

Seminario de Ciencia de Datos

Diego Fernández Slezak - Alejo Salles

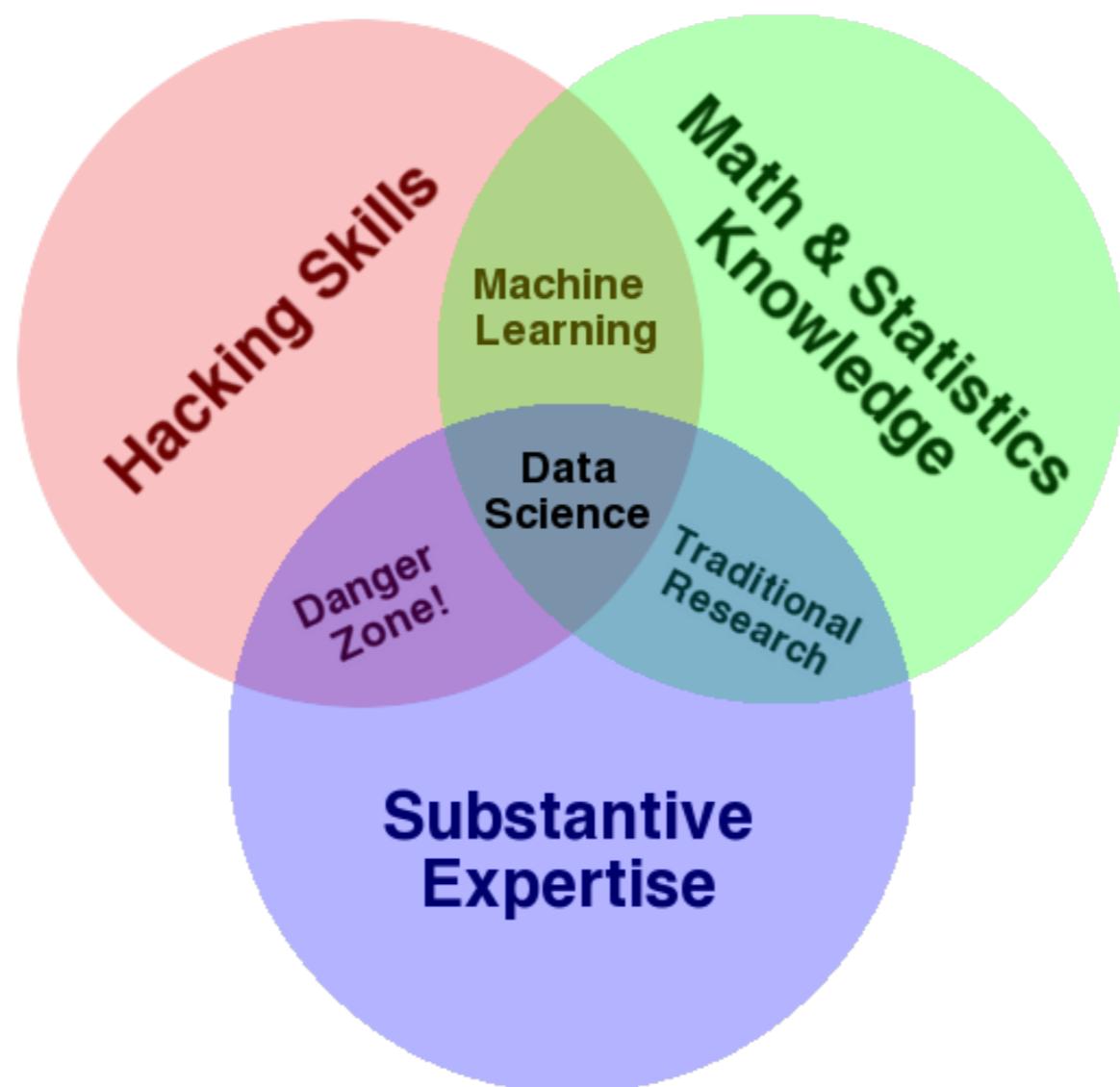
Operativas

- 6 Módulos - 6 Teóricas - 6 Prácticas
- Python (6 librerías!)
- Orientación práctica
- Evaluación

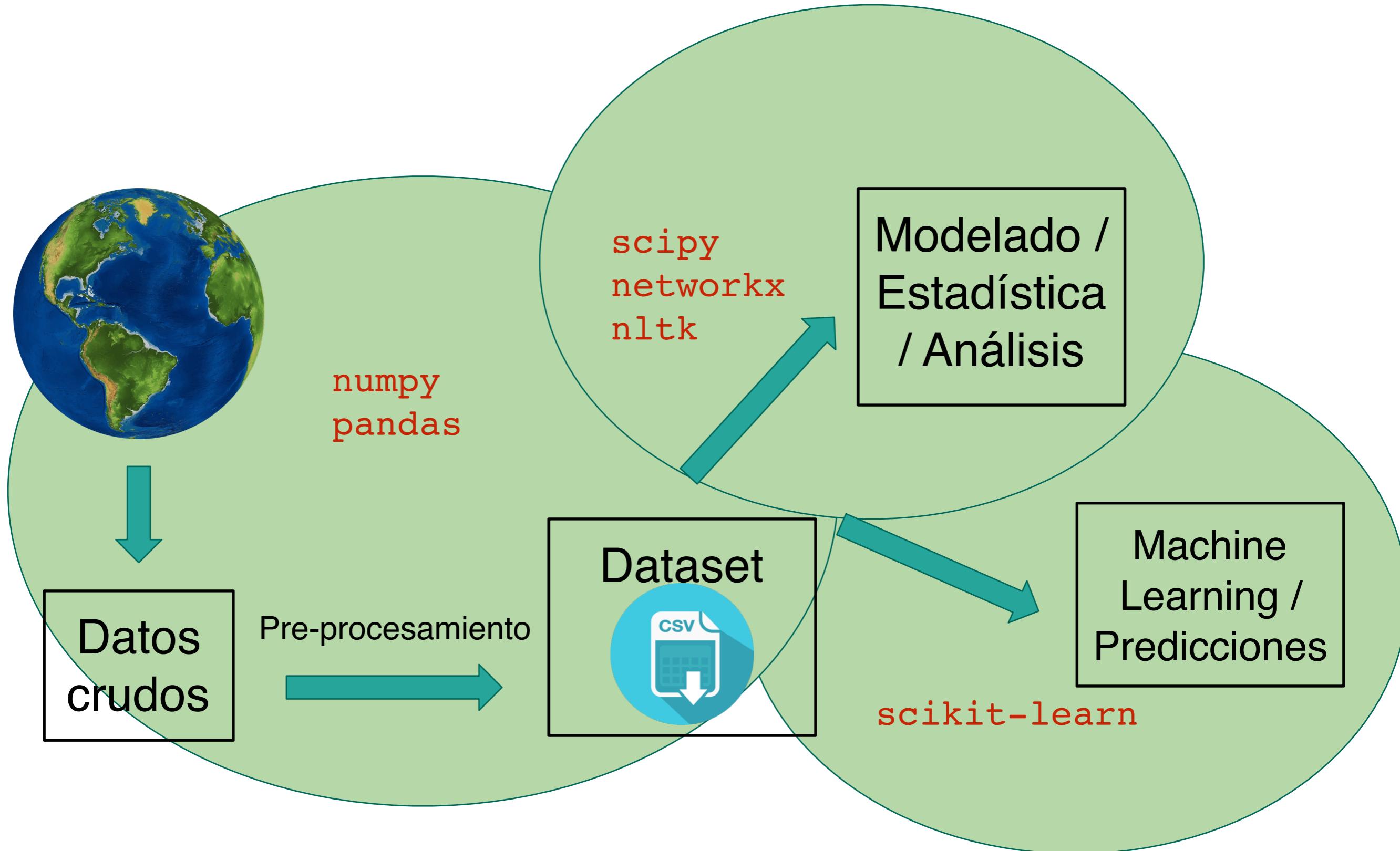
Programa

Módulo	Librería	Duración
Introducción/ Estadística	numpy/scipy	2 Semanas
Series Temporales	pandas	2 Semanas
Probabilidad e Inferencia	-	3 Semanas
Machine Learning	scikit-learn	3 Semanas
Redes Complejas	networkx	3 Semanas
Análisis de Texto	nltk	3 Semanas

¿Qué es Data Science?



¿Qué tareas realiza un Data Scientist?



¿Qué habilidades requiere un Data Scientist?

- Hacer la pregunta correcta
- Saber interpretar los datos y su estructura
- Sintetizar y visualizar conclusiones

Módulo 1: Estadística

Tirada	Resultado
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
...	...



¿Está cargada la moneda?

Hipótesis/Modelo

Estadística: nos permite sacar conclusiones del mundo exterior

Librerías del Módulo

numpy

```
import numpy as np
```

- Manipulación de datos numéricos
- Define **array** (¡no confundir con listas!)
- Subyace muchas librerías

scipy

```
from scipy import stats
```

- Librería “científica”
- Basada en **numpy**
- Special functions, integration, optimization, interpolation, Fourier transforms, signal processing, linear algebra, spatial data structures and algorithms, statistics, ...

Librerías Extra

matplotlib

```
import matplotlib.pyplot as plt
```

- Gráficos
- Originalmente imitaba Matlab, hoy estándar para graficar en python

seaborn

```
import seaborn as sns
```

- ¡Gráficos lindos!
- No hay que hacer nada, sólo importar y usar **matplotlib**

statsmodels Estadística avanzada

Test Binomial

scipy.stats.binom_test

Hipótesis: resultados distribuidos binomialmente

$$p(k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Hipótesis/Modelo

¿Cuál es la probabilidad de obtener m o más caras de n tiradas?

$$\sum_{k \geq m}^n p\left(k \middle| \theta = \frac{1}{2}\right)$$

¿Cuál es la probabilidad de obtener un resultado *tan o más extremo*?

$$\sum_{k=m}^n p\left(k \middle| \theta = \frac{1}{2}\right) + \sum_{k=0}^{n-m} p\left(k \middle| \theta = \frac{1}{2}\right)$$

¡El *p*-valor!

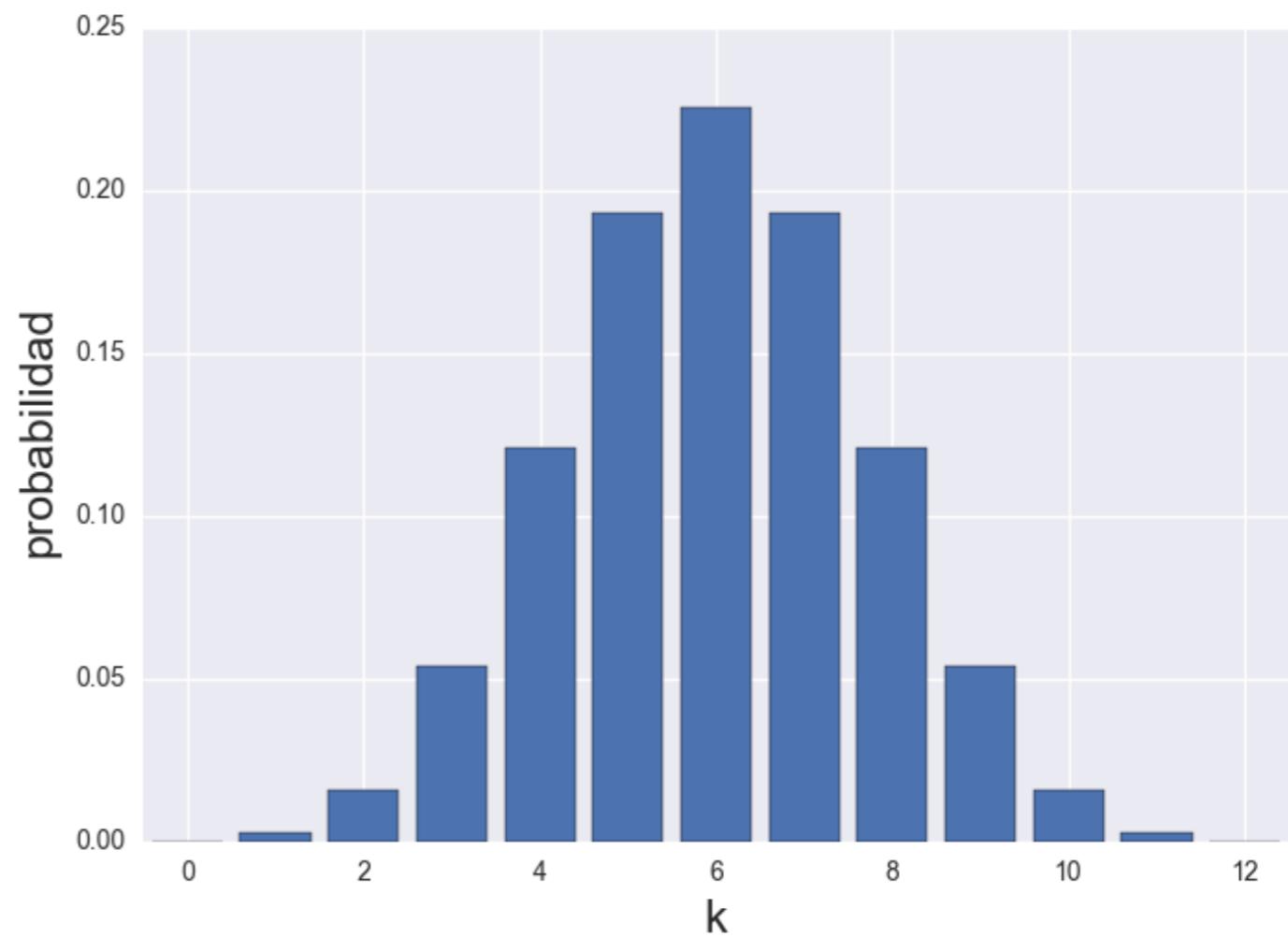
Test Binomial

scipy.stats.binom_test

$$p = \sum_{k=m}^n p\left(k \middle| \theta = \frac{1}{2}\right) + \sum_{k=0}^{n-m} p\left(k \middle| \theta = \frac{1}{2}\right)$$

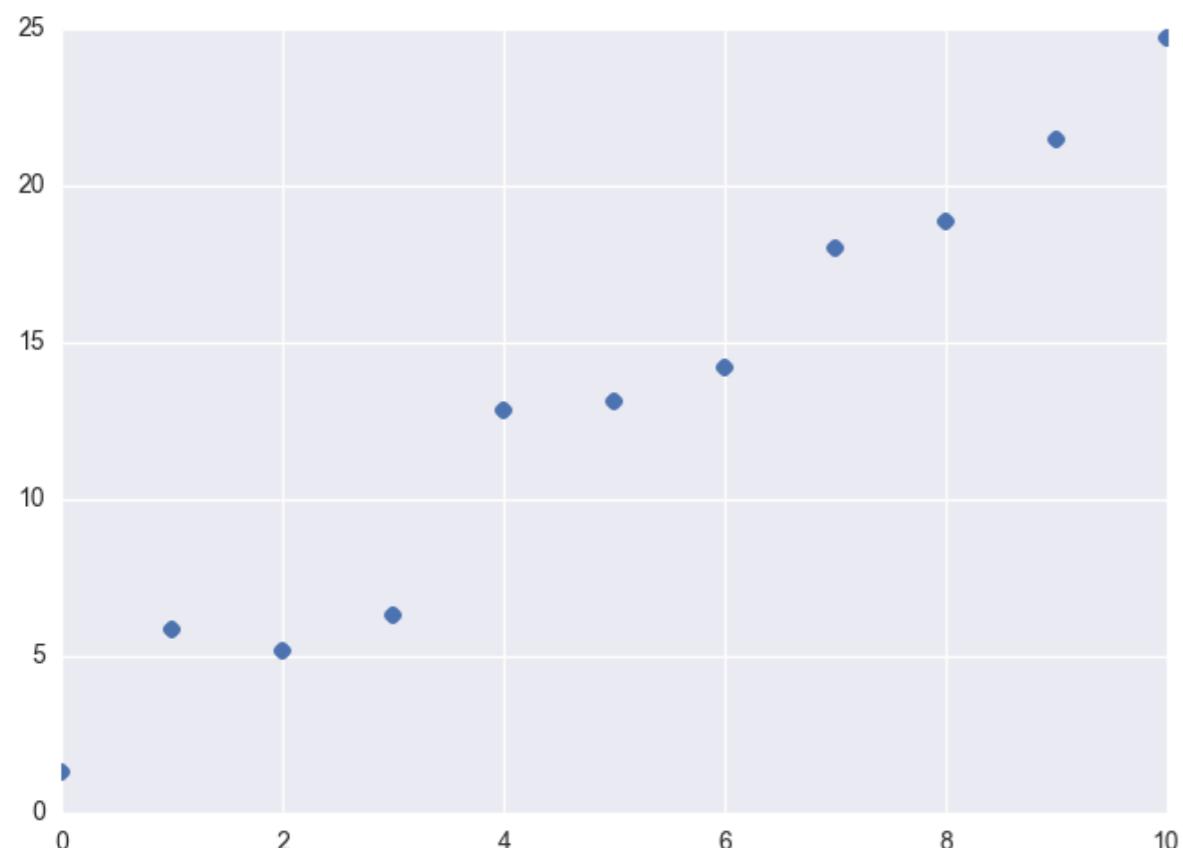
p-valor para 12 caras de 12 tiradas...

$$\begin{aligned} p &= \sum_{k=12}^{12} p\left(k \middle| \theta = \frac{1}{2}\right) \\ &+ \sum_{k=0}^0 p\left(k \middle| \theta = \frac{1}{2}\right) \\ &= 0.00049 \end{aligned}$$

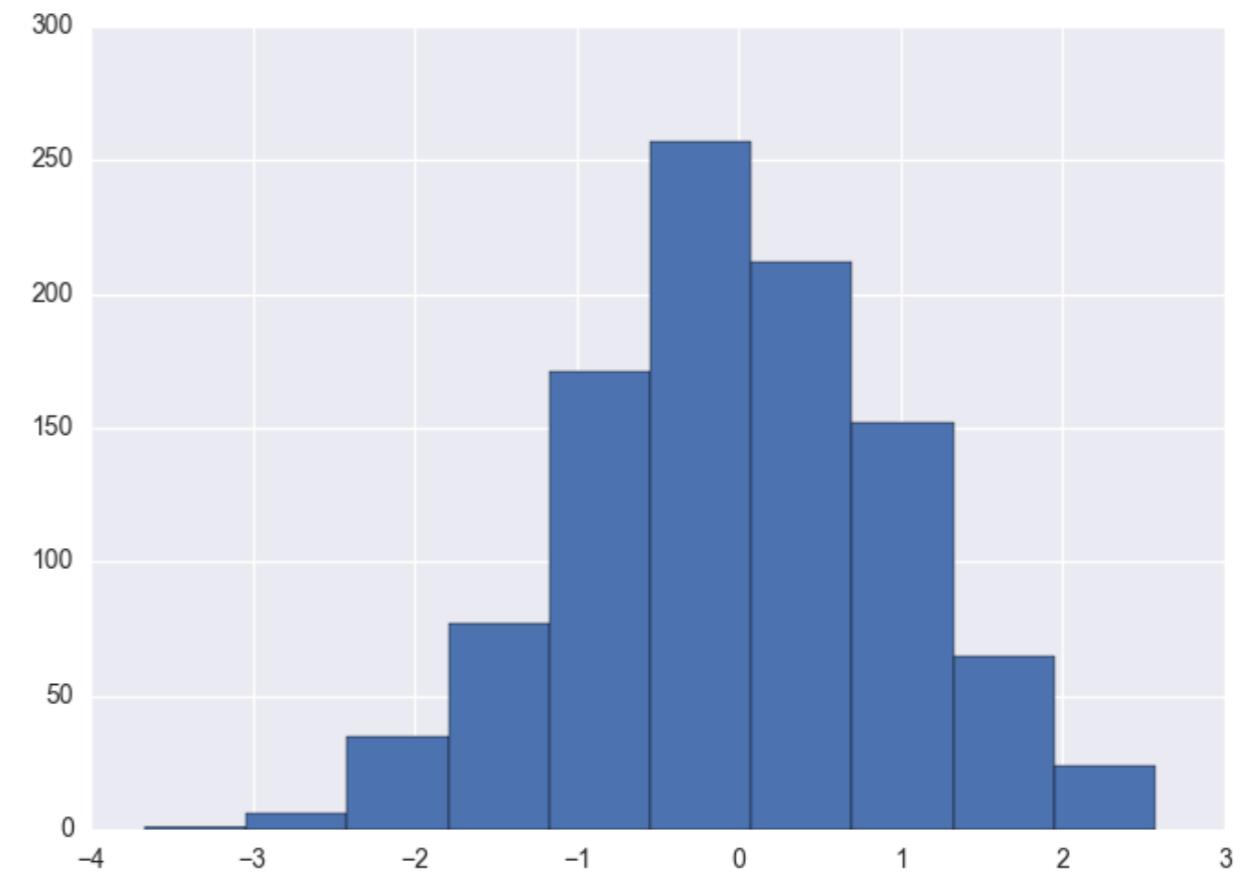


Cómo ver datos

Scatter plots



Histogramas



t-test(s) de Student

- Muestras independientes

`scipy.stats.ttest_ind`

- 1 muestra

`scipy.stats.ttest_1samp`

- Muestras apareadas

`scipy.stats.ttest_rel`

Como antes: una cola o dos colas, según el caso...

t-test(s) - Hipótesis

- Variables distribuídas normalmente (Gaussianas)
 - Shapiro-Wilk o Kolmogorov-Smirnov
- Igual varianza
 - F-test, y si no, Welch t-test
- Muestras independientes

Generalmente, cuanto más fuertes las hipótesis, más poderoso el test

t-test

Muestras independientes

scipy.stats.ttest_ind

Alturas (cm)

mujeres	varones
162	181
171	187
157	161
165	
174	
...	

¿Cuál es la probabilidad de que las muestras vengan de una distribuciones con igual media?

$$t = 1.445, p = 0.199$$

Estadístico

Distribución conocida

t-test

Muestra única

`scipy.stats.ttest_1samp`

Alturas (cm)

161

171

157

181

187

...

Sabemos que la altura media de la gente en Noruega es 175 cm.
¿Es distinta nuestra altura?

¿Cuál es la probabilidad de que las muestras vengan de una distribución con media distinta de cero?
(u otro valor fijo de referencia)

t-test

Muestras apareadas

scipy.stats.ttest_rel

Alumno	Prueba 1	Prueba 2
1	6	8
2	7	9
3	10	9
4	9	10
5	5	5
...

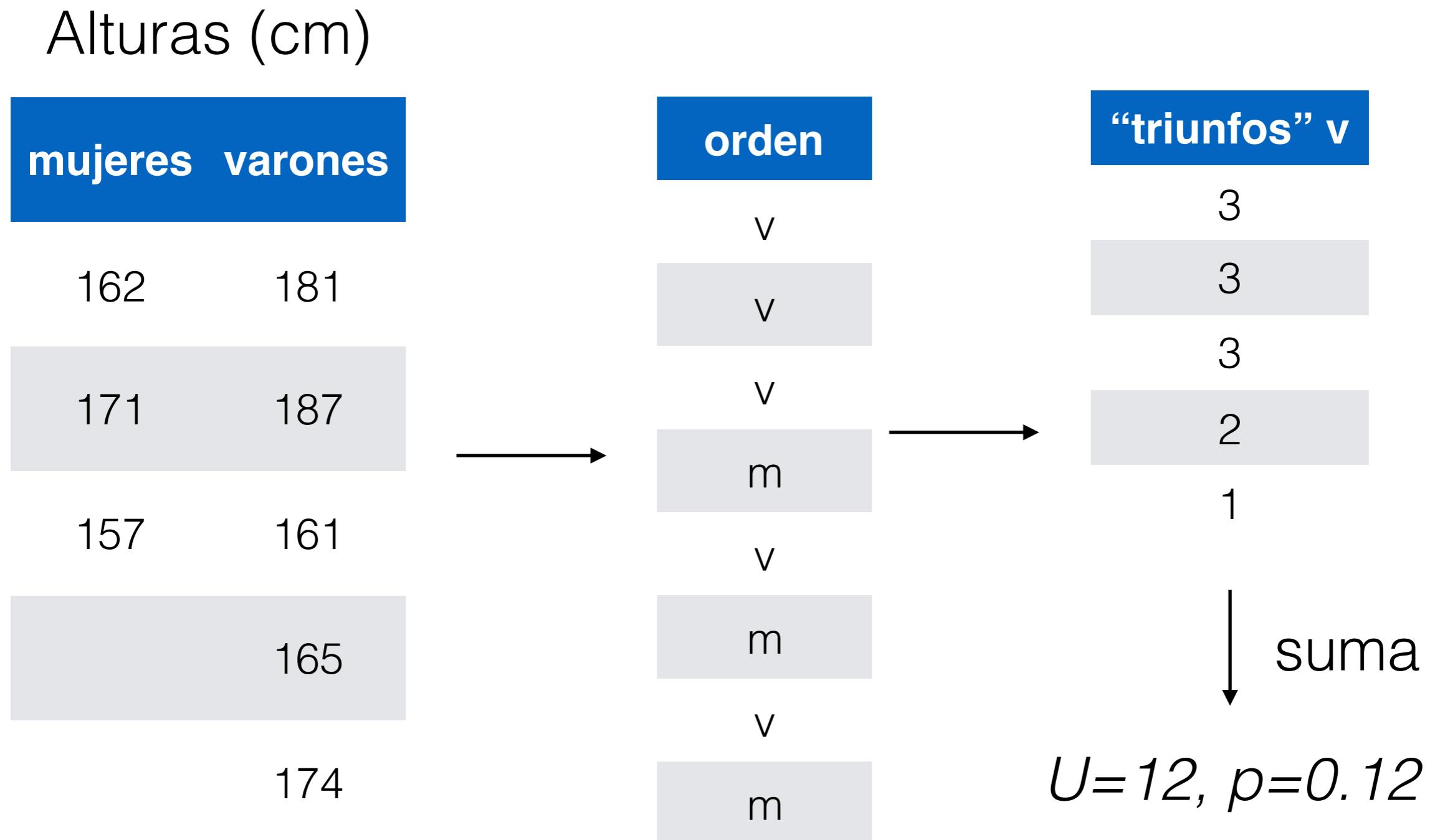
¿Cuál es la probabilidad de que las muestras vengan de distribuciones con igual media?

Más poderoso que el de 2 muestras independientes

Tiene que haber una “constante” entre muestras (aquí el alumno)

¿Y si no valen las hipótesis? ¡Tests no-paramétricos!

Tests de rank sum, Mann-Whitney U, Wilcoxon



Tests de rank sum, Mann-Whitney U, Wilcoxon

- Muestras independientes
- Muestras apareadas

`scipy.stats.ranksums`

`scipy.stats.mannwhitneyu`

`scipy.stats.wilcoxon`

Test de Permutaciones

Alturas (cm)

mujeres	varones
162	181
171	187
157	161
165	
174	

altura	etiqueta
162	m
171	m
157	m
181	v
187	v
161	v
165	v
174	v

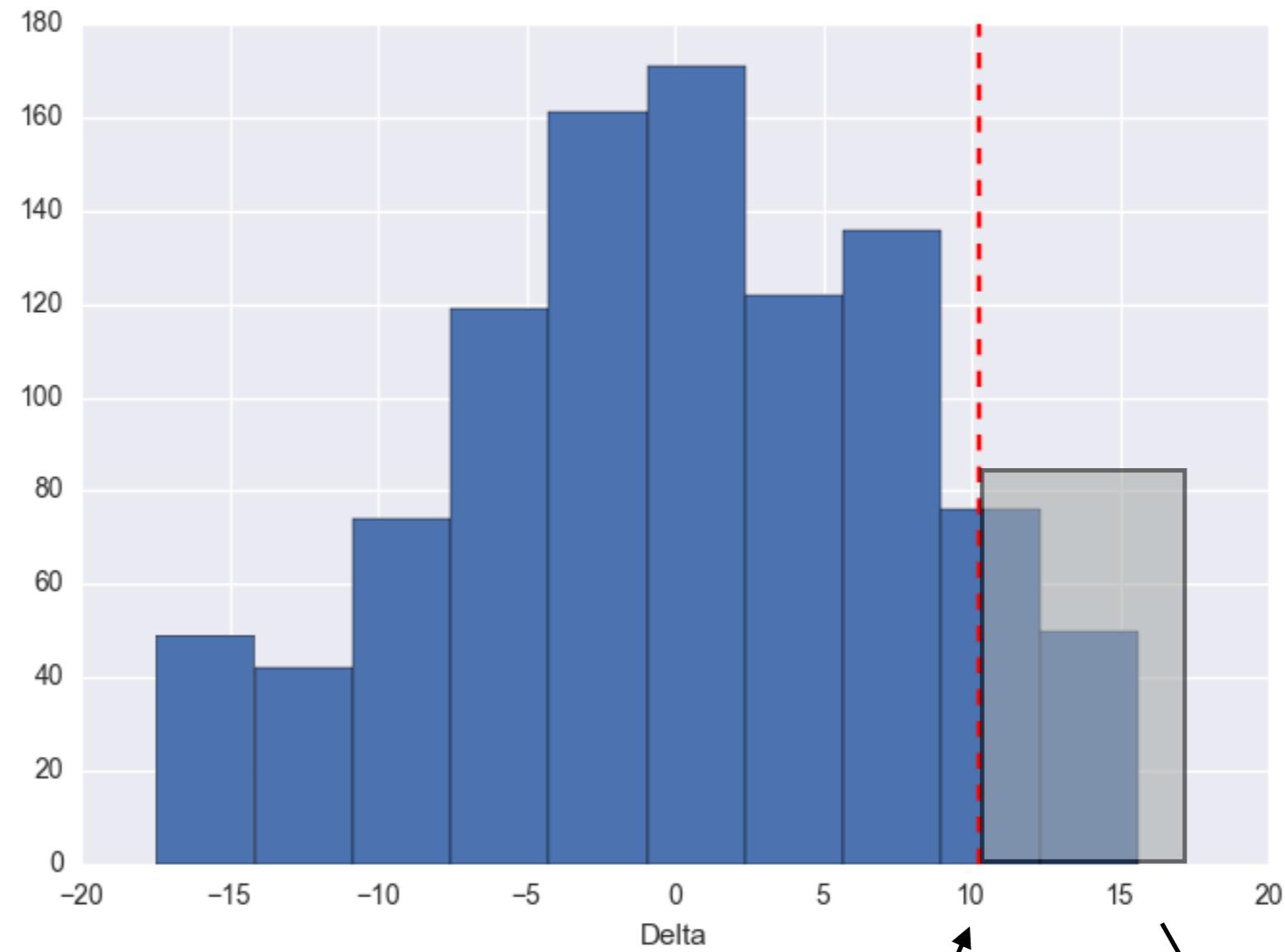
altura	etiqueta
162	m
171	v
157	v
181	m
187	v
161	v
165	v
174	m

$$\Delta_0 = \bar{h}_v - \bar{h}_m = 10.27$$

$$\Delta_1 \ \Delta_2 \ \dots$$

Test de Permutaciones

altura	etiqueta
162	m
171	m
157	m
181	v
187	v
161	v
165	v
174	v



$$\Delta_0 = \bar{h}_v - \bar{h}_m = 10.27$$

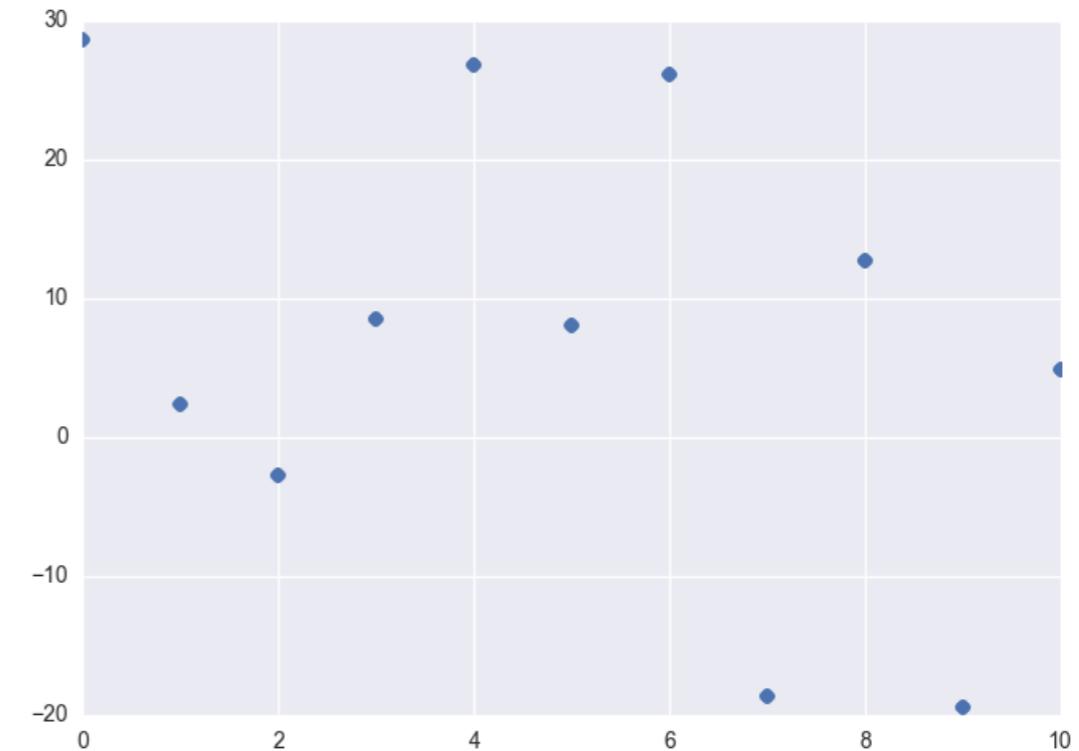
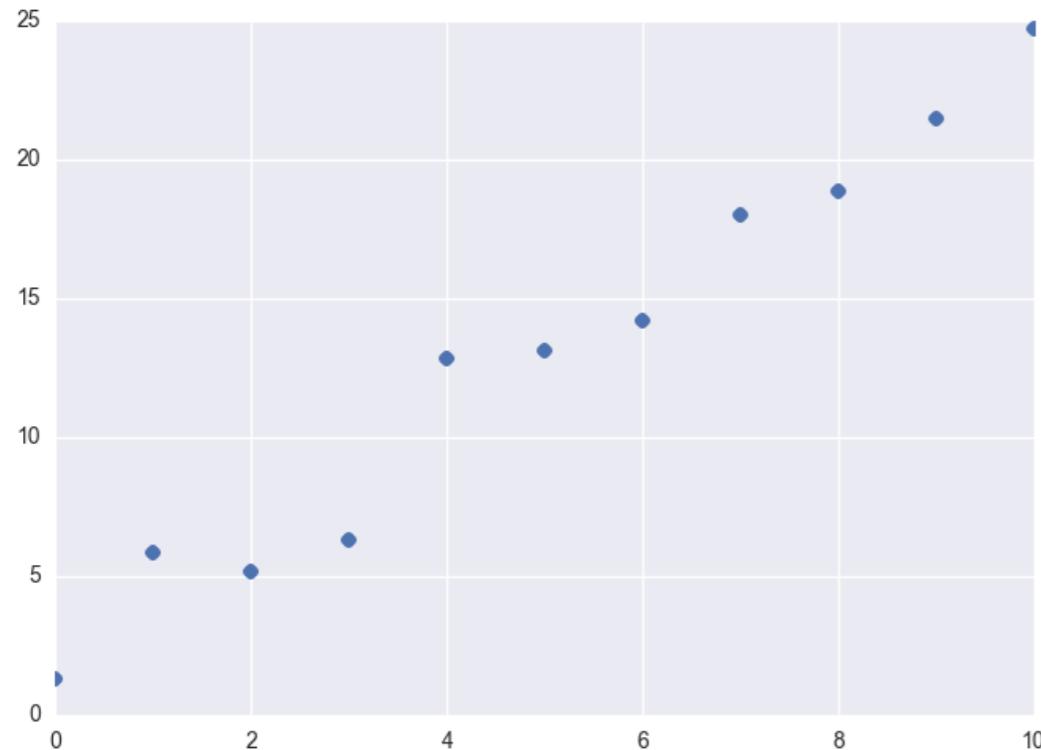
$$p=0.11$$

Correlación de Pearson

scipy.stats.pearsonr

$$\rho \in [-1, 1]$$

Mide correlación *lineal* de los datos, asumidos normales



$$\rho = 0.98$$

$$p < 10^{-6}$$

$$\rho = -0.39$$

$$p = 0.24$$

ANOVA

`scipy.stats.f_oneway`

mujeres varones otros

162 181 167

171 187 188

157 161 172

165

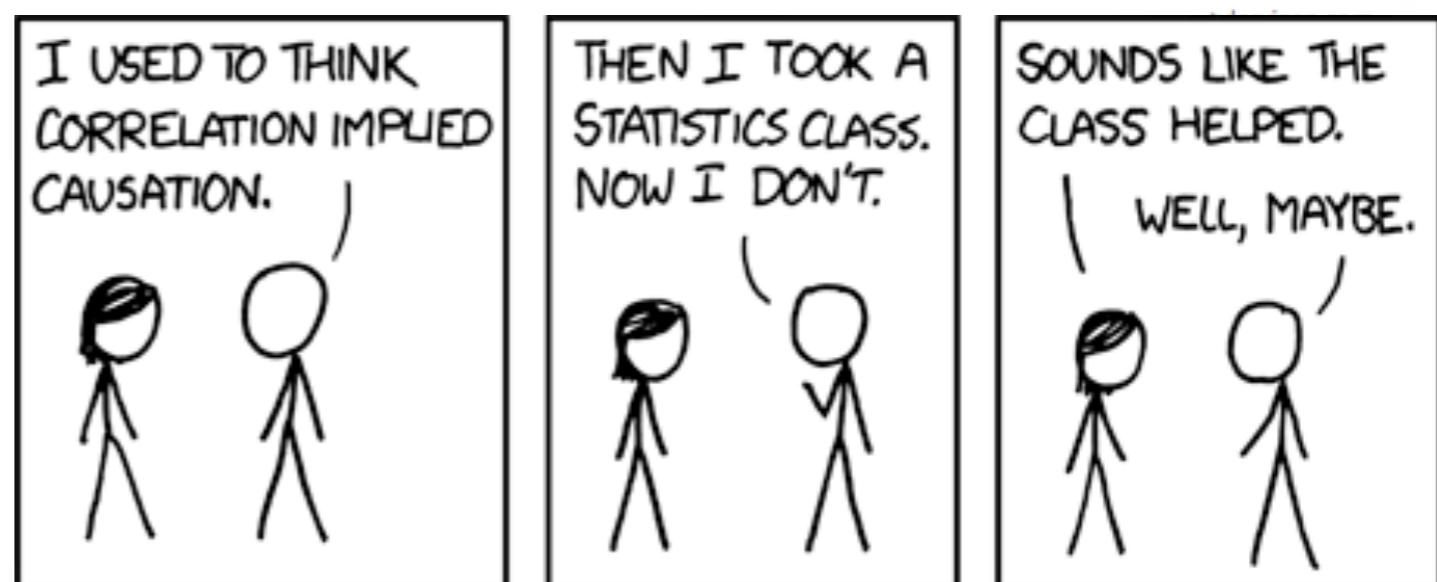
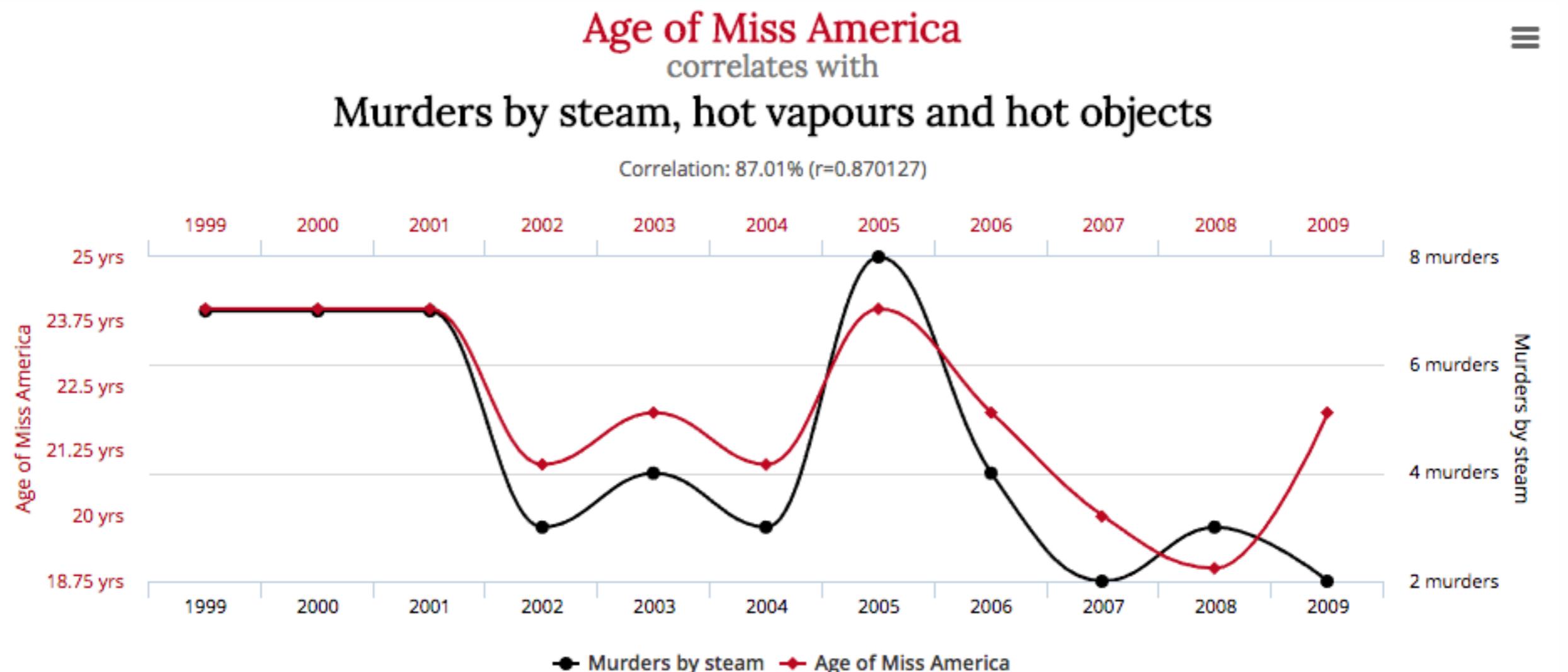
174

Test para una variable
categórica con
múltiples valores

Generalización del t-test a
muchos grupos

Variantes: N-way, ANCOVA, ...

¡Atención! correlación $\not\Rightarrow$ causalidad



¡Atención!

Falta de evidencia \neq inexistencia del efecto

mujeres varones

162 181

171 187

157 161

165

174

...

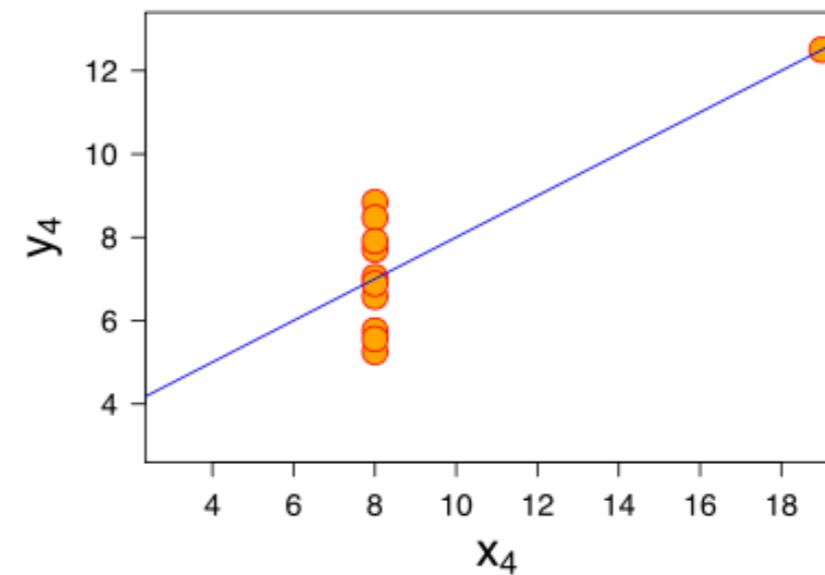
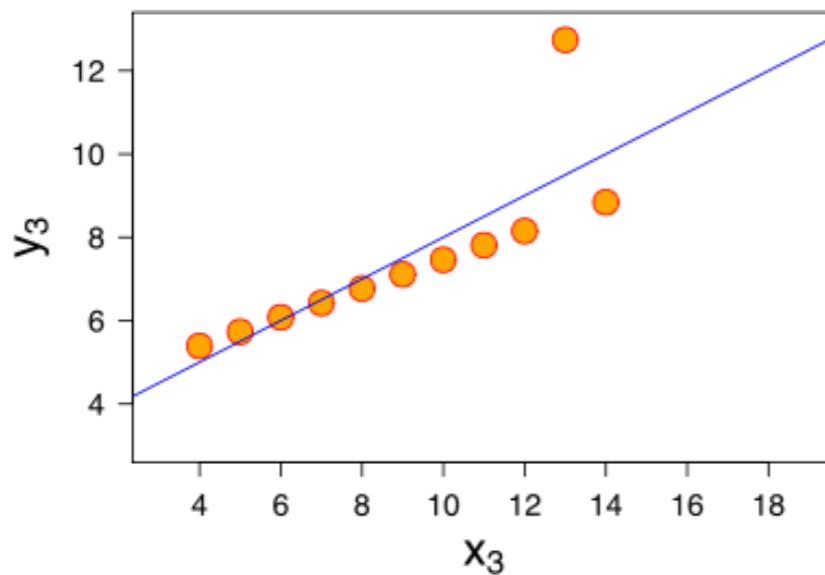
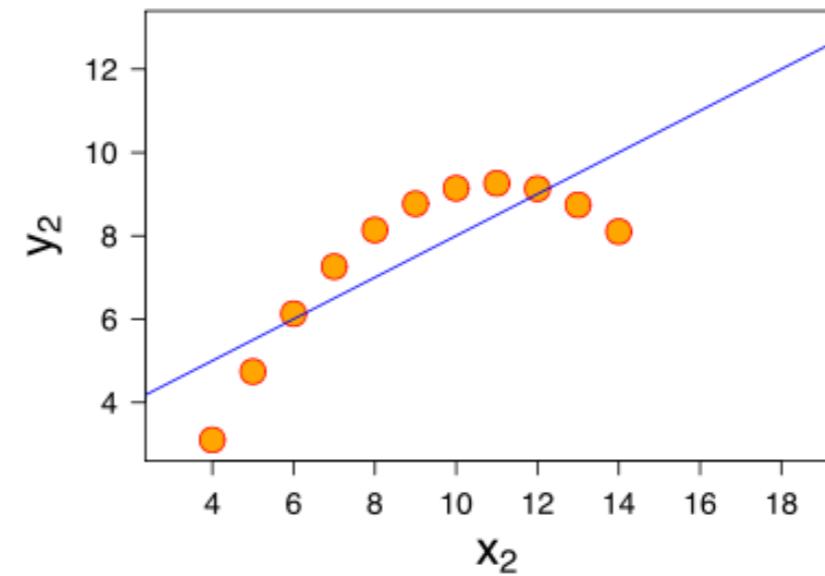
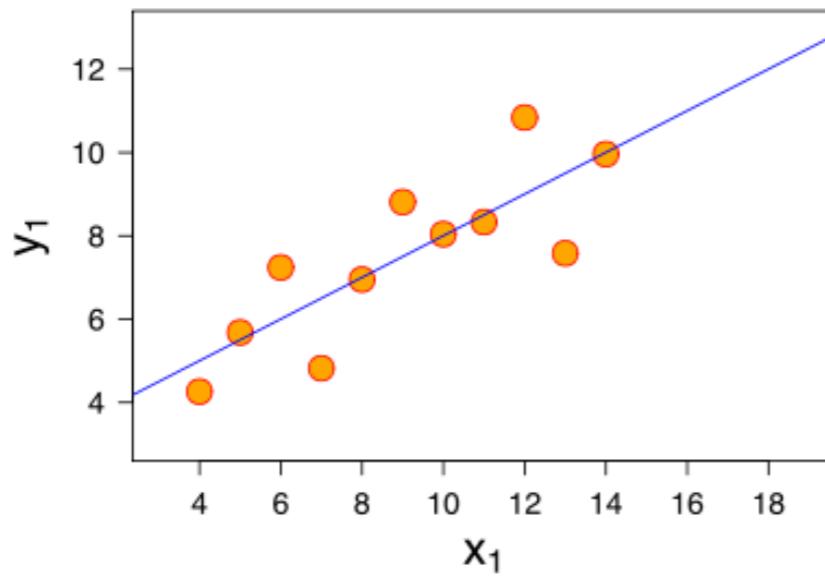
$$t = 1.445, p = 0.199, t\text{-test}$$

¿Qué significa este resultado?

¿Que los varones no son más altos que las mujeres?

No: que no tenemos suficiente evidencia para afirmar que sí lo son

Anscombe's Quartet



Misma x media, y media, varianza de x, varianza de y,
correlación entre x e y, regresión lineal

Cheat sheet

- `numpy.loadtxt(...)`
- Para acceder `numpy.array`: `data[i]`,
`data2D[i,j]`, `data2D[i][j]`
- Operador de slice: `data[:]`, `data2D[:, i]`
- `plt.plot()`, `plt.bar()`, `plt.hist()`
- `plt.show()`!