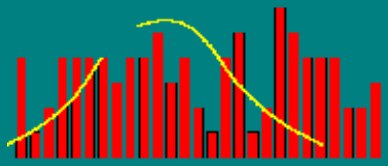
ANOVA



ANÁLISIS DE LA VARIANZA (UN FACTOR)

POR: JUAN MARTÍNEZ de LEJARZA ESPARDUCER

ir a HipEstat (hipertexto de Estadística)

UNIVERSIDAD DE VALENCIA

DEPARTAMENTO DE ECONOMÍA APLICADA

"LECCIÓN EN FORMATO HTML"

pulsar para ver la página numerada que se desea

PLANTEAMIENTO TEÓRICO

EJEMPLO PARA LA COMPRESIÓN DE LA SITUACIÓN:

Supongamos una población de las notas $y_{i.j}$ de un universo de 9 alumnos de tres grupos distintos , así:

grupo 1	grupo 2	grupo 3
5	5	5
5	5	5
5	5	5

evidentemente en este caso la media global es 5 y la de cada grupo también $\mathbf{y_{i.j} = \mu}$ cada valor es igual a la media general. NO HAY DIFERENCIAS ENTRE GRUPOS NI DENTRO DE LOS GRUPOS

Supongamos que aplicamos un método de enseñanza (factor) que afecta : subiendo las notas del grupo 1 en 1 punto , las del grupo dos en 2 puntos y no modificando las del grupo 3 . Así.

grupo 1	grupo 2	grupo 3
5+1=6	5+2=7	5
5+1=6	5+2=7	5
5+1=6	5+2=7	5

ahora la nota de un alumno sería $\mathbf{y_{i.j} = \mu + \alpha_j}$ en el que los α_j son 1, 2, 0 los EFECTOS QUE PRODUCEN EL FACTOR (MÉTODO) EN CADA NIVEL . PARECE CLARO QUE EL FACTOR INFLUYE EN ESTABLECER DIFERENCIAS ENTRE GRUPOS; PERO NO DENTRO

EJEMPLO PARA LA COMPRESIÓN DE LA SITUACIÓN II

Lo más habitual es que haya alumnos que rindan más que otros (por diversas razones aleatorias o que en principio no dependan de un factor) ,son por tanto, comportamientos aleatorios individuales que denominamos $\varepsilon_{i,i}$ implantando algunos en el ejemplo ,sería:

grupo 1	grupo 2	grupo 3
5+1-1=5	5+2+2=9	5+0+3=8
5+1-2=4	5+2+0=7	5+0+4=9
5+1+0=6	5+2+1=8	5+0+0=5

en el que los efectos aleatorios $\varepsilon_{i,i}$ serían -1,-2,0,2,0,1,3,4,0 que fomentan la variablidad **dentro** de los grupos INTRA-GRUPOS

Entonces para cada valor tendremos el modelo $y_{i,i} = \mu + \alpha_i + \varepsilon_{i,i}$

TENEMOS DOS TIPOS DE VARIABILIDAD : LA ENTRE GRUPOS (DEBIDA AL FACTOR) Y LA INTRA GRUPOS (DEBIDA A LA ALEATORIEDAD....) .

PARA PODER AFIRMAR QUE EL FACTOR PRODUCE EFECTOS. LA VARIABILIDAD ENTRE LOS GRUPOS HA DE SER *SIGNIFICATIVAMENTE* GRANDE RESPECTO A LA INTRA GRUPOS

Sea Y una variable aleatoria sobre la que se han tomado N muestras ;de manera que obtenemos k muestras correspondientes a las k categorías(niveles) del factor. Si el tamaño de la muestra para cada categoría es el mismo (n) para todas, estaremos ante un modelo **BALANCEADO** en el que $N = nk$ Y sigue una $N(\mu_i, \sigma)$ para $i = 1,2,3,\ldots,k$.

	NIVELES DEL FACTOR			
	1	2	....	k
1	$Y_{1.1}$	$Y_{2.1}$	.....	$Y_{k.1}$
2	$Y_{1.2}$	$Y_{2.2}$	.....	$Y_{k.2}$
j	$Y_{1.i}$	$Y_{2.i}$ ..... $Y_{i.i}$		$Y_{k.i}$
n	$Y_{1.n}$	$Y_{2.n}$	.....	$Y_{k.n}$

$i = 1,2,3,\ldots.k$ $j = 1,2,3,\ldots.n$ (en el caso de balanceado)

media muestral correspondiente al nivel i del factor $= 1/n \sum_{j=1}^{n} Y_{ij} = \bar{y}_{i.}$

media general $= 1/N \sum_{i=1}^{k} \sum_{j=1}^{n} Y_{ij} = \bar{y}$

MODELO E HIPÓTESIS

EL MODELO SERÍA EL SIGUIENTE : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

donde :

Y_{ij} = observación j-ésima del nivel i μ = media general

α_i = efecto el i-ésimo nivel del factor ε_{ij} = error aleatorio independiente N(0,σ)

PANTEAMOS LA SIGUIENTE HIPÓTESIS NULA

$H_O = \alpha_1 = \alpha_2 = \alpha_3 = ... = \alpha_\kappa$

o bien si consideramos $\mu_i = \mu + \alpha_i$

entonces $H_o = \mu_1 = \mu_2 = \mu_3 = ...= \mu_\kappa$

en definitiva se quiere comprobar la **no influencia** del factor α

DE OTRA FORMA : SI TODAS LAS MUESTRAS PROCEDEN DE LA MISMA POBLACIÓN

CONDICIONES GENERALES DE APLICACIÓN.

A- **INDEPENDENCIA DE LOS ERRORES** . Los errores experimentales han de ser independientes . Se consigue si los sujetos son asignados aleatoriamente. Es decir , se consigue esta condición si las elementos de los diversos grupos han sido elegidos ***por muestreo aleatorio*** .

B- **NORMALIDAD** . Se supone que los errores experimentales se distribuyen normalmente.Lo que supone que cada una de las puntuaciones $y_{i.i}$ se distribuirá normalmente . Para comprobarlo se puede aplicar un test de ajuste a la distribución Normal como el de **Kolmogov-Smirnov**

C- **HOMOGENEIDAD DE VARIANZAS (HOMOSCEDASTICIDAD).** La varianza de los subgrupos ha de ser homogénea $\sigma^2_1 = \sigma^2_2 = ...= \sigma^2_\kappa$ ya que están debidas al error. Se comprobarán mediante los test de : Razón de varianzas (máx/mín) , C de Cochran , Barlett-Box,

DESCOMPOSICIÓN DE LA VARIABILIDAD TOTAL

que quedaría establecida de la siguiente forma:

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

$$\text{S.C.T.} = \text{S.C.I.} + \text{S.C.E.}$$

Donde : S.C.T. = SUMA DE CUADRADOS TOTAL

S.C.I. = SUMA DE CUADRADOS INTRA-GRUPOS (within-groups)

S.C.E. = SUMA DE CUADRADOS ENTRE-GRUPOS (between-groups)

En el caso de NO ser BALANCEADO , no habría una solo n sino k distintos valores.

la descomposición quedaría entonces así:

$$\sum\sum (Y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{i=1}^{n_i} (Y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y})^2$$

$$\text{S.C.T.} = \text{S.C.I.} + \text{S.C.E.}$$

S.C.T. será la suma de las desviciones de cada $Y_{..}$ respecto a la media general $\bar{y}$

S.C.I. será la variación total de las observaciones alrededor de cada una de las medias muestrales

S.C.E. será la variación de las medias muestrales de cada grupo alrededor de la media general

SI DIVIDIMOS LA DESCOMPOSICIÓN DE LA VARIABILIDAD POR LA VARIANZA

$$\frac{S.C.T.}{\sigma^2} = \frac{S.C.I.}{\sigma^2} + \frac{S.C.E.}{\sigma^2}$$

Dado que las observaciones de las k muestras son independientes , cada uno de los tres sumatorios desde i = 1 hasta i = k es la suma de k variables aleatorias que tienen distribuciones χ^2

de tal manera que

$S.C.T. / \sigma^2 \longrightarrow \chi^2$ con (N-1) grados de libertad

$S.C.I. / \sigma^2 \longrightarrow \chi^2$ con k(n-1) grados de libertad

$S.C.E. / \sigma^2 \longrightarrow \chi^2$ con (k-1) grados de libertad

en el caso de diseño NO BALANCEADO $S.C.I. / \sigma^2 \longrightarrow \chi^2$ con $(n_1-1)+(n_2-1)...(n_k-1)$ grados de libertad

Bajo la hipótesis nula; la variabilidad entre grupos no deberá superar significativamente a la variabilidad intra grupo luego el cociente **S.C.E / S.C.I. no deberá ser significativamente grande** . Esa ***Significabilidad*** nos la dará una distribución de probabilidad asociada. Así:

$$\frac{S.C.E.}{S.C.I.} \longrightarrow \frac{\chi^2_{(k-1)}}{\chi^2_{k(n-1)} \text{ ó bien } \chi^2_{(N-k)}}$$

si multiplicamos el conciente por N-k / k-1

$$\frac{S.C.E. \cdot (N-k)}{S.C.I \cdot (k-1)} \longrightarrow = \frac{\chi^2_{(k-1)}/(k-1)}{\chi^2_{(N-k)}/(N-k)} \longrightarrow F_{(k-1),(N-k)} \Rightarrow \mathbf{F}$$

Así para un nivel de significación **α**

Si $\mathbf{F} > F_\alpha$ se rechaza la H_0 medias de los grupos no son iguales .

Si $\mathbf{F} < F_\alpha$ no se puede rechazar la H_0 ; luego medias de los grupos son iguales

no influye el factor

no hay diferencias significativa entre los niveles

TABLA " ANOVA " DE UN FACTOR .

FUENTE DE VARIACIÓN	G.L	SUMA DE CUADRADOS	$F_{(k-1)(N-k)}$
ENTRE-GRUPOS	k-1	S .C. E.	
INTRA-GRUPOS	N-1 $(n_1-1)+(n_2-1)+.(n_k-1)$ no balanceado	S. C. I.	
TOTAL		S. C. T.	$F = \frac{S.C.I\ (N-k)}{S.C.E\ (k-1)}$

COMPARACIONES MULTIPLES ENTRE MEDIAS.

Si se ha **rechazado la hipótesis nula** de igualdad de medias esto supone : EXISTE AL MENOS UNA DIFERENCIA. . PERO NO SABEMOS CUÁNTAS,NI ENTRE QUE MEDIAS.

Para ello se establecen pruebas de comparación múltiple (a posteriori) ; aquí vamos a ver dos la de TUKEY y la de SCHEFFE.

PRUEBA DE TUKEY (Honestly Significant Difference).

A) Se estiman las diferencias entre las medias de los grupos (Ψ). Así

	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_k$
$\overline{X}_1$	.	$\Psi = \overline{X}_1 - \overline{X}_2$	$\Psi = \overline{X}_1 - \overline{X}_k$
$\overline{X}_2$	.	.	$\Psi = \overline{X}_2 - \overline{X}_k$
$\overline{X}_k$	.	.	.

B) Se calculará la desviacione típica según Tukey

$$\sigma_\Psi = \sqrt{\frac{\text{M.C.I (media cuadrática I.}}{\text{n.}}}$$

C) Los cocientes entre los diversos Ψ_i / σ_Ψ se compararán con la tabla de Tukey que nos indicará que diferencias son significativamente distintas para un α prefijado

PRUEBA DE SCHEFFÈ

Está más generalizada y **es más** conservadora que la de Tukey

. Se realiza de la siguiente manera.

A) Se estiman las diferencias de medias (como en el caso Tukey).

	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_k$
$\overline{X}_1$	.	$\Psi = \overline{X}_1 - \overline{X}_2$	$\Psi = \overline{X}_1 - \overline{X}_k$
$\overline{X}_2$	.	.	$\Psi = \overline{X}_2 - \overline{X}_k$
$\overline{X}_k$	.	.	.

B) Se elabora la desviación típica $S_j = \sqrt{\text{M.C.I.} .(1/n_j + 1/n_{j'})}$.

siendo n_j = número de observaciones uno de los grupos que forman la diferencia de medias

siendo $n_{j'}$ = número de observaciones del otro grupo que forma la diferencia de medias

C) Se divide cada diferencia por su correspondiente desviación típica . Cada valor resultante se compara con una $F_{\alpha, (k-1), (N-k)}$ ó $(n_1-1)+(n_2-1)+..+(n_k-1)$ g.l.

METODOLOGÍA INFORMÁTICA CON SPSS/PC+

ONEWAY VAR = *(VARIABLE DEPENDIENTE)* BY *(FACTOR)* (MÍNIMO,MÁXIMO).

Mediante esta formulación obtendremos una tabla ANOVA de la variable dependiente y el factor entre los valores (niveles) que le hayamos planteado . Se podrá conocer si rechazamos o no la hipótesis de igualdad de medias.

ONEWAY VAR=*(VARIABLE DEPENDIENTE)*BY*(FACTOR)* (MÍNIMO,MÁXIMO)/

STATISTICS=3.

Mediante la inclusión de "STATISTICS=3" Conseguimos verificar o no la hipótesis de Homoscedasticidad mediante los test de Cochran , Barlett y F máxima de Hartley

ONEWAY VAR = *(VARIABLE DEPENDIENTE)* BY *(FACTOR)* (MÍNIMO,MÁXIMO)/

/RANGES=*(nombre del procedimiento).*

Mediante la inclusión de "RANGES= ...." conseguimos las comparaciones multiples entre medias una vez se ha rechazado la hipótesis de igualdad de medias

PRUEBAS DE COMPARACIONES MULTIPLES CON SPSS/PC+

YA SE VIO LA APLICACIÓN DEL COMANDO

.....RANGES = *nombre del procedimiento*

Los procedimientos y sus comandos son:

PRUEBA	nombre del procedimiento
Student-Newman-Keuls	SNK
Tukey	TUKEY
B de Tukey	BTUKEY
diferencia mínima significativa	LSD(p)
diferencia mínima significativa modificada	MODLSD(p)
recorrido multiple de Duncan	DUNCAN(p)
comparaciones de Scheffè	SCHFFÈ(p)

(p) donde p es el nivel de significación explicitado : 0.01 , 0.05 , 0.1