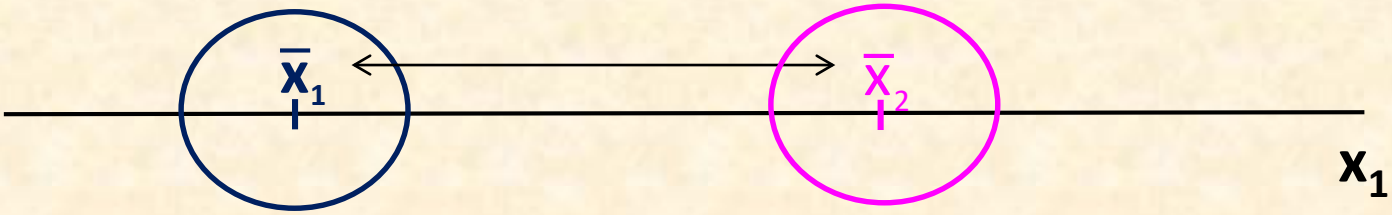
MASTER EN ANÁLISIS AVANZADO DE DATOS MULTIVARIANTES

ANALISIS MULTIVARIANTE DE LA VARIANZA (MANOVA)

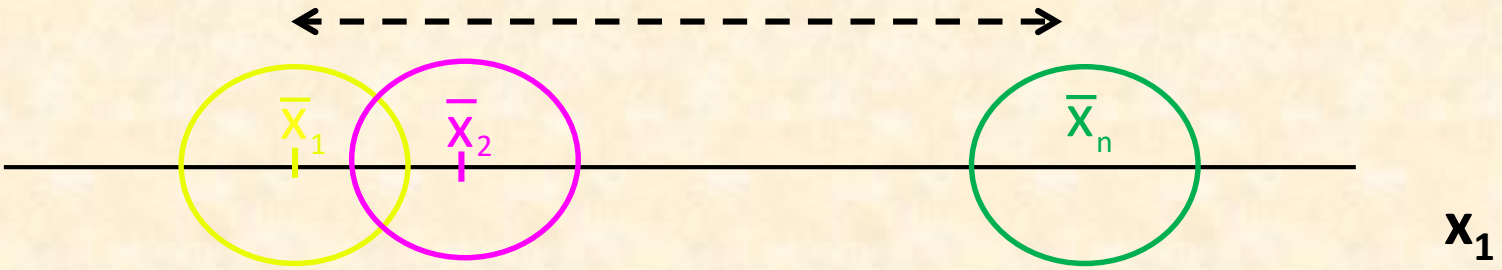
DEPARTAMENTO DE ESTADÍSTICA
UNIVERSIDAD DE SALAMANCA

Estudio de las diferencias entre grupos

Una variable dependiente



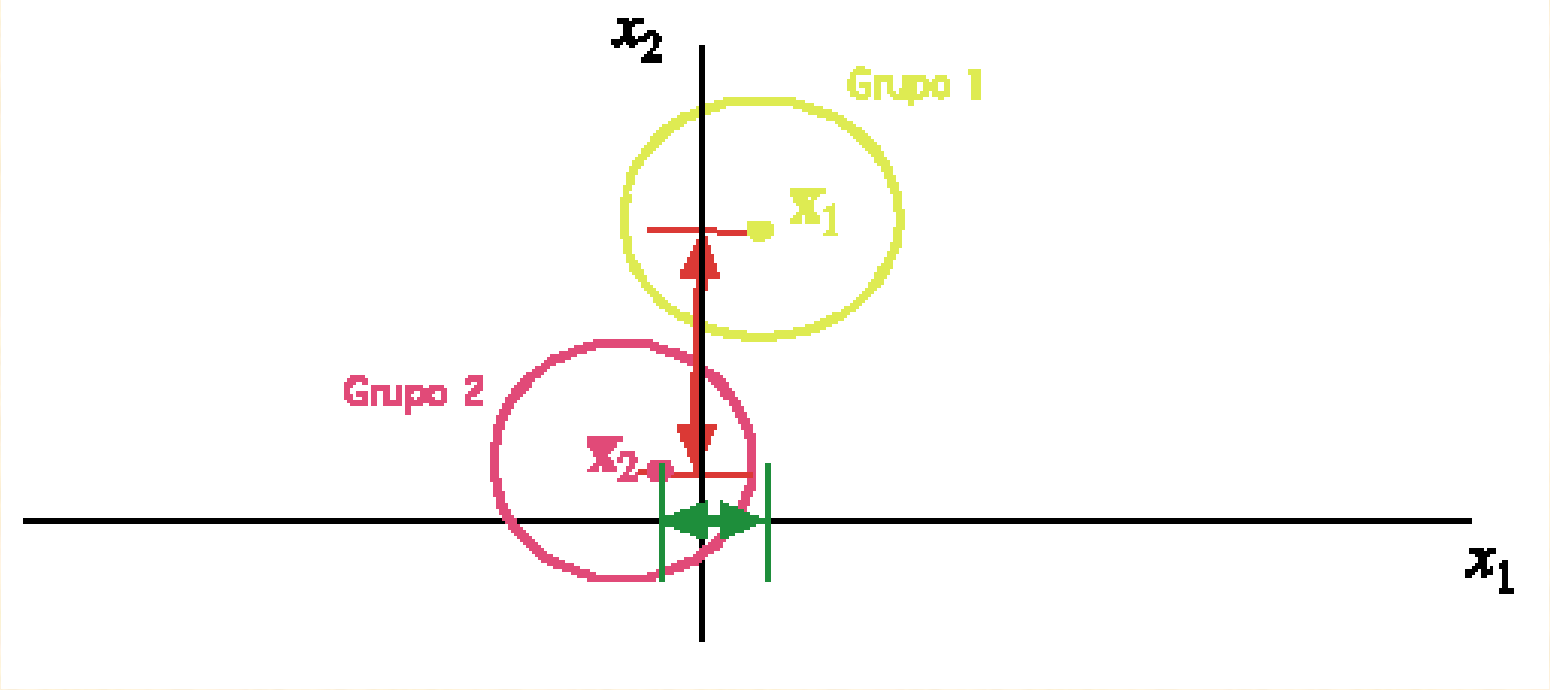
T-Student



ANOVA DE UN FACTOR O MÁS FACTORES
(Interacción, medidas repetidas….)

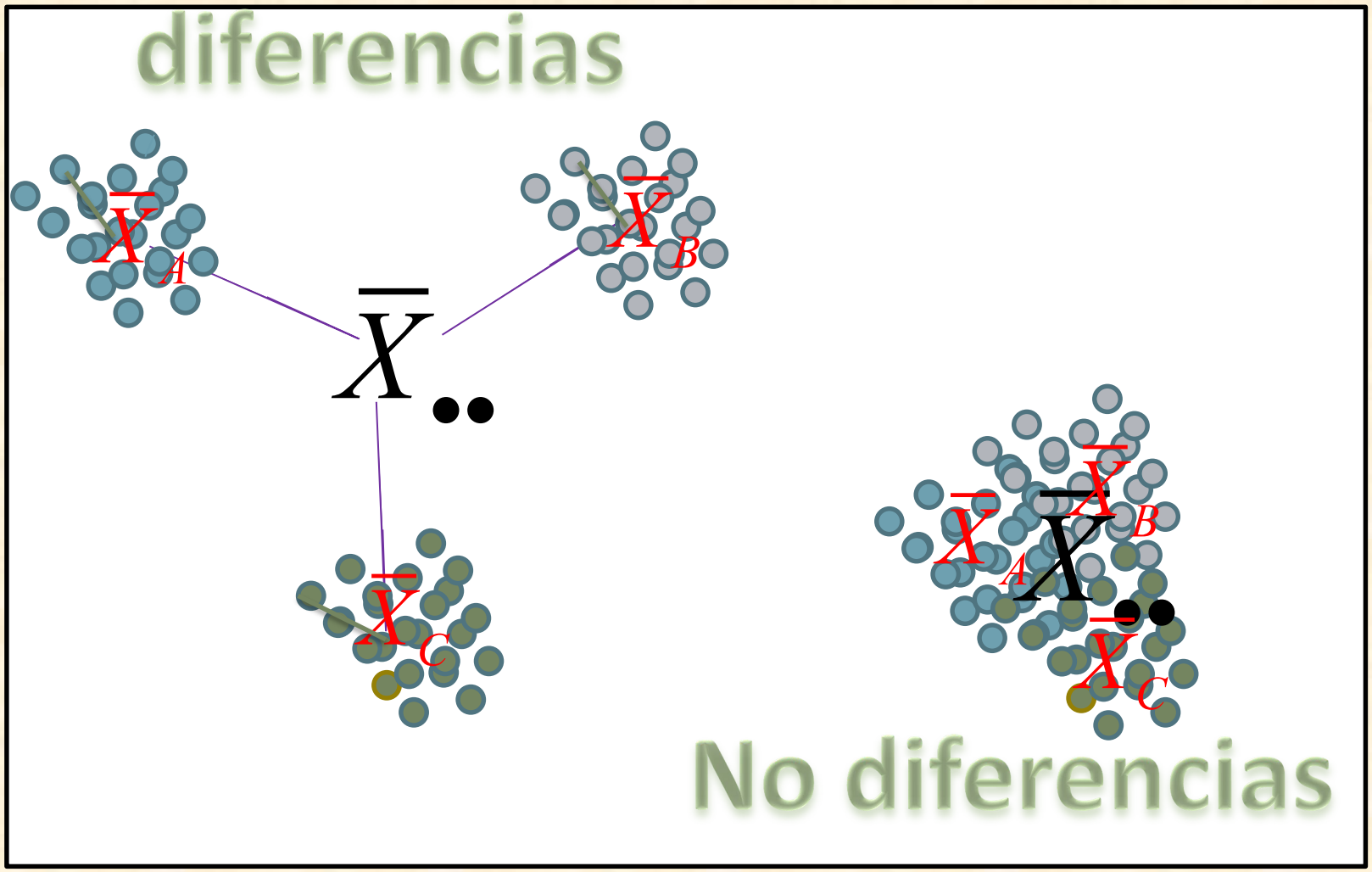
Estudio de las diferencias entre grupos

Dos variables dependientes y dos grupos



Estudio de las diferencias entre grupos

Dos variables dependientes y tres grupos



Estudio de las diferencias entre grupos

MANOVA

La comparación de los g grupos a través de sus vectores de medias.

MÉTODOS RELACIONADOS:

ANÁLISIS DISCRIMINANTE

- Ver las diferencias entre grupos y la clasificación de un individuo en un grupo de entre varios definidos a priori.

ANÁLISIS CANÓNICO DE POBLACIONES

- La representación de la estructura de los grupos en dimensión reducida.

MANOVA BIPLOT

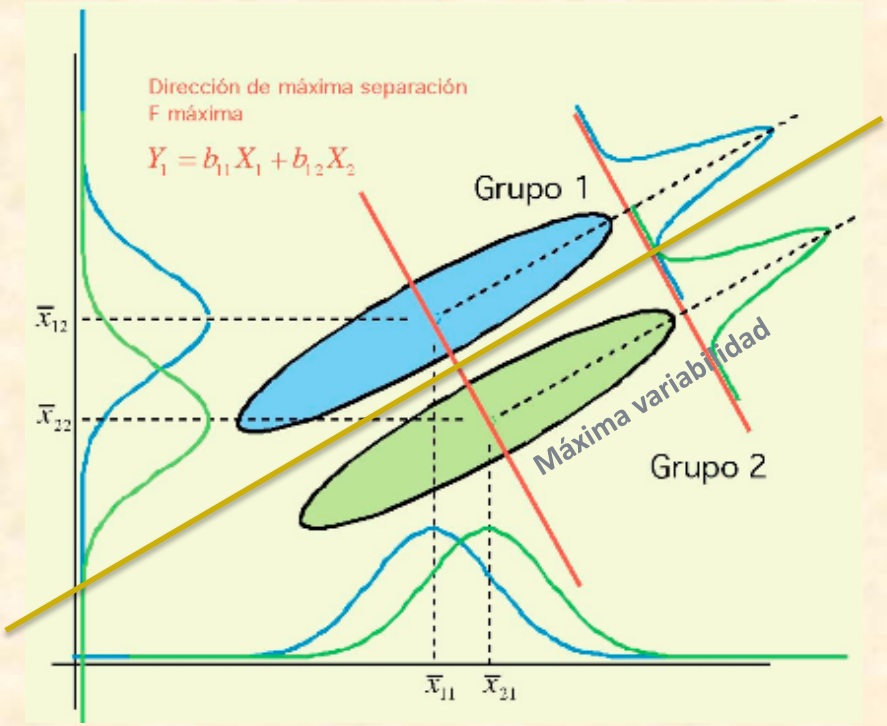
- La comparación de los g grupos y añadir información directa sobre las variables responsables de la separación de los grupos.

Análisis de la estructura de los grupos

OBJETIVO GENERAL

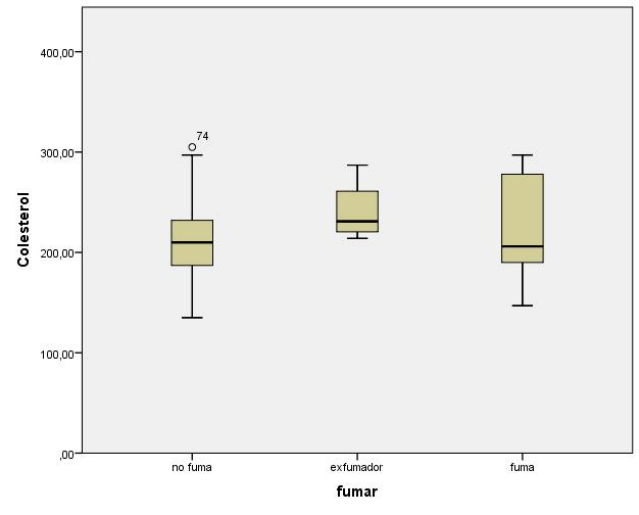
ESTUDIAR LAS DIFERENCIAS ENTRE LOS GRUPOS Y CARACTERIZARLAS MEDIANTE TÉCNICAS MULTIVARIANTES.

Se necesitan un nuevo grupo de técnicas ya que , en general, las Direcciones de máxima variabilidad no coinciden con las direcciones de máxima separación entre grupos.



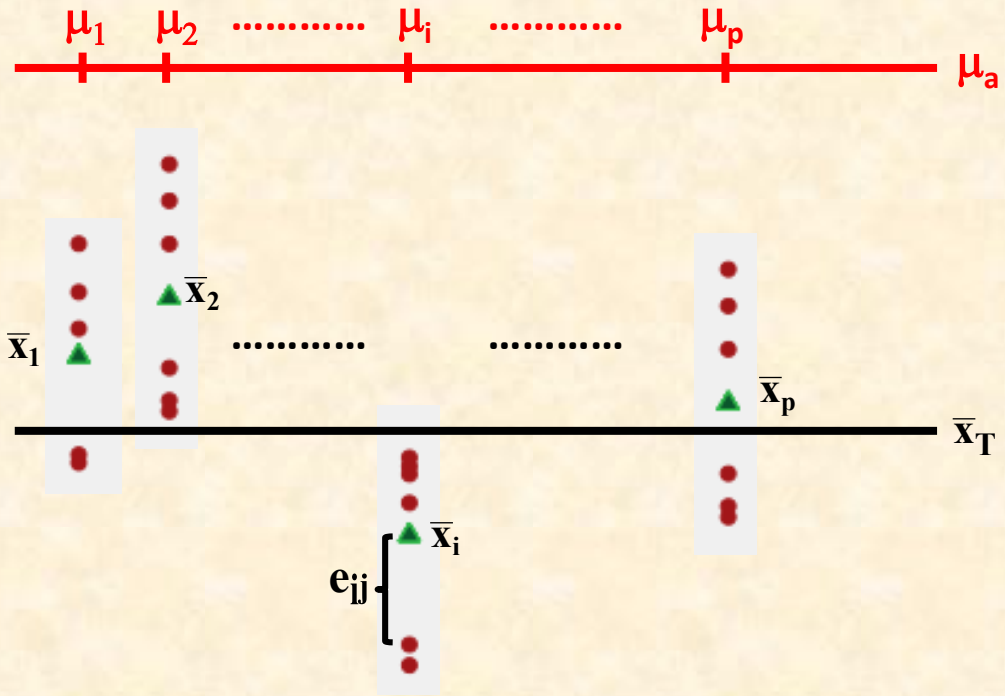
Revisión Análisis de la varianza de un factor

CARACTERÍSTICAS:
- **Independencia**
- **Normalidad**
- **Homocedasticidad**



$$H_0: \mu_1 = \cdots = \mu_i = \cdots = \mu_p = \mu_a$$

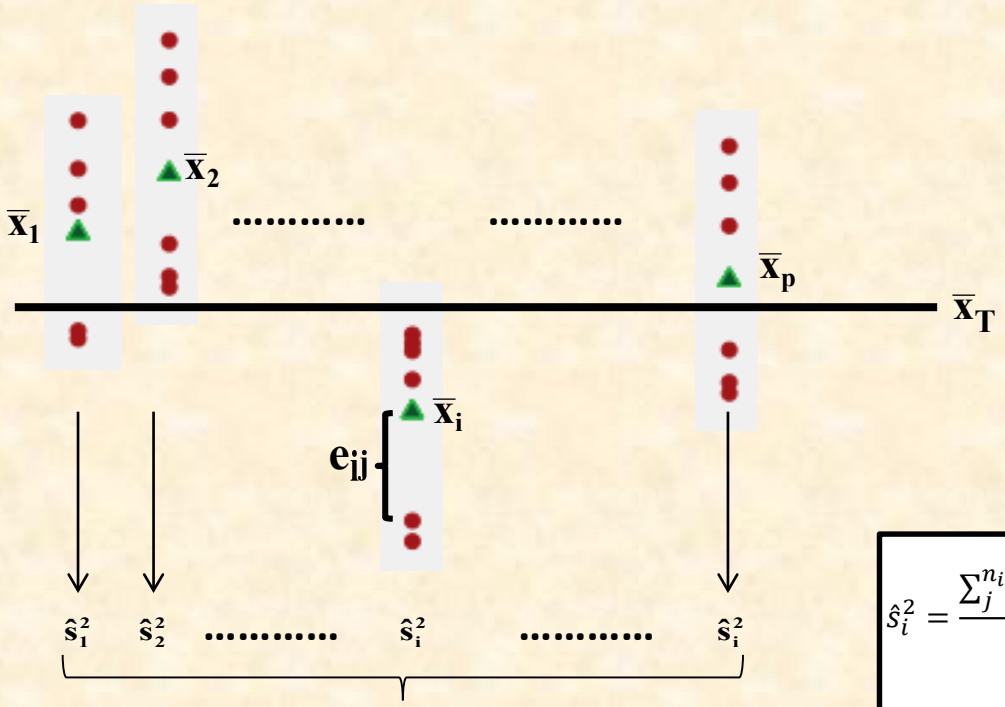
$$H_1: \exists kk' / \ \mu_k \neq \mu_{k'}$$



$$\mathbf{SCT = SCE + SCR}$$

Revisión Análisis de la varianza de un factor

ESTIMADORES VARIABILIDAD DENTRO



$$\hat{s}_i^2 = \frac{\sum_j^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}$$

$$\hat{s}_R^2 = \frac{\sum_i^p w_i \hat{s}_i^2}{\sum_i^p w_i} = \frac{\sum_i^p (n_i - 1)\hat{s}_i^2}{\sum_i^p (n_i - 1)} = \frac{Q_R}{N - p}$$

$$N = n_1 + n_2 + \cdots + n_i + \cdots + n_p$$

Revisión Análisis de la varianza de un factor

ESTIMADORES

$$Q_T = \sum_i^p \sum_j^{n_i} \left(x_{ij} - \bar{x}_T\right)^2 \longrightarrow \hat{s}_T^2 = \frac{Q_T}{N-1}$$

$$Q_E = \sum_i^p n_i \left(\bar{x}_i - \bar{x}_T\right)^2 \longrightarrow \hat{s}_E^2 = \frac{Q_E}{p-1}$$

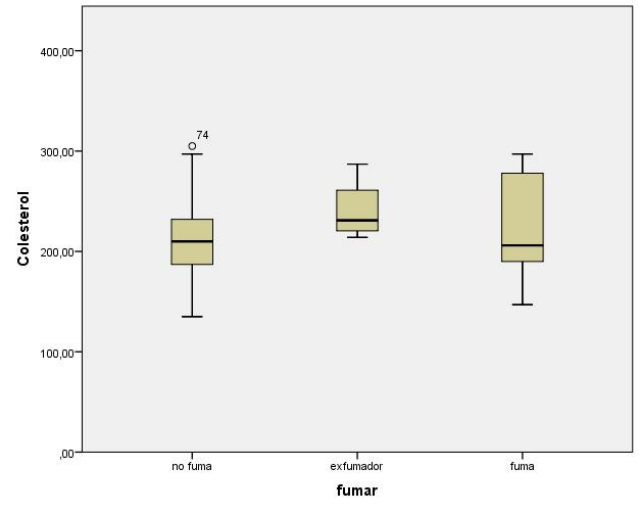
$$Q_R = \sum_i^p \sum_j^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 \longrightarrow \hat{s}_R^2 = \frac{Q_R}{N-p}$$

ESTADISTICO DE CONTRASTE $\longrightarrow$ $F_{p-1,N-p} = \dfrac{S_E^2}{S_R^2}$

Revisión Análisis de la varianza de un factor

CARACTERÍSTICAS:
- **Independencia**
- **Normalidad**
- **homocedasticidad**

$$H_0: \mu_1 = \cdots = \mu_i = \cdots = \mu_p = \mu_a$$
$$H_1: \exists kk' / \ \mu_k \neq \mu_{k'}$$



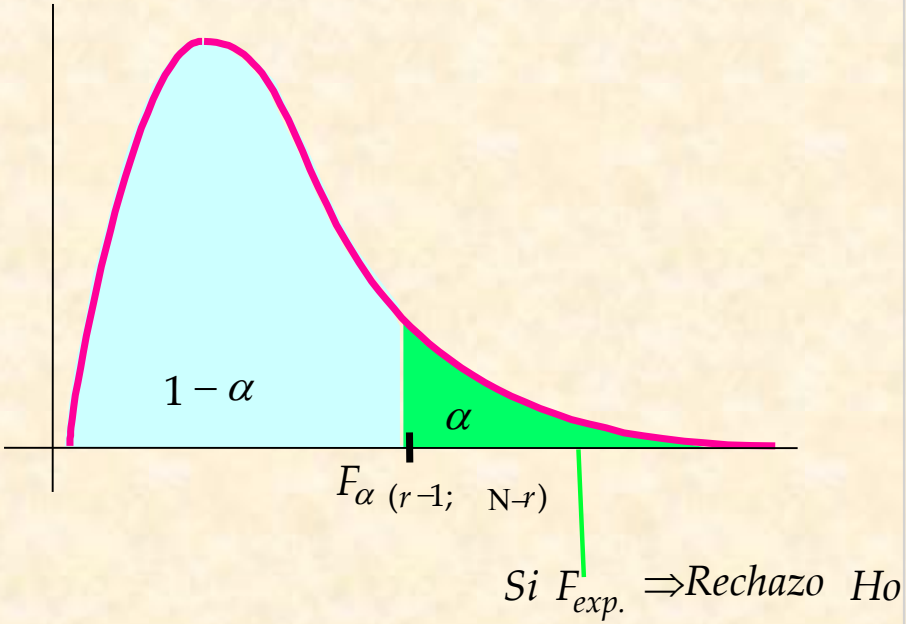
ANOVA				
Fuente	*Sumas de Cuadrados*	*g.l.*	*Estimadores*	*Fexp.*
Entre	$Q_{Entre} = \sum_{i=1}^{r} n_i \left(x_{i\bullet} - \bar{x}\right)^2$	*r-1*	$S^2_{Entre} = Q_{Entre} / r - 1$	$F_{\exp} = \frac{S^2_{Entre}}{S^2_{Dentro}}$
Residual	$Q_{Dentro} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_{ij} - x_{i\bullet}\right)^2$	*N-r*	$S^2_{Dentro} = Q_{Dentro} / N - r$	
Total	$Q_{Total} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}\right)^2$	*N-1*		

Revisión Análisis de la varianza de un factor

Regla de decisión

Si el valor experimental F_{exp} ***supera*** el valor crítico de una F de Snedecor con r-1 y N-r g.l. al nivel de significación elegido, se ***rechazará*** la Ho de igualdad de medias poblacionales y se aceptará la alternativa de que al menos algún par de medias es diferente.

Contraste ***Unilateral superior***



Análisis Multivariante de la Varianza

Este método fue planteado inicialmente mediante trabajos de **Hotelling, Wishart y Wilks** en 1930. **Morrison** explicó de forma clara este método en 1978 y Rencher en **1998** explica bajo su punto de vista qué criterio es mejor en cada caso dependiendo de la hipótesis alternativa. Véase Rao (1952), Scheffé (1959), Searle (1971), Seber (1984) y Huberty (1994) como lecturas complementarias de este método.

Supongamos que tenemos **p variables** observables y **una muestra de tamaño n**.
Los datos los escribimos en forma de matriz a la que llamaremos **X** que previamente han sido **centrados por columnas.**

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots & x_{1p} \\ x_{21} & x_{22} \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \cdots & x_{np} \end{pmatrix}$$

Análisis Multivariante de la Varianza

¿Pórqué utilizar el MANOVA y no "p" ANOVAS individuales?

- El MANOVA, trabaja con todas las variables simultáneamente y trata de encontrar **la combinación lineal** de ellas que nos de una F de Snedecor que **maximice la diferencia entre los grupos** a comparar.

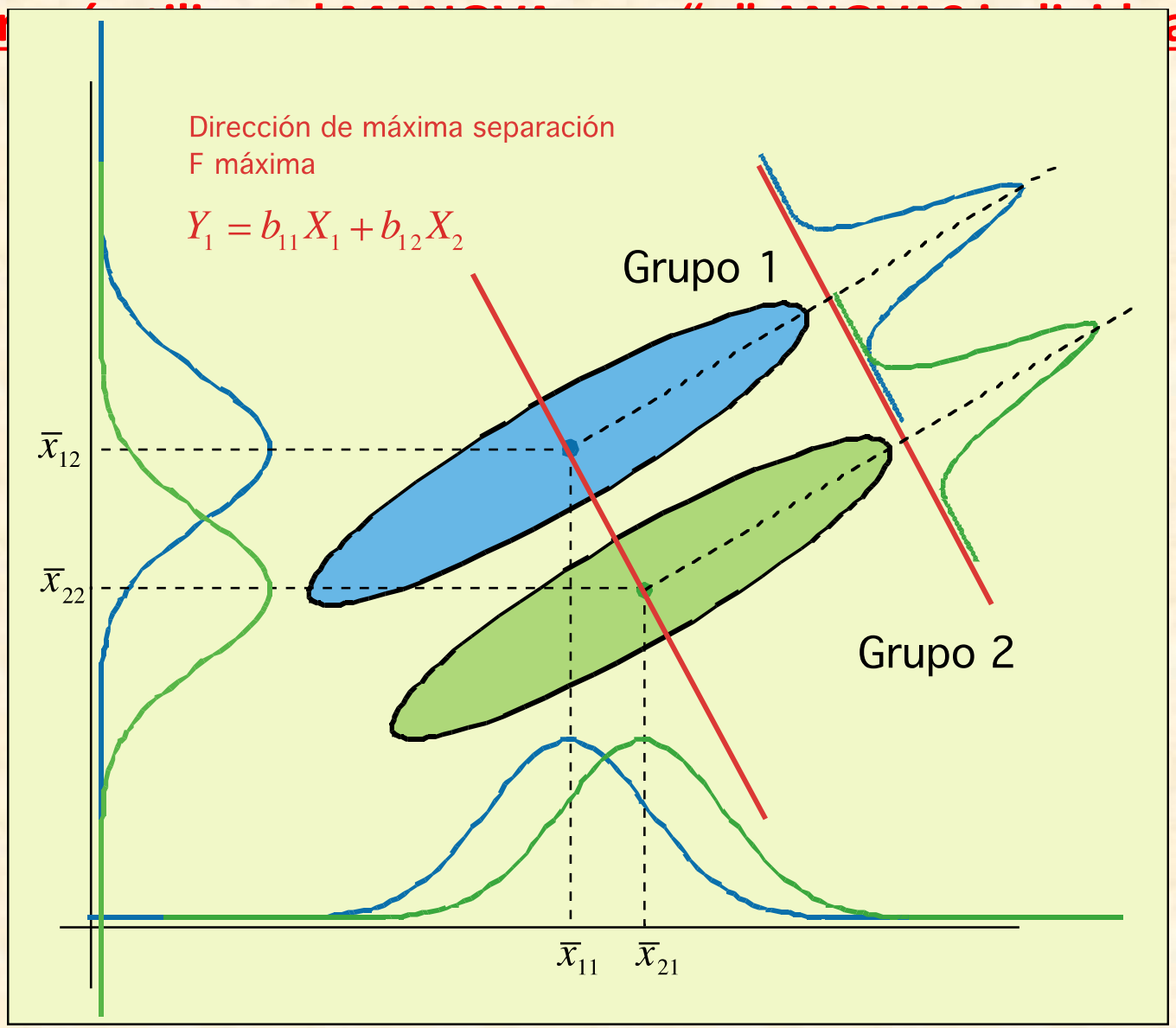
- En el MANOVA **controlamos mejor el riesgo tipo I** cosa que no ocurre cuando hacemos ANOVAS individuales.

- En el MANOVA, al trabajar con la distancia de Mahalanobis, nos permite eliminar la información redundante ya que controla mejor las correlaciones entre las variables.

- Si hacemos ANOVAS individuales puede que en ninguna de las variables encontremos Diferencias, pero que puedan existir diferencias en alguna combinación lineal de ellas.

Análisis Multivariante de la Varianza

¿Pór ales?



Análisis Multivariante de la varianza

La matriz de datos **X** está dividida en **"g" grupos** mutuamente excluyentes de tal forma que la **matriz de vectores de medias** se puede escribir como:

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots & x_{1p} \\ x_{21} & x_{22} \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \cdots & x_{np} \end{pmatrix} \longrightarrow \begin{pmatrix} X_1 = (x_{ij}^1)_{G_1} \\ X_2 = (x_{ij}^2)_{G_2} \\ \vdots \\ X_k = (x_{ij}^k)_{G_{1k}} \\ \vdots \\ X_g = (x_{ij}^g)_{G_g} \end{pmatrix} \longrightarrow \begin{pmatrix} \bar{X}_1 = (\bar{x}_{11}, \ldots, \bar{x}_{1p}) \\ \bar{X}_2 = (\bar{x}_{21}, \ldots, \bar{x}_{2p}) \\ \vdots \\ \bar{X}_k = (\bar{x}_{k1}, \ldots, \bar{x}_{kp}) \\ \vdots \\ \bar{X}_g = (\bar{x}_{g1}, \ldots, \bar{x}_{gp}) \end{pmatrix}$$

Análisis Multivariante de la varianza

MATRIZ DE MEDIAS

$$\bar{X}_1 = (\bar{x}_{11}, \ldots, \bar{x}_{1p}) \quad \vdots \quad \bar{X}_g = (\bar{x}_{g1}, \ldots, \bar{x}_{gp}) \qquad \bar{X} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_g \end{pmatrix} = \begin{pmatrix} (\bar{x}_{11} & \cdots & \bar{x}_{1p}) \\ \vdots & \vdots & \vdots \\ (\bar{x}_{g1} & \cdots & \bar{x}_{gp}) \end{pmatrix}$$

$\bar{x}_{kj}$ = media del grupo **k** en la variable **j**.

MATRIZ DE TAMAÑOS DE LOS GRUPOS

$$\mathbf{D_g} = \begin{bmatrix} \mathbf{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \cdots & \cdots & \mathbf{0} \\ \vdots & \cdots & \mathbf{n_k} & \cdots & \mathbf{0} \\ \vdots & \cdots & \cdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{n_g} \end{bmatrix}$$

Análisis Multivariante de la varianza

DESCOMPOSICIÓN DE LA VARIABILIDAD TOTAL

VARIABILIDAD TOTAL:

$$\mathbf{Q_T = x' x}$$

VARIABILIDAD ENTRE GRUPOS:

$$\mathbf{Q_h = \bar{x}' D_g \bar{x}}$$

VARIABILIDAD RESIDUAL:

$$\mathbf{Q_r = Q_T - Q_h}$$

$$\mathbf{Q_T = Q_h + Q_r}$$

Análisis Multivariante de la varianza

MATRIZ DE COVARIANZAS DENTRO

$$\begin{pmatrix} X_1 = (x_{ij}^1)_{G_1} \\ X_2 = (x_{ij}^2)_{G_2} \\ \vdots \\ X_k = (x_{ij}^k)_{G_1 k} \\ \vdots \\ X_g = (x_{ij}^g)_{G_g} \end{pmatrix} \rightarrow \begin{cases} s_1 = \frac{1}{n_1 - 1} x_1' x_1 \\ s_2 = \frac{1}{n_2 - 1} x_2' x_2 \\ \vdots \\ s_k = \frac{1}{n_k - 1} x_k' x_k \\ \vdots \\ s_g = \frac{1}{n_g - 1} x_g' x_g \end{cases}$$

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

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

MATRIZ DE COVARIANZAS ENTRE LOS GRUPOS

$$\mathbf{H} = \frac{1}{g-1} \bar{x}' D_g \bar{x} = \frac{Q_h}{g-1}$$

Análisis Multivariante de la varianza

$$H_0 = \mu_1 = \mu_2 = \cdots = \mu_k = \cdots = \mu_g = \mu$$
$$H_1 = \exists k \in \{1, \ldots, g\}\ \mu_k \neq \mu$$

Fuentes	Suma de cuadrados	g.l.	Estimadores	Cociente
Entre/ Hipótesis	$\mathbf{Q_h = \bar{x}' D_g \bar{x}}$	$\mathbf{g-1}$	$\mathbf{H = Q_h / g-1}$	$\boldsymbol{H S^{-1}}$
Dentro Error	$\mathbf{Q_r = \sum_{k=1}^{g} (n_k - 1) S_k = Q_T - Q_h}$	$\mathbf{n-g}$	$\mathbf{S = Q_r / n-g}$	
Total	$\mathbf{Q_T = x' x}$	$\mathbf{n-1}$		

El cálculo de la significación del Análisis Multivariante de la Varianza es mucho mas complejo que en el caso del ANOVA

Análisis Multivariante de la varianza

Cálculo de la significación

•En el Manova, al existir varias variables dependientes, surge el **problema de la interpretación de los resultados**.

•En la prueba global se pueden realizar **análisis posteriores** para estudiar la importancia de las variables dependientes. No existe unanimidad en cuanto a las técnicas más apropiadas a emplear.

•Se pueden emplear **técnicas alternativas** para estudiar tanto las diferencias específicas entre niveles de las variables independientes como la importancia de cada una de las variables dependientes.

•La aplicación de estas técnicas alternativas dependerán de las **diferentes razones que puedan tener los investigadores** al utilizar el Manova.

Análisis Multivariante de la varianza

Test Multivariantes

Estos test estadísticos son válidos para muestras de **poblaciones normales** y cuyas **matrices de covarianzas** dentro de los grupos (**S**) son **homogéneas**.
Si la muestra es grande no es necesario que se cumpla el supuesto de normalidad

Roy statistic:

λ_i^2 = La mayor raíz característica de $|H - \lambda S|$ que es el mayor valor propio de HS^{-1}

Lawley y Hotelling:

$$T = \text{traza}\,(HS^{-1}) = \sum_{i=1}^{s} \lambda_i$$

$$s = \text{Min.}\,(g-1,\ p).$$

Análisis Multivariante de la varianza

Test Multivariantes

Wilk's Λ-statistic:

$$\Lambda_{p,\, g-1,\, n-g} = \frac{|S|}{|H+S|} = \\ = |S(H+S)^{-1}| = \prod_{i=1}^{s} \lambda_i$$

Pillai statistic:

$$\mathbf{V = trace}\left[\mathbf{H(H+S)^{-1}}\right] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

Análisis Multivariante de la varianza

Técnicas Estadísticas Posteriores a un Manova Significativo

Cuando se encuentran resultados significativos en la prueba global se pueden realizar análisis posteriores para estudiar la importancia de las variables dependientes.

- **Realizar análisis ANOVA de las variables:** Se basa en el llamado procedimiento de la F protegida (Cramer y Bock, 1966) que mantiene el error global de tipo 1 al nivel nominal especificado.

  -Para explicar las diferencias entre los grupos se tendrán en cuenta solo las variables significativas en el ANOVA.

  -Tintm (1975) propuso un método más conservador para controlar el error de tipo 1, que consiste en dividir el error alfa por el número de variables dependientes.

PROBLEMÁTICA: No tienen en cuenta las correlaciones entre las variables dependientes (multidimensionalidad), lo que indica que cuanto más correlacionadas estén las variables dependientes más sesgadas son las pruebas.

Análisis Multivariante de la varianza

Técnicas Estadísticas Posteriores a un Manova Significativo

•**Hacer un Análisis Discriminante:** Este procedimiento halla las combinaciones lineales de las p variables dependientes que mejor separan los k grupos, maximizando la razón entre la varianza inter y la varianza intra de las combinaciones lineales.

-Una vez obtenidas las combinaciones lineales (funciones discriminantes), también llamadas (funciones canónicas) se interpretan sólo las significativas.

-Se suelen utilizar los coeficientes tipificados de las funciones discriminantes. Estos coeficientes indican la importancia relativa de cada variable.

PROBLEMÁTICA: Al igual que con los coeficientes de la regresión múltiple, las intercorrelaciones entre las variables dependientes afectan en gran medida la interpretación de los coeficientes y hay que tener en cuenta posibles efectos como la multicolinealidad o existencia de variables redundantes.

Análisis Multivariante de la varianza

Condiciones a tener en cuenta en un Manova

- **Normalidad multivariante**: Se establece que todas las variables sigan una distribución normal. Esta condición no es muy robusta.

- **Homocedasticidad:** Se comprueba con la **M de Box** (generalización del test de Barlett para la comprobación de la homogeneidad de varianzas univariantes).

- **Linealidad entre las variables dentro de cada grupo:** Se puede verificar mediante diagramas de dispersión o mediante los coeficientes de correlación lineal de Pearson.

- **Ausencia de Multicolinealidad y Singularidad:** Es necesaria para que sean válidos los resultados si hiciésemos un **Análisis Discriminante**, ya que es necesario invertir las matrices en el cálculo de los coeficientes para las funciones discriminantes y si hay multicolinealidad estos coeficientes son inestables.

Análisis Multivariante de la varianza

Resumen de las Principales características

Este método se utiliza cuando el interés del estudio es la comparación de los grupos a través de sus vectores de medias, es decir, las diferencias entre los vectores de medias mediante contrastes estadísticos.

El MANOVA se puede considerar un Modelo Lineal General Multivariante en el que las variables continuas son las dependientes y los grupos son las regresoras.

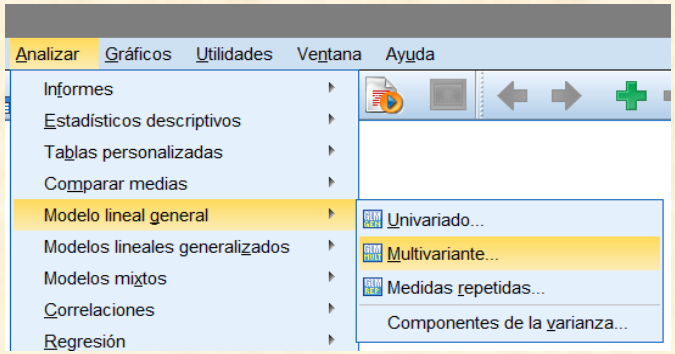
Los resultados se muestran en forma de test estadísticos para el conjunto de todos los grupos y comparaciones por parejas de los mismos o contrastes que incluyan varios grupos.

Trabaja con todas las variables simultáneamente buscando una combinación lineal de éstas que tenga la F de Snedecor univariante máxima, puede que ninguna de las variables originales sea significativa y existir diferencias en una combinación lineal de ellas y elimina la información redundante controlando las correlaciones entre las variables mediante las distancias de Mahalanobis.

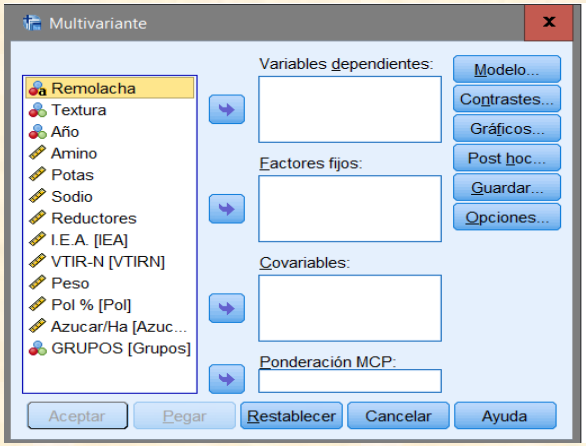
Análisis Multivariante de la varianza

Aplicación con SPSS

1.- Al seleccionar este procedimiento nos aparecerá un menú como el de la figura:



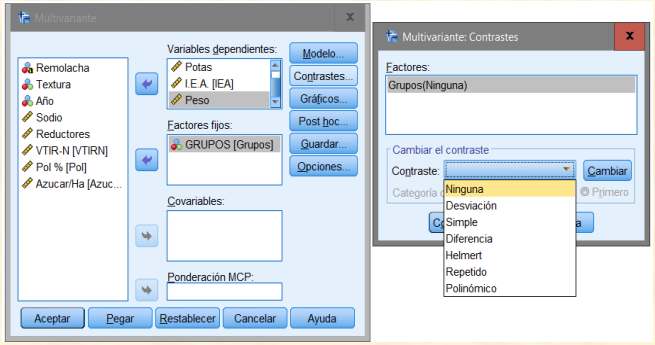
2.- Hemos de introducir las variables y el factor o factores que definen los grupos:



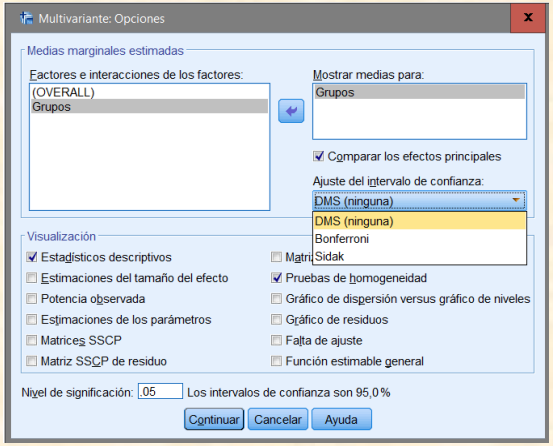
Análisis Multivariante de la varianza

Aplicación con SPSS

3.- Al seleccionar Contrastes:



4.- Al seleccionar Opciones:



Análisis Multivariante de la varianza

Aplicación con SPSS

- **La prueba de Box:**

La hipótesis nula de esta prueba supone que las k matrices de varianza covarianza generadas de los grupos son idénticas a nivel poblacional.

- **Estadísticos de contraste:**

La prueba de cuadro de la igualdad de matrices de covarianzas[a]

M de Box	40,588
F	,670
df1	50
df2	6607,318
Sig.	,964

Pruebas multivariante[a]

Efecto		Valor	F	Gl de hipótesis	gl de error	Sig.
Intersección	Traza de Pillai	,986	1037,682[b]	4,000	57,000	,000
	Lambda de Wilks	,014	1037,682[b]	4,000	57,000	,000
	Traza de Hotelling	72,820	1037,682[b]	4,000	57,000	,000
	Raíz mayor de Roy	72,820	1037,682[b]	4,000	57,000	,000
Grupos	Traza de Pillai	,529	1,828	20,000	240,000	,019
	Lambda de Wilks	,549	1,886	20,000	189,997	,015
	Traza de Hotelling	,687	1,907	20,000	222,000	,013
	Raíz mayor de Roy	,359	4,311[c]	5,000	60,000	,002

a. Diseño : Intersección + Grupos

b. Estadístico exacto

c. El estadístico es un límite superior en F que genera un límite inferior en el nivel de significación.

Análisis Multivariante de la varianza

¿Qué estadístico hemos de usar?

A modo de guía práctica podemos señalar que **si se viola el supuesto de homogeneidad**, lo más adecuado es elegir **la prueba de Pillai**, en caso de que dicho supuesto se verifique deberemos tener en cuenta si estamos ante **una estructura concentrada o difusa**.
Si tenemos dos niveles de tratamiento, todas las pruebas conducen al mismo resultado ya que son equivalentes al estadístico **T^2 de Hotelling**.

Decimos que una estructura es **concentrada** si para explicar la varianza existente entre tratamientos necesitamos pocas combinaciones lineales de las variables dependientes, en el caso contrario nos encontramos con una estructura **difusa**.

Análisis Multivariante de la varianza

¿Cómo distingo la estructura?

Para saber que tipo de estructura presentan nuestros datos es preciso llevar a cabo un análisis de **reducción dimensional**.

Este proceso en SPSS lo hemos de realizar vía sintaxis de la siguiente manera:

A continuación de la orden “MANOVA" debemos introducir las variables dependientes que utilizamos y después de "by" la variable independiente, especificando los dos niveles de tratamiento extremos.

MANOVA V1..... VP by Grupos(1,g)/
print signif(dimenr).

Dicho análisis nos permite conocer cuantas combinaciones lineales de las variables dependientes son necesarias para explicar toda la variabilidad existente en la matriz SCPC de la hipótesis.

Análisis Multivariante de la varianza

Procedimiento de reducción de dimensionalidad

- En primer lugar contrasta la hipótesis nula del análisis multivariante mediante la prueba de Wilks, en el caso de rechazarla ello implica que al menos existe una combinación lineal de las variables dependientes que explica una porción significativa de varianza entre grupos, a partir de ello elimina la varianza explicada por dicha combinación y aplica de nuevo la prueba a la varianza entre grupos residual.

- Si la prueba continúa siendo significativa implica que como mínimo existe otra combinación lineal, la cual es extraída, continuando el proceso hasta que llegue un momento en que aceptemos la hipótesis nula.

Dimension Reduction Analysis

Roots	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
1 TO 4	,54851	1,88578	20,00	190,00	**,015**
2 TO 4	,74553	1,50395	12,00	153,75	,128
3 TO 4	,95606	,44689	6,00	118,00	,846
4 TO 4	,98658	,40797	2,00	60,00	,667

Análisis Multivariante de la varianza

Procedimiento de reducción de dimensionalidad

Dimension Reduction Analysis

Roots	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
1 TO 4	,54851	1,88578	20,00	190,00	**,015**
2 TO 4	,74553	1,50395	12,00	153,75	,128
3 TO 4	,95606	,44689	6,00	118,00	,846
4 TO 4	,98658	,40797	2,00	60,00	,667

Como vemos la prueba rechaza la hipótesis nula global en primera instancia para, una vez extraída la primera combinación aceptar las hipótesis nulas.

Existe una sola combinación lineal de las variables dependientes significativa y, que por lo tanto, la estructura es concentrada.

En este caso lo más recomendable es utilizar la prueba de **Pillai o la de Hotelling**.

Si la estructura es difusa es preferible utilizar la prueba de **Wilks.**

Algunos autores prefieren utilizar la **prueba de Roy** como la más adecuada ante estructuras concentradas.

No existe un estadístico exacto de contraste para esta prueba con lo que los valores que nos proporciona son aproximados.