Las bases del análisis de varianza

EDWIN GARRO

PXS

blog.pxsglobal.com

El Análisis de Varianza es, para muchos, la técnica más importante de toda la estadística industrial. La usamos en pruebas de hipótesis de tres o más medias, en análisis de regresión, en diseño de experimentos. Vamos a hacer un repaso de las bases del ANOVA y así estar en una mejor posición de explicar los resultados e implicaciones de su aplicación.

¿Quién inventó el análisis de varianza?

Esta joya de las técnicas estadísticas se la debemos al gran biólogo cuantitativo Sir Ronald Fisher. Fisher tomó la posición de estadístico en la mítica Estación Experimental de Rothamsted en el año 1911 (hoy llamado Rothamsted Research) y se encontró con una enorme cantidad de datos recolectados desde 1842 relacionados con agricultura y biología. La aplicación, y en muchos casos invención, de métodos estadísticos permitió un rico análisis de la información existente.

¿Por qué inventó Fisher en ANOVA y otras técnicas?

Fisher sintió que analizar datos ya tomados era como hacer autopsias. Era necesario intervenir desde la planeación de la toma de datos, o como lo conocemos hoy, hacer el diseño del experimento. Así nació la serie de investigaciones "Studies in Crop Variation" y fue precisamente en uno de estos estudios "Study of Crop Variation II: The manurial response of different potato varieties",

en 1923, que se introdujo por primera vez el análisis de varianza como una forma de estudiar las diferencias entre más de dos medias. Note que la ANOVA se creó con la intención de tener un "antes" y un "después". El "antes" es la planeación científica de la recolección de la información, y el "después" el análisis estadístico del resultado. Puede encontrar más detalle en la publicación "Guinness, Gosset, Fisher, and Small Samples", de 1987, escrita por Joan Fisher Box (hija de Fisher y primera esposa de otro grande de la estadística George Box).

Con el Análisis de Varianza el uso de la estadística pasó de simple reporteo sofisticado de datos a herramienta de carácter científico. Y en palabras del mismo Fisher:

> "Consultar al estadístico después de que el experimento ha finalizado es muchas veces meramente conducir un examen post mortem. Tal vez pueda decir de qué murió el experimento"

.

La genialidad de ANOVA, analizar varianzas para describir diferencias entre medias

Vamos a usar en ejemplo obvio (¿o no lo será tanto?). Vamos a comparar las estaturas de jugadores de tres deportes: baloncesto, futbol, y futbol americano. Podríamos partir de la hipótesis inicial que los jugadores de baloncesto son más altos, en promedio, que los futbolistas. Los datos son las estaturas de los cinco jugadores mejor pagados del mundo de los tres deportes en el año 2015.

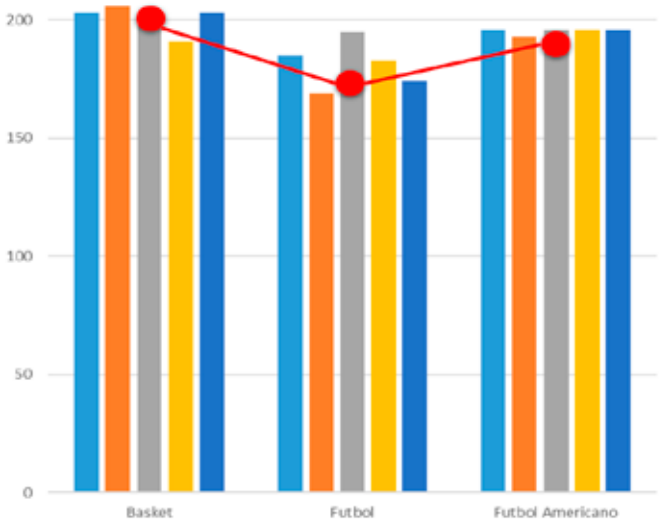
Jugador	Deporte	Salario (millones de $)	Estatura cm
Lebron James	Basket	64.8	203
Kevin Durant	Basket	54.2	206
Kobe Bryant	Basket	49.5	198
Derrick Rose	Basket	33.9	191
Carmelo Anthony	Basket	30.5	203
Cristiano Ronaldo	Futbol	79.6	185

Jugador	Deporte	Salario (millones de $)	Estatura cm
Lionel Messi	Futbol	73.8	169
Zlatan Ibrahimovic	Futbol	39.1	195
Gareth Bale	Futbol	35	183
Neymar	Futbol	31	174
Ben Roethlisberger	Futbol Americano	48.9	196
Ndamukong Suh	Futbol Americano	38.6	193

Para efectos del análisis la información que necesitamos es solamente el deporte y la estatura.

Basket	Futbol	Futbol Americano
203	185	196
206	169	193
198	195	196
191	183	196
203	174	196
promedio basket	promedio futbol	promedio futbol americano
200.2	181.2	195.4

Hagamos un rápido análisis descriptivo. Las barras representan las estaturas individuales, y los puntos (o círculos) el promedio de cada categoría. En principio los jugadores de baloncesto parecen más altos. ¿Cómo lo validamos estadísticamente?



Fisher notó que comparar los promedios en parejas se vuelve muy lento y conlleva problemas de representatividad estadística. Cada par debería ser analizado con un tamaño de muestra muy grande (puede ser que de varios cientos de datos) para poder "confiar" en todo el estudio. Así que fue necesario crear un análisis estadístico-matemático que permitiera utilizar muestras relativamente pequeñas. Para este ejemplo con solamente 3 poblaciones, si cada una tuviera un nivel de confianza de 95%, el análisis por parejas sería:

PAREJA ANALIZADA	NIVEL DE CONFIANZA
Basket – Futbol	95%
Basket – Futbol americano	95%
Futbol – Futbol americano	95%

$$Nivel\ de\ confianza\ del\ estudio = 0.95^3 = 0.86 = 86\%$$

Entre más poblaciones se analicen menor el nivel de confianza combinado.

El truco está en analizar dos fuentes de variación, la variación entre las poblaciones y la variación dentro de las poblaciones (que en ANOVA se llaman tratamientos). Si la variación entre tratamientos es igual a la variación dentro de tratamientos, entonces se puede concluir que no hay diferencias significativas entre las medias. Si ambas variaciones son diferentes es porque al menos una de las medias es diferente.

Además la variación total es igual a la suma de ambas variaciones con lo que tenemos nuestra primera ecuación:

$$Variación\ total = Variación\ entre\ tratamientos + Variación\ dentro\ de\ tratamientos$$

Cada una de las variaciones se calcula mediante una suma de cuadrados lo que nos lleva a depurar la ecuación de la siguiente forma:

$$Suma\ de\ cuadrados\ total = Suma\ de\ cuadrados\ entre\ tratamientos + Suma\ de\ cuadrados\ dentro\ de\ tratamientos$$

O simplemente:

$$\boldsymbol{SST = SS\ entre + SS\ dentro}$$

Se usa SS para mantener la nomenclatura clásica en inglés "Sum of Squares".

Los cálculos son la base para luego construir la famosa tabla de ANOVA.

La suma de cuadrados dentro de tratamientos

Usando los datos de estatura vamos a calcular la suma de cuadrados dentro de tratamientos.

1. Calcule el promedio de cada grupo.
2. A cada dato réstele el promedio de su grupo.
3. Eleve al cuadrado el resultado de cada resta.
4. Sume los cuadrados de cada grupo.
5. Por último sume los resultados de cada grupo.

Los cálculos para el ejemplo se ven de la siguiente manera:

Basket	Promedio	Dif	Cuadrado
203	200.20	2.80	7.84
206	200.20	5.80	33.64
198	200.20	-2.20	4.84
191	200.20	-9.20	84.64
203	200.20	2.80	7.84
200.20		**Suma**	**138.80**

Futbol	Promedio	Dif.	Cuadrado
185	181.20	3.80	14.44
169	181.20	-12.20	148.84
195	181.20	13.80	190.44
183	181.20	1.80	3.24
174	181.20	-7.20	51.84
181.20		**Suma**	**408.80**

Futbol Americano	Promedio	Dif	Cuadrado
196	194.80	1.20	1.44
193	194.80	-1.80	3.24
193	194.80	-1.80	3.24
196	194.80	1.20	1.44
196	194.80	1.20	1.44
194.80		**Suma**	**10.80**

La suma de cuadrados dentro de tratamientos es:

$$SS\ dentro = 138.80 + 408.80 + 10.80 = 558.40$$

La suma de cuadrados total

Se acostumbra calcular de seguido la suma de cuadrados total. Este cálculo se hace la siguiente forma:

1. Calcule el promedio de todos los datos.
2. A cada dato réstele el promedio general.
3. Eleve al cuadrado el resultado de cada resta.
4. Sume todos los cuadrados.

Para el ejemplo:

	ESTATURAS	PROMEDIO	DIF	CUADRADO
	203	192.07	10.93	119.46
	206	192.07	13.93	194.04
	198	192.07	5.93	35.16
	191	192.07	-1.07	1.14
	203	192.07	10.93	119.46
	185	192.07	-7.07	49.98
	169	192.07	-23.07	532.22
	195	192.07	2.93	8.58
	183	192.07	-9.07	82.26
	174	192.07	-18.07	326.52
	196	192.07	3.93	15.44
	193	192.07	0.93	0.86
	193	192.07	0.93	0.86
	196	192.07	3.93	15.44
	196	192.07	3.93	15.44
PROMEDIO	**192.07**		**SUMA**	**1516.93**

$$SST = 1516.93$$

La suma de cuadrados entre tratamientos

Como ya tenemos dos de los tres componentes de la ecuación, el tercero se puede hacer por despeje, sin embargo para efectos de entender su procedencia, y para confirmar el resultado, lo haremos desde cero.

1. Calcule el promedio de cada grupo.
2. Calcule el gran promedio.
3. Réstele a cada promedio el gran promedio.
4. Eleve el resultado de cada resta el cuadrado.
5. Multiplique el cuadrado de cada grupo por el número de datos en cada grupo.
6. Sume los resultados.

Para las estaturas:

DEPORTE	PROMEDIO	GRAN PROM	DIF	CUADRADO	CUADRADO X 5
BASKET	200.20	192.07	8.13	66.0969	330.4845
FUTBOL	181.20	192.07	-10.87	118.1569	590.7845
FUT. AMER.	194.80	192.07	2.73	7.4529	37.2645
PROMEDIO	**192.07**			**SUMA**	**958.5335**

$$SS\ entre = 958.53$$

!

!

Verificamos ahora el resultado final con la ecuación de la suma de cuadrados totales.

$$SST = SS\,entre + SS\,dentro$$

$$1516.93 = 958.53 + 558.40$$

$$SST = 1516.93$$

Se confirma el resultado.

TABLA DE ANOVA Y ESTIMACIONES DE LA VARIANZA (MEDIAS DE CUADRADOS)

Tabla de ANOVA y estimaciones de la varianza (medias de cuadrados)

El lector estará familiarizado con la fórmula de la varianza para muestras:

$$\frac{\sum(x-\bar{x})^2}{n-1}$$

Donde el numerador es una de las sumas de cuadrados y el denominador corresponde a los grados de libertad de la muestra. SSentre y SSdentro son sumas de cuadrados que va en el numerador. Veamos ahora cómo se agregan los grados de libertad para completar cada uno de los estimados de varianza.

Grados de libertad de la variación entre grupos = 3 grupos – 1 = 2

Grados de libertad de la variación dentro de grupos = 15 datos – 3 grupos = 12

Otra forma de entender los grados de libertad dentro de grupos es:

$$\sum_{i=1}^{\text{número de grupos}} (datos\ por\ grupo - 1)$$

En el caso del ejemplo sería: (5-1)+(5-1)+(5-1)=12
Es importante entender esta segunda forma de la ecuación porque algunas veces los grupos no tienen la misma cantidad de datos.
La estimación final de las varianzas mediante las medias de cuadrados queda de la siguiente forma:

$$MS\ entre = \frac{958.53}{2} = 479.27$$

$$MS\ dentro = \frac{558.4}{12} = 46.53$$

Las sumas de cuadrados y los grados de libertad son la base para construir la tabla de ANOVA. La primera parte de la tabla es la siguiente:

Fuente	GL	Suma cuadrados	Media cuadrados
Deporte (entre)	2	958.53	479.27
Error (dentro)	12	558.40	46.53
Total	14	1516.93	

La lógica de la ANOVA

El estimador de la varianza del error (Media de Cuadrados del Error) no se ve afectado por diferencias entre las medias, mientras que el estimador de la varianza entre grupos se infla si alguna de las medias es diferente. Fisher desarrolló la distribución F, y con ella una forma muy sencilla de comparar varianzas. La comparación consiste en calcular un valor F que proviene de los datos mediante la fórmula:

$$F = \frac{MS\ grupos\ (entre)}{MS\ error\ (dentro)}$$

Para el ejemplo:

$$F = \frac{479.27}{46.53} = 10.30$$

Luego se compara este un con valor teórico o crítico (de tabla) que se obtiene buscando el nivel de significancia, los grados de libertad del numerador y los grados de libertad del denominador. Por ejemplo para el caso de las estaturas, con un nivel de significancia del 5%, 2 grados de libertad en el numerador y 12 grados de libertad en el denominador, la lectura se haría de la siguiente forma:

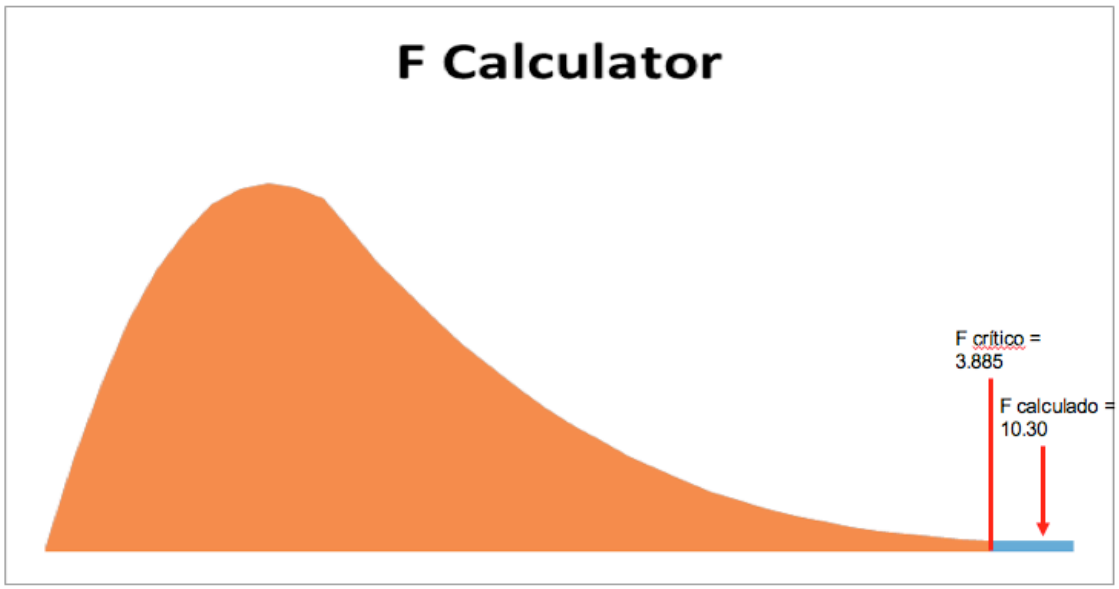
0.050	1	2	3	4
1	161.448	199.500	215.707	224.583
2	18.513	19.000	19.164	19.247
3	10.128	9.552	9.277	9.117
4	7.709	6.944	6.591	6.388
5	6.608	5.786	5.409	5.192
6	5.987	5.143	4.757	4.534
7	5.591	4.737	4.347	4.120
8	5.318	4.459	4.066	3.838
9	5.117	4.256	3.863	3.633
10	4.965	4.103	3.708	3.478
11	4.844	3.982	3.587	3.357
12	4.747	3.885	3.490	3.259
13	4.667	3.806	3.411	3.179
14	4.600	3.739	3.344	3.112

El valor de tabla, o valor crítico es 3.885.

La tabla completa de ANOVA luce de la siguiente forma:

Fuente	GL	Suma cuadrados	Media cuadrados	F	p-Value	Critical Value
Deporte (entre)	2.0000	958.5333	479.2667	10.2994	0.0025	3.8853
Error (dentro)	12.0000	558.4000	46.5333			
Total	14.0000	1516.9333				

El F crítico se compara con el F calculado. Si el calculado es mayor que el crítico entonces quiere decir que efectivamente al menos una media es diferente. Para el ejemplo la comparación luce de la siguiente forma:



Como 10.30 (F calculado) es mayor que 3.885 (valor crítico) se concluye que la media de al menos un grupo es diferente de las demás.

El P value

Solamente falta explicar un valor de la tabla de ANOVA, el llamado p value. El p value es el valor del nivel de significancia asociado al F calculado. Los valores menores al nivel de significancia de la prueba indican que al menos una media es diferente. En este caso, el valor de 0.0025 comparado con 0.05 de la prueba, indica que alguna de las medias de estaturas es diferente. El cálculo a mano del p value para la distribución de F requiere del uso de curvas características de operación que no se explican aquí.

CONCLUSIÓN

La comparación entre dos estimados de la varianza de los grupos en un experimento lleva a determinar si al menos una media, de alguno de los grupos, es estadísticamente diferente de los otros. La media de cuadrados del error (dentro de grupos) se compara con la media de cuadrados entre tratamientos (entre grupos). La prueba F indica si ambos estimadores son diferentes, y si es así quiere decir que al menos una media es diferente de las demás. En el caso del ejemplo podemos concluir que los futbolistas son más pequeños que los jugadores de

baloncesto. La comparación con los jugadores de futbol americano es un poco más compleja y requiere de otras técnicas que trataremos en otra ocasión. Los cálculos que llevan a la construcción de la tabla de ANOVA son sencillos y solamente requieren de conocimiento aritmético básico. A pesar que de que los diferentes softwares estadísticos construyen la tabla de forma automática, siempre es conveniente repasar y recordar el origen de los cálculos para poder explicar los resultados y tener una mejor visión de la técnica.