MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull

Índice

Inferencia estadística

Diseño de experimentos

Tipos de diseños según variables y recogida

- Según tipo de variable
  - Numérica. Comparación de medias o de variancias.
    [Ej: comparar tiempos de entrega de envíos]
  - Dicotómica. Comparación de proporciones.
    [Ej: comparar probabilidad de envío entregado]
- Según recogida de datos
  - Muestras independientes*:* cada caso es una **medida independiente**
    [Ej: Cada envío se realiza por una de las dos compañías]
  - Muestras pareadas*:* cada caso da lugar a dos medidas, **pares de medidas**
    [Ej: Se emparejan los envíos y se hacen con las dos compañías al unísono y se considera su diferencia]
  - Siempre que sea viable, es preferible un diseño apareado para mejorar la eficiencia

Comparación de medias (μ_1=μ_2). Muestras independientes

Proceso

- Prueba formal

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Bajo H_0 se cumple

$$\hat{t} = \frac{(\bar{y}_1 - \bar{y}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- **Premisas**: Variables a comparar normales y varianzas iguales en ambas muestras.
- Como las varianzas se considerar iguales, se realiza una estimación conjunta (s)
- R: `t.test(YA, YB, var.equal=TRUE)`

Comparación de medias ($\boldsymbol{\mu_1 = \mu_2}$). Muestras pareadas

Proceso

- Cada unidad proporciona información sobre la diferencia. Se define una nueva variable diferencia **D** para cada unidad.
- El estudio se basa en la variable D:

$$H_0: \mu_A - \mu_B = 0 \;\rightarrow H_0: \mu_D = 0$$

- Por tanto, el estadístico es equivalente al de la PH de μ para una muestra

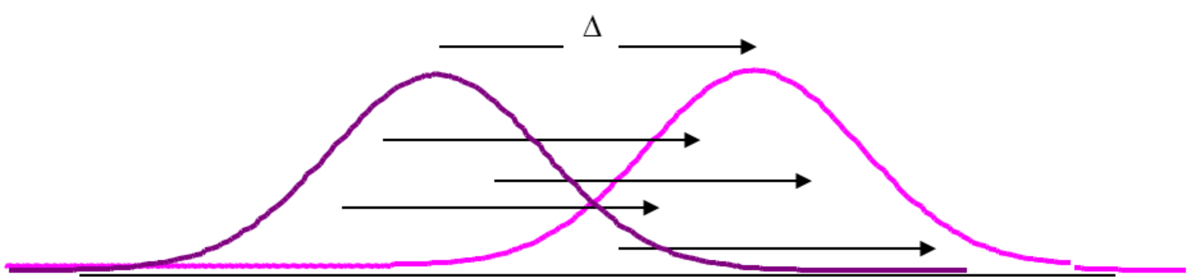
$$\hat{t} = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1}$$

  - $\overline{D}$: media de la variable diferencia (D)
  - S_D : desviación típica de D
  - μ_D : valor a contrastar (0 en caso de igualdad)
- **R:** `t.test(Y_A, Y_B, paired=TRUE)`

Comparación de medias ($\mu_1=\mu_2$). Muestras independientes

Efecto aditivo o multiplicativo

- En la prueba sobre los parámetros μ_1 i μ_2 a partir de les dos medias $\bar{y}_1$ i $\bar{y}_2$ de las dos muestras, al hacer la diferencia se mide un desplazamiento Δ. Este desplazamiento puede dar lugar a un:
  - Efecto aditivo: cada caso tiene el mismo desplazamiento (la compañía A entrega los envíos, en promedio, 30' más tarde que la compañía B)
  - Efecto multiplicativo: el tiempo es proporcional entre compañías (la compañía A entrega los envíos con un 10% de retraso respecto a la compañía B)
- Un **efecto aditivo constante** implica que la diferencia $(\bar{y}_1 - \bar{y}_2)$ puede resumir el efecto en **cada individuo**



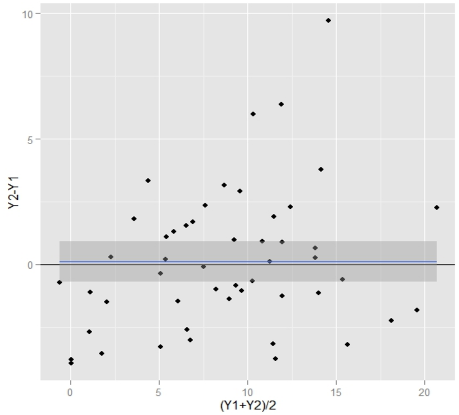
Comp. de medias apareadas. Premisa de efecto aditivo

Gráfico de Bland – Altman (BA)

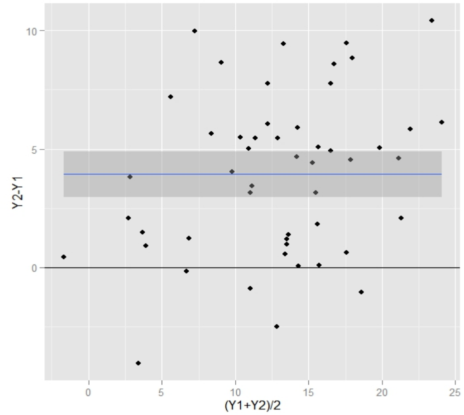
- El gráfico de BA representa para cada pareja, su diferencia (eje y) en función de su media (eje x). Se evalúa si cambia su diferencia en función de la magnitud. De esta forma, se observa si el efecto es aditivo o no.
- Para el tipo de análisis explicado, se requiere que el efecto sea aditivo, porque de lo contrario carece de sentido estudiar un efecto común para todos los rangos de valores.
- Al hacer transformación logarítmica, un efecto multiplicativo se convierte en aditivo
- **R**: *paired.plotBA* (del paquete PairedData)

Comp. de medias apareadas. Premisa de efecto aditivo

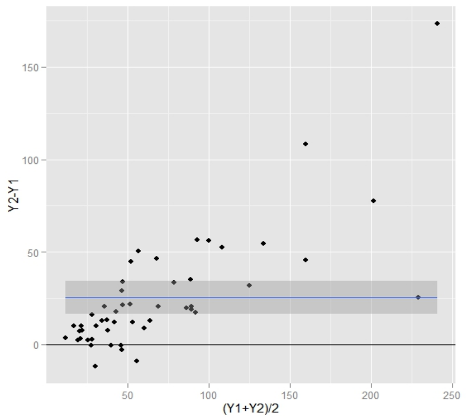
Gráfico de Bland – Altman (BA)



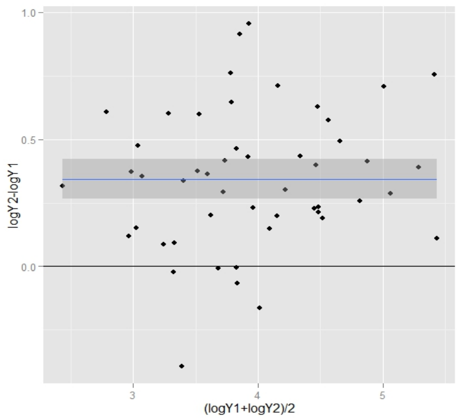
Sin efecto



Efecto aditivo



Efecto multiplicativo



Efecto aditivo después de Log

Se observan 4 situaciones:

- Ausencia de efecto: las diferencias son cero en promedio
- Efecto aditivo: existe un efecto aditivo que no depende de las magnitudes de la variable
- Efecto multiplicativo: Cuánto mayor, son los valores, mayor es el efecto
- Efecto multiplicativo corregido con logaritmos: al sacar logaritmos de las variables, el efecto se convierte en aditivo y se puede proceder con el método habitual

Comparación de variancias ($\boldsymbol{\sigma_1^2 = \sigma_2^2}$). Muestras independientes

Proceso

- Prueba formal:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 \\ H_1: \sigma_1 \neq \sigma_2 \end{cases}$$

- Bajo H_0, se tiene:

$$\hat{F} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

- Premisas: Las variables originales son normales
- Es indiferente que la varianza se ponga en el numerador si la prueba es bilateral. Para el cálculo del p-valor en una prueba bilateral siempre se debe escoger la cola más pequeña (izquierda o derecha) y multiplicarla por 2
- Si es unilateral, todo el riesgo α se acumula en el lado correspondiente al sentido de H_1
- **R:** `var.test`

Comparación de prop ($\boldsymbol{\pi_1} = \boldsymbol{\pi_2}$). Muestras independientes

Proceso

- Prueba formal:

$$\begin{cases} H_0: \pi_1 = \pi_2 \\ H_1: \pi_1 \neq \pi_2 \end{cases}$$

- Bajo H_0, se tiene:

$$\hat{z} = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2}}} \sim N(0,1) \quad \text{con} \quad P = \frac{n_1 \cdot P_1 + n_2 \cdot P_2}{n_1 + n_2} \ ; \ P_1 = \frac{X_1}{n_1} \ ; \ P_2 = \frac{X_2}{n_2}$$

- Premisas: muestras no muy pequeñas

- Existe un estadístico alternativo para la comparación de proporciones equivalente al anterior (estadístico de ji-cuadrado o prueba de Pearson) que tiene la desventaja que no permite estimar el efecto de la diferencia por intervalo de confianza.

- **R:** `prop.test`

Tipos de errores en una prueba de hipótesis

Tipo I

- En los ejemplos, se han expresado las conclusiones como “rechazo de H_0” o “no rechazo de H_0”.
- Si el objetivo es tomar una decisión, se debe tener un criterio **a priori** que fije un límite α por debajo del cual, el p-valor aporte evidencia de rechazo de H_0
- Si se repite la decisión ‘n’ veces, como en un proceso de control de calidad, α informa de una frecuencia de errores determinada:

En un 100α % de los casos que rechazamos H_0, esta es cierta.

- **Error de tipo I.** Al poner a prueba una hipótesis sobre los parámetros poblacionales, se puede cometer el error de <u>actuar como si la hipótesis nula (H_0) fuera falsa cuando no lo es</u>. La probabilidad de este error es:

α = P(“concluir H_1 siendo H_0 cierta”)

[Por diseño, esta probabilidad está fijada]

Tipos de errores en una prueba de hipótesis

Tipo II

- También se puede producir la situación complementaria.
- **Error de tipo II.** En la misma situación, se puede cometer el error de no encontrar evidencia en contra de la hipótesis nula (H_0) cuando realmente es falsa. La probabilidad de este error se expresa como:

$$\beta = P(\text{“concluir } H_0 \text{ siendo } H_1 \text{ cierta”})$$

- En general esta probabilidad no es controlable ya que su valor depende del valor real del parámetro testeado (que es desconocido).

Tipos de errores		**Decisión**	
		A_0	A_1
Realidad	H_0	**Decisión correcta**	**Error Tipo I (α)**
	H_1	**Error Tipo II (β)**	**Decisión correcta**

Árbol de decisiones y consecuencias

Tipos de errores

H_0: Un método nuevo es efectivo?

Análisis de los dos tipos de errores

SÍ es efectivo

- **Rechazar H_0** (se implanta el nuevo)
  - + Beneficios
- **No rechazar H_0** (No se implanta el nuevo)
  - - Ausencia de beneficios potenciales
  - - Costes de I+D sin retorno
  - - Perdida de competitividad
  - - Frustración por no lograr objetivos

NO es efectivo

- **Rechazar H_0** (se implanta el nuevo)
  - - Malestar de los usuarios
  - - Pérdidas a largo plazo
  - - Dilema: invertir más para resolver el problema, o parar la inversión
- **No rechazar H_0** (No se implanta el nuevo)
  - - Decepción inicial
  - + Protección derechos usuarios
  - + Oportunidades para buscar competitividad por otras vías

Realidad desconocida

Decisión

Consecuencias

MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull