Fundamentos de Estadística

Silvia N. Pérez
Especialización en Ciencia de Datos - UNO

Pruebas de comparación

POBLACIÓN DE ESTUDIO	CONDICION	DATOS CUANTITATIVOS		DATOS CUALITATIVOS
		PRUEBA PARAMETRICA	PRUEBA NO PARAMETRICA	PRUEBA NO PARAMETRICA
DOS GRUPOS	INDEPENDIENTES	T-student para muestras indepenientes	U- Mann Whitney	chi^2 (Independencia)
	APAREADOS	T-student para muestras relacionadas	Wilcoxon	Mc. Nemar (dicotómicas)
MÁS DE DOS GRUPOS	INDEPENDIENTES	ANOVA (Análisis de varianza)	Kruskal Wallis	chi^2

Análisis de Varianza (ANOVA)

Qué datos tenemos?

Quiero analizar

cuanti ~ cuali

- Un factor (tratamientos) **Cualitativa**
- Una variable de respuesta **Cuantitativa**
- Pregunta de interés:

¿Las medias de cada grupo difieren o están “afectadas” por la variable cuanti??

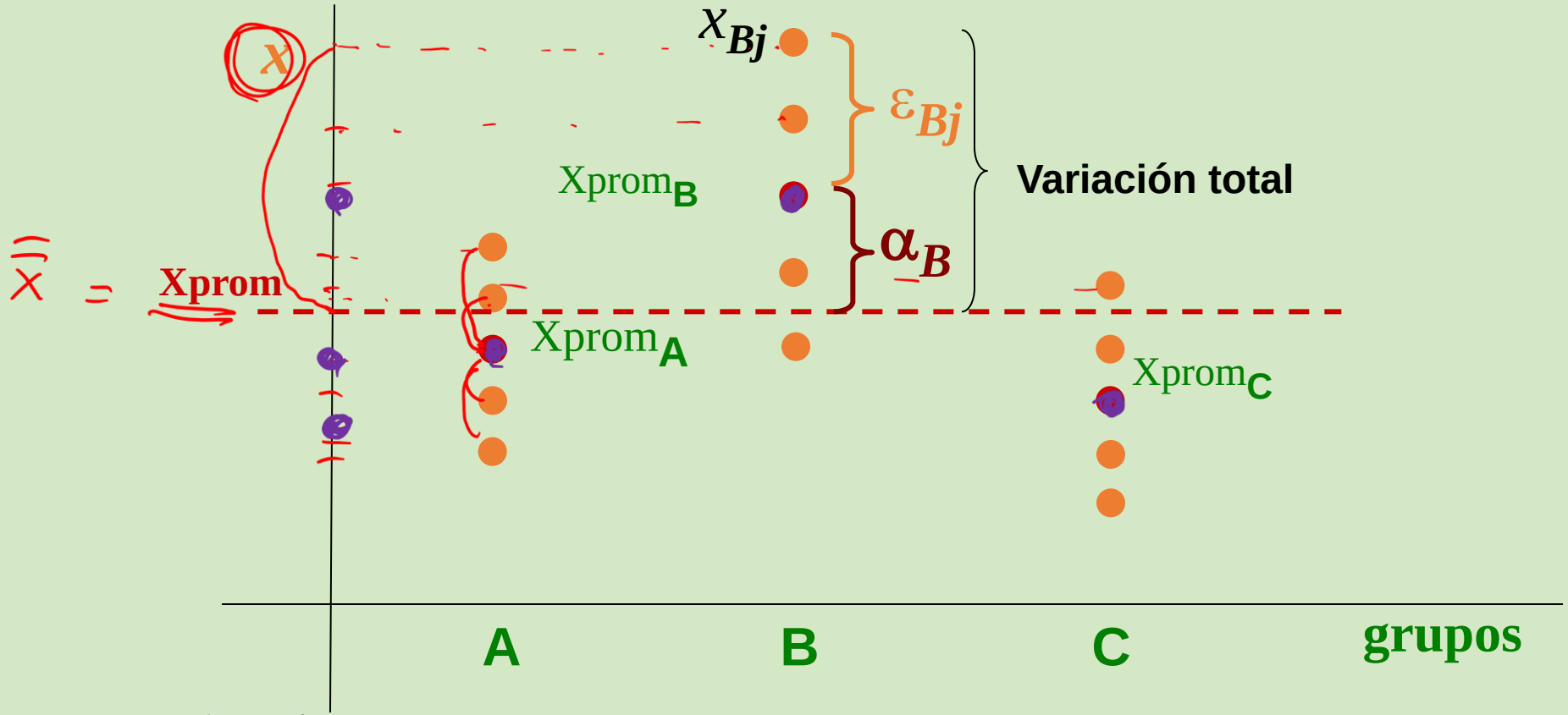
- Número de grupos:

- Si son dos grupos: utilizamos test de t
- más de 2 grupos: comparaciones multiples con ANOVA

ANOVA (ANalysis Of Variance)

Modelo lineal de efectos fijos

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$



promedio de grupo i

promedio general

$$\sum_{ij}\left(x_{ij}-\bar{\bar{x}}\right)^2=\sum_{ij}(\bar{x}_i-\bar{\bar{x}})^2+\sum_{ij}\left(x_{ij}-\bar{x}_i\right)^2$$

SCT TOTAL

dispersión s/ discriminar grupos

SCF ENTRE grupos (a)

SCE DENTRO de grupos (residual)

$$CM_{entre}=\frac{SCentre}{gl_{entre}}=\frac{\sum_{ij}(\bar{x}_{ij}-\bar{\bar{x}})^2}{a-1}$$

Mira la dispersión de los promedios de cada grupo

$$CM_{dentro\ \ o\ \ residual}=\frac{SCdentro}{gl_{dentro}}=\frac{\sum_{ij}(x_{ij}-\bar{x}_i)^2}{n-a}$$

Mira la dispersión dentro de cada grupo

F = CM_entre/ CM_dentro

tiene distribución $\mathcal{F}$ con (a-1) y (n-a) grados de libertad

Supuestos para la validez del test

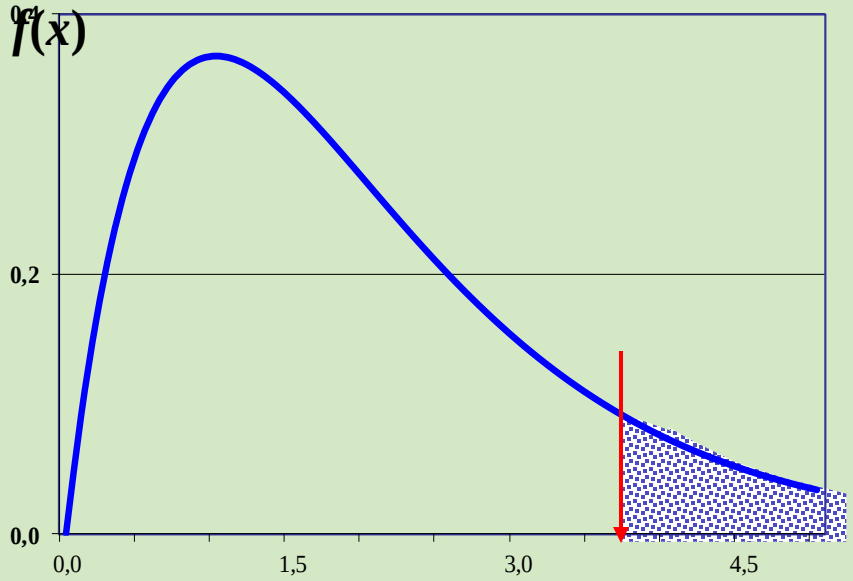
- Normalidad
- Homocedasticidad (= varianzas)
- Independencia de las observaciones

Tabla ANOVA

FUENTE DE VARIACION	SUMA DE CUADRADOS	GL	CUADRADOS MEDIOS	F
ENTRE GRUPOS	SCF entre	a-1	$\text{SC entre}/(\text{a}-1)$	$\frac{CM\ \text{entre}}{CM\ \text{dentro}}$
DENTRO DE GRUPOS	SCE dentro	n-a	$\text{SC dentro}/(\text{n}-\text{a})$	
TOTAL	SCT total	n-1		

El $F_{calculado}$ se compara con

el $F_{tabulado}$ con (a-1) y (n-a) GL



Estoes si Y a la variable cuanti

y G: variable cuali

Si G = {a o b} (2 valores posibles) ⇒ el Test t puede:

$H_0: \mu_y^a = \mu_y^b$ $H_1: \neq$ $>$ $<$

Si G = {a, b, c ...} (+2 valores) ⇒ ANOVA puede

$H_0: \mu_y^a = \mu_y^b = \mu_y^c = \dots$

H_1: "no son todos iguales"
o sea hay alguna ≠

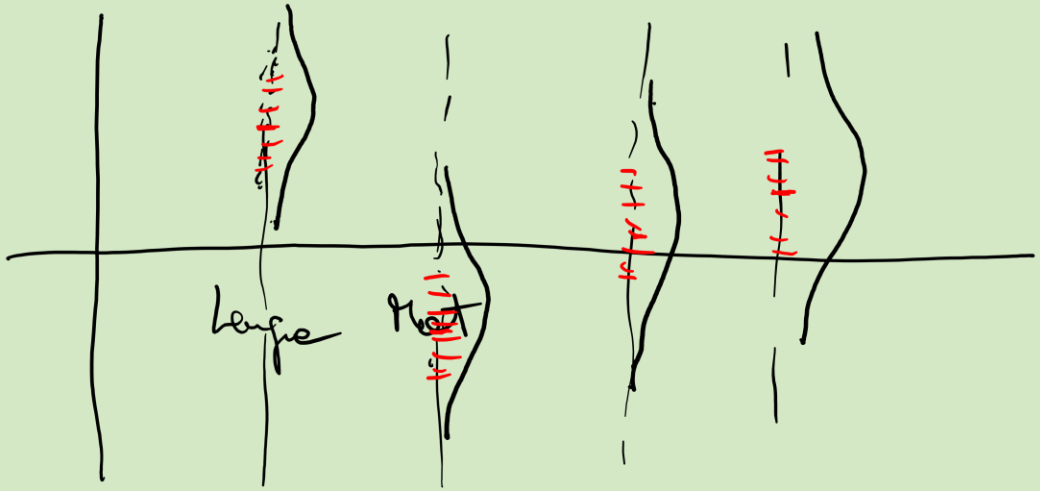
Ej: Y = forraje y G = {trat1, trat2, trat3}

$H_0: \mu_y^1 = \mu_y^2 = \mu_y^3$ H_1: alguna ≠

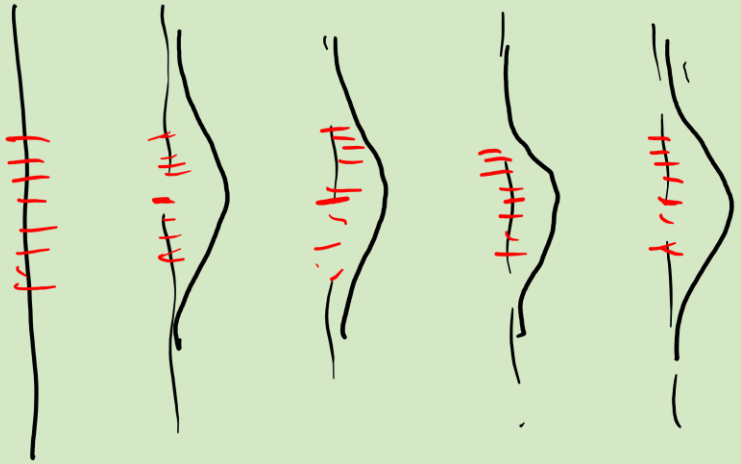
En los datos puede que tenga

mean(Y) = $\bar{X}_1 = 101.7$ $\bar{X}_3 = 100.2$

$\bar{X}_2 = 99.98$



moyens ne =



moyens sont =

Ejemplo en R

Se quiere comparar las medias de notas de estudiantes según el área de interés que declararon.
Los datos corresponden a estudiantes_completa.xlsx

POST HOC

<u>Comparaciones a posteriori:</u> Esto es, si ANOVA indica Rechazo H_0 $\Rightarrow$
$\Rightarrow$ hacemos estas pruebas p/decidir entre pares de medias si hay $\neq$

- Tuckey: si hay muchas categorías de las cualis.
- Bonferroni: Se usa si son pocas las categorías de las cualis.

Se recomienda hacer varios test a posteriori, y decidir mirando todo.

Tablas de contingencia

Tablas de contingencia

- Analiza dos variables cualitativas, X e Y, buscando asociación entre ellas.
- Permite probar la hipótesis de independencia entre ellas, esto es:

H_0: X e Y independientes

H_1: dependientes, esto es, relacionados

Tablas de contingencia

Tabla de frecuencias observadas

X // Y	y1	y2	…	Frec. Marginal de X
x1				
x2				
…			n_{ij}	
Frec. Marginal de Y				n

Se comparan las frecuencias observadas en cada casilla versus frecuencias esperadas bajo la hipótesis de independencia

Ejemplo:

Se consulta sobre acuerdo con la construcción de Metrobus y nivel de educación.
Hay relación?

Tabla de frecuencias observadas

	Muy /bastante	Poco / nada	Total
Sin estudios / primaria	1130	1123	2253
Secundaria	860	599	1459
Universitarios	480	340	820
Total	2470	2062	4532

Tablas de contingencia

frecuencias observadas (en %)

	Muy /bastante	Poco / nada	Total
Sin estudios / primaria	50.2%	49.8%	100%
Secundaria	58.9%	41.1%	100%
Universitarios	58.5%	41.5%	100%
Total	54.5%	45.5%	100%

¿Se podría hablar de que existe asociación entre nivel de estudios y acuerdo con Metrobus?

Hipótesis:

H_0: hay indep entre 'acuerdo' y 'nivel de estudios'

H_1: no indep (o sea, hay alguna relación o asociación)

Para explorar si existe asociación se compara esta tabla con una tabla en la que no existiría asociación o también llamada tabla de frecuencias esperadas (suponiendo independencia)

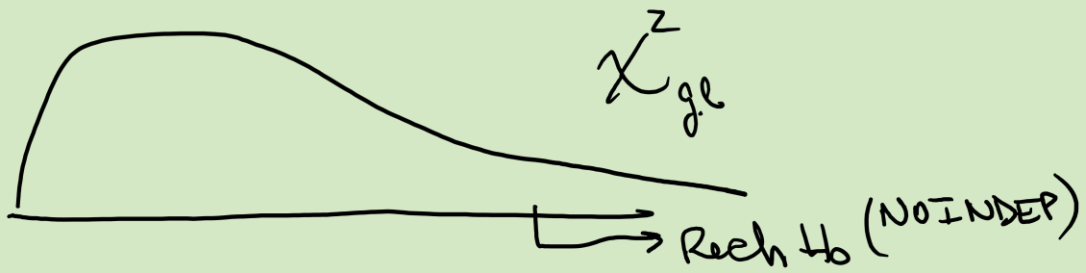
Tablas de contingencia

Las hipótesis

- H0: independencia (no asociación)
- H1: existe asociación

Estadístico de prueba para comprobar la hipótesis:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad \sim \chi^2_{(r-1)(s-1)} \text{ si } H_0 \text{ cierta}$$



Tablas de contingencia

Estadístico de prueba:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad \sim \chi^2_{(r-1)(s-1)} \text{ si } H_0 \text{ cierta}$$

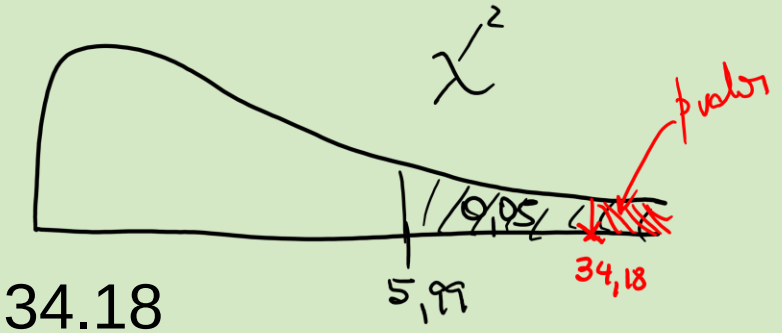
Con nuestros datos, se calculan las frec esperadas:

OBSERVADAS	Muy /bastante	Poco / nada	Total
Sin estudios / primaria	1130	1123	2253
Secundaria	860	599	1459
Universitarios	480	340	820
Total	2470	2062	4532

ESPERADAS	Muy /bastante	Poco / nada	Total
Sin estudios / primaria	1127.9	1025.1	2253
Secundaria	795.2	663.8	1459
Universitarios	446.9	373.1	820
Total	2470	2062	4532

Se calcula el valor del estadístico de prueba, que aquí da **34.18**

Tablas de contingencia



- el valor obtenido en la prueba de chi cuadrado es 34.18
- es mayor que el valor de chi-cuadrado obtenido en la tabla para un nivel de significación de 0.05 con dos grados de libertad (5.99).
- Es decir, 34.18 pertenece a la Región de Rechazo,
- $\chi 2$ observado > 5.99 por lo que rechazamos la H0 (hipótesis nula)

rechazo la independencia

- **Esto permite decir que existe asociación entre nivel educativo y acuerdo con Metrobus,** con un nivel de significación del 5%.

Tablas de contingencia

¡Ojo! Para poder realizar la prueba ji-cuadrado se debe cumplir un supuesto:

Como máximo un 20% de las celdas debe tener una frecuencia esperada menor a 5.
(no debe haber demasiadas celdas con pocos casos).

Medidas de asociación

Observación:
-La prueba de contingencia no indica la dirección de la asociación
-No informa sobre el grado de asociación

Coeficiente de Contingencia

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

V de Cramer

$$V=\sqrt{\frac{\chi^2}{n\cdot t}}$$

- t =min{#filas-1, #col-1}
- V de Cramer entre 0 y 1.

<0.05: asociación débil
0.05-0.25 moderado
0.25-1 fuerte

Medidas de asociación

χ^2= 34.18
n= 4532

En los datos anteriores:

Coeficiente de Contingencia = **0.08652**

V de Cramer = **0.08684**

CONCLUSIÓN: La asociación detectada entre nivel de estudios y acuerdo, aunque significativa, **no es fuerte** (indices cercanos a 0).

Ejemplo en R

Se quiere decidir si hay independencia entre el sexo del estudiante y el interés en un área en particular.
Los datos corresponden a estudiantes_completa.xlsx