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

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

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.

Licencia

¿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

$\Rightarrow$ 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? ¿Categóricas? ¿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?

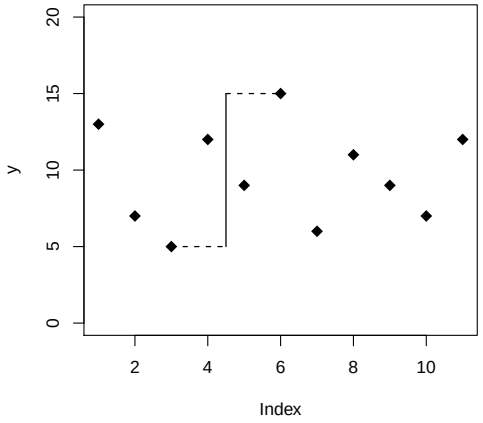
¡Porque Todo varia!

Mucha variabilidad temporal, espacial y entre individuos:

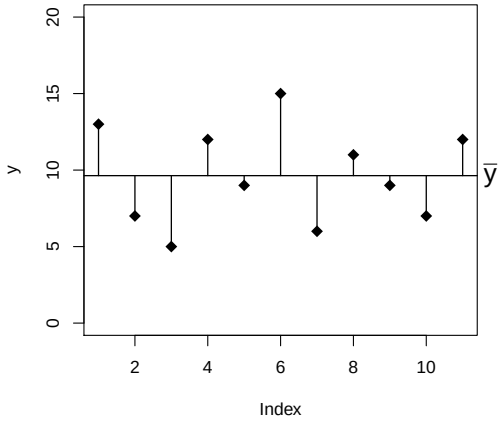
- Genética
- Factores ambientales
- Azar
- Errores de observación y medida

¿Como medir la variabilidad?

- Rango: $[5, 15]$



¿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)

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
- $MS = \dfrac{\sum(y-\bar{y})^2}{n}$
- Desviación cuadrática media (Mean square deviation MS)

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

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

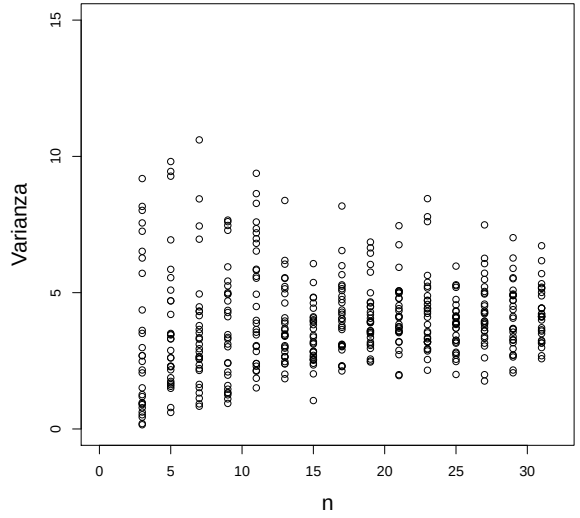
- Medida cuantitativa de la variabilidad:

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

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

Varianza y tamaño de muestra

Media: 10, Varianza: 4



Una medida de fiabilidad

¡Error estándar de la media!

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

Intervalos de confianza

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

Diseño experimental

Conceptos claves

Replicación: aumenta fiabilidad

Aleatorización: reduce sesgo

- Si replican y randomizan correctamente, ¡no hay problema!
- Diseño inadecuado $\not\rightarrow$ 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
  $\Rightarrow$ Indicación sobre varianza base y magnitud de la respuesta al tratamiento
- Método práctico (en general): $\geqslant 30$

Poder y réplicas

- Poder: probabilidad de rechazar H_0 cuando es falsa
- ¿Cuantas réplicas para detectar un efecto δ con 80% probabilidad de no cometer un error?
- Experiencia y/o estudio piloto
  $\Rightarrow$ Primera estimación del efecto δ 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 $\rightarrow$ seudoreplicación temporal
- Varias medidas del mismo lugar $\rightarrow$ seudoreplicación spacial
- ¿Cuántos grados de libertad?

¿Qué hacer con seudoreplicación?

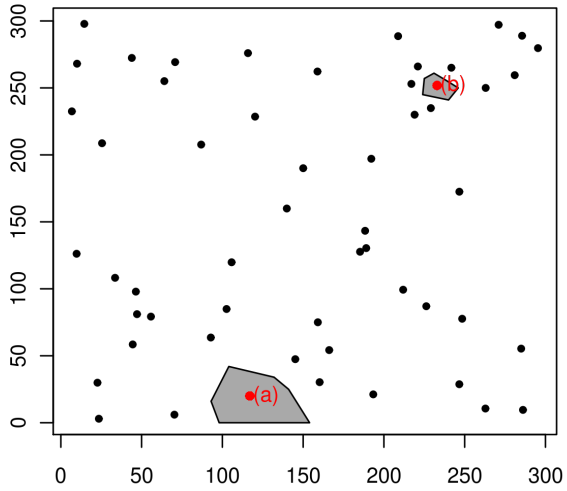
- Promediar seudoreplicación y hacer análisis sobre medias
- Hacer análisis separados por cada período de tiempo
- Usar análisis de series de tiempo o modelos de efectos mixtos

Aleatorización

- ¿Cómo seleccionar un árbol al azar en una selva?
- ¿Hojas accesibles?
- ¿Cerca del laboratorio?
- ¿Parece sano?
- ¿Sin insectos?

$\Rightarrow$ ¡Sesgo en la fotosíntesis!

Selección aleatoria de un árbol



Controles

- No controles, no conclusiones

¿Cuánto tiempo?

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

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

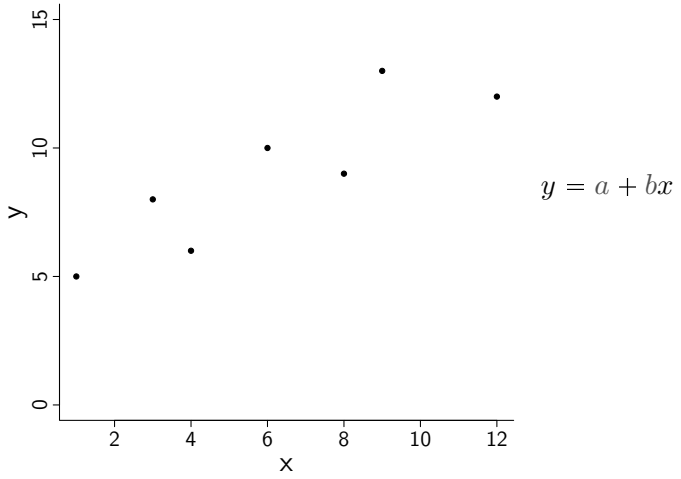
(Maximum Likelihood: ML)

- Dado los datos
- Y dado un modelo
- ¿Qué valores de parámetros hacen a los datos observados más probables?

$\Rightarrow$ Estimadores sin sesgo que minimizan la varianza

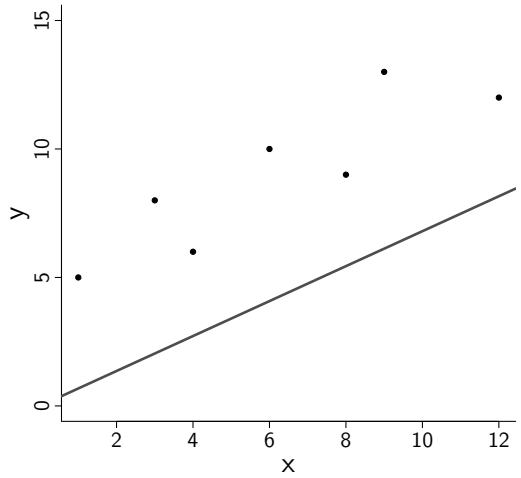
Máximo de verosimilitud

Ejemplo: regresión $y = a + bx$



Máximo de verosimilitud

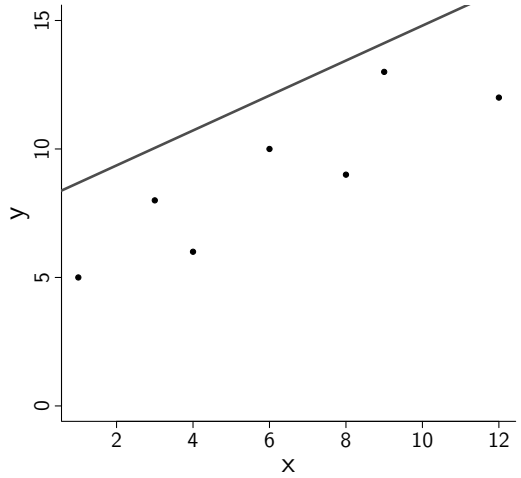
Ejemplo: regresión $y = a + bx$



$y = 0 + 0.68x$

Máximo de verosimilitud

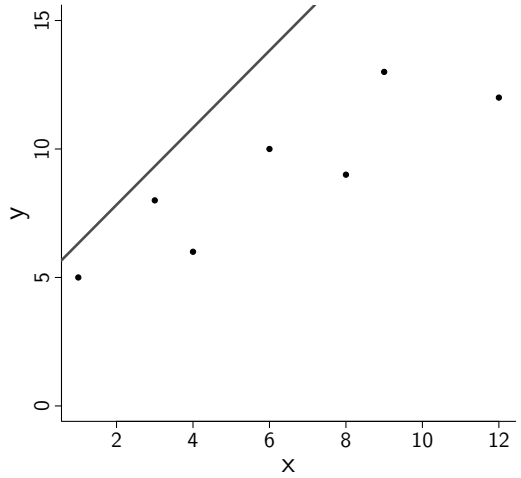
Ejemplo: regresión $y = a + bx$



$y = 8 + 0.68x$

Máximo de verosimilitud

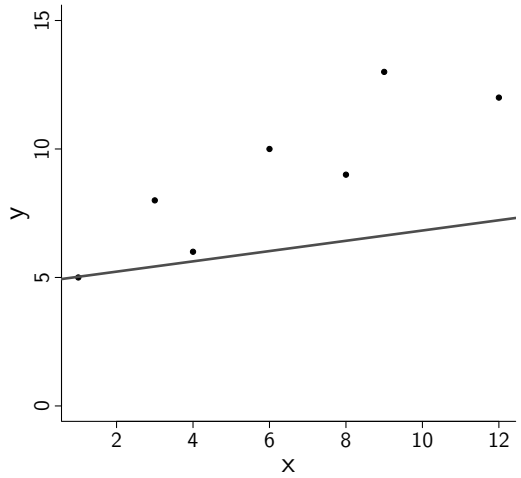
Ejemplo: regresión $y = a + bx$



$y = 4.827 + 1.5x$

Máximo de verosimilitud

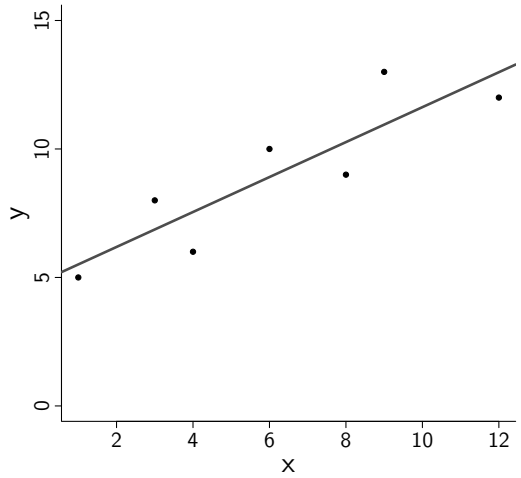
Ejemplo: regresión $y = a + bx$



$y = 4.827 + 0.2x$

Máximo de verosimilitud

Ejemplo: regresión $y = a + bx$



$$y = 4.827 + 0.68x$$