Introducción a R en investigación biomédica

Sesión 2: **Introducción a la bioestadística en R**

Unidad de Apoyo Metodológico, IIS Biogipuzkoa
Jone Renteria Aguirregabiria & Lore Zumeta-Olaskoaga

Donostia, 24 y 26 de noviembre de 2025

Contenido

Introducción a la bioestadística en R

A. Inferencia estadística en R

- Contrastes de hipótesis y el p-valor
- Tests estadísticos para la comparación de variables cuantitativas y categóricas
- Casos de uso en R

B. Introducción a la modelización estadística en R

- ¿Qué es un modelo?
- Regresión lineal simple/múltiple
- Inferencia sobre los parámetros de regresión
- Casos de uso en R

Introducción a la bioestadística - ¿Qué es?

- **Estadística**: ciencia con base matemática que se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, encontrar patrones y analizarlos, así como hacer inferencias a partir de ellos. Consiste en:
  - evaluar la incertidumbre de las estimaciones y cómo esa incertidumbre afecta a las inferencias y la toma de decisiones,
  - dar sentido a los datos sin engañarnos a nosotros mismos en el proceso.

- **Bioestadística**:
  - Estadística aplicada a problemas biomédicos.
  - Toma de decisiones (fundamentadas) ante la incertidumbre o la variabilidad.
  - Intento de eliminar sesgos o encontrar explicaciones alternativas a las planteadas por investigadores con intereses creados.
  - Diseño experimental, medición, descripción, gráficos estadísticos, análisis de datos, inferencia, predicción…

Introducción a la bioestadística - ¿Qué es?

- La (bio)estadística no es una caja de herramientas y fórmulas matemáticas, sino una forma de pensar basada en la evidencia.

- Es muy importante:
  - comprender el problema;
  - formular correctamente la pregunta para abordarlo;
  - comprender y optimizar las mediciones;
  - comprender las fuentes de variabilidad;
  - y mucho más.

- "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H.G. Wells (1866-1946)

- MacKay & Oldford (2000) desarrollaron una representación en 5 etapas del **método estadístico aplicado a la investigación científica**: Problema, Plan, Datos, Análisis, Conclusión.

Ciencia Estadística - Enfoque *Resolución de Problemas*

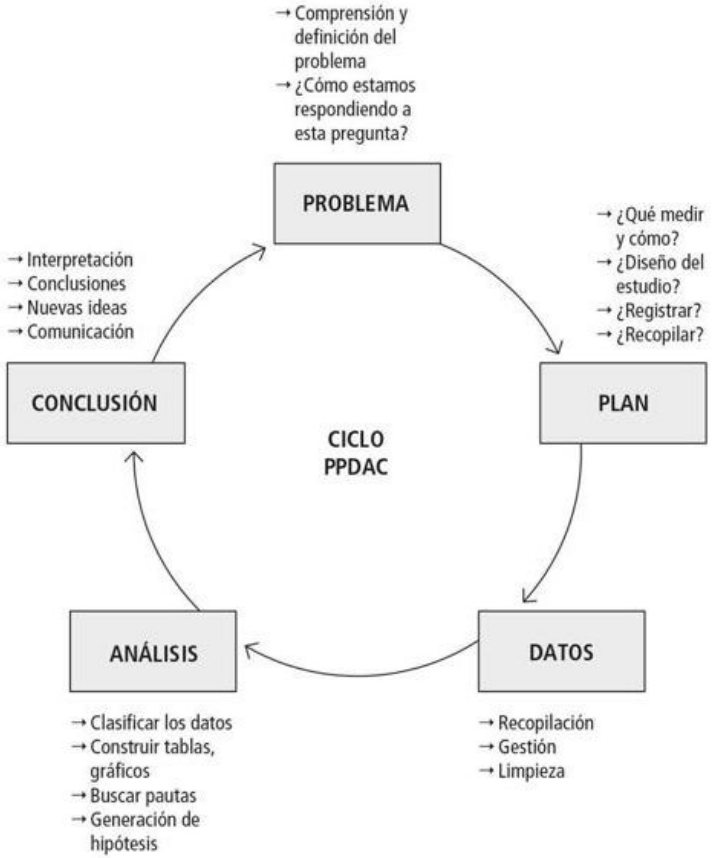
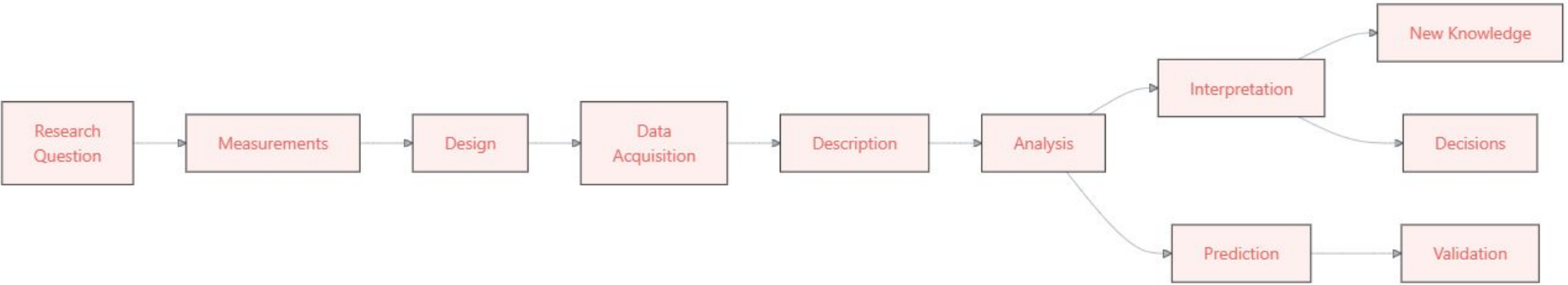


Figura 0.3. Ciclo PPDAC de resolución de problemas, que va del Problema, el Plan, los Datos, el Análisis, a las Conclusiones y la comunicación, y vuelta a empezar en un nuevo ciclo.

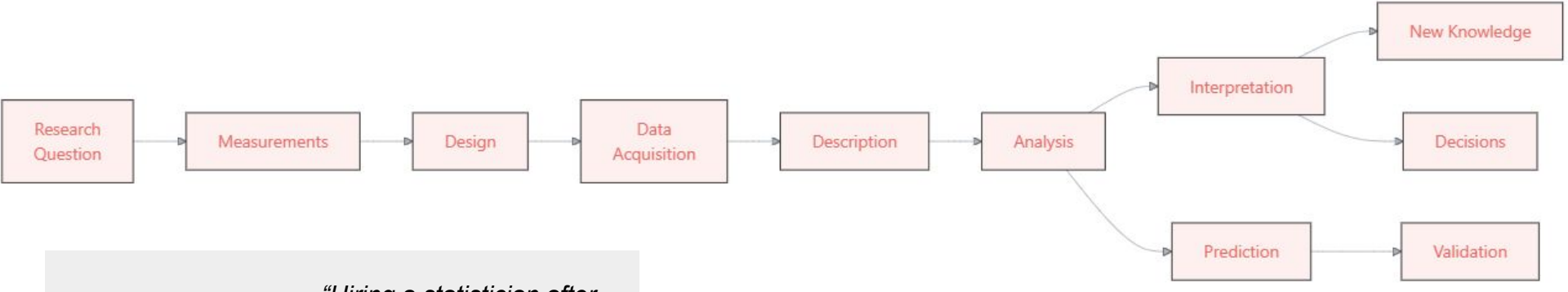
Fuente: *The Art of Statistics*, David Spiegelhalter

Ciencia Estadística - Método científico



Fuente: F. Harrell, Biostatistics for Biomedical Research

Ciencia Estadística - Método científico



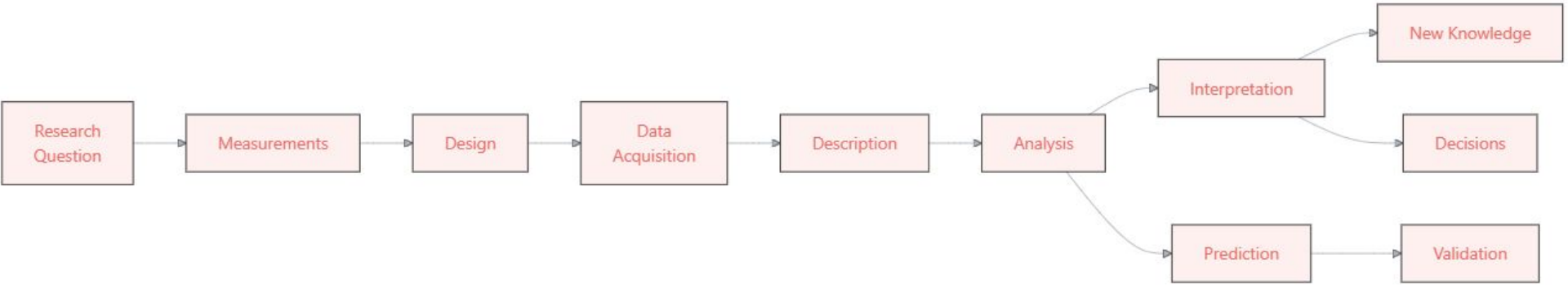
Fuente: F. Harrell, Biostatistics for Biomedical Research

F. Yates, R. Fisher & W. Cochran
Fuente: Wikipedia Commons

"Hiring a statistician after the data has been collected is like hiring a physician when a patient is in the morgue: she might be able to tell you what went wrong, but she is unlikely to be able to fix it"

Ronald Fisher, Statistician and Biologist

Ciencia Estadística - Método científico



Fuente: F. Harrell, Biostatistics for Biomedical Research

"The best thing about being a statistician is that you get to play in everyone's backyard"

Fuente: Wikipedia

John Tukey, Mathematician and Statistician

Bloque A
Inferencia estadística en R

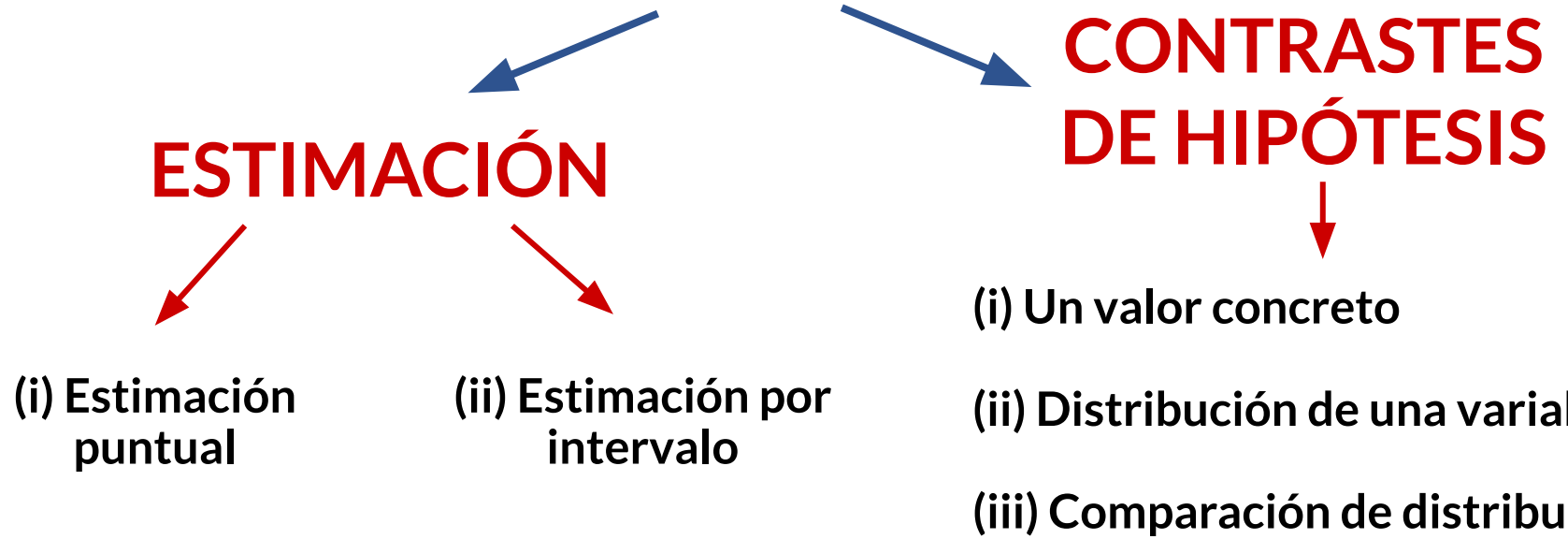
Inferencia Estadística

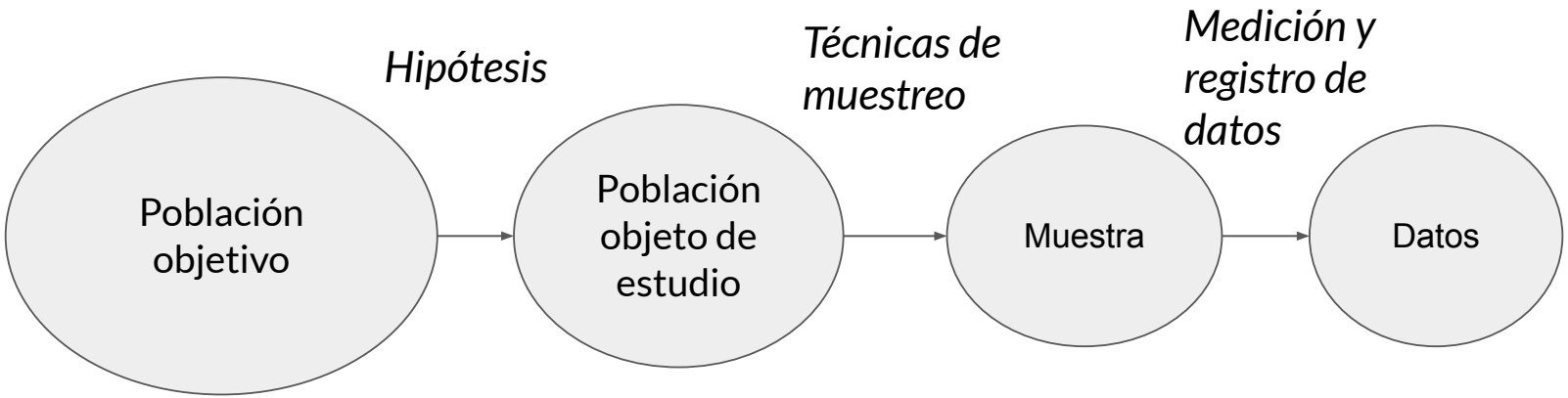
Definición: el conjunto de métodos y técnicas que permiten **inducir**, a partir de la información empírica proporcionada por una muestra, cuál es el **comportamiento de una determinada población** con un **riesgo de error** medible en términos de probabilidad.

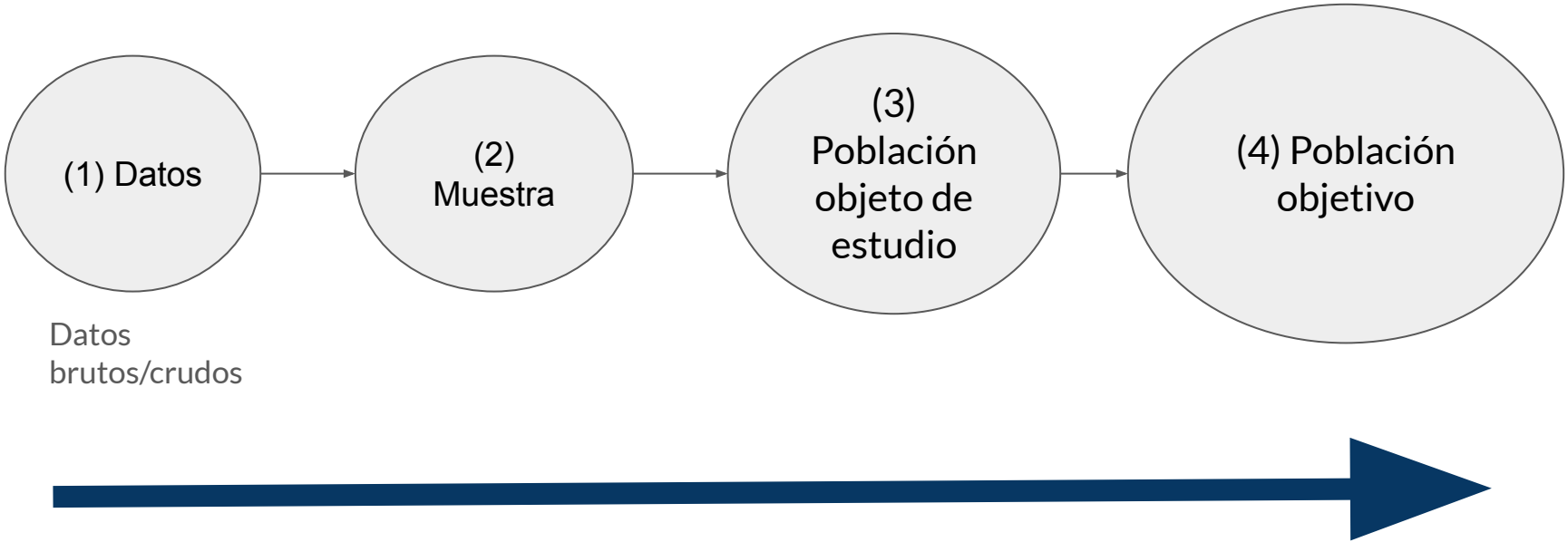
Los métodos básicos de la estadística inferencial son la **estimación** y **el contraste de hipótesis**, que juegan un papel fundamental en la investigación.

Inferencia Estadística

Definición: el conjunto de métodos y técnicas que permiten **inducir**, a partir de la información empírica proporcionada por una muestra, cuál es el **comportamiento de una determinada población** con un **riesgo de error** medible en términos de probabilidad.




Población objetivo
Hipótesis
Población objeto de estudio
Técnicas de muestreo
Muestra
Medición y registro de datos
Datos



Inferencia inductiva

Contraste de hipótesis

- Se emplean los contrastes de hipótesis cuando el objetivo es tratar de corroborar o invalidar una afirmación.

- Se plantea una hipótesis y se utiliza la información proporcionada por la muestra para ver si se "*acepta*" o se rechaza dicha hipótesis.

- Toda hipótesis constituye un juicio, es decir, una afirmación o una negación de algo. Por lo tanto, habrá dos enunciados, donde una hipótesis (un enunciado) contradice a la otra.

- El análisis estadístico de los datos sirve para determinar qué hipótesis "*aceptas*" o cuál rechazas.

H_0: Hipótesis nula **+ conservadora**

H_1: Hipótesis alternativa **Lo que se quiere probar**

Un ejemplo

H_0: En el tratamiento de las lesiones tendinosas, no hay diferencias entre un tratamiento activo o pasivo. **(H_0 o Hipótesis nula)**

H_1: El tratamiento activo es mejor que el pasivo en las lesiones tendinosas. **(H_1 o Hipótesis alternativa)**

Tipos de contrastes de hipótesis

- Un valor concreto

$$\begin{cases} \mathrm{H}_0 : \mu_{\text{peso}} = 65, \\ \mathrm{H}_1 : \mu_{\text{peso}} \neq 65. \end{cases}$$

- Para contrastar la distribución de una variable

$$\begin{cases} \mathrm{H}_0 : \text{peso} \sim \text{Distr. Normal}, \\ \mathrm{H}_1 : \text{peso} \not\sim \text{Distr. Normal}. \end{cases}$$

- Para establecer la igualdad de *medias* entre las distribuciones de dos o más variables

$$\begin{cases} \mathrm{H}_0 : \mu_{\text{peso hombres}} = \mu_{\text{peso mujeres}}, \\ \mathrm{H}_1 : \mu_{\text{peso hombres}} \neq \mu_{\text{peso mujeres}}. \end{cases}$$

Tipos de contrastes de hipótesis

- Un valor concreto

$$\begin{cases} \mathrm{H}_0 : \mu_{\text{peso}} = 65, \\ \mathrm{H}_1 : \mu_{\text{peso}} \neq 65. \end{cases}$$

El promedio (la media) del peso no es igual a 65kg

- Para contrastar la distribución de una variable

$$\begin{cases} \mathrm{H}_0 : \text{peso} \sim \text{Distr. Normal}, \\ \mathrm{H}_1 : \text{peso} \not\sim \text{Distr. Normal}. \end{cases}$$

La variable peso no sigue una distribución normal

- Para establecer la igualdad de *medias* entre las distribuciones de dos o más variables

$$\begin{cases} \mathrm{H}_0 : \mu_{\text{peso hombres}} = \mu_{\text{peso mujeres}}, \\ \mathrm{H}_1 : \mu_{\text{peso hombres}} \neq \mu_{\text{peso mujeres}}. \end{cases}$$

El promedio del peso de los hombres es diferente al promedio del peso de las mujeres

Contraste de hipótesis como *toma de decisiones*

- El contraste de hipótesis trata de evaluar si los datos de la muestra son compatibles con la hipótesis, suponiendo que las muestras difieren de las poblaciones por azar.

- Se decide en función de la información proporcionada por la muestra de tamaño N obtenida de la población.

- Decisión elegida:

<u>Decisión 1</u>:

Rechazar H_0
Aceptar H_1

<u>Decisión 2</u>:

No rechazar (“*aceptar)*” H_0
Rechazar H_1

Contrastes de hipótesis - analogía de juicio penal

$H_0 \rightarrow$ Inocente

$H_1 \rightarrow$ Culpable

4 posibilidades

Verdad	Resultado de la prueba de hipótesis	
	No se rechaza H_0	Rechazamos H_0
H_0 es cierta	Decisión correcta	Error
H_0 es falsa	Error	Decisión correcta (la que buscamos)

Contrastes de hipótesis - analogía de juicio penal

$H_0 \rightarrow$ **Inocente**

$H_1 \rightarrow$ **Culpable**

4 posibilidades

Verdad	**Resultado de la prueba de hipótesis**	
	No se rechaza H_0	**Rechazamos H_0**
H_0 es cierta	**Decisión correcta**	**Error**
H_0 es falsa	**Error**	**Decisión correcta (la que buscamos)**

…ante todo, **prevalece la presunción de inocencia (H_0 es cierta)**

Debemos demostrar que la H_0 no es cierta, porque la inocencia se asume y la culpabilidad hay que demostrarla

		Veredicto	
		No se rechaza la hipótesis nula (el sospechoso es declarado "no culpable")	**Se rechaza la hipótesis nula en favor de la alternativa (el sospechoso es declarado culpable)**
Verdad	**Hipótesis nula (el sospechoso es inocente)**	Se acierta en no rechazar la hipótesis nula. Se acierta en declara a un inocente "no culpable".	Error de tipo I: rechazar incorrectamente la hipótesis nula. Condenar erróneamente a un inocente. "Inocente a la cárcel"
	Hipótesis alternativa (el sospechoso es culpable)	Error de tipo II: no rechazar incorrectamente la hipótesis nula. No condenar a una persona culpable. "Criminal a la calle"	Rechazar correctamente la hipótesis nula. Condenar correctamente a una persona culpable.

Error (Tipo I) =P (Rechazar H_0 | H_0 es cierta)

Error (Tipo II) = P (No rechazar H_0 | H_0 es falsa)

Inocente a la cárcel

Criminal a la calle

¡¡¡ El más importante a evitar !!!

		Veredicto	
		Inocente	**Culpable**
Verdad	**Inocente**	**1 - α** (Probabilidad de retención adecuada)	**Riesgo α** (Error Tipo I)
	Culpable	**Riesgo β** (Error Tipo II)	**1- β** (Poder estadístico)

Error (Tipo I) =P (Rechazar H_0 | H_0 es cierta)

Error (Tipo II) = P (No rechazar H_0 | H_0 es falsa)

Inocente a la cárcel

Criminal a la calle

¡¡¡ El más importante a evitar !!!

		Veredicto	
		Inocente	**Culpable**
Verdad	**Inocente**	**1 - α** (Probabilidad de retención adecuada)	**Riesgo α** (Error Tipo I)
	Culpable	**Riesgo β** (Error Tipo II)	**1- β** (Poder estadístico)

Valores que suelen tomar "α" y "β":

Normalmente, los científicos hacemos uso de de los niveles "α": 0.1, **0.05** y 0.01

Normalmente, el valor de "β" es: **0.20**

Contrastes de hipótesis - analogía de juicio penal

- La aceptación o rechazo de la hipótesis nula depende, en última instancia, de lo que se observe en la muestra.

- Para cada muestra, se dará una estimación a partir de la cual se tomará la decisión:
  si la estimación que calculamos con la muestra difiere demasiado del valor que esperaríamos obtener si la hipótesis H_0 fuese cierta, entonces se rechazará H_0 y, en caso contrario, diremos que no tenemos evidencia suficiente para rechazar la H_0.

- La lógica que guía la decisión es la de mantener la hipótesis nula a no ser que en la muestra haya pruebas contundentes de su falsedad.

Contrastes de hipótesis - analogía de juicio penal

- La aceptación o rechazo de la hipótesis nula depende, en última instancia, de lo que se observe en la muestra.

- Para cada muestra, se dará una estimación a partir de la cual se tomará la decisión: si la estimación que calculamos con la muestra difiere demasiado del valor que esperaríamos obtener si la hipótesis H_0 fuese cierta, entonces se rechazará H_0 y, en caso contrario, diremos que no tenemos evidencia suficiente para rechazar la H_0.

- La lógica que guía la decisión es la de mantener la hipótesis nula a no ser que en la muestra haya pruebas contundentes de su falsedad.

Siguiendo con el símil del juicio, se trataría de mantener la inocencia mientras no haya pruebas claras de culpabilidad.

p-valor en un contraste de hipótesis

Definición: el *"p-valor"* es la probabilidad de obtener un resultado al menos tan extremo como el que hemos obtenido, si la hipótesis nula (y todas las demás asunciones del modelo) fuera realmente cierta.

p-valor en un contraste de hipótesis

Definición: el *"p-valor"* es la probabilidad de obtener un resultado al menos tan extremo como el que hemos obtenido, si la hipótesis nula (y todas las demás asunciones del modelo) fuera realmente cierta.

→ Indica el **grado de incompatibilidad de los datos con la hipótesis** que se está evaluando.

p-valor en un contraste de hipótesis

Definición: el *"p-valor"* es la probabilidad de obtener un resultado al menos tan extremo como el que hemos obtenido, si la hipótesis nula (y todas las demás asunciones del modelo) fuera realmente cierta.

→ Indica el **grado de incompatibilidad de los datos con la hipótesis** que se está evaluando.

Si un p-valor es lo suficientemente pequeño, entonces decimos que los resultados son **estadísticamente significativos**.

- Cuanto **más pequeño** sea el *p-valor* obtenido, entonces: o bien ha ocurrido algo muy sorprendente, o la hipótesis nula no es cierta → más evidente resulta que la hipótesis nula puede ser una asunción inapropiada → Rechazar la H_0

- Cuanto **más grande** sea el *p-valor* obtenido, entonces: no hay evidencia contraria a la hipótesis nula

Ausencia de la evidencia ≠ **Evidencia de la ausencia**

p-valor en un contraste de hipótesis

- Esta probabilidad...
  - Depende del tipo de estudio
  - Depende del tipo de variable
  - Depende de la influencia de otras variables

- Es una probabilidad condicionada.

- Se adopta un proceso paso a paso para tomar la decisión sobre qué hipótesis se considera verdadera.

¿Qué es exactamente el p-valor?
https://www.youtube.com/watch?v=oqS12S60bqY&t=16s

p-value explained according to the ASA statement
https://www.youtube.com/watch?v=7KPOlBgK8Sk

Pasos del contraste de hipótesis

- Establecer la hipótesis nula (H_0).
- Establecer la hipótesis alternativa (H_1).
- Seleccionar la prueba estadística para calcular la probabilidad bajo la hipótesis nula.
- Tomar una muestra y calcular el valor de la prueba.
- Comparar el valor de la prueba con un valor crítico y decidir
  - si se debe rechazar la hipótesis nula
  - o si no hay pruebas suficientes para rechazar la hipótesis nula

Pasos del contraste de hipótesis

- Establecer la hipótesis nula (H_0).
- Establecer la hipótesis alternativa (H_1).
- Seleccionar la prueba estadística para calcular la probabilidad bajo la hipótesis nula.
- Tomar una muestra y calcular el valor de la prueba.
- Comparar el valor de la prueba con un valor crítico y decidir
  - si se debe rechazar la hipótesis nula
  - o si no hay pruebas suficientes para rechazar la hipótesis nula

Un test estadístico que estime algo que, si resulta ser suficientemente extremo, nos llevaría a dudar de la hipótesis nula

p-valor - y la significatividad estadística

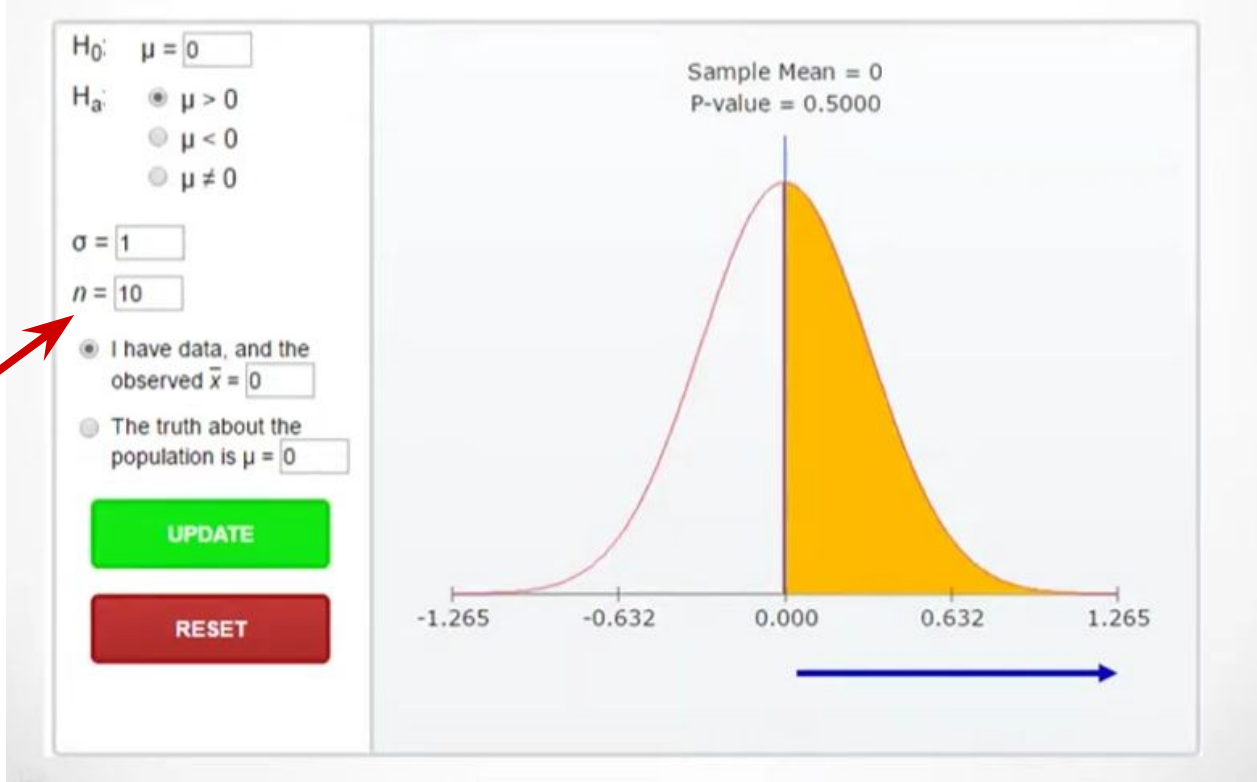
Una vez fijado **el riesgo α**, la regla de decisión para realizar un contraste también puede expresarse de la siguiente manera:

$$\begin{aligned} &\text{Si p-valor} < \alpha \;\rightarrow\; \text{Rechazar } H_0 \\ &\text{Si p-valor} \geq \alpha \;\rightarrow\; \text{No rechazar } H_0 \end{aligned}$$

- **Si p-valor < α**: decimos que los resultados son ***estadísticamente significativos (statistically significant)***

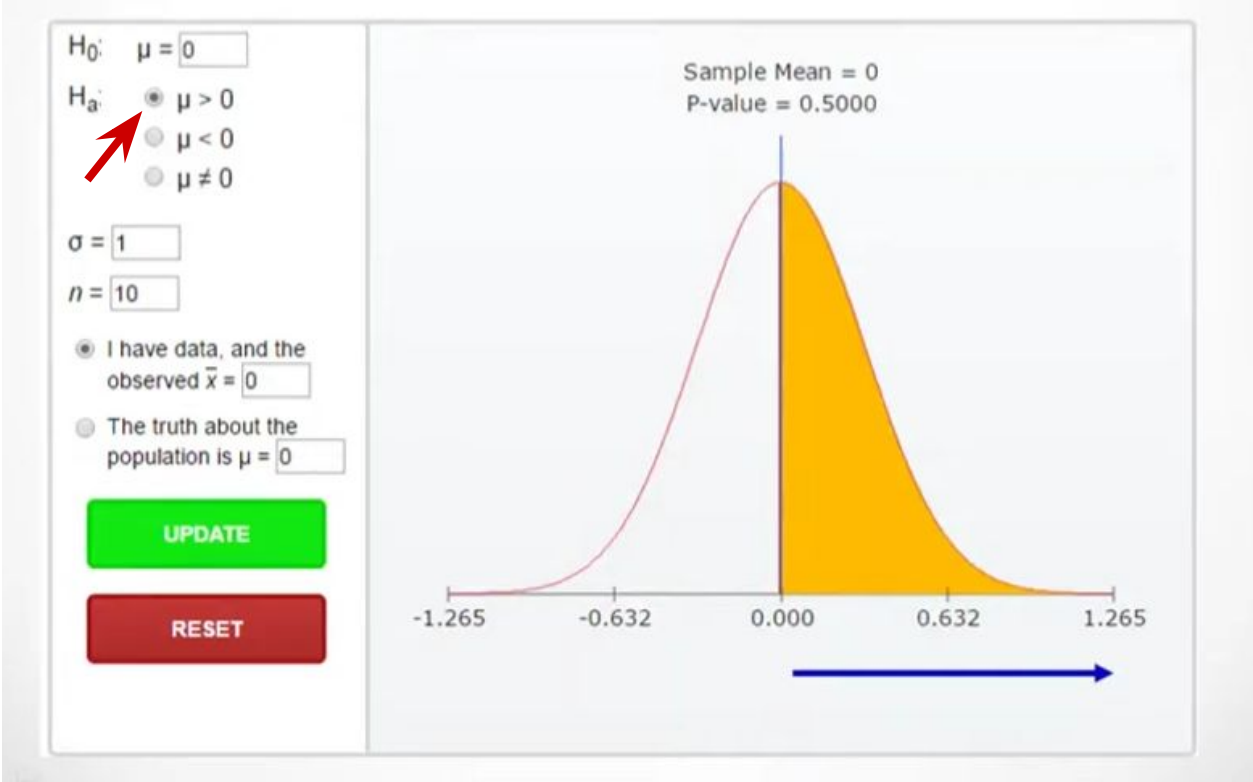
- **Si p-valor ≥ α**: decimos que los resultados ***no son estadísticamente significativos (statistically not significant)***

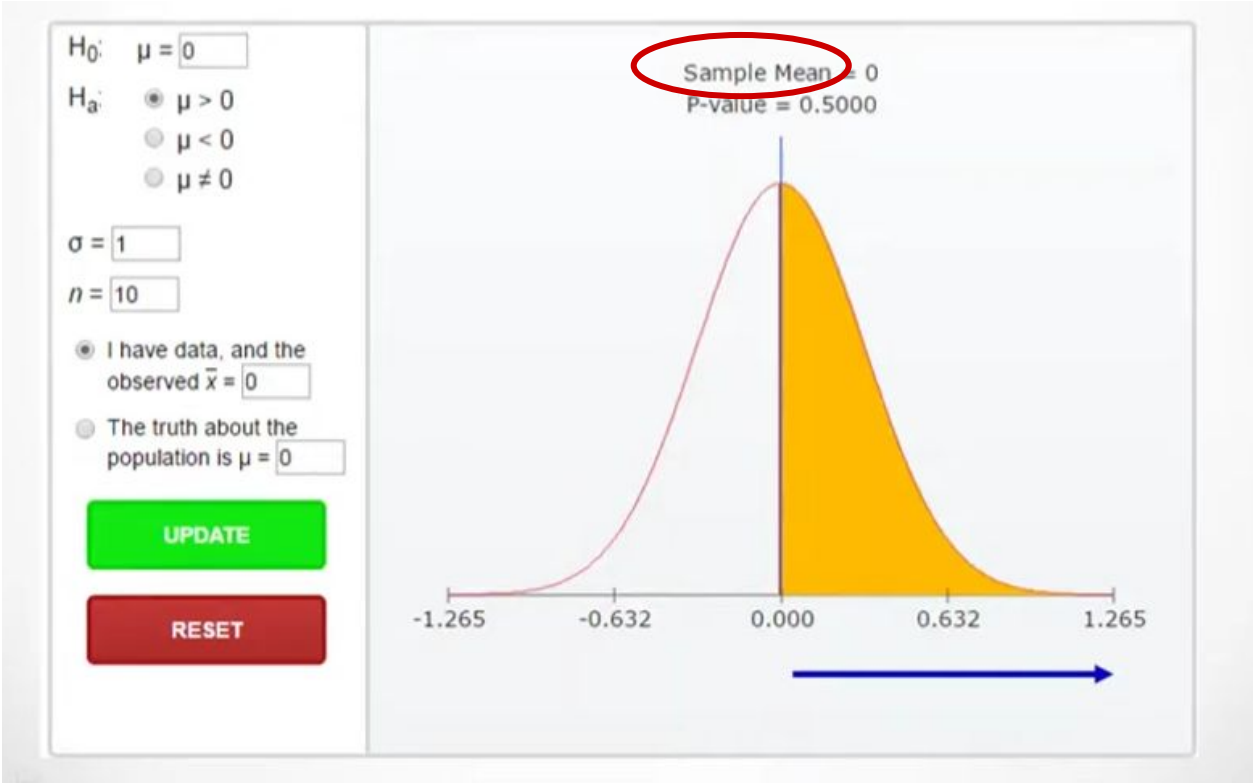
p-value explained according to the ASA statement
https://www.youtube.com/watch?v=7KPOIBgK8Sk

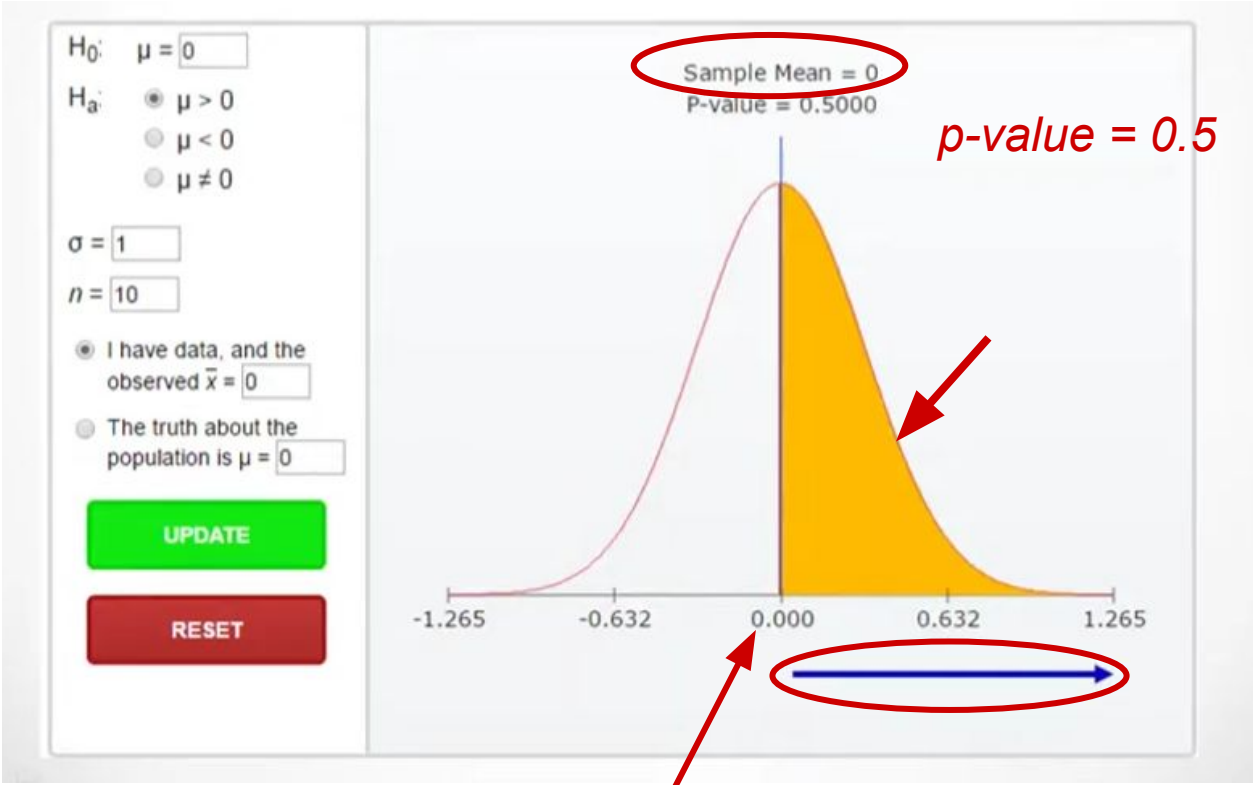


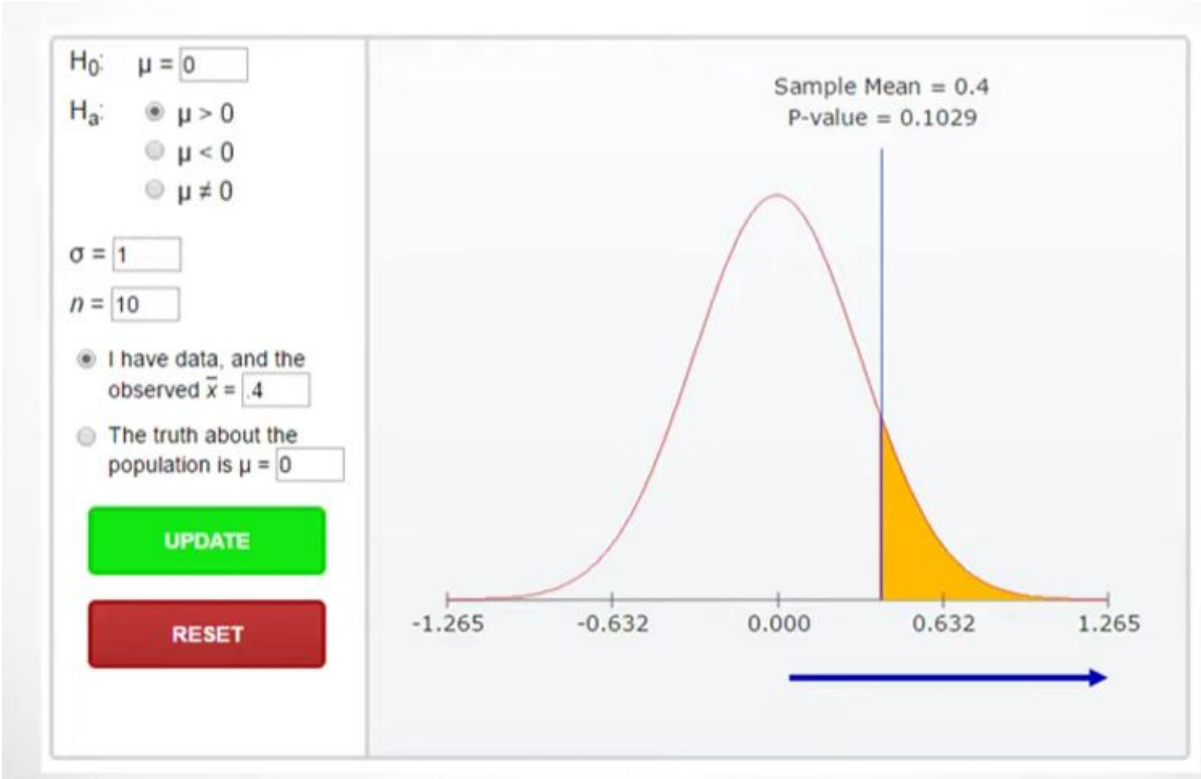
p-value explained according to the ASA statement
https://www.youtube.com/watch?v=7KPOlBgK8Sk

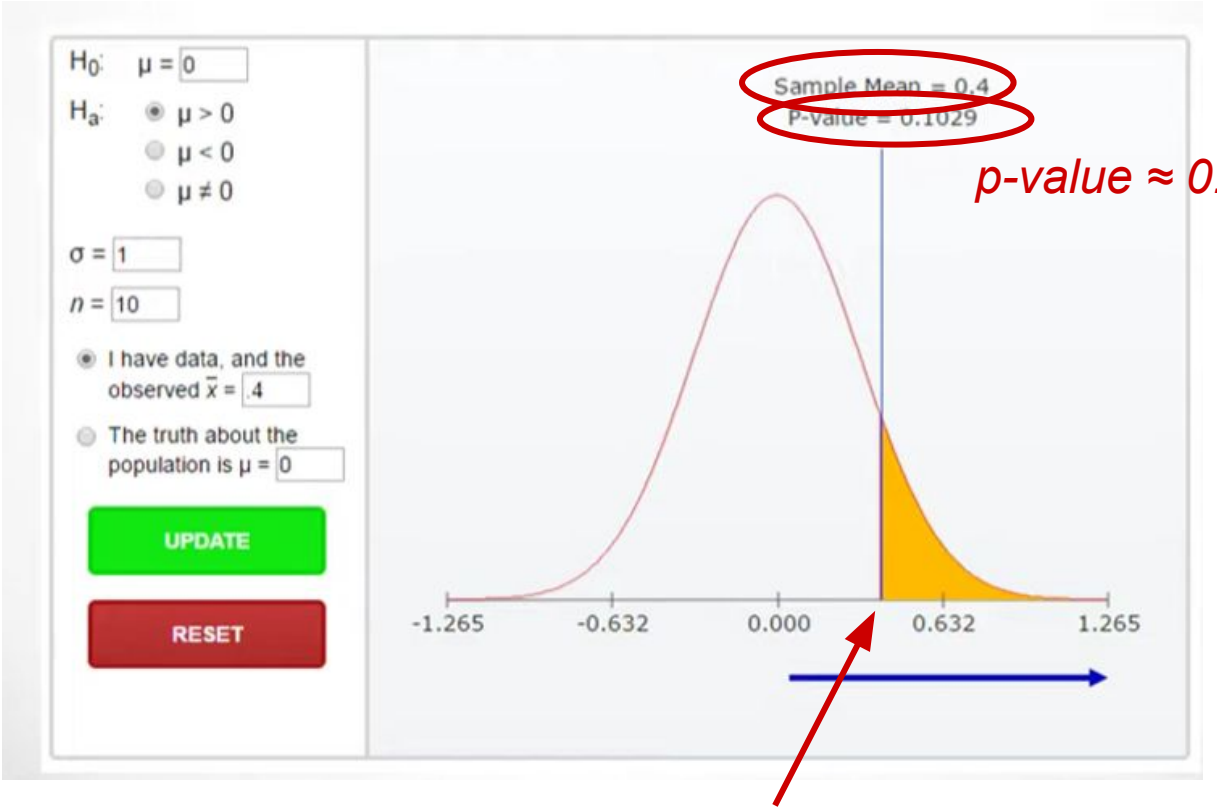

H_0: μ = 0
H_a:
μ > 0
μ < 0
μ ≠ 0
σ = 1
n = 10
I have data, and the observed $\bar{x}$ = 0
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0
P-value = 0.5000
-1.265
-0.632
0.000
0.632
1.265


H0: μ = 0
Ha: μ > 0
μ < 0
μ ≠ 0
σ = 1
n = 10
I have data, and the observed x̄ = 0
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0
P-value = 0.5000
-1.265
-0.632
0.000
0.632
1.265


H0: μ = 0
Ha: μ > 0
μ < 0
μ ≠ 0
σ = 1
n = 10
I have data, and the observed x̄ = 0
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0
P-value = 0.5000
-1.265
-0.632
0.000
0.632
1.265

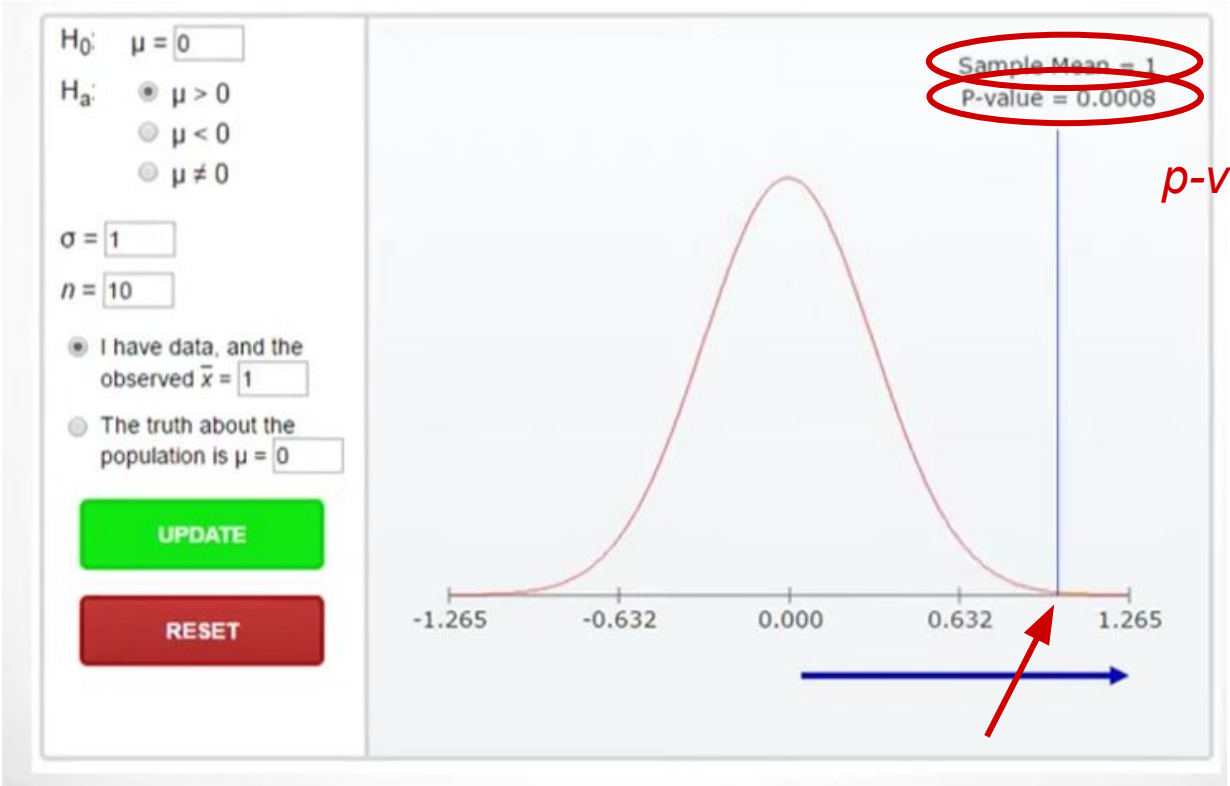

H0: μ = 0
Ha: μ > 0
μ < 0
μ ≠ 0
σ = 1
n = 10
I have data, and the observed x̄ = 0
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0
P-value = 0.5000
p-value = 0.5
-1.265
-0.632
0.000
0.632
1.265


H_0: μ = 0
H_a: $\mu > 0$
$\mu < 0$
$\mu \neq 0$
σ = 1
n = 10
I have data, and the observed $\bar{x}$ = .4
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0.4
P-value = 0.1029
-1.265
-0.632
0.000
0.632
1.265


H0: μ = 0
Ha: μ > 0
μ < 0
μ ≠ 0
σ = 1
n = 10
I have data, and the observed x̄ = .4
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 0.4
P-value = 0.1029
-1.265
-0.632
0.000
0.632
1.265

p-value ≈ 0.10


H_0: μ = 0
H_a: $\mu > 0$
$\mu < 0$
$\mu \neq 0$
σ = 1
n = 10
I have data, and the observed $\bar{x}$ = 1
The truth about the population is μ = 0
UPDATE
RESET
Sample Mean = 1
P-value = 0.0008
-1.265
-0.632
0.000
0.632
1.265



p-value ≈ 0.0008

Caso práctico I

La base de datos `titanic.csv` proporciona información sobre los pasajeros del fatídico viaje inaugural del transatlántico “Titanic”.

`PassengerId`	Passenger ID
`Survived`	Passenger Survival Indicator. 0 = No, 1 = Yes
`Pclass`	Passenger Class, 1=1st , 2=2nd, 3=3rd
`Name`	Name
`Sex`	Sex. Female or Male
`Age`	Age in years
`SubSp`	Number of Siblings/Spouses Aboard
`Parch`	Number of Parents/Children Aboard
`Ticket`	Ticket Number
`Fare`	Passenger Fare
`Cabin`	Cabin
`Embarked`	Port of Emarbakation. One of, Cherbourg, Queenstown or Southampton

Fuente: https://www.kaggle.com/c/titanic/data
Disponible en **R**: `Titanic::titanic train`
Nota: `titanic.csv` ligeramente modificado

Caso práctico I

La base de datos `titanic.csv` proporciona información sobre los pasajeros del fatídico viaje inaugural del transatlántico “Titanic”.

- La edad media de los pasajeros/as y tripulación del Titanic fue de 35 años.
- Los/as pasajeros/as que sobrevivieron pagaron, en promedio, tickets más caros que los que no sobrevivieron.
- La tipología de pasajeros/as (viajeros/as solos, parejas y familias) difirió según el puerto de embarque.
- Las mujeres tuvieron mayores tasas de supervivencia.
- Los niños/as tuvieron mayores tasas de supervivencia.

Fuente: https://www.kaggle.com/c/titanic/data
Disponible en **R**: `Titanic::titanic train`
Nota: `titanic.csv` ligeramente modificado

Tipos de contrastes de hipótesis

- **Tests paramétricos**: se supone que la variable objeto de estudio sigue una distribución concreta y se contrastan los valores de los parámetros.
  - La distribución de la recurrencia del cáncer de mama sigue una distribución binomial.
  - La media de edad es igual entre las pacientes que reciben terapia hormonal y las que no (asumiendo normalidad o simetría en la variable).
  - ...

- **Tests no-paramétricos**: no se asume ninguna distribución para los datos. Las pruebas se refieren a la distribución, no a sus parámetros (p.ej. se refieren a la forma de la distr., basándose en estadísticos de posición).
  - La supervivencia en el Titanic no está relacionada con la clase en la que viajaban los/as pasajeros.
  - La distribución del número de nodos es la misma entre las pacientes que sobrevivieron y las que no.

Tipos de contrastes de hipótesis

- Test de normalidad
- Test de una media de una población
- Test de dos medias de dos muestras independientes
- Test de más de dos medias de muestras independientes
- Test de independencia o asociación
- …

Tipos de contrastes de hipótesis

- **Test de normalidad**
- Test de una media de una población
- Test de dos medias de dos muestras independientes
- Test de más de dos medias de muestras independientes
- Test de independencia o asociación
- ...

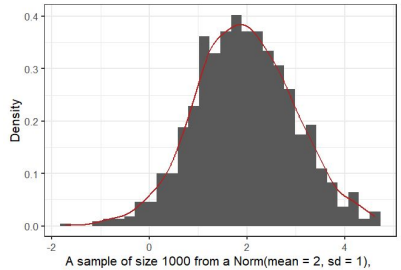
Tipos de contrastes de hipótesis

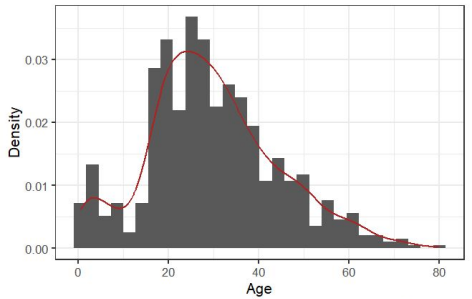
Test de normalidad

- Algunas pruebas paramétricas asumen que los datos provienen de una población normal. ¿Cómo podemos verificar esta suposición? ¿Qué podemos hacer si la suposición es falsa?
- Podemos usar métodos gráficos (histogramas, q-q plots) o contrastes de hipótesis.
- Los tests más usados son: Kolmogorov-Smirnov y Shapiro-Wilks test.
- El contraste se formula:
  - H_0: Los datos siguen una distribución normal
  - H_1: Los datos **NO** siguen una distribución normal
- En **R:**

```
shapiro.test(x = my_var)
```

donde `x` es un vector numérico de valores de datos.





Tipos de contrastes de hipótesis

- Test de normalidad
- **Test de una media de una población**
- Test de dos medias de dos muestras independientes
- Test de más de dos medias de muestras independientes
- Test de independencia o asociación
- ...

Tipos de contrastes de hipótesis

Test de una media de una población

- No lo utilizamos muy a menudo.
- Muy similar a las preguntas de estimación. Se puede resolver calculando un intervalo de confianza.
- Queremos verificar a partir de una muestra una hipótesis previa sobre la media de una población. $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$
- Se supone que la muestra se extrae de una población en la que los valores se distribuyen normalmente (en realidad, la normalidad no es necesaria).
- El test a utilizar se llama **t-test** y se basa en el *estadístico t*:

$$t = \frac{\overline{x} - \mu_0}{SE}$$

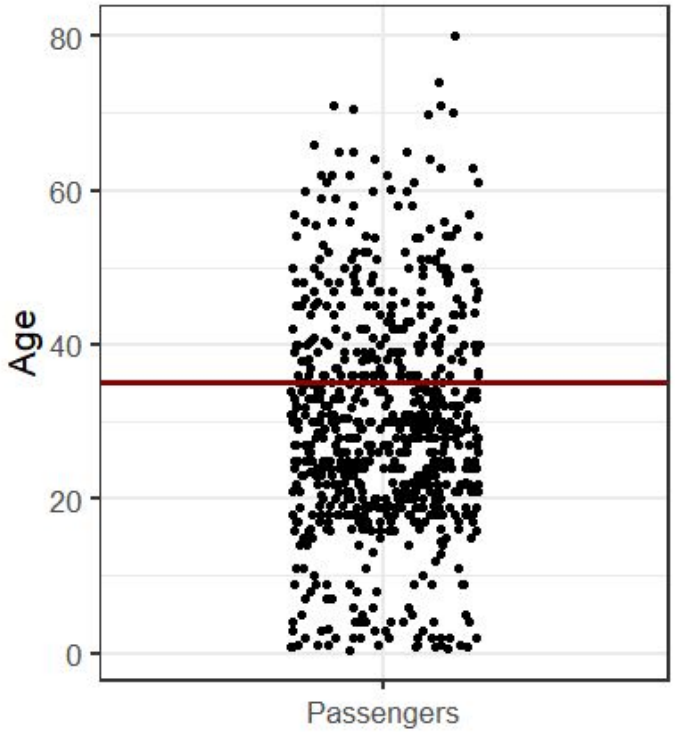
- En **R**:

```
t.test(x = my_var, mu = mu0,
       conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"))
```

Test para la media de una población

La edad media de los pasajeros y tripulación del Titanic fue de 35 años.

$$\begin{cases} \mathrm{H}_0 : \mu = 35 \text{ años}, \\ \mathrm{H}_1 : \mu \neq 35 \text{ años}. \end{cases}$$



Nota: *asumiremos que la variable* `Age` *sigue una distribución normal.*

Test para la media de una población

La edad media de los pasajeros y tripulación del Titanic fue de 35 años.

```
t_res <- t.test(titanic$Age, mu = 35)
t_res
```

```
##
##     One Sample t-test
##
## data:  titanic$Age
## t = -9.7507, df = 713, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 35
## 95 percent confidence interval:
##  28.63179 30.76645
## sample estimates:
## mean of x
##  29.69912
```

Test para la media de una población

La edad media de los pasajeros y tripulación del Titanic fue de 35 años.

```
t_res <- t.test(titanic$Age, mu = 35)
t_res
```

```
##
##     One Sample t-test
##
## data:  titanic$Age
## t = -9.7507, df = 713, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 35
## 95 percent confidence interval:
##  28.63179 30.76645
## sample estimates:
## mean of x
##  29.69912
```

Test para la media de una población

La edad media de los pasajeros y tripulación del Titanic fue de 35 años.

```
t_res <- t.test(titanic$Age, mu = 35)
t_res
```

```
##
##     One Sample t-test
##
## data:  titanic$Age
## t = -9.7507, df = 713, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 35
## 95 percent confidence interval:
##  28.63179 30.76645
## sample estimates:
## mean of x
##  29.69912
```

Tipos de contrastes de hipótesis

- Test de normalidad
- Test de una media de una población
- **Test de dos medias de dos muestras independientes**
- Test de más de dos medias de muestras independientes
- Test de independencia o asociación
- ...

Tipos de contrastes de hipótesis

Test de igualdad de dos medias de muestras independientes

- Dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Son mucho más comunes que las pruebas con una sola muestra.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

- Hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**
    (aunque esto con una muestra suficientemente grande no es completamente necesario).
  - Contrastar la variabilidad de los dos grupos. **Igualdad de varianzas**.

- Si se cumplen los supuestos, el test a utilizar se llama **t-test** y se basa en el *estadístico t*:

- En **R**:

```
t.test(x = vars_group1, y = vars_group2,
       conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"),
       paired = FALSE,
       var.equal = TRUE)
```

Tipos de contrastes de hipótesis

Test de igualdad de dos medias de muestras independientes

- Dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Son mucho más comunes que las pruebas con una sola muestra.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

- Hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**
    (aunque esto con una muestra suficientemente grande no es completamente necesario).
  - Contrastar la variabilidad de los dos grupos. **Igualdad de varianzas**.

- Si se cumplen los supuestos, el test a utilizar se llama **t-test** y se basa en el *estadístico t*:

- En **R**:

```
t.test(formula = y_vars ~ x_vars,
       conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"),
       paired = FALSE,
       var.equal = TRUE)
```

Tipos de contrastes de hipótesis

Test de igualdad de dos medias de muestras independientes

- Dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Son mucho más comunes que las pruebas con una sola muestra.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

- Hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**
    (aunque esto con una muestra suficientemente grande no es completamente necesario).

- Si **NO** se cumple el supuesto de la normalidad → test No-Paramétrico → test U de Mann-Whitney

O Wilcoxon Rank Sum Test

- En **R**:

```
wilcox.test(x = vars_group1, y = vars_group2,
       conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"),
       paired = FALSE)
```

Tipos de contrastes de hipótesis

Test de igualdad de dos medias de muestras independientes

- Dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Son mucho más comunes que las pruebas con una sola muestra.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

- Hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**
    (aunque esto con una muestra suficientemente grande no es completamente necesario).
  - Contrastar la variabilidad de los dos grupos. **Igualdad de varianzas.**

```
var.test(x = vars_group1, y = vars_group2,
         ratio = 1)
```

```
car::leveneTest(y = y_vars, group = x_vars,
                data = data)
```

- Si **NO** se cumple el supuesto de la homogeneidad de varianzas → test t de Welch

- En **R**:

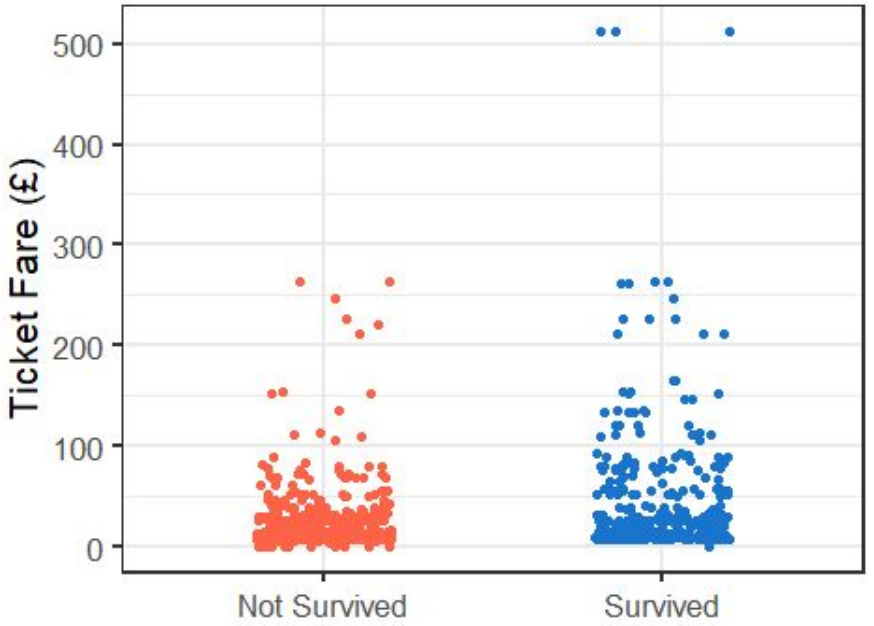
```
t.test(x = vars_group1, y = vars_group2,
       conf.level = 0.95,
       alternative = c("two.sided", "less", "greater"),
       paired = FALSE,
       var.equal = FALSE)
```

Test de igualdad de dos medias de muestras independientes

Los/as pasajeros/as que sobrevivieron pagaron, en promedio, tickets más caros que los que no sobrevivieron.

$$\begin{cases} H_0 : \mu_{Surv} = \mu_{NotSurv}, \\ H_1 : \mu_{Surv} \neq \mu_{NotSurv}. \end{cases}$$

$$\begin{cases} H_0 : \mu_{Surv} \leq \mu_{NotSurv}, \\ H_1 : \mu_{Surv} > \mu_{NotSurv}. \end{cases}$$



Test de igualdad de dos medias de muestras independientes

Los/as pasajeros/as que sobrevivieron pagaron, en promedio, tickets más caros que los que no sobrevivieron.

```
shapiro.test(x = titanic$Fare[which(titanic$Survived == 0)])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  titanic$Fare[which(titanic$Survived == 0)]
## W = 0.51304, p-value < 2.2e-16
```

```
shapiro.test(x = titanic$Fare[which(titanic$Survived == 1)])
```

```
##
##     Shapiro-Wilk normality test
##
## data:  titanic$Fare[which(titanic$Survived == 1)]
## W = 0.59673, p-value < 2.2e-16
```

Test de igualdad de dos medias de muestras independientes

Los/as pasajeros/as que sobrevivieron pagaron, en promedio, tickets más caros que los que no sobrevivieron.

```
Fare_NSurv <- titanic$Fare[which(titanic$Survived_cat == "Not Survived")]
Fare_Surv  <- titanic$Fare[which(titanic$Survived_cat == "Survived")]

u_res <- wilcox.test(x = Fare_NSurv,
                     y = Fare_Surv,
                     paired = FALSE,
                     alternative = "less")
u_res
```

```
##
##     Wilcoxon rank sum test with continuity correction
##
## data:  Fare_NSurv and Fare_Surv
## W = 57807, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Tipos de contrastes de hipótesis

- Test de normalidad
- Test de una media de una población
- Test de dos medias de dos muestras independientes
- **Test de más de dos medias de muestras independientes**
- Test de independencia o asociación
- ...

Tipos de contrastes de hipótesis

Test de igualdad de más de dos medias de muestras independientes

- Más de dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ vs. $H_1: \exists\, i, j\ \mu_i \neq \mu_j$

- Como antes, hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**

  - Contrastar la variabilidad de los dos grupos. **Igualdad de varianzas**.

- Si se cumplen los supuestos, el test a utilizar es **ONE-Way ANOVA**.

- En **R**:

```
aov(formula = y_vars ~ x_vars,
    data = data)
```

```
oneway.test(formula = y_vars ~ x_vars,
            data = data, var.equal = TRUE)
```

Tipos de contrastes de hipótesis

Test de igualdad de más de dos medias de muestras independientes

- Más de dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ vs. $H_1: \exists\, i, j\ \mu_i \neq \mu_j$

- Como antes, hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**

- Si **NO** se cumple el supuesto de la normalidad → test No-Paramétrico → test de Kruskal-Wallis

- En **R:**

```
kruskal.test(formula = y_vars ~ x_vars,
             data = data)
```

Tipos de contrastes de hipótesis

Test de igualdad de más de dos medias de muestras independientes

- Más de dos grupos de sujetos (no emparejados)
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Comprobamos si las medias poblacionales son iguales.
  $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ vs. $H_1: \exists\, i, j\ \mu_i \neq \mu_j$

```
var.test(x = vars_group1, y = vars_group2,
         ratio = 1)
```

```
car::leveneTest(y = y_vars, group = x_vars,
                data = data)
```

- Como antes, hay que comprobar varios supuestos para saber qué test utilizar:
  - Contrastar si los datos provienen de una **distribución normal**
  - Contrastar la variabilidad de los dos grupos. **Igualdad de varianzas.**

- Si **NO** se cumple el supuesto de la homogeneidad de varianzas → ANOVA de Welch

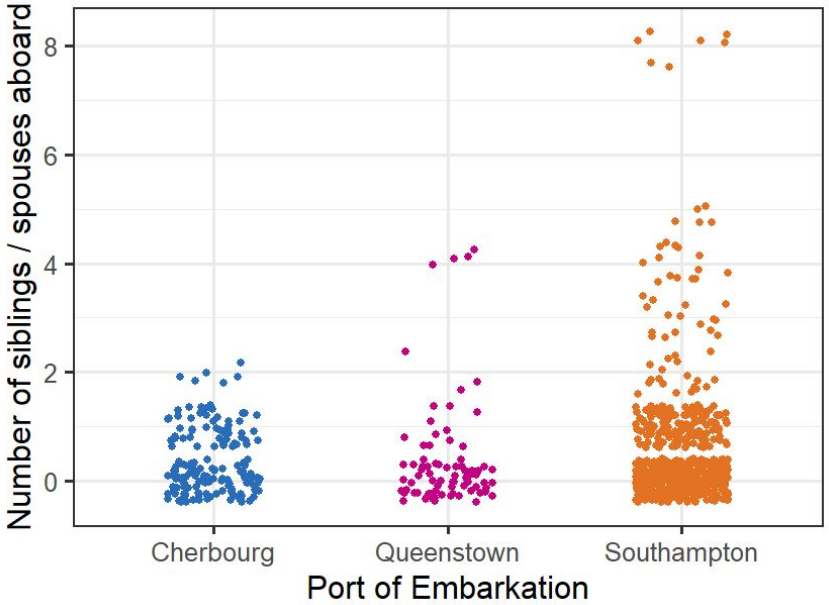
- En **R**:

```
oneway.test(formula = y_vars ~ x_vars,
            data = data, var.equal = FALSE)
```

Contraste de hipótesis para la igualdad de más de dos medias

La tipología de pasajeros/as (viajeros/as solos, parejas y familias) difirió según el puerto de embarque.

$$\begin{cases} \mathrm{H}_0 : \mu_\mathrm{C} = \mu_\mathrm{Q} = \mu_\mathrm{S}, \\ \mathrm{H}_1 : \text{al menos uno } \neq \end{cases}$$



Contraste de hipótesis para la igualdad de más de dos medias

La tipología de pasajeros/as (viajeros/as solos, parejas y familias) difirió según el puerto de embarque.

```
SibSbC <- titanic$SibSp[which(titanic$Embarked == "Cherbourg")]
SibSbQ <- titanic$SibSp[which(titanic$Embarked == "Queenstown")]
SibSbS <- titanic$SibSp[which(titanic$Embarked == "Southampton")]

shapiro.test(SibSbC)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  SibSbC
## W = 0.65462, p-value < 2.2e-16
```

```
# shapiro.test(SibSbQ)
# shapiro.test(SibSbS)
```

Contraste de hipótesis para la igualdad de más de dos medias

La tipología de pasajeros/as (viajeros/as solos, parejas y familias) difirió según el puerto de embarque.

```
krus_res <- kruskal.test(SibSp ~ Embarked, data = titanic)
krus_res
```

```
##
##     Kruskal-Wallis rank sum test
##
## data:  SibSp by Embarked
## Kruskal-Wallis chi-squared = 2.562, df = 2, p-value = 0.2778
```

Tipos de contrastes de hipótesis

- Test de normalidad
- Test de una media de una población
- Test de dos medias de dos muestras independientes
- Test de más de dos medias de muestras independientes
- **Test de independencia o asociación**
- …

Tipos de contrastes de hipótesis

Test de independencia o asociación

- Dos variables cualitativas.
  Nota: si las muestras son dependientes (pareadas) se debe utilizar otro test.

- Comprobamos en una población la posible dependencia de dosd variables cualitativas
  H_0: las variables son independientes vs. H_1: son dependientes

- Como antes, hay que comprobar varios supuestos para saber qué test utilizar:
  - Si hay suficiente representatividad de todas las cateogorías (todas las celdas > 5): **Test del Chi-cuadrado**
  - Si no: **test exacto de Fisher.**

- En **R**:

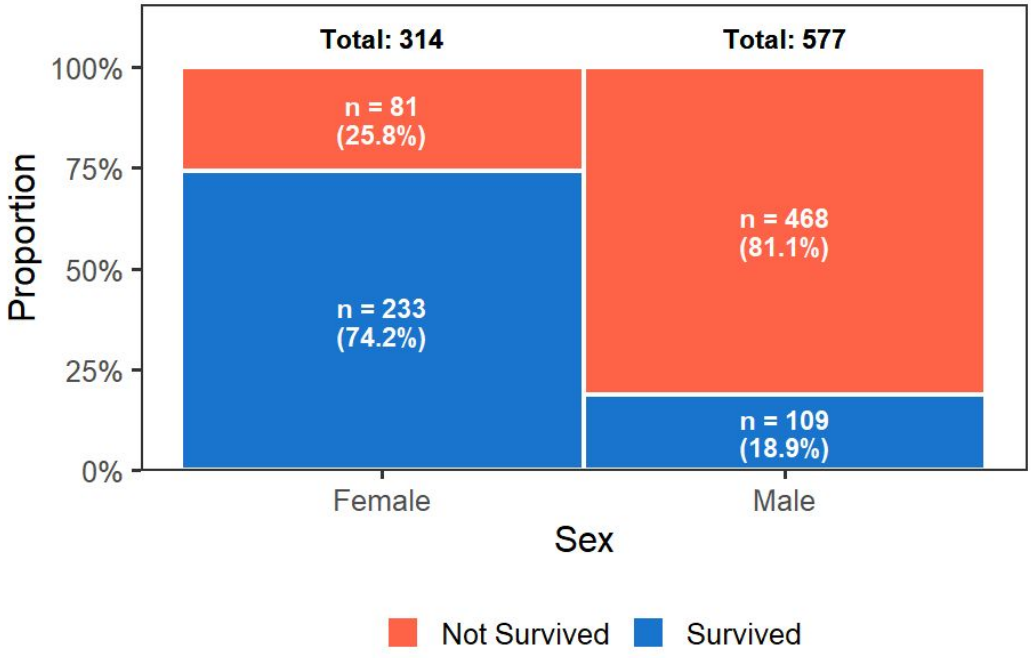
```
chisq.test(x = my_contingency_table,
           conf.level = 0.95,
           alternative = c("two.sided", "less", "greater"))
```

```
fisher.test(x = my_contingency_table,
            conf.level = 0.95,
            alternative = c("two.sided", "less", "greater"))
```

Contraste de hipótesis para la igualdad de proporciones

Las mujeres tuvieron mayores tasas de supervivencia.

$$\begin{cases} \mathrm{H}_0 : p_{\mathrm{Muj}} = p_{\mathrm{Homb}}, \\ \mathrm{H}_1 : p_{\mathrm{Muj}} \neq p_{\mathrm{Homb}}. \end{cases}$$



Contraste de hipótesis para la igualdad de proporciones

Las mujeres tuvieron mayores tasas de supervivencia.

```
table(titanic$Sex, titanic$Survived_cat)
```

```
##
##          Not Survived Survived
##   Female           81      233
##   Male            468      109
```

```
round(prop.table(table(titanic$Sex, titanic$Survived_cat), margin = 1)*100, 1)
```

```
##
##          Not Survived Survived
##   Female         25.8     74.2
##   Male           81.1     18.9
```

Contraste de hipótesis para la igualdad de proporciones

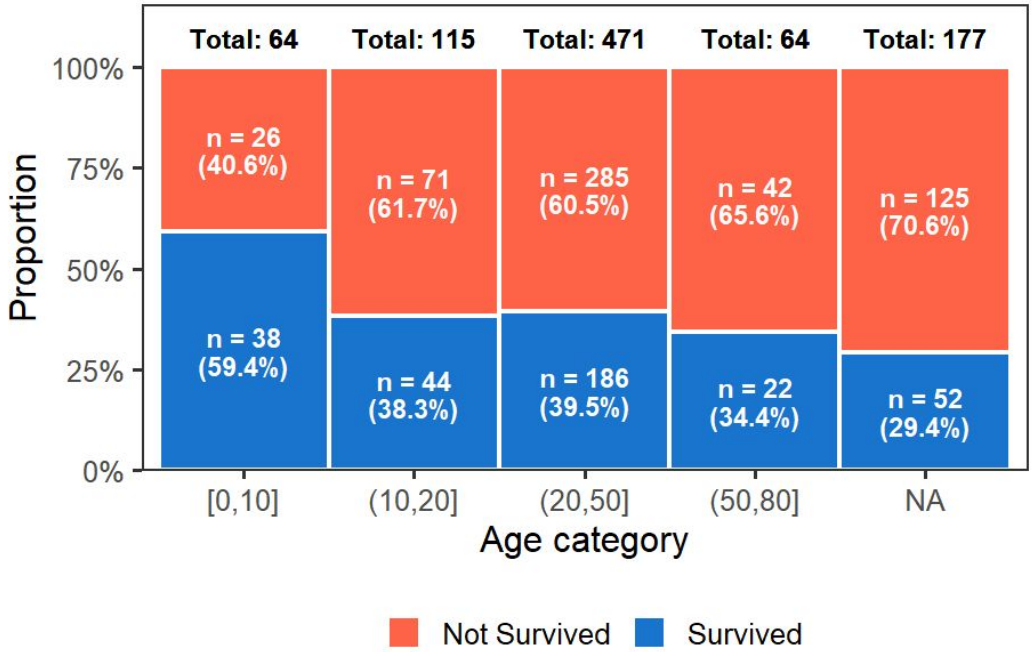
Las mujeres tuvieron mayores tasas de supervivencia.

```
chisq.test(titanic$Sex, titanic$Survived_cat)
```

```
##
##     Pearson's Chi-squared test with Yates' continuity correction
##
## data:  titanic$Sex and titanic$Survived_cat
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Contraste de hipótesis para la igualdad de proporciones

Los niños/as tuvieron mayores tasas de supervivencia.



Contraste de hipótesis para la igualdad de proporciones

Los niños/as tuvieron mayores tasas de supervivencia.

```
table(titanic$Age_cat, titanic$Survived_cat)
```

```
##
##           Not Survived Survived
##   [0,10]            26       38
##   (10,20]           71       44
##   (20,50]          285      186
##   (50,80]           42       22
```

```
round(prop.table(table(titanic$Age_cat, titanic$Survived_cat), margin = 1)*100, 1)
```

```
##
##           Not Survived Survived
##   [0,10]          40.6     59.4
##   (10,20]         61.7     38.3
##   (20,50]         60.5     39.5
##   (50,80]         65.6     34.4
```

```
chisq.test(titanic$Age_cat, titanic$Survived_cat)
```

```
##
##     Pearson's Chi-squared test
##
## data:  titanic$Age_cat and titanic$Survived_cat
## X-squared = 10.883, df = 3, p-value = 0.01238
```

Resumen: α y *p-valor*

α y el *p-valor* están relacionados, pero no son lo mismo….

Sobre el α:

- Se prefija antes del experimento.
- Generalmente bajo (p.ej. la convención del 0.05).
- Vinculado con el valor crítico (“*al conocer uno, se conoce automáticamente el otro*”).
- No se ve afectado por el proceso de muestreo.

Sobre el ***p-valor***:

- Se calcula después del experimento.
- Puede tomar cualquier valor entre [0, 1].
- Tras el cálculo, podemos conocer el nivel de significancia alcanzado.
- Depende del proceso de muestreo.

Resumen: α y *p-valor*

α y el *p-valor* están relacionados, pero no son lo mismo….

¿Qué sé? (Conocimiento/Inferencia)

Sobre el α:

¿Qué hago? (Acción)

- Se prefija antes del experimento.
- Generalmente bajo (p.ej. la convención del 0.05).
- Vinculado con el valor crítico (“*al conocer uno, se conoce automáticamente el otro*”).
- No se ve afectado por el proceso de muestreo.

Sobre el ***p-valor***:

- Se calcula después del experimento.
- Puede tomar cualquier valor entre [0, 1].
- Tras el cálculo, podemos conocer el nivel de significancia alcanzado.
- Depende del proceso de muestreo.

Para ser exactos, la significatividad y la potencia son irrelevantes una vez que se ha llevado a cabo el experimento.

Fisher vs. Neyman-Pearson

Resumen: conceptos erróneos (comunes) sobre el *p-valor*

- El *p-valor* **no** es la probabilidad de que la hipótesis nula sea cierta, ni es la probabilidad de que la hipótesis alternativa sea falsa; o la probabilidad de que los datos sean resultado exclusivamente del azar.

- El *p-valor* **no puede usarse** para calcular la probabilidad de que una hipótesis sea cierta.

- El *p-valor* **no** es la probabilidad de rechazar erróneamente la hipótesis nula.

- El *p-valor* **no** es la probabilidad de que, al replicar el experimento, se obtuviera la misma conclusión.

- El *p-valor* **no** indica la magnitud o importancia del efecto observado. Sin embargo, ambas cosas varían juntas: cuanto mayor es el efecto (tamaño del efecto), menor será el tamaño de muestra requerido para obtener un valor p significativo.

Greenland et al. 2016, *"Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations"*
Goodman 2008, "*A Dirty Dozen: Twelve P-Value Misconceptions* "

Resumen: conceptos erróneos (comunes) sobre el *p-valor*

Declaración de la ASA:
Item 4: Una inferencia apropiada requiere informar todo con transparencia.
¿Qué indica el *p-valor*?
Item 1: Los *p-valores* pueden indicar cómo son de incompatibles los datos con un determinado modelo estadístico

¿Qué no mide o no proporciona el *p-valor*?
Item 2: Los *p-valores* no miden la probabilidad de que la hipótesis estudiada sea cierta, o la probabilidad de que los datos sean resultado exclusivamente del azar.
Item 3: Las conclusiones científicas y las decisiones económicas o políticas no deberían basarse en si un *p-valor* cruza un determinado umbral.
Item 5: Un *p-valor*, o significatividad estadística, no mide el tamaño de un efecto o la importancia de un resultado.
Item 6: En sí mismo, el *p-valor* no proporciona una buena medida de la evidencia referida a un modelo o hipótesis.

Cortés Martínez et al. (2015), Importancia de la potencia y la hipótesis en el valor p.
Wasserstein et al. 2016, *"The ASA Statement on p-Values: Context, Process, and Purpose"*
Vilaró et al. 2019, "*Adherence to reporting guidelines increases the number of citations: the argument for including a methodologist in the editorial process and peer-review* "

Resumen: conceptos erróneos (comunes) sobre el *p-valor*

Declaración de la ASA:
Item 4: Una inferencia apropiada requiere informar todo con transparencia.
¿Qué indica el *p-valor*?
Item 1: Los *p-valores* pueden indicar cómo son de incompatibles los datos con un determinado modelo estadístico

¿Qué no mide o no proporciona el *p-valor*?
Item 2: Los *p-valores* no miden la probabilidad de que la hipótesis estudiada sea cierta, o la probabilidad de que los datos sean resultado exclusivamente del azar.
Item 3: Las conclusiones científicas y las decisiones económicas o políticas no deberían basarse en si un *p-valor* cruza un determinado umbral.
Item 5: Un *p-valor*, o significatividad estadística, no mide el tamaño de un efecto o la importancia de un resultado.
Item 6: En sí mismo, el *p-valor* no proporciona una buena medida de la evidencia referida a un modelo o hipótesis.

Vilaró *et al. BMC Medical Research Methodology* (2019) 19:112
https://doi.org/10.1186/s12874-019-0746-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Adherence to reporting guidelines increases the number of citations: the argument for including a methodologist in the editorial process and peer-review

Marta Vilaró[1,12*], Jordi Cortés[1], Albert Selva-O'Callaghan[2,3,4], Agustín Urrutia[2,3,5], Josep-Maria Ribera[2,3,6], Francesc Cardellach[2,7], Xavier Basagaña[8,9,10], Matthew Elmore[1], Miquel Vilardell[2,3,4], Douglas Altman[11], José-Antonio González[1] and Erik Cobo[1,2]

Scientifically tested ≠ p-valor < 0.05

<u>Recomendaciones</u>:
- Reportar los tamaños de efecto (i.e. la magnitud del efecto, el *effect size*) con sus intervalos de incertidumbre (Guía CONSORT 1996, item 17a)
- Seguir las guías de publicación (p.ej. CONSORT, STROBE, TRIPOD....)

Cortés Martínez et al. (2015), Importancia de la potencia y la hipótesis en el valor p.
Wasserstein et al. 2016, "*The ASA Statement on p-Values: Context, Process, and Purpose*"
Vilaró et al. 2019, "*Adherence to reporting guidelines increases the number of citations: the argument for including a methodologist in the editorial process and peer-review*"

Comparaciones múltiples - *Multiple testing*

- El peligro de llevar a cabo muchas pruebas de significatividad… Si realizamos muchas pruebas simultáneamente, la probabilidad de que haya, por casualidad,al menos un falso positivo aumenta y ya no coincide con la probabilidad de error de tipo I.

- Este aumento en la probabilidad de error de tipo I debe compensarse de alguna manera → ajustes por pruebas múltiples

- Hay diversas propuestas en la literatura para “corregir” el *p-valor*. P.ej. Bonferroni, Benjamini Hochberg, Holm…

- Ejemplo de: *Basic Statistics for Biomedical Research Course*, Vall d'Hebron Institut d’Oncologia (VHIO) Edition 2019

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Imagina que eres un aventurero que tiene la opción de cruzar un puente para escapar del peligro y encontrar un tesoro, y que hay un cartel delante del puente que dice:

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Imagina que eres un aventurero que tiene la opción de cruzar un puente para escapar del peligro y encontrar un tesoro, y que hay un cartel delante del puente que dice:

“Este puente solo se ha roto una de cada 100 veces”.

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Imagina que eres un aventurero que tiene la opción de cruzar un puente para escapar del peligro y encontrar un tesoro, y que hay un cartel delante del puente que dice:

“Este puente solo se ha roto una de cada 100 veces”.

Por tanto, el *p-valor* de nuestra metáfora es 0.01

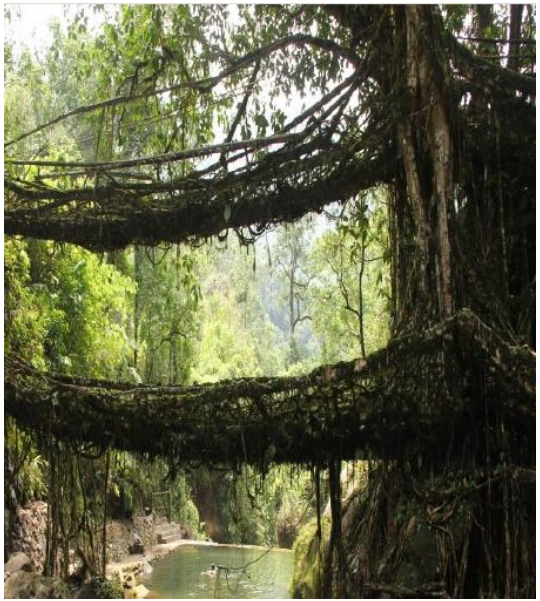
Podrías aceptar que el 1 % es un riesgo lo suficientemente pequeño como para cruzar el puente y perseguir tu objetivo . OK

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Pero, ¿qué decides si para alcanzar tu objetivo tienes que cruzar cientos de puentes de ese tipo?

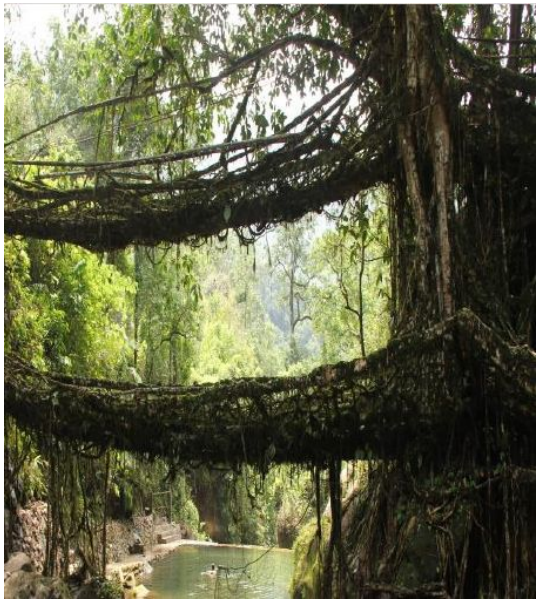


https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Pero, ¿qué decides si para alcanzar tu objetivo tienes que cruzar cientos de puentes de ese tipo?



En este caso, la probabilidad de caerse al cruzar uno de los puentes es obviamente demasiado alta (porque solo tenemos una vida).

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

- Pero, ¿qué decides si para alcanzar tu objetivo tienes que cruzar cientos de puentes de ese tipo?

En este caso, la probabilidad de caerse al cruzar uno de los puentes es obviamente demasiado alta (porque solo tenemos una vida).

$$\begin{aligned}\text{Prob}(\text{Caerse en uno de los 100 puentes}) &= 1 - \text{Prob}(\text{No caerse en ninguno de los 100 puentes}) \\ &= 1 - (\text{Prob}(\text{No caerse en el puente}))^{100} \\ &= 1 - (1 - 0.01)^{100} = 0.634\end{aligned}$$

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Comparaciones múltiples - *Multiple testing*

to cross or not to cross

Por lo tanto, en este caso (multiple testing), el *p-valor* por sí solo no es una buena referencia para aceptar o no la significación estadística.

Debemos aplicar algún tipo de ajuste a los *p-valores* que nos "*permita cruzar*" todos los puentes con seguridad.

https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/tree/master/Sessio_7

Caso práctico II

La base de datos `rotterdam.xlsx` incluye 2982 pacientes con cáncer de mama primario cuyos registros se incluyeron en el banco de tumores de Rotterdam (Royston and Altman (2013)).

`pid`	Patient identifier
`year`	Year of surgery
`age`	Age at surgery
`meno`	Menopausal status (0 = premonopausal, 1 = postmenopausal)
`size`	Tumor size, a factor with levels <=20, 20-50, >50
`grade`	Differentiation grade
`nodes`	Number of positive lymph nodes
`pgr`	Progesterone receptors (fmol/l)
`er`	Estrogen receptors (fmol/l)
`hormon`	Hormonal treatment (0 = no, 1 = yes)
`chemo`	Chemotherapy
`rtime`	Days to relapse or last follow-up
`recur`	0 = no relapse, 1 = relapse
`dtime`	Days to death or last follow-up
`death`	0 = alive, 1 = dead

Fuente: Patrick Royston and Douglas Altman, *External validation of a Cox prognostic model: principles and methods*. BMC Medical Research Methodology 2013, 13:33
Disponible en **R**: `survival::rotterdam`

Caso práctico II

La base de datos `rotterdam.xlsx` incluye 2982 pacientes con cáncer de mama primario cuyos registros se incluyeron en el banco de tumores de Rotterdam (Royston and Altman (2013)).

- El tamaño medio del tumor difiere de 25 mm.
- Hay alguna diferencia en la edad entre las pacientes que reciben terapia hormonal y las que no.
- La distribución del tamaño de los tumores difiere entre los grupos que reciben terapia hormonal.
- El estado de recurrencia está asociado con el número de categorías de ganglios positivos.

Fuente: Patrick Royston and Douglas Altman, *External validation of a Cox prognostic model: principles and methods*. BMC Medical Research Methodology 2013, 13:33
Disponible en **R**: `survival::rotterdam`

Bloque B

Introducción a la modelización estadística en R

¿Qué es un modelo?

- En el mundo físico: una simplificación de cosas del mundo real.
- En estadística: un modelo trata de captar la estructura de los datos de la forma más sencilla posible.
- En ambos casos: una ficción conveniente.
- La “estructura básica” de un modelo estadístico es

$$\text{Datos} = \underbrace{\text{Modelo}}_{\text{Expresa los valores que esperamos que adopten los datos, dado nuestro conocimiento}} + \underbrace{\text{Error}}_{\text{Diferencia entre las estimaciones del modelo y los datos observados}}$$

- Nos interesa minimizar al máximo el Error.

“*All models are wrong but some are useful*.” George Box

¿Qué es un modelo?

- En el mundo físico: una simplificación de cosas del mundo real.
- En estadística: un modelo trata de captar la estructura de los datos de la forma más sencilla posible.
- En ambos casos: una ficción conveniente.
- La “estructura básica” de un modelo estadístico es

Un modelo simple utilizando **un parámetro**

$$y_i = \beta + \epsilon$$

Dato/observación i-ésima

Parámetro
No sabemos su valor real/verdadero. Tenemos que estimarlo a partir de los datos ($\hat{\beta}$)

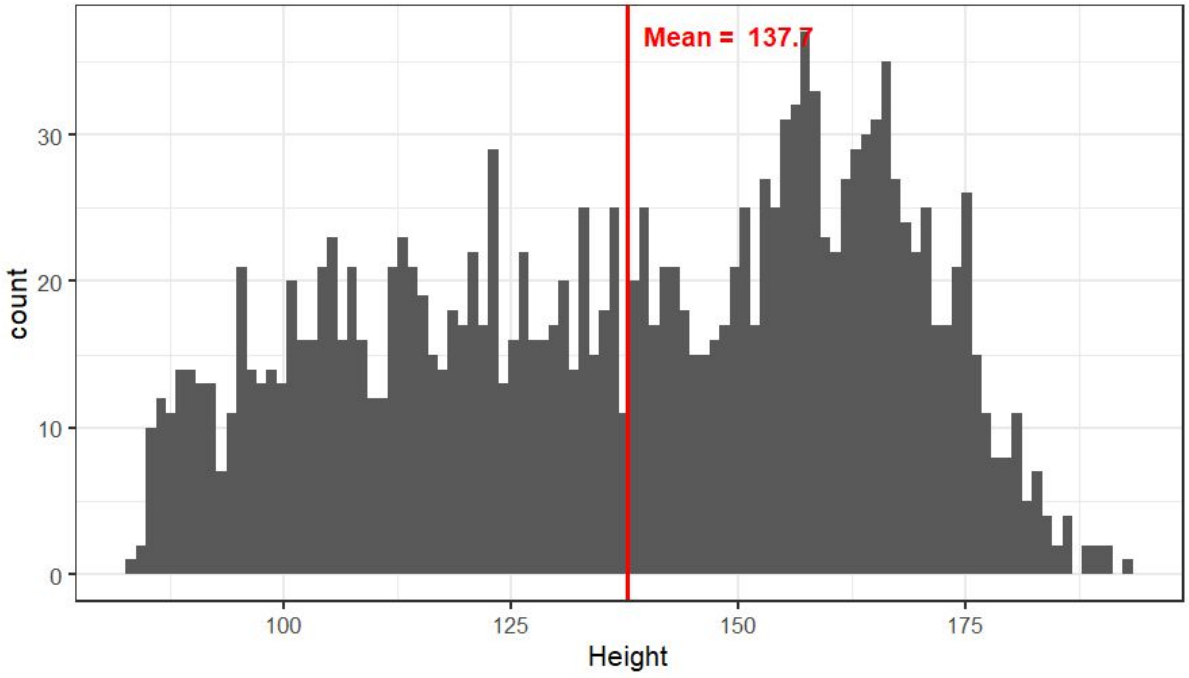
Este modelo no depende de la observación i. La estimación es la misma para todos los datos.

- Nos interesa minimizar al máximo el Error.

“*All models are wrong but some are useful*.” George Box

Ejemplo

Altura de las niñas y niños de la muestra NHANES



Muestra: niños/as (< 18 años) de la muestra NHANES

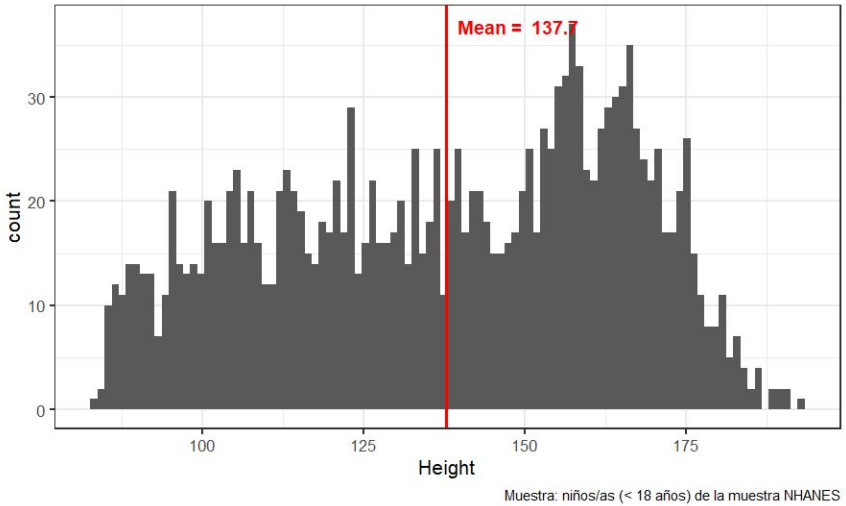
$$\hat{y}_i = 137.7$$

Fuente: US National Health and Nutrition Examination Survey (NHANES, 2009-2012)
Disponible en **R**: `NHANES::data(NHANES)`

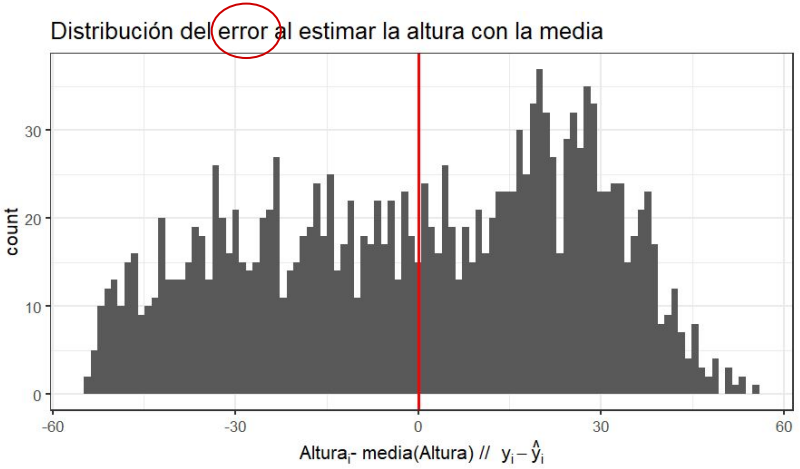
Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Ejemplo

Altura de las niñas y niños de la muestra NHANES



Muestra: niños/as (< 18 años) de la muestra NHANES



Muestra: niños/as (< 18 años) de la muestra NHANES

$$\hat{y}_i = 137.7$$

¿CÓMO MEDIR EL ERROR?

$$\text{error}_i = y_i - \hat{y}_i$$

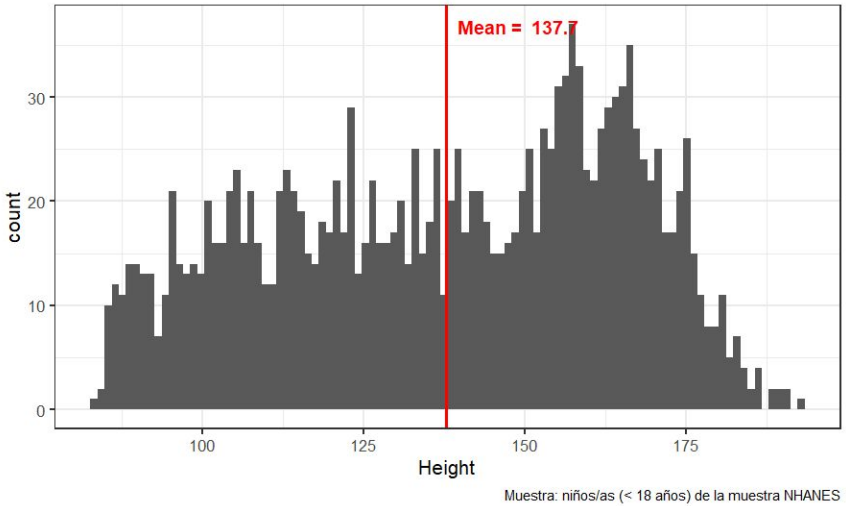
$$\text{error}_i = (y_i - \hat{y}_i)^2$$

$$\text{error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

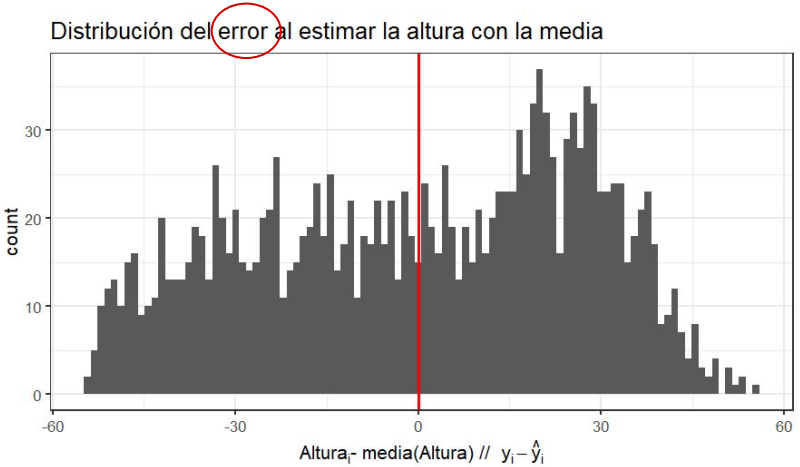
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Ejemplo

Altura de las niñas y niños de la muestra NHANES



Muestra: niños/as (< 18 años) de la muestra NHANES



Muestra: niños/as (< 18 años) de la muestra NHANES

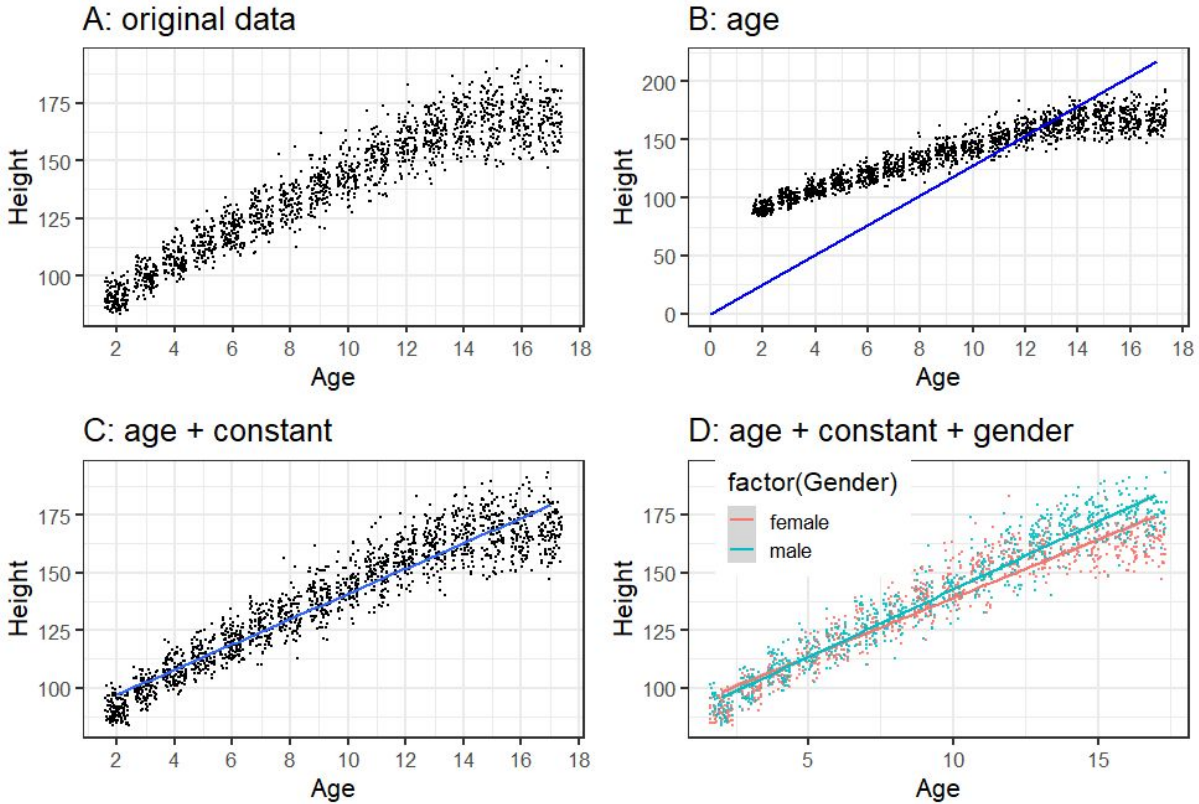
$$\hat{y}_i = 137.7$$

¿CÓMO MEDIR EL ERROR?

$$RMSE = 26.9$$

Ejemplo

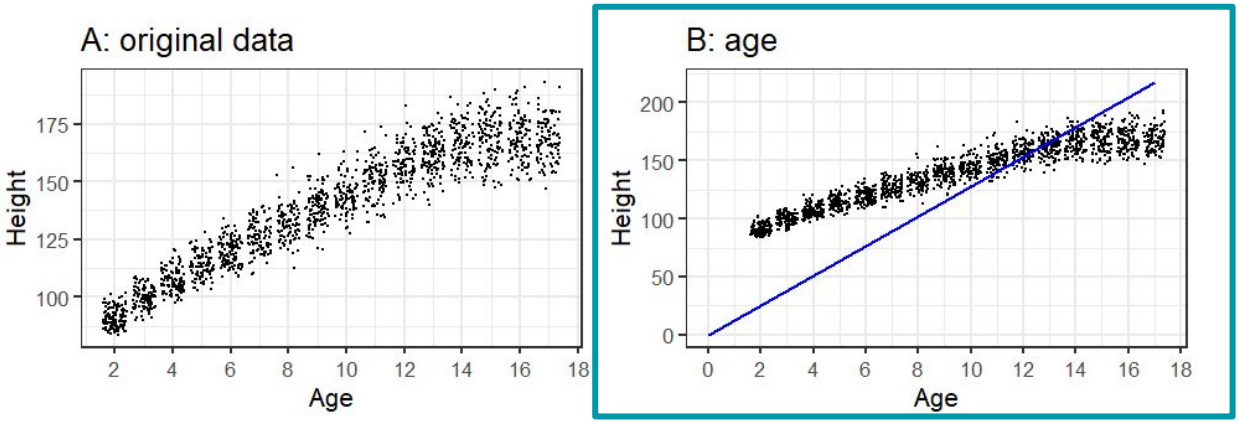
Altura de las niñas y niños de la muestra NHANES

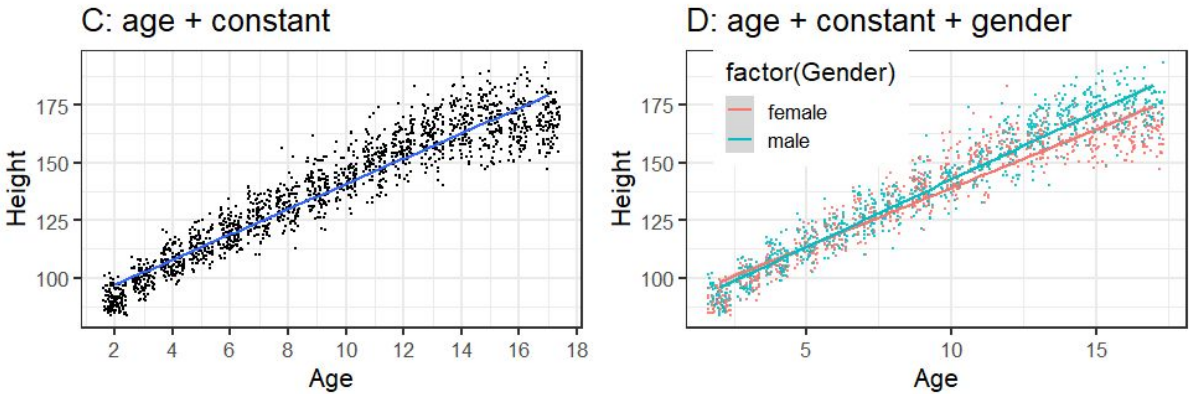


Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Ejemplo

Altura de las niñas y niños de la muestra NHANES





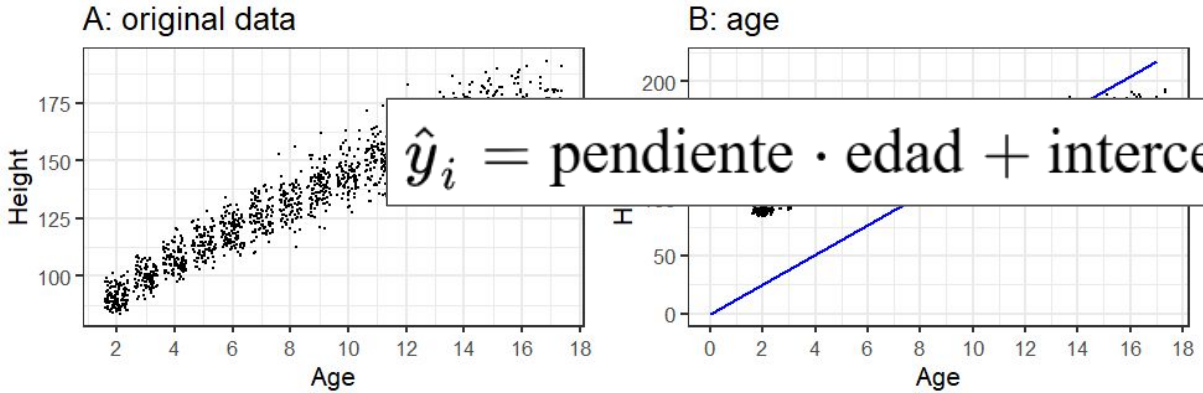
Modelo B

$$\hat{y}_i = \hat{\beta} \cdot \text{edad}$$

Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

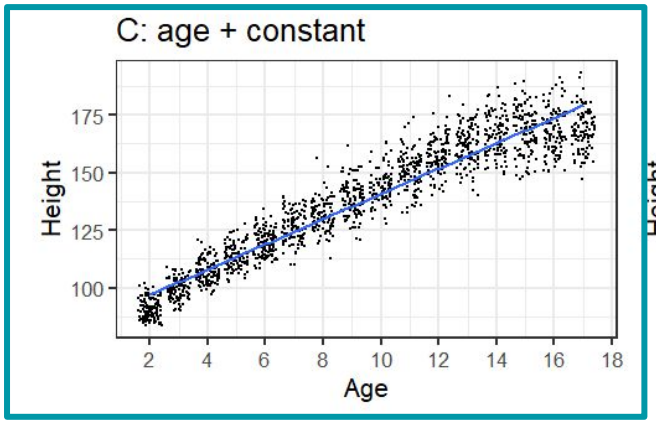
Ejemplo

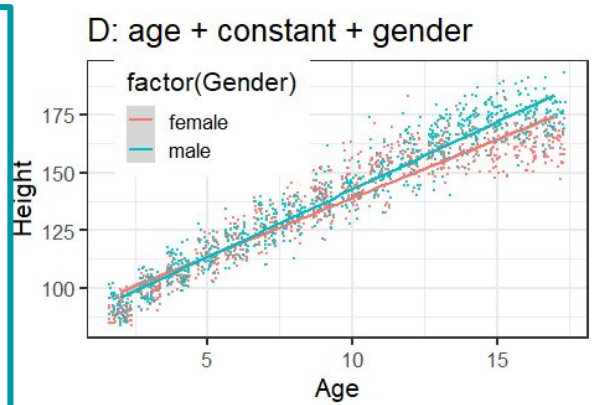
Altura de las niñas y niños de la muestra NHANES



Modelo C

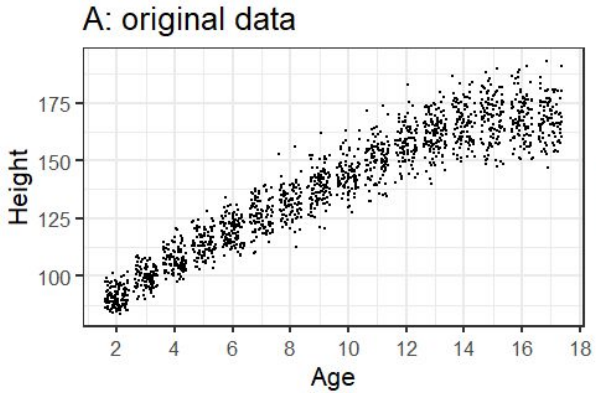
$$\hat{y}_i = \text{pendiente} \cdot \text{edad} + \text{intercepto} = \hat{\beta}_1 \cdot \text{edad} + \hat{\beta}_0$$

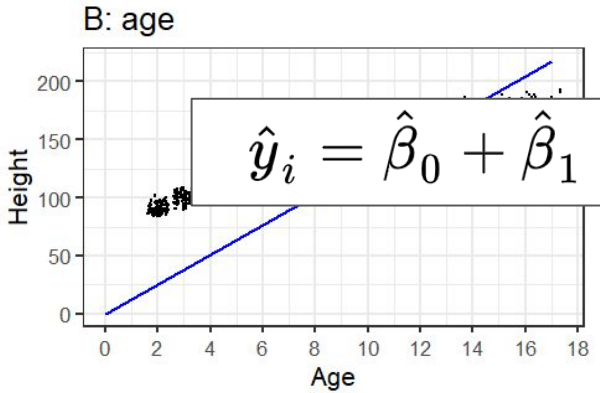




Ejemplo

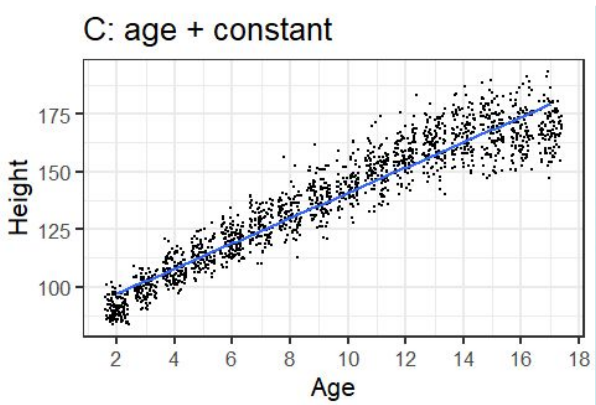
Altura de las niñas y niños de la muestra NHANES

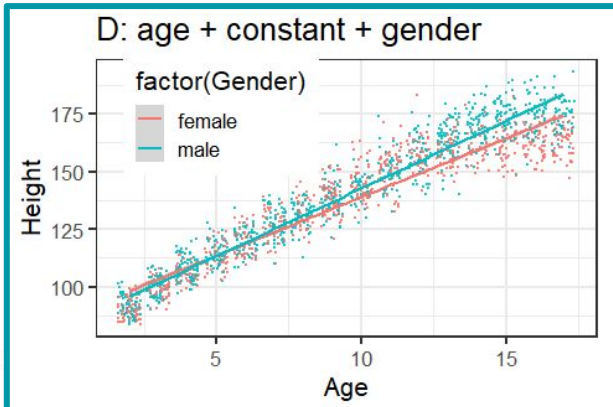




Modelo D

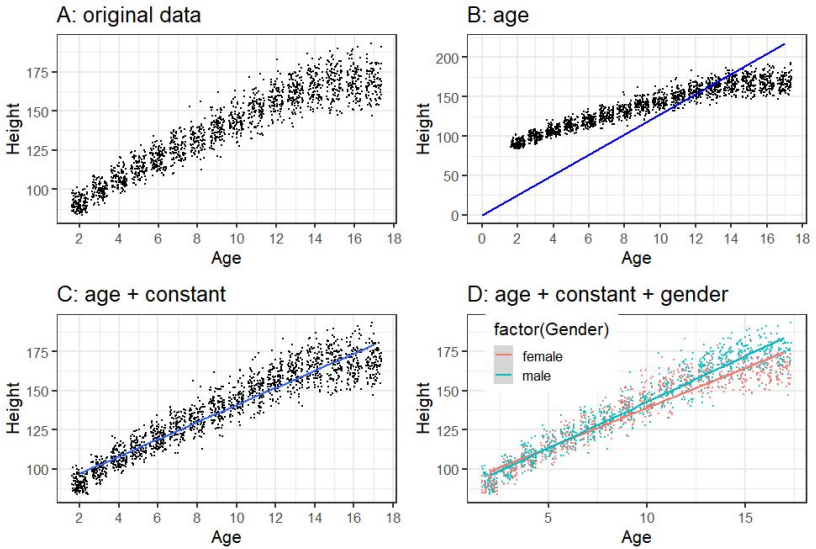
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{edad} + \hat{\beta}_2 \text{Genero}_{\text{chico}}$$





Ejemplo

Altura de las niñas y niños de la muestra NHANES

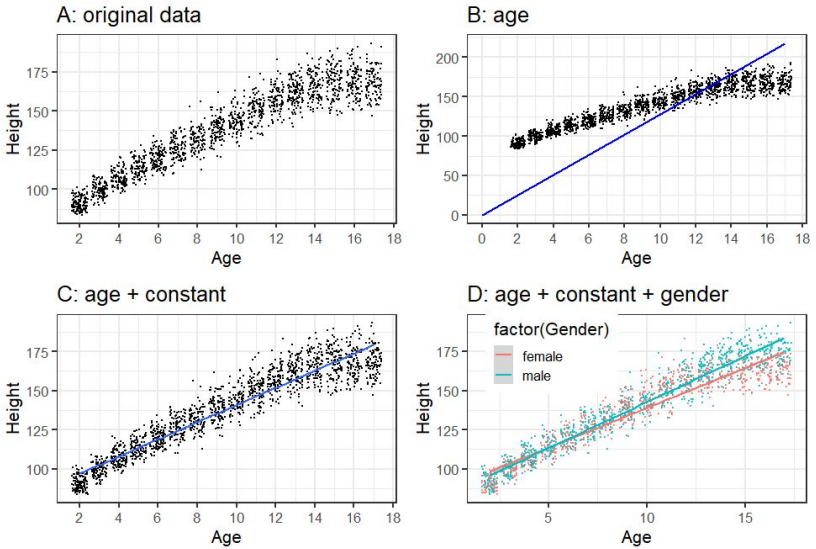


Model	RMSE
mean	26.9
age	39.2
constant + age	8.4
constant + age + gender	8.2

Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Ejemplo

Altura de las niñas y niños de la muestra NHANES



En **R**, las funciones más importante `lm()` y `summary()`:

```
my_model <- lm(formula = y ~ x, data)

summary(my_model)
```


Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Regresión lineal simple

Definición: Un modelo de **regresión lineal** es aquel en el que el modelo para la variable dependiente **Y** se compone de una combinación lineal de variables independientes **X**, cada una de las cuales se multiplica por un peso (e.g. β), que determina la contribución relativa de esa variable independiente a la estimación del modelo.

$$Y \sim X$$

Variable dependiente/ Variable respuesta/ Outcome

Variable independiente

- La **mayoría de los modelos** utilizados en estadística pueden enmarcarse en términos del modelo lineal general o una extensión del mismo.
- La **mayoría de los tests estadísticos** son casos especiales del modelo lineal.

Regresión lineal simple

Definición: Un modelo de **regresión lineal** es aquel en el que el modelo para la variable dependiente **Y** se compone de una combinación lineal de variables independientes **X**, cada una de las cuales se multiplica por un peso (e.g. β), que determina la contribución relativa de esa variable independiente a la estimación del modelo.

$$Y \sim X$$

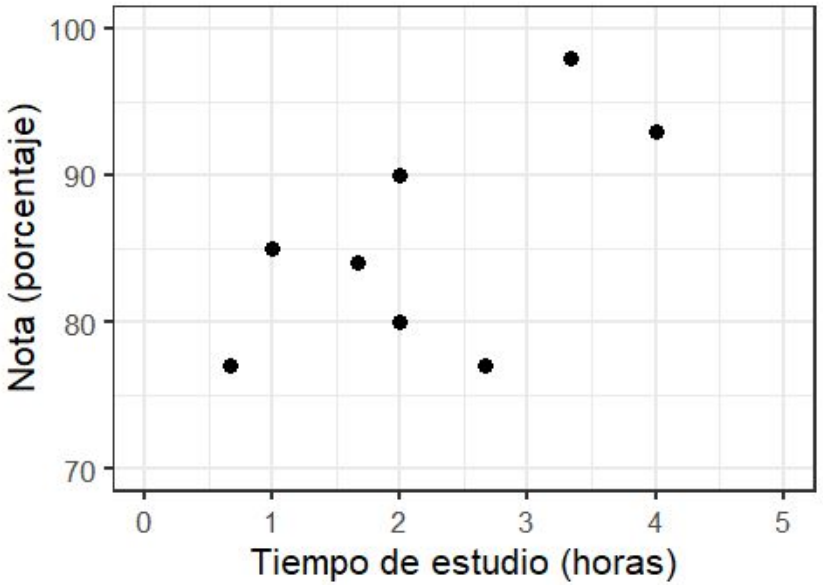
Variable dependiente/ Variable respuesta/ Outcome

Variable independiente

- **Asunciones** de la regresión lineal simple:
  - Linealidad y aditividad
  - Errores independientes
  - Homocedasticidad (varianza constante de los errores)
  - Normalidad de los errores (los errores deben seguir la distribución normal)

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas**



Describir: ¿Qué tan fuerte es la relación entre la calificación y el tiempo de estudio?

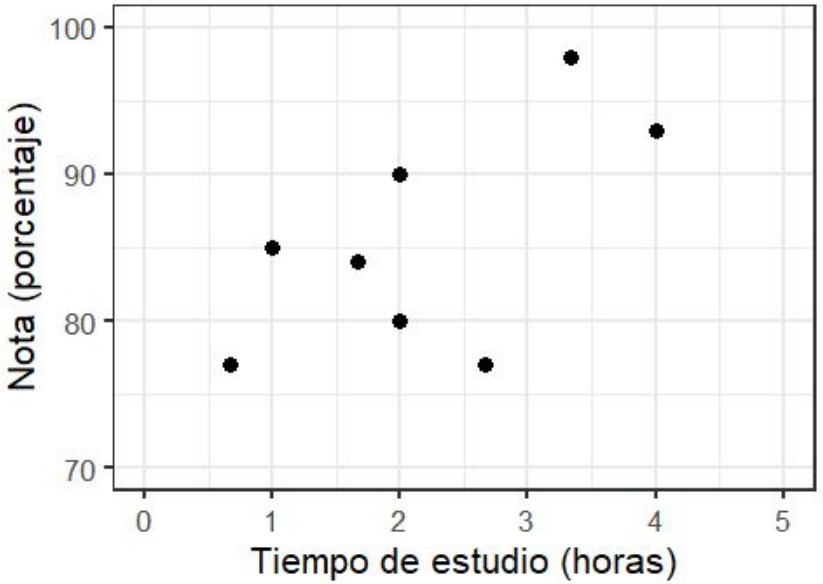
Decidir: ¿Existe una relación estadísticamente significativa entre la calificación y el tiempo de estudio?

Predecir: Dada una cantidad determinada de tiempo de estudio, ¿qué calificación esperamos obtener?

Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas**



Lo que queda una vez que se ha ajustado el modelo; a menudo nos referimos a estos como los residuos del modelo.

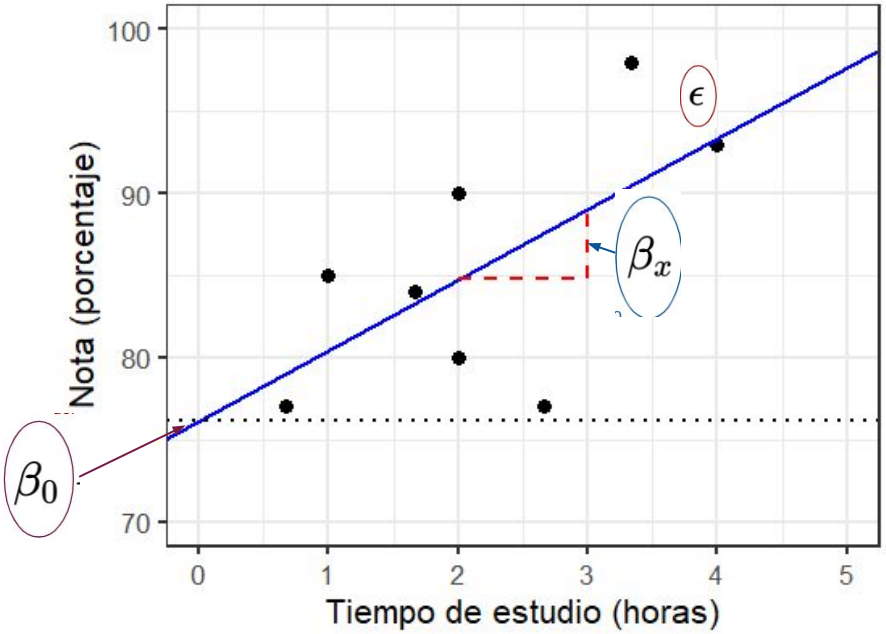
$$y = x \cdot \beta_x + \beta_0 + \epsilon$$

Cuánto esperaríamos que cambiara **y** dado un cambio de una unidad en **x**.

El intercepto es un offset global, que nos indica qué valor esperaríamos que tuviera **y** cuando **x = 0**

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas**



Lo que queda una vez que se ha ajustado el modelo; a menudo nos referimos a estos como los residuos del modelo.

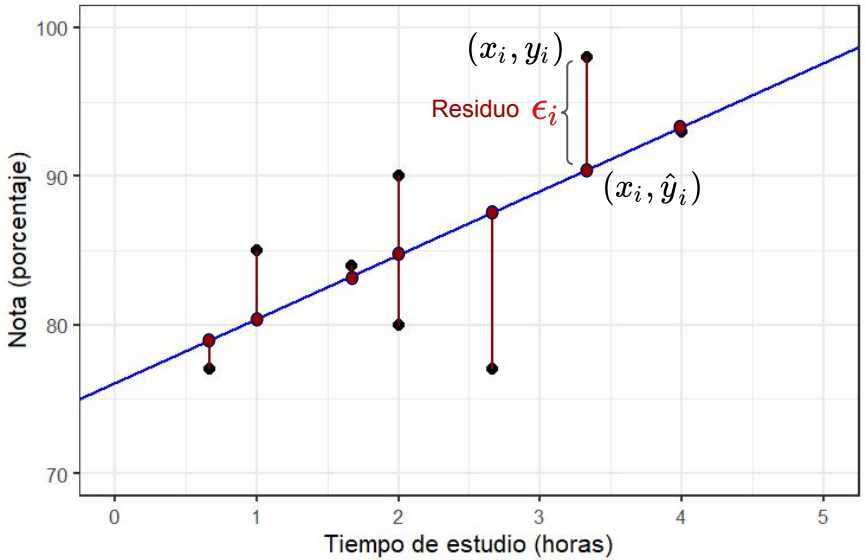
$$y = x \cdot \beta_x + \beta_0 + \epsilon$$

Cuánto esperaríamos que cambiara **y** dado un cambio de una unidad en **x**.

El intercepto es un offset global, que nos indica qué valor esperaríamos que tuviera **y** cuando **x = 0**

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas**



Lo que queda una vez que se ha ajustado el modelo; a menudo nos referimos a estos como los residuos del modelo.

$$y = x \cdot \beta_x + \beta_0 + \epsilon$$

Cuánto esperaríamos que cambiara **y** dado un cambio de una unidad en **x**.

El intercepto es un offset global, que nos indica qué valor esperaríamos que tuviera **y** cuando **x = 0**

Técnica de mínimos cuadrados: *la recta que hace que estos residuos sean los más pequeños posibles*

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas. En R:**

```
mylinearModel <- lm(grade ~ studyTime, data = grades)

summary(mylinearModel)
```

```
##
## Call:
## lm(formula = grade ~ studyTime, data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.656  -2.719   0.125   4.703   7.469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.156      5.161  14.756 6.09e-06 ***
## studyTime      4.313      2.142   2.013   0.0907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.386 on 6 degrees of freedom
## Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3037
## F-statistic: 4.054 on 1 and 6 DF,  p-value: 0.09073
```

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas. En R:**

```
mylinearModel <- lm(grade ~ studyTime, data = grades)

summary(mylinearModel)
```

```
##
## Call:
## lm(formula = grade ~ studyTime, data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.656  -2.719   0.125   4.703   7.469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.156      5.161  14.756 6.09e-06 ***
## studyTime      4.313      2.142   2.013   0.0907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.386 on 6 degrees of freedom
## Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3037
## F-statistic: 4.054 on 1 and 6 DF,  p-value: 0.09073
```

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas. En R:**

```
mylinearModel <- lm(grade ~ studyTime, data = grades)

summary(mylinearModel)
```

```
##
## Call:
## lm(formula = grade ~ studyTime, data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.656  -2.719   0.125   4.703   7.469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.156      5.161  14.756 6.09e-06 ***
## studyTime      4.313      2.142   2.013   0.0907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.386 on 6 degrees of freedom
## Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3037
## F-statistic: 4.054 on 1 and 6 DF,  p-value: 0.09073
```

$$\hat{y} = \underset{\hat{\beta}_0}{76.16} + \underset{\hat{\beta}_1}{4.31} * \text{studyTime}$$

Regresión lineal simple

Ejemplo simulado, **relación entre tiempo de estudio y notas. En R:**

```
mylinearModel <- lm(grade ~ studyTime, data = grades)

summary(mylinearModel)
```

```
##
## Call:
## lm(formula = grade ~ studyTime, data = grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.656  -2.719   0.125   4.703   7.469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.156      5.161  14.756 6.09e-06 ***
## studyTime      4.313      2.142   2.013   0.0907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.386 on 6 degrees of freedom
## Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3037
## F-statistic: 4.054 on 1 and 6 DF,  p-value: 0.09073
```

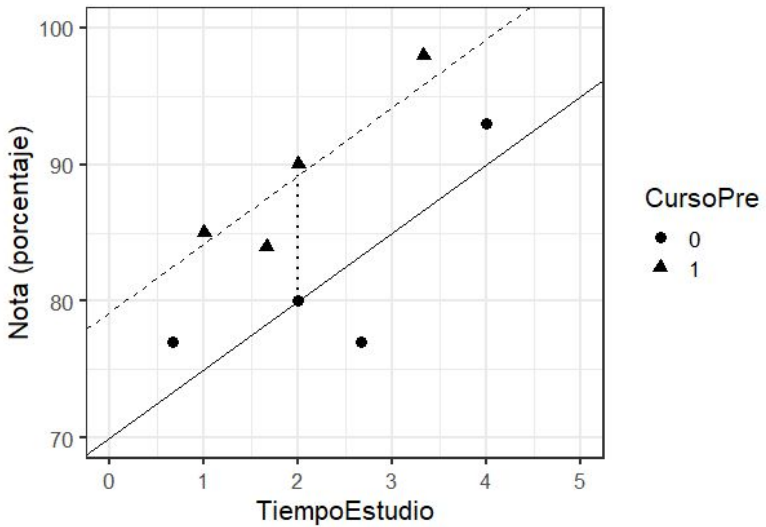
Regresión lineal **múltiple**

Siguiendo con el ejemplo simulado.. Supongamos que descubrimos que algunos de los estudiantes habían cursado previamente una asignatura sobre el tema.
Quemos estudiar si aquellos que habían cursado previamente la asignatura obtienen mejores resultados que los que no lo habían hecho, con el mismo tiempo de estudio:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{TiempoEstudio} + \hat{\beta}_2 \cdot \text{CursoPre}$$

Regresión lineal **múltiple**

Siguiendo con el ejemplo simulado.. Supongamos que descubrimos que algunos de los estudiantes habían cursado previamente una asignatura sobre el tema.
Quemos estudiar si aquellos que habían cursado previamente la asignatura obtienen mejores resultados que los que no lo habían hecho, con el mismo tiempo de estudio:



Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*" .

Regresión lineal **múltiple**

Siguiendo con el ejemplo simulado.. Supongamos que descubrimos que algunos de los estudiantes habían cursado previamente una asignatura sobre el tema.
Quemos estudiar si aquellos que habían cursado previamente la asignatura obtienen mejores resultados que los que no lo habían hecho, con el mismo tiempo de estudio:

```
mylinearModel2 <- lm(grade ~ studyTime + priorClass, data = grades)
summary(mylinearModel2)
```

```
##
## Call:
## lm(formula = grade ~ studyTime + priorClass, data = grades)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  3.58333  0.75000 -3.58333 -0.08333  0.75000 -6.41667  2.08333  2.91667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.083      3.768  18.600 8.27e-06 ***
## studyTime      5.000      1.366   3.661   0.0146 *
## priorClass1    9.167      2.879   3.184   0.0244 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.021 on 5 degrees of freedom
## Multiple R-squared:  0.8028,    Adjusted R-squared:  0.724
## F-statistic: 10.18 on 2 and 5 DF,  p-value: 0.01726
```

Regresión lineal **múltiple**

Siguiendo con el ejemplo simulado.. Supongamos que descubrimos que algunos de los estudiantes habían cursado previamente una asignatura sobre el tema.
Quemos estudiar si aquellos que habían cursado previamente la asignatura obtienen mejores resultados que los que no lo habían hecho, con el mismo tiempo de estudio:

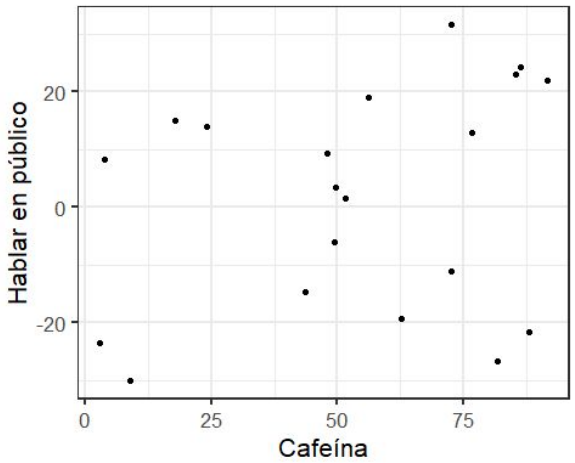
```
mylinearModel2 <- lm(grade ~ studyTime + priorClass, data = grades)
summary(mylinearModel2)
```

```
##
## Call:
## lm(formula = grade ~ studyTime + priorClass, data = grades)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  3.58333  0.75000 -3.58333 -0.08333  0.75000 -6.41667  2.08333  2.91667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.083      3.768  18.600 8.27e-06 ***
## studyTime      5.000      1.366   3.661   0.0146 *
## priorClass1    9.167      2.879   3.184   0.0244 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.021 on 5 degrees of freedom
## Multiple R-squared:  0.8028,    Adjusted R-squared:  0.724
## F-statistic: 10.18 on 2 and 5 DF,  p-value: 0.01726
```

Regresión lineal **múltiple** : interacción entre variables

Hemos asumido que el efecto del tiempo de estudio en la calificación (es decir, la pendiente de regresión) era el mismo para ambos grupos. Sin embargo, en algunos casos podríamos imaginar que el efecto de una variable podría diferir dependiendo del valor de otra variable → **interacción**

Otro ejemplo que plantea la siguiente pregunta: ¿Qué efecto tiene la cafeína en hablar en público?



Parece que no hay una relación aparente entre la Cafeína y el "Hablar en público"

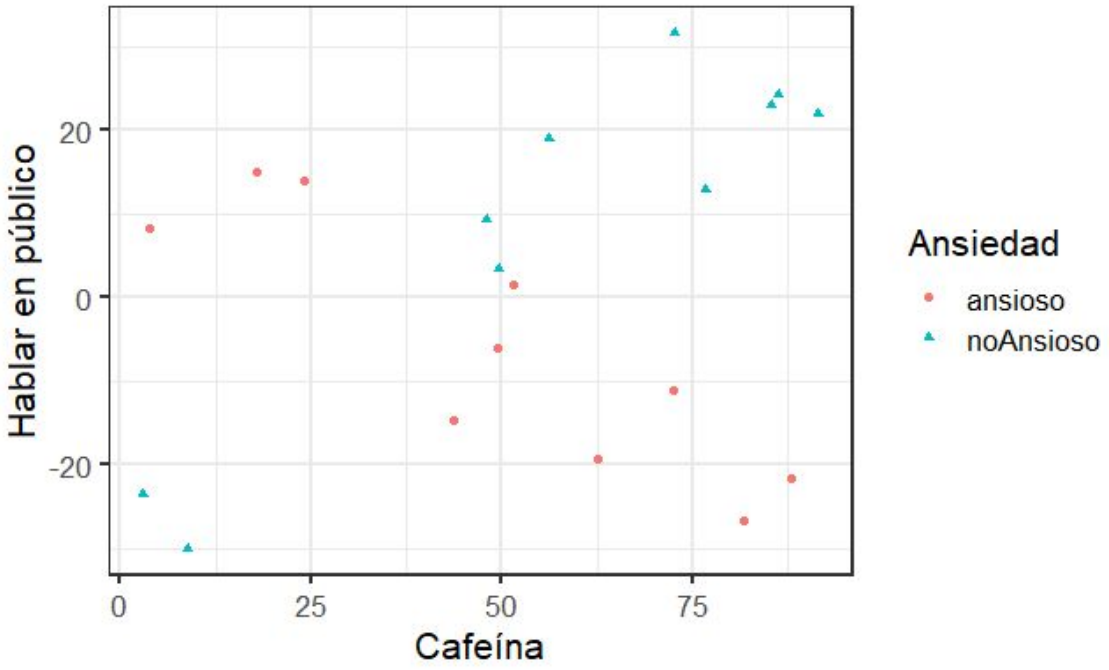
Fuente: Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*"

Regresión lineal **múltiple** : interacción entre variables

```
lmResultCafeina <- lm(HablarPubl ~ Cafeina, data = caffeine)
summary(lmResultCafeina)
```

```
##
## Call:
## lm(formula = HablarPubl ~ Cafeina, data = caffeine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.096 -16.024   5.014  16.453  26.979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.4132     9.1653  -0.809    0.429
## Cafeina       0.1676     0.1508   1.111    0.281
##
## Residual standard error: 19.19 on 18 degrees of freedom
## Multiple R-squared:  0.06419,    Adjusted R-squared:  0.0122
## F-statistic: 1.235 on 1 and 18 DF,  p-value: 0.2811
```

Regresión lineal **múltiple** : interacción entre variables



Parece que la relación entre el habla y la cafeína es diferente para los dos grupos, ya que la cafeína mejora el rendimiento de las personas sin ansiedad y lo degrada en las personas con ansiedad.

Regresión lineal **múltiple** : interacción entre variables

```
lmResultCafAnx <- lm(HablarPubl ~ Cafeina + Ansiedad, data = df)
summary(lmResultCafAnx)
```

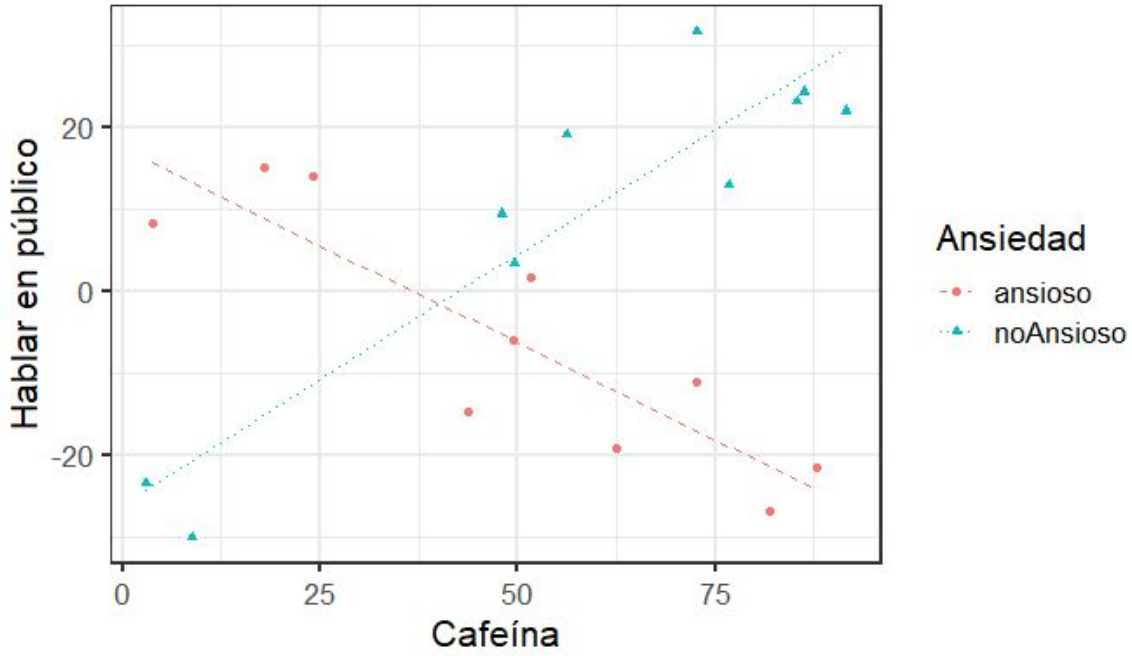
```
##
## Call:
## lm(formula = HablarPubl ~ Cafeina + Ansiedad, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.968  -9.743   1.351  10.530  25.361
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12.5812     9.1967  -1.368    0.189
## Cafeina             0.1313     0.1446   0.908    0.377
## AnsiedadnoAnsioso  14.2328     8.2324   1.729    0.102
##
## Residual standard error: 18.21 on 17 degrees of freedom
## Multiple R-squared:  0.2041,    Adjusted R-squared:  0.1105
## F-statistic:  2.18 on 2 and 17 DF,  p-value: 0.1436
```

No hay efectos significativos ni de la cafeína ni de la ansiedad, lo que puede resultar un poco confuso.

El problema: el modelo intenta utilizar la misma pendiente que relaciona el habla con la cafeína para ambos grupos.

Regresión lineal **múltiple** : interacción entre variables

Si queremos ajustarlos utilizando líneas con pendientes separadas, necesitamos incluir una interacción en el modelo.
Equivale a ajustar líneas diferentes para cada uno de los dos grupos.



Regresión lineal **múltiple** : interacción entre variables

```
lmResultInteraccion <- lm(
  HablarPubl ~ Cafeina + Ansiedad + Cafeina * Ansiedad,
  data = df
)
summary(lmResultInteraccion)
```

```
##
## Call:
## lm(formula = HablarPubl ~ Cafeina + Ansiedad + Cafeina * Ansiedad,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.385  -7.103  -0.444   6.171  13.458
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                17.43085    5.43012   3.210 0.005461 **
## Cafeina                    -0.47416    0.09664  -4.906 0.000158 ***
## AnsiedadnoAnsioso         -43.44873    7.79141  -5.576 4.17e-05 ***
## Cafeina:AnsiedadnoAnsioso   1.08395    0.12931   8.382 3.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.085 on 16 degrees of freedom
## Multiple R-squared:  0.8524,    Adjusted R-squared:  0.8247
## F-statistic:  30.8 on 3 and 16 DF,  p-value: 7.014e-07
```

Tanto la cafeína como la ansiedad tienen efectos significativos (lo que denominamos **efectos principales/**"*main effects*") y que existe una **interacción** entre la cafeína y la ansiedad.

Regresión lineal **múltiple** : interacción entre variables

- Debemos ser muy cautelosos a la hora de interpretar un efecto principal significativo si también existe una interacción significativa.

- ya que la interacción sugiere que el efecto principal difiere según los valores de otra variable y, por lo tanto, no es fácil de interpretar.

¿Qué hace que un modelo sea "bueno"?

- Queremos que describa bien nuestros datos → que tenga el menor error posible al modelarlos
- Queremos que sea generalizable a nuevos conjuntos de datos → que su error sea lo más bajo posible cuando lo aplicamos a un nuevo conjunto de datos

Estas dos características a menudo pueden contraponerse

¿Puede un modelo ser "demasiado bueno"?

- Normalmente preferiremos un modelo que tenga un error menor.
- **Pero**, a veces resulta que el modelo con el error más bajo suele ser mucho peor a la hora de generalizar a nuevos conjuntos de datos → ***overfitting***

¿Qué significa realmente “**predecir** ”?

- Cuando hablamos de “predicción” en la vida cotidiana, generalmente nos referimos a la capacidad de estimar el valor de alguna variable **antes de ver los datos**.

- Sin embargo, el término se utiliza a menudo en el contexto de la regresión lineal para referirse al ajuste de un modelo a los datos→ los valores **estimados** ($\hat{y}$) se denominan a veces “*predicciones*” y las variables independientes se denominan “*predictores*”.

- Esto tiene una **connotación poco acertada**, ya que implica que nuestro modelo también debería ser capaz de predecir los valores de nuevos puntos de datos en el futuro.

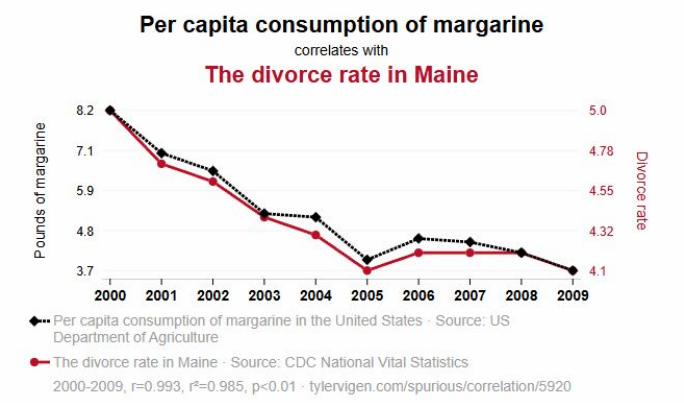
- Los modelos complejos pueden ajustarse excesivamente a los datos, de tal manera que se pueden observar predicciones aparentemente buenas incluso cuando no hay una señal real que predecir.

- Debe considerar con mucho escepticismo las afirmaciones sobre la precisión de las predicciones, a menos que se hayan realizado utilizando los métodos adecuados.

Correlación vs. Causalidad

- Correlación, asociación… no implican causalidad

- Es fácil creer en un vínculo causal, cuando en realidad otros factores están influyendo en los resultados.



https://tylervigen.com/spurious-correlations

- Es difícil establecer causalidad, pero los estudios controlados aleatorizados bien diseñados son el mejor marco disponible → **inferencia causal**

- Los datos observacionales podrían ocultar factores subyacentes que influyan en la aparente relación observada → **variables de confusión** o **factores no observados**

- Los métodos estadísticos permiten **ajustar** con base en otros factores.

Correlación vs. Causalidad

- Correlación, asociación… no implican causalidad

- Es fácil creer en un vínculo causal, cuando en realidad otros factores están influyendo en los resultados.

https://tylervigen.com/spurious-correlations

- Es difícil establecer causalidad, pero los estudios controlados aleatorizados bien diseñados son el mejor marco disponible → **inferencia causal**

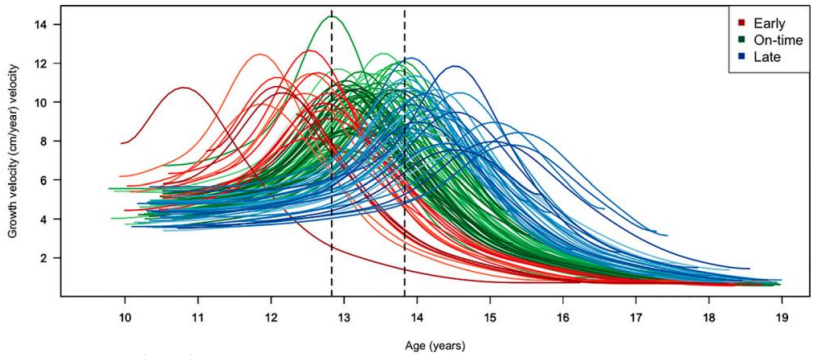
- Los datos observacionales podrían ocultar factores subyacentes que influyan en la aparente relación observada → **variables de confusión** o **factores no observados**

- Los métodos estadísticos permiten **ajustar** con base en otros factores.

Más allá del modelo de regresión lineal

- Es habitual analizar datos cuyos resultados (variables respuesta) son binarios (Sí/No), conteos (número de casos), en lugar de continuos. → *Generalized Linear Models GLM*
  (p.ej. regresión logística, GLM Poisson, GLM Negative Binomial etc.)
  O del tipo "tiempo-hasta-evento" → Análisis de Supervivencia
  (*p.ej. Cox Proportional Hazards Model, Frailty models,*
  Joint models, PAMMs etc.)

- Podemos utilizar la misma maquinaria para modelizar efectos no lineales, i.e. que no siguen una línea recta (como las curvas).



Monasterio et al. (2023), The burden of injuries according to maturity status and timing: A two-decade study with 110 growth curves in an elite football academy.

- Existe una amplia gama de métodos alternativos y sofisticados.
  p.ej. técnicas de regularización, árboles de clasificación y regresión (CARTs, Random Forests), Boosting, máquinas de vectores de soporte, redes neuronales etc.

Caso práctico

La base de datos `rotterdam.xlsx` incluye 2982 pacientes con cáncer de mama primario cuyos registros se incluyeron en el banco de tumores de Rotterdam (Royston and Altman (2013)).

- Investiga la relación entre la edad al diagnóstico (variable `age`) y el tamaño del tumor (variable `size_sim`).
  ¿Existe una relación lineal significativa?¿Cómo cambia el tamaño del tumor por cada año adicional de edad? ¿Qué porcentaje de la variabilidad en el tamaño del tumor explica la edad?

- Analiza cómo múltiples factores influyen en el tamaño del tumor (variable `size_sim`). Ajusta un modelo de regresión múltiple que incluya: edad (`age`), número de nodos positivos (`nodes`) y estado menopáusico (`meno`)
  ¿Mejora la capacidad predictiva al incluir variables adicionales?¿Qué variable tiene mayor efecto sobre el tamaño del tumor?¿Cómo interpretas el coeficiente del estado menopáusico?

- El número de nodos positivos (`nodes`) tiene una distribución sesgada. Explora el efecto de transformar esta variable. Crea la variable `lognodes = log(nodes + 1)`

Fuente: Patrick Royston and Douglas Altman, *External validation of a Cox prognostic model: principles and methods*. BMC Medical Research Methodology 2013, 13:33
Disponible en **R**: `survival::rotterdam`

Referencias I

- Grolemund G., & Wickham H. (2017). *"R for Data Science"*. O'Reilly Media. URL: https://r4ds.hadley.nz/
- McAleer P. (2022). "*A Handy Workbook for Research Methods & Statistics* (0.0.9012)". https://psyteachr.github.io/handyworkbook/
- Harrell F. (2025) "*Biostatistics for Biomedical Research*". https://hbiostat.org/bbr/
- Russell A. Poldrack (2024) "*Statistical Thinking for the 21st Century*" . https://statsthinking21.github.io/statsthinking21-core-site
- Bofill M., Cortés J., Pérez-Hoyos S. & Sánchez A (2019). "Basics Statistics for Biomedical Research" UEB-VHIR & GRBIO. https://github.com/uebvhir/Course_Basic_Statistics-VHIO-2019/

Referencias II

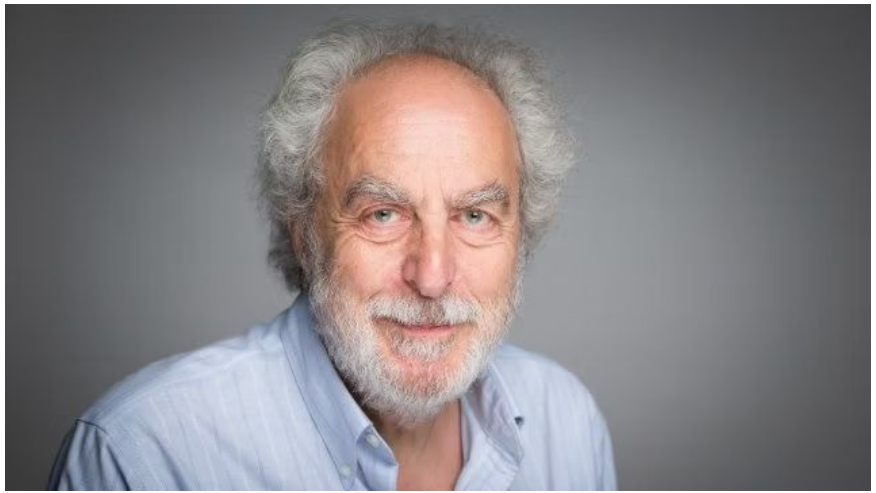
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). "*Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*". European journal of epidemiology, 31(4), 337-350.

- Goodman, S. (2008, July). "*A dirty dozen: twelve p-value misconceptions*". In Seminars in hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.

- Cortés Martínez, J., Casals, M., Langohr, K., & González Alastrué, J. A. (2015). "*Importancia de la potencia y la hipótesis en el valor p*". Medicina clínica, 146(4), 178-181.

- Wasserstein, R. L., & Lazar, N. A. (2016). "*The ASA statement on p-values: context, process, and purpose*". The American Statistician, 70(2), 129-133.

- Vilaró, M., Cortés, J., Selva-O'Callaghan, A., Urrutia, A., Ribera, J. M., Cardellach, F., ... & Cobo, E. (2019). "*Adherence to reporting guidelines increases the number of citations: the argument for including a methodologist in the editorial process and peer-review*". BMC medical research methodology, 19(1), 112.



“We need less research, better research, and research done for the right reasons”at BMJ https://www.bmj.com/content/308/6924/283 “The scandal of poor medical research”

Doug Altman,
Medical statistician & Statistics educator