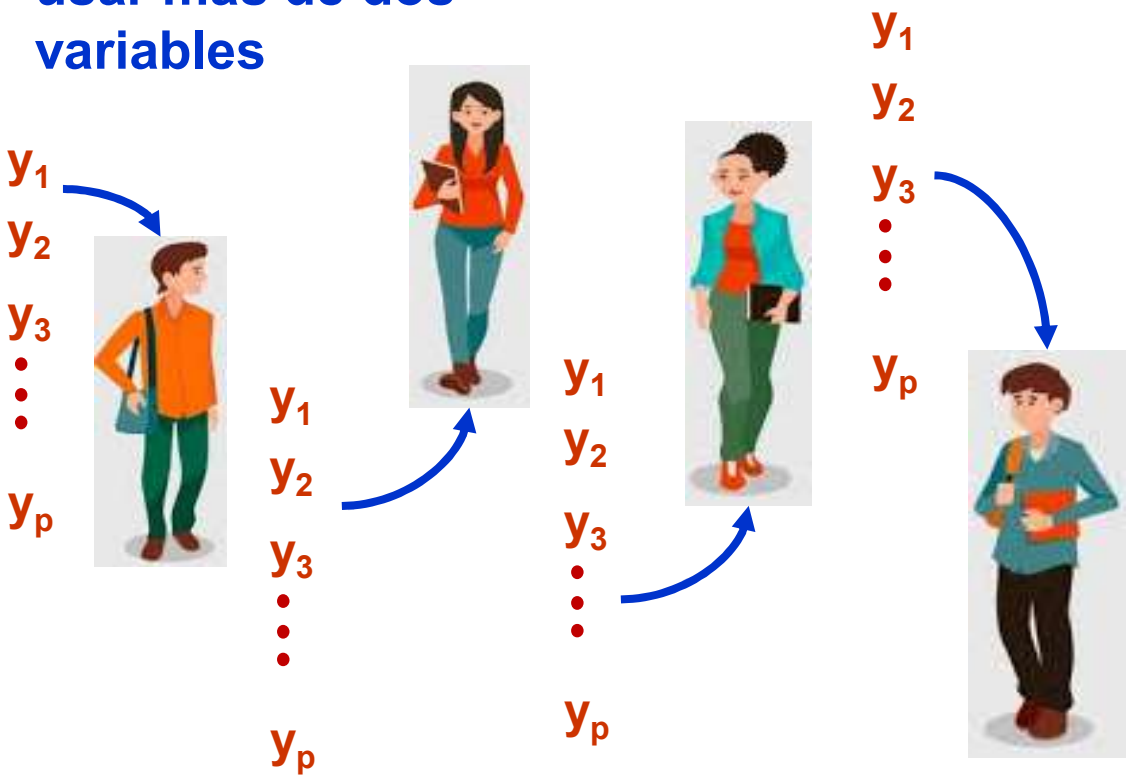
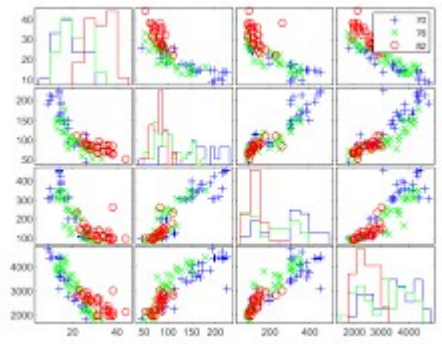
ANÁLISIS DE VARIANZA MULTIVARIADO (MANOVA)

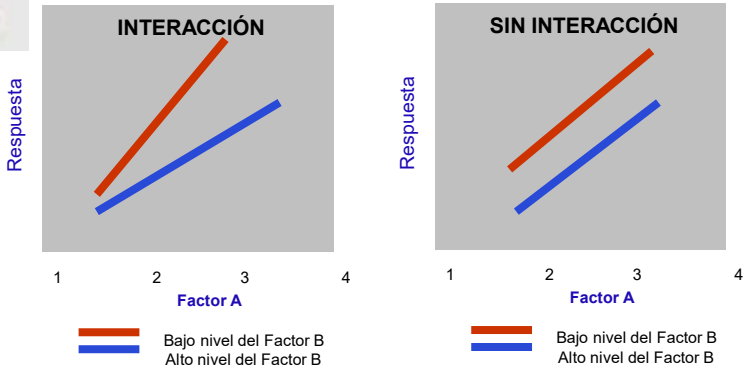
Razones para usar más de dos variables





SEXO: Masculino, Femenino

TRATAMIENTO: Sí, No

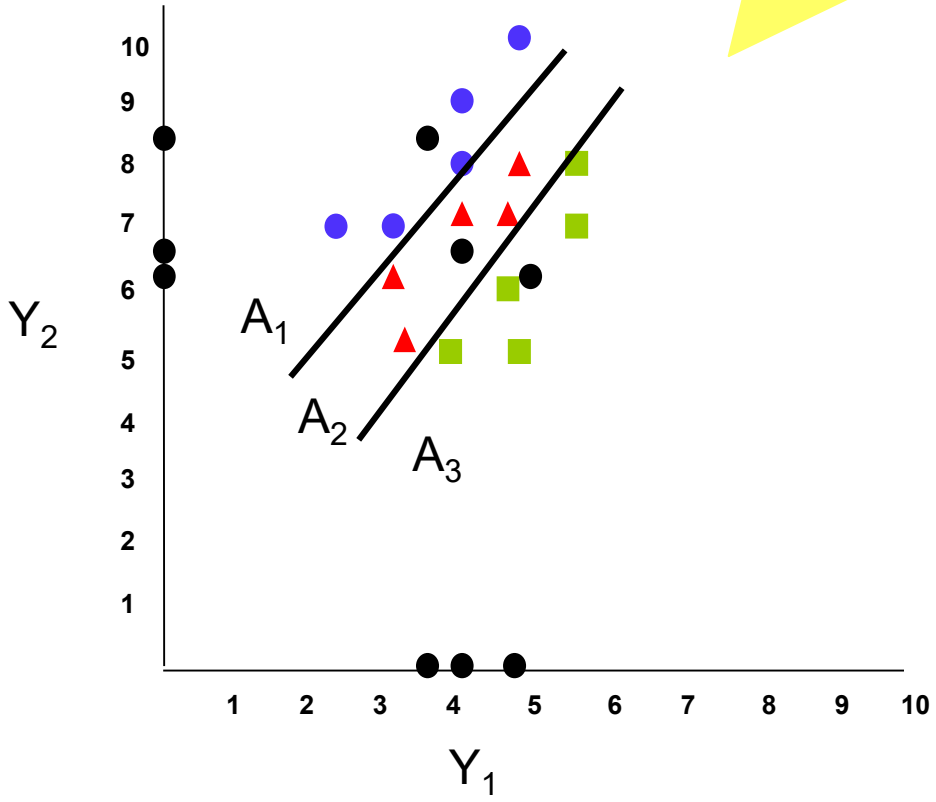


MANOVA es significativa al 0.001
Ninguna prueba univariada es significativa al 0.05

	Grupo I	Grupo II	Grupo III
Y_1	4.6	5.0	6.2
Y_2	8.2	6.6	6.2

KERLINGER & PEDHAZUR (1973) Multiple regression in behavioral research. New York. Holt, Renehart & Winston

Distancia de las medias entre grupos es mayor que en los ejes



A_1		A_2		A_3	
1	2	1	2	1	2
3	7	4	5	5	5
4	7	4	6	6	5
5	8	5	7	6	6
5	9	6	7	7	7
6	10	6	8	7	8

Las pequeñas diferencias no detectables en cada variable combinadas producen una diferencia acumulada significativa cuando las variables se analizan en forma conjunta

CASOS: Cuestionamientos que pueden ser analizados por MANOVA

¿Los estudiantes a quienes se ha asignado aleatoriamente en cada uno de los tres centros de preparación pre-universitaria tienen diferentes rendimientos en aptitud matemática, verbal y analítica?

¿Existen diferencias entre los escolares provenientes de los centros de educación pública y privada en cuanto al uso del lenguaje en términos de expresión oral, expresión escrita y grado de elaboración?

Un investigador está interesado en el efecto de los tratamientos asignados aleatoriamente sobre diferentes tipos de ansiedad: prueba de ansiedad, ansiedad en reacción al estrés y ansiedad inercial. El investigador también está interesado si existen diferencias por sexo, incorporando en tal sentido el efecto factorial.

Sobre la base del rol de los sexos se ha clasificado a las personas como masculino fuerte y débil y femenino fuerte y débil. Un investigador desea saber si los cuatro grupos (masculino y femenino cruzados factorialmente) difieren colectivamente en término de autoestima, introversión-extroversión y neuroticismo.

CUATRO RAZONES CON SUSTENTO ESTADÍSTICO PARA PREFERIR EL MANOVA

1. El uso fragmentado (independiente) de pruebas univariadas eleva el error tipo I, es decir, la probabilidad de rechazar la hipótesis planteada cuando en realidad es verdadera. Por ejemplo,

   Caso.- Comparación de dos grupos con 10 variables de investigación

   $$\underbrace{(.95)(.95)\ldots(.95)}_{\text{10 veces}} \approx .60$$ (nivel de confianza)

2. Las pruebas estadísticas univariadas ignoran información importante. Al contrario, las pruebas multivariadas incorporan la relación de las variables en el coeficiente de correlación.

CUATRO RAZONES CON SUSTENTO ESTADÍSTICO PARA PREFERIR EL MANOVA

3. Aunque los grupos podrían no ser significativamente diferentes en alguna de sus variables individualmente podrían serlo en un análisis multivariado

¡la prueba multivariada puede ser más potente en este caso!

4. El análisis multivariado de las pruebas estadísticas individuales puede reflejar diferencias significativas que no sería posibles detectar con un enfoque univariado. Por ejemplo: "No hubieron diferencias significativas en la prueba multivariada" y, sin embargo, en las pruebas individuales se obtuvo:

Prueba parcial I	**Grupo 1 muy superior a Grupo 2**	Prueba parcial III	**Grupo 1 no hay diferencia Grupo 2**
Prueba parcial II	**Grupo 1 algo superior Grupo 2**	Prueba parcial IV	**Grupo 1 muy inferior al Grupo 2**

Por otro lado, exagerar en el uso del MANOVA sin una razón empírica o teórica, entonces, pequeñas o no significativas diferencias podrían ocultar reales diferencias univariadas

	Univariado	Multivariado
Dos Grupos	ANOVA (prueba t)	MANOVA T^2 Hotelling
Dos o Más Grupos	ANOVA (Prueba F)	MANOVA (Prueba F)
	ANOVA Factorial	MANOVA (Factorial)

MANOVA

Supuestos

SUPUESTOS

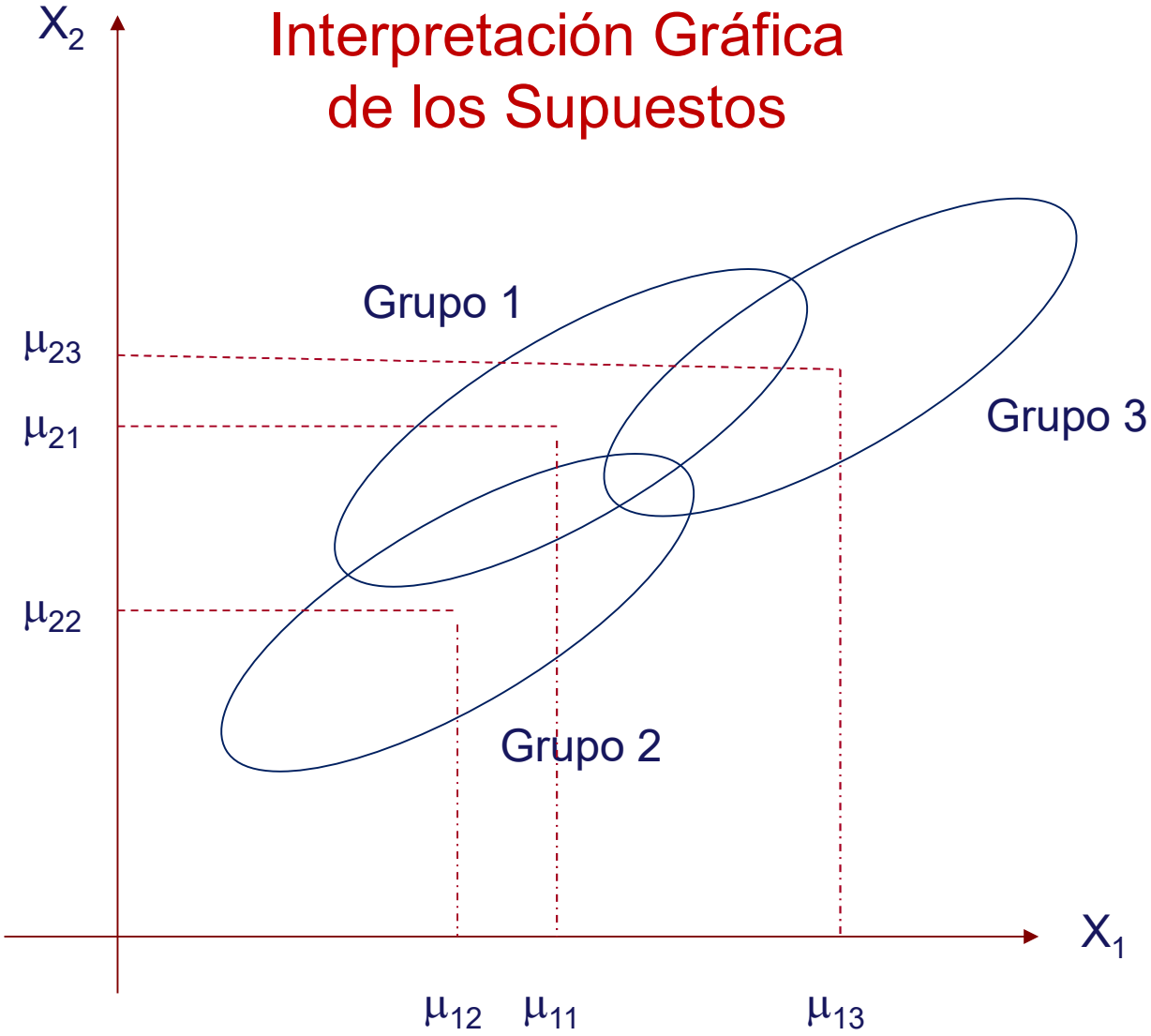
1. $X_{l1}, X_{l2}, ..., X_{lnl}$ es una muestra aleatoria de tamaño n_l de una población con media μ_l; l=1, 2, ..., g

2. **Muestras Aleatorias de diferentes poblaciones son independientes**
   (La violación de este supuesto es un problema serio)

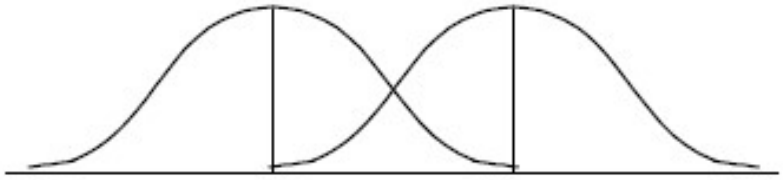
3. **Todas las poblaciones tienen matriz varianza-covarianza común igual a** Σ
   (Condicionalmente robusta muestra más pequeña/muestra más grande<1.5; robusta si los tamaños de muestra son aproximadamente iguales)

4. **Cada población es normal multivariada**
   (robusto con respecto al error tipo I; no existen estudios con respecto al efecto de las asimetrias, pero una distribución ´platicurtica atenua la potencia)

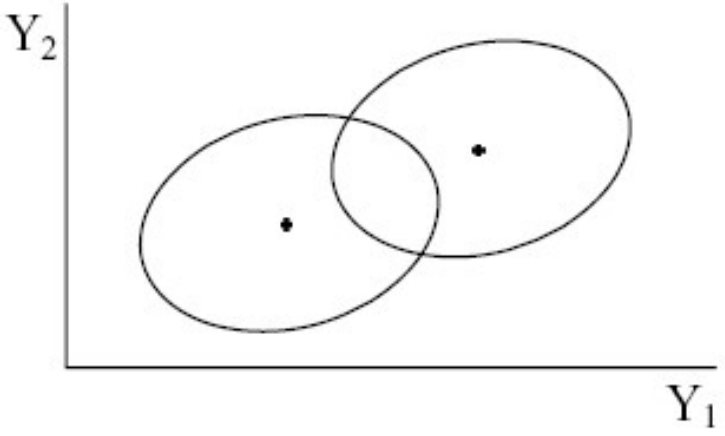
Interpretación Gráfica de los Supuestos



SUPUESTOS ESPECIALES DEL MANOVA (Y ANÁLISIS DISCRIMINANTE)

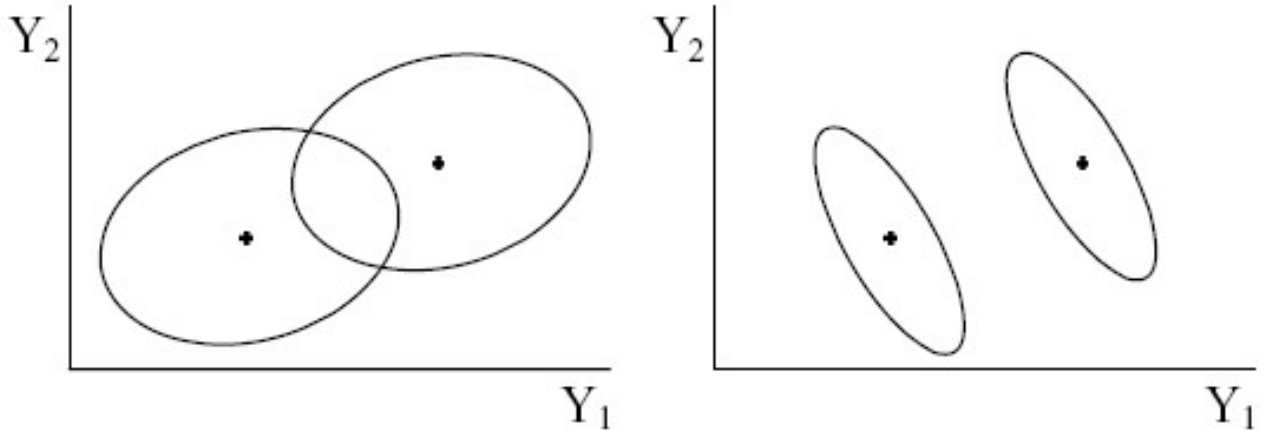


ANOVA
"Homogeneidad de Varianzas"

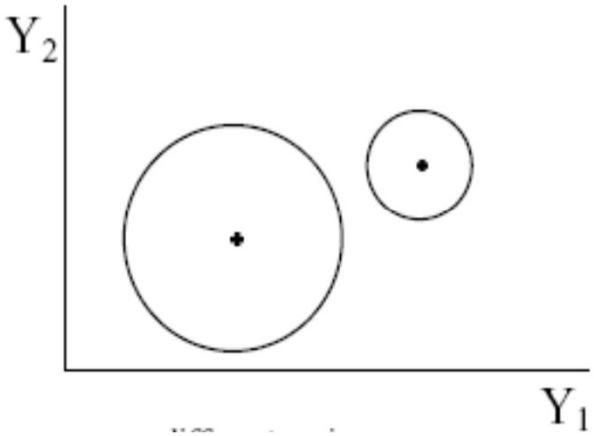


MANOVA
"Homogeneidad de Dispersión"
(varianzas y covarianzas)

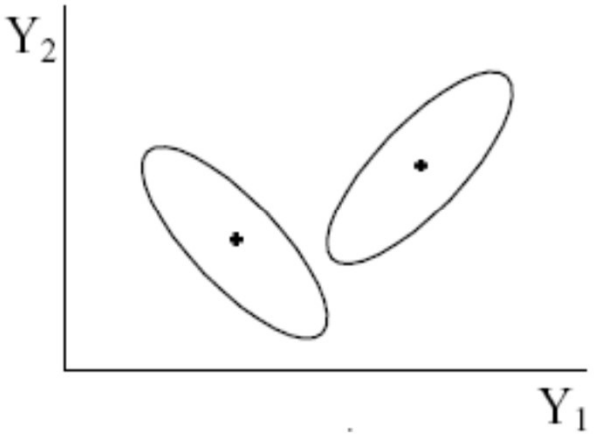
HOMOGENEIDAD DE DISPERSIÓN



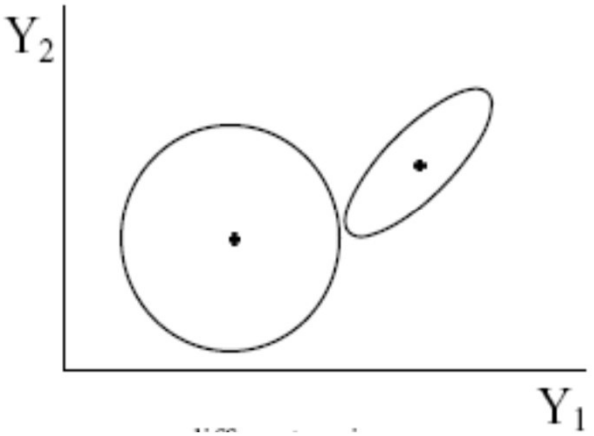
HETEROGENEIDAD DE DISPERSIÓN



Varianzas Diferentes
Correlaciones Iguales



Varianzas Iguales
Correlaciones Diferentes



Varianzas Diferentes
Correlaciones Diferentes

Modelos de Análisis

- Análisis Multivariado No Estructurado
- Análisis Multivariado Estructurado (un factor o un sentido)
- Análisis Multivariado Estructurado (dos factores o dos sentido)
- Análisis de Medidas Repetidas
- Análisis de Un Factor Entre Grupos
- Análisis de Dos Factores Entre Grupos
- Análisis de Tres Factores Entre Grupos
- Covariación

MANOVA

Un factor o un sentido

El Modelo

Los Datos

Muestras Aleatorias recolectadas de "g" poblaciones:

Población 1:	X_{11}	X_{12}	$X_{1\,n1}$
Población 2:	X_{21}	X_{22}	$X_{2\,n2}$
.		.	
.		.	
.		.	
Población g:	X_{g1}	X_{g2}	$X_{g\,ng}$

" MANOVA nos permite averiguar si el vector de medias poblacional es el mismo en cada población "

El Modelo

$$X_{lj} = \mu + \tau_i + \varepsilon_{ij}$$

$j = 1, 2, ..., n_l$

$i = 1, 2, ..., g$

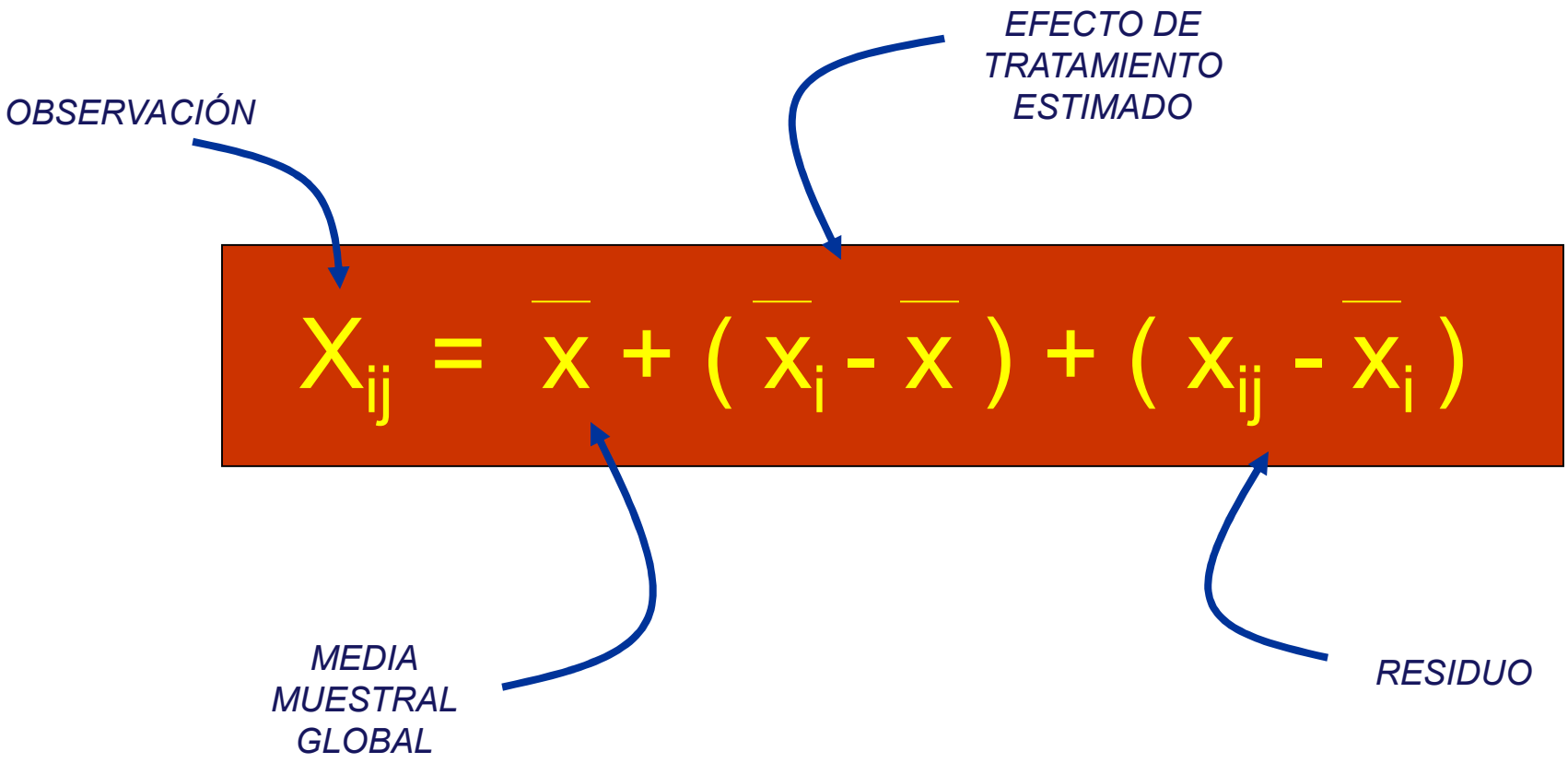
$\varepsilon_{ij} \sim N_p(0,\Sigma)$

μ : Media Poblacional Global

τ_i : i-ésimo efecto del tratamiento

Condición: $\sum_{l=1}^{g} n_i \tau_i = 0$

Descomposición de una observación



SUMA DE CUADRAOS CORREGIDO

$$\sum_{I}\sum_{j}(X_{ij}-\bar{x})(X_{ij}-\bar{x})^t =$$

$$= \sum_{i} n_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^t + \sum_{i}\sum_{j}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^t$$

SUMA DE CUADRADO DE TRATAMIENTO

SUMA DE CUADRADOS RESIDUAL

SUMA DE CUADRADOS RESIDUAL

$$\sum_i \sum_j (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^t =$$

$$= (n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g$$

Prueba de Hipótesis

HIPÓTESIS DE NO EFECTO DE TRATAMIENTOS O NO DIFERENCIA ENTRE GRUPOS

$$H_O: \quad \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

$$H_O: \quad \mu_1 = \mu_2 = \ldots = \mu_g$$

DADO QUE: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$

Prueba del Ratio de Verosimilitud

Ho: $\theta \in \Omega_o$

H_1: $\theta \in \Omega_1$

Teorema

Si $\Omega_1 \subset$ **Rq es un espacio q-dimensional y si** $\Omega_o \subset \Omega_1$ **es un subespacio r-dimensional, entonces, bajo condiciones de regularidad:**

$$\forall \theta \in \Omega_o: \ -2\log\lambda \rightarrow \chi^2_{(q-r)} \ \text{cuando} \ n \rightarrow \infty$$

Región de rechazo R:

$$R = \{X: -2\log \lambda_{(X)} > \chi^2_{1-\alpha;(q-r)} \}$$

Donde λ se determina como:

$$\lambda = \frac{\text{Max } L(\Omega_0)}{\text{Max } L(\Omega_1)}$$

En el contexto del modelo de regresión lineal:

$$\lambda^{-\frac{2}{n}} = \frac{\hat{\varepsilon}'\hat{\varepsilon}_{\Omega_0}}{\hat{\varepsilon}'\hat{\varepsilon}_{\Omega_1}}$$

$$\Rightarrow \quad \left(\lambda^{-\frac{2}{n}} - 1 \right) \frac{n-k}{q} \sim F_{(q\,,\,n-k)}$$

Prueba de Λ Wilks

A partir de una muestra aleatoria $\mathbf{x}_i$ de tamaño n de una población con distribución normal $\boldsymbol{f}(x;\theta)$:

$$L(\mathbf{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta}).$$

$$\Rightarrow \quad l(\mathbf{X};\boldsymbol{\theta}) = \log L(\mathbf{X};\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i;\boldsymbol{\theta}). \qquad (1)$$

Donde, $L(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\}$, tomando logaritmos a L

$$\Rightarrow \quad l(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log L(\mathbf{X};\boldsymbol{\theta}) = -\frac{n}{2}\log|2\pi\boldsymbol{\Sigma}| - \tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}).$$

Analizando el argumento de la sumatoria se tiene:

$$(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = (\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) + (\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) + 2(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}), \qquad (2)$$

Aplicando sumatoria a ambos miembros de (2):

$$\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}).$$, además, $(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) = \operatorname{tr}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'.$

De allí que: $\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = \operatorname{tr}\boldsymbol{\Sigma}^{-1}\left\{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'\right\} + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}).$ y $\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})' = n\mathbf{S}$

$$\Rightarrow \quad l(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{n}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{n}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}). \qquad (3)$$

En un caso de **g** muestras independientes de tamaño $\mathbf{n_i}$ se tiene la siguiente estructura de datos:

$$\underset{(nxp)}{\mathbf{X}} = \begin{bmatrix} X_1 \\ \vdots \\ X_g \end{bmatrix}$$

Donde:

$n^{-1}\mathbf{W}$ es el estimador MV de la varianza común y $\mathbf{W} = \sum n_i S_i$

$n = \sum n_i$

Suma de cuadrados dentro de grupos (SCP)

De (3) y asumiendo que: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$

$$l = -\tfrac{1}{2}\sum_i [n_i \log |2\pi\boldsymbol{\Sigma}| + n_i \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\mathbf{S}_i + \mathbf{d}_i\mathbf{d}_i')], \qquad (4)$$

Donde S_i es la matriz de covarianzas de la i-esima muestra y $d_i = x_i - \mu_i$.

MANOVA

FUENTE DE VARIABILIDAD	MATRIZ SUMA DE CUADRADOS Y CRUZADOS	GRADOS DE LIBERTAD
GRUPOS	$B = \sum_{I} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t$	$g - 1$
RESIDUAL	$W = \sum_{I} \sum_{j} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^t$	$\sum_{1}^{g} n_i - g$
TOTAL (corregido en media)	**B + W**	$\sum_{1}^{g} n_i - 1$

IV.
La Prueba de Wilks

Rechazamos Ho si Λ es muy pequeño

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^t}{\sum_i \sum_j (X_{ij} - \bar{x})(X_{ij} - \bar{x})^t}$$

$$\Lambda^* = \prod \frac{1}{1+\lambda_j} \qquad \Lambda^* \text{ (Lamda de Wilks)}$$

*"A mayor valor de **(1 - Λ*)**, mayor la proporción de varianza generalizada que puede ser atribuida a la variación entre grupos"*

Donde los λ_j son los valores característicos de $\mathbf{BW^{-1}}$

Veamos la distribución de $\Lambda^{2/n}=\frac{|W|}{|B+W|}=|I+W^{-1}B|^{-1}$

Bajo **Ho**, $X_{(nxp)} \sim N_p(\mu,\Sigma)$
Cochran (teorema 3.4.4) y
Craig (teorema 3.4.5)

$W = X'C_1X \sim W_p(\Sigma, n-k)$

$B = X'C_2X \sim W_p(\Sigma, k-1)$

$T = B + W$

B y W son independientes

$$|I+W^{-1}B|^{-1} \sim \Lambda(p,n-g,g-1)$$

Ver Teorema 3.7.3, Mardia et al. (1982)

Distribución de T=W+B:

Sean $\mathbf{X}_i(n_i x p)$ ($\forall i=1,\ldots k$) muestras aleatorias independientes de una población normal con media μ y matriz de covarianzas Σ. Además, si $\mathbf{1}_i$ es un vector nx1 con unos en los lugares correspondientes a la i-esima muestra, $\mathbf{I}_i = \text{diag}(\mathbf{1}_i)$ y la matriz de centralización $\mathbf{H}_i=\mathbf{I}_i-n_i^{-1}\mathbf{1}_i\mathbf{1}'_i$ para la i-ésima muestra. Si $\mathbf{X}(nxp)=(X_1, X_2, \ldots, X_k)'$ determine la distribución de $\mathbf{Z} = \mathbf{X}'C_1\mathbf{X}+\mathbf{X}'C_2\mathbf{X}$ sabiendo que: $C_1=\Sigma\mathbf{H}_i$ y $C_2=\Sigma\, n_i^{-1}\mathbf{1}_i\mathbf{1}'_i - n^{-1}\mathbf{1}\mathbf{1}'$.

$$\underset{(nxp)}{\mathbf{X}} = \begin{bmatrix} X_1 \\ \vdots \\ X_g \end{bmatrix} \begin{matrix} (n_1 xp) \\ \\ (n_g xp) \end{matrix}$$

$$\underset{nx1}{\mathbb{1}_i} = \begin{bmatrix} 0 \\ \cdots \\ \mathbb{1}_{n_i} \\ \cdots \\ 0 \end{bmatrix}, \quad \underset{nxn}{\mathbb{I}_i} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbb{I}_{n_i} & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ y } \underset{nxn}{H_i} = \underset{nxn}{\mathbb{I}_i} - \underset{nxn}{n_i^{-1}\mathbb{1}_i\mathbb{1}_i'} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbb{I}_{n_i} - n_i^{-1}\mathbb{J}_{n_i} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Se debe entender que: $\sum_i \mathbb{1}_i = \mathbb{1}_n$ y $\sum_i \mathbb{I}_{n_i} = \mathbb{I}_n$

Definiendo C_1: $$C_1 = \sum H_i = diag\left(\mathbb{I}_{n_i} - n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'\right)$$

$$C_1 = \begin{bmatrix} \mathbb{I}_{n_1} - n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}_{n_1}' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbb{I}_{n_k} - n_k^{-1}\mathbb{1}_{n_k}\mathbb{1}_{n_k}' \end{bmatrix}$$

Se verifica que C_1 es simétrica e idempotente

$$\mathrm{r}(C_1) = Tr(C_1) = n - k \text{ dado que:}$$
$$Tr\left(\mathbb{I}_{n_i} - n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}'_{n_i}\right) = n_i - 1 \text{ para } i = \overline{1,k}$$
$$X'C_1X \sim W_p(\Sigma, n-k)$$

Definiendo C_2:

$$C_2 = \sum_i n_i^{-1}\mathbb{1}_i\mathbb{1}'_i - n^{-1}\mathbb{1}\mathbb{1}'$$
$$= n_1^{-1}\mathbb{1}_1\mathbb{1}'_1 + n_2^{-1}\mathbb{1}_2\mathbb{1}'_2 + \ldots + n_k^{-1}\mathbb{1}_k\mathbb{1}'_k - n^{-1}\mathbb{1}_n\mathbb{1}_n$$

$$= \begin{bmatrix} n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} & 0 & \ldots & 0 \\ 0 & n_2^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & n_k^{-1}\mathbb{1}_{n_k}\mathbb{1}'_{n_k} \end{bmatrix} - \begin{bmatrix} n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} & n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_2} & \ldots & n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_k} \\ n^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_1} & n^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_2} & \ldots & n^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ n^{-1}\mathbb{1}_{n_k}\mathbb{1}'_{n_1} & n^{-1}\mathbb{1}_{n_k}\mathbb{1}'_{n_2} & \ldots & n^{-1}\mathbb{1}_{n_k}\mathbb{1}'_{n_k} \end{bmatrix}$$

Se verifica que C_2 es simétrica

Se verifica que C_2 es simétrica e idempotente

Para el bloque (1,1)

$$= \left[n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} - n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} \quad -n^{-1}\mathbb{1}_{n_1}\mathbb{1}_{n_2} \quad \dots \quad -n^{-1}\mathbb{1}_{n_1}\mathbb{1}_{n_k} \right] \begin{bmatrix} n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} - n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} \\ -n^{-1}\mathbb{1}_{n_1}\mathbb{1}_{n_2} \\ \vdots \\ -n^{-1}\mathbb{1}_{n_1}\mathbb{1}_{n_k} \end{bmatrix}$$

$$= n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} - n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1}$$

Para el bloque (1,2)

$$= \left[n_1^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} - n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_1} \quad -n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_2} \quad \dots \quad -n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_k} \right] \begin{bmatrix} -n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_2} \\ n_2^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_2} - n^{-1}\mathbb{1}_{n_2}\mathbb{1}'_{n_2} \\ \vdots \\ -n^{-1}\mathbb{1}_{n_k}\mathbb{1}'_{n_2} \end{bmatrix}$$

$$= -2n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_2} + \frac{n_1 + n_2 + \cdots + n_k}{n^2}\mathbb{1}_{n_1}\mathbb{1}'_{n_2}$$

$$= -n^{-1}\mathbb{1}_{n_1}\mathbb{1}'_{n_2}$$

Se verifica que C_2 es idempotente

$$Tr\left(n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}'_{n_i} + n^{-1}\mathbb{1}_{n_i}\mathbb{1}'_{n_i}\right) = 1 - \frac{n_i}{n} \implies \sum_{i=1}^{k}\left(1 - \frac{n_i}{n}\right) = k - 1 \Rightarrow \mathrm{r}(C_2) = k - 1$$

$$X'C_2X \sim W_p(\Sigma, k-1)$$

Aplicamos el teorema de Craig.
Calculamos $C_1'C_2$.Como C_1 es bloque diagonal solo calcularemos un bloque genérico i-ésimo y a partir de allí generalizamos.

$$\left(I_{n_i} - n_1^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'\right)\left(n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' - n^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'\right)$$
$$= n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' - n_1^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' - n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' + n_1^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'n^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}'$$
$$= n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' - n_i^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' - n^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' + n^{-1}\mathbb{1}_{n_i}\mathbb{1}_{n_i}' = 0$$

Entonces, $C_1'C_2 = 0$ y concluimos que $X'C_1X$ y $X'C_2X$ son independientes.

Además, $X'C_1X + X'C_2X = W_p(\Sigma, n-k) + W_p(\Sigma, k-1) = W_p(\Sigma, n-1)$

DISTRIBUCIÓN DE LAMBDA WILKS

No.Variables	No.Grupos	Distribución Muestral Datos Normales
p = 1	g ≥ 2	$\frac{\sum n_l - g}{g - 1}\left\{\frac{1 - \Lambda^*}{\Lambda^*}\right\} \sim F(g-1, \Sigma n_l - g)$
p = 2	g ≥ 2	$\frac{\sum n_l - g - 1}{g - 1}\left\{\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right\} \sim F(2(g-1), \Sigma n_l - g - 1)$
p ≥ 1	g = 2	$\frac{\sum n_l - p - 1}{p}\left\{\frac{1 - \Lambda^*}{\Lambda^*}\right\} \sim F(p, \Sigma n_l - p - 1)$
p ≥ 1	g = 3	$\frac{\sum n_l - p - 2}{p}\left\{\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right\} \sim F(2p, 2(\Sigma n_l - p - 2))$

I.
La Traza de Pillai-Barttlet

$$Tp = \sum_j \frac{\lambda_j}{1 + \lambda_j}$$

Donde los λ_j son los valores característicos de $\mathbf{B(B+W)^{-1}}$

II.
La Traza de Hotelling-Lawley

$$Th = \sum_j \lambda_j$$

Donde los λ_j son los valores característicos de $\mathbf{BW^{-1}}$

III.
La Prueba de Roy

$$Tr = \lambda_1$$

$$Max\{\lambda_j\} = \lambda_1$$

$$\forall j,\ j=1,\ldots,p$$

Donde los λ_j son los valores característicos de $\mathbf{W(B+W)^{-1}}$

HIPÓTESIS DE IGUALDAD DE MATRICES DE COVARIANZAS ENTRE GRUPOS

$$H_O: \quad \Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$$

$$H_1: \quad \Sigma_i \neq \Sigma_j$$

PRUEBA M DE BOX

$$X_K \sim N_k(\mu_k, \Sigma_k) \quad ; k = 1, 2, \ldots, g$$

$$M = \frac{\left(\prod_{k=1}^{g} |S_k| \right)^{(n_k - 1)/2}}{|\overline{S}|^{(n-g)/2}}$$

S_k : matriz de covarianzas muestral del grupo "k"

$$S = \sum_{k=1}^{g} (n_k - 1) S_k / (n - g)$$

$$n = \sum_{k=1}^{g} n_k$$

Distribuciones aproximadas de "M"

A) $-2\,(1 - c_1)\,\ln M \sim \chi^2_{\left(\frac{1}{2}p(p+1)(g-1)\right)}$

$$c_1 = \frac{(2p^2 + 3p - 1)}{6(p+1)(g-1)} \left[\sum_{k=1}^{g} \frac{1}{(n_k - 1)} - \frac{1}{(n - g)}\right]$$

B) $-2b \log M \sim F(v_1, v_2)$

$$V_1 = \frac{1}{2} p(p+1)(g-1) \qquad b = (1 - c_1 - v_1/v_2)/v_1$$

$$V_2 = (v_1 + 2) / \left(c_2 - c_1^2\right)$$

$$c_2 = \frac{(p-1)(p+2)}{6(g-1)} \left[\sum_{k=1}^{g} \frac{1}{(n_k - 1)}^2 - \frac{1}{(n - g)}^2\right]$$

MANOVA

En la Investigación Social

POTENCIA DE PRUEBA Y DETERMINACIÓN A PRIORI DEL TAMAÑO DE MUESTRA

Tamaños de Muestra por Grupo de 3 a 6 Variables

(Potencia=70% y α=5%)

Efecto	Cantidad de Variables			
	3	4	5	6
Muy grande	12–16	14–18	15–19	16–21
Grande	25–32	28–36	31–40	33–44
Mediano	42–54	48–62	54–70	58–76
Pequeño	92–120	105–140	120–155	130–170

FUENTE: Stevens, James. Applied Multivariate Statistics for the Social Sciences. LEA, 2002. London

POTENCIA DE PRUEBA Y DETERMINACIÓN A PRIORI DEL TAMAÑO DE MUESTRA

Parámetros para la tabla de Lauter para una desviación mínima de la hipótesis nula multivariada:

1. Existe una variable ***i*** tal que: $\frac{1}{\sigma^2} \sum_{j=1}^{j} (\mu_{ij} - \mu_i) \geq q^2$ donde μ_i es la media total y σ^2 la varianza
2. Existe una variable ***i*** tal que $\frac{1}{\sigma_i} \left| \mu_{i,j1} - \mu_{i,j2} \right| \geq d$ para dos grupos j_1 y j_2
3. Existe una variable ***i*** tal que para todo par de los grupos ***1*** y ***m*** tenemos que:

$$\frac{1}{\sigma_i} \left| \mu_{i1} - \mu_{im} \right| \geq c$$

Tamaño de muestra necesario en **tres** grupos MANOVA para potencia 70%, 80% y 90% para α=0.05 y α=0.01

			α = .05			α = .01		
		Power =	.70	.80	.90	.70	.80	.90
Number of Variables		2	11	13	16	15	17	21
Effect Size		3	12	14	18	17	20	24
	q^2=1.125	4	14	16	19	19	22	26
Very Large	d=1.5	5	15	17	21	20	23	28
	c=0.75	6	16	18	22	22	25	29
		8	18	21	25	24	28	32
		10	20	23	27	27	30	35
		15	24	27	32	32	35	42
Large	q^2=0.5	2	21	26	33	31	36	44
	d=1	3	25	29	37	35	42	50
	c=0.5	4	27	33	42	38	44	54
		5	30	35	44	42	48	58
		6	32	38	48	44	52	62
		8	36	42	52	50	56	68
		10	39	46	56	54	62	74
		15	46	54	66	64	72	84
Moderate	q^2=0.2813	2	36	44	58	54	62	76
	d=0.75	3	42	52	64	60	70	86
	c=0.375	4	46	56	70	66	78	94
		5	50	60	76	72	82	100
		6	54	66	82	76	88	105
		8	60	72	90	84	98	120
		10	66	78	98	92	105	125
		15	78	92	115	110	125	145
Small	q^2=0.125	2	80	98	125	115	140	170
	d=0.5	3	92	115	145	135	155	190
	c=0.25	4	105	125	155	145	170	210
		5	110	135	170	155	185	220
		6	120	145	180	165	195	240
		8	135	160	200	185	220	260
		10	145	175	220	200	230	280
		15	170	210	250	240	270	320

FUENTE: Stevens, James. Applied Multivariate Statistics for the Social Sciences. LEA, 2002. London

Tamaño de muestra necesario en **cuatro** grupos MANOVA para potencia 70%, 80% y 90% para α=0.05 y α=0.01

			α = .05			α = .01		
		Power =	*.70*	*.80*	*.90*	*.70*	*.80*	*.90*
Number of Variables		2	12	14	17	17	19	23
Effect Size		3	14	16	20	19	22	26
	q^2=1.125	4	15	18	22	21	24	28
Very Large	d=1.5	5	16	19	23	23	26	30
	c=0.4743	6	18	21	25	24	27	32
		8	20	23	28	27	30	36
		10	22	25	30	29	33	39
		15	26	30	36	35	39	46
Large	q^2=0.5	2	24	29	37	34	40	50
	d=1	3	28	33	42	39	46	56
	c=0.3162	4	31	37	46	44	50	60
		5	34	40	50	48	54	64
		6	36	44	54	50	58	70
		8	42	48	60	56	64	76
		10	46	52	64	62	70	82
		15	54	62	76	72	82	96
Moderate	q^2=0.2813	2	42	50	64	60	70	86
	d=0.75	3	48	58	72	68	80	96
	c=0.2372	4	54	64	80	76	88	105
		5	58	70	86	82	94	115
		6	62	74	92	86	100	120
		8	70	84	105	96	115	135
		10	78	92	115	105	120	145
		15	92	110	130	125	145	170
Small	q^2=0.125	2	92	115	145	130	155	190
	d=0.5	3	105	130	165	150	175	220
	c=0.1581	4	120	145	180	165	195	240
		5	130	155	195	180	210	250
		6	140	165	210	190	220	270
		8	155	185	230	220	250	300
		10	170	200	250	240	270	320
		15	200	240	290	280	320	370

FUENTE: Stevens, James. Applied Multivariate Statistics for the Social Sciences. LEA, 2002. London

Tamaño de muestra necesario en **cinco** grupos MANOVA para potencia 70%, 80% y 90% para α=0.05 y α=0.01

			α = .05			α = .01		
		Power =	*.70*	*.80*	*.90*	*.70*	*.80*	*.90*
Number of Variables		2	13	15	19	18	20	25
Effect Size		3	15	17	21	20	23	28
	q^2=1.125	4	16	19	23	22	26	30
Very Large	d=1.5	5	18	21	25	24	28	33
	c=0.3354	6	19	22	27	26	30	35
		8	22	25	30	29	33	39
		10	24	27	33	32	36	42
		15	28	33	39	38	44	50
Large	q^2=0.5	2	26	32	40	37	44	54
	d=1	3	31	37	46	44	50	60
	c=0.2236	4	34	42	50	48	56	66
		5	37	44	54	52	60	70
		6	40	48	58	56	64	76
		8	46	54	66	62	70	84
		10	50	58	72	68	78	90
		15	60	70	84	80	90	110
Moderate	q^2=0.2813	2	46	56	70	66	76	92
	d=0.75	3	54	64	80	74	86	105
	c=0.1677	4	60	72	88	82	96	115
		5	64	78	96	90	105	125
		6	70	82	105	96	110	135
		8	78	92	115	110	125	145
		10	86	105	125	120	135	160
		15	105	120	145	140	160	185
Small	q^2=0.125	2	100	125	155	145	170	210
	d=0.5	3	120	145	180	165	195	240
	c=0.1118	4	130	160	195	185	210	260
		5	145	170	220	200	230	280
		6	155	185	230	220	250	300
		8	175	210	260	240	280	330
		10	190	230	280	260	300	360
		15	230	270	330	310	350	420

FUENTE: Stevens, James. Applied Multivariate Statistics for the Social Sciences. LEA, 2002. London

Tamaño de muestra necesario en **cinco** grupos MANOVA para potencia 70%, 80% y 90% para α=0.05 y α=0.01

			α = .05			α = .01		
		Power =	*.70*	*.80*	*.90*	*.70*	*.80*	*.90*
Number of Variables		2	14	16	20	19	22	26
Effect Size		3	16	18	23	22	25	29
	q^2=1.125	4	18	21	25	24	27	32
Very Large	d=1.5	5	19	22	27	26	30	35
	c=0.2535	6	21	24	29	28	32	37
		8	23	27	33	31	35	42
		10	25	30	36	34	39	46
		15	30	35	42	42	46	54
Large	q^2=0.5	2	28	34	44	40	46	56
	d=1	3	33	39	50	46	54	64
	c=0.1690	4	37	44	54	52	60	70
		5	40	48	60	56	64	76
		6	44	52	64	60	68	82
		8	50	58	70	68	76	90
		10	54	64	78	74	84	98
		15	64	76	90	88	98	115
Moderate	q^2=0.2813	2	50	60	76	70	82	98
	d=0.75	3	58	70	86	80	94	115
	c=0.1268	4	64	76	96	90	105	125
		5	70	84	105	98	115	135
		6	76	90	110	105	120	145
		8	86	100	125	120	135	160
		10	94	110	135	130	145	175
		15	115	135	160	155	175	210
Small	q^2=0.125	2	110	135	170	155	180	220
	d=0.5	3	130	155	190	180	210	250
	c=0.0845	4	145	170	220	200	230	280
		5	155	185	230	220	250	300
		6	170	200	250	230	270	320
		8	190	230	280	260	300	350
		10	210	250	300	290	330	390
		15	250	290	360	340	380	460

FUENTE: Stevens, James. Applied Multivariate Statistics for the Social Sciences. LEA, 2002. London

EJERCICIOS

Un investigador tiene que analizar cuatro grupos por MANOVA con 5 variables dependientes. Se desea una potencia de 80% y α=0.05. de investigaciones previas y su conocimiento acerca de la naturaleza de los tratamientos anticipa un moderado efecto del tamaño de la muestra. ¿Cuántas observaciones por grupo necesitará?

RPTA.-

Según la tabla para cuatro grupos necesitará 70 observaciones por grupo.

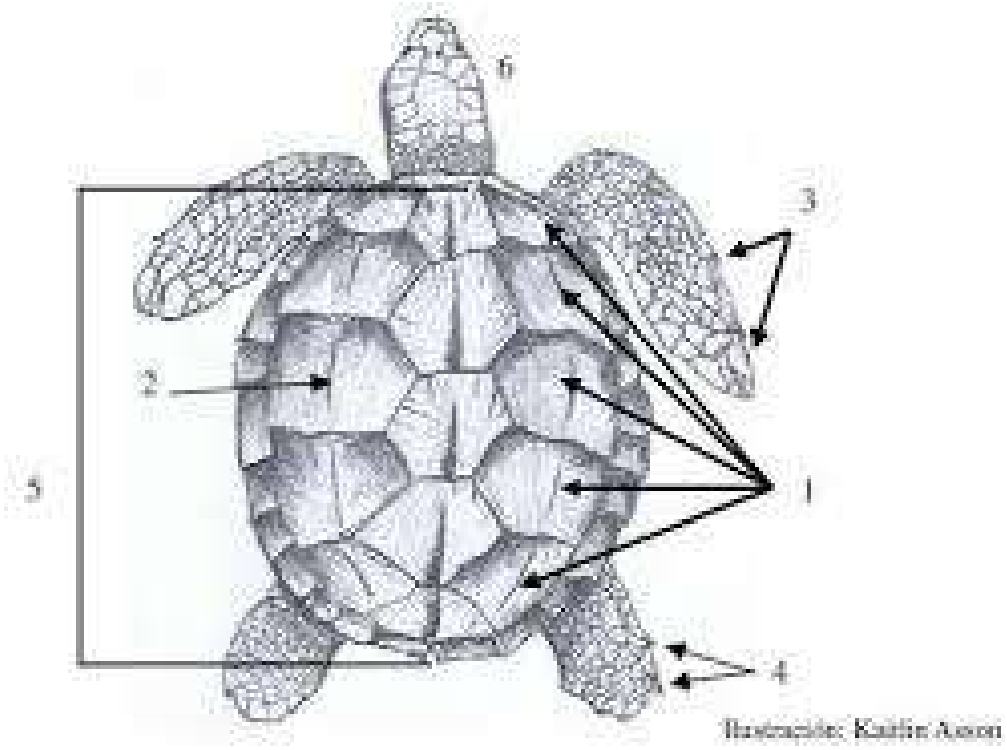
EJERCICIOS
(Continuación)

Un equipo de investigadores tiene cinco grupos y 7 variables dependientes para analizar por MANOVA. Ellos desean una potencia del 70% y α=0.05. De estudios anteriores anticipan un gran efecto del tamaño de la muestra. ¿Cuántas observaciones por grupo necesitan?

RPTA.-
Interpolando en la tabla para 5 grupos entre 6 y 8 variables vemos que son necesarias 43 observaciones por grupo , haciendo un total de 215 observaciones.

ESTUDIO DE LAS TORTUGAS "CARETTA CARETTA" Quelónidos

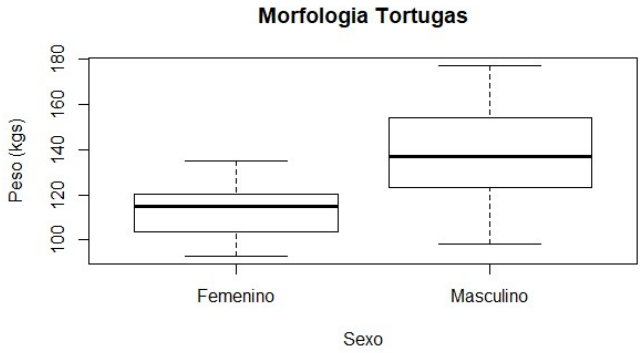
X_1: Peso (kgs)
X_2: Patas anteriores (cm)
X_3: Patas posteriores (cm)

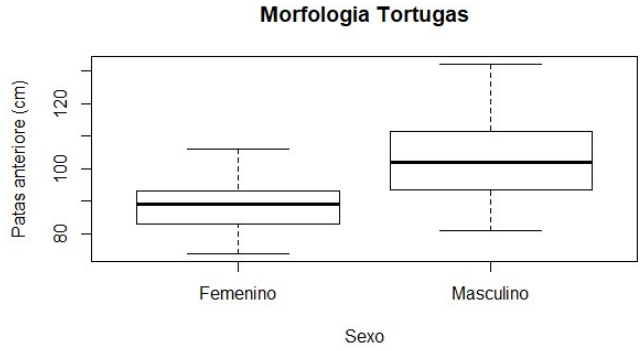


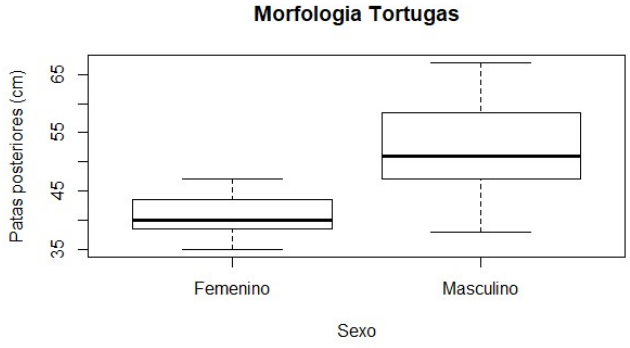
X_1: Peso (kgs)
X_2: Patas anteriores (cm)
X_3: Patas posteriores (cm)

La muestra:
n= 48
n_1=24
n_2=24

id	x1	x2	x3	sexo
1	98	81	38	Masculino
2	103	84	38	Masculino
3	103	86	42	Masculino
4	105	86	42	Masculino
5	109	88	44	Masculino
6	123	92	50	Masculino
7	123	95	46	Masculino
8	133	99	51	Masculino
9	133	102	51	Masculino
10	133	102	51	Masculino
11	134	100	48	Masculino
12	136	102	49	Masculino
13	138	98	51	Masculino
14	138	99	51	Masculino
15	141	105	53	Masculino
16	147	108	57	Masculino
17	149	107	55	Masculino
18	153	107	56	Masculino
19	155	115	63	Masculino
20	155	117	60	Masculino
21	158	115	62	Masculino
22	159	118	63	Masculino
23	162	124	61	Masculino
24	177	132	67	Masculino
25	93	74	37	Femenino
26	94	78	35	Femenino
27	96	80	35	Femenino
28	101	84	39	Femenino
29	102	85	38	Femenino
30	103	81	37	Femenino
31	104	83	39	Femenino
32	106	83	39	Femenino
33	107	82	38	Femenino
34	112	89	40	Femenino
35	113	88	40	Femenino
36	114	86	40	Femenino
37	116	90	43	Femenino
38	117	90	41	Femenino
39	117	91	41	Femenino
40	119	93	41	Femenino
41	120	89	40	Femenino
42	120	93	44	Femenino
43	121	95	42	Femenino
44	125	93	45	Femenino
45	127	96	45	Femenino
46	128	95	45	Femenino
47	131	95	46	Femenino
48	135	106	47	Femenino



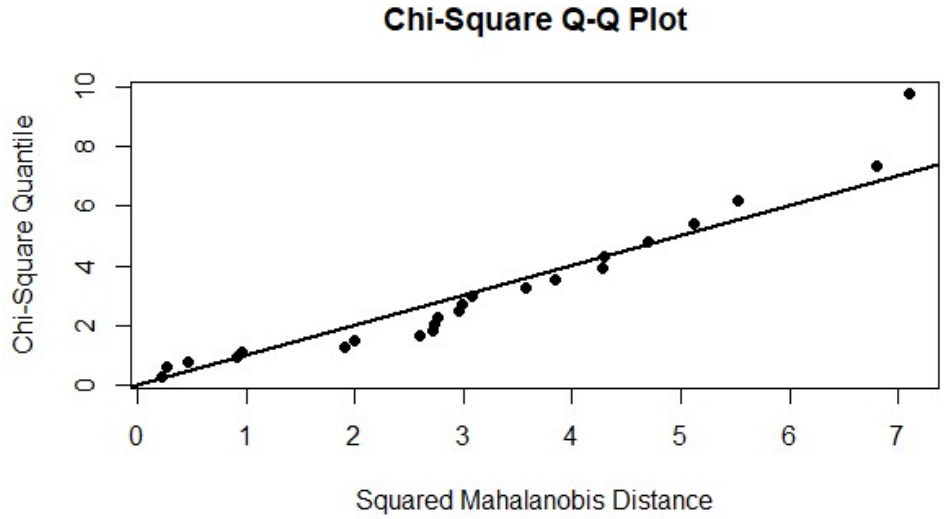




Prueba de normalidad

A. Tortugas masculinas

Ho: La muestra de tortugas masculinas proviene de una población con distribución normal

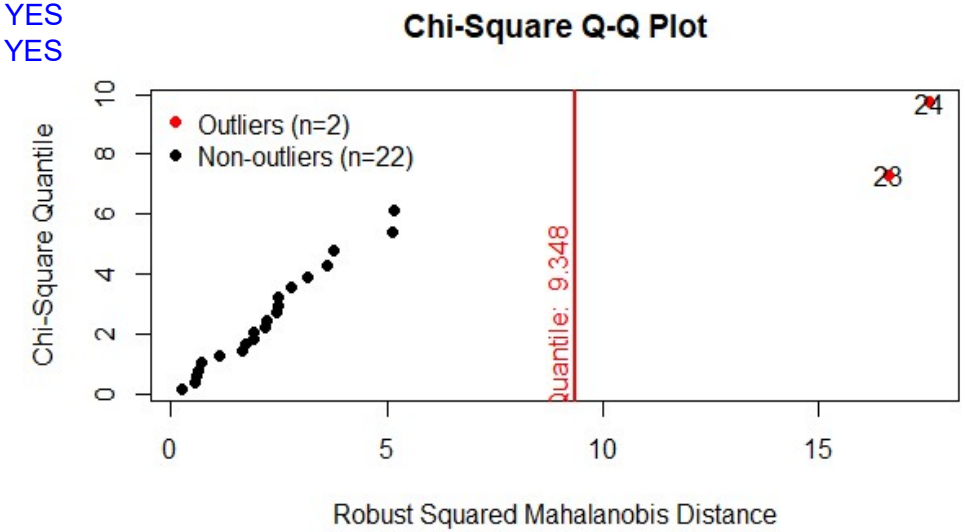


Prueba de Mardia

Test	Statistic		p value	Result
1 Mardia	Skewness	11.239711247582	0.339149885307009	YES
2 Mardia	Kurtosis	-1.00790375015034	0.313500671138852	YES
3	MVN	<NA>	<NA>	YES

Prueba de Royston

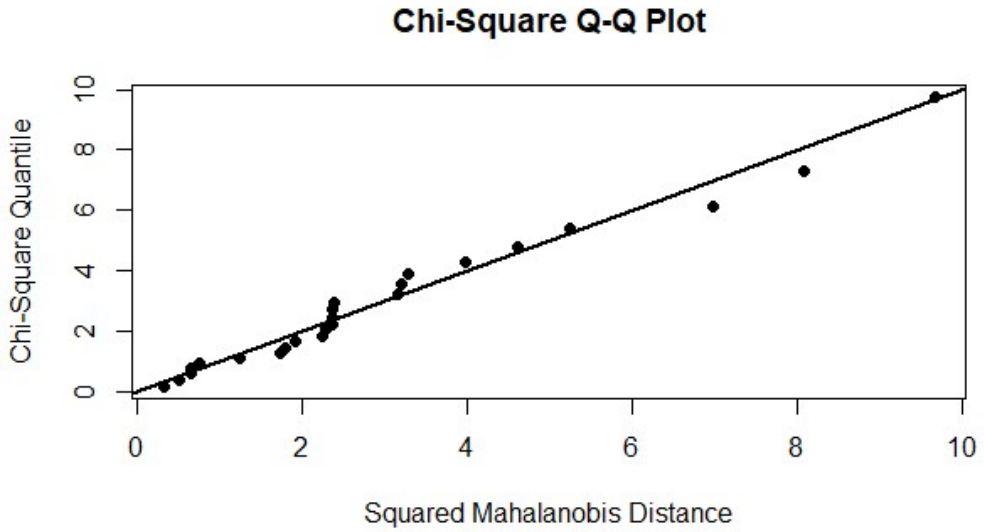
Test	H	p value	MVN
1 Royston	0.534647	0.5587516	YES



Prueba de normalidad

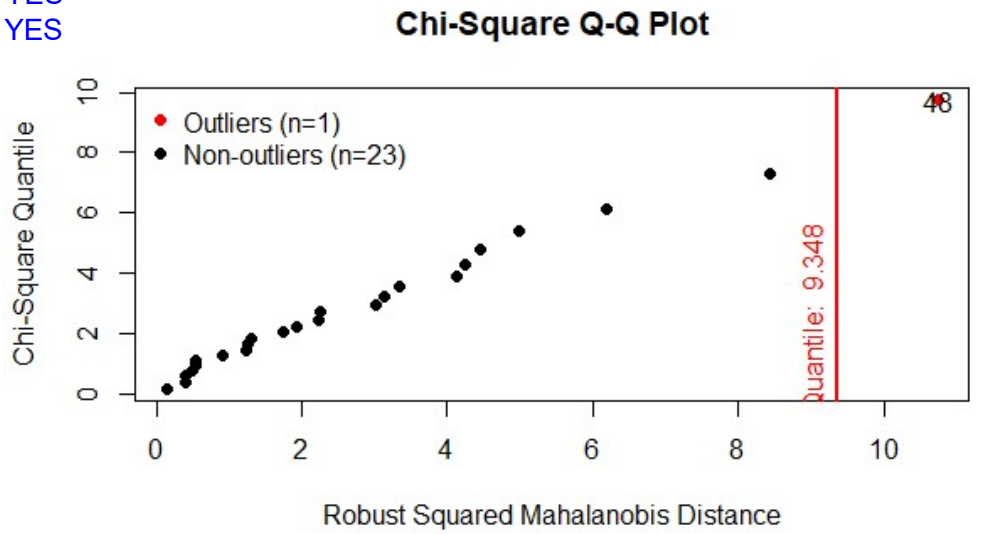
B. Tortugas femeninas

Ho: La muestra de tortugas femeninas proviene de una población con distribución normal



Prueba de Mardia

	Test	Statistic	p value	Result
1	Mardia Skewness	5.82683513935361	0.829595853447198	YES
2	Mardia Kurtosis	-0.184710642782656	0.853455998087885	YES
3	MVN	<NA>	<NA>	YES



Prueba de Royston

	Test	H	p value	MVN
1	Royston	0.4508114	0.6799438	YES

Prueba de igualdad de varianzas

$$H_O: \quad \Sigma_{macho} = \Sigma_{hembra}$$

La matriz de covarianzas de las poblaciones de tortugas machos y hembras son iguales

Prueba de M de Box

Box's M-test for Homogeneity of Covariance Matrices

data: X[, 2:4]
Chi-Sq (approx.) = 23.405, df = 6, **p-value = 0.0006716**

```
# Cargar datos
X = read.delim("clipboard")

#Observar la distribución de los datos
boxplot(x1~sexo,data=X, main="Morfologia Tortugas", xlab="Sexo", ylab="Peso (kgs)")
boxplot(x2~sexo,data=X, main="Morfologia Tortugas", xlab="Sexo", ylab="Patas anteriore (cm)")
boxplot(x3~sexo,data=X, main="Morfologia Tortugas", xlab="Sexo", ylab="Patas posteriores (cm)")

#Instalacion de paquetes prueba de normalidad multivariada
install.packages("MVN")
library(MVN)

#Distribución normal de datos Tortugas masculinas
mvn(X[1:24,2:4],mvnTest="mardia", multivariatePlot = "qq",multivariateOutlierMethod = "quan")
mvn(X[1:24,2:4],mvnTest="royston", multivariatePlot = "qq",multivariateOutlierMethod = "quan")

#Distribución normal de datos Tortugas femeninas
mvn(X[25:48,2:4],mvnTest="mardia", multivariatePlot = "qq",multivariateOutlierMethod = "quan")
mvn(X[25:48,2:4],mvnTest="royston", multivariatePlot = "qq",multivariateOutlierMethod = "quan")

#Prueba de igualdad de varianzas
install.packages("biotools")
library(biotools)

#Prueba M de Box
boxM(X[,2:4],X[,5])
```

MANOVA

En este caso es equivalente a T^2 por solo tener dos poblaciones

```
library(mvtnorm)
MV1 <- manova(cbind(x1,x2,x3)~sexo,data=X)
summary(data, test="Wilks")
summary(data, test="Roy")
summary(data, test="Pillai")
summary(data, test="Hotelling-Lawley")
```

```
 Df   Wilks approx F num Df den Df    Pr(>F)
sexo       1 0.38857   23.078      3     44 3.967e-09 ***
Residuals 46
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(MV1, test="Roy")
         Df    Roy approx F num Df den Df    Pr(>F)
sexo       1 1.5735   23.078      3     44 3.967e-09 ***
Residuals 46
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(MV1, test="Pillai")
         Df  Pillai approx F num Df den Df    Pr(>F)
sexo       1 0.61143   23.078      3     44 3.967e-09 ***
Residuals 46
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(MV1, test="Hotelling-Lawley")
         Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
sexo       1          1.5735   23.078      3     44 3.967e-09 ***
Residuals 46
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```