Acerca del análisis de varianza

1. Es un modelo de análisis de covariación asimétrico, para variable independiente nominal (obviamente también ordinales o intervalares) sin limitación del número de categorías y variable dependiente intervalar. Las submuestras de la variable dependiente determinadas por cada categoría de la independiente (también conocidas como los "grupos" de la dependiente) deben ser seleccionadas de manera aleatoria e independiente entre grupos. La distribución de la dependiente en cada grupo debe ser normal en la población, y la varianza intragrupal debe ser igual para todos los grupos en la población (homocedasticidad). El modelo no requiere relación lineal entre las variables, y esta es una diferencia importante con el modelo de regresión–correlación de Pearson.

1.1. El test de Levene, aplicado en el práctico de método III, es un análisis de varianza especial para estimar si se cumple homocedasticidad para un determinado nivel de significación (ver explicación adjunta).

2. En un archivo Excel adjunto se muestra en un ejemplo el esquema del modelo cuando se disponen los datos para cálculos manuales. El trabajo práctico previsto en el curso consiste en la resolución de otro ejemplo a partir de los datos ya procesados por el programa SPSS, pero es necesario que como parte de la teoría del modelo conozcamos el proceso de cálculo y sus fundamentos.

3. Se puede ver en el esquema que los valores de la variable dependiente son clasificados por las categorías de la independiente. La distribución de estos valores en cada grupo son representados en los cálculos globales por sus respectivas medias, esto es por las medias de la submuestra de la variable dependiente para una determinada categoría de la independiente. También se calcula una media total, que es la media de toda la distribución de la variable dependiente sin tomar en cuenta su relación con la independiente. Esta media total cumple una función similar a la del marginal de la dependiente en los cuadros bivariados.

4. Si todas las medias de grupo fueran iguales entre sí, la incidencia de la independiente en la dependiente sería nula, sería un caso de independencia estadística, ya que (en el ejemplo) sería indiferente para la vida útil de las piezas, que hayan sido fabricadas en uno u otro turno. En consecuencia, en este modelo la hipótesis nula se expresa así: $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$. La letra griega μ (mu) es el símbolo que indica valores de medias en la población. La letra k indica la cantidad de grupos (en nuestro caso k = 3).

5. La influencia de la independiente sobre la dependiente será mayor cuanto más difieran entre sí las medias de grupo. Si se logra medir cúanto difieren tendremos una medida de la covariación. Para ello partimos de considerar que las k medias (3 en nuestro caso) forman una distribución de valores como cualquier otra. Son k valores (3 en nuestro caso) cuya media es la media total, y de lo que se trata es de medir la dispersión de estos k valores en torno de la media total. Una medida adecuada sería la varianza, cuya expresión general es $s^2 = \Sigma(x - \bar{x})^2 / n$. Por sencillas razones matemáticas que aquí omitiremos es preferible utilizar otra medida de dispersión, que tiene la misma expresión de la varianza, salvo que no está normalizada por n: es la variación o también llamada 'suma de cuadrados': $SC = \Sigma(x - \bar{x})^2$. Aplicada a nuestro caso y ponderando adecuadamente dado que cada media de grupo representa a una cantidad de casos que puede diferir de un grupo a otro, resulta: $SCE = \Sigma(\bar{x}_k - \bar{x}_T)^2 . * n_k$, donde SCE es la 'suma de cuadrados entre', k en nuestro ejemplo toma sucesivamente los valores 1, 2 y 3, $\bar{x}_k$ es la media del grupo respectivo, x_T es la media total y n_k es la cantidad de caso del respectivo grupo. No es difícil comprender que la SCE es la variación explicada por la variable independiente.

6. Para lograr una medida normalizada de correlación se divide la variación explicada por la total. Esta última mide la dispersión de cada valor individual de la variable dependiente (distinguir de la distribución de medias): $SCT = \Sigma(X - \bar{X})^2$. El coeficiente $Eta^2 = SCE / SCT$ mide la proporción de la variación de la variable dependiente que es explicada por la variación de la independiente. Puede expresarse esto mismo en porcentajes multiplicando por 100 el valor de Eta^2. Extrayéndole la raíz cuadrada se obtiene el coeficiente Eta, que no mide proporción de explicación sino la fuerza de la relación entre ambas variables. Eta y Eta^2 no pueden ser negativos y varían entre 0 y 1.

7. La prueba de significación correspondiente al modelo puede comprenderse fácilmente si se tiene en cuenta que todas las pruebas de significación que estamos estudiando, además de tener un núcleo conceptual común, también coinciden esencialmente en su operatoria y difieren en la forma específica que toma la hipótesis nula y en el estadístico de contraste aplicado. Simplificando, el estadístico de contraste es un instrumento especial en función del cual pueden compararse dos valores: uno calculado a partir de datos empíricos y otro teórico del cual se conoce su distribución de muestreo que ha sido volcada en tablas de sencilla utilización. Entonces podremos aprovechar los conocimientos que ya hemos adquirido sobre la prueba de significación con el estadístico de prueba χ^2 para relaciones entre dos variables nominales u ordinales. En el modelo de análisis de varianza el estadístico de prueba es F de Snedecor, siendo $F = (SCE / (k - 1)) / (SCD / (n - k))$.

8. SCD es la 'suma de cuadrados dentro', o sea la variación de los valores de la dependiente respecto de la media de su propio grupo, luego sumadas acumulativamente para todos los grupos. Puede verse sin mayor dificultad que la SCD es la varianza inexplicada por la variable independiente. Su expresión teórica es $SCD = \Sigma(X - \bar{X}_k)^2$, aunque su cálculo puede simplificarse porque las variaciones se relacionan entre sí según la expresión $SCT = SCE + SCD$ (esta relación es válida para las variaciones y no para las varianzas, y es ésta la razón por la que todo el procedimiento aplicado hasta ahora está referido a variaciones). Entonces puede calcularse $SCD = SCT - SCE$.

9. El valor teórico de F se obtiene entrando a las tablas respectivas, en primer término por el nivel de significación adoptado a priori, y en la página correspondiente por dos tipos de grados de libertad: en la cabeza de la tabla (n_1), elegir la columna correspondiente a los grados de libertad asociados a la SCE, que es k-1, y en las filas (n_2) elegir la correspondiente a los grados de libertad asociados a la SCD, que es n-k. En la intersección se encuentra F teórico. Se rechaza la hipótesis nula si $F_e \geq F_t$. Si se dispone de un programa informático (como en la guía de trabajos prácticos) también, como en otras pruebas de significación, se pueden comparar los p-valores: si $p_e \leq p_t$ se rechaza la hipótesis nula.

10. Este texto trata de reseñar conceptos mínimos, no es exhaustivo y no reemplaza a la bibliografía obligatoria.

EJEMPLO DE ANALISIS DE VARIANZA

Vida útil de piezas de aluminio según turno de fabricación

VIDA ÚTIL DE LAS PIEZAS DE ALUMINIO (días)

TURNO

MAÑANA	$x - x_1$	$(x - x_1)^2$	TARDE	$x - x_2$	$(x - x_2)^2$	NOCHE	$x - x_3$	$(x - x_3)^2$	TOTAL
520	62.7	3,927.1	460	2.7	7.1	410	-47.3	2,240.4	
480	22.7	513.8	420	-37.3	1,393.8	420	-37.3	1,393.8	
475	17.7	312.1	410	-47.3	2,240.4	395	-62.3	3,885.4	
520	62.7	3,927.1	510	52.7	2,773.8	440	-17.3	300.4	
			480	22.7	513.8	460	2.7	7.1	
			460	2.7	7.1				

4	$=n_1$	$n_2=$	6	$n_3=$	5	$n_T=$ 15
1995	$=\Sigma_1$	$\Sigma_2=$	2740	$\Sigma_3=$	2125	$\Sigma_T=$ ####
498.8	$= x_1$	$x_2=$	456.7	$x_3=$	425.0	$x_T=$ ####

SCT= $\Sigma(x - x_T)^2=$ 23443.333

SCE= $\Sigma(x_k - x_T)^2 x n_k=$ 12091.25

SCD= SCT-SCE= 11352.083

$\eta^2=SCE/SCT=$ 0.515765

$\eta = \sqrt{\eta^2}=$ 0.7181678

PRUEBA DE SIGNIFICACION

	Σ CUAD	G LIB	ESTIM VAR	$F_{empírico}$
TOTAL	23443.333	14 =n-1		
EXPLICADA	12091.25	2 =k-1	6045.625	6.3906772
NO EXPLICADA	11352.083	12 =n-k	946.0069444	

ENTRANDO A LA TABLA DE DISTRIBUCION DE MUESTREO DE F:

"n1"= 2

"n2"= 12 Ft = 3.68

$F_{empírico} > F_{teórico}$ Se rechaza la Ho

Acerca del test de Levene de uniformidad de varianzas

Comentarios en torno del ejercicio sobre análisis de varianza de la guía de trabajos prácticos de Metodo III

El test de Levene es una prueba F en la que cada valor de la variable dependiente es el valor absoluto de la diferencia entre cada valor original de la variable independiente y la media de su grupo. Está especialmente diseñada para contrastar el supuesto de homocedasticidad. En este texto nos centramos en la interpretación de la misma, obviando los detalles de su estructura interna.

El cuadrito siguiente –tomado de la guía de trabajos prácticos– informa sobre los valores mínimamente necesarios para tomar una decisión sobre la hipótesis nula en esa especial prueba F.

Test of Homogeneity of Variances

EL AJUSTE FORMA PARTE DE UN ATAQUE MAS GENERAL A LA UNIV

Levene Statistic	df1	df2	Sig.
2.270	2	758	.104

Dice lo siguiente:

Fempírico = 2.270

Grados de libertad asociados con la mayor de las estimaciones de la varianza = 2

Grados de libertad asociados con la menor de las estimaciones de la varianza = 758

En la tabla de la distribución de F, para p = 0.05, df1 =2 y df2 = 758 es Fteórico = 3.00

Es Fe < Ft por lo que no se puede rechazar la Ho.

Lo mismo ocurre si comparamos significaciones (p-valor): pe > pt (0.104 > 0.05) y coincidentemente no se puede rechazar la hipótesis nula.

Ocurre que en la prueba de Levene estamos interesados en NO rechazar la hipótesis nula para que resulte corroborado el supuesto de homocedasticidad y en consecuencia el resultado al que llegamos, de no rechazar la Ho, es favorable (a la inversa de lo que estamos habituados a considerar en las pruebas de significación "típicas").

Es decir que en la prueba de Levene estamos interesados en altos valores de pe, cuanto más altos estaremos corroborando el supuesto de homocedasticidad con mayor probabilidad.

El resultado de nuestro ejercicio es en alguna medida satisfactorio, porque hubiera sido totalmente descalificatorio para el supuesto de homocedasticidad si hubiéramos rechazado la Ho.

Pero estrictamente, si nuestro criterio sigue siendo hacer afirmaciones con probabilidad (confianza) de 0.95, hubiéramos necesitado una pe mayor o igual que 0.95. Estamos lejos de eso, pero al menos 0.104 es más que 0.05 y en consecuencia nos "salvamos" de lo peor, que hubiera sido NO RECHAZAR Ho, y seguimos adelante con nuestro análisis principal, pero sabemos que estrictamente hablando no se cumple el supuesto de homocedasticidad con un 95% de confianza. Por suerte, los estadísticos dicen que el análisis de varianza es una prueba estadísticamente robusta en relación con el supuesto de homocedasticidad.