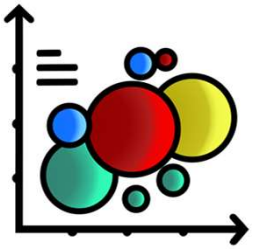
# Simulaciones para la enseñanza de la estadística: desafíos y nuevas oportunidades

Adriana Pérez y Gerardo Cueto
Grupo de Bioestadística Aplicada
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires



VI Jornadas Argentinas de Educación Estadística

# Simulaciones

- Constituyen una poderosa herramienta para propiciar el aprendizaje activo de conceptos abstractos y anti-intuitivos
- Permiten a los estudiantes utilizar problemas auténticos y crear un entorno de aprendizaje para practicar y facilitar la adquisición de habilidades
- Permiten evaluar la adecuación de los modelos

# Objetivos del taller

1. Generar casos auténticos para el aprendizaje de técnicas estadísticas
2. Facilitar la comprensión de conceptos estadísticos complejos mediante simulaciones
3. Validar modelos mediante simulaciones



# 1. Simulaciones para generar casos auténticos

Generar un problema para una materia de bioestadística. ej: alumnos de Biología
Bajo un contexto real obtenido de un paper.

Definir:

- Contexto biológico.
- Las variables y sus parámetros
- Simular la base de datos
- Escribir el texto del problema

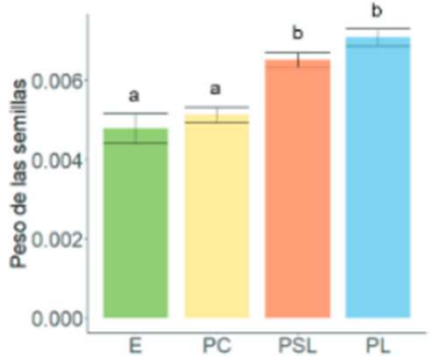
Quiero hacer un ejercicio para una materia de bioestadística para biólogos.
Quiero simular los datos del ejercicio.
Quiero obtener las variables, los parámetros para la simulación y el contexto biológico del ejercicio de una tabla de un paper.
¿Qué cosas tengo que darte para que generes el texto del ejercicio y simules los datos?

El contexto del ejercicio sale de este resumen, que es de un paper de la revista Ecología Austral:

"La polinización entomófila incrementa la calidad y la cantidad de frutos y semillas de la mayoría de los principales cultivos del mundo. San Juan es una de las principales provincias argentinas productoras de semilla hortícola; sin embargo, no se cuenta con información que permita asegurar un servicio ecosistémico de polinización óptimo a los cultivos. En este estudio se propuso determinar la influencia de la polinización entomófila en la formación de frutos y en la cantidad y calidad de las semillas de cultivos destinados a producción de semillas de achicoria, cebolla, rabanito, repollo y zapallo. Además, se evaluó si los cultivos exhibían déficit polínico en los lotes estudiados. Para ello, en cada uno de los lotes se evaluó el servicio de polinización de los cultivos a través de un experimento comparativo en el que se contrastó la formación de frutos, el número de semillas formadas por fruto, el peso y el poder germinativo de las semillas mediante diferentes tratamientos de polinización. Se encontró que la polinización entomófila actúa de manera diferencial sobre los parámetros medidos y que tiene un efecto positivo sobre la polinización de los cultivos. Además, los resultados indicaron que la formación de frutos y semillas no estuvo afectada por déficit polínico. El conocimiento alcanzado contribuye a implementar prácticas de manejo orientadas a conservar la entomofauna polinizadora, a mejorar el servicio de polinización de los cultivos y a promover la economía de los agricultores locales."

Quiero que el ejercicio se base en un Anova de un factor y sus contrastes que genere resultados como el de la figura adjunta.

¿Podrías generar el texto del ejercicio y la base de datos?




Resultados presentados en promedio ± EE. Referencias: E=Exclusión. PC=Polinización suplementada cruzada. PSL=Polinización suplementada libre. PL=Polinización libre. Medias con una letra común no son significativamente diferentes (P>0.05).

sí, quiero que generes la tabla con los datos, pero la variable la quiero en miligramos

gracias! ahora quiero agregar a la base de datos otras variables que están correlacionadas con el peso de semillas. quiero que las variables y los parámetros de las simulaciones salgan de esta tabla

**Table 1**. Mean value and standard deviation (SD) of fruit set, number of seeds, weight and germination power of seeds obtained per treatment (see references in the text), cultivated species and agroecosystem.

| Especie cultivada | | Fruit set | | N° semillas | | Peso | | Poder germinativo | |
|---|---|---|---|---|---|---|---|---|---|
| | Trat. | Media | DE | Media | DE | Media | DE | Media | DE |
| *Cichorium intybus* L. (sitio 1) | E | 0.7583 | 0.3765 | 10.1481 | 5.2675 | 0.0040 | 0.0013 | 0.3313 | 0.2809 |
| | PC | 0.8158 | 0.2986 | 15.0323 | 4.3473 | 0.0054 | 0.0011 | 0.4188 | 0.2183 |
| | PL | 0.8710 | 0.0727 | 13.3300 | 2.1930 | 0.0070 | 0.0013 | 0.6500 | 0.1242 |
| | PSL | 0.7333 | 0.3689 | 14.8276 | 3.1061 | 0.0063 | 0.0013 | 0.4500 | 0.0612 |

Gracias! Quiero que me ayudes en la generación del código en R para generar la simulación de la base

gracias.
La variable Num_semillas es de conteos, podrías simularla con un modelo Poisson?
La variable Poder_germinativo es una proporción, podrías simularla con un modelo Binomial?

# Link al script generado a partir de la interacción con chat GPT

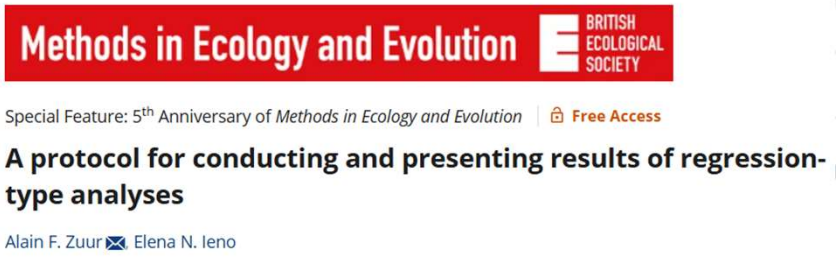
https://drive.google.com/file/d/12qxSEZrNEYnc4hpgSU6xwnXWhbcWc9K1/view?usp=sharing

# 2. Simulaciones para facilitar la comprensión de conceptos estadísticos complejos

https://drive.google.com/file/d/17JXiMu7c30jQTVfqgUxdsAtDAQpDfVtH/view?usp=sharing

# 3. Simulaciones para evaluar la adecuación de un modelo

# Protocolo para realizar y presentar resultados de análisis tipo regresión




- Plantear preguntas adecuadas
- Visualizar el diseño experimental
- Realizar la exploración de datos
- Identificar la estructura de dependencia en los datos
- Presentar el modelo estadístico
- Ajustar el modelo
- Evaluar supuestos del modelo
- Interpretar y presentar los resultados numéricos del modelo
- Crear una representación visual del modelo
- Simular a partir del modelo

# Caso de estudio

En el Monte Austral el sobrepastoreo y la extracción de hidrocarburos son los disturbios antrópicos que mayor impacto ecológico producen.

Se llevó a cabo un estudio con el objetivo de determinar el impacto de la actividad ganadera sobre la biomasa vegetal en estancias ubicadas en el Monte Austral, en las provincias de Neuquén y Río Negro

Las estancias ganaderas del área fueron clasificadas según la carga ganadera en pastoreo bajo, medio y alto.

De cada grupo se seleccionaron al azar 5 estancias. En cada una de ellas se trazaron al azar 3 transectas de 100 m lineales cada una, y se midió la biomasa vegetal total (en gramos).

Los datos se encuentran en el archivo Monte_Austral3.csv

# Modelo

- La variable respuesta es la biomasa vegetal total (medida en gramos). Se trata de una variable cuantitativa continua.
- Las variables explicativas son:
  - VE1: carga ganadera, tipo cualitativa con tres niveles (bajo, medio y alto), de efectos fijos.
  - VE2: estancia ganadera, tipo cualitativa y de efectos aleatorios. Anidada en la variable explicatoria “carga ganadera”.

$$Yijk=\mu+\alpha i+Bj+\varepsilon ijk \quad i=1 \text{ a } 3,\ j=1 \text{ a } 5,\ k=1 \text{ a } 3$$

$$\varepsilon ijk \backsim N(0,\sigma^2)$$

$$Bj \backsim N(0,\sigma^2 estancias)$$

$$\varepsilon ij,\ Bj \text{ independientes}$$

```
modelo1 <- lmer(biomasa ~ pastoreo + (1|estancia), data = datos)
```

# Simulaciones para evaluar la adecuación de un modelo

`simulate()` genera datos bajo el modelo ajustado (usa estimaciones del modelo)

```
datos_simulados <- simulate(modelo, nsim = 100, seed= 123)
```

```
Y_simulado <- b0 + b1*X + rnorm(n, mean = 0, sd = sigma)
```

Se pueden compara los datos simulados con los observados mediante gráficos y métricas.

https://drive.google.com/file/d/17dyM4zwBkrL5qAgTe1tnLC3vGkuPYGEE/view?usp=sharing