


ANOVA

IRONHACK

Intro

La técnica de análisis de varianza (ANOVA) constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la **media de una variable continua** (variable cuantitativa de interés). Es por lo tanto uno de los test estadísticos que se pueden emplear para comparar las medias de dos o más grupos.

- **Factor:** variable cualitativa.
 - **Niveles del factor:** valores del factor.
 - **Tratamientos:** combinaciones de los niveles de cada factor.
- 

Intro

La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que, al menos dos medias, difieren de forma significativa. ANOVA permite comparar múltiples medias, pero lo hace mediante el estudio de las varianzas.

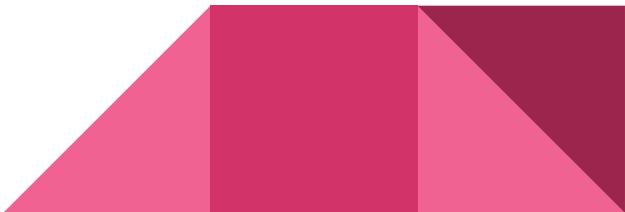
- $H_0: \mu_1 = \mu_2 = \mu_3 = \dots \mu_k$
- H_a : Al menos dos son diferentes

k = número de tratamientos del experimento



Intro

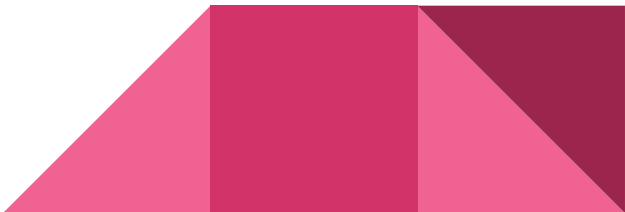
La idea intuitiva detrás del funcionamiento básico de un ANOVA es la siguiente:

- Calcular la media de cada uno de los grupos.
 - Calcular la varianza de las medias de los grupos. Esta es la varianza explicada por la variable grupo, y se le conoce como **intervarianza**.
 - Calcular las varianzas internas de cada grupo y obtener su promedio. Esta es la varianza no explicada por la variable grupo, y se le conoce como **intravarianza**.
- 

Intro

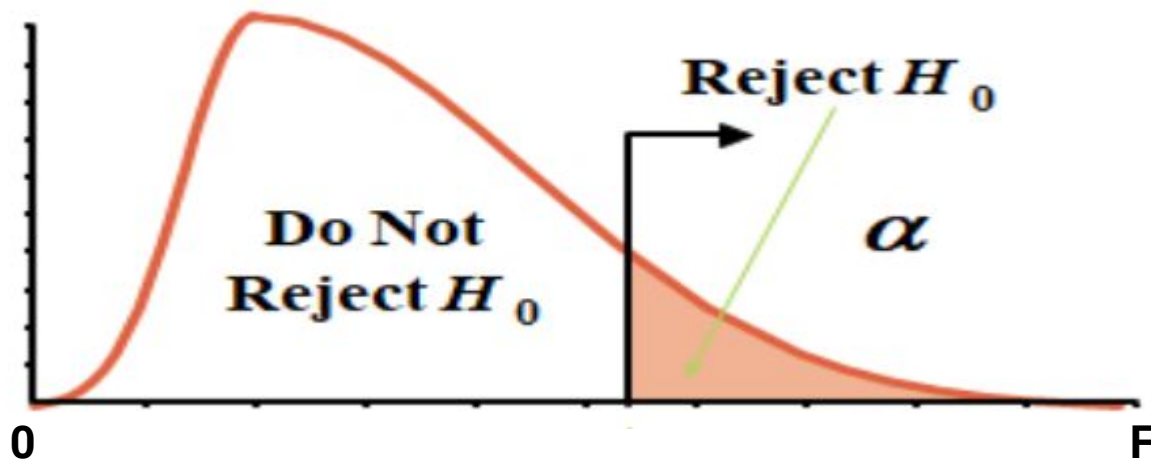
Acorde a la hipótesis nula de que todas las observaciones proceden de la misma población (tienen la misma media y varianza), es de esperar que la **intervarianza y la intravarianza sean iguales**. A medida que las medias de los grupos se alejan las unas de las otras, **la intervianza aumenta y deja de ser igual a la intravarianza**.

El estadístico estudiado en este test es el **estadístico f**, y se calcula como el ratio entre la varianza de las medias de los grupos (intervarianza, MST) y el promedio de la varianza dentro de los grupos (intravarianza, MSE).

$$F \text{ statistic} = \frac{MST}{MSE}$$


Intro

Este estadístico se distribuye como una variable F de Fisher-Snedecor. Esta distribución cubre todos los posibles valores del estadístico f.



Intro

- El resultado **no puede ser inferior a 0**.
- Si el resultado es **inferior a 1**, significa que hay **más varianza dentro de los grupos que entre ellos**. Por tanto, **no rechazamos** la hipótesis nula de que las medias de los tratamientos son iguales.
- Si el **resultado es 1**, significa que hay **exactamente la misma varianza entre los grupos que dentro de ellos**. Por tanto, **no rechazamos la H0** de que las medias de los tratamientos son iguales.
- Si el resultado es **mayor que el valor crítico**, significa que **la intervarianza es mucho mayor que la intravarianza**. En este caso, **rechazamos la hipótesis nula** y decimos que no todas las medias de los tratamientos son iguales. La prueba, por tanto, **es siempre unilateral de cola superior**.

Intro

En este caso, vamos a utilizar los pvalues para obtener conclusiones del test, por simplificar.

En caso de que queráis aprender cómo obtener los valores críticos de f , podéis preguntarme. Lo suyo es utilizar Python, la tabla de f o una calculadora online para obtenerlos.

Para ello, hay que tener en cuenta los grados de libertad de la intravarianza y de la intervianza que van a ser distintos en función del experimento.




ANOVA DE UNA VÍA

ANOVA DE UNA VÍA

El ANOVA de una vía, también conocido como ANOVA con un factor o modelo factorial de un solo factor, permite estudiar si existen diferencias significativas entre la media de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor, cuando los datos no están pareados. Es una extensión de los t-test independientes para más de dos grupos.

Las hipótesis contrastadas en un ANOVA de un factor son:

- $H_0: \mu_1 = \mu_2 = \mu_3 = \dots \mu_k$
 - H_a : Al menos dos son diferentes
- 

ANOVA DE UNA VÍA

Experimental Cancer treatment		
Treatment 1	Treatment 2	Treatment 3
patient 5 pain level	patient 3 pain level	patient 7 pain level
patient 1 pain level	patient 8 pain level	patient 2 pain level
patient 9 pain level	patient 4 pain level	patient 6 pain level

- Factor?
- k?
- Niveles del factor?
- Variable de interés
- Unidades experimentales?

ANOVA DE UNA VÍA EN PYTHON

La mejor opción que he visto en Python es con la librería statsmodels. Tiene dos pasos. Lo primero es construir el modelo y lo segundo obtener los resultados del test ANOVA.

- Creamos el modelo con la función ols de statsmodels. `ols('variable interés ~ factor', data = dataset fuente).fit`
- Obtenemos los resultados de el test anova con la función anova_lm de statsmodels



ANOVA DE UNA VÍA EN PYTHON. Ejemplo

Supón que se un estudio quiere comprobar si existe una diferencia significativa entre el ratio de bateos exitosos de los jugadores de béisbol dependiendo de la posición en la que juegan.

Utilizamos el dataset 'datosbaseball'.



ANOVA DE UNA VÍA EN PYTHON. Ejemplo

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('bateo ~ posicion', data=datosbaseball).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

	sum_sq	df	F	PR(>F)
posicion	0.007557	3.0	1.994349	0.114693
Residual	0.407984	323.0	NaN	NaN

ANOVA DE BLOQUES O MEDIDAS REPETIDAS

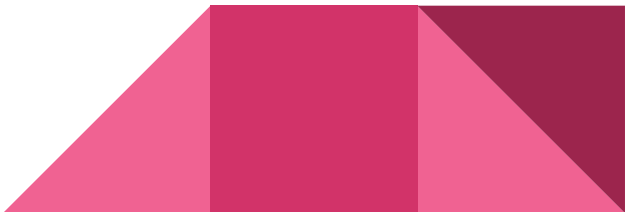
ANOVA DE UNA VÍA. Variables dependientes

Cuando las variables a comparar son variables con mediciones distintas pero sobre los mismos sujetos o sujetos del mismo bloque, no se cumple la condición de independencia, por lo que se requiere un ANOVA específico que realice comparaciones considerando que los datos son pareados (de forma similar como se hace en los t-test pareados pero para comparar más de dos grupos).



ANOVA DE UNA VÍA. Variables dependientes

Block	Treatment 1	Treatment 2	Treatment 3
Stage 1	patient 2 stage 1	patient 5 stage 1	patient 3 stage 1
Stage 2	patient 4 stage 2	patient 2 stage 2	patient 6 stage 2
Stage 3	patient 7 stage 3	patient 10 stage 3	patient 1 stage 3

- Factor?
 - k?
 - Bloque
 - Niveles del factor?
 - Variable de interés
 - Unidades experimentales?
- 

ANOVA DE UNA VÍA Variables dependientes. PYTHON

Vamos a utilizar la misma librería y funciones. Lo único que cambia es la forma en la cual escribimos es modelo en ols.

- `ols('variable interés ~ factor + bloque', data = dataset fuente).fit`



ANOVA DE UNA VÍA Variables dependientes. PYTHON

Supóngase un estudio en el que se quiere comprobar si el precio de la compra varía entre 4 cadenas de supermercado. Para ello, se selecciona una serie de elementos de la compra cotidiana y se registra su valor en cada uno de los supermercados ¿Existen evidencias de que el precio medio de la compra es diferente dependiendo del supermercado?

Utiliza el dataset 'datostienda' para obtener una conclusión de este contraste.



ANOVA DE UNA VÍA Variables dependientes. PYTHON

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('precio ~ tienda + producto ', data=datos).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

	sum_sq	df	F	PR(>F)
tienda	5.737207	3.0	13.025212	1.897557e-05
producto	139.479418	9.0	105.553625	1.102247e-18
Residual	3.964224	27.0	NaN	NaN

ANOVA DE DOS VÍAS

ANOVA DE DOS VÍAS

El análisis de varianza de dos vías, también conocido como plan factorial con dos factores, sirve para estudiar la relación entre una variable dependiente cuantitativa y dos variables independientes cualitativas (factores), cada uno con varios niveles. El ANOVA de dos vías permite estudiar cómo influyen por sí solos cada uno de los factores sobre la variable dependiente (modelo aditivo) así como la influencia de las combinaciones que se pueden dar entre ellas (modelo con interacción).



ANOVA DE DOS VÍAS

	No study time	A little	Medium	A lot
Professor A	Student 307 score Student 279 score Student 38 score	Student 10 score Student 387 score Student 169 score	Student 138 score Student 224 score Student 8 score	Student 83 score Student 99 score Student 157 score
Professor B	Student 22 score Student 57 score Student 90 score	Student 180 score Student 101 score Student 263 score	Student 3 score Student 392 score Student 125 score	Student 239 score Student 205 score Student 347 score

Factor A: Study time
Factor B: Professor
Response: Score

Value of b? 2

Value of k? 8

Value of n? 24

Value of a? 4

Experimental units:
students

Value of r? 3

Nota: el número de unidades experimentales por tratamiento se conoce como r (réplicas)

ANOVA DE DOS VÍAS PYTHON

Vamos a utilizar la misma librería y funciones. Lo único que cambia es la forma en la cual escribimos es modelo en ols.

- `ols('variable interés ~ factor1 * factor2', data = dataset fuente).fit`



ANOVA DE DOS VÍAS PYTHON. Ejemplo

Una empresa de materiales de construcción quiere estudiar la influencia que tienen el grosor y el tipo de templado sobre la resistencia máxima de unas láminas de acero. Para ello miden el estrés hasta la rotura (variable cuantitativa dependiente) para dos tipos de templado (lento y rápido) y tres grosores de lámina (8mm, 16mm y 24 mm).

Utiliza el dataset 'datosconstruct' para obtener una conclusión de este contraste.



ANOVA DE DOS VÍAS PYTHON

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('resistencia ~ templado * grosor ', data=datos).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

	sum_sq	df	F	PR(>F)
templado	112.675320	1.0	401.572490	2.479871e-17
grosor	10.296125	1.0	36.695175	2.122221e-06
templado:grosor	1.540125	1.0	5.488973	2.706347e-02
Residual	7.295217	26.0	NaN	NaN

POSTHOC TESTS

POSTHOC TESTS

Con el test de ANOVA podemos concluir que, al menos, dos de las medias difieren. Es decir, al menos dos de los tratamientos tienen un efecto en la variable continua.


Pero...¿cuáles?¿cuánto difieren?¿son solo dos o más?

Para descubrir esta info hacemos posthoc tests.



POSTHOC TESTS

Los posthoc tests, básicamente, construyen intervalos que estiman la diferencia de dos medias poblacionales. Hacen esto por cada combinación de pares, es decir, si hemos rechazado la hipótesis de que, por ejemplo, 4 medias poblacionales son iguales (al menos dos difieren), el posthoc test construiría los siguientes intervalos:

- IC para estimar $(\mu_1 - \mu_2)$
 - IC para estimar $(\mu_1 - \mu_3)$
 - IC para estimar $(\mu_1 - \mu_4)$
 - IC para estimar $(\mu_2 - \mu_3)$
 - IC para estimar $(\mu_2 - \mu_4)$
 - IC para estimar $(\mu_3 - \mu_4)$
- 

POSTHOC TESTS

Tendríamos que interpretar cada uno de los intervalos para ser capaces de 'rankear' las medias poblacionales. Lo hacemos, siguiendo estas reglas:

- **Si el intervalo de confianza de la diferencia del par de medias contiene 0:** las verdaderas medias no difieren (no tenemos evidencia estadística para demostrarlo)
- **Si ambos extremos del intervalo de confianza son positivos:** La primera media supera a la segunda
- **Si ambos extremos del intervalo de confianza son negativos:** La segunda media supera a la primera



```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

```
tukey = pairwise_tukeyhsd(endog = datos['bateo'], groups = datos['posicion'], alpha = 0.05)  
print(tukey)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject
C	DH	0.0252	0.1065	-0.0034	0.0538	False
C	IF	0.0089	0.5009	-0.0075	0.0254	False
C	OF	0.0116	0.287	-0.0053	0.0286	False
DH	IF	-0.0163	0.3588	-0.0419	0.0094	False
DH	OF	-0.0135	0.5289	-0.0395	0.0124	False
IF	OF	0.0027	0.9	-0.0085	0.0139	False

```
=====
```

POSTHOC TESTS

Con el test de ANOVA podemos concluir que, al menos, dos de las medias difieren. Es decir, al menos dos de los tratamientos tienen un efecto en la variable continua.

Pero...¿cuáles?¿cuánto difieren?¿son solo dos o más?

Para descubrir esta info hacemos posthoc tests.





Vuestro turno

House Prices Dataset

¿La clasificación de la zona (MSZoning) tiene efecto sobre el precio de venta?

- Lee la descripción de los datos para entender cuáles son los niveles del factor.
- Resuelve la pregunta utilizando un ANOVA de UNA VÍA

¿Cuál es la zona con el precio media de venta más alto? ¿Y el más bajo?

- Utiliza el posthoc test de Tukey para responder a esta pregunta

