Análisis de datos

Sugerencias

1. Acomoden los datos (Matlab, R, Python)
2. Grafiquen variables conocidas
3. Filtren datos "malos"
4. Miren las distribuciones

Hipótesis nula

Definición: hipótesis que el investigador trata de refutar, rechazar o anular.

Ejemplo:

- Quiero probar que algo tiene una mayor tasa de crecimiento cuando cambio X parámetro.
- Hipótesis nula: ese algo no presenta una mayor tasa de crecimiento cuando se cambia el parámetro.

Student t-test

Sirva para comparar dos distribuciones normales

One-sample t-test
Hipótesis nula: la media poblacional es μ_0 $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Paired-sample t-test (X,Y)
Hipótesis nula: la media de X-Y es μ_0

Two-sample t-test (X,Y)
Hipótesis nula: la media de X-Y es μ_0

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Matlab

```
[h,p] = ttest(x,mu)
[h,p] = ttest(x,y)
[h,p] = ttest2(x,y)
```

p-value

Wikipedia:

- the **p-value** is the **probability** of obtaining a **test statistic result** at least as extreme or as close to the one that was actually observed, assuming that the **null hypothesis is true.**

Wilcoxon Rank Sum test

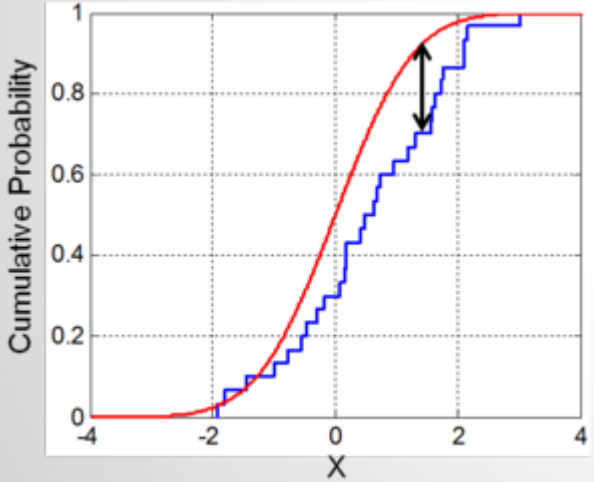
Test no-paramétrico. Sirve para comparar dos distribuciones, si sus medias son iguales.

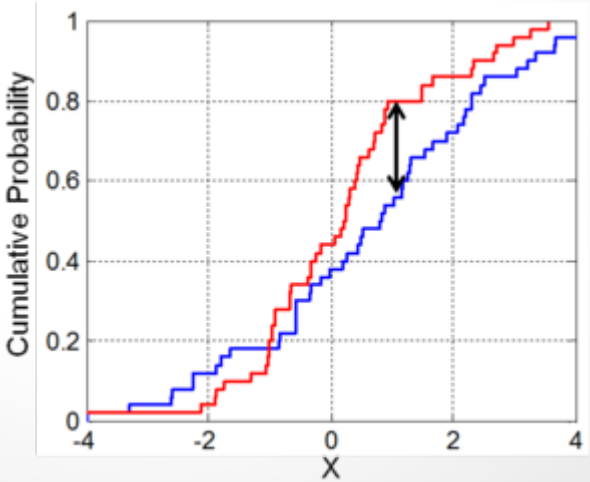
Alternativa cuando las distribuciones no son normales.

Matlab: `p = ranksum(x,y)`

Kolmogorov-Smirnov

Test no-paramétrico. Sirve para comparar dos distribuciones





Matlab

```
[h,p] = kstest(x)

[h,p] = kstest2(x,y)
```

ANOVA

Técnica para comparar medias de dos o más grupos de poblaciones

Precondiciones:

- Residuos con distribución normal
- Muestras independientes
- Misma varianza igual entre poblaciones.
- Dentro de un grupo, las respuestas son independientes

ANOVA de dos vías

Técnica para comparar influencia de dos variables independientes con otra dependiente.

Analiza la interacción

Precondiciones:

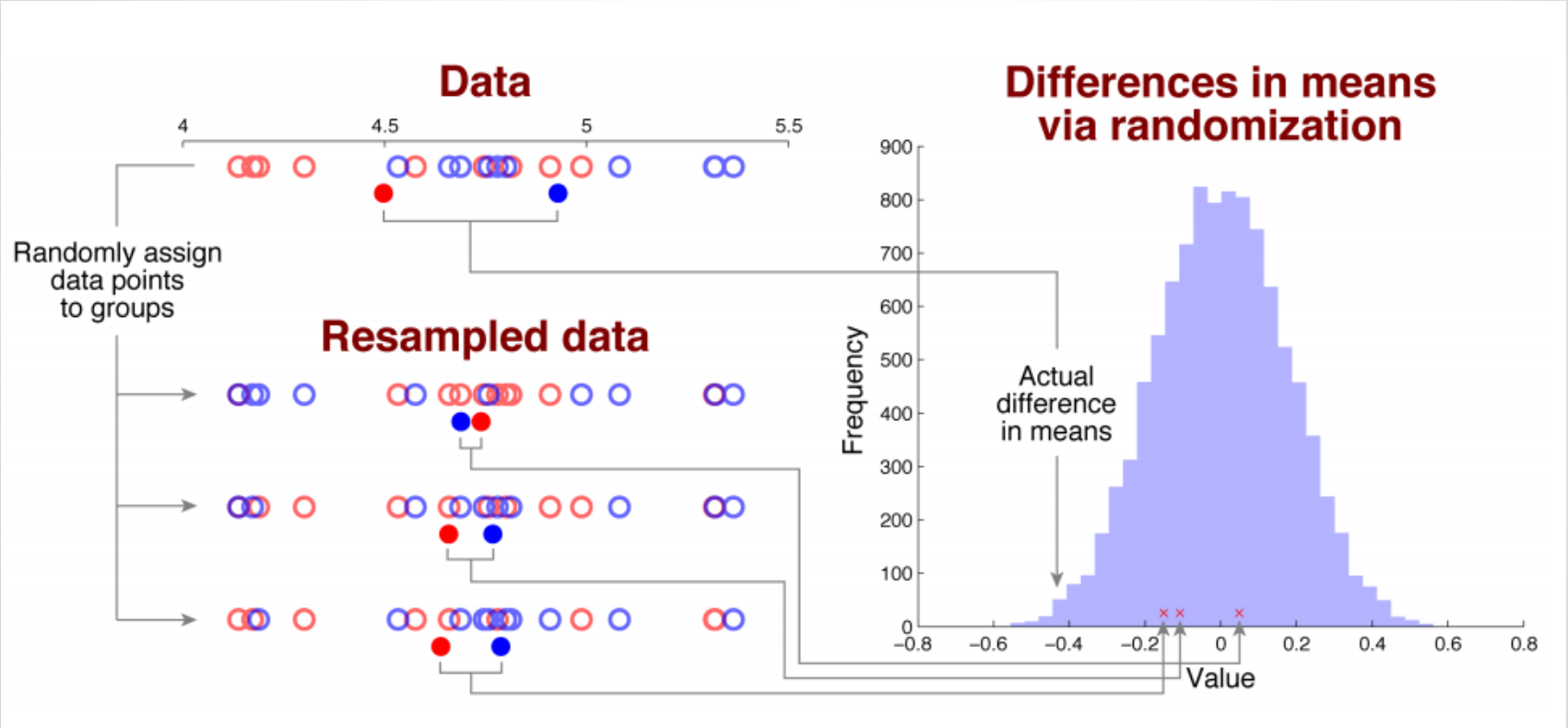
- Residuos con distribución normal
- Muestras independientes
- Misma varianza igual entre poblaciones.
- Dentro de un grupo, las respuestas son independientes

Bootstrapping

Método estadístico de muestreo al azar.

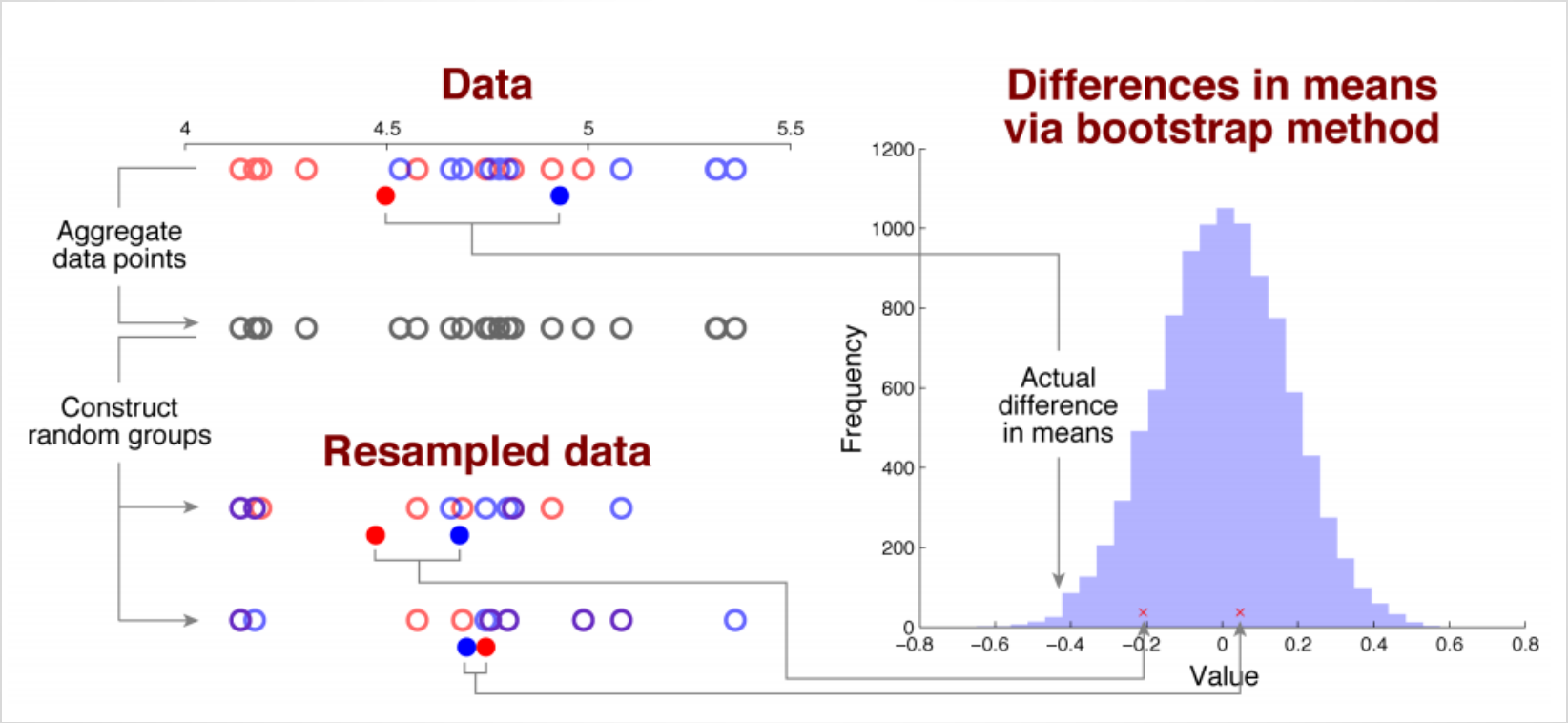
Sirve para medir precisión o robustez de estimaciones.

Randomization



https://talks.stanford.edu/psychology/psych-216a/

Resampling



https://talks.stanford.edu/psychology/psych-216a/

Ejemplo Bootstrapping

Atributos	**SxMxC** KruskallWallis	**SxM** Willcoxon	**SxC** Willcoxon	**MxC** Willcoxon
N	0.0132	0.0017	0.1234	0.2345
E	0.0128	0.0016	0.0985	0.2786
LCC	0.0017	0.0002	0.0058	0.0769
LSC	0.0464	0.0281	0.0789	0.3282
ATD	0.0213	0.0148	0.0207	0.7209
PE	0.0017	0.0002	0.0580	0.0769
L1	0.3166	1	0.4667	0.4667
L2	0.4941	0.2513	0.4104	0.9310
L3	0.2810	0.1958	0.7273	0.1935

Tab. 3.3: Tests sobre el grafo *naive*, los valores en rojo son p-values significativos ($p < 0{,}05$).

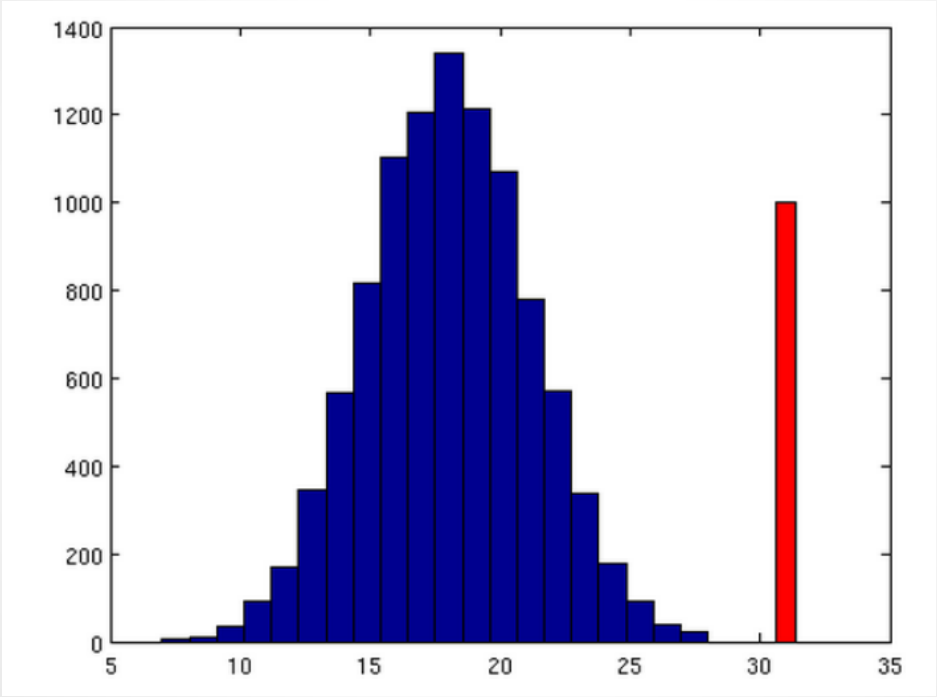
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	1	0
1	1	1	0
1	0	0	1
0	0	0	0
1	1	0	0

trabajo original

1	1	0	0
1	1	0	0
1	1	1	0
1	1	0	0
1	1	1	0
1	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0

naive

Ejemplo Bootstrapping



Regresión lineal

Se busca crear un modelo lineal:
respuestas *y* como función lineal de variables independedientes as Xi (predictores).

$$y = \beta 0 + \sum \beta i Xi + \varepsilon i$$

donde β representa los parámetros lineales estimados y ε los términos de error.

Matlab:

```
b = regress(y,X)
```

```
mdl = fitlm(X,y)
```

returns a linear model of the responses y, fit to the data matrix X.

Principal Component Analysis

1. Recolectar los datos a analizar.
2. Restar a los datos obtenidos la media.
3. Calcular la matriz de covarianza entre los datos
4. Calcular los autovalores y autovectores de la matriz de covarianza.
5. Seleccionar los autovectores (componentes principales) que se deseen

Matlab:

```
coeff = pca(X)
```