Introducción a la estadística

Bases indispensables y uso de R

Olivier Devineau
olivier.devineau@fcdarwin.org.ec

Fundación Charles Darwin

Taller interno, 27–30 abril 2010

Análisis de varianza

Comparar $\geqslant 2$ 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
  $\Rightarrow$ Media: μ_i, desviación estándar: σ_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. $H_0 : \mu_3 = \mu_6$
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?
- $1 - Pr(no\ error\ de\ tipo\ I)$
- $1 - 0.95 \cdot 0.95 = 0.0975$

¿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 $\Rightarrow$ precisión baja
- No es posible considerar estructuras complicadas (*e.g.* 2 factores experimentales)

  $\Rightarrow$ El análisis de varianza se encarga de estos problemas

Concepto del Anova

- Variables explicativas categóricas = factores
- $\geqslant 2$ 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

Varios tipos de anova

- 1 factor, 2 niveles $\rightarrow$ test t
- 1 factor, $\geqslant 3$ niveles $\rightarrow$ anova simple (one-way anova)
- $\geqslant 2$ 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

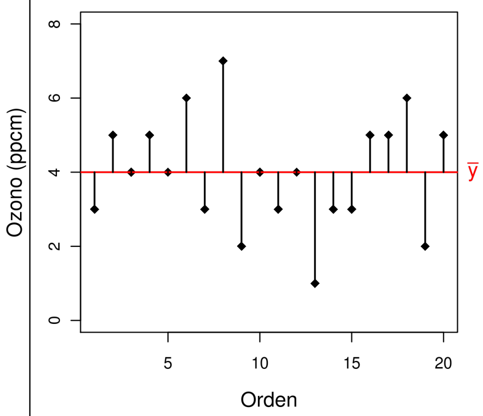
Análisis de varianza ¿para comparar medias?

Ejemplo: Cantidad de ozono

- Variable dependiente Y: concentración de ozono
- Variable explicativa: 1 factor $\textsc{Jardín}$, 2 niveles A y B
- 10 réplicas por jardín

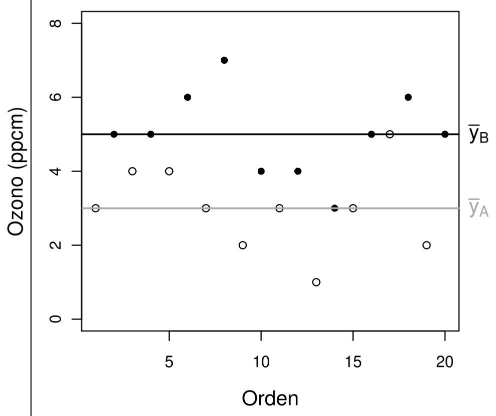
- ¿La concentración de ozono es la misma?

Principio del Anova (1)



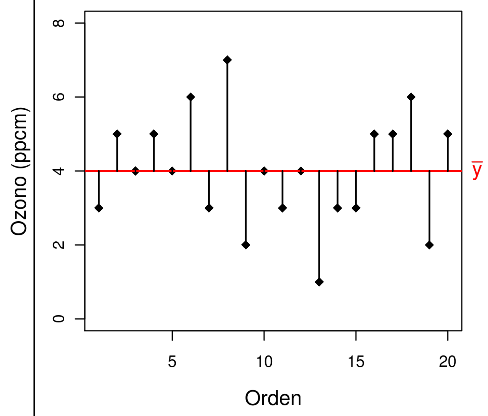
- Mucha dispersión
- Concentración media
- $SSY = \sum(y_i - \bar{y})^2$
- Residuales: suma total de los cuadrados (total sum of squares SSY)
- Variación entre los tratamientos

Principio del Anova (2)



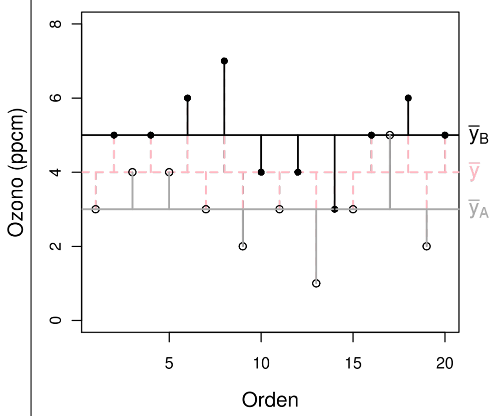
- Jardín A
- Jardín B
- $C_B > C_A$
- ¿La diferencia es significativa o no?

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

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

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

Condiciones del anova

¡Las mismas que por la regresión!

- Independencia
- Homogeneidad de las varianzas
- Normalidad

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

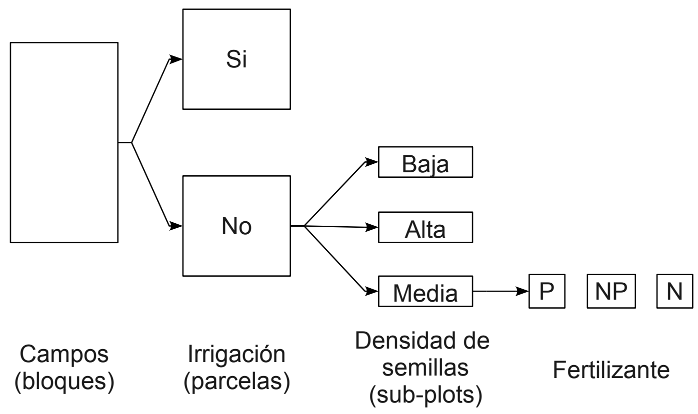
Diseños factoriales

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

Reconocer diseños complicados para evitar seudoreplicación
(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

Un ejemplo de diseño "split plot"



Factores fijos
(Fixed effects)

- Todos los niveles estan 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 ...

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, ...