¿Qué es Análisis de la Varianza (ANOVA)?

**Análisis de la Varianza** ( **ANOVA** ) es una fórmula estadística que se utiliza para comparar las varianzas entre las medias (o el promedio) de diferentes grupos. Una variedad de contextos lo utilizan para determinar si existe alguna diferencia entre las medias de los diferentes grupos.

Por ejemplo,

Para estudiar la efectividad de diferentes medicamentos para la diabetes, se diseñan y experimentan para explorar la relación entre el tipo de medicamento y el nivel de azúcar sanguínea resultante.

La población de la muestra es un conjunto de personas. Dividimos la población de la muestra en varios grupos y cada grupo recibe un medicamento en particular durante un período de prueba. Al final del período de prueba, se miden los niveles de azúcar sanguínea para cada uno de los participantes individuales.

Luego, para cada grupo, se calcula el nivel medio de azúcar sanguínea. ANOVA ayuda a comparar las medias de estos grupos para averiguar si son estadísticamente diferentes o si son similares.

El resultado de ANOVA es la 'estadística F'. Este ratio muestra la diferencia entre la varianza dentro del grupo y la varianza entre grupos, lo que finalmente produce una cifra que permite concluir que la hipótesis nula es respaldada o rechazada.

Si hay una diferencia significativa entre los grupos, la hipótesis nula no es compatible y la razón F será mayor.

Terminología ANOVA

**Variable dependiente**: este es el elemento que se está midiendo y que se teoriza como afectado por las variables independientes.

**Variable(s) independiente(s)**: estos son los elementos que se están midiendo y que pueden tener un efecto sobre la variable dependiente.

**Una hipótesis nula (H0)**: Sucede cuando no existe diferencia entre los grupos o medias. Dependiendo del resultado de la prueba ANOVA, la hipótesis nula será aceptada o rechazada.

**Una hipótesis alternativa (H1)**: Cuando se teoriza que existe una diferencia entre los grupos y las medias.

**Factores y niveles**: En la terminología ANOVA, una variable independiente se denomina factor que afecta a la variable dependiente. El nivel denota los diferentes valores de la variable independiente que se utilizan en un experimento.

**Modelo de factor fijo**: algunos experimentos utilizan solo un conjunto discreto de niveles para los factores. Por ejemplo, una prueba de factor fijo evaluaría tres dosis diferentes de un medicamento y no buscaría ninguna otra dosis.

**Modelo de factor aleatorio**: este modelo extrae un valor aleatorio de nivel de todos los valores posibles de la variable independiente.

¿Cuál es la diferencia entre ANOVA de un factor y de dos factores?

Existen dos tipos de ANOVA.

ANOVA unidireccional

El análisis de la varianza unidireccional también se conoce como ANOVA de un solo factor o ANOVA simple. Como sugiere el nombre, ANOVA de una vía es adecuado para experimentos con una sola variable independiente (factor) con dos o más niveles. Por ejemplo, una variable dependiente puede ser en qué mes del año hay más flores en el jardín. Habrá doce niveles. Un ANOVA unidireccional asume:

* Independencia: el valor de la variable dependiente para una observación es independiente del valor de cualquier otra observación.
* Normalidad: el valor de la variable dependiente se distribuye normalmente
* Varianza: la varianza es comparable en diferentes grupos de experimentos.
* Continuo: la variable dependiente (cantidad de flores) es continua y se puede medir en una escala que se puede subdividir.

ANOVA factorial completo (también llamado ANOVA bidireccional)

ANOVA factorial completo se utiliza cuando existen dos o más variables independientes. Cada uno de estos factores puede tener varios niveles. ANOVA factorial completo solo se puede utilizar en el caso de un experimento factorial completo, donde se utilizan todas las posibles combinaciones de los factores y sus niveles. Este podría ser el mes del año en el que existen más flores en el jardín y luego la cantidad de horas de sol. Este ANOVA bidireccional no solo mide la variable independiente frente a la independiente, sino también si los dos factores se afectan entre sí. ANOVA bidireccional asume:

* Continuo: Al igual que un ANOVA unidireccional, la variable dependiente deberá ser continua.
* Independencia: cada muestra es independiente de otras muestras, sin combinaciones.
* Varianza: la varianza de los datos entre los diferentes grupos es la misma.
* Normalidad: las muestras son representativas de una población normal.
* Categorías: Las variables independientes deberán estar en categorías o grupos separados.

¿Para que sirve ANOVA?

Algunas personas cuestionan la necesidad de ANOVA; después de todo, los valores medios se podrían evaluar con solo mirarlos. Pero ANOVA hace más que comparar los valores medios.

Aunque los valores medios de varios grupos parecen ser diferentes, esto podría ser debido a un error de muestreo más que al efecto de la variable independiente sobre la variable dependiente. Si se debe a un error de muestreo, la diferencia entre las medias de los grupos no tendrá sentido. ANOVA ayuda a averiguar si la diferencia en los valores medios es estadísticamente significativa.

ANOVA también revela indirectamente si una variable independiente está influyendo en la variable dependiente.

Por ejemplo, en el experimento de nivel de azúcar sanguíneo anterior, suponga que ANOVA encuentra que las medias de los grupos no son estadísticamente significativas y que la diferencia entre las medias de los grupos se debe solo a un error de muestreo.

Este resultado infiere que el tipo de medicación (variable independiente) no es un factor significativo que influya en el nivel de azúcar sanguíneo.

Limitaciones de ANOVA

ANOVA solo establecerá si existe una diferencia significativa entre las medias de al menos dos grupos, pero no explicará cual es el par que difiere en sus medias. Si existe un requisito de datos detallados, el despliegue de los procesos estadísticos de seguimiento adicionales ayudará a descubrir qué grupos difieren en el valor medio. Normalmente, ANOVA se utiliza en combinación con otros métodos estadísticos.

ANOVA también presupone que el conjunto de datos se distribuye uniformemente, ya que solo compara medias. Si los datos no se distribuyen a lo largo de una curva normal **y existen valores atípicos**, ANOVA no será el procedimiento adecuado para interpretar los datos.

De manera similar, ANOVA asume que las desviaciones estándar son iguales o similares entre los grupos. Si existe una gran diferencia en las desviaciones estándar, la conclusión de la prueba podría ser inexacta.

¿Cómo se utiliza ANOVA en la ciencia de datos?

Uno de los mayores desafíos del aprendizaje automático es la selección de las características más fiables y útiles que se utilizan para formar un modelo.

ANOVA ayuda a seleccionar las mejores características para formar un modelo.

ANOVA minimiza el número de variables de entrada para reducir la complejidad del modelo.

ANOVA ayuda a determinar si una variable independiente está influyendo en una variable objetivo.

Un ejemplo del uso de ANOVA en la [ciencia de datos](https://www.tibco.com/es/reference-center/what-is-data-science) es la detección de correo no deseado. ANOVA y las pruebas F se despliegan para identificar características que son importantes para identificar correctamente qué correos electrónicos son spam y cuáles no.

Preguntas que ANOVA ayuda a responder

Aunque ANOVA implica pasos estadísticos complejos, es una técnica beneficiosa para las empresas mediante el uso de IA. Las organizaciones utilizan ANOVA para tomar decisiones sobre qué alternativa elegir entre muchas posibilidades. Por ejemplo, ANOVA podrá ayudar a:

* Comparar la producción de dos variedades de trigo diferentes con tres marcas de fertilizantes diferentes.
* Comparar la efectividad de varios anuncios en las redes sociales sobre las ventas de un producto en particular.
* Comparar la efectividad de diferentes lubricantes en diferentes tipos de vehículos.