¿Realmente son Diferentes Mis Productos?

Jeanette Castillo Balderas
Factum

Análisis de Varianza de un Sólo Factor

El estudio de las diferencias entre las opiniones, preferencias, hábitos y perfiles de los consumidores, es una tarea cotidiana de la investigación de mercados. Y dado que nuestras sugerencias serán accionables por nuestros clientes, es de suma importancia de-terminar cuando dichas diferencias son realmente significativas y cuando no.

Entre los métodos para determinar la diferencia entre dos o más cantidades, se encuentra el análisis de varianza, el cual se utiliza para contrastar la hipótesis de que las medias de varias muestras son iguales. Por lo general, cada conjunto muestral está afectado por un tratamiento específico, que puede influir en los valores que tome la variable que se estudia.

Se denominará *factor* a la variable que se supone ejerce una influencia sobre la variable de estudio, la cual es conocida como variable dependiente. Por las características del factor, éste puede tener varios niveles o categorías.

El modelo de varianza con un sólo factor supone un modelo lineal simple en el cual la variable dependiente es igual a una media teórica más una variable aleatoria. Esta última se supone normalmente con media cero. Sean G niveles o diferentes categorías del factor analizado. De esta forma el modelo expuesto sería:

$$Y_g = \mu_g + \varepsilon_g \quad g = 1,\ldots,G$$

Y_g es la variable dependiente.

μ_g es la media teórica.

ε_g es la variable aleatoria o el error de la estimación (también llamado residual).

Con respecto a los datos, los valores de la variable factor deben ser enteros y la variable dependiente debe ser cuantitativa (nivel de medida de intervalo).

La hipótesis que queremos demostrar es la siguiente:

Hipótesis nula: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_g$ vs. Hipótesis alternativa: al menos dos medias son diferentes

Con la ayuda de un paquete estadístico generamos un resumen del modelo, en una tabla (1) que se llama ANOVA, y se presenta a continuación:

ANOVA

Tabla 1

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Factor	G-1	SCF	MCF=SCF/G-1	
Residual	n-G	SCR	MCR=SCR/n-G	F=MCF/MCR
Total	n-1	SCT	MCT=SCT/n-1	

en donde:

SCF = *Suma de cuadrados del factor* . (Es la desviación de la media muestral de cada grupo respecto a la media global, es decir, lo que explica el factor).

SCR = *Suma de cuadrados del Residual* (desviación no explicada por el factor y debida a un error aleatorio).

SCT = *Suma de cuadrados del Total* (desviación de cada observación con respecto a la media global).

F es nuestro estadístico de prueba y la regla de decisión es la siguiente: se rechazará la hipótesis nula (igualdad en todas las medias) si F de la tabla ANOVA es mayor que el valor obtenido en tablas de la función $F_{(G-1,\ n-G)}$ con el nivel de confianza elegido por el investigador.

Existe una prueba para la bondad del ajuste y para la cual se usa el coeficiente de determinación que está dado por: R^2 = SCF / SCT. En otras palabras un valor próximo a 1 indica que la mayor parte de la variabilidad total puede atribuirse al factor, mientras que un valor próximo a cero significa que el factor explica muy poco a esa variabilidad total.

Una vez realizado el análisis y en caso de haber aceptado la hipótesis nula de que los distintos grupos tienen la misma media, puede darse por concluido el análisis, pero en caso contrario será conveniente determinar que grupos presentan las diferencias.

Uno de los métodos de análisis útiles a este fin es el de *comparaciones múltiples*, el cual consiste en contrastar cada pareja de medias y que, dependiendo del paquete estadístico utilizado, generan una matriz que indica las medias de grupo significativamente diferentes. Existen varias pruebas, pero las más usadas y en orden de menor a mayor exigencia para determinar diferencias significativas son, Ducan, SNK, Tukey y Scheffé.

Ejemplo de Aplicación

Supongamos que se desea lanzar al mercado un nuevo sabor para una bebida, se dispone de cuatro nuevos sabores, por lo que es necesario determinar si son igualmente preferidos entre los consumidores, para cada sabor se levantó información en seis spots diferentes. Los resultados para cada sabor se presentan en la tabla 2.

Tabla 2

SABORES	CONSUMIDORES QUE LO PREFIEREN					
A	16	11	20	21	14	7
B	37	32	20	29	37	32
C	21	12	14	17	13	20
D	45	59	48	46	38	47

Análisis

Lo primero es generar la tabla ANOVA con la opción de que el análisis se haga para un sólo factor, lo cual nos genera los siguientes resultados (tabla 3).

ANOVA

Tabla 3

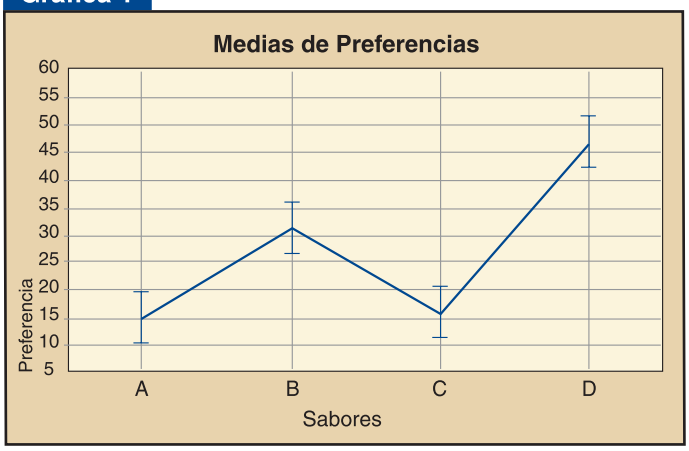
FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Factor	3	4133.5	1377.8	42.84
Residual	20	643.2	32.16	
Total	23	4776.7		

Comparando valor de tablas para $F_{(3,20)} = 3.86$ y dado que F obtenida en el análisis de varianza es mayor, rechazamos la hipótesis nula, en otras palabras: hay evidencias suficientes para concluir que la preferencia entre los cuatro sabores es diferente, y que dicha diferencia es significativa.

Si realizamos la bondad del ajuste del modelo observamos que $R^2 = 0.8653$, lo cual indica que el 86.53% de la variación puede atribuirse al factor, en es-te caso, al sabor de la bebida.

Por lo tanto, debemos ahora determinar el sabor o sabores que son diferentes, para lo cual analizaremos las medias de los sabores.

Gráfica 1



La gráfica 1 muestra diferencias entre los sabores, especialmente D, pero habrá que corroborarlo.

PRUEBA DE SCHEFFE

Probabilidades para Pruebas Post Hoc

Tabla 4

SABORES	A	B	C	D
A		0.000881	0.982407	0.000000
B	0.000881		0.002112	0.001096
C	0.982707	0.002112		0.000000
D	0.000000	0.001096	0.000000	

La tabla anterior muestra los valores de las compa-raciones de cada sabor contra los demás, haciendo comparaciones dos a dos. Los valores en rojo indican diferencia significativa en los sabores fila-columna correspondientes a la intersección.

Ahora podemos identificar claramente las diferencias entre los sabores: D es preferido sobre los sabores A, B y C; B es preferido sobre A y C; la preferencia entre A y C no es estadísticamente diferente. Por la naturaleza del problema el análisis puede ayudar a la sugerencia de lanzar el sabor D, o bien el producto B, ya que su frecuencia de preferencia es significativamente mayor a C y D.

Este método es muy útil, ya que de acuerdo a las necesidades del investigador, puede extenderse a más de un factor e incluso, utilizar varios niveles en cada factor, por ejemplo, evaluar en cada sabor, la intensidad de color, cantidad de gas, nivel de concentración del sabor, etcétera, obteniendo así información más detallada.

Bibliografía

Draper, N.; Smith, H (1981) *Applied Regression Analysis*; J. Wiley, New York. Second edition.

Montgomery, D.C.; Peck, E. (1992) *Introduction to linear regression analysis*. Wiley, New York. Second edition.