scidata

ANOVA

De una y dos vías

Instructor: Juan Luis Palacios Soto

palacios.s.j.l@gmail.com

Contenido

Uso del ANOVA de una vía

Se trata del Análisis de la varianza (de ahí su nombre). En particular, para ANOVA de una vía tenemos:

- Se utliliza para comparar la varianza entre diferentes muestras elegidas de una misma población.
- La idea es comparar dos o más muestras basados en la diferencia de las varianzas y se trata de una prueba de hipótesis sobre las medias poblacionales.
- Puede ser usado cuando tenemos al menos dos variables, donde una es categórica y la otra es continua. A la categórica la llamamos variable independiente o predictiva y a la continua la llamamos variable dependiente o de respuesta.

Por ejemplo, podrías probar si el sexo influye en el salario, si la terapia influye en la tensión arterial o si el campo de estudio influye en la duración de los estudios. El salario, la tensión arterial y la duración de los estudios son entonces las variables dependientes (respuesta). En todos estos casos, compruebas ahora si el factor influye en la variable dependiente.
Tomado de:
`https://datatab.es/tutorial/two-factorial-anova-without-repeated-measures`

Condiciones para ser aplicable

Supongamos que tenemos la variable independiente X y la variable de respuesta Y . Para poder aplicar correctamente el ANOVA de una vía se debe cumplir:

- Los grupos son independientes.
- Y debe ser aproximadamente normal en cada grupo (siendo menos estricta esta condición cuanto mayor sea el tamaño de cada grupo).
- Todos los grupos tienen la misma varianza (esta condición es más importante cuanto menor es el tamaño de los grupos).
- No tener datos atípicos.
- Es preferible tomar muestras del mismo tamaño porque esto minimiza el error tipo II (No rechazar H_0 cuando es falsa.)

Hipótesis

Supongamos que la muestra se divide, según la variable independiente, en los grupos $G1, G2, ..., Gk$. Sea μ_i la media poblacional del grupo i. Entonces la prueba de hipótesis se establece como:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs } H_1 : \mu_i \neq \mu_j \text{ para algún par } i \neq j$$

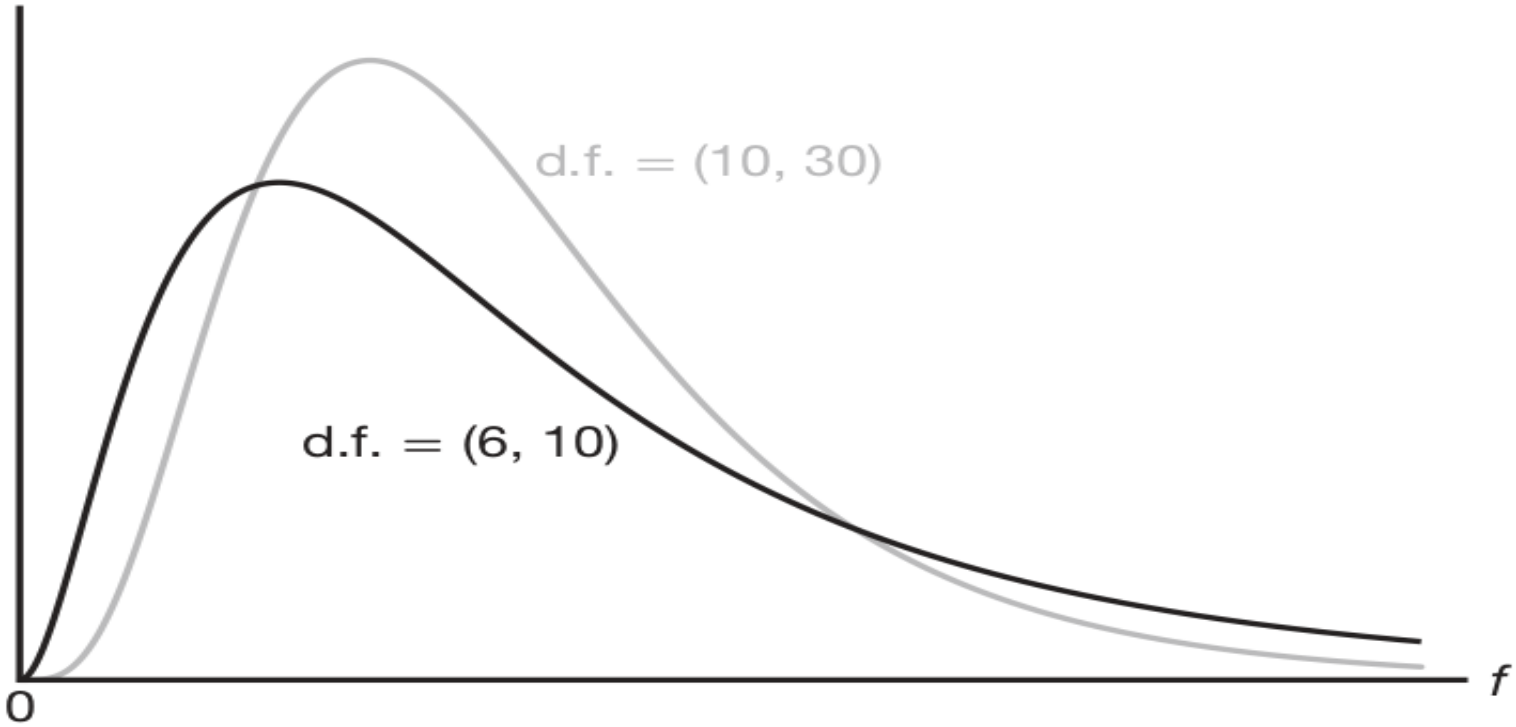
Por lo tanto, en caso de rechazar H_0, se tiene que proceder a lo que se conoce como pruebas post hoc, siendo la prueba por parejas la más conocida. Ver https://github.com/scidatmath2020/Inferencia- Estadistica/blob/master/C08. %20ANOVA.ipynb

Proceso de investigación

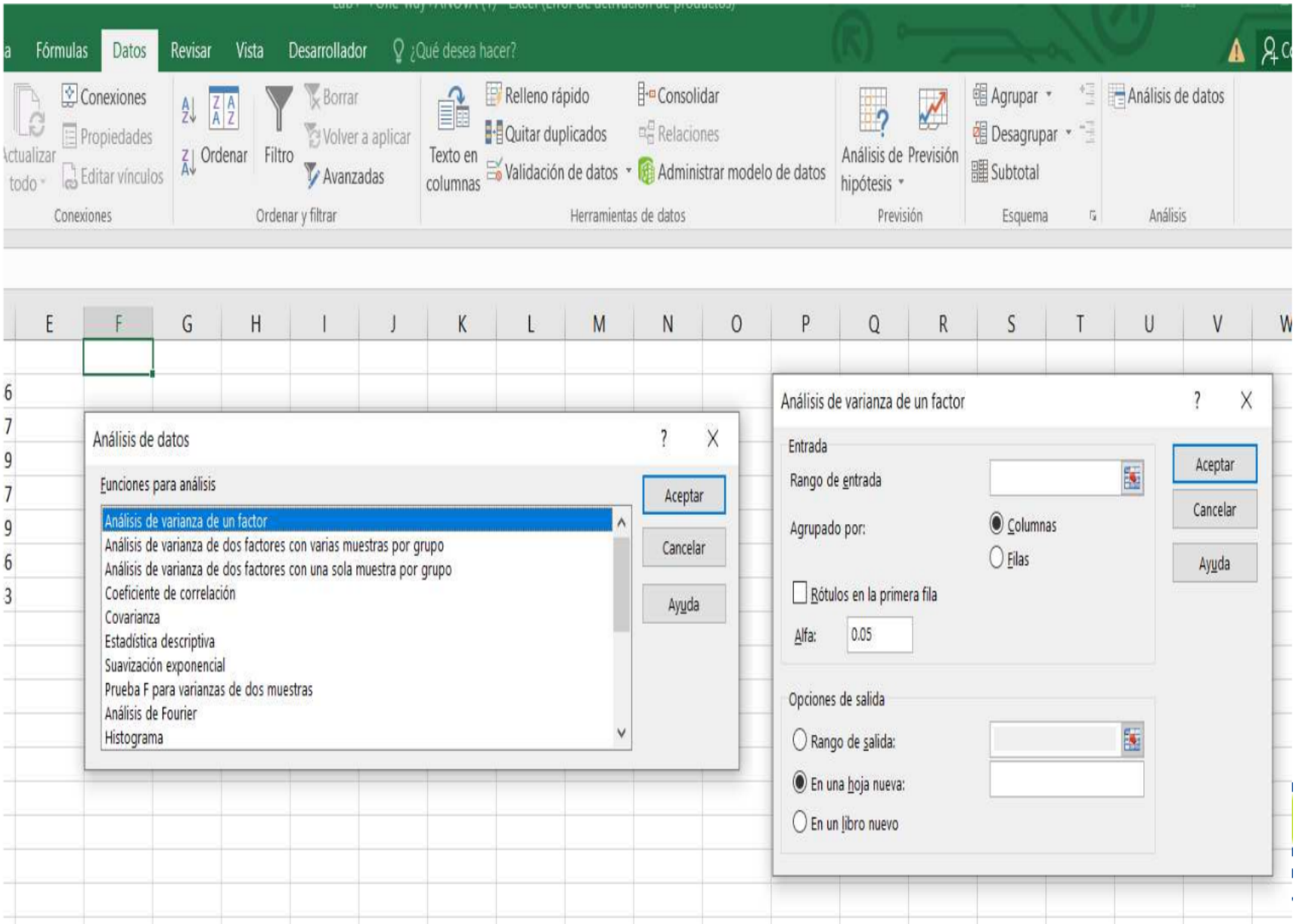
Los pasos para la aplicación del ANOVA de una vía son:

1. **Seleccionar el nivel de significación**. Este es denotado por α. Generalmente $\alpha = 0.01$, $\alpha = 0.05$ o $\alpha = 0.10$.
2. **Encontrar el valor crítico**.Este es denotado por f. Antiguamente se usaba una tabla de F . Se seleccionaba la columna basado en α y la fila basado en los grados de libertad.
3. **Calcular el valor** F. Se refiere a calcular el número F con los datos de la muestra como la media de la suma de los cuadrados.
4. **Comparar y decidir**. Si $F \geq f$, rechazamos H_0. En caso contrario, aceptamos la hipótesis nula.

Distribución F



Proceso en Excel



Ejemplo

Suponga que en un experimento industrial a un ingeniero le interesa la forma en que la absorción media de humedad del concreto varía para 5 agregados de concreto diferentes. Las muestras se exponen a la humedad durante 48 horas y se decide que para cada agregado deben probarse 6 muestras, lo que hace que se requiera probar un total de 30 muestras. En la tabla de abajo se presentan los datos registrados.

Absorción de humedad en agregados para concreto.

Agregado:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Media	553.33	569.33	610.50	465.17	610.67	561.80

Ejemplo

Parte de un estudio realizado en Virginia Tech se diseñó para medir los niveles de actividad de la fosfatasa alcalina sérica (en unidades de Bessey-Lowry) en niños con trastornos convulsivos que recibían terapia de anticonvulsivantes bajo el cuidado de un médico privado. Se reclutaron 45 sujetos para el estudio y se clasificaron en cuatro grupos de medicamentos: G-1 Control (no recibieron anticonvulsivantes ni tenían historia de trastornos convulsivos), G-2 Fenobarbital, G-3 Carbamazepina y G-4 otros anticonvulsivantes.

Nivel de actividad de la fosfatasa alcalina sérica.

G-1		G-2	G-3	G-4
49.20	97.50	97.07	62.10	110.60
44.54	105.00	73.40	94.95	57.10
45.80	58.05	68.50	142.50	117.60
95.84	86.60	91.85	53.00	77.71
30.10	58.35	106.60	175.00	150.00
36.50	72.80	0.57	79.50	82.90
82.30	116.70	0.79	29.50	111.50
87.85	45.15	0.77	78.40	
105.00	70.35	0.81	127.50	
95.22	77.40			

Uso del ANOVA de dos vías

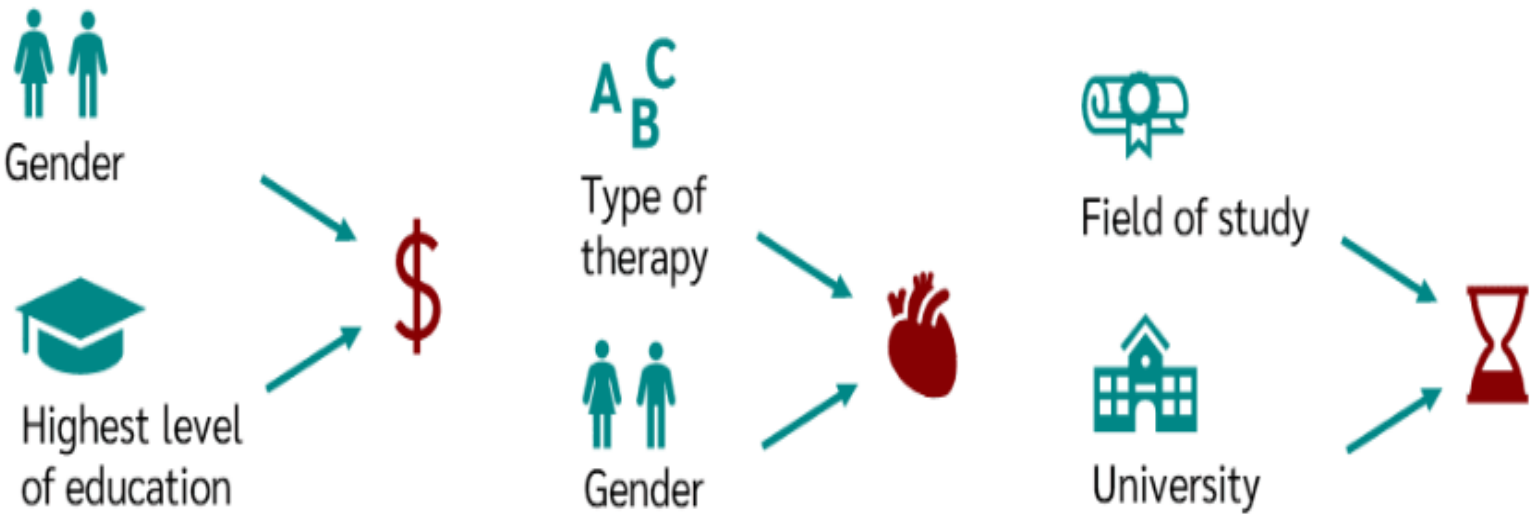
El análisis de varianza de dos vías (o dos factores) permite estudiar simultáneamente los efectos de dos fuentes de variación (dos variables o factores).

- En un anova de dos vías se clasifica a los individuos de acuerdo a dos factores (o vías) para estudiar simultáneamente sus efectos.
- Sirve para estudiar la relación entre una variable de respuesta (dependiente) cuantitativa y dos o más variables predictivas (independientes) cualitativas (factores) cada uno con varios niveles.

Ejemplos de ANOVA de dos vías

Ahora puede que tengas otra variable categórica que también quieras incluir. Puede que te interese saber si

- además del sexo, el nivel de estudios más alto también influye en el salario.
- además de la terapia, el sexo también influye en la tensión arterial.
- además del campo de estudio, la universidad a la que se asiste también influye en la duración de los estudios



Hipótesis

Con el ANOVA de 2 factores se prueban tres afirmaciones, por lo que hay 3 hipótesis nulas y, por tanto, 3 hipótesis alternativas, asaber:

Hipóstesis

Hipótesis nulas

- **NO** hay diferencias significativas en la media entre los grupos (niveles de los factores) del primer factor.
- **NO** hay diferencias significativas en la media entre los grupos (niveles de los factores) del segundo factor.
- Un factor **NO** influye en el efecto del otro factor.

- Hay una diferencia significativa en la media entre los grupos (niveles de los factores) del primer factor.
- Hay una diferencia significativa en la media entre los grupos (niveles de los factores) del segundo factor.
- Un factor tiene un efecto sobre el efecto del otro factor.

Condiciones para ser aplicable

Para que pueda calcularse un análisis de varianza de dos factores sin medidas repetidas, deben cumplirse los siguientes supuestos:

- Los grupos son independientes.
- La variable de respuesta (dependiente) debe ser cuantitativa y las variables de respuesta (independientes) deben ser cualitativas (categóricas).
- Homogeneidad: Todos los grupos tienen aproximadamente la misma varianza.
- Distribución normal: Los datos dentro de los grupos deben distribuirse normalmente.
- Es más eficiente (los errores disminuyen) si el tamaño de la muestra entre las diferentes combinaciones es la misma.

Uso del P valor para interpretar el resultado ANOVA de dos vías

En los tres casos la decisión de rechazar o no la hipótesis nula H_0 dependerá del P valor. Como hay tres hipótesis, si el P valor es menor al nivel de significancia $\alpha = 0.5, 0.01$, se toma la decisión de rechazar la hipótesis nula. En caso contrario aceptamos la hipótesis nula. El P valor se encuentra regularmente en la última columna del resultado del ANOVA usando algún software estadístico. En este ejemplo es la columna $Pr > F$, donde system y type son los componentes principales (los dos factores o vías) y system*type es la interacción entre los factores sytem y type.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
system	2	14.52333333	7.26166667	5.84	0.0169
type	3	40.08166667	13.36055556	10.75	0.0010
system*type	6	22.16333333	3.69388889	2.97	0.0512

Ejemplo

En un experimento realizado para determinar cuál de 3 sistemas de misiles distintos es preferible, se midió la tasa de combustión del propulsor para 24 arranques estáticos. Se emplearon 4 tipos de combustible diferentes y el experimento generó observaciones duplicadas de las tasas de combustión para cada combinación de los tratamientos. Los datos, ya codificados, se presentan en la tabla de abo. Pruebe las siguientes hipótesis: a) H_0: no hay diferencia en las tasas medias de combustión del propulsor cuando se emplean diferentes sistemas de misiles, b) H_0': no existe diferencia en las tasas medias de combustión de los 4 tipos de propulsor, c) H_0'': no hay interacción entre los distintos sistemas de misiles y los diferentes tipos de propulsor.

Sistema de misiles	**Tipo de propulsor**			
	b_1	b_2	b_3	b_4
a_1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
a_2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
a_3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1