# 1 Introducción

Métodos empíricos 2

05/04/2022

# Hoy

- Organización del curso
- Diseño de análisis y estadística inferencial
- Replicabilidad y reproducibilidad
- R
- Estadística descriptiva

# Organización

# Información básica

- 5 ECTS (~125-150 horas)
- Material y comunicación: https://aulaglobal.upf.edu/
- Equipo
  - Thomas Brochhagen (thomas.brochhagen@upf.edu)
  - Guillermo Montaña (guillermo.montana@upf.edu)
- Estructura
  - Martes: Conceptos y discusión
  - Jueves: Práctica y más discusión

# Observaciones sobre lengua

1. Instrucción
2. Ejercicios
3. Comunicación personal

# Evaluación

- Ejercicios semanales (20%, no recuperable)
- Ejercicio práctico (20%, recuperable)
- Revisión por pares (20%, no recuperable)
- Informe de diseño de análisis y su ejecución (40%, recuperable)

# Evaluación

- Ejercicios semanales (20%, no recuperable)
  - A través de Aula Global
  - Corrección automática
  - Máximo 2 intentos
  - 5-6 en total

# Evaluación

- Revisión por pares (20%, no recuperable)
  - Solamente opción de aprobar, o no
  - No se aprueba si no se entregan los dos documentos; o si cualquiera de ellos no se adecúa a un estándar mínimo de calidad

# Evaluación

- Ejercicio práctico (20%, recuperable)
  - A través de Aula Global
  - A mitad de curso
  - Corrección automática
  - Un solo intento

# Evaluación

Diseño y ejecución de análisis (40%)

- Individual o en grupos (división de trabajo debe ser aprobada)
- Evaluación en función a tema / tamaño del grupo
- Máximo dos páginas

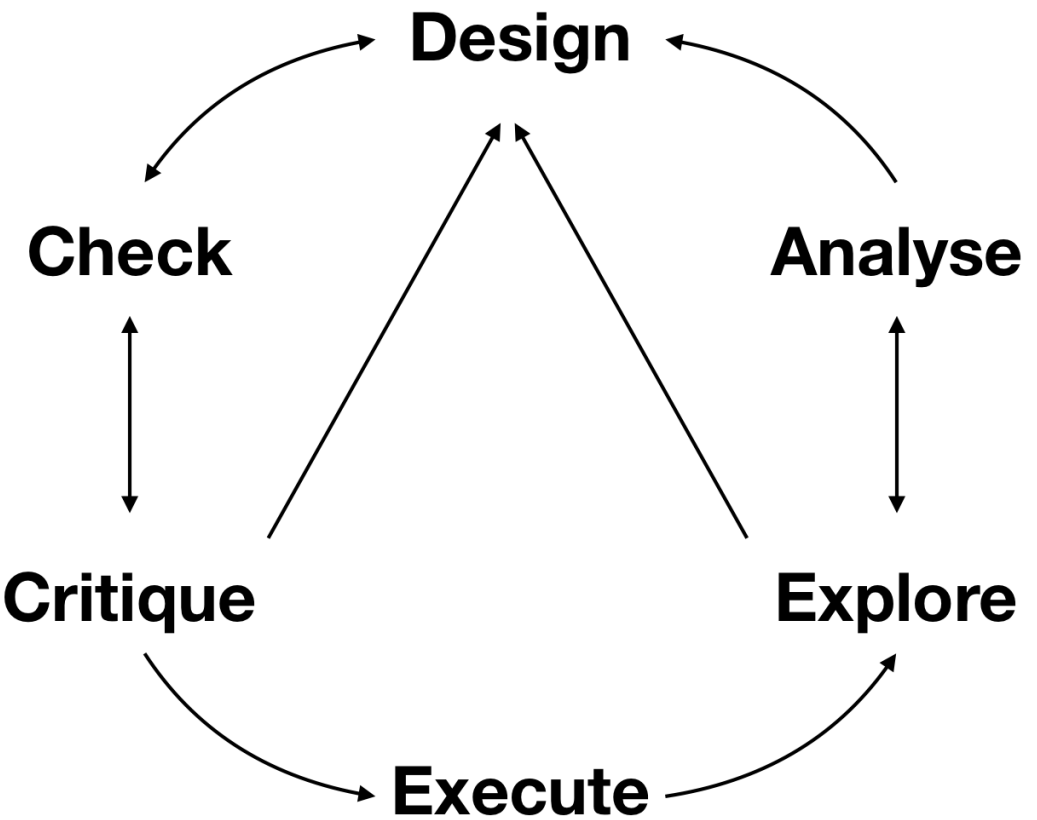
---

**Criterios** (max. 100 puntos)

- Claridad (30): descripción clara y al punto; uso adecuado de visualización
- Replicabilidad (15)
- Reproducibilidad (15)
- Contenido (40): el análisis está bien motivado y ejecutado con métodos adecuados
- Creatividad (20): supera expectativas en una o múltiples dimensiones, como visualización, temática, metodología, documentación

# Contenido del curso y motivación general

# Ciclo de análisis



# Sesiones

- Diseño (1-3)
- Control (3)
- Crítica (3-9)
- Ejecución (3)
- Exploración (3-9)
- Análisis (4-9)

1. Introducción
2. Diseño de análisis
3. Recolección de datos y muestras
4. Introducción a la regresión
5. Regresión multivariada
6. Regresión generalizada I
7. Regresión generalizada II
8. Corpus I
9. Corpus II
10. Revisión y temas avanzados

# Algunas motivaciones

- Manipulación e interpretación de datos
  - Crítica
  - Limpieza
  - Modelado
  - Visualización
  - (Programación)
- Indispensable para ciencias del lenguaje
- Cada vez más indispensable en el siglo 21

# A efectos prácticos

- Manipulación e interpretación de datos
- Comparación de grupos
- Predicción
- Visualización
- (Programación)

… aplicado a ciencias del lenguaje

# Análisis inferencial

# Análisis inferencial (vs. descriptiva)

- Inferencia de propiedades (más allá de la muestra)
- Predicción
- Comparación
- Causa-efecto

| x | y | x | y | x | y | x | y |
|---|---|---|---|---|---|---|---|
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.50 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |
| **Mean** | 9.00 | 7.50 | 9.00 | 7.50 | 9.00 | 7.50 | 9.00 | 7.50 |
| **Stdev** | 3.32 | 2.03 | 3.32 | 2.03 | 3.32 | 2.03 | 3.32 | 2.03 |
| **R** | 0.82 | | **R** | 0.82 | **R** | 0.82 | **R** | 0.82 |

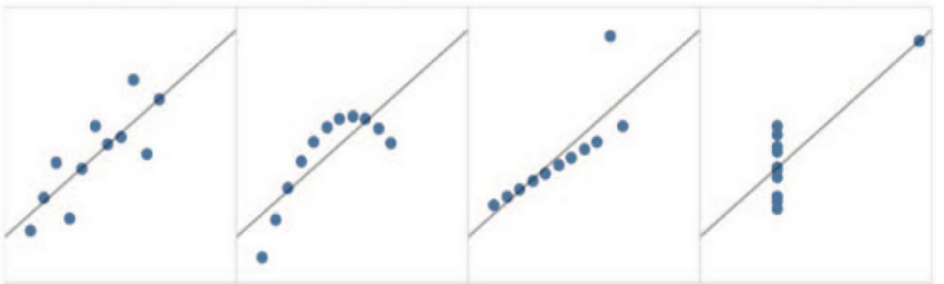


Fig. 1 de Franconeri et al. 2021 The Science of Visual Data Communication: What Works

# Análisis inferencial: Centrado en modelos

All models are wrong, but some are useful

-George E.P. Box

# Análisis inferencial: Centrado en datos

We should only let data speak for themselves when they have learned to clean themselves

-Erik van Zwet

# Regresión

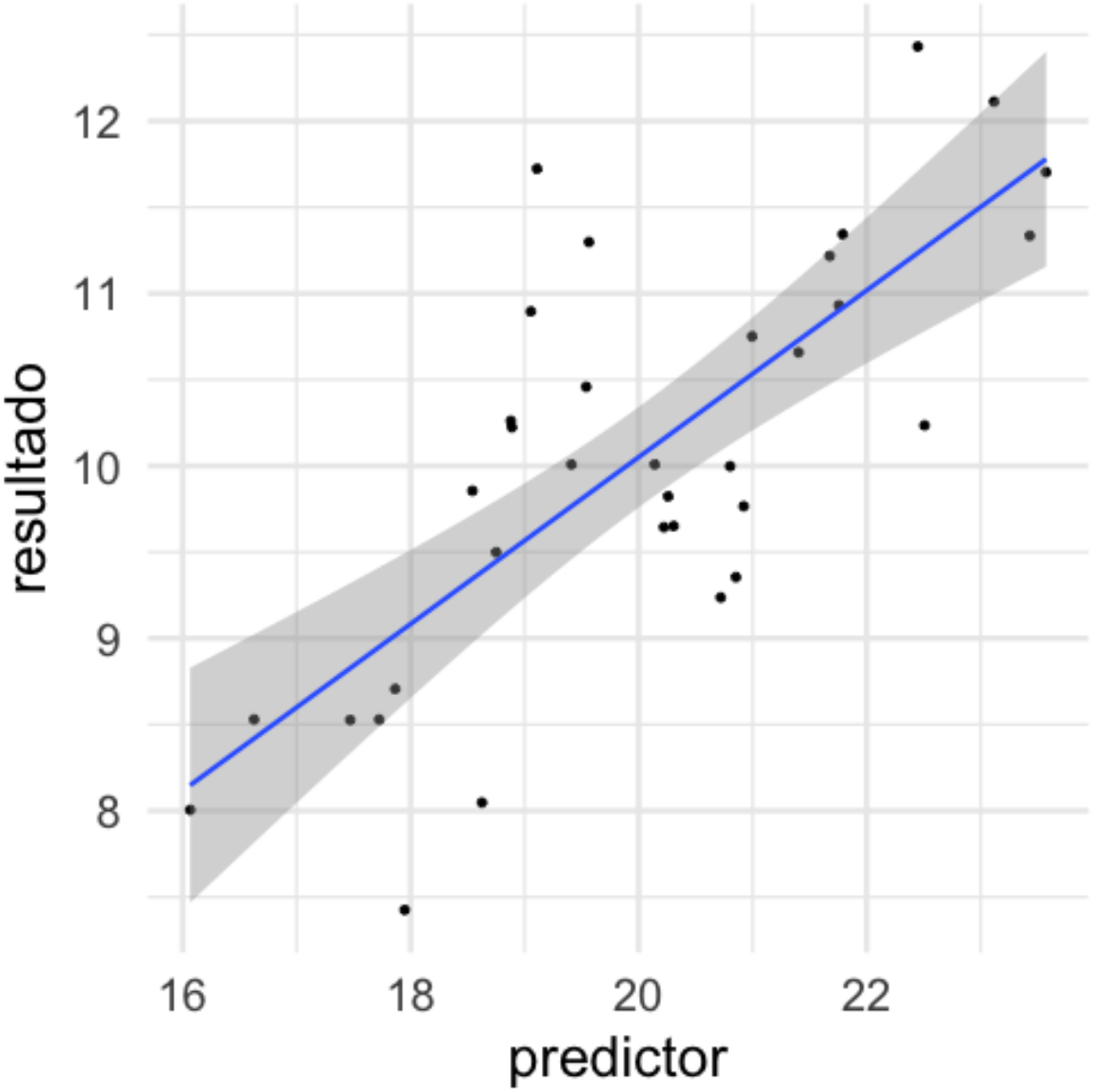
Estimación del efecto de uno o más *predictores* en un *resultado*

- Fluidez en una segunda lengua a base de la primera
- Probabilidad del uso de un pronombre a base de su entorno lingüístico
- Tamaño de una palabra en función a su frecuencia
- …

---

Estimación de diferencias entre grupos

- Comparación de un método de aprendizaje de la lengua contra otro
- Comparación de políticas lingüísticas
- …


resultado
12
11
10
9
8
16
18
20
22
predictor

# Replicabilidad y reproducibilidad

# Crísis de replicabilidad y reproducibilidad

Replication is one of the central issues in any empirical science.

To confirm results or hypotheses by a repetition procedure is at the basis of any scientific conception. A replication experiment to demonstrate that the same findings can be obtained in any other place by any other researcher is conceived as an operationalization of objectivity.

It is the proof that the experiment reflects knowledge that can be separated from the specific circumstances (such as time, place, or persons) under which it was gained.

-Stefan Schmidt

## Replicabilidad

Que se puedan obtener resultados consistentes con los mismo datos de entrada; pasos computacionales; métodos; código; y condiciones de análisis

---

## Reproducibilidad

Que se puedan obtener resultados consistentes en diferentes análisis que buscan responder la misma pregunta, cada cual con sus propios datos

# Crisis de replicabilidad y reproducibilidad

- Una cantidad importante de resultados no se han podido ni replicar ni reproducir (50%-80%, dependiendo del campo)
- Conllevó a un gran cambio en metodología y documentación (estamos en ello)
- Independientemente, es central documentar todas tus decisiones y manipulaciones; de principio a fin
- Lenguas de programación son una herramienta ideal para
  - conducir análisis empírico, y
  - asegurarse de que sea replicable

# R

```r
library(ggplot2) #librería gráfica
set.seed(123)    #semilla aleatorea

x   <-   rnorm(n = 34, mean = 20, sd = 2) #valores predictor
err <-   rnorm(n = 34, mean = 0,sd=1)     #valores variacion
y   <-   x/2 + err                        #valores resultado

df  <-   data.frame(resultado = y,        #formato conjunto
                    predictor = x)

ggplot(df, aes(x = predictor, y = resultado)) + #gráfico
  geom_point() +                                #scatter plot
  geom_smooth(method='lm') +                    #regresión
  theme_minimal(base_size=25)                   #apariencia
```

# En este curso

- Uso rudimentario e interactivo
- Principalmente: intuición sobre estructuras de datos y su manipulación
- Manera de asegurar replicabilidad
- Manera de mejorar comunicación

# Estadística descriptiva

# Estadística descriptiva

- Tendencias centrales (promedio; mediana; moda)
- Dispersión (varianza; desviación típica)
- Variación conjunta (correlación, co-varianza)

---

Si falta una revisión: Franke (2021) *An Introduction to Data Analysis*, capítulo 5

# Próxima sesión

- Entrega de "Assignment 1" (08:00 AM 19/04)
- Si hay, preguntas generales sobre el curso
- Si hay, preguntas generales sobre el informe
- Si hay, preguntas generales sobre R

---

- **Diseño de análisis**