

A close-up photograph of a blue coin-operated binocular viewer. The viewer has two eyepieces at the top, a coin slot in the middle, and a small instruction panel at the bottom. The background is a blurred city skyline at night with warm, bokeh lights.

Estadística II

Maestría en Generación y Análisis de Información Estadística
Universidad Nacional de Tres de Febrero

A dark blue, irregular ink splash or blotch serves as the background for the text. The splash has a textured, watercolor-like appearance with some lighter blue and white areas around the edges. The text is centered within the dark blue area.

Análisis de la varianza

Test para comparación de dos medias normales

x_{11}, \dots, x_{1n} Muestra de una población normal con media μ_1 y desvío σ_1

x_{21}, \dots, x_{2m} Muestra de una población normal con media μ_2 y desvío σ_2

$$H_a: \mu_1 - \mu_2 \neq 0$$

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

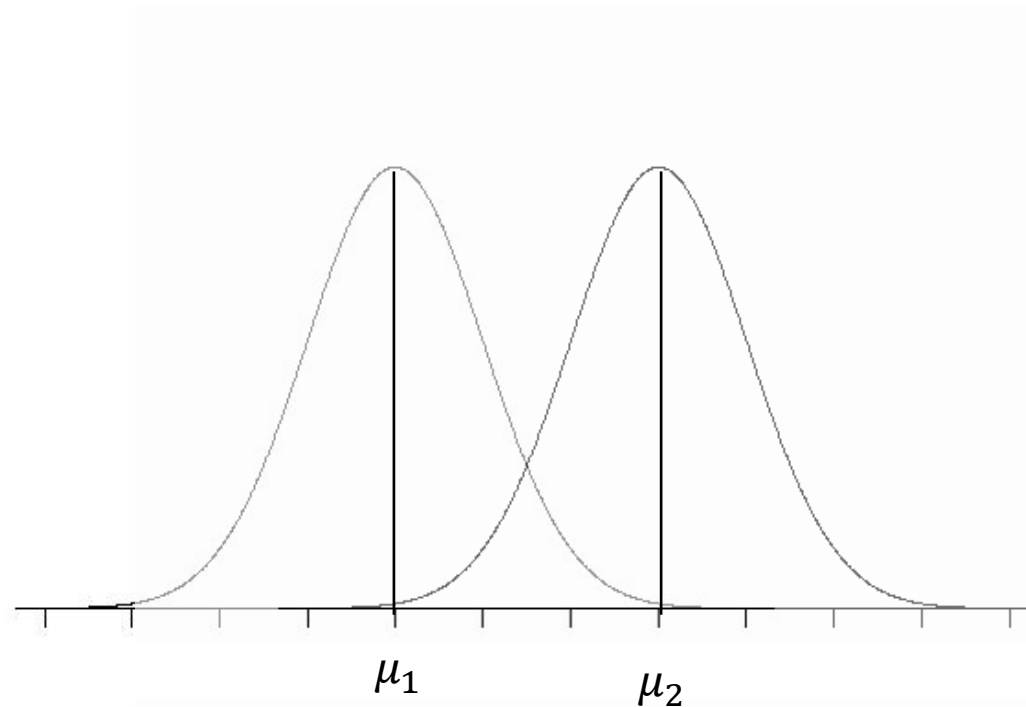
$$H_a: \mu_1 - \mu_2 < 0$$

Test de t para comparación de dos medias normales

Muestras independientes

Sigmas desconocidos
e iguales

$$E = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$



CONCEPTO



Hasta ahora hemos visto como podemos resolver respecto a la veracidad de una afirmación para una o dos poblaciones pero ¿qué pasa si tenemos tres o más?



Tenemos que usar el análisis de varianza también llamado ANOVA

Análisis de la varianza

El análisis de varianza no constituye un método o procedimiento único



Según los diseños y datos disponibles existen diversos modelos de análisis de varianza.

Análisis de la varianza
de un factor

Análisis de la varianza
de varios factores

Diseños factoriales

Análisis de la varianza
de medidas repetidas

Análisis de la varianza



Por qué utilizamos el análisis de varianza en vez de la t de Student



¿No se podrían comparar todos los grupos de dos en dos con la t de Student?



Al hacer muchas comparaciones de dos en dos, aumenta la probabilidad de que algunas diferencias resulten significativas por azar



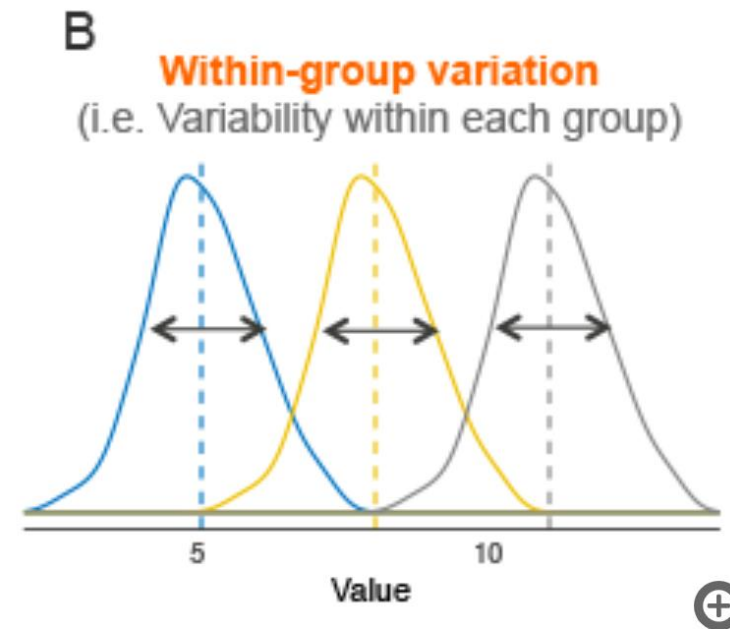
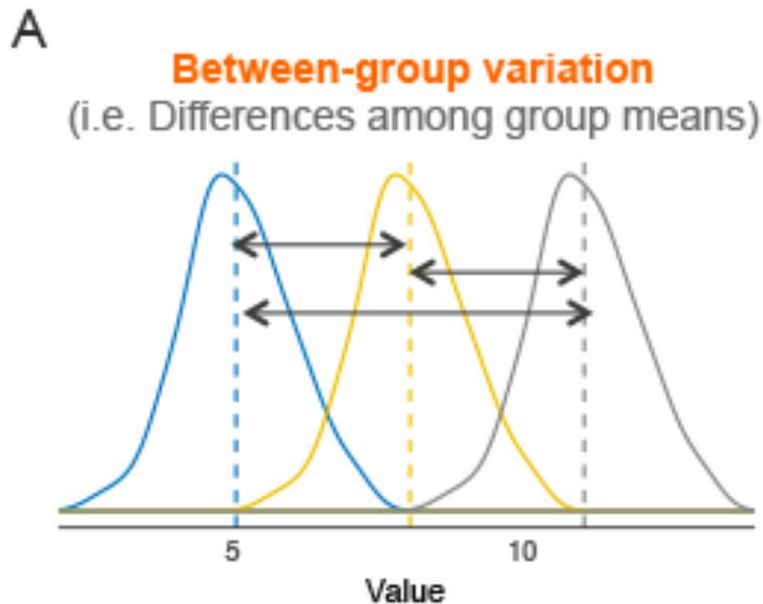
Si por ejemplo tenemos tres grupos podríamos hacer tres comparaciones: entre el 1º y el 2º, entre el 1º y el 3º y entre el 2º y el 3º. Operando con un nivel de confianza de $\alpha = .05$, la probabilidad de encontrar al menos una diferencia significativa por azar es de hecho del 14.26% y no del 5%



Una prueba estadística basada en todos los datos utilizados simultáneamente, es más estable que la prueba o análisis que parcializa los datos y no los examina todos juntos.

Concepto

- El concepto básico del ANOVA es muy SIMPLE: compara la varianza que hay entre todas las unidades dentro de cada grupo con la que hay entre el promedio de los grupos.
- Si el primero es mayor entonces la variación entre los grupos o muestras no representa una variación real.



SUPUESTOS

El ANOVA es una técnica de prueba de hipótesis

Se requiere que los datos tengan una distribución NORMAL

Asimismo, se supone que las poblaciones que proveen las muestras tienen varianzas iguales.

La tercera suposición es que las muestras son Independientes.

Análisis de la varianza

Se seleccionan muestras aleatorias de tamaño n de cada una de las k poblaciones

Se supone que las poblaciones son normalmente distribuidas e independientes
Con medias $\mu_1, \mu_2, \dots, \mu_k$, y varianza común σ^2 .

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : *Al menos dos de las medias no son iguales*

Análisis de la varianza

Tratamiento		1	2	...	i	...	k
		y_{11}	y_{21}	...	y_{i1}	...	y_{k1}
		y_{12}	y_{22}	...	y_{i2}	...	y_{k2}
		\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
		y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}
Total	$Y_{..}$	$Y_{1.}$	$Y_{2.}$...	$Y_{i.}$...	$Y_{k.}$
Media	$\bar{Y}_{..}$	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$...	$\bar{Y}_{i.}$...	$\bar{Y}_{k.}$

Análisis de la varianza

$$y_{ij} = \mu_i + \epsilon_{ij}$$

O bien

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Donde μ es la media general,

$$\mu_i = \mu + \alpha_i \quad \sum_{i=1}^k \alpha_i = 0$$

Análisis de la varianza

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

H_1 : Al menos una de las α_i no es igual a cero

$$\sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2$$

$$SST = SSEntre + SSIntra$$

Análisis de la varianza

- La variación TOTAL, SST es la que toma en cuenta la variación entre TODAS las unidades tomando en cuenta la diferencia a la gran media
- La varianza INTRA GRUPOS, SSIntra considera la variación que hay dentro de cada grupo
- La Varianza ENTRE GRUPOS, SSEntre compara las medias de cada Grupo con la gran Media

Tabla de ANOVA



Los datos de las varianzas se resumen en lo que se llama “LA TABLA DE ANÁLISIS DE VARIANZA”



Que reúne los valores y los llamados grados de libertad.

Tabla de ANOVA

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados medios	F
Entre Grupos	K-1	SCEntre	CME= SCE/(k-1)	CME/CMI
Intra Grupos	N-K	SCIntra	CMI= SCI/(n-k)	
TOTAL	N -1	SCT		

La distribución “F de Snedecor”

La distribución de F es aquella que se usa para estimar cualquier cociente de Varianzas.

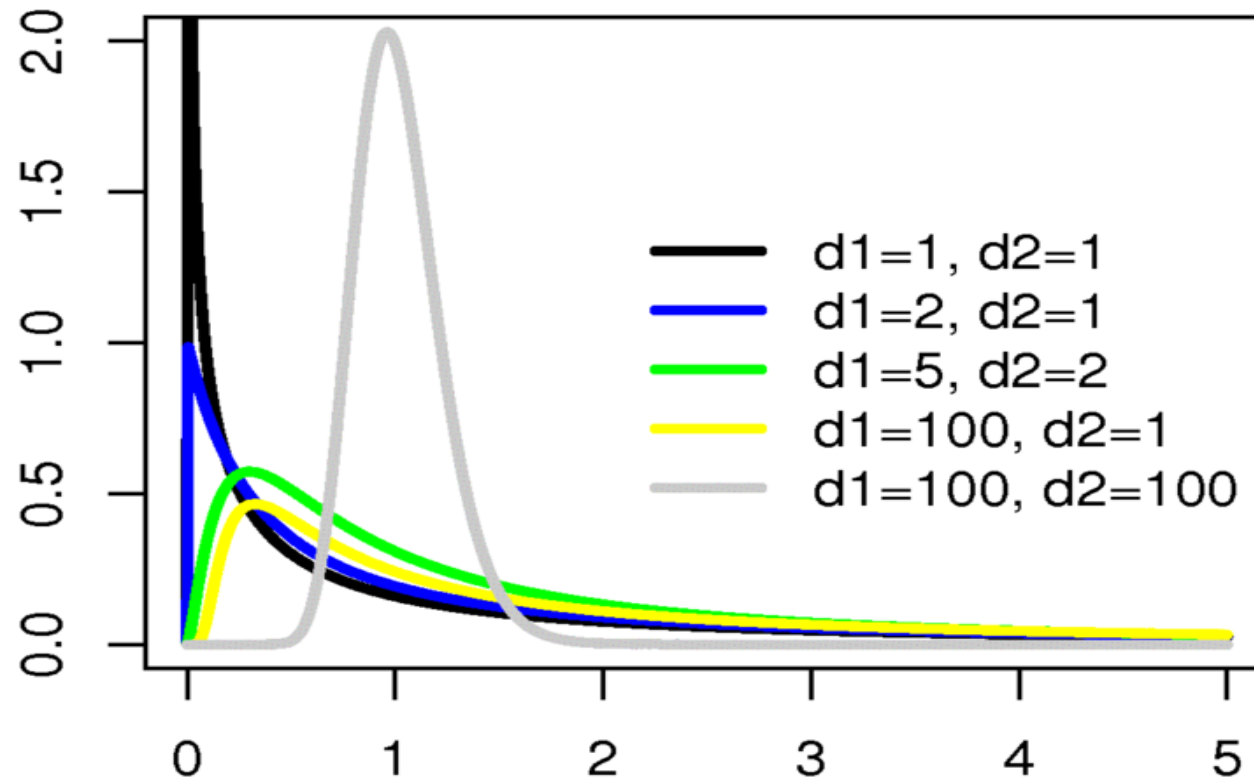


Al igual que la T es una familia de Curvas cuya curva exacta a usar esta determinada por dos grados de libertad.

Grados de libertad del numerador

Grados de libertad del denominador

Familia de distribuciones F



ANOVA como modelo lineal



Tratamiento		1	2	...	i	...	k
		y_{11}	y_{21}	...	y_{i1}	...	y_{k1}
		y_{12}	y_{22}	...	y_{i2}	...	y_{k2}
		\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
		y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}
Total	$Y_{..}$	$Y_{1.}$	$Y_{2.}$...	$Y_{i.}$...	$Y_{k.}$
Media	$\bar{Y}_{..}$	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$...	$\bar{Y}_{i.}$...	$\bar{Y}_{k.}$

Análisis de la varianza

$$y_{ij} = \mu_i + \epsilon_{ij}$$

O bien

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Donde μ es la media general,

$$\mu_i = \mu + \alpha_i \quad \sum_{i=1}^k \alpha_i = 0$$

Análisis de la varianza

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

H_1 : Al menos una de las α_i no es igual a cero

$$\sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2$$

$$SST = SSEntre + SSIntra$$

Modelo Lineal

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

Descomposición de y_{ij}

$$y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}) + (y_{ij} - \bar{Y}_i)$$

Suma de cuadrados

$$\sum y_{ij}^2 = n\bar{Y}^2 + \sum n_i(\bar{Y}_i - \bar{Y})^2 + \sum \sum (y_{ij} - \bar{Y}_i)^2$$

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

```
one.way <- aov(PrimerNoche ~ fedad, data = Insomnio_2)
summary(one.way)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## fedad         2   69.3    34.66    8.588 0.000371 ***
## Residuals    96  387.5     4.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_apa(one.way)
```

```
##           Effect
## 1 (Intercept) F(1, 96) = 1124.93, p < .001, petasq = .92 ***
## 2          fedad F(2, 96) =    8.59, p < .001, petasq = .15 ***
```

Test Post hoc en ANOVA

Anova

- Anova no dice lo que realmente necesitamos
 - Estamos interesados en diferencias específicas y no en rechazar una hipótesis general
- Por lo tanto, es de poca utilidad si no seguimos indagando

La situación

- ANOVA
- Qué nos dice?
 - Los grupos tienen medias diferentes
 - Cuáles grupos?
 - No se sabe
- Qué se debe hacer para conocer cuáles difieren?
 - Comparaciones múltiples

Problema

- Hacer múltiples tests del mismo tipo incrementa el error de tipo I
- Ejemplo: 4 grupos
 - 6 comparaciones posibles
 - $\alpha = 1 - (1-.05)^6 = .265$!!!!

Situación general

Estadístico general

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SE_1}{n_1} + \frac{SE_2}{n_2}}}$$

Estadístico para muestras de igual tamaño y varianzas iguales

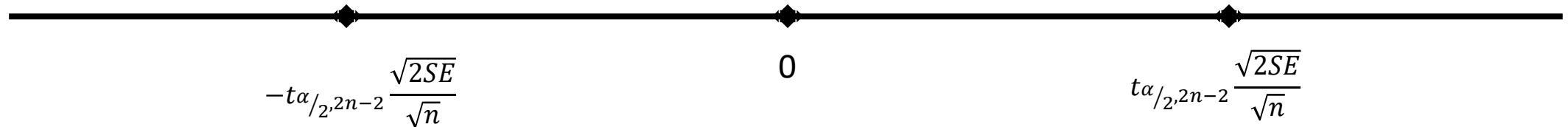
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2SE}{n}}}$$

Intervalo de Confianza

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, 2n-2} \frac{\sqrt{2SE}}{\sqrt{n}}$$

Situación general

- De acuerdo donde caiga $\bar{X}_1 - \bar{X}_2$ se rechaza o no la hipótesis de igualdad de medias



Elementos que influyen en la comparación

- La distribución del estadístico
- El nivel de significación fijado
- La variabilidad de cada grupo
- El tamaño muestral de cada grupo

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, 2n-2} \frac{\sqrt{2SE}}{\sqrt{n}}$$

Posibilidades

- Vieja escuela
 - Mínima diferencia significativa (DMS)
 - Bonferroni
 - Sidak
- Otras opciones standard
 - Tukey
 - Student Newman-Keuls
 - Ryan
 - Scheffe
- Situaciones especiales
 - Violación de la hipótesis de homocedasticidad
 - Grupos de tamaños desiguales
- Procedimientos de a pasos
 - Holm's
 - Hochberg
- Nuevos enfoques: FDR, ICI

Posibilidades

- Vieja escuela
 - Mínima diferencia significativa (LSD): Es el clásico uso del t-test. Recomendado para hasta tres grupos

$$LSD = t_{\alpha/2, n-2} \frac{\sqrt{2SE}}{\sqrt{n}}$$

- Se rechaza cuando la diferencia de medias es mayor a LSD en valor absoluto

Bonferroni and Sidak test

- Bonferroni
 - Usa $\alpha' = \alpha/c$ donde c es el número de comparaciones a realizar
- Sidak
 - Utiliza $\alpha' = 1 - (1 - \alpha)^{1/c}$
 - Ejemplo 3 comparaciones
 - Bonferroni $\alpha = .05/3 = .0167$
 - Sidak $\alpha = 1 - (1 - .05)^{1/3} = .0170$

Sidak es menos estricto

Posibilidades

- Otras opciones basadas en el rango estudentizado
 - Student Newman-Keuls
 - Tukey

Rango Studentizado

$$Q = \frac{\max(\bar{X}_i - \mu_i) - \min(\bar{X}_j - \mu_j)}{\sqrt{\frac{MS_{error}}{n}}} \sim q_{t,n-t}$$

$$HSD = q_{t,n-t}(\alpha) \frac{\sqrt{MS_{error}}}{\sqrt{n}}$$

Tests para situaciones específicas

- Heterocedasticidad
 - Games-Howell
- n desiguales

- $\bar{n}_h = \frac{k}{\sum \frac{1}{n_i}}$

Cuál usar?

- Se cumplen los supuestos:
 - Tukey's
 - REWQ
- N desiguales:
 - Gabriel
 - Hochberg
- Varianzas desiguales:
 - Games-Howell
- Se compara contra un grupo control
 - Dunnett

ANOVA DE DOS FACTORES




Tratamiento		1	2	...	i	...	k
		y_{11}	y_{21}	...	y_{i1}	...	y_{k1}
		y_{12}	y_{22}	...	y_{i2}	...	y_{k2}
		\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
		y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}
Total	$Y_{..}$	$Y_{1.}$	$Y_{2.}$...	$Y_{i.}$...	$Y_{k.}$
Media	$\bar{Y}_{..}$	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$...	$\bar{Y}_{i.}$...	$\bar{Y}_{k.}$

Anova de un factor

Modelo Lineal

Media general

Efecto del factor

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


Descomposición de y_{ij}

$$y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}) + (y_{ij} - \bar{Y}_i)$$

Suma de cuadrados

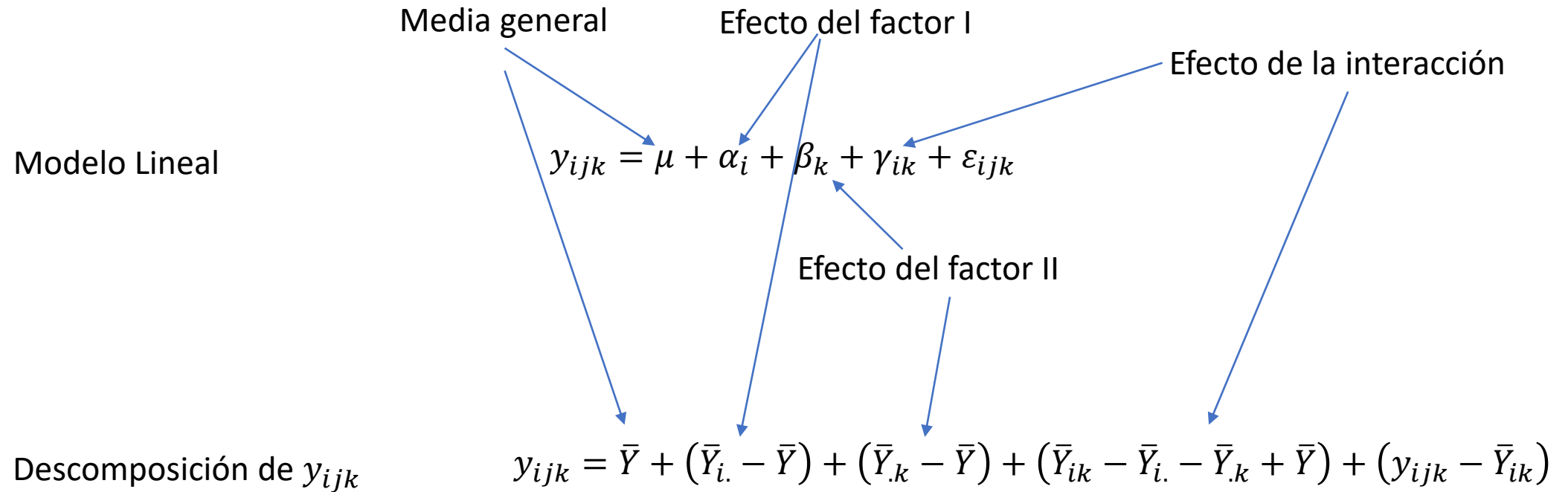
$$\sum y_{ij}^2 = n\bar{Y}^2 + \sum n_i(\bar{Y}_i - \bar{Y})^2 + \sum \sum (y_{ij} - \bar{Y}_i)^2$$

Anova de dos factores

	Menor que 20	Entre 20 y 25	Mayor que 25	
Varón	y_{111} \vdots y_{11n_1}	y_{121} \vdots y_{12n_2}	y_{131} \vdots y_{13n_3}	$\bar{Y}_{1..}$
Mujer	y_{211} \vdots y_{21n_4}	y_{221} \vdots y_{22n_5}	y_{231} \vdots y_{23n_6}	$\bar{Y}_{2..}$
	$\bar{Y}_{.1.}$	$\bar{Y}_{.2.}$	$\bar{Y}_{.3.}$	$\bar{Y}_{...}$

$\bar{Y}_{ij.}$ Media de cada subgrupo

Anova de dos factores



$$\sum y_{ijk}^2 = n\bar{Y}^2 + \sum n_i(\bar{Y}_{i.} - \bar{Y})^2 + \sum n_k(\bar{Y}_{.k} - \bar{Y})^2 + \sum_i \sum_k n_{ik} (\bar{Y}_{ik} - \bar{Y}_{i.} - \bar{Y}_{.k} + \bar{Y})^2 + \sum \sum (y_{ijk} - \bar{Y}_{ik})^2$$

ANOVA de dos factores

Test de hipótesis **paramétrico** para evaluar hipótesis sobre el valor de medias poblacionales (**parámetro**) de varios grupos (muestras **independientes**), controlando por otra variable

SUPUESTOS:

- Variable cuantitativa
- Muestreos probabilísticos
- Grupos donde la variable cuantitativa distribuye normal y con igual varianza (homocedasticidad) en cada grupo

ANOVA de dos factores: test de hipótesis

- | | |
|--------|---|
| Test 1 | <ul style="list-style-type: none">• H_0: Las medias son iguales en todas las categorías del factor 1 (controlando por factor 2) \leftrightarrow controlando por f2, NO hay asociación entre f1 y la VD• H_1: Las medias de al menos dos de las categoría del f1 son distintas (controlando por factor 2) \leftrightarrow controlando por f2, hay asociación entre f1 y la VD |
| Test 2 | <ul style="list-style-type: none">• H_0: Las medias son iguales en todas las categorías del factor 2 (controlando por factor 1) \leftrightarrow controlando por f1, NO hay asociación entre f2 y la VD• H_1: Las medias de al menos dos de las categoría del f2 son distintas (controlando por factor 1) \leftrightarrow controlando por f1, hay asociación entre f2 y la VD |
| Test 3 | <ul style="list-style-type: none">• H_0: no hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es igual en las categorías de otro factor• H_1: hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es distinta en las categorías de otro factor |

ANOVA de dos factores: Estadístico del Test y distribución nula

Medias de cuadrados: Suma de cuadrado/gl

- *Media de cuadrados totales (MCT): $\frac{SCT}{n-1}$*
- *Media de cuadrados explicados por f1 (MCE1): $\frac{SCE1}{I-1}$* I numero de grupos en f1
- *Media de cuadrados explicados por f2 (MCE2): $\frac{SCE2}{K-1}$* K numero de grupos en f2
- *Media de cuadrados explicados por interacción f1 y f2 (MCE12): $\frac{SCE12}{(I-1)*(K-1)}$*
- *Media de cuadrados residual (MCR): $\frac{SCR}{(n-I*K)}$*

ANOVA de dos factores: Estadístico del Test y distribución nula

- Test 1
- H_0 : Las medias son iguales en todas las categorías del factor 1 (controlando por factor 2) \leftrightarrow controlando por f2, NO hay asociación entre f1 y la VD
 - H_1 : Las medias de al menos dos de las categorías del f1 son distintas (controlando por factor 2) \leftrightarrow controlando por f2, hay asociación entre f1 y la VD

$$\text{Estadístico } F = \frac{MCE1}{MCR} \sim F_{I-1, n-1K} gl$$

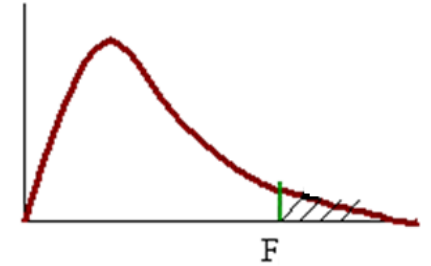
- Test 2
- H_0 : Las medias son iguales en todas las categorías del factor 2 (controlando por factor 1) \leftrightarrow controlando por f1, NO hay asociación entre f2 y la VD
 - H_1 : Las medias de al menos dos de las categorías del f2 son distintas (controlando por factor 1) \leftrightarrow controlando por f1, hay asociación entre f2 y la VD

$$\text{Estadístico } F = \frac{MCE2}{MCR} \sim F_{K-1, n-1K} gl$$

- Test 3
- H_0 : no hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es igual en las categorías de otro factor
 - H_1 : hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es distinta en las categorías de otro factor

$$\text{Estadístico } F = \frac{MCE12}{MCR} \sim F_{(I-1)(K-1), n-1K} gl$$

ANOVA de dos factores: Valor P



- Test 1
 - H_0 : Las medias son iguales en todas las categorías del factor 1 (controlando por factor 2) \leftrightarrow controlando por f2, NO hay asociación entre f1 y la VD
 - H_1 : Las medias de al menos dos de las categoría del f1 son distintas (controlando por factor 2) \leftrightarrow controlando por f2, hay asociación entre f1 y la VD
- Test 2
 - H_0 : Las medias son iguales en todas las categorías del factor 2 (controlando por factor 1) \leftrightarrow controlando por f1, NO hay asociación entre f2 y la VD
 - H_1 : Las medias de al menos dos de las categoría del f2 son distintas (controlando por factor 1) \leftrightarrow controlando por f1, hay asociación entre f2 y la VD
- Test 3
 - H_0 : no hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es igual en las categorías de otro factor
 - H_1 : hay interacción entre f1 y f2 \leftrightarrow la diferencia entre las medias de un factor, es distinta en las categorías de otro factor

$$P(F_{I-1, n-1K \text{ gl}} > F_{obs} = \frac{MCE1}{MCR})$$

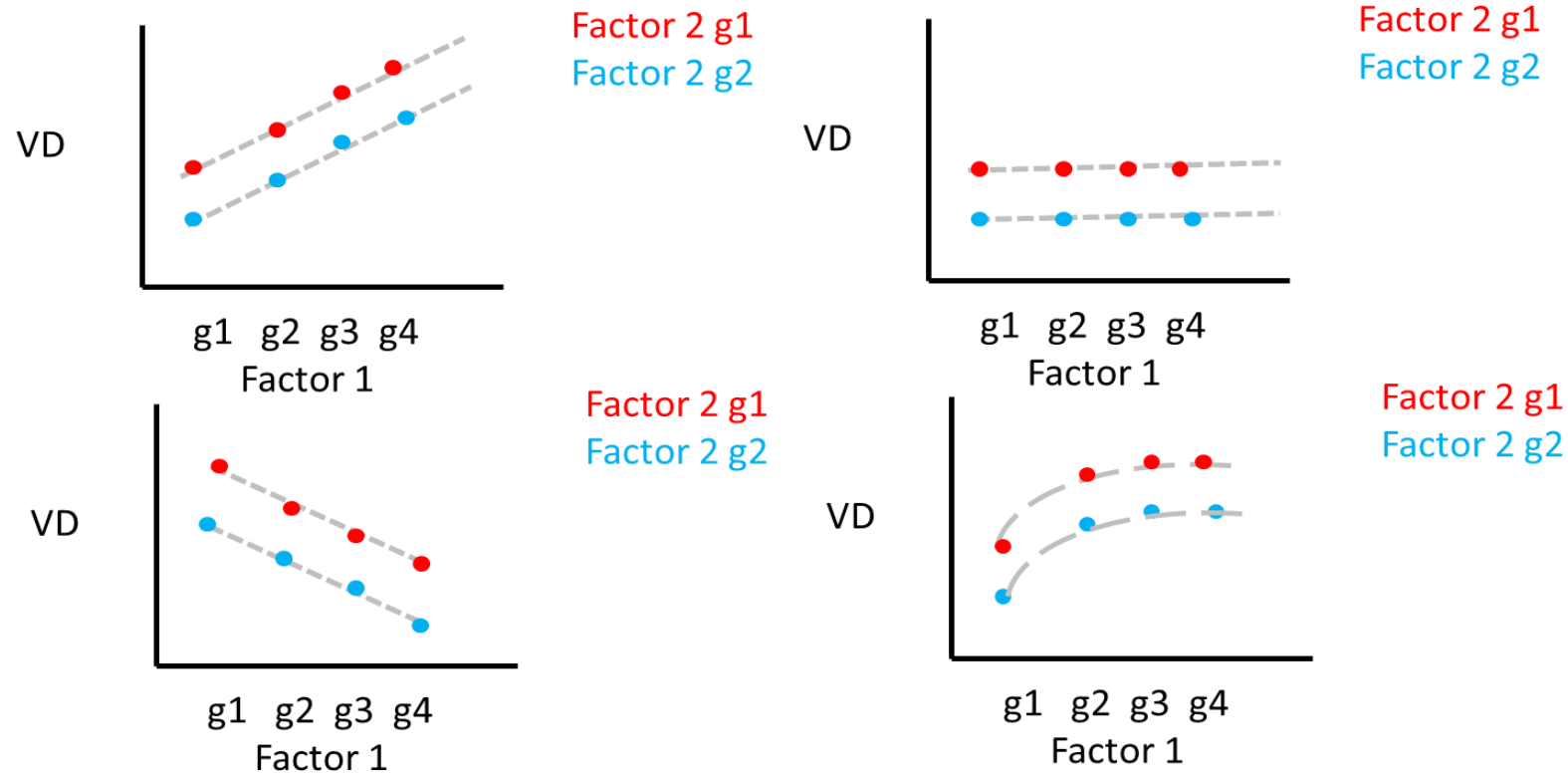
$$P(F_{K-1, n-1K \text{ gl}} > F_{obs} = \frac{MCE2}{MCR})$$

obs

$$P(F_{(I-1)(K-1), n-1K \text{ gl}} > F_{obs} = \frac{MCE12}{MCR})$$

ANOVA de dos factores: Ausencia de Interacción

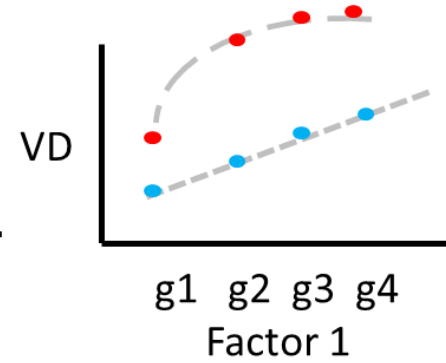
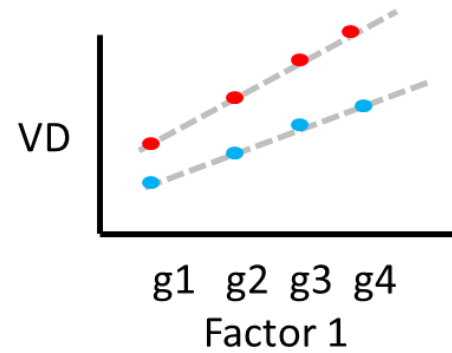
Estima cuánta de la variabilidad observada en los datos (VD) puede ser explicada por cada factor (VD) y la interacción de ambos.



ANOVA de dos factores: Interacción

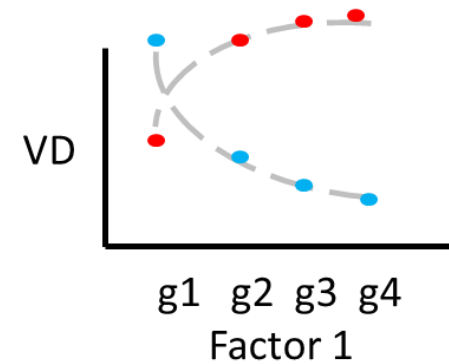
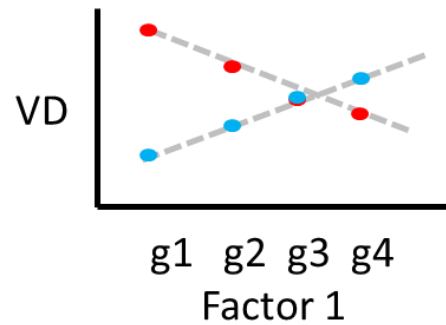
Con interacción

Factor 2 g1
Factor 2 g2



Factor 2 g1
Factor 2 g2

Factor 2 g1
Factor 2 g2



Factor 2 g1
Factor 2 g2

The background of the slide features a collection of light-colored wooden rings and stars scattered across a matching light surface. The rings are circular with a central hole, and the stars are five-pointed. Some objects are in sharp focus, while others are blurred in the background, creating a sense of depth. The overall lighting is soft and even.

Tamaño del efecto en Anova

Particiones de la varianza

- Anova de un factor
- $SST = SSEntre + SSIntra$
- Anova de dos factores
- $SST = SSfactor1 + SSfactor2 + SSinteracción + SSerror$

Tamaño del efecto

- Eta cuadrado

$$\eta^2 = \frac{SS_{efecto}}{SS_{total}}$$

- Eta cuadrado parcial

$$\eta_p^2 = \frac{SS_{efecto}}{SS_{efecto} + SS_{error}}$$

- Eta cuadrado generalizado

Tamaño del efecto

- Eta cuadrado

$$\text{Tamaño del efecto} = \begin{cases} \leq 0.04 & \text{débil} \\ 0.04 - 0.34 & \text{moderado} \\ > 0.34 & \text{fuerte} \end{cases}$$

- y el Eta cuadrado parcial

$$\text{Tamaño del efecto} = \begin{cases} \leq 0.02 & \text{débil} \\ 0.02 - 0.09 & \text{moderado} \\ > 0.09 & \text{fuerte} \end{cases}$$