ANÁLISIS DE LA VARIANZA (ANOVA)

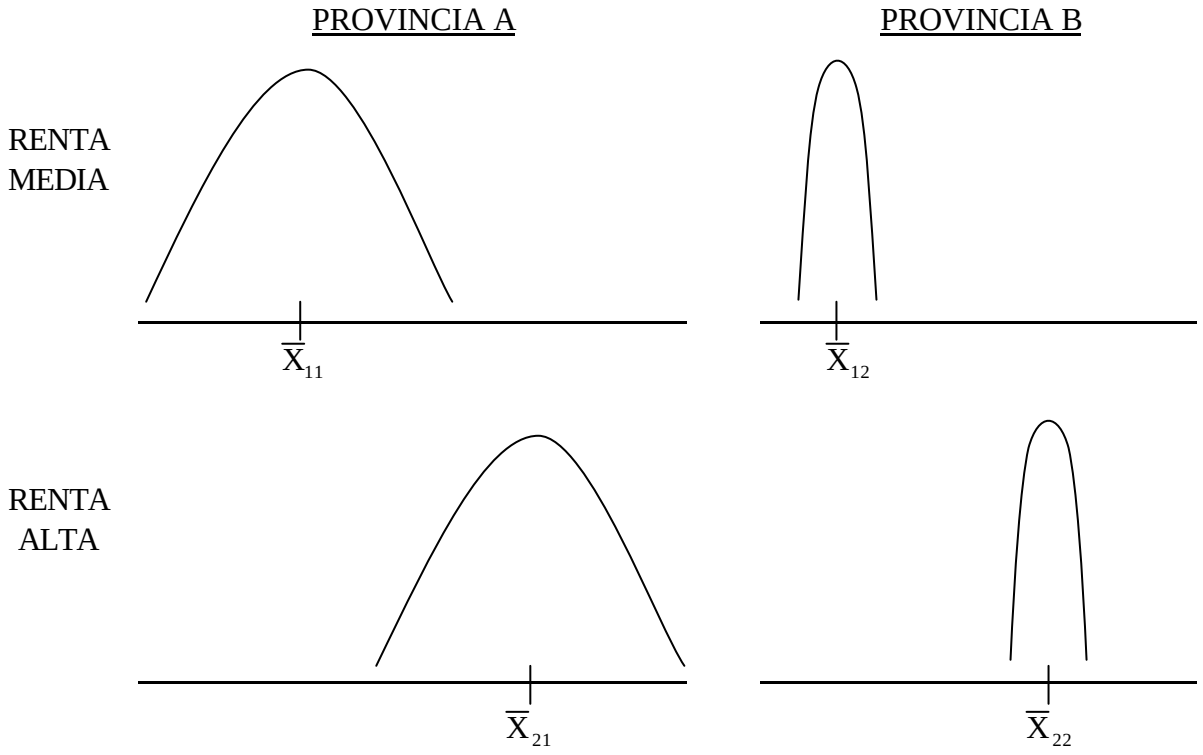
José Vicéns Otero
Ainhoa Herrarte Sánchez
Eva Medina Moral

Enero 2005

1.- INTRODUCCIÓN

En múltiples ocasiones el analista o investigador se enfrenta al problema de determinar si dos o más grupos son iguales, si dos o más cursos de acción arrojan resultados similares o si dos o más conjuntos de observaciones son parecidos. Pensemos por ejemplo en el caso de determinar si dos niveles de renta producen consumos iguales o diferentes de un determinado producto, si las notas de dos grupos en una asignatura son similares, si tres muestras de análisis químico de una sustancia son iguales, o si los municipios de cuatro provincias colindantes tienen el mismo nivel de paro.

Una aproximación simple sería comparar las medias de estos grupos y ver si las medias aritméticas de la variable estudiada son parecidas o diferentes. Pero tal aproximación no es válida ya que la dispersión de las observaciones influirá en la posibilidad de comparar los promedios o medias de cada grupo. Así, supongamos que tenemos una variable X (consumo) y dos grupos (nivel de renta alto y medio) y que tenemos dos resultados distintos correspondientes a dos provincias



Es evidente que la conclusión de que con renta alta el consumo es mayor que con renta media es más rotundo en la provincia B que en la A. En la provincia A existen familias de renta media con un consumo superior a otras familias de renta alta aunque en promedio $\overline{X}_{21} > \overline{X}_{11}$. Esta situación no se produce en la provincia B donde todas las observaciones de

renta alta son superiores a las de renta media. En consecuencia la dispersión deberá tenerse en cuenta para realizar una comparación de medias o de grupos y esto es lo que se pretende con el Análisis de la Varianza.

El Análisis de la Varianza puede contemplarse como un caso especial de la modelización econométrica, donde el conjunto de variables explicativas son variables ficticias y la variable dependiente es de tipo continuo. En tales situaciones la estimación del modelo significa la realización de un análisis de la varianza clásica (ANOVA), de amplia tradición en los estudios y diseños experimentales. Una ampliación a este planteamiento es cuando se dispone de una variable de control que nos permite corregir el resultado del experimento mediante el análisis de la covariación con la variable a estudiar. En tal situación nos encontramos frente a un análisis de la covarianza (ANCOVA). A continuación se expondrán ambos procedimientos, ANOVA, ANCOVA, precedidos de un ejemplo que facilita su compresión.

2.- ANÁLISIS DE LA VARIANZA: ANOVA

Ejemplo: Pretendemos medir la influencia que tiene en la venta de un producto de alimentación, la posición en que se exhibe al público dentro del establecimiento.

Las posiciones establecidas son:

- ALTA: por encima de los ojos.
- MEDIA: nivel de los ojos.
- BAJA: por debajo del nivel de los ojos.

Para la realización del experimento se han seleccionado 12 autoservicios de dimensiones similares. Los autoservicios se agrupan en tres conjuntos de cuatro elementos cada uno, procediendo de forma aleatoria en su asignación. Con ello suponemos que los tres conjuntos son de características similares, colocándose el producto en cada uno de ellos, de una de las formas anteriormente descritas y registrando sus ventas durante veinte días. Las ventas resultantes, en unidades, quedan recogidas en el cuadro I. Se pretende responder a las siguientes preguntas:

1º.¿Tiene alguna influencia el posicionamieno del producto en la venta del mismo?.

2º.¿Qué posicionamiento es más eficaz?

3º.¿Son estadísticamente significativas las diferencias obtenidas?

Cuadro I. Ventas en autoservicios por tipo de tratamiento

POSICIÓN PRODUCTO	ESTABLECIMIENTO	VENTAS (unidades)
ALTA	A	663
	B	795
	C	922
	D	1056
MEDIA	E	798
	F	926
	G	1060
	H	1188
BAJA	I	528
	J	660
	K	792
	L	924

Este sencillo ejemplo nos presenta el caso de tener un único factor a analizar (posición del producto) y tres niveles del factor (ALTO, MEDIO, BAJO). Sin embargo, podemos encontrarnos con múltiples factores a estudiar simultáneamente. Al mismo tiempo, podemos distinguir tres tipos de modelos según sean de:

-Efectos fijos: donde sólo estudiamos determinados niveles del factor (es el caso de las tres alturas) y únicamente perseguimos sacar conclusiones para éstos.(Situación más común en las Ciencias Sociales).

-Efectos aleatorios: en este caso los niveles son infinitos y estudiamos una muestra de los mismos. Sus resultados también serán aleatorios.

-Efectos mixtos: cuando nos encontramos con uno o más factores de las clases anteriores.

Como vemos, ANOVA será especialmente útil en aquellos supuestos en los que queramos analizar distintas situaciones o alternativas de actuación y donde de alguna forma podemos intervenir en la realización del experimento. A diferencia del análisis econométrico habitual, donde las series históricas son dadas y no podemos repetir la situación, ni modificar alguna de las condiciones o variables (pensemos en el P.I.B., inflación, etc.) para estudiar sus efectos, en el contexto ANOVA y ANCOVA nos encontraremos la mayoría de las veces ante datos experimentales (controlables y/o repetibles en mayor o menor grado).

Si bien los desarrollos clásicos de ANOVA y ANCOVA se han efectuado desde el análisis de variación de las variables y su descomposición (variaciones entre - intragrupos), podemos efectuar una sencilla aproximación desde el análisis de regresión múltiple, con idénticos resultados.

Dado que suponemos al alumno familiarizado con la aproximación tradicional de ANOVA, en base a explicaciones de otras asignaturas, aquí nos limitaremos a un breve recuerdo de la misma.

El modelo ANOVA tradicional tiene la expresión:

$$Y_{ij} = m + t_j + e_{ij}$$

Y_{ij} = es la variable objeto de estudio y que en nuestro caso es la venta para el establecimiento i del nivel j.

m = es una constante e indica la respuesta media de todos los niveles.

t_j = es el efecto diferencial del nivel j. Recoge la importancia de cada tratamiento y es el objetivo del análisis. Dado que los t_j son efectos diferenciales sobre m tenemos que $\sum t_j = 0$.

e_{ij} = es un término de error, considerado como variable aleatoria $N \sim (0, s^2)$

Por tanto, las ventas de un autoservicio, Y_{ij} se descomponen en una parte que es común a todos los tratamientos, *(m)*, o en otras palabras el efecto medio de todos ellos y otra parte, (t_j) que es el efecto diferencial de poner el producto a una determinada altura y que es propio de ese nivel. Un tercer componente es lo no recogido por los dos anteriores y que denominamos error.

No olvidemos que el subíndice i nos indica el elemento o autoservicio:

$i = 1,2,.....n_j$

para cada nivel j.

$j = 1,2,.....g$

En nuestro ejemplo, g es igual a tres niveles (ALTO, MEDIO Y BAJO) y n_j es igual a cuatro para cualquier nivel j (cuatro establecimientos para cada nivel).

El ANOVA tradicional parte de descomponer la variación total de la muestra, en dos componentes:

VARIACIÓN TOTAL	=	VARIACIÓN ENTRE	+	VARIACIÓN INTRA

Esta igualdad básica nos indica que la variación total es igual a la suma de la variación o dispersión entre los grupos, más la variación o dispersión dentro de cada grupo. Los grupos están definidos por los niveles de factor.

La anterior igualdad puede expresarse por:

$$\underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2}_{\text{V. TOTAL}} = \underbrace{\sum_{j=1}^{g} n_j(\overline{Y}_{\bullet j}-\overline{Y}_{\bullet\bullet})^2}_{\text{V. ENTRE}} + \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{\bullet j})^2}_{\text{V. INTRA}}$$

Correspondiendo cada término de la suma a las anteriores variaciones y siendo $\overline{Y}_{\bullet\bullet}$ la media total e $\overline{Y}_{\bullet j}$ la media de grupo o nivel j.

Los grados de libertad (número de observaciones – parámetros a estimar) correspondientes a cada uno de los componentes de la variación total son:

- Variación ENTRE: g – 1
- Variación INTRA: n – g
- Variación TOTAL: n – 1

Dado que a través del Análisis de la Varianza se persigue saber si los distintos niveles de un factor influye en los valores de una variable continua (en nuestros ejemplo queremos saber si la posición de un producto en una estantería influye en las ventas de ese producto), para que efectivamente sí haya diferencias en los valores de la variable continua según el nivel del factor, se tiene que dar simultáneamente que el comportamiento de la variable continua sea lo más distinto posible para los distintos niveles del factor, y a su vez, que dentro de cada

grupo (determinado por los niveles del factor) los valores sean lo más homogéneos posibles. En otras palabras, se tiene que dar que la variación intragrupos sea mínima, y que la variación entre-grupos sea máxima.

Por tanto el análisis de la varianza se va a basar no sólo en la descomposición de la variación total, sino además en la comparación de la variación ENTRE-grupos y la variación INTRA-grupos, teniendo en cuenta sus correspondientes grados de libertad.

Se demuestra que:

$$E\left[\frac{VARIACIÓN \quad ENTRE / g-1}{VARIACIÓN \quad INTRA / n-g}\right] \approx F_{g-1,n-g}$$

Por tanto, un valor elevado de este cociente significará que mayores son las diferencias entre los distintos grupos (niveles del factor), cumpliéndose asimismo que la variación dentro de cada grupo sea mínima, y por tanto la probabilidad de que los niveles del factor influyan en los valores de la variable continua será mayor.

Dado que dicho cociente se distribuye como una F de Snedecor con g-1,n-g grados de libertad, el valor para el cual podremos asumir que sí existen efectos diferenciales entre los niveles dependerá del valor de tablas de la función F para un nivel de significación de al menos el 5%. Si el valor calculado es mayor que el valor de tablas significará que sí hay efectos diferenciales entre los grupos y por tanto aceptaremos la hipótesis de que existe dependencia entre las variables.

Por el contrario, si el valor calculado es inferior al valor de tablas de una $F_{g-1,n-g}$ aceptaremos que no existen efectos diferenciales entre los grupos, o en otras palalabras:

$$t_1 = t_2 = \ldots = t_n = 0$$

Así, la hipótesis nula a contrastar a través del Análisis de la Varianza puede ser establecida como igualdad de efectos:

$$H_0 = t_1 = t_2 = \ldots = t_g = 0$$

siendo la hipótesis alternativa (H_1) que alguno de los efectos diferenciales sea distinto de cero.

Resumiendo diremos:

Si $F > F_{g-1,n-g} \rightarrow H_1$ (Existen diferencias entre los tratamientos)
Si $F = F_{g-1,n-g} \rightarrow H_0$ (No existen diferencias entre los tratamientos)

En nuestro ejemplo los resultados de la aproximación tradicional se presentan en el cuadro adjunto. Recordemos que la fuente de variación "explicada" corresponde a la *entre* grupos y la "residual" a la *intra* grupos. Los grados de libertad correspondientes son:

$g - 1 = 2$ (g = 3 niveles)
$n - g = 9$ (n = 12 establecimientos)

Corregido por los grados de libertad podemos obtener por cociente el ratio F que en este caso es 2,492. Si comparamos este valor con el obtenido en las tablas, encontramos que para un 95% de probabilidad $F_\varepsilon = 4{,}26$ luego aceptaríamos la hipótesis nula de que todos los efectos del factor altura son iguales.

VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRÁTICA	F
ENTRE (Explicada)	142578.667	2	71289.333	2.492
INTRA (Residual)	257438.000	9	28604.222	
TOTAL	400016.667	11	36365.152	

3.- UTILIZACIÓN DEL PROGRAMA SPSS

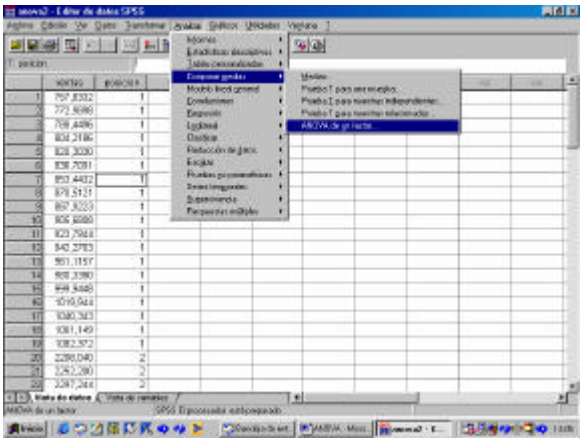
A continuación se describirán cuales son los pasos necesarios para realizar el Análisis de la Varianza utilizando la aplicación del SPSS para Windows. Para nuestra aplicación utilizaremos el ejemplo en el que se intenta determinar si el posicionamiento del producto influye o no en sus ventas, por lo que generamos una nueva variable que denominaremos posición y que diferencia los niveles del factor.

Establecimiento	Ventas	Posicionamiento del Producto
A	663	1
B	795	1
C	922	1
D	1056	1
E	798	2

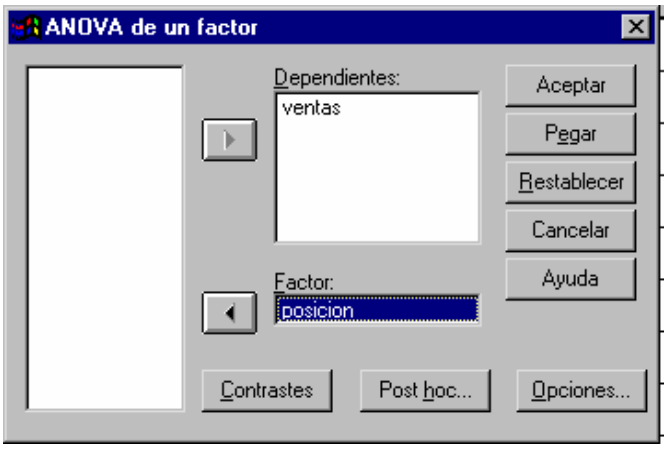
F	926	2
G	1060	2
H	1188	2
I	528	3
J	660	3
K	792	3
L	924	3

Análisis de la Varianza con un solo factor

Opción recomendable cuando deseamos aplicar un Análisis de la Varianza en el que utilizamos un sólo factor como variable explicativa. Para ello, una vez abierto nuestro archivo de datos e introducidas las variables "posición" y "ventas", nos introducimos en la opción de **"Analizar"** y pinchamos en **"Comparar Medias"**, seleccionando la opción **"ANOVA de un factor"** que nos permitirá realizar el Análisis de la Varianza.



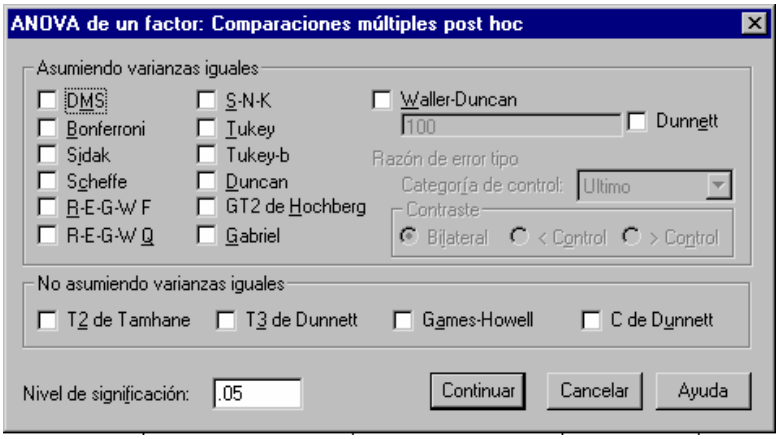
Una vez seleccionada esta opción aparece el cuadro de diálogo del Anova de un Factor, donde debemos especificar cuál es la variable dependiente (Ventas) y el Factor o variable independiente (Posición). Inicialmente las variables aparecerán en el cuadro blanco de la parte izquierda de la imagen; nosotros deberemos desplazar dichas variables a su casilla correspondiente utilizando los iconos de las flechas. En nuestro ejemplo deberemos introducir la variable "Ventas" en la casilla correspondiente a "Variables dependientes", y la variable "Posición" en la casilla que dice "Factor", tal y como se muestra en la imagen.



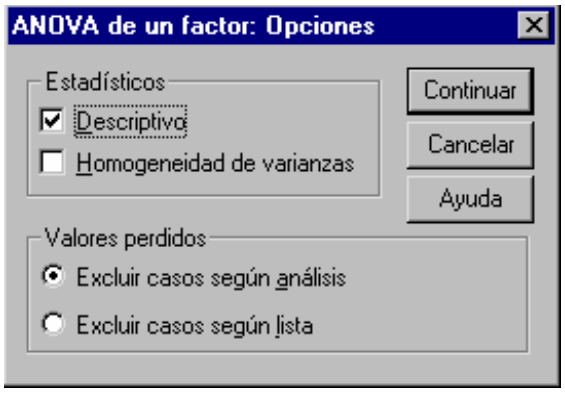
A continuación podemos seleccionar una serie de opciones, pulsando en cada uno de los tres botones del cuadro de dialogo inicial (Contrastes, Post hoc y Opciones).

Pulsando el botón **Contrastes** permite dividir la suma de cuadrados entre-grupos en componentes de tendencia o especificar contrastes a priori para que se contrasten mediante el estadístico t.

Cuando el ANOVA rechace la hipótesis nula (es decir cuando aceptemos la hipótesis de que los niveles del factor influyen sobre la variable endógena) será interesante realizar un análisis ex-post. Este tipo de análisis se basa en comparaciones múltiples por parejas entre las medias de los distintos grupos, para así conocer a qué grupos exactamente se deben las diferencias observadas entre ellos. El botón **Post Hoc** nos permite seleccionar distintas pruebas para realizar este tipo de análisis, entre las que se encuentran el test de la Diferencia Mínima Significativa (DMS), Bonferroni, Sidak, etc...



Pulsando el botón **Opciones** aparece la siguiente pantalla, cuyas distintas alternativas se explican a continuación:



- **Descriptivos:** Muestra el número de casos, la media, la desviación típica, el error típico, los valores mínimo y máximo y el intervalo de confianza al 95% para cada variable dependiente en cada grupo.

- **Homogeneidad de varianzas:** Contrastan las violaciones del supuesto de igualdad de varianzas utilizando la prueba de Levene.

- **Excluir casos según análisis:** Excluye los casos que tienen valores perdidos en la variable implicada en esa prueba.

- **Excluir casos según lista:** Excluye los casos que tienen valores perdidos en cualquiera de las variables utilizadas en cualquiera de los análisis.

Una vez seleccionadas todas las opciones que consideremos necesarias para nuestro análisis ya estaremos en condiciones para realizar al análisis de la varianza (ANOVA), pulsando la tecla Aceptar. Los resultados del ANOVA aparecerán en el Navegador de resultados de SPSS.

A continuación se muestran la salida de SPSS correspondiente al Análisis de la Varianza con los datos propuestos en el ejemplo habiendo seleccionado únicamente las opción de Estadísticos descriptivos en el botón de Opciones:

ANOVA de un factor

Descriptivos

		N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
						Límite inferior	Límite superior		
VENTAS POSICION	ALTA	4	859,0000	168,6120	84,3060	590,7046	1127,2954	663,00	1056,00
	MEDIA	4	993,0000	168,3528	84,1764	725,1170	1260,8830	798,00	1188,00
	BAJA	4	726,0000	170,4113	85,2056	454,8416	997,1584	528,00	924,00
	Total	12	859,3333	190,6965	55,0493	738,1706	980,4961	528,00	1188,00

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
VENTAS	Inter-grupos	142578,67	2	71289,333	2,492	,138
	Intra-grupos	257438,00	9	28604,222		
	Total	400016,67	11			

La primera tabla muestra la media, la desviación típica, el error típico, y los valores máximo y mínimo para cada uno de los grupos. Los valores de esta tabla nos permiten conocer en qué posición sobre la estantería, las ventas del producto son mayores. Dados

estos resultados se puede observar a primera vista que las ventas en la posición media son mayores que las ventas en las posiciones baja y alta, y que cuando el producto se coloca en la posición baja las ventas del producto son las menores.

La siguiente tabla es la salida básica de un Análisis de la Varianza: a través de los datos que muestra podremos saber si realmente existe una relación de dependencia entre las variables objeto de estudio o no, podremos saber si los distintos niveles de la variables cualitativa o factor (posición del producto sobre la estantería) determinan el valor de la variable cuantitativa (ventas del producto).

Lo que en la tabla aparece como "Inter-grupos" es el valor de la VARIACIÓN ENTRE, y el valor de "Intra-grupos", es la VARIACIÓN INTRA. También aparece el valor de la VARIACIÓN TOTAL. A continuación, la salida muestra los grados de libertad, que para el caso de la "Variación Entre" son g – 1 = 2 y en el caso de la "Variación Intra" son n – g = 9. La columna "Media cuadrática" muestra los valores del cociente de la Variación Entre y la Variación Intra por sus correspondientes grados de libertad. Recordemos que cuanto más se aproximen la media cuadrática factorial (Variación Entre/g-1) y la media cuadrática residual (Variación Intra/n-g) mayor será la probabilidad de aceptar la hipótesis nula (H_0) o no influencia del factor.

Por último la salida del SPSS nos muestra el valor calculado del estadístico F y su nivel de significación. El nivel de significación nos va a permitir aceptar o rechazar la hipótesis nula (independencia entre las variables) sin necesidad de tener que comparar el valor de la F con su valor real de las tablas estadísticas de una F de Snedecor.

El valor que nos sirve de referencia a la hora de aceptar o rechazar la hipótesis nula es el nivel de significación. Si el nivel de significación es **mayor** que 0,05, aceptaremos la hipótesis nula de independencia entre las variables (no existen efectos diferenciales entre los tratamientos). Si el nivel de significación es **menor** que 0,05 rechazaremos la hipótesis nula y aceptaremos la hipótesis alternativa, es decir, concluiremos que existe una relación de dependencia entre las variables, y en este caso podremos decir que los distintos niveles del factor sí influyen sobre los valores de la variable cuantitativa. El nivel de significación como se expuso en el capítulo anterior es la probabilidad de rechazar la hipótesis nula siendo cierta.

En nuestro caso, dado que el valor del nivel de significación es 0,138 y este valor es mayor que 0,05 aceptaremos la hipótesis nula de que no existen efectos diferenciales entre los tratamientos. Esto querrá decir que la posición del producto sobre la estantería no hace que las ventas del mismo sean estadísticamente diferentes.