3 Recolección de datos y muestras

Métodos empíricos 2

26/04/2022

Hoy

- Muestras
- Control
- ¿Dónde están los datos?
- Distribuciones

Muestras

Tipos de muestras

Muestra completa: toda la población de interés

Muestra representativa/sin sesgo: tomada de la muestra completa con un método que no depende de la muestra que se está tomando

Muestra no representativa/con sesgo: los datos son influenciados por el método de toma

Tipos de muestras

¿Es la l-velarizada (vs. no velarizada) un fonéma del Catalán?

Tamaño de muestra

¿Es la l-velarizada (vs. no velarizada) un fonéma del Catalán?

Tamaño de muestra (p = 0.52)

Asume que la probabilidad que una palabra corta sea usada para (i) un significado frecuente es 0.52 vs. (ii) un significado menos frecuente.

Tamaño de muestra (p = 0.65)

Asume que la probabilidad que una palabra corta sea usada para (i) un significado frecuente es 0.65 vs. (ii) un significado menos frecuente.

Tamaño de muestra (p = 0.9)

Asume que la probabilidad que una palabra corta sea usada para (i) un significado frecuente es 0.9 vs. (ii) un significado menos frecuente.

- Tanto sesgo como tamaño importan
- Pero cuánto importan se responde en función a la pregunta y el efecto que esperas, a priori

Control

Estudios piloto

Versión a pequeña escala de tu análisis

- Comprueba si el plan de análisis es realizable
- Comprueba la calidad del plan de análisis, pero no necesariamente su sensibilidad (en función al tamaño del efecto)
- Ahorra tiempo y dinero

Simulaciones

Versión idealizada de tu análisis

- Comprueba si el plan de análisis (sin recolección) es realizable
- Comprueba la calidad del plan de análisis y su sensibilidad, pero es ciego a problemas de recolección y es sólo tan bueno como tus suposiciones
- Ahorra (más) tiempo y dinero

¿Dónde están los datos?

Experimentales

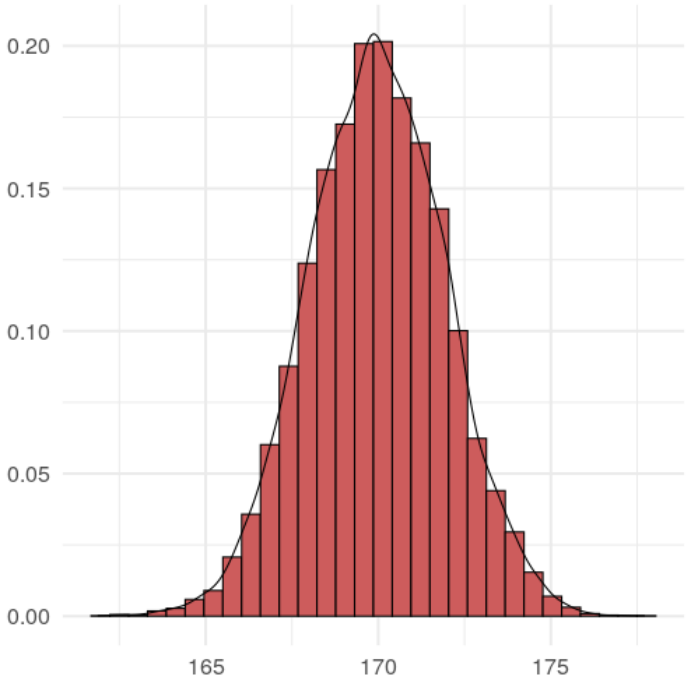
- Laboratorio (eye-tracking, fMRI)
- Plataformas online (MTurk, Prolific)
- Recolección "en la calle"
- Datos de previos estudios (SWOW, NoRaRe, SimLex-999, VisualGenome)

Datos no/semi estructurados

- Corpora (Wikipedia, Twitter)
- Scraping
- Datos de previos estudios (CLICS)
- Modelos (word embeddings; language models)

Distribuciones

Distribucion normal (Gaussiana)



y ~ Normal(μ, σ **)**

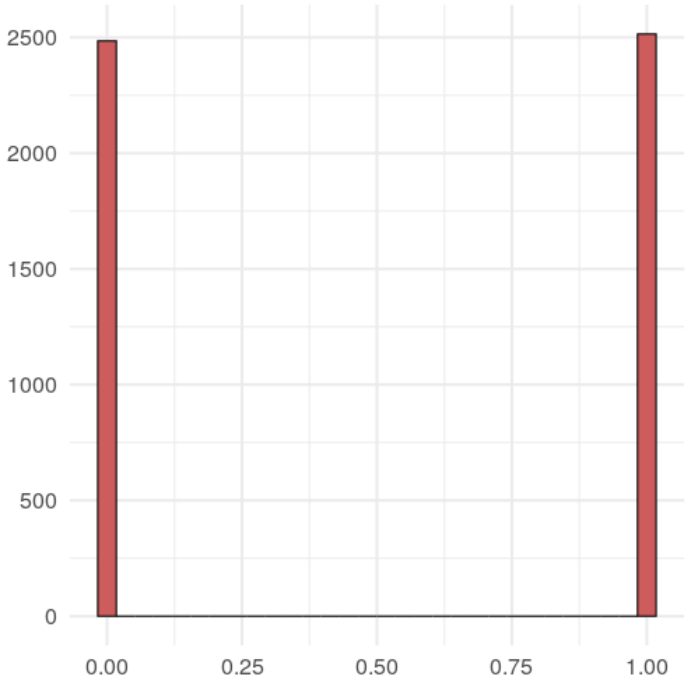
y ~ Normal(promedio, desv. est. **)**

y ~ Normal(170, 2)

Distribucion normal (Gaussiana)

- **Valores posibles:** Números reales
- **Parámetros:** promedio, desviación estándar
- Común "en la naturaleza" y epistémicamente liviana

Distribucion Bernoulli (Binomial)



y ~ Bernoulli(p **)**

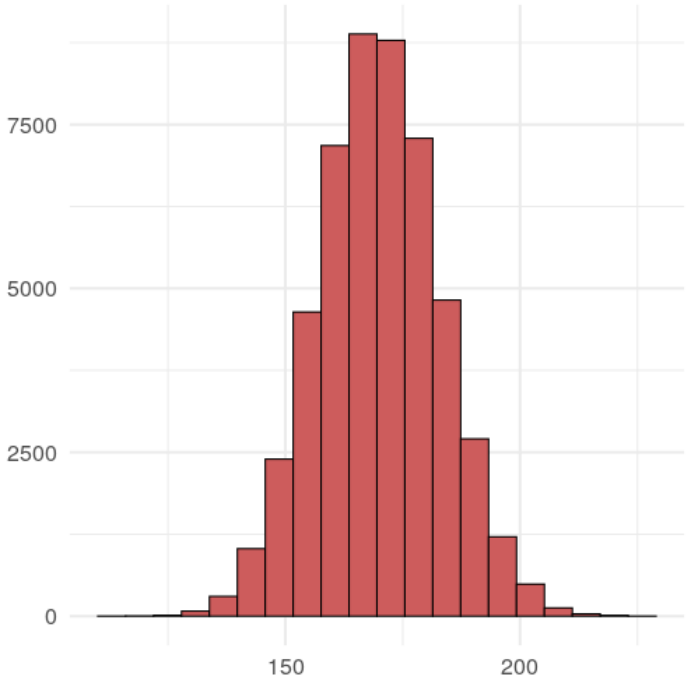
y ~ Bernoulli(prob. de éxito **)**

y ~ Bernoulli(0.5)

Distribucion Bernoulli (Binomial)

- **Valores posibles:** 0 o 1
- **Parámetros:** probabilidad de éxito (1)
- Común experimentos y ciencias sociales

Distribucion Poisson



y ~ Poisson(λ)

y ~ Poisson(ritmo **)**

y ~ Poisson(170)

Distribucion Poisson

- **Valores posibles:** números naturales + 0
- **Parámetros:** ritmo (expectativa de promedio)
- Común en ciencias sociales, cuando contamos eventos
- Su varianza es igual a su ritmo/promedio $\Rightarrow$
  su desviación estándar es $\sqrt{\lambda}$

Próxima sesión

- Entrega de "Assignment 3" (08:00 AM 03/05)

- **Introducción a la regresión**