

# Introducción a la estadística

Bases indispensables y uso de 

Olivier Devineau

Fundación Charles Darwin

Taller interno, 27–30 abril 2010

# Introducción y conceptos importantes

# Cosas importantes

- Teoría estadística: 8:30–10:00, 10:30–12:00
- Práctica con *R*: 13:30–15:00, 15:30–17:00
- Café: 10:00–10:30 y 15h00–15h30
- Por favor, apagan los celulares

¡Preguntas bienvenidas en cualquier momento!

# Agradecimientos

Use material amablemente provisto por:

- Claude-Pierre Guillaume, EPHE, Montpellier, Francia
- Damien Caillaud, UT, Austin, Texas, USA
- Julien Dutheil, CNRS, Montpellier, Francia
- Vladimir Grosbois, CIRAD, Montpellier, Francia

Correcciones, comentarios y sugerencias por

- Eliana Bontti, FCD

## Agradecimientos

Se uso también:

- Crawley, M.J. 2005. *Statistics, an introduction using R*. John Wiley & Sons.
- Quinn, G.P., and Keough, M.J. 2002. *Experimental design and data analysis for biologists*. Cambridge University Press.

## ¿Qué es la estadística?

Definición

- Principios y métodos para recoger, clasificar, resumir y analizar datos
- Aprender, hacer conclusiones y tomar decisiones

## La verdadera estadística . . .

Evolución de salarios y empleados en una empresa

		Obreros	Ejecutivos	Promedio
Salario	2004	200	2000	1100
	2006	180	1800	990
Empleados	2004	1000	100	550
	2006	600	500	550

Periódico Salarios bajaron en un 10%

Empresa Salario promedio por empleado aumentó de \$363.6 a \$916.3

Periódico Hubo despidos en la empresa

Empresa Igual número de empleados y reclutamiento

## La estadística . . .

### Puede

- Proveer criterios objetivos para probar hipótesis
- Optimizar esfuerzos
- Evaluar razonamiento de manera crítica

### NO puede

- Decir la verdad
- Compensar ausencia de controles o mala planificación
- Indicar importancia que no es probabilística

## Primer paso para entender datos: ¡describirlos!

- Distribución normal, poisson, binomial ...
- Media, mediana
- Varianza, desviación estándar y error estándar

⇒ Estadística *descriptiva* informa sobre forma, centro y amplitud de los datos

## Describir no es suficiente

- No es suficiente averiguar que hay variación
- ¿Variación científicamente interesante o variación natural?

Estadística inferencial permite:

- Distinguir entre señal y ruido
- Deducir información y llegar a conclusiones

## Lo más difícil es empezar

- ¿Qué tipo de análisis?
- Depende de los datos y de la pregunta inicial
- ¿Cómo saber que hacer? ¡habiéndolo hecho miles de veces!

## ¿Estadística paramétrica o no?

### Paramétrica

- Intervalos regulares
- Hipótesis de distribución *normal*
- Media y error/desviación estándar

### No paramétrica

- Cualquier tipo de escala
- No hipótesis de distribución (independencia)
- Mediana y desviación mediana

## ¿Qué preguntarse para empezar?

- ¿Cuál es la variable dependiente?
- ¿De qué tipo es? ¿Medida continua, número, proporción, categoría?
- ¿Cuáles son las variables independientes?
- ¿Son continuas? ¿Categorías? ¿Ambos?

## ¿Qué análisis? Guía de decisión

### 1) Variables independientes

- |                                 |           |
|---------------------------------|-----------|
| • Todas continuas               | Regresión |
| • Todas categóricas             | Anova     |
| • Ambas continuas y categóricas | Ancova    |

## ¿Qué análisis? Guía de decisión

### 2) Variable dependiente

- |                          |                                 |
|--------------------------|---------------------------------|
| • Continua               | Regresión normal, Anova, Ancova |
| • Proporción             | Regresión logística             |
| • Número                 | Regresión log-lineal            |
| • Binaria                | Análisis logístico binario      |
| • Tiempo hasta la muerte | Análisis de sobrevivencia       |

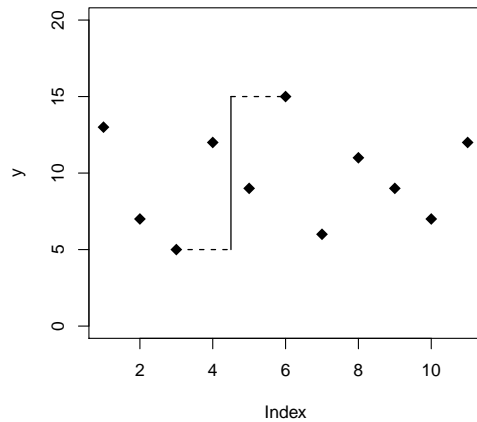
## Por qué la estadística?

¡Porque Todo varia!

Mucha variabilidad temporal, espacial y entre individuos:

- Genética
- Factores ambientales
- Azar
- Errores de observación y medida

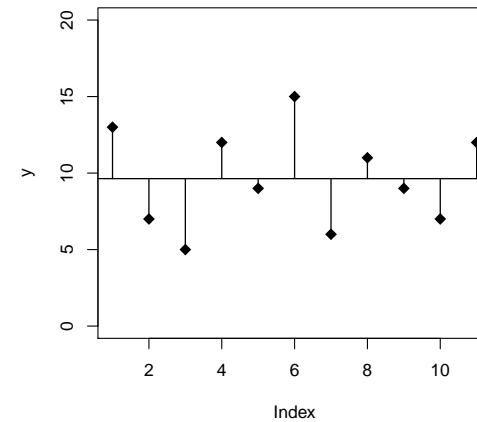
## ¿Como medir la variabilidad?



- Rango: [5, 15]
- Media y desviaciones de la media
- Residuales
- $\sum(y - \bar{y}) = 0$
- $SS = \sum(y - \bar{y})^2$
- Suma de los cuadrados (sum of squares)

17 / 1

## ¿Como medir la variabilidad?



- Rango: [5, 15]
- Media y desviaciones de la media
- Residuales
- $\sum(y - \bar{y}) = 0$
- $SS = \sum(y - \bar{y})^2$
- Suma de los cuadrados (sum of squares)

18 / 1

## Una mejor medida de la variabilidad

- $SS = \sum(y - \bar{y})^2$ ,  $n = 11$
- ¿Que pasa con  $SS$  si se agrega un punto?
- $SS$  aumenta por cada nuevo punto
- $MS = \frac{\sum(y - \bar{y})^2}{n}$
- Desviación cuadrática media (Mean square deviation  $MS$ )

19 / 1

## Grados de libertad

- Muestra de 5 números:  $\bar{y} = 4$ ,  $\sum y = 20$

2	7	4	0	7
---	---	---	---	---

- Total libertad en la selección de números 1 – 4  
 $\Rightarrow$  4 grados de libertad (degrees of freedom  $d.f.$ )
- $df = n - p$
- $n$  = número de muestras,  $p$  = número de parámetros estimados por el modelo

20 / 1

## Varianza (1)

Medida de la variabilidad

- $MS = \frac{\sum(y - \bar{y})^2}{n}$
- No se puede calcular  $MS$  antes de conocer  $\bar{y}$
- ¿De donde se obtiene  $\bar{y}$ ?
- $\bar{y}$  es un parámetro estimado de los datos
- Se pierde un grado de libertad

## Varianza (2)

Formalización y definición

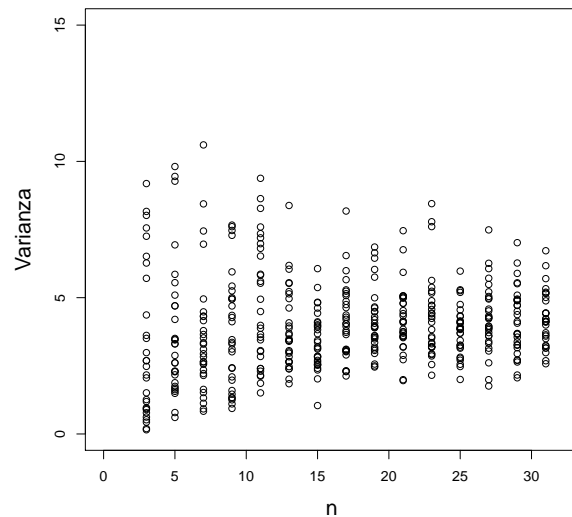
- Medida cuantitativa de la variabilidad:

$$\text{Varianza} = \frac{\text{Suma de cuadrados}}{\text{Grados de libertad}} = \frac{SS}{df}$$

$$s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

## Varianza y tamaño de muestra

Media: 10, Varianza: 4



## Una medida de fiabilidad

¡Error estándar de la media!

- ¿Fiabilidad de estimaciones cuando  $s^2 \nearrow$  ?
- Fiabilidad  $\propto s^2$
- ¿Y qué tal del tamaño de la muestra?
- Fiabilidad  $\propto \frac{s^2}{n}$
- Qué son las unidades?
- $SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$

## Intervalos de confianza

- Muestreo repetido → rango de valores
- Intervalo de confianza  $\propto$  Fiabilidad
- Distribución  $t$  de Student
- Nivel de confianza  $\alpha$  y grados de libertad  $df$
- Número de errores estándar que se espera
- $CI_{95\%} = \bar{y} \pm t_{\alpha, df} \sqrt{\frac{s^2}{n}}$

## Diseño experimental

Conceptos claves

Replicación: aumenta fiabilidad

Aleatorización: reduce sesgo

- Si replican y randomizan correctamente, ¡no hay problema!
- Diseño inadecuado  $\nrightarrow$  buenos resultados

## Replicación

- Permite aumentar la fiabilidad y cuantificar la variabilidad dentro de un tratamiento
- Medidas repetidas deben:
  - Ser independientes (individuos distintos)
  - No formar una serie temporal
  - No estar agrupadas juntas en un lugar
  - Tener escala espacial adecuada

## Replicación (2)

- Idealmente: una réplica de cada tratamiento debe estar agrupada en un bloque y cada tratamiento debe estar repetido en varios bloques

## ¿Cuántas réplicas?

- Tantas como sea posible 😊
- ¿Cómo saber? Estudios pilotos y experiencia  
⇒ Indicación sobre varianza base y magnitud de la respuesta al tratamiento
- Método práctico (en general):  $\geq 30$

## Poder y réplicas

- Poder: probabilidad de rechazar  $H_0$  cuando es falsa
- ¿Cuántas réplicas para detectar un efecto  $\delta$  con 80% probabilidad de no cometer un error?
- Experiencia y/o estudio piloto  
⇒ Primera estimación del efecto  $\delta$  y de la varianza  $s^2$

$$n \approx \frac{8 * s^2}{\delta^2}$$

## Seudoreplicación

Condición importante: independencia de los errores

- Medidas repetidas del mismo individuo → seudoreplicación temporal
- Varias medidas del mismo lugar → seudoreplicación espacial
- ¿Cuántos grados de libertad?

## ¿Qué hacer con seudoreplicación?

- Promediar seudoreplicación y hacer análisis sobre medias
- Hacer análisis separados por cada período de tiempo
- Usar análisis de series de tiempo o modelos de efectos mixtos

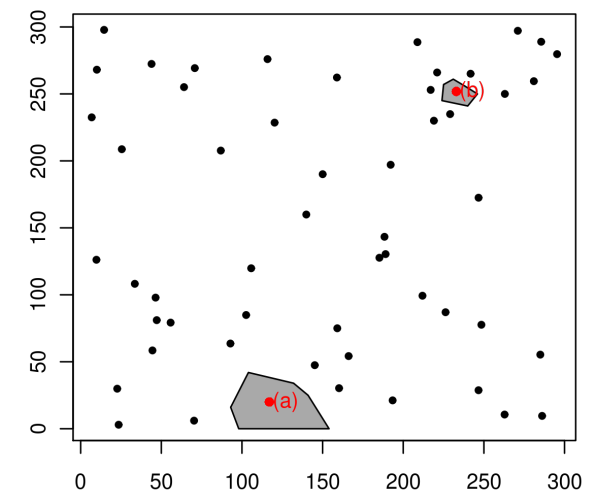


## Aleatorización

- ¿Cómo seleccionar un árbol al azar en una selva?
  - ¿Hojas accesibles?
  - ¿Cerca del laboratorio?
  - ¿Parece sano?
  - ¿Sin insectos?
- ⇒ ¡Sesgo en la fotosíntesis!

33 / 1

## Selección aleatoria de un árbol



34 / 1

## Controles

- No controles, no conclusiones

35 / 1

## ¿Cuánto tiempo?

- Idealmente: determinar duración por adelantado
- NO seguir experimento hasta que se obtenga un “buen” resultado

36 / 1

## Inferencia fuerte

- Formular una hipótesis clara
- Diseñar un test aceptable
- Sin replicación, aleatorización y controles, no hay progreso

## Modelaje estadístico

- Datos: lo que pasó
- Descripción → patrones → mecanismos
- Modelo para explicar y predecir
- Varios (muchos) modelos están ajustados a los datos
- → Modelo mínimo y adecuado

## Modelaje estadístico

Mínimo: Suficientemente simple

Adecuado: ¿Por qué usar modelo que no describe los datos?

Mejor modelo: La menor proporción de varianza que no sea explicada (desviación residual mínima)

## La navaja de Occam

Principio de parsimonia

- Con varias explicaciones igualmente válidas
- Correcta: la más simple

En estadística significa que:

- Tan pocos parámetros como sea posible
- Modelos lineales > no lineales
- Pocas condiciones > muchas
- Pocas variables > muchas
- 1 explicación simple > varias explicaciones complicadas

## La navaja de Einstein

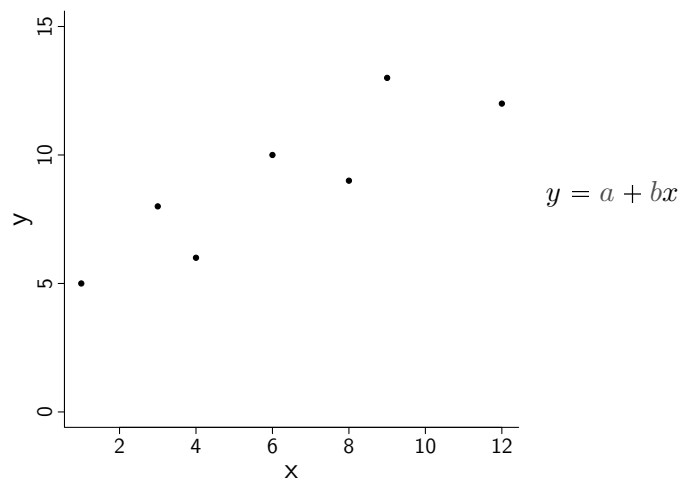
Einstein: “Un modelo debe ser tan simple como posible.  
Pero no más simple”

## Máximo de verosimilitud (Maximum Likelihood: ML)

- Dado los datos
  - Y dado un modelo
  - ¿Qué valores de parámetros hacen a los datos observados más probables?
- ⇒ Estimadores sin sesgo que minimizan la varianza

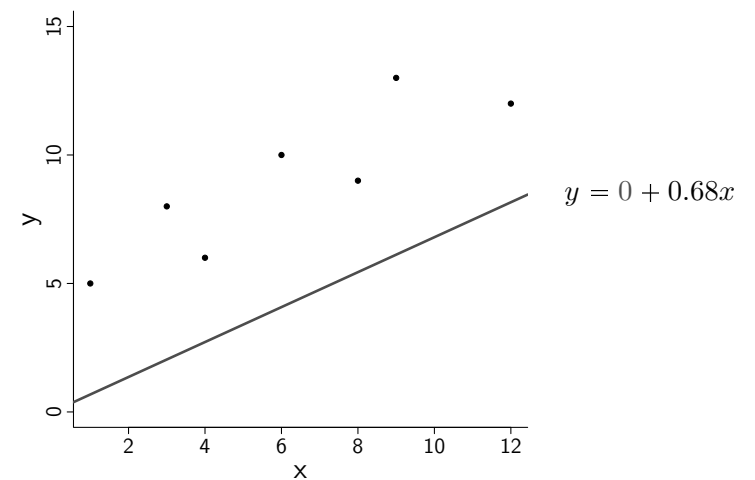
## Máximo de verosimilitud

Ejemplo: regresión  $y = a + bx$



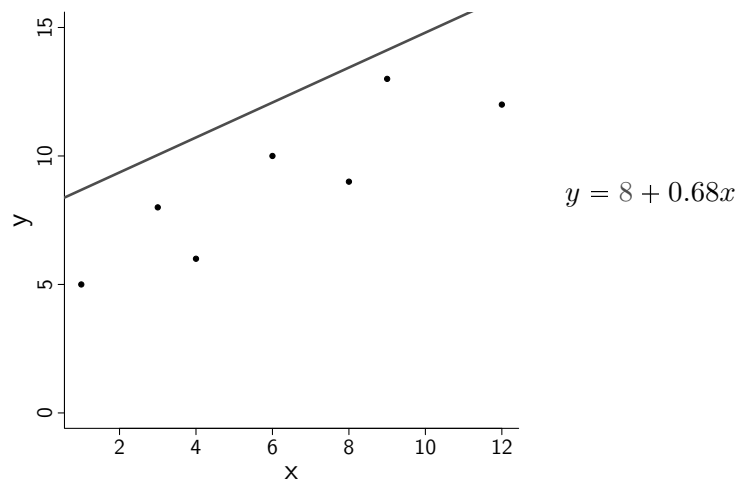
## Máximo de verosimilitud

Ejemplo: regresión  $y = a + bx$



## Máximo de verosimilitud

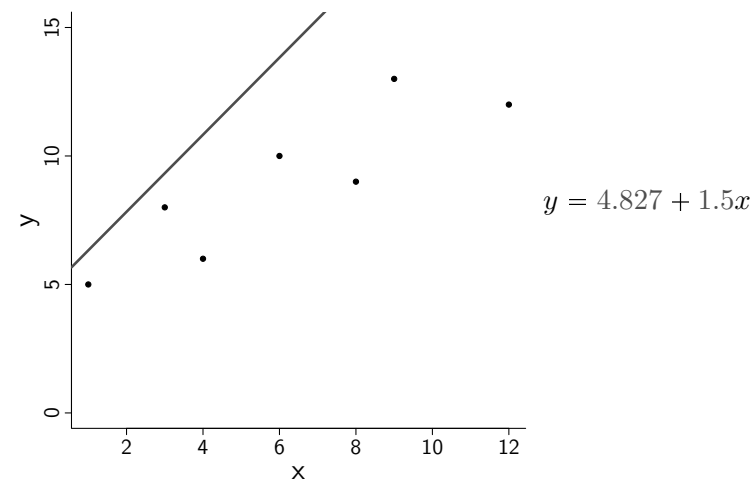
Ejemplo: regresión  $y = a + bx$



45 / 1

## Máximo de verosimilitud

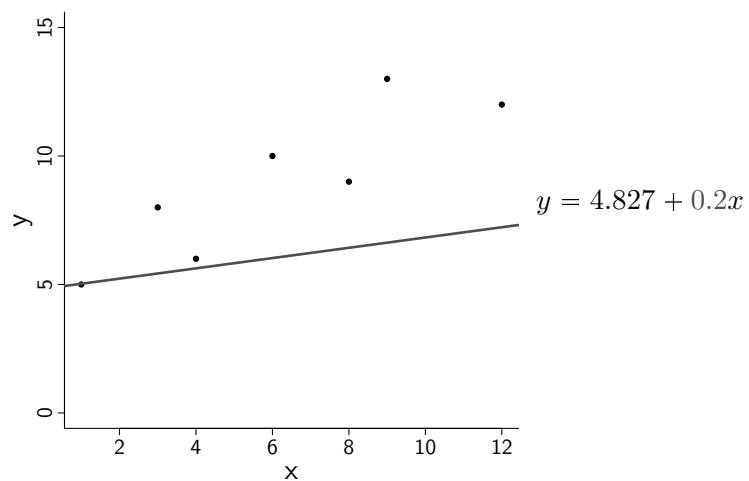
Ejemplo: regresión  $y = a + bx$



46 / 1

## Máximo de verosimilitud

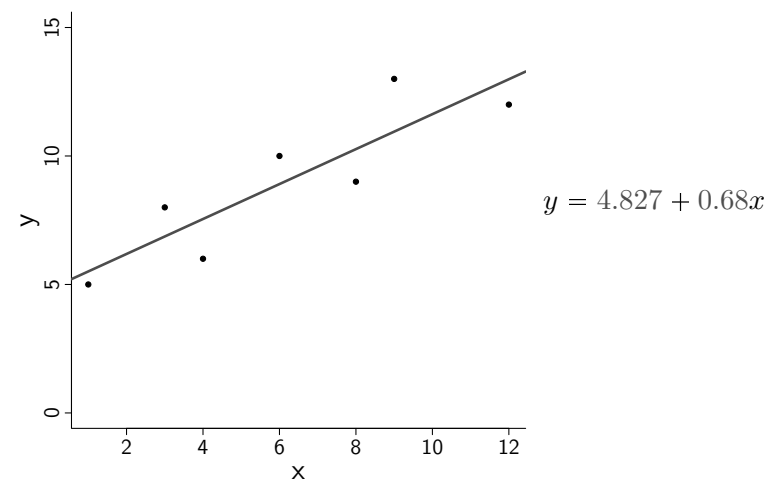
Ejemplo: regresión  $y = a + bx$



47 / 1

## Máximo de verosimilitud

Ejemplo: regresión  $y = a + bx$



48 / 1

## Noción de test estadístico

## Distribución de probabilidad

- Representación de las probabilidades asociadas con los estados posibles de una variable aleatoria

Ejemplo:  $X$  = número de hijos en una familia de 2 niños

- $2\varnothing, (1\sigma, 1\varnothing), (1\varnothing, 1\sigma), 2\sigma$
  - $p(X = 0 \sigma) = 1/4$
  - $p(X = 1 \sigma) = 1/4 + 1/4$
  - $p(X = 2 \sigma) = 1/4$
- }  $\sum p(X) = 1$

## Distribución binomial

Definición

- Serie de  $n$  intentos independientes
- Cada intento  $\rightarrow$  Éxito / Fracaso
- Probabilidad de éxito:  $p$

- Distribución discontinua
- $X \sim \mathcal{B}(n, p)$
- $P(r) = \binom{n}{r} p^r (1-p)^{n-r}$

## Distribución Binomial (2)

- 39% de los habitantes tienen ojos azules
- $X \sim \mathcal{B}(3, 0.39)$



## Distribución binomial

¿Cuándo se aplica?

- Porcentaje de mortalidad
- Tasa de infección
- Proporción: sexos, respuesta a un tratamiento, intenciones de voto . . .

Se necesita saber cuantos individuos hay en categoría *éxito* y cuantos hay en categoría *fracaso*

## Distribución de Poisson

Definición

- Cuantas veces un evento raro ocurre por unidad de tiempo/espacio

- Distribución discontinua

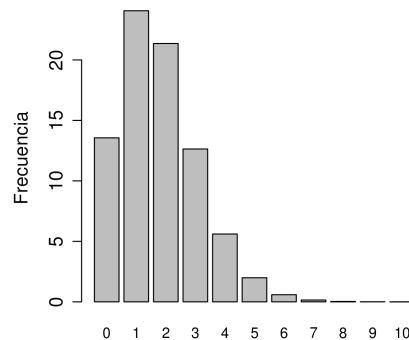
- $X \sim \mathcal{P}(\lambda)$

- $P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$

## Distribución de Poisson

¿Cuándo se aplica?

- Plantas en una parcela
- Semillas comidas por una ave por minuto
- Bebés naciendo por hora en un hospital
- Errores en un texto
- Degradación de sustancia radioactiva



## Distribución normal

Definición

- Teorema del límite central
- Suficientes muestras  $\rightarrow$  medias  $\rightarrow$  distribución normal

- Distribución continua

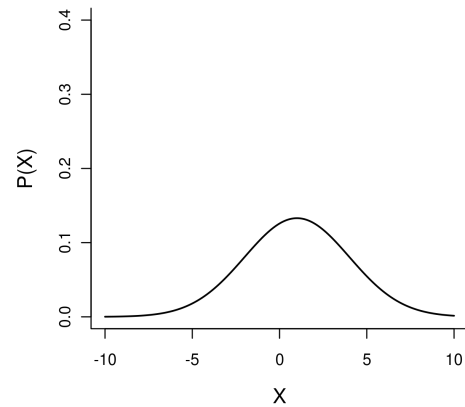
- $X \sim \mathcal{N}(\mu, \sigma)$

- $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$

## Distribución normal

¿Cuándo se aplica?

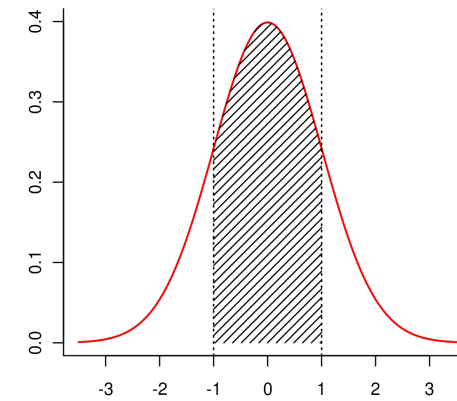
- ¡Todo el tiempo!
- Regresión lineal, análisis de varianza ...



57 / 1

## Distribución Normal Estándar

$X \sim \mathcal{N}(0, 1)$

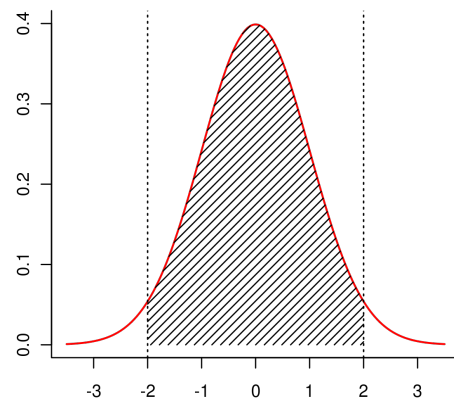


- $\pm 1 \sigma \sim 68\%$
- $\pm 2 \sigma \sim 95\%$
- $\pm 3 \sigma \sim 99\%$

58 / 1

## Distribución Normal Estándar

$X \sim \mathcal{N}(0, 1)$

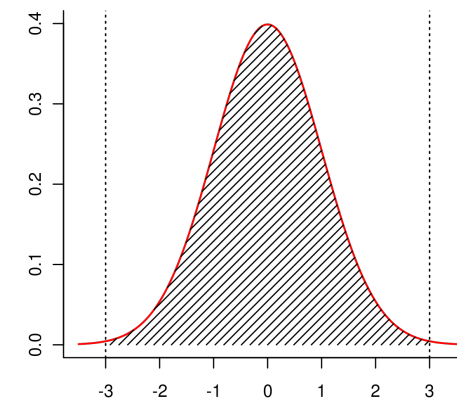


- $\pm 1 \sigma \sim 68\%$
- $\pm 2 \sigma \sim 95\%$
- $\pm 3 \sigma \sim 99\%$

59 / 1

## Distribución Normal Estándar

$X \sim \mathcal{N}(0, 1)$



- $\pm 1 \sigma \sim 68\%$
- $\pm 2 \sigma \sim 95\%$
- $\pm 3 \sigma \sim 99\%$

60 / 1

## Otras distribuciones de variables

- Lognormal (largo, peso ...)
- Exponencial (Tiempo de fracaso)
- Gamma
- Distribución de Weibull
- Beta

## Distribuciones de estadísticos

- Distribución  $z$
- Distribución  $t$  de Student
- Distribución del  $\chi^2$
- Distribución  $F$  de Fischer

## ¿Qué es un test estadístico?

Herramienta para tomar decisión

- Calcular un estadístico  $T_{obs}$  de una muestra
- Comparar  $T_{obs}$  con la distribución de  $T_{teo}$  cuando la hipótesis es verdadera
- La posición de  $T_{obs}$  informa sobre la probabilidad de que la hipótesis sea verdadera

## Test estadístico: procedimiento

- 1 Pregunta biológica: ¿Hay cóndores en el parque?
- 2 Pregunta estadística: Hipótesis  $H_0$
- 3 Elección del test estadístico: ¿Cuál usar?
- 4 Criterios de decisión: ¿Qué riesgo de error? ¿Qué nivel de confianza?



## Test estadístico: procedimiento

- ⑤ ¡Colección de los datos!
- ⑥ Cálculo de el estadístico del test
- ⑦ Decisión estadística: ¿Se puede rechazar  $H_0$  o no?
- ⑧ Inferencia y explicación biológica

## Buenas y malas hipótesis

- Una buena hipótesis se puede rechazar/falsear
- ① Hay cóndores en el parque
- ② No hay cóndores en el parque
- ¡Ausencia de prueba no es prueba de ausencia!

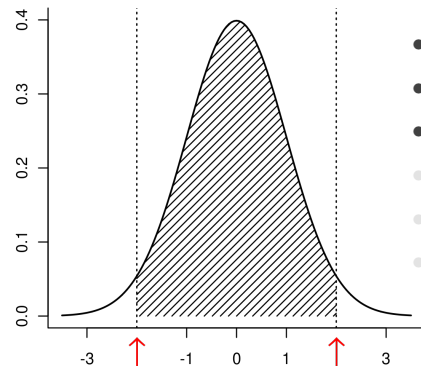
## Hipótesis nula

- “Nada está pasando”
  - “Las medias de dos muestras son las mismas”
  - “La pendiente de la relación es cero”
- ⇒ La hipótesis nula se puede falsear. Rechazar cuando los datos muestran que es suficientemente improbable

## Elección del test

- Tipo de variables: cualitativas, cuantitativas ...
- Número y tamaño de las muestras
- Condiciones de cada test

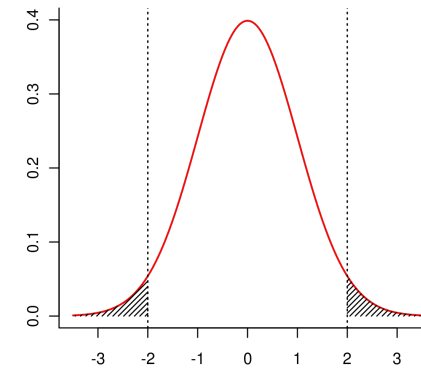
## Criterios de decisión (1)



- $\pm 2 \sigma \sim 95\%$
- Valores umbrales
- Región de aceptación
- 5% menos probable
- Región de rechazo
- Riesgo  $\alpha$

69 / 1

## Criterios de decisión (1)



- $\pm 2 \sigma \sim 95\%$
- Valores umbrales
- Región de aceptación
- 5% menos probable
- Región de rechazo
- Riesgo  $\alpha$

70 / 1

## Criterios de decisión (2)

- 2 errores posibles :  
 Tipo I : Rechazar  $H_0$  cuando es verdadera  
 Tipo II : Aceptar  $H_0$  cuando es falsa

Hipótesis nula	Situación real	
	Verdadera	Falsa
Acepta	Decisión correcta Poder $1 - \beta$	Tipo II Riesgo $\beta$
Rechaza	Tipo I Riesgo $\alpha$	Decisión correcta

71 / 1

## Hay que comprometer ...

Poder: Probabilidad de rechazar  $H_0$  cuando es falsa

- Error I: rechazar  $H_0$  cuando es verdadera  $\alpha$
- Error II: aceptar  $H_0$  cuando es falsa  $\beta$
- Poder:  $1 - \beta$
- $\alpha$  y  $\beta$  relacionados
- Cuando  $\alpha \searrow \beta \nearrow$

72 / 1

## ¿Cuándo $\alpha$ debe ser alto?

### Ejemplo: Efectos secundarios de una droga

- Test final antes de comercializar
- Grupo A: droga | Grupo B: placebo
- $H_0$ : no hay diferencia entre grupos A y B
- $H_1$ : A tiene mayor frecuencia de anomalías que B

73 / 1

## ¿Cuándo $\alpha$ debe ser alto?

Aceptar riesgo  $\alpha$  más alto para reducir riesgo  $\beta$

### $\alpha$ alto: error de tipo I

- $H_0$  rechazada pero verdadera
- No se comercializa
- Más estudios para determinar efecto real

### $\beta$ alto: error de tipo II

- $H_0$  “aceptada” pero falsa
- Comercialización
- ¡Mucha gente sufre de los efectos secundarios!

74 / 1

## Colección de los datos

### ¡Acuérdense!

- Aleatorización
- Replicación

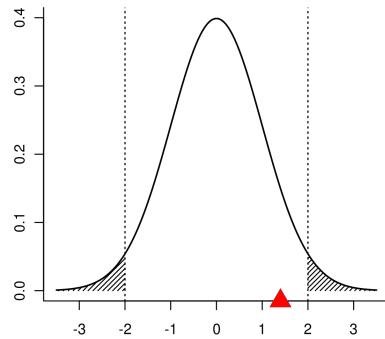
75 / 1

## Computación del estadístico del test

### Ejemplo: Prevalencia de la malaria

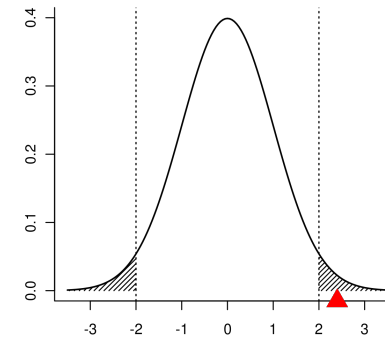
- “La prevalencia es la misma en A y en B”
- $H_0 : \mu_A = \mu_B$
- El estadístico del test representa la diferencia de prevalencia:  $T = f(\text{prev}_A - \text{prev}_B)$
- Distribución de  $T$  corresponde a  $H_0$  verdadera

76 / 1

Comparación de  $T$  con la  
distribución teórica

- $T_{obs}$  no está en la región de rechazo
- No se puede rechazar  $H_0$
- No es posible afirmar que hay una diferencia de prevalencia entre A y B

77 / 1

Comparación de  $T$  con la  
distribución teórica

- $T_{obs}$  está en la región de rechazo
- Se puede rechazar  $H_0$
- Se concluye que la prevalencia de la malaria es diferente entre A y B
- El riesgo de que esta conclusión sea falsa es  $\alpha = 5\%$

78 / 1

Valor  $P$ 

- Medida de la credibilidad de la hipótesis nula

## Ejemplo

- $H_0 : \mu_A = \mu_B$
- $p < 0.05 \Rightarrow$  improbable que  $H_0$  sea verdadera:  $\mu_A \neq \mu_B$
- $p = 0.23 \Rightarrow$  No hay suficiente evidencia para rechazar  $H_0$

79 / 1

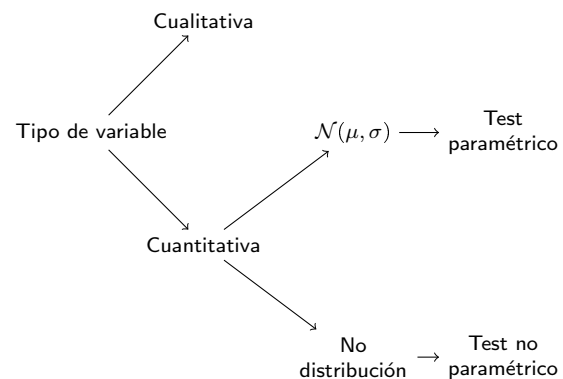
## Significancia

- ¿Qué significa "Resultado significativo"?
- Diccionario: Que tiene sentido
- Estadística: Improbable que haya ocurrido por azar si la hipótesis nula es verdadera
- Improbable: Ocurre menos de 5% de las veces

80 / 1

## ¿Como elegir el test adecuado?

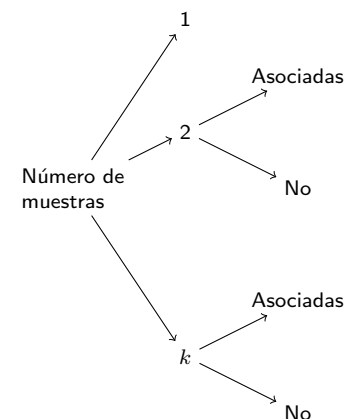
Algunas directrices (1)



81 / 1

## ¿Como elegir el test adecuado?

Algunas directrices (2)



82 / 1

## Dependencia – Asociación

Tests asociados

- Muestras asociadas: vienen del mismo grupo
- Relacionadas por correlación o por regresión
- Conexión espacial
- Conexión temporal

⇒ Usar tests específicos: e.g., “paired t-test”

83 / 1

## Comparar una muestra con una distribución teórica

⇒ Test de conformidad

- Test  $t$  de conformidad
- Test de Wilcoxon
- Test binomial
- Test  $\chi^2$  de conformidad
- ...

84 / 1

## Comparar dos muestras

⇒ Test de comparación (de homogeneidad)

- Test  $t$  (posiblemente “asociado”)
- Test de Mann-Whitney
- Test de Fisher
- Test  $\chi^2$
- ...

85 / 1

Comparar *más* de dos muestras

⇒ Test de comparación (continuación)

- Anova / Manova
- Test de Kruskal-Wallis
- Test de Friedman
- Test  $\chi^2$
- ...

86 / 1

Evaluar el grado de asociación  
entre variables

Muestras independientes

⇒ Correlación y regresión

- Correlación de Pearson / de Spearman ( $n = 2$ )
- Regresión simple / regresión logística ( $n=2$ )
- Regresión no paramétrica
- Regresión múltiple / regresión logística múltiple ( $n > 2$ )
- ...

87 / 1

Comparar un grupo con una  
distribución teórica

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial

## Comparar 2 grupos no asociados

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$

## Comparar 2 grupos asociados

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar

## Comparar $\geq 3$ grupos no asociados

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar
Anova simple	Test de Kruskal-Wallis	Test $\chi^2$

## Comparar $\geq 3$ grupos asociados

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar
Anova simple	Test de Kruskal-Wallis	Test $\chi^2$
Anova con medidas repetidas	Test de Friedman	Test $Q$ de Cochran

## Cuantificar asociación entre 2 variables

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar
Anova simple	Test de Kruskal-Wallis	Test $\chi^2$
Anova con medidas repetidas	Test de Friedman	Test $Q$ de Cochran
Correlación de Pearson	Correlación de Spearman	Coefficientes de contingencia

## Predecir valor desde 1 variable

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar
Anova simple	Test de Kruskal-Wallis	Test $\chi^2$
Anova con medidas repetidas	Test de Friedman	Test $Q$ de Cochran
Correlación de Pearson	Correlación de Spearman	Coefficientes de contingencia
Regresión (no)lineal simple	Regresión no paramétrica	Regresión logística simple

## Predecir valor desde varias variables

Medidas $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$	Categoría, grado, sin distribución	Binomial
Test $t$ 1 muestra	Test de Wilcoxon	Test $\chi^2$ , test binomial
Test $t$ no asociado	Test de Mann-Whitney	Test de Fisher, test $\chi^2$
Test $t$ asociado	Test de Wilcoxon	Test de McNemar
Anova simple	Test de Kruskal-Wallis	Test $\chi^2$
Anova con medidas repetidas	Test de Friedman	Test $Q$ de Cochran
Correlación de Pearson	Correlación de Spearman	Coefficientes de contingencia
Regresión (no)lineal simple	Regresión no paramétrica	Regresión logística simple
Regresión (no)lineal múltiple	_____	Regresión logística múltiple

## Más recursos para elegir un test

- *Handbook of Biological Statistics:*  
<http://udel.edu/~mcdonald/statbigchart.html>
- *Statistics Online Computational Resources:*  
[www.socr.ucla.edu/Applets.dir/ChoiceOfTest.html](http://www.socr.ucla.edu/Applets.dir/ChoiceOfTest.html)
- *GraphPad / Intuitive Biostatistics:*  
[www.graphpad.com/www/Book/Choose.htm](http://www.graphpad.com/www/Book/Choose.htm)
- *Social Research Methods:*  
[www.socialresearchmethods.net/selstat/ssstart.htm](http://www.socialresearchmethods.net/selstat/ssstart.htm)
- *James D. Leeper, University of Alabama:*  
<http://bama.ua.edu/~jleeper/627/choosestat.html>
- *S. Holttum, B. Blizard, Canterbury Christ Church University:*  
[www.whichtest.info/index.html](http://www.whichtest.info/index.html)



## Correlación y regresión

## Dos categorías de tests estadísticos

Tests de comparación : 1 variable,  $\geq 2$  poblaciones

Tests de relación :  $\geq 2$  variables, 1 población

## $\geq 2$ variables es común en biología

### 2 variables para el mismo individuo

- Presión sanguínea  $X_1$ , peso  $X_2$
- Abundancia de una especie de planta  $X_1$ , nivel del pH en el suelo  $X_2$ , temperatura  $X_3$

- Datos bivariados o multivariados

⇒ ¿Cuál es la relación entre las variables?

## Relación entre $\geq 2$ variables

La estadística correlacional

### Varios tipos de relación

- No conexión
- Relación *handout* :  $1 > 0$  /  $< 0$ , causal / no
- Conexión funcional  $\rightarrow$  predicción

### Objetivo de la estadística correlacional

- Determinar validez y fuerza de la relación entre las variables
- Determinar la dirección de la relación

## Estadística correlacional

Correlación: ¿Cómo 2 variables varían juntas?

Regresión: Relación entre 1 variable dependiente y  
 $\geq 1$  variable independiente

Análisis multivariados: Relación entre  $\geq 2$  variables  
independientes / dependientes / ambos

## Noción de correlación

### Ejemplo

- 1 población: 2 variables continuas
- Presión sanguínea  $X_1$ , peso  $X_2$
- Cada muestra  $i$  : 1 valor por cada variable:  $x_{i1}$  y  $x_{i2}$
- ¿La presión sanguínea y el peso son correlativas?

## Noción de correlación (2)

Definición

Correlación se define en terminos de:

- Varianza de  $X_1$ :  $var(X_1)$
- Varianza de  $X_2$ :  $var(X_2)$
- ¿Como  $X_1$  y  $X_2$  varían juntas? Covarianza:  $cov(X_1, X_2)$

⇒ Coeficiente de correlación

$$r = \frac{cov(X_1, X_2)}{\sqrt{var(X_1) \cdot var(X_2)}}$$

## El coeficiente de correlación $r$

Correlación de Pearson (paramétrica)

- No unidad
- $r \in [-1, 1]$
- Magnitud: fuerza de la relación
- Signo: dirección de la relación
- Muestra:  $r$ , Población:  $\rho$

## ¿Qué test para chequear la correlación?

$X_1$ : Presión sanguínea y  $X_2$ : peso

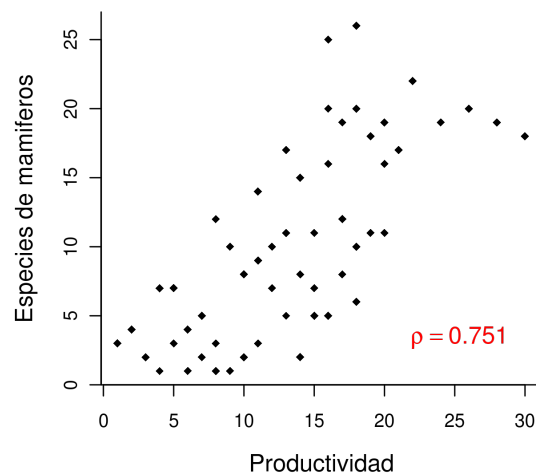
- ¿Hipótesis nula?
- No hay una relación lineal entre la presión sanguínea y el peso
- $H_0 : \rho = 0$
- Cuando  $H_0$  es verdadera,  $r \sim \mathcal{N}(\mu, \sigma)$   
⇒ uso de test  $t$  de Student

## Correlación no paramétrica

- ¿Qué hacer cuando los requisitos no se cumplen?
- ⇒ Coeficiente de correlación de rango
- de Spearman:  $\rho$
  - de Kendall:  $\tau$
- ¡Más conservadores!

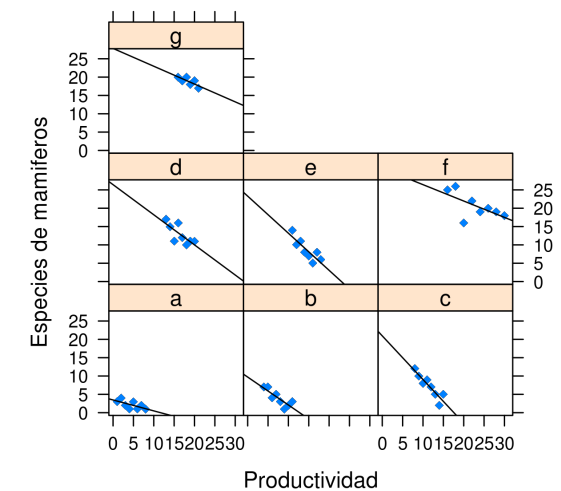
## La correlación depende de la escala

¡Las cosas no son siempre como parecen!



## La correlación depende de la escala

¡Las cosas no son siempre como parecen!



## Modelo lineal: concepto general

- Se puede identificar:
  - 1 variable respuesta / dependiente  $Y$
  - $\geq 1$  variable explicativa / predictiva / independiente / covariable  $X_1, X_2, \dots$
- Cada unidad de muestra:  $y_i, x_{1i}, x_{2i} \dots$
- Explicar el patrón de  $Y$  con  $X$

109 / 1

## Modelo lineal

Forma general de los modelos estadísticos

- *Variable dependiente = modelo + error*
- Modelo: covariables y parámetros
- Covariables: continuas / categóricas / ambos
- Error: parte de la variable dependiente que no está explicada por el modelo
- Se supone una distribución para el componente del error, y de ahí para la variable dependiente  $Y$

110 / 1

## ¿Qué significa lineal?

- Relación de línea recta entre 2 variables
- Combinación lineal de parámetros
- No exponente, no multiplicación por otro parámetro
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

111 / 1

## Análisis de regresión lineal

Contexto

- Usar datos de una muestra para estimar valores de parámetros y sus errores estándar
- ¿Cuándo se usa?
- Variables explicativa y dependiente son continuas
- Altura, peso, volumen, temperatura ...
- Nube de puntos  $\rightarrow$  regresión lineal

112 / 1

## Análisis de regresión lineal

### Objetivos

- Describir la relación lineal entre  $Y$  y  $X$
- Determinar cuánto de la variación en  $Y$  se explica por la relación lineal con  $X$  y cuánto de esta variación no se puede explicar
- Predecir nuevos valores de  $Y$  a partir de valores de  $X$

113 / 1

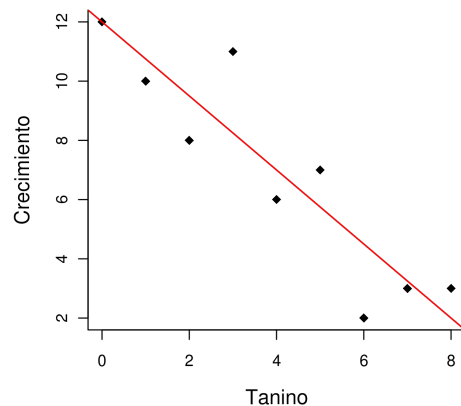
## Análisis de regresión lineal

### Varios tipos de regresión

- Regresión lineal: lo más simple y frecuente
- Regresión polinomial: chequear si una relación es no lineal
- Regresión no lineal
- Regresión no paramétrica: si no hay forma funcional

114 / 1

## Principio de la regresión lineal



- Datos
- Modelo:  $y = a + bx$
- ¿Cambio en  $y$ ?  
 $\delta y = -10$
- ¿Cambio en  $x$ ?  
 $\delta x = +8$
- Pendiente  
 $b = \delta y / \delta x = -1.25$
- ¿Ordenada al origen?  
 $a = 12$
- $y = 12 - 1.25x$

115 / 1

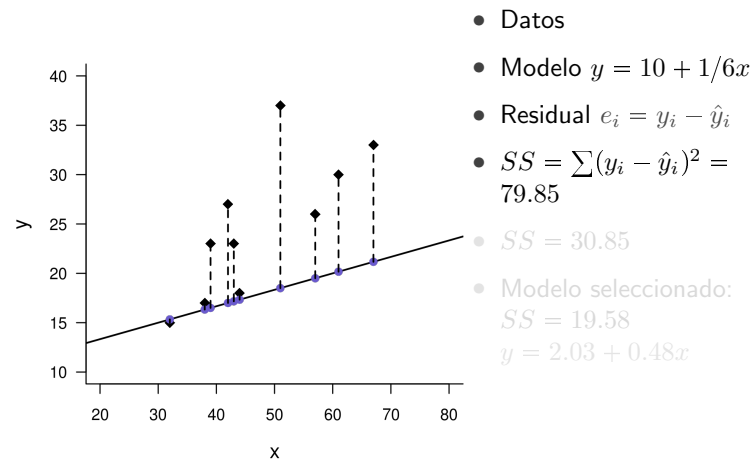
## Principio de la regresión lineal (2)

- Ajustar un modelo a los datos
- Estimar los parámetros del modelo
- Probar varios valores de parámetros hasta encontrar el mejor modelo
- Máxima verosimilitud (Maximum Likelihood ML)
- Mínimos cuadrados (Ordinary Least Square OLS)

116 / 1

## Cuadrados mínimos: principio

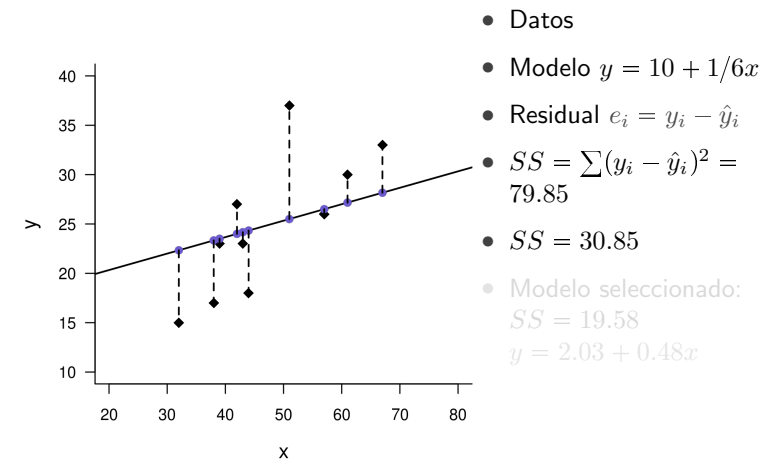
OLS: Ordinary Least Squares



117 / 1

## Cuadrados mínimos: principio

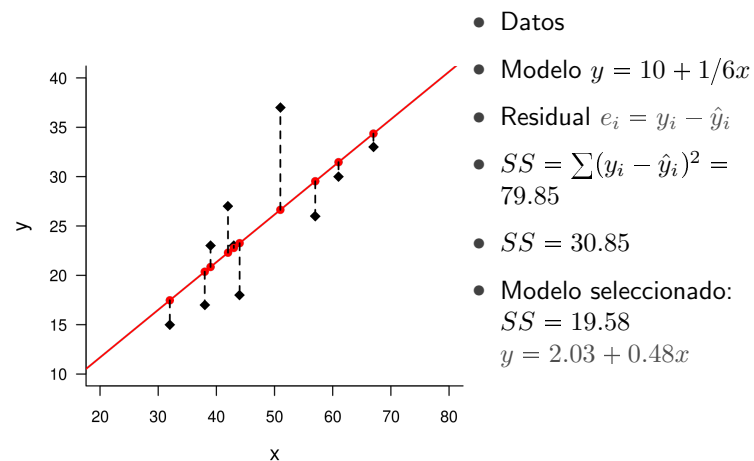
OLS: Ordinary Least Squares



118 / 1

## Cuadrados mínimos: principio

OLS: Ordinary Least Squares



119 / 1

## Hipótesis nula en regresión

- ¿Cuál sería  $H_0$ ?
- No hay una relación lineal entre las variables
- Pendiente  $b = 0$ 
  - Test de Fisher:  $F$
  - Test de Student:  $t$

120 / 1

## Varianza explicada

$r^2$ : coeficiente de determinación

- Variación de  $Y$  explicada por la relación con  $X$
- (coeficiente de correlación)<sup>2</sup>
- $r^2 \in [0, 1]$
- ¿Como se mejora el ajuste del modelo con pendiente comparado a un modelo sin pendiente?
- $r^2$  inadecuado para comparar modelos con números de parámetros diferentes

## Comparar varios modelos

- Evaluar varias hipótesis → varios modelos
- $H_0$ : modelo simple,  $H_1$ : modelo más complejo
- Hay que comparar los modelos

## Comparar modelos de regresión

### Minimos cuadrados (OLS)

- Ajuste: proporción de varianza explicada
- No-ajuste: proporción de varianza residual

⇒ Análisis de varianza

### Máxima verosimilitud (ML)

- Ajuste: tamaño de la verosimilitud

⇒ Prueba de la razón de verosimilitud (Likelihood Ratio Test o AIC)

## Comparar modelos de regresión (2)

Siempre la misma lógica

- Medir el ajuste de cada modelo
- Comparar los ajustes de diferente modelos para examinar hipótesis sobre los parámetros

### Ejemplo: presión sanguínea y peso

- Modelo 1:  $P = \beta_0 + \varepsilon$
- Modelo 2:  $P = \beta_0 + \beta_1 * peso + \varepsilon$
- Comparar  $M_1$  y  $M_2$  es equivalente a evaluar  $H_0 : \beta_1 = 0$

## Condiciones del análisis de regresión (1)

- Involucran de los términos de errores ( $\varepsilon_i$ )
- De la variable dependiente  $Y$
- Importantes para intervalos de confianza
- Importantes para tests de hipótesis con distribución  $t$  o  $F$
- Residuales importantes para chequear condiciones

125 / 1

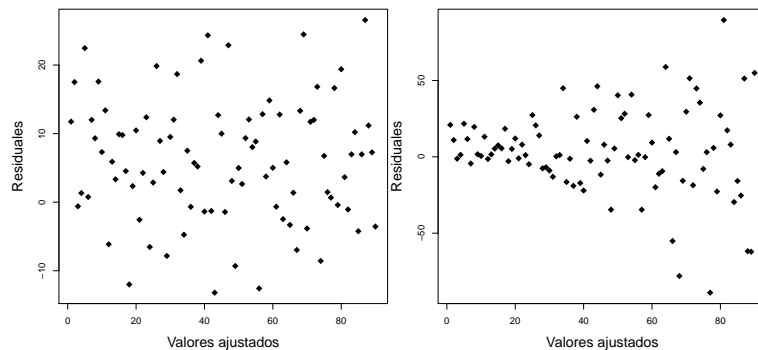
## Condiciones del análisis de regresión (2)

- Normalidad:  $\varepsilon$  tiene una distribución normal
- Homogeneidad de la varianza:  $\varepsilon$  tiene la misma varianza por cada  $x_i$ :  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 = \dots = \sigma_\varepsilon^2$
- Independencia:  $\varepsilon$  son independientes: Los valores de  $Y$  para cualquier  $x_i$  no influyen los valores de  $Y$  para otra  $x_i$

126 / 1

## Homogeneidad de la varianza

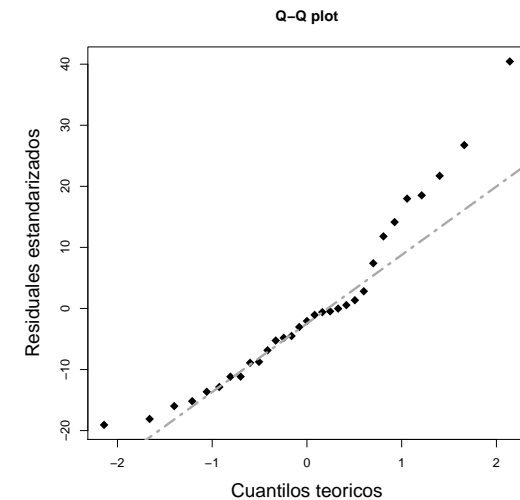
- No tendencia
- Heteroscedasticidad



- Test de Levene, test de Bartlett

127 / 1

## Normalidad de los residuales



- Test de Shapiro-Wilk

128 / 1



## ¿Qué hacer si las condiciones no cumplen?

- Residuales no son independientes:
  - Modelos con efectos aleatorios (random effect models)
- Residuales no son normales:
  - Alternativa no paramétrica
  - Transformación de los datos *log*, *sqrt*, *exp* ...
  - Modelo lineal generalizado (Generalized Linear Model GLM)
- Heterogeneidad de la varianza:
  - GLM

129 / 1

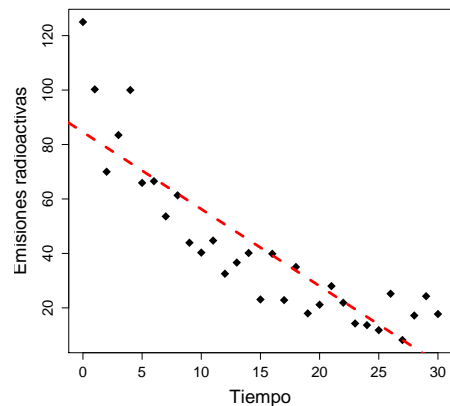
## Si el modelo es inadecuado, se puede...

- Transformar variable dependiente
- Transformar  $\geq 1$  variable explicativa
- Probar otras variables explicativas
- Usar una estructura de error diferente (GLM)
- Usar alternativa no paramétrica (smoothing)
- Usar pesos diferentes por diferentes valores de  $y$

130 / 1

## Regresión polinomial

Ejemplo: Desintegración radioactiva

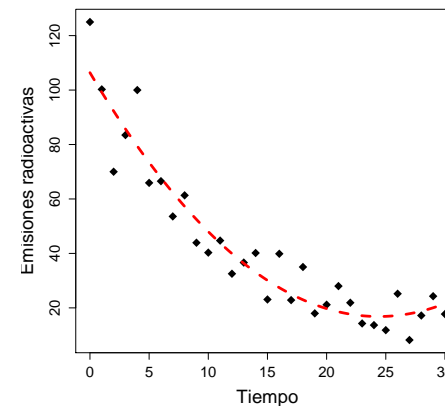


- Regresión lineal:  
 $y = ax + b$
- Regresión polinómica
- $X_2 = X^2$
- $y = ax^2 + bx + c$
- $y = ae^{-bx}$
- ¡Descripción, no explicación!

131 / 1

## Regresión polinomial

Ejemplo: Desintegración radioactiva

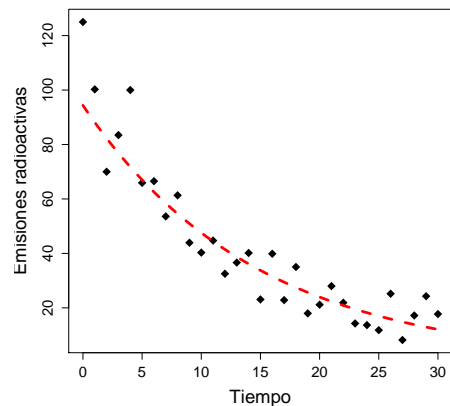


- Regresión lineal:  
 $y = ax + b$
- Regresión polinómica
- $X_2 = X^2$
- $y = ax^2 + bx + c$
- $y = ae^{-bx}$
- ¡Descripción, no explicación!

132 / 1

## Regresión polinomial

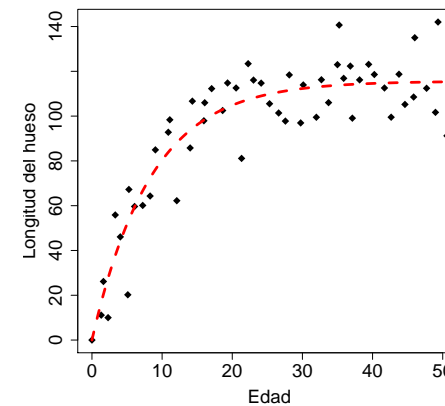
Ejemplo: Desintegración radioactiva




- Regresión lineal:  
 $y = ax + b$
- Regresión polinómica  
 $X_2 = X^2$
- $y = ax^2 + bx + c$
- $y = ae^{-bx}$
- ¡Descripción, no explicación!

133 / 1

## Regresión no lineal y GAM



- : `nls()`
- Teoría:  
 $y = a - be^{-cx}$
- No información:  
Modelos Aditivos  
Generalizados  
(Generalized Additive  
Models GAM)

134 / 1

## Recordatorio de vocabulario

- Normalidad de los errores:
  - Modelos lineales
- Normalidad + var. descriptivas continuas/categorías:
  - Modelos lineales generales
- Errores no normales y/o varianza no homogénea:
  - Modelos lineales generalizados (GLM)

135 / 1

## Modelos lineales generalizados (2)

Varianza no constante / residuales no normales

⇒ Se puede especificar la distribución de los errores

- Proporciones (regresión logística) → Binomial
- Conteos (modelo log-lineal) → Poisson
- Variable dependiente binaria (vivo/muerto) → Binomial
- Tiempo hasta muerte (varianza aumenta) → Exponencial

136 / 1

## (No) enamorarse de su modelo ...

- Todos los modelos son incorrectos
- Algunos modelos son mejores que otros
- El modelo correcto nunca se puede conocer con certeza
- Cuanto mas simple el modelo mejor

## Análisis de varianza

## Comparar $\geq 2$ muestras

Control biológico de las plagas del maíz

### Ejemplo: 5 tratamientos

- Nematodos del suelo
- Avispas parásitas
- Nematodos y avispas
- Bacterias
- Control

## Control biológico (2)

- Muestra aleatoria por cada tratamiento
- Medida del peso de las mazorcas  
⇒ Media:  $\mu_i$ , desviación estándar:  $\sigma_i$
- ¿Cuál tratamiento produce más choclo?
- ¿Como comparar las medias entre tratamientos?

¿Tests  $t$  repetidos?

- ①  $H_0 : \mu_1 = \mu_2$
  - ②  $H_0 : \mu_1 = \mu_3$
  - ③  $H_0 : \mu_1 = \mu_4$
  - ④  $H_0 : \mu_1 = \mu_5$
  - ⑤  $H_0 : \mu_2 = \mu_3$
  - ⑥  $H_0 : \mu_2 = \mu_4$
  - ⑦  $H_0 : \mu_2 = \mu_5$
  - ⑧  $H_0 : \mu_3 = \mu_4$
  - ⑨  $H_0 : \mu_3 = \mu_5$
  - ⑩  $H_0 : \mu_4 = \mu_5$
- Cada hipótesis: riesgo de error de tipo I
  - Con 1 hipótesis:  $\alpha = 0.05$
  - ¿Valor de  $\alpha$  con 2 hipótesis?
  - ¿0.025, 0.05, 0.0725, 0.0975, 0.10?
  - $1 - Pr(\text{no error de tipo I})$
  - $1 - 0.95 \cdot 0.95 = 0.0975$

¿Tests  $t$  repetidos?

¡Amplifica el riesgo de error de tipo I!

número de muestras $i$	número de hipótesis $j$	Riesgo total $1 - 0.95^j$
2	1	0.05
3	3	0.14
4	6	0.26
5	10	0.40
6	15	0.54
10	45	0.90

El problema con tests  $t$  multiples

- Riesgo de error de tipo I más grande
- Solo considera variación para 2 muestras al mismo tiempo  
⇒ precisión baja
- No es posible considerar estructuras complicadas (e.g. 2 factores experimentales)  
⇒ El análisis de varianza se encarga de estos problemas

## Concepto del Anova

- Variables explicativas categóricas = factores
- $\geq 2$  niveles / grupos / tratamientos
- Dividir entre variación no explicada y variación explicada por las variables explicativas
- Ajustar modelos lineales para explicar o predecir valores de la variable dependiente

## Objetivos del Anova

- Examinar la contribución relativa de diferentes fuentes de variación sobre la cantidad total de variación de la variable dependiente
- Evaluar la hipótesis  $H_0$  que las medias de los grupos / tratamientos son iguales

145 / 1

## Varios tipos de anova

- 1 factor, 2 niveles → test  $t$
- 1 factor,  $\geq 3$  niveles → anova simple (one-way anova)
- $\geq 2$  factores → anova de 2 or 3 factores (two/three-way anova)
- Replicación por cada nivel → diseño factorial  $\Rightarrow$  permite estudiar las interacciones entre variables

146 / 1

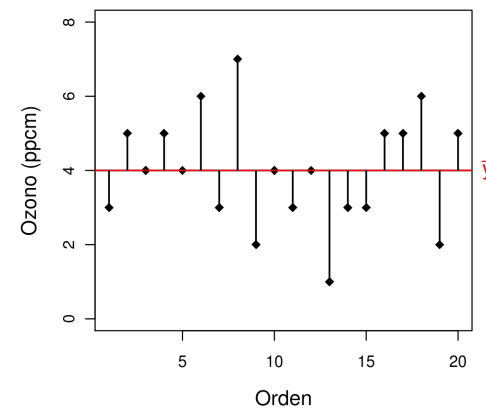
## Análisis de varianza ¿para comparar medias?

### Ejemplo: Cantidad de ozono

- Variable dependiente  $Y$ : concentración de ozono
- Variable explicativa: 1 factor JARDÍN, 2 niveles  $A$  y  $B$
- 10 réplicas por jardín
- ¿La concentración de ozono es la misma?

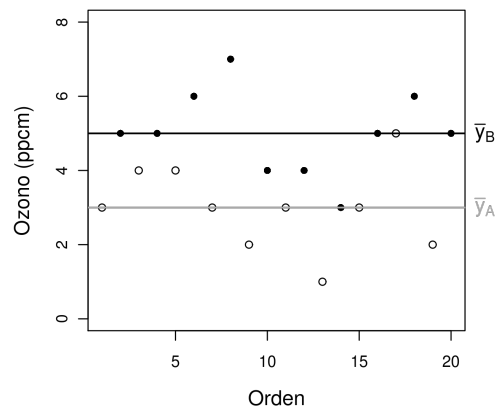
147 / 1

## Principio del Anova (1)



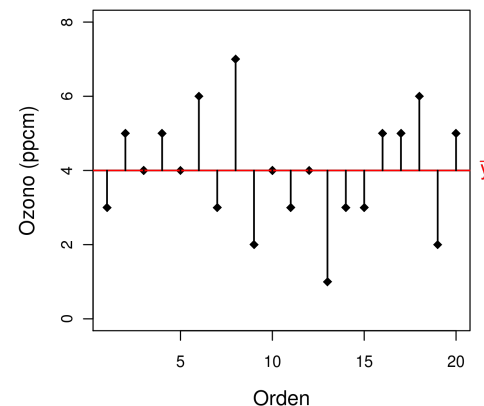
- Mucha dispersión
- Concentración media
- $SSY = \sum (y_i - \bar{y})^2$
- Residuales: suma total de los cuadrados (total sum of squares SSY)
- Variación entre los tratamientos

## Principio del Anova (2)



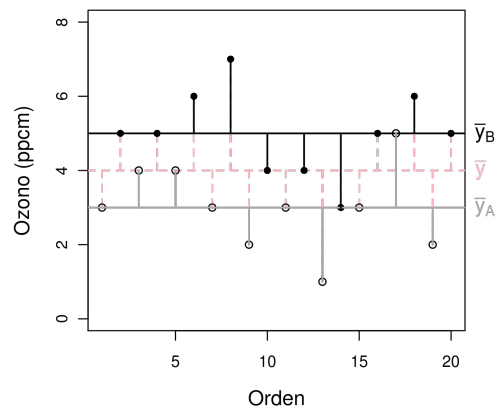
- Jardín A
- Jardín B
- $C_B > C_A$
- ¿La diferencia es significativa o no?

## Principio del Anova (3)



- ¿Qué pasa con los residuales si  $\bar{y}_A = \bar{y}_B$ ?
- ¿Y si  $\bar{y}_A \neq \bar{y}_B$ ?
- $SSE = \sum_{j=1}^k \sum (y_{ij} - \bar{y}_j)^2$
- Suma de cuadrados del error (Error sum of squares SSE)
- Variación dentro de los tratamientos
- ¿SSE versus SSY ?
- ¡SSE < SSY!

## Principio del Anova (3)



- ¿Qué pasa con los residuales si  $\bar{y}_A = \bar{y}_B$ ?
- ¿Y si  $\bar{y}_A \neq \bar{y}_B$ ?
- $SSE = \sum_{j=1}^k \sum (y_{ij} - \bar{y}_j)^2$
- Suma de cuadrados del error (Error sum of squares SSE)
- Variación dentro de los tratamientos
- ¿SSE versus SSY ?
- ¡SSE < SSY!

## Para resumir

Análisis de varianza para comparar medias

- Cuando  $\bar{y}_A \neq \bar{y}_B$ ,  $SSE < SSY$
- Variación total = modelo + error
- $SSY = SSA + SSE$
- SSA: proporción de varianza explicada
- Si  $SSE < SSY \Rightarrow \bar{y}_A \neq \bar{y}_B$

## De vuelta al jardín ...

- $SSY = 44$
- ¿Cuanto es atribuible a la diferencia entre  $\bar{y}_A$  y  $\bar{y}_B$ ?
- Jardín A:  $SSE_A = 12$ , Jardín B:  $SSE_B = 12$
- Suma de cuadrados de error  
 $SSE = SSE_A + SSE_B = 12 + 12 = 24$
- Suma de cuadrados del tratamiento:  
 $SSA = SSY - SSE = 44 - 24 = 20$

153 / 1

## Tabla de Anova

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón-F
Jardín	$SSA = 20.0$	1	20.0	15.0
Error	$SSE = 24.0$	18	$s^2 = 1.33$	
Total	$SSY = 44.0$	19		

- $F_{teo} = 4.41$ , ¿Qué se puede concluir?
- No se puede aceptar  $H_0$
- $\bar{y}_A \neq \bar{y}_B$
- Concentración de ozono es diferente entre los jardines A y B

154 / 1

## Condiciones del anova

¡Las mismas que por la regresión!

- Independencia
- Homogeneidad de las varianzas
- Normalidad

¡Condiciones sobre los residuales!  $\Rightarrow$  hacer los tests después del análisis

155 / 1

## Diseños factoriales

- $\geq 2$  factores
- $\geq 2$  niveles per factor
- Replicación para cada combinación de niveles
- Interacciones: respuesta a un factor depende del nivel de otro factor

156 / 1

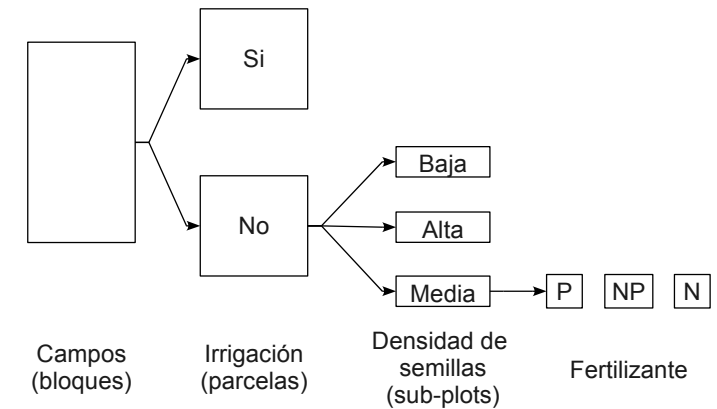
## Reconocer diseños complicados para evitar pseudoreplicación

(Nested design and Split plots)

- Muestreo jerárquico: medidas repetidas del mismo individuo o estudios con varias escalas espaciales
- Parcelas subdivididas: diferentes tratamientos en diferentes parcelas de diferentes tamaños

157 / 1

## Un ejemplo de diseño “split plot”



158 / 1

## Factores fijos

(Fixed effects)

- Todos los niveles están incluidos
- No extrapolación fuera de estos niveles
- Si se repite el estudio → mismos niveles
- Modelos con efectos fijos (fixed effects models)
- Anova tipo I
- Ejemplo: nivel de zinc (Fondo, bajo, medio alto), fertilizantes ...

159 / 1

## Factores aleatorios

(Random effects)

- Muestra aleatoria de los niveles posibles
- Inferencia (extrapolación) sobre todos los grupos
- Si se repite el estudio → otros niveles
- Modelos de efectos aleatorios (random effect models)
- Anova tipo II
- Ejemplo: Sitios de estudio, ...

160 / 1