA y PESO
```

```
cor.test(ESTATURA, PESO, use = "pairs", conf.level =  
0.95)
```

Pearson's product-moment correlation

data: ESTATURA and PESO

t = 11.916, df = 122, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6397976 0.8055378

sample estimates:

cor

0.7333834

Contraste bilateral

H0 : r = 0

H1 : r ≠ 0

Ejemplo de test para r

Resumiendo



Pearson's product-moment correlation

data: ESTATURA and PESO

t = 11.916, df = 122, p-value < 2.2e-16

P-valor

alternative hypothesis: true correlation is not equal to 0

90 percent confidence interval:

0.6397976 0.8055378 Intervalo de confianza para el p al 95%

sample estimates: valor observado del coeficiente de correlación

cor

0.7333834

Ejemplo de test para r

Resumiendo



Contraste bilateral

$H_0 : r = 0$

$H_1 : r \neq 0$

- Resolución del contraste mediante IC al 95%:

$$IC_r^{95\%} = [0.6397976, \quad 0.8055378]$$

0 $\notin IC_P^{90\%}$  La hipótesis de que la r es 0 NO es aceptable

- Resolución mediante el p-valor:

(p-value = **0**) < 0,001  Fuerte evidencia contra la H_0

¡ Conclusión: la Hipótesis de que la correlación (Pearson) entre ESTATURA y PESO es 0 NO es aceptable!

Ejemplo 2 de test para r



```
# TEST de hipótesis, H0: r=0 H1: p<>0 Nivel de  
confianza 95% para ACCESOS y GASTO
```

```
cor.test(ACCESOS, GASTO, use = "pairs", conf.level =  
0.95)
```

Pearson's product-moment correlation

data: ACCESOS and GASTO

t = 1.6711, df = 122, p-value = 0.09727

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.02744955 0.31753261

sample estimates:

cor

0.1495911

Contraste bilateral

H0 : r = 0

H1 : r ≠ 0

Ejemplo 2 de test para r

Resumiendo



Pearson's product-moment correlation

data: ACCESOS and GASTO

t = 1.6711, df = 122, p-value = 0.09727

P-valor

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.02744955 0.31753261

sample estimates:

cor

0.1495911



Ejemplo 2 de test para r

Resumiendo



Contraste bilateral

H0 : r = 0

H1 : r ≠ 0

- Resolución del contraste mediante IC al 95%:

$$IC_r^{95\%} = [-0.02744955, 0.31753261]$$

0 $\in IC_P^{90\%}$ ✉ La hipótesis de que la r es 0 es aceptable

- Resolución mediante el p-valor:

0,005 < (p-value = **0.09727**) < 0,01 ✉ Poca evidencia contra la H0

¡ Conclusión: la Hipótesis de que la correlación (Pearson) entre GASTO y ACCESOS es 0 es aceptable!

¿Qué es y para qué se usa la inferencia?

Resumiendo

- **Consiste en:** inferir el modelo probabilístico que ha generado los datos a partir de las frecuencias observadas de la variable.
- **Se usa para:**
 - **Describir** una variable o las relaciones entre un conjunto de éstas (**muestreo**).
 - **Contrastar** relaciones entre las variables (**diseño**)
 - **Predecir** valores esperados de una variable o un conjunto de ellas.
- Uno de los problemas básicos de la **Inferencia Estadística** es el de obtener **conclusiones** sobre la pauta de variabilidad de una variable aleatoria en una determinada **población**, a partir de la información contenida en una **muestra** aleatoria de individuos de dicha población, así como medir su **significación**, esto es, la confianza que nos merecen.

Test de hipótesis e IC

Resumiendo

Paso 1

- Conocimiento del problema y objetivos

Paso 2

- Plantear H_0 y H_1

Paso 3

- Preparar el experimento y tomar la muestra **Muestreo, preproceso**

Paso 4

- Analizar la muestra: descriptivo (normalidad, datos anómalos,...)

Paso 5

- Identificar el tipo de prueba (estadísticos)

Paso 6

- Determinar nivel de significación α

Nivel de confianza vs
nivel de significación, ...

Paso 7

- Realizar la prueba (t, F, ...) y/o obtener IC

t.test(), prop.test, ... 

Paso 8

- Decidir: Aceptar o Rechazar hipótesis

P-valor

Paso 9

- Interpretar resultados y redactar conclusiones

- Los estadísticos que aparecen en el apartado de los IC y/o en el de los tests de hipótesis son los mismos para un determinado parámetro o comparación de éstos.
- También son los mismos sea el contraste bilateral o unilateral, lo que cambia es la forma de la región crítica y la interpretación de la decisión del test o IC
- En este curso sólo hemos visto los más importantes y utilizados, pero hay otros muchos, aunque la lógica, tanto de la construcción de los IC, como de la lógica de los contrastes es la misma. Lo que cambia es el estadístico.

Aceptar o Rechazar una hipótesis

Resumiendo

- El hecho de aceptar o rechazar una hipótesis no significa que ésta se CIERTA o FALSA, respectivamente.
- En vez de aceptar o rechazar, deberíamos hablar de la probabilidad de haber obtenido los datos que hemos obtenido en la muestra asumiendo que la H_0 es cierta.
- Nunca estaremos completamente seguros de si una hipótesis es cierta o falsa.

La mayoría de los tests paramétricos (test t, ANOVA, etc.) se basan en una serie de suposiciones:

- **NORMALIDAD**: las variables deben seguir una distribución Normal en cada uno de los grupos que se comparan. Aunque puede asumirse que se cumple para muestras grandes ($n > 100$), debe explorarse siempre, con gráficos y/o pruebas de normalidad.
- **HOMOCEDASTICIDAD**: en el caso de comparación de poblaciones, las varianzas de dichas poblaciones deben ser iguales. Los tests basados en esta hipótesis son menos exigentes respecto de su incumplimiento, aunque hay tests específicos.

Big Data

- **INDEPENDENCIA**: es la más difícil de verificar y quizás la más importante. La única forma de evitarlo es conseguir que la muestra sea representativa, seleccionada al azar (muestra aleatoria simple), o por un procedimiento equivalente a juicio de los expertos en el proceso.

- En ocasiones es difícil aceptar la hipótesis previa de que los datos son normales al disponerse, por ejemplo, de distribuciones muy asimétricas.
- En estos casos los contrastes anteriores no detectan claras diferencias en el comportamiento de las poblaciones, debido a que la dispersión es muy grande o debido a que la medida de tendencia central utilizada (la media) no es la correcta porque está afectada por los valores extremos.
- Los contrastes paramétricos descritos antes son especialmente sensibles a valores extremos de la variable.

Test no paramétricos

- Para solucionar el problema se utiliza la mediana en lugar de la media construyéndose los que se denominan **contrastes no paramétricos** al no referirse ya a parámetros de una distribución concreta.
 - Comparación de medianas de dos poblaciones con datos independientes: el contraste U de Mann-Withney
 - Comparación de medianas de dos poblaciones con datos apareados: el test de Wilcoxon

- En el caso de comparación de múltiples poblaciones, podríamos pensar en realizar el contraste para la comparación de medias 2 a 2.
- En general, la práctica de analizar los resultados de este tipo de experimentos comparando 2 a 2 (mediante las técnicas ya vistas) todas las parejas posibles de tratamientos no es recomendable:
 - es muy laboriosa
 - incrementa la probabilidad global de cometer un error de 1ª especie




- **Análisis de la Varianza ANOVA:** Técnica estadística muy poderosa para el análisis de observaciones que dependen, o pueden depender, simultáneamente de uno o más factores.

- Tablas de contingencia ✉ test de proporciones
- Regresión ✉ test de medias ✉ parámetros de regresión
- -----

La lógica de los contrastes es la misma

Métodos de inferencia



- **Paramétricos**: se supone que los datos provienen de una distribución que puede caracterizarse por un grupo de parámetros (m , σ , λ , p , ...) que pueden estimarse a partir de éstos.  se supone que la forma de la distribución es conocida (Normal, Poisson, ...). Cuando los datos analizados cumplen las asunciones para la aplicación de los tests paramétricos **es preferible usarlos SIEMPRE**, ya que son más potentes, en el sentido que tienen mayor capacidad para rechazar la hipótesis nula cuando ésta es falsa.
- **No paramétricos**: se suponen aspectos muy generales de la distribución (continua, simétrica, ...) y tratan de estimar su forma o estructura. Dentro del enfoque paramétrico estas pruebas suelen usarse para contrastar hipótesis sobre la forma de la distribución (*test de normalidad*, ...)
- **Bayesiana**



Métodos de inferencia



Cuando los datos con los que tenemos que trabajar:

- Son variables nominales u ordinales (cualitativas o discretas con pocos valores)
- Se incumple alguna hipótesis de los tests paramétricos (no normalidad, ...)
- Las transformaciones en las variables (logarítmica, ...) o eliminación de datos extremos no son una solución al incumplimiento de estas hipótesis



Hay que recurrir a los tests no paramétricos
Generalmente son tests o contrastes



Glosario

Aceptar H_0	Modelo
Ajuste	Muestreo
Chi-2	Nivel de Confianza
Distribución muestral	Nivel de significación
Error Estándar	Potencia
Estadístico	Predicción
Estimación	P-valor
Estimación puntual	Rechazar H_0
F de Snedecor	Región de aceptación
Grados de Libertad	Región de rechazo
Heterocedasticidad	Riesgo de 1ª especie alfa
Hipótesis Alternativa H_1	t de Student
Hipótesis Nula H_0	Tamaño muestra
Homocedasticidad	Test bilateral
Inferencia	Test o contraste de hipótesis
Inferir	Test unilateral
Intervalo de Confianza	Valor crítico
Métodos paramétricos	

Herramientas Estadísticas para Big Data

Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

3- Inferencia en muestras grandes



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

www.upv.es

E. Vázquez
Dto. De Estadística e Investigación Operativa,
Aplicadas y Calidad