UNIVERSIDAD DEL SALVADOR
FACULTAD DE CUENCIAS SOCIALES
Estadística y Técnicas de Computación Aplicadas a las Ciencias Sociales

Horacio Chitarroni*

EL ANÁLISIS SIMPLE DE LA VARIANZA

¿Los ingresos de los asalariados difieren mucho según el tamaño de los establecimientos en los que trabajan?

Tradicionalmente se ha entendido que el tamaño de las empresas era un factor asociado a la productividad y, por ello, a la retribución de los trabajadores. Sin embargo, los cambios tecnológicos hacen que no necesariamente sea así.

En algunas ramas, empresas que ocupan muy pocos trabajadores altamente calificados podrían tener productividad más alta y pagar más que otras de mayor tamaño, pertenecientes a otros sectores de actividad, más intensivos en mano de obra. Así, el nivel de remuneraciones de las empresas podría tender a independizarse del tamaño.

¿Conserva vigencia la relación entre tamaños y remuneraciones o es, a esta altura, poco importante? El análisis simple de la varianza (ANOVA) es una herramienta estadística adecuada para explorar la cuestión.

Al ocuparnos de las medidas de estadística descriptiva –y en particular las de dispersión– tuvimos ocasión de examinar la varianza de una distribución. Recordaremos que ella se obtenía calculando primero los residuos de cada valor o puntuación a la media aritmética. Luego, estos residuos se elevaban al cuadrado (con lo que se tornaban todos ellos positivos) y se sumaban[1]. Esta suma se denominaba *suma de cuadrados* o *suma cuadrática.* Si luego se la dividía por los grados de libertad de la distribución (n-1), entonces se obtenía la varianza.

$$\sigma^2 = S^2 = \frac{\sum \left(x - \bar{x}\right)2}{n-1}$$

La varianza expresa, entonces, la variación de las puntuaciones de los diferentes casos en torno a la media. Algunos se sitúan por encima, en tanto que otros lo hacen por debajo, a mayor o menor distancia del promedio. Podemos preguntarnos qué es lo que los hace variar. Analizar la varianza es, precisamente, eso: encontrar algún criterio o factor que nos de cuenta de esas variaciones, Que nos explique –al menos en parte– por qué tienen lugar.

Supongamos que examinamos la distribución de los ingresos de un conjunto de trabajadores. Esa distribución tendrá una media aritmética. Y habrá algunos que ganaran sueldos próximos a ella. Pero otros se situarán por encima o por debajo y a diferentes distancias.

---

* Profesor asociado

[1] La suma algebraica de los residuos a la media sin cuadrar, obviamente, vale cero.

Uno podría pensar que, al menos en parte, la variación obedece a la calificación de las tareas que desempeñan: los de calificación profesional ganarán, en promedio, más que los técnicos, éstos que los operativos y éstos últimos que los no calificados.

También puede ser que en ciertas ramas se gane más: en la industria que en los servicios y en éstos últimos que en el comercio, por caso.

O bien que en las empresas de mayor tamaño se obtengan retribuciones medias más elevadas que en las firmas medianas. Y en éstas se gane más que en la microempresas.

Es posible que cada uno de estos criterios contribuya a explicar al menos alguna parte de la variación de los ingresos. Y seguramente se combinarán entre sí para dar cuenta de ella. Por ejemplo, tal vez los que más ganen sean los trabajadores de calificación profesional de las grandes empresas industriales.

Pero de cualquier modo sería muy interesante poder determinar cuánto explica cada uno de estos criterios y, por lo tanto, cuál de ellos lo hace en mayor medida. El análisis simple de la varianza (ANOVA) es una herramienta estadística apta para ello.

¿Cómo lo hace?

Para ello, se vale de una propiedad que podemos denominar como *propiedad aditiva de la varianza*, que en rigor se cumple para el numerador de la fórmula: es decir, para la suma cuadrática. Esta suma cuadrática puede descomponerse dividiendo la distribución en segmentos en función de alguna otra variable.

Imaginemos que la distribución de los ingresos de los asalariados está representada por el siguiente rectángulo:

Xi	$\overline{Xt}$	Xi

Los diversos valores de X (genéricamente Xi se distribuyen en torno a la media total de la distribución. La suma de cuadrados total puede expresarse entonces como sumatoria de los residuos elevados al cuadrado de Xi a $\overline{Xt}$:

SCT = $\sum \left(Xi - \overline{Xt}\right)^2$

Pero si dividimos la distribución en segmentos –por ejemplo según el tamaño de la empresa en la que los trabajadores se desempeñan: pequeñas, medianas y grandes– tendríamos:

Xij $\overline{X}\,jp$ Xij	Xij $\overline{X}\,jm$ Xij	Xij $\overline{X}\,jg$ Xij

Al interior de cada segmento tendremos, a su vez, una media del segmento ($\overline{X}\,j$) en torno a la cual se distribuirán los valores de X de los casos incluidos en ese segmento (*Xij)*. Y podremos calcular residuos entre cada observación y la media de su segmento o grupo. Así, resultarán sumas de cuadrados posibles de calcular dentro de los grupos –tantas como grupos haya– que luego se pueden adicionar entre sí. La suma de cuadrados intragrupos o dentro de los grupos (*within-groups*) es, entonces, una doble sumatoria:

$$SCD = \sum\sum\left(Xij - \overline{Xj}\right)^2$$

Por fin, hay otra parte de la suma cuadrática, que podemos denominar suma de cuadrados intergrupos o entre grupos *(between-groups*). Ella está determinada por las dispersiones o residuos entre cada media grupal y la media total:

$$SCE = \sum\left(\overline{Xij} - \overline{Xt}\right)_2$$

La distancia de una observación cualquiera a la media total puede descomponerse, asimismo, en dos segmentos: distancia de esa observación a la media de su grupo más distancia de la media grupal a la media total:

$$Xij - \overline{Xt} = (Xij - \overline{Xj}) + (\overline{Xj} - \overline{Xt})$$

Por fin, podemos decir que la suma de cuadrados total (que serían una expresión de la variación total de la variable), resulta de sumar estas dos ultimas:

SCT = SCD + SCE

¿De qué nos sirve segmentar la suma de cuadrados. Razonemos así: si el criterio o factor empleado para formar los grupos estuviera estrechamente relacionado con la variación de los salarios, entonces las medias grupales serían muy diferentes entre sí, en tanto que al interior de los grupos quedaría escasa dispersión: los salarios serían muy homogéneos al interior de cada tamaño de empresas. Es más: si lo único que hiciera variar los salarios fuese el tamaño de la empresa, entonces no existiría dispersión dentro de los grupos. La suma de cuadrados dentro sería cero.

Por el contrario, si el factor elegido nada tuviera que ver con la dispersión de los salarios, entonces ocurriría que las medias grupales serían muy semejantes entre sí (no habría diferencias apreciables entre trabajadores que se desempeñaran en establecimientos de diferente tamaño) y, en cambio, subsistiría mucha dispersión interna en los grupos definidos por tamaño, determinada por otros factores (calificación, antigüedad, rama de actividad, etc.). En ese caso, sería la suma de cuadrados entre grupos la que tendería a cero.

Por esta razón, la suma de cuadrados entre los grupos corresponde a la varianza “explicada” por el factor –puesto que podemos imputarla a él– en tanto que la suma de cuadrados residual, la que queda al interior de los grupos, es la correspondiente a la varianza “no explicada” por el factor, puesto que este no se relaciona con ella: depende de otros factores.

De manera que cuando la suma de cuadrados entre grupos es una proporción muy grande de la total, entonces el factor “explica mucho”. Al revés, si es una proporción pequeña, entonces “explica poco”. Un simple cociente nos permite determinarlo. Se trata de un coeficiente que se llama razón de correlación (E^2), cuya interpretación es el porcentaje de varianza (en nuestro ejemplo, del ingreso laboral) explicada por el factor de agrupación (el tamaño de las empresas):

$$E^2 = \frac{SCE}{SCT}$$

El cálculo de las varianzas

Para el cálculo de las respectivas varianzas (intragrupos e intergrupos), a las que se suele denominar "cuadrados medios", es preciso dividir cada suma de cuadrados (intra e Inter) por los respectivos grados de libertad:

Grados de libertad para la varianza total: N – 1

Grados de libertad para la varianza intragrupos: N – K

Grados de libertad para la varianza intergrupos: K – 1

Donde:

N: número total de observaciones

K: cantidad de grupos (categorías del factor)

Varianza total (VT) = $\frac{SCT}{N-1}$

Varianza dentro (VD) = $\frac{SCD}{N-K}$

Varianza entre (VE) = $\frac{SCE}{K-1}$

La comparación de medias

Fundamentalmente, el análisis de la varianza consiste en un procedimiento de comparación de medias, que puede concebirse como una extensión de la prueba t de Student[2] a un número mayor que dos grupos. En tal sentido, conserva la estructura de un diseño experimental donde en vez de contar con dos grupos de comparación tenemos varios, en los que el factor o estímulo operaría con diferentes intensidades.

En nuestro ejemplo, queremos saber si los ingresos medios de los asalariados difieren de un modo significativo si se los clasifica según el tamaño de la empresa en la que se desempeñan.

Pero, supongamos que efectivamente difieren, estamos comparando medias de muestras. ¿No podrá suceder que las diferencias observadas sean por entero casuales e imputables al azar del muestreo? Una hipótesis de nulidad nos dirá precisamente eso: en las respectivas poblaciones de las que provienen las muestras, las medias no difieren.

Para contrastar esta hipótesis nula, se emplea una prueba de significación: la F de Snedecor. F se calcula mediante la fórmula que sigue:

F = $\frac{VE}{VD}$

Véase que F se obtiene como ratio entre la varianza "explicada" o entre grupos y la varianza residual o "no explicada" (dentro de los grupos). El valor así obtenido debe ser contrastado con un valor crítico proporcionado por la tabla correspondiente a la distribución de muestreo de este estadístico (tabla IV del anexo, en la que n_1 corresponde a los grados de libertad de la varianza entre grupos y n_2 a los de la varianza dentro de los grupos). Toda vez que el F calculado sea

[2] A la que nos hemos referido en el capítulo 7 de este texto.

igual o mayor que el tabular, para el nivel de significación elegido, hemos de rechazar la hipótesis nula, lo que nos permitirá afirmar que la diferencia entre las medias muestrales es significativa: estas muestras no pueden provenir de poblaciones cuyas medias son iguales. Si F es un número inferior a la unidad (Blalock) carece de sentido la comparación con la tabla y será preciso aceptar la hipótesis nula.

Los supuestos del ANOVA

El análisis de varianza exige asumir ciertos supuestos estadísticos con referencia a la distribución de los datos.

En primer lugar, ha de suponerse que las muestras cuyas medias se comparan son muestras al azar obtenidas de las respectivas poblaciones en forma independiente.

En segundo término habremos de suponer normalidad: esto quiere decir que al interior de cada grupo o segmento, en la población, los residuos de cada puntuación a la media (que expresan las varianzas no explicadas por el factor) debieran distribuirse en forma aproximadamente normal.

En el caso de la homocedasticidad, de lo que se trata es que las varianzas al interior de cada uno de estos segmentos debieran ser similares.

En lo que atañe a la normalidad, es posible observar los histogramas correspondientes a cada una de las muestras, para ver si se apartan sensiblemente de la distribución normal. Hay, asimismo, pruebas específicas que nos dicen si unos datos muestrales provienen o no de una población normal, como el test de Kolmogorov-Smirnov, que no abordaremos aquí.

Con respecto a la homocedasticidad, el test de Levene para la igualdad de varianzas –que ya mencionamos al tratar acerca de la prueba t de Student– permite someter a prueba la hipótesis nula que afirma que las varianzas de las poblaciones de las que provienen las muestras no difieren. Si se rechaza esta hipótesis nula, nos encontramos con un problema de heterocedasticidad: las varianzas no son iguales.

El incumplimiento de los supuestos torna menos consistentes los resultados. Sin embargo, se considera que el análisis de varianza con un solo factor es un procedimiento robusto, que solamente se ve afectado si las varianzas son muy desiguales o bien el apartamiento de los residuos con respecto a la normalidad es muy extremo.

Los contrastes post-hoc

Ahora bien, una vez que se ha rechazado la hipótesis nula, lo que se está afirmando es que las medias son significativamente diferentes entre sí. Ello surge de la prevalencia de la varianza explicada sobre la residual o inexplicada: tal lo que expresa F de Snedecor.

Pero si tenemos varias medias grupales, ¿cuáles de ellas serán las que difieren de manera significativa entre sí? En principio no lo sabemos: es preciso hacer comparaciones de a pares. Si contamos con tres grupos, hemos de comparar el 1 con el 2, el 1 con el 3 y el 2 con el 3.

Para ello existen diversas pruebas de contraste, algunas de ellas más conservadoras y otras más permisivas en cuanto a posibilitar el rechazo de la hipótesis nula. Las hay aptas para el caso en que puede asumirse el supuesto de igualdad de varianzas y también para los casos en que tenemos heterocedasticidad. Los programas de procesamiento de datos estadísticos como el SPSS disponen de un gran número de ellas: en todas ellas la presentación de los resultados es análoga. Se muestra la diferencia entre las medias tomadas de a dos, el error estándar de la estimación de dicha diferencia, la significación con que es posible rechazar la hipótesis nula

que afirma que las medias no difieren y, por fin, el intervalo de confianza en el que se encuentra la verdadera diferencia entre las medias poblacionales con un 95% de confianza.

Y los ingresos de los trabajadores, ¿dependen mucho del tamaño de la empresa...?

Podemos volver ahora la mirada a la cuestión inicial. Veamos cuánto de la variación registrada por los ingresos de los asalariados puede explicarse a través del tamaño de las empresas en las que trabajan.
La salida inicial del SPSS muestra los resultados del análisis:

Tamaño	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
hasta 5	3104	587,4	514,11	9,22	569,37	605,55
6 a 40	5802	954,4	738,32	9,69	935,39	973,40
más de 41	6245	1260,6	906,33	11,46	1238,10	1283,07
Total	15151	1005,4	815,5	6,62	992,44	1018,41

La primera tabla de salida muestra algunos resultados descriptivos: los tamaños muestrales en cada grupo o categoría del factor y las respectivas medias (grupales y total). Vemos que difieren en el "sentido esperado": la remuneración promedio va creciendo a medida que aumenta el tamaño de las empresas empleadoras. Luego tenemos el desvío estándar de cada grupo, así como el total. Y el error estándar de la estimación de las verdaderas medias grupales y total en la población. Si se suma y se resta 1,96 veces dicho error estándar a la media estimada, obtenemos los límites superior e inferior del intervalo de confianza dentro del cual se encuentran dichas medias poblacionales con un 95% de probabilidad, cosa que encontramos en las últimas dos columnas.

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
235,28	2	15148	0,0000

Luego tenemos los resultados del test de levene para la homogeneidad de varianzas entre los grupos. La significación del mismo nos indica que, lamentablemente, el riesgo de cometer error de tipo I si rechazamos la hipótesis nula que manifiesta que las varianzas no difieren es muy pequeño (no es cero, por supuesto, pero el primer dígito viene antecedido de, al menos, cuatro ceros...). Estamos forzados, pues, a aceptarla. Tenemos un problema de heterocedasticidad. Pero sigamos...

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	963.947.191,4	2	481.973.595,7	801,29	0,0000
Within Groups	9.111.466.366,2	15148	601.496,3		
Total	10.075.413.557,6	15150			

Ahora tenemos aquí la tabla de ANOVA propiamente dicha. Primero nos muestra las sumas de cuadrados: entre los grupos (*between*) y dentro de los grupos (*within*). También la total. Un sencillo cociente entre la suma cuadrática entre grupos y la total nos arroja el valor de E^2:

$$E^2 = \frac{963.947.191,4}{10.075.413.557,6} = 0,10$$

Esto nos revela que solamente el 10% de la variación de los sueldos puede explicarse por el distinto tamaño de las empresas. Previsiblemente, el resto se debe a otros factores. Véase que la mayor parte de la dispersión se conserva dentro de los grupos.

Pero además vemos los grados de libertad asociados a cada segmento de la varianza y luego, los cuadrados medios (las varianzas) que resultan de dividir cada suma cuadrática (entre grupos y dentro de los grupos) por los grados de libertad. Por fin, el cociente entre estos dos segmentos de la varianza es el valor de F de Snedecor. Y la última columna de esta tabla, crucial para la toma de la decisión, es la significación de F: el riesgo de error en que se incurre al rechazar la hipótesis de nulidad que afirma la igualdad de las medias. Como esta probabilidad es muy pequeña, no vacilaremos en rechazar esa hipótesis nula.

Ahora, faltaría saber cuáles son las medias grupales que difieren significativamente una de otra. Para eso, hemos aplicado dos test post-hoc, adecuados para cuando no se cumple –como aquí sucede– el supuesto de homocedasticidad.

Dependent Variable: P21

	(I) tamaño del establecimiento	(J) tamaño del establecimiento	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tamhane	hasta 5	6 a 40	-366,94	13,38	0,0000	-398,9	-335
		más de 41	-673,12	14,72	0,0000	-708,3	-638
	6 a 40	hasta 5	366,94	13,38	0,0000	335,0	398,9
		más de 41	-306,19	15,02	0,0000	-342,0	-270,3
	más de 41	hasta 5	673,12	14,72	0,0000	638,0	708,3
		6 a 40	306,19	15,02	0,0000	270,3	342
Dunnett T3	hasta 5	6 a 40	-366,94	13,38	0,0008	-399,0	-334,9
		más de 41	-673,12	14,72	0,0006	-708,3	-637,9
	6 a 40	hasta 5	366,94	13,38	0,0008	334,9	399
		más de 41	-306,19	15,02	0,0000	-341,9	-270,5
	más de 41	hasta 5	673,12	14,72	0,0006	637,9	708,3
		6 a 40	306,19	15,02	0,0000	270,5	341,9

Se trata de los test de Tamhane y Dunnett. Tenemos allí la comparación de las medias grupales una a una, las diferencias entre ellas y el error estándar en la estimación de la verdadera diferencia en la población. Luego, el nivel de significación con el que se puede rechazar la hipótesis nula que predicaría que tal diferencia es igual a cero. Aquí, ambos test permiten rechazar en todos los casos esas hipótesis de nulidad, con probabilidades de error muy reducidas: eso es lo esencial. Por fin, las últimas columnas nos muestran los intervalos de confianza dentro de los cuales se situará cada una de las diferencias con un 95% de probabilidad. ¡Si alguno de estos intervalos pasara por cero (límite inferior negativo y límite superior positivo) no vacilaríamos en aceptar la hipótesis nula!

En definitiva, los salarios medios sí que difieren entre empresas de diferente tamaño. Y no cabe duda de que siempre conviene trabajar en establecimientos más grandes. Pero no obstante, solo una pequeña proporción de la gran diversidad que muestran las remuneraciones obedece a esta factor (apenas un 10%). El resto se relaciona con otras características, tanto propias de los mismos trabajadores como de sus puestos de trabajo.