

# 3 Recol·lecció de dades i mostres

Mètodes empírics 2

25/04/2023

# Avui

- Mostres
- Control
- On són les dades?
- Distribucions

# Mostres

**Què és una mostra? Dóna un exemple**

# Tipus de mostres

**Mostra completa:** tota la població d'interès

**Mostra representativa/sense biaix:** presa de la mostra completa amb un mètode que no depèn de la mostra que s'està prenent

**Mostra no representativa/amb biaix:** les dades són influenciades pel mètode de presa

# Tipus de mostres

**Per la pregunta**

**"És la l-velaritzada (vs. no velaritzada) un fonéma del Català?"**

**dóna un exemple d'una mostra representativa i una no representativa**

# Grandària de mostra

**Per la pregunta**

**"És la l-velaritzada (vs. no velaritzada) un fonéma del Català?"**

**quina diferència fa si tinc una mostra de 5 persones o una de 1000?**

# Grandària de mostra ( $p = 0.52$ )

Assumeix que la probabilitat que una paraula curta sigui usada per a (i) un significat freqüent és 0.52 vs. (ii) un significat menys freqüent.

**Què tan gran ha de ser la mostra per detectar aquesta diferència?**





# Grandària de mostra ( $p = 0.65$ )

Assumeix que la probabilitat que una paraula curta sigui usada per a (i) un significat freqüent és 0.65 vs. (ii) un significat menys freqüent.



# Grandària de muestra ( $p = 0.9$ )

Assumeix que la probabilitat que una paraula curta sigui usada per a (i) un significat freqüent és 0.9 vs. (ii) un significat menys freqüent.



- Tant biaix com grandària importen
- Però quant importen es respon en funció de la pregunta i l'efecte que esperes, a priori

# Control

# Estudis pilot

Versió a petita escala de la teva anàlisi

- Comproveu si el pla d'anàlisi és realitzable
- Comprova la qualitat del pla d'anàlisi, però no necessàriament la seva sensibilitat (en funció de la mida de l'efecte)
- Estalvia temps i diners



# Simulacions

Versió idealitzada de la teva anàlisi

- Comprova si el pla d'anàlisi (sense recol·lecció) és realitzable
- Comprova la qualitat del pla d'anàlisi i la seva sensibilitat
- Cec a problemes de recol·lecció; només és tan bo com les teves suposicions
- Estalvia (més) temps i diners

# On són les dades?

# Experimentales

- Laboratori (eye-tracking, fMRI)
- Plataformes online (MTurk, Prolific, Google Forms)
- Recol·lecció "al carrer"
- Dades de previs estudis (SWOW, NoRaRe, SimLex-999, VisualGenome)

# Dades no/semi estructurades

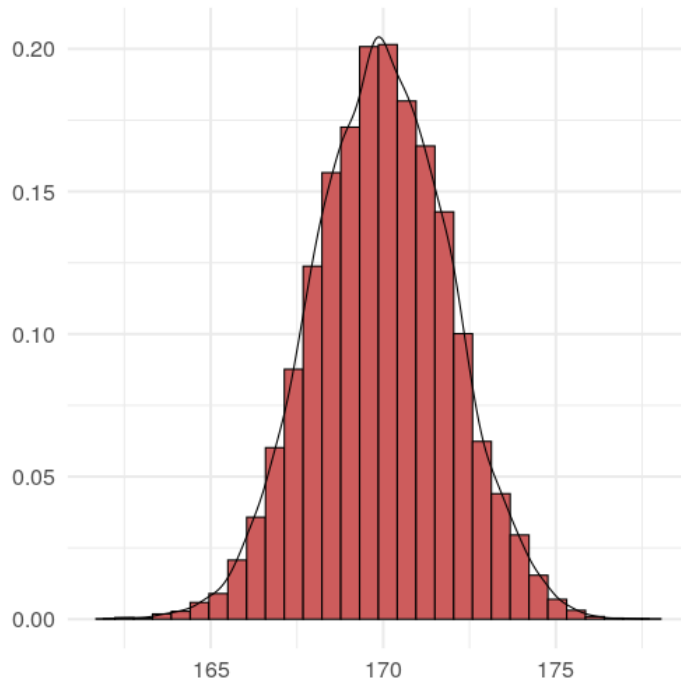
- Corpora (Wikipèdia, Twitter)
- Scraping
- Dades de previs estudis (CLICS)
- Models (word embeddings; language models)

# Manipulació de dades experimentals amb R

# Distribucions

## **Què és una distribució?**

# Distribucio normal (Gaussiana)



$y \sim \text{Normal}(\mu, \sigma)$

$y \sim \text{Normal}(\text{mitja, desv. est.})$

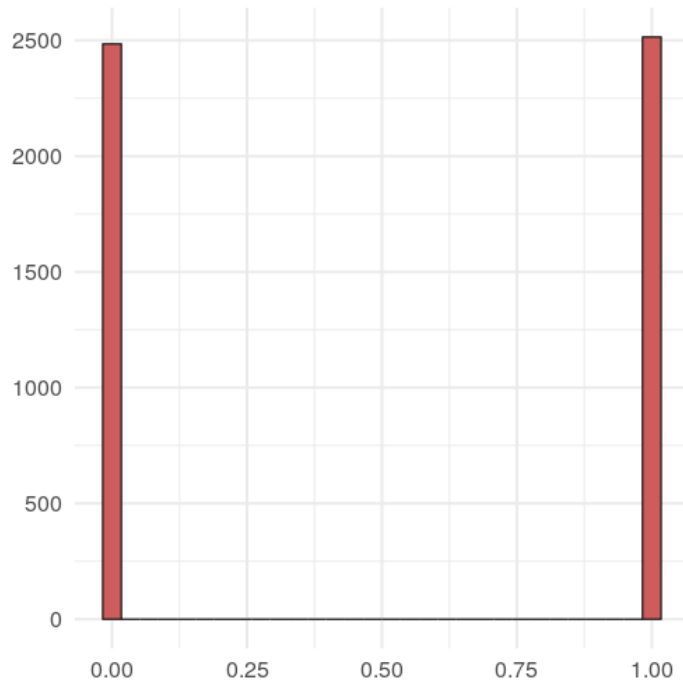
$y \sim \text{Normal}(170, 2)$



# Distribucio normal (Gaussiana)

- **Valors possibles:** Nombres reals
- **Paràmetres:** mitjana, desviació estàndard
- Comú "a la natura" i epistèmicament lleugera

# Distribucio Bernoulli (Binomial)



$y \sim \text{Bernoulli}(p)$

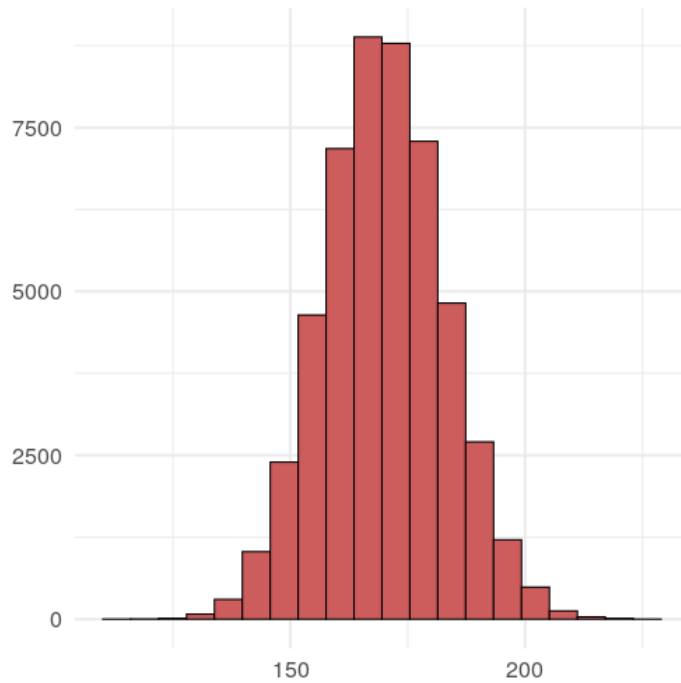
$y \sim \text{Bernoulli}(\text{prob. d'èxit})$

$y \sim \text{Bernoulli}(0.5)$

# Distribucio Bernoulli (Binomial)

- **Valors possibles:** 0 o 1
- **Paràmetres:** probabilitat d'èxit (  $p$  )
- Comú en experiments i ciències socials

# Distribucio Poisson



$y \sim \text{Poisson}(\lambda)$

$y \sim \text{Poisson}(\text{ritme})$

$y \sim \text{Poisson}(170)$

# Distribucio Poisson

- **Valors possibles:** nombres naturals + 0
- **Paràmetres:** ritme ( expectativa de mitja )
- Comú en ciències socials, quan comptem esdeveniments
- La seva variància és igual al seu ritme/mitjana  $\Rightarrow$   
la seva desviació estàndard és  $\sqrt{\lambda}$

# Propera sessió

- Lliurament de "Assignment 2" (08:00 AM 02/05)
- 

- **Introducció a la regressió**