

Introducción a la estadística

Bases indispensables y uso de 

Olivier Devineau

`olivier.devineau@fcdarwin.org.ec`

Fundación Charles Darwin

Taller interno, 27–30 abril 2010

1 / 161

Introducción y conceptos importantes

2 / 161

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!

3 / 161

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

4 / 161

Agradecimientos

Y también:

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

5 / 161

Licencia 

- Este documento está bajo la licencia Creative Commons: *Reconocimiento - No comercial - Compartir bajo la misma licencia 3.0 Ecuador*
- Para ver una copia de esta licencia, visite:
<http://creativecommons.org/licenses/by-nc-sa/3.0/ec/>
- Código \LaTeX a petición

6 / 161

¿Qué es la estadística?

Definición

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

7 / 161

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

8 / 161

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

9 / 161

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

10 / 161

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

11 / 161

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!

12 / 161

¿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

13 / 161

¿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?

14 / 161

¿Qué análisis? Guía de decisión

1) Variables independientes

- Todas continuas
- Todas categóricas
- Ambas continuas y categóricas

Regresión

Anova

Ancova

15 / 161

¿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 supervivencia |

16 / 161

Por qué la estadística?

¡Porque Todo varia!

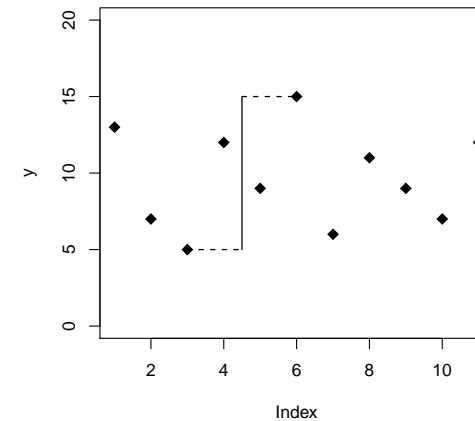
Mucha variabilidad temporal, espacial y entre individuos:

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

17 / 161

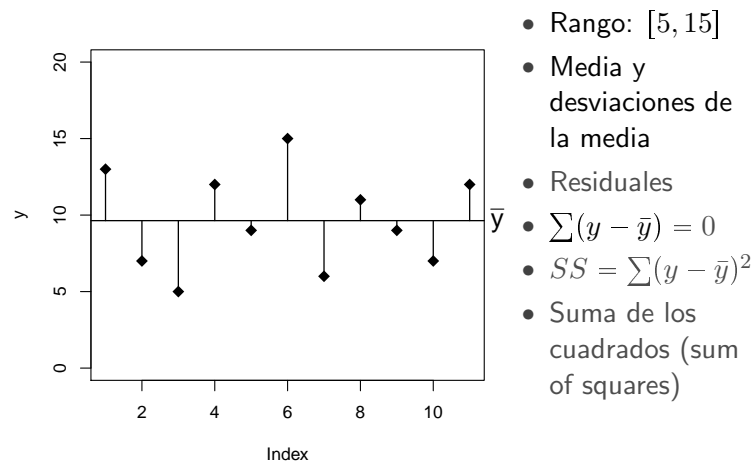
¿Como medir la variabilidad?

- Rango: [5, 15]



18 / 161

¿Como medir la variabilidad?



19 / 161

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)

20 / 161

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

21 / 161

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

22 / 161

Varianza (2)

Formalización y definición

- Medida cuantitativa de la variabilidad:

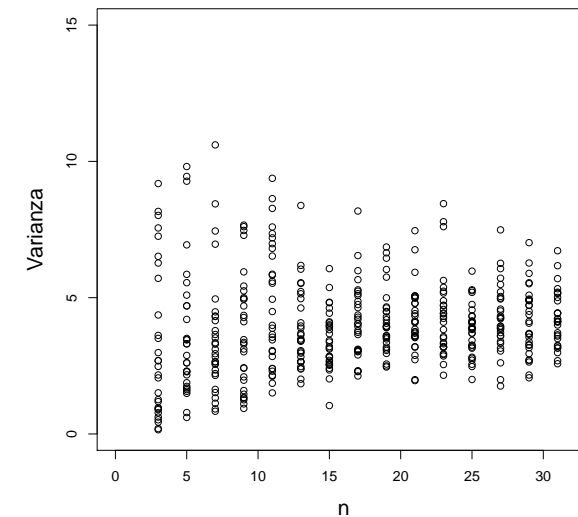
$$\text{Varianza} = \frac{\text{Suma de cuadrados}}{\text{Grados de libertad}} = \frac{SS}{df}$$

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

23 / 161

Varianza y tamaño de muestra

Media: 10, Varianza: 4



24 / 161

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?
- $SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$

25 / 161

Intervalos de confianza

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

26 / 161

Diseño experimental

Conceptos claves

Replicación: aumenta fiabilidad

Aleatorización: reduce sesgo

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

27 / 161

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

28 / 161

Replicación (2)

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

29 / 161

¿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): ≥ 30

30 / 161

Poder y réplicas

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

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

31 / 161

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?

32 / 161

¿Qué hacer con pseudoreplicación?

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

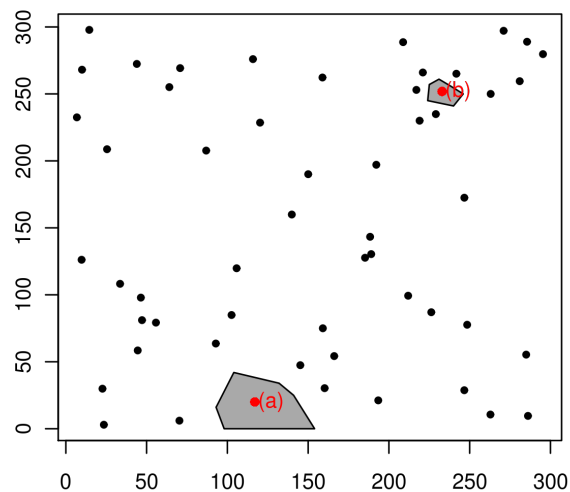
33 / 161

Aleatorización

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

34 / 161

Selección aleatoria de un árbol



35 / 161

Controles

- No controles, no conclusiones

36 / 161

¿Cuánto tiempo?

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

37 / 161

Inferencia fuerte

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

38 / 161

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

39 / 161

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)

40 / 161

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

41 / 161

La navaja de Einstein

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

42 / 161

Máximo de verosimilitud

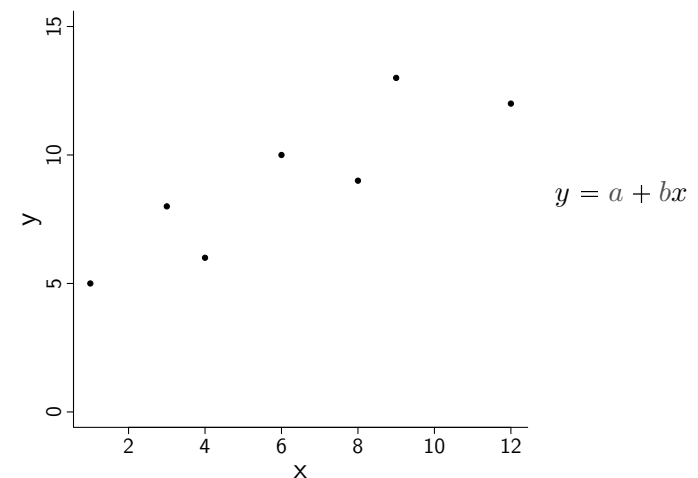
(Maximum Likelihood: ML)

- 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

43 / 161

Máximo de verosimilitud

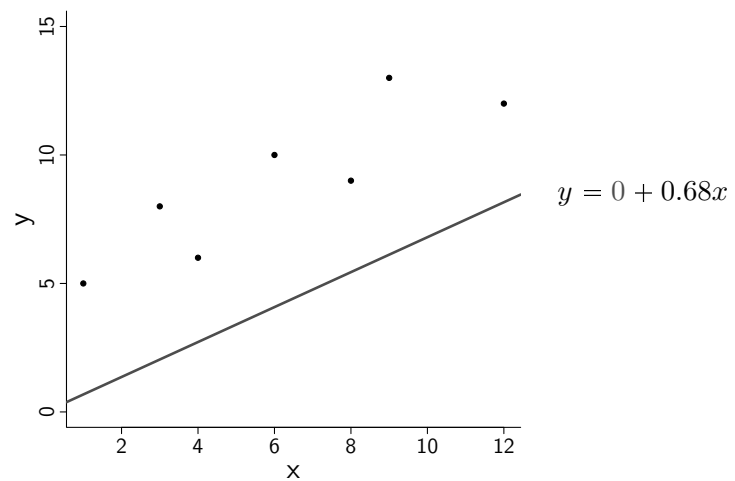
Ejemplo: regresión $y = a + bx$



44 / 161

Máximo de verosimilitud

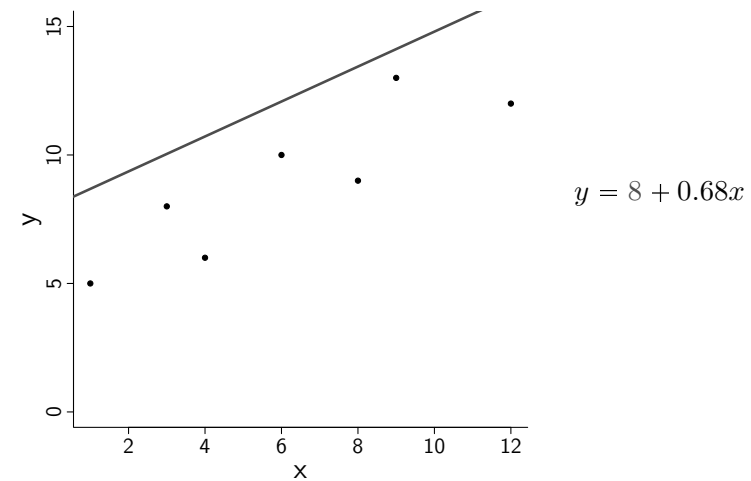
Ejemplo: regresión $y = a + bx$



45 / 161

Máximo de verosimilitud

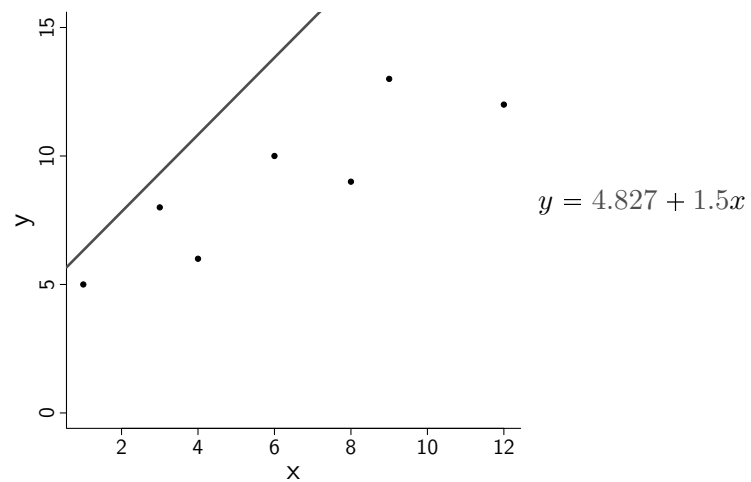
Ejemplo: regresión $y = a + bx$



46 / 161

Máximo de verosimilitud

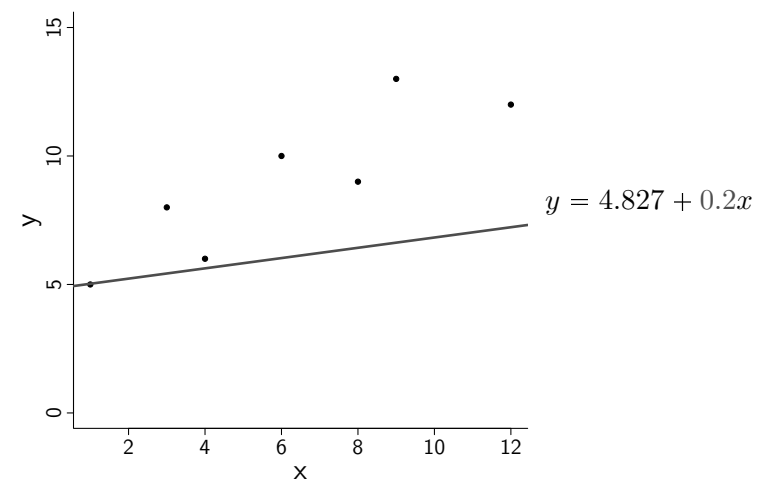
Ejemplo: regresión $y = a + bx$



47 / 161

Máximo de verosimilitud

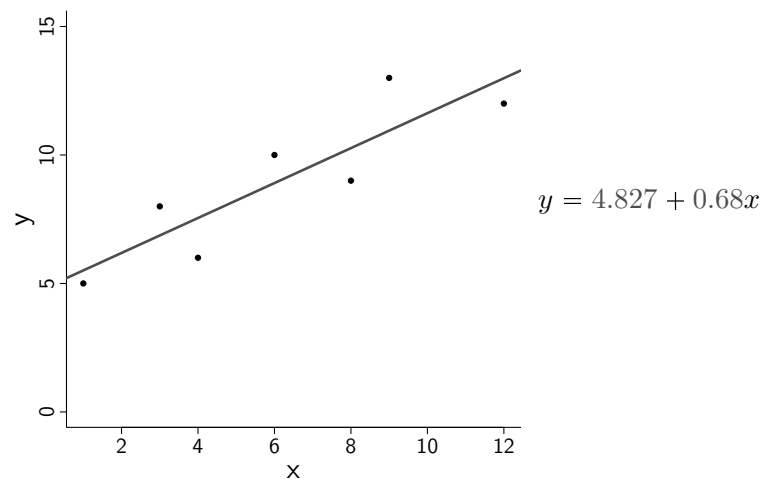
Ejemplo: regresión $y = a + bx$



48 / 161

Máximo de verosimilitud

Ejemplo: regresión $y = a + bx$



49 / 161

Noción de test estadístico

50 / 161

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

- $2\varnothing, (1\sigma, 1\varnothing), (1\varnothing, 1\sigma), 2\sigma$
 - $p(X = 0 \sigma) = 1/4$
 - $p(X = 1 \sigma) = 1/4 + 1/4$
 - $p(X = 2 \sigma) = 1/4$
- $\left. \vphantom{\begin{matrix} p(X = 1 \sigma) = 1/4 + 1/4 \\ p(X = 2 \sigma) = 1/4 \end{matrix}} \right\} \sum p(X) = 1$

51 / 161

Distribución binomial

Definición

- Serie de n intentos independientes
- Cada intento \rightarrow Éxito / Fracaso
- Probabilidad de éxito: p
- Distribución discontinua
- $X \sim \mathcal{B}(n, p)$
- $P(r) = \binom{n}{r} p^r (1 - p)^{n-r}$

52 / 161

Distribución Binomial (2)

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



53 / 161

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*

54 / 161

Distribución de Poisson

Definición

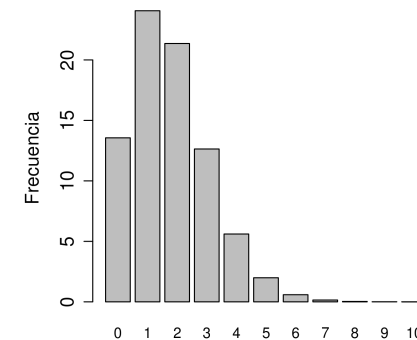
- Cuántas veces un evento raro ocurre por unidad de tiempo/espacio
- Distribución discontinua
- $X \sim \mathcal{P}(\lambda)$
- $P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$

55 / 161

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



56 / 161

Distribución normal

Definición

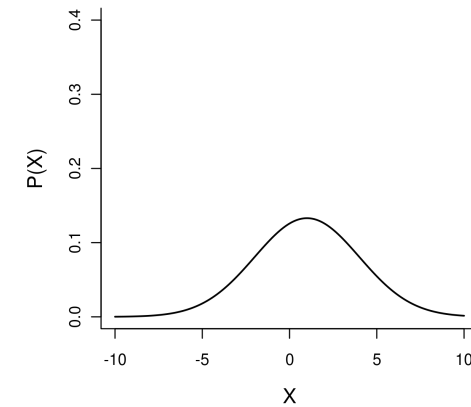
- Teorema del límite central
- Suficientes muestras \rightarrow medias \rightarrow distribución normal
- Distribución continua
- $X \sim \mathcal{N}(\mu, \sigma)$
- $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$

57 / 161

Distribución normal

¿Cuándo se aplica?

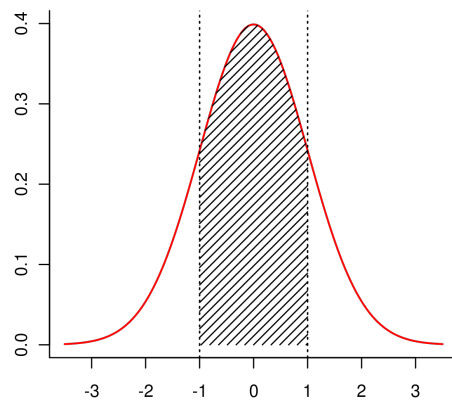
- ¡Todo el tiempo!
- Regresión lineal, análisis de varianza ...



58 / 161

Distribución Normal Estándar

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

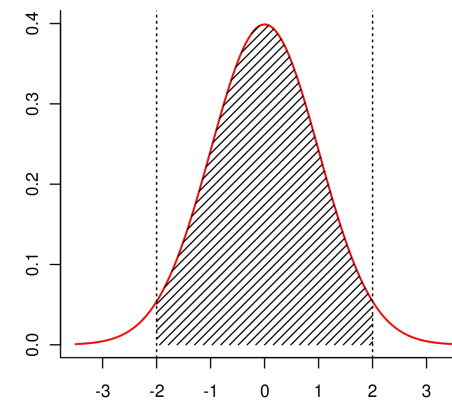


- $\pm 1 \sigma \sim 68\%$

59 / 161

Distribución Normal Estándar

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

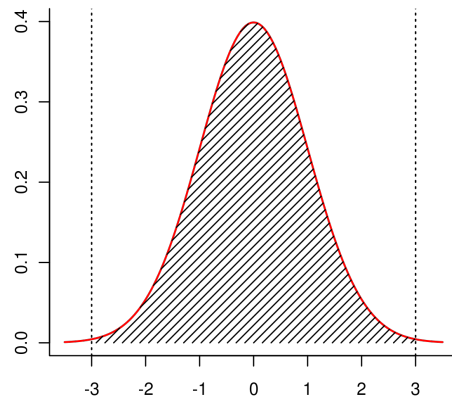


- $\pm 2 \sigma \sim 95\%$

60 / 161

Distribución Normal Estándar

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



- $\pm 3 \sigma \sim 99\%$

61 / 161

Otras distribuciones de variables

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

62 / 161

Distribuciones de estadísticos

- Distribución z
- Distribución t de Student
- Distribución del χ^2
- Distribución F de Fischer

63 / 161

¿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

64 / 161

Test estadístico: procedimiento

- ❶ Pregunta biológica: ¿Hay cóndores en el parque?
- ❷ Pregunta estadística: Hipótesis H_0
- ❸ Elección del test estadístico: ¿Cuál usar?
- ❹ Criterios de decisión: ¿Qué riesgo de error? ¿Qué nivel de confianza?

65 / 161

Test estadístico: procedimiento

- ❺ ¡Colección de los datos!
- ❻ Cálculo de el estadístico del test
- ❼ Decisión estadística: ¿Se puede rechazar H_0 o no?
- ❽ Inferencia y explicación biológica

66 / 161

Buenas y malas hipótesis

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

67 / 161

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

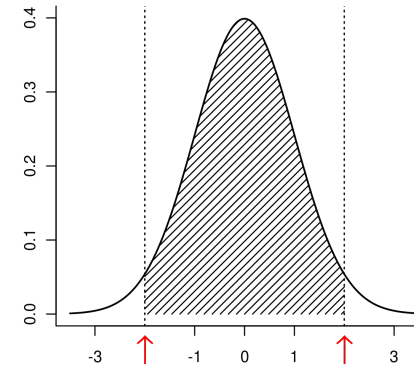
68 / 161

Elección del test

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

69 / 161

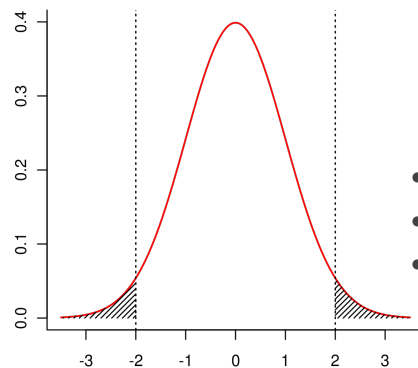
Criterios de decisión (1)



- $\pm 2 \sigma \sim 95\%$
- Valores umbrales
- Región de aceptación

70 / 161

Criterios de decisión (1)



- 5% menos probable
- Región de rechazo
- Riesgo α

71 / 161

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 β
Rechaza	Tipo I Riesgo α	Decisión correcta

72 / 161

Hay que comprometer ...

Poder: Probabilidad de rechazar H_0 cuando es falsa

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

73 / 161

¿Cuando α 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

74 / 161

¿Cuándo α debe ser alto?

Aceptar riesgo α más alto para reducir riesgo β

α alto: error de tipo I

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

β alto: error de tipo II

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

75 / 161

Colección de los datos

¡Acuérdense!

- Aleatorización
- Replicación

76 / 161

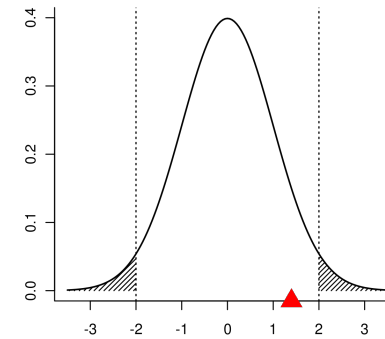
Computación del estadístico del test

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

77 / 161

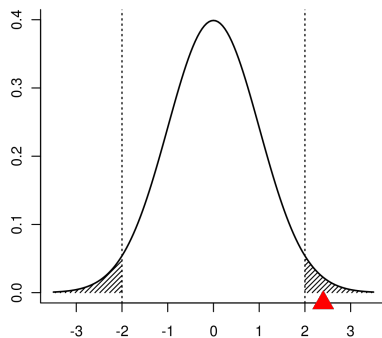
Comparación de T con la distribución teórica



- 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

78 / 161

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 $\alpha = 5\%$

79 / 161

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

80 / 161

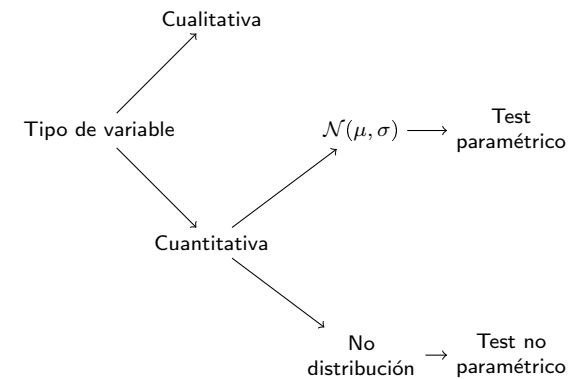
Significancia

- ¿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

81 / 161

¿Como elegir el test adecuado?

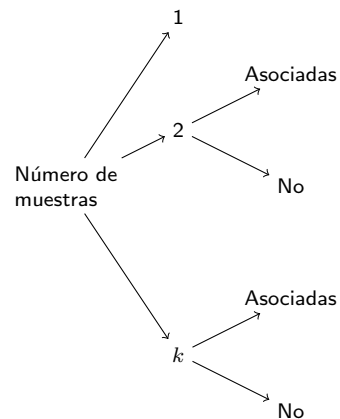
Algunas directrices (1)



82 / 161

¿Como elegir el test adecuado?

Algunas directrices (2)



83 / 161

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”

84 / 161

Comparar una muestra con una distribución teórica

- ⇒ Test de conformidad
- Test t de conformidad
 - Test de Wilcoxon
 - Test binomial
 - Test χ^2 de conformidad
 - ...

85 / 161

Comparar dos muestras

- ⇒ Test de comparación (de homogeneidad)
- Test t (posiblemente “asociado”)
 - Test de Mann-Whitney
 - Test de Fisher
 - Test χ^2
 - ...

86 / 161

Comparar *más* de dos muestras

- ⇒ Test de comparación (continuación)
- Anova / Manova
 - Test de Kruskal-Wallis
 - Test de Friedman
 - Test χ^2
 - ...

87 / 161

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 ($n|handout : 1 > 2$)
 - ...

88 / 161

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 χ^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 χ^2 , test binomial
Test t no asociado	Test de Mann-Whitney	Test de Fisher, test χ^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 χ^2 , test binomial
Test t no asociado	Test de Mann-Whitney	Test de Fisher, test χ^2
Test t asociado	Test de Wilcoxon	Test de McNemar

Comparar ≥ 3 grupos no asociados

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

Comparar ≥ 3 grupos asociados

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

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
- *GraphPad / Intuitive Biostatistics:*
www.graphpad.com/www/Book/Choose.htm
- *Social Research Methods:*
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

97 / 161

Correlación y regresión

98 / 161

Dos categorías de tests estadísticos

Tests de comparación : 1 variable, ≥ 2 poblaciones

Tests de relación : ≥ 2 variables, 1 población

99 / 161

≥ 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

\Rightarrow ¿Cuál es la relación entre las variables?

100 / 161

Relación entre ≥ 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

101 / 161

Estadística correlacional

Correlación: ¿Cómo 2 variables varían juntas?

Regresión: Relación entre 1 variable dependiente y ≥ 1 variable independiente

Análisis multivariados: Relación entre ≥ 2 variables independientes / dependientes / ambos

102 / 161

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?

103 / 161

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: $cov(X_1, X_2)$

\Rightarrow Coeficiente de correlación

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

104 / 161

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: ρ

105 / 161

¿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, $r \sim \mathcal{N}(\mu, \sigma)$
 \Rightarrow uso de test t de Student

106 / 161

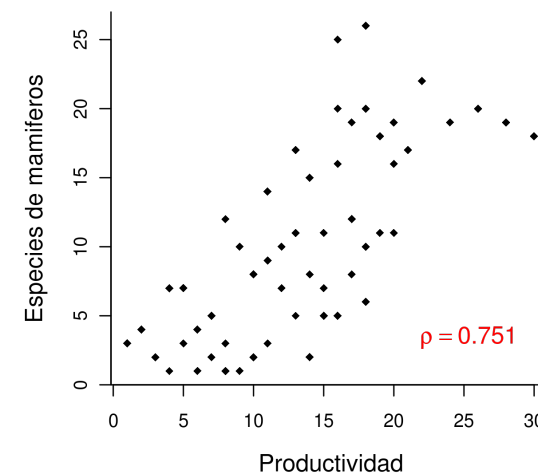
Correlación no paramétrica

- ¿Qué hacer cuando los requisitos no se cumplen?
- \Rightarrow Coeficiente de correlación de rango
- de Spearman: ρ
 - de Kendall: τ
 - ¡Más conservadores!

107 / 161

La correlación depende de la escala

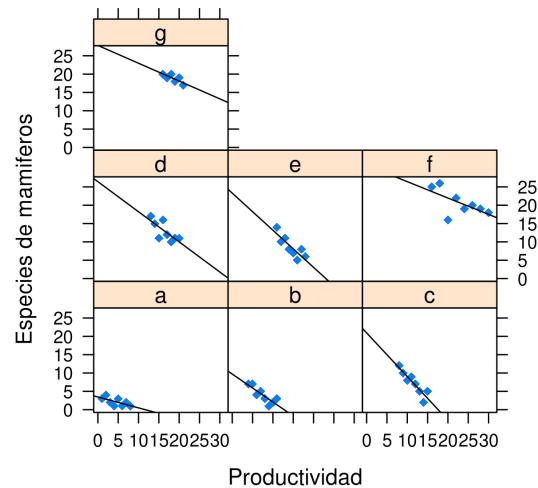
¡Las cosas no son siempre como parecen!



108 / 161

La correlación depende de la escala

¡Las cosas no son siempre como parecen!



109 / 161

Modelo lineal: concepto general

- Se puede identificar:
 - 1 variable respuesta / dependiente Y
 - ≥ 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

110 / 161

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

111 / 161

¿Qué significa lineal?

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

112 / 161

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 → regresión lineal

113 / 161

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

114 / 161

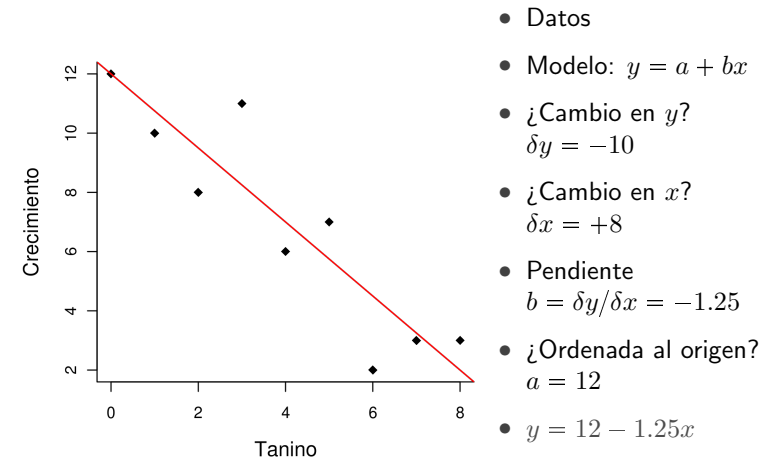
Análisis de regresión lineal

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

115 / 161

Principio de la regresión lineal



116 / 161

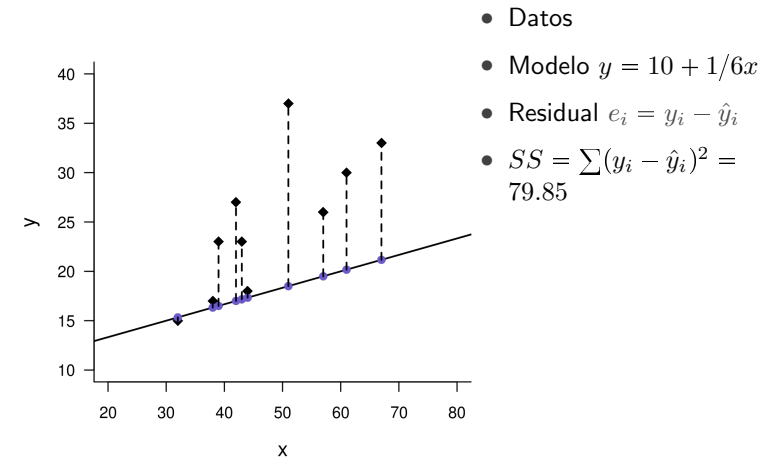
Principio de la regresión lineal (2)

- 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)

117 / 161

Cuadrados mínimos: principio

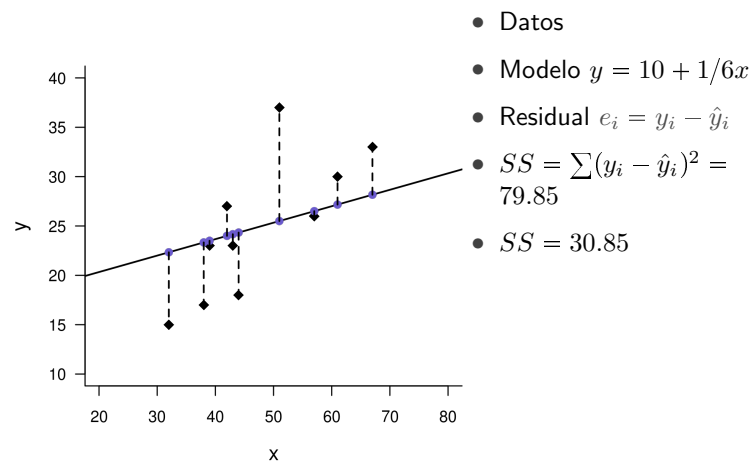
OLS: Ordinary Least Squares



118 / 161

Cuadrados mínimos: principio

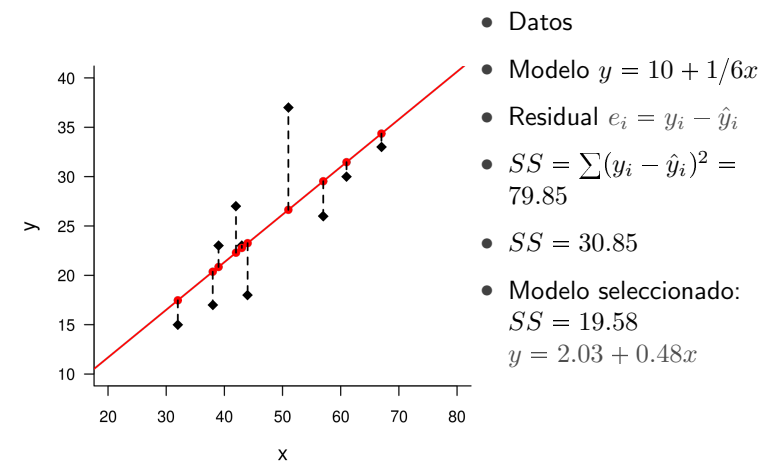
OLS: Ordinary Least Squares



119 / 161

Cuadrados mínimos: principio

OLS: Ordinary Least Squares



120 / 161

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

121 / 161

Varianza explicada

r^2 : coeficiente de determinación

- Variación de Y explicada por la relación con X
- (coeficiente de correlación)²
- $r^2 \in [0, 1]$
- ¿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

122 / 161

Comparar varios modelos

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

123 / 161

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)

124 / 161

Comparar modelos de regresión (2)

Siempre la misma lógica

- Medir el ajuste de cada modelo
- Comparar los ajustes de diferentes 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 $H_0 : \beta_1 = 0$

125 / 161

Condiciones del análisis de regresión (1)

- Involucran de los términos de errores (ε_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

126 / 161

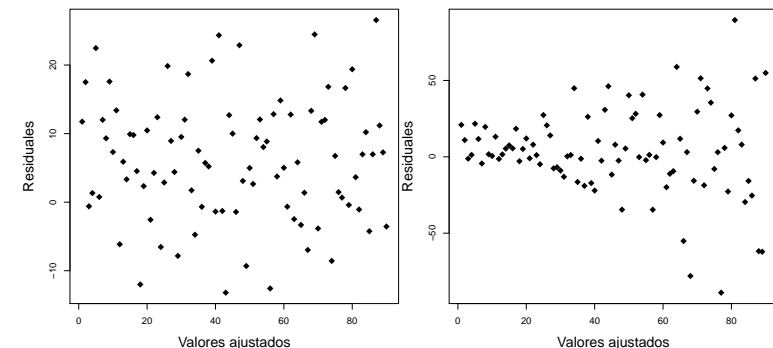
Condiciones del análisis de regresión (2)

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

127 / 161

Homogeneidad de la varianza

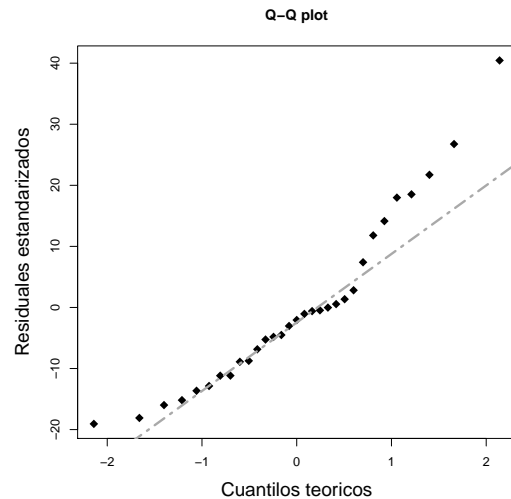
- No tendencia
- Heteroscedasticidad



- Test de Levene, test de Bartlett

128 / 161

Normalidad de los residuales



- Test de Shapiro-Wilk

129 / 161

¿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

130 / 161

Si el modelo es inadecuado, se puede...

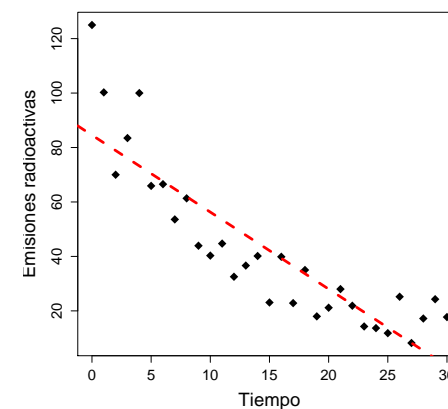
- Transformar variable dependiente
- Transformar ≥ 1 variable explicativa
- Probar otras variables explicativas
- Usar una estructura de error diferente (GLM)
- Usar alternativa no paramétrica (smoothing)
- Usar pesos diferentes por diferentes valores de y

131 / 161

Regresión polinomial

Ejemplo: Desintegración radioactiva

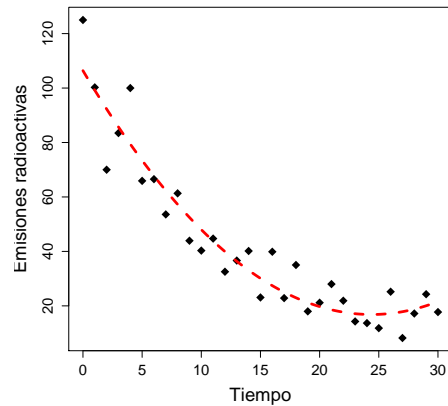
- Regresión lineal:
 $y = ax + b$



132 / 161

Regresión polinomial

Ejemplo: Desintegración radioactiva

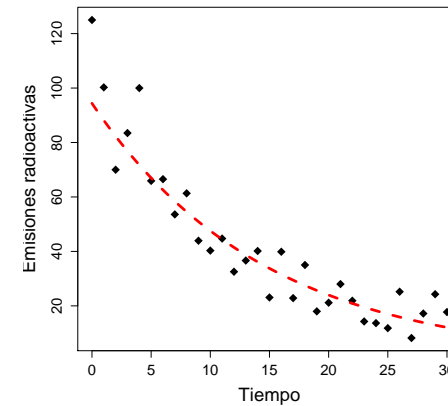


- Regresión polinómica
- $X_2 = X^2$
- $y = ax^2 + bx + c$

133 / 161

Regresión polinomial

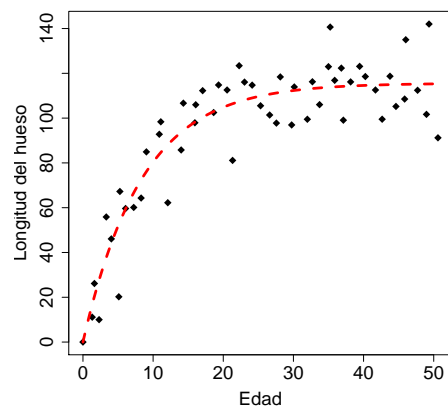
Ejemplo: Desintegración radioactiva




- $y = ae^{-bx}$
- ¡Descripción, no explicación!

134 / 161

Regresión no lineal y GAM



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

135 / 161

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)

136 / 161

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

137 / 161

(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

138 / 161

Análisis de varianza

139 / 161

Comparar ≥ 2 muestras

Control biológico de las plagas del maíz

Ejemplo: 5 tratamientos

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

140 / 161

Control biológico (2)

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

141 / 161

¿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 $H_0 : \mu_3 = \mu_6$
 - ⑩ $H_0 : \mu_4 = \mu_5$
- Cada hipótesis: riesgo de error de tipo I
 - Con 1 hipótesis: $\alpha = 0.05$
 - ¿Valor de α con 2 hipótesis?
 - ¿0.025, 0.05, 0.0725, 0.0975, 0.10?
 - $1 - Pr(\text{no error de tipo I})$
 - $1 - 0.95 \cdot 0.95 = 0.0975$

142 / 161

¿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

143 / 161

El problema con tests t multiples

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

144 / 161

Concepto del Anova

- Variables explicativas categóricas = factores
- ≥ 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

145 / 161

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

146 / 161

Varios tipos de anova

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

147 / 161

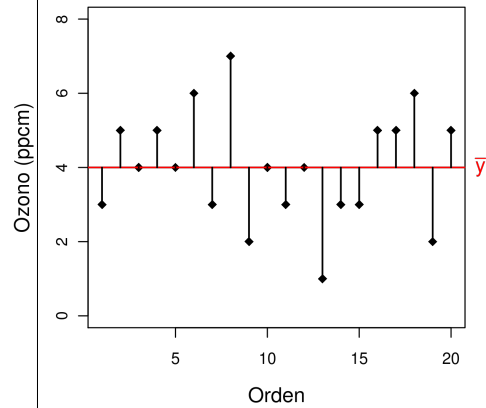
Análisis de varianza ¿para comparar medias?

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?

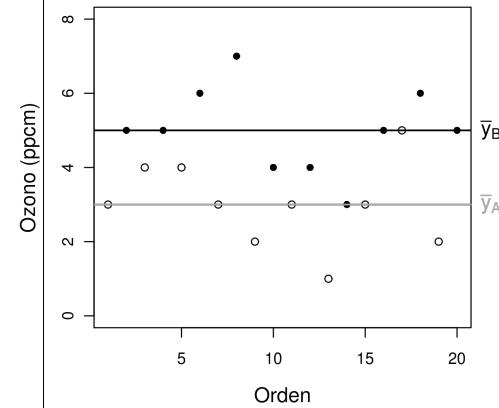
148 / 161

Principio del Anova (1)



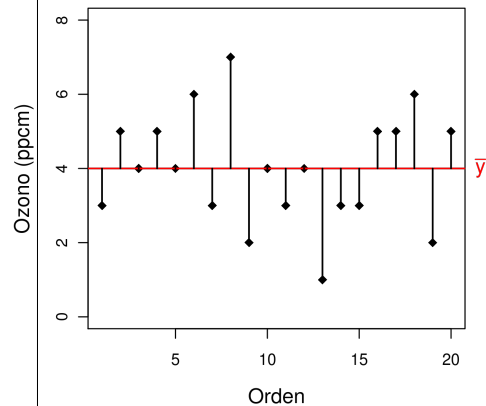
- Mucha dispersión
- Concentración media
- $SSY = \sum (y_i - \bar{y})^2$
- Residuales: suma total de los cuadrados (total sum of squares SSY)
- Variación entre los tratamientos

Principio del Anova (2)



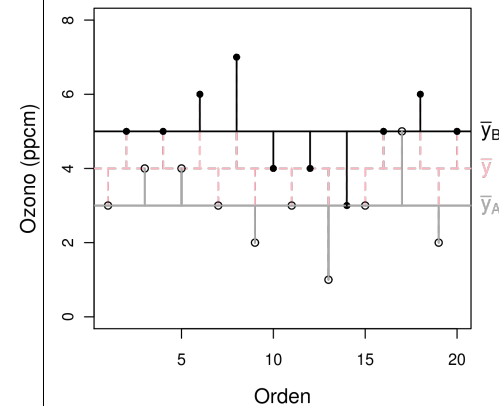
- Jardín A
- Jardín B
- $C_B > C_A$
- ¿La diferencia es significativa o no?

Principio del Anova (3)



- ¿Qué pasa con los residuales si $\bar{y}_A = \bar{y}_B$?

Principio del Anova (3)



- ¿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 $SSE < SSY \Rightarrow \bar{y}_A \neq \bar{y}_B$

153 / 161

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:
 $SSA = SSY - SSE = 44 - 24 = 20$

154 / 161

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

155 / 161

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

156 / 161

Diseños factoriales

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

157 / 161

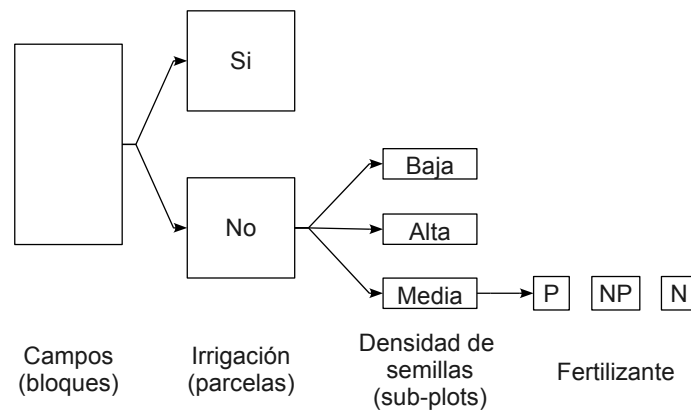
Reconocer diseños complicados para evitar pseudoreplicación

(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

158 / 161

Un ejemplo de diseño “split plot”



159 / 161

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 ...

160 / 161

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, . . .