Contenidos de Aprendizaje

Tema 2

ANOVA

Introducción

🎯 Objetivos

- Identificar que es el análisis de varianza y sus usos
- Comprender el concepto de diferencia de media entre grupos e intra grupos
- Analizar el funcionamiento del estadístico F
-

⚙️ Setup

A continuación te mostramos las instalaciones necesarias para tu clase:

- Python 3
- Google Collab

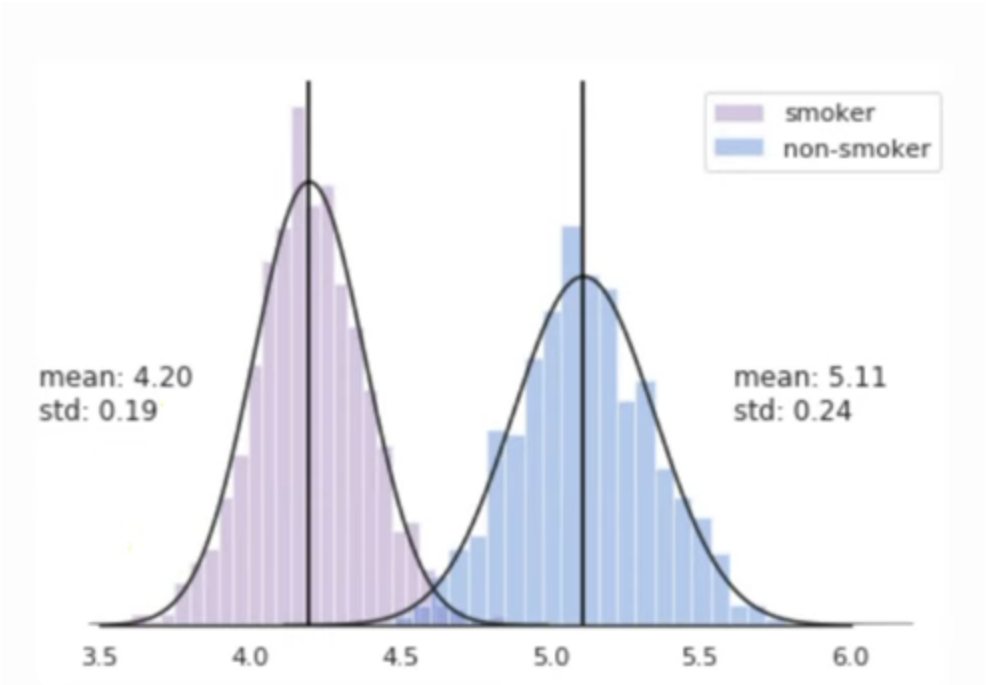
📚Tema 2.1. Diferencia de media entre grupos

⏳ 60 minutos de Clase.

Muchos análisis incluyen experimentos que prueban si existen diferencias en las medias de más de dos grupos. La evaluación de las diferencias entre grupos puede considerarse un experimento de un factor, también conocido como diseño completamente aleatorio.

Se le conoce como **factor** a la variable que define los grupos que se utilizan para las pruebas de las diferencias de las medias entre múltiples grupos. Los valores de la variable que sirve de factor se denominan niveles, y un conjunto de niveles puede ser un conjunto de valores numéricos discretos o un conjunto de categorías.

Una de las herramientas más comunes para comparar medias es la **prueba t** , esta prueba nos permite determinan si las medias entre dos grupos son iguales entre sí. La prueba t parte del supuesto de que los grupos han sido tomados de distribuciones de datos normales con varianzas iguales.



Una de las limitantes de la prueba t, es que únicamente es posible comparar dos grupos a la vez. En el caso de tener más de dos grupos a la vez, es posible realizar comparaciones entre las diferentes combinaciones de grupos, sin embargo. A medida que se incrementa la cantidad de grupos se aumentan los errores de tipo 1.

ANOVA

La técnica ANOVA nos permite comparar de forma simultánea diferencias entre las medias poblacionales de más de dos grupos en experimentos de un solo factor.

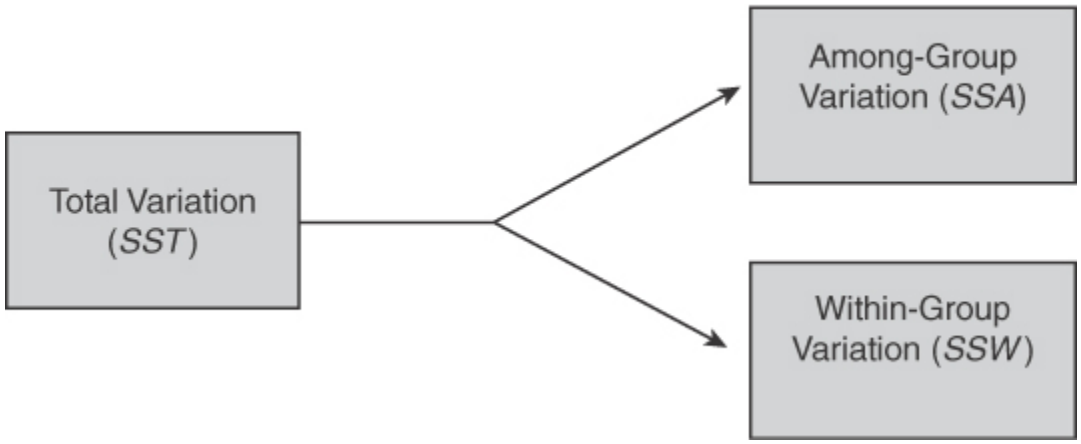
A diferencia de la prueba t, ANOVA nos permite comparar múltiples grupos a la vez sin el riesgo de incrementar el error de tipo 1. Si bien ANOVA es un acrónimo para análisis de varianza, la intuición detrás de este método es la comparación de las medias entre los diferentes grupos y no su varianza.

Las hipótesis nula y alternativa para esta prueba son:

Ho: *Las medias de las poblaciones son las mismas.*
H1: *Al menos dos medias de los grupos difieren de forma significativa.*

En análisis de varianza en ANOVA sigue la intuición del modelo linear general en el que la varianza se calcula por medio de la suma de cuadrados totales y en la que el objetivo es explicar que tanta variación puede atribuirse al modelo y que tanta puede atribuirse al error experimental.

$$SST = SSA + SSW$$



En nuestro modelo ANOVA la variación entre grupos SSW corresponde al error experimental, mientras que la variación dentro del grupo SSA representa la variación que corresponde al factor de interés.

La suma de cuadrados totales es la variación total entre cada valor y la media de todos los valores.

$$SST = \sum_{j=1}^{n} (\bar{X}_j - \bar{X})^2$$

La variación dentro del grupo es la suma de cuadrados de las diferencias entre la media de cada grupo y la media de todos los grupos, ponderada por el tamaño de la muestra en cada grupo.

$$SSB = \sum_{j=1}^{k} (\bar{X}_j - \bar{X})^2$$

La variación entre grupos corresponde a la diferencia entre cada valor y la media del grupo en que se encuentra para cada valor.

$$SST = \sum_{j=1}^{n} (\bar{X}_j - \bar{X})^2$$

El estadístico F

Las pruebas F han sido nombras en honor a Sir Ronald Fisher. El estadístico F determina la proporción entre dos varianzas. Recordemos que la varianza es una medida de dispersión que indica que tan alejados están los datos de su media, un valor grande de varianza nos indica una mayor dispersión.

La varianza se calcula mediante la obtención del cuadrado de la de la desviación estándar. En términos humanos, la desviación estándar resulta mucho más sencilla de comprender, pues esta se encuentra en las mismas unidades que los datos a analizar.

El estadístico F se basa en la proporción del valor cuadrático medio. El estadístico F puede usarse en diversas situaciones, por ejemplo, para evaluar la cualidad de varianzas o para evaluar la significancia de un modelo de regresión, entre otros usos.

En ANOVA el estadístico F calcula el valor cuadrático medio entre el valor cuadrático entre grupos. Esto es la relación entre la variación entre en las muestras, y la variación dentro de la muestra.

$$F = \frac{MSA}{MSW}$$

MSA se calcula mediante la obtención de la suma de cuadrados entre grupos (SS) divido entre el número de grupos menos .

MSW se calcula mediante la obtención de la suma de cuadrados entre grupos divido por el tamaño de la muestra menos 1.

Suposiciones realizadas por ANOVA

Para utilizar la prueba F de ANOVA se deben hacer tres suposiciones principales:

- Aleatoriedad e independencia
- Normalidad
- Homogeneidad de la varianza.

El primer supuesto, aleatoriedad e independencia, debe cumplirse siempre, porque la validez del experimento depende del muestreo aleatorio o de la asignación aleatoria de elementos o sujetos a los grupos.

La normalidad, establece que los valores de cada grupo se seleccionan de poblaciones distribuidas normalmente. El estadístico F no es muy sensible a las desviaciones de este supuesto de normalidad. Mientras las distribuciones no sean muy asimétricas, el nivel de significación del estadístico F no se verá muy afectado por la falta de normalidad, especialmente en el caso de muestras grandes. Si se viola gravemente el supuesto de normalidad, existen alternativas no paramétricas al estadístico F.

La igualdad de varianzas, establece que la varianza dentro de una población debe ser homogénea para todas las poblaciones. Las desviaciones de este supuesto pueden afectar seriamente al nivel de significación y a la potencia de la prueba.

📚Tema 2.3 Pruebas Post-Hoc

⏳ 20 minutos de Clase.

Una vez que realizamos el análisis ANOVA y hemos encontrado una diferencia significativa entre los grupos, debemos responde a la pregunta de cuáles son los grupos que difieren del resto. Para determinar esto debemos hacer pruebas entre todos los posibles pares de grupos. Existen diferentes métodos para realizar esto.

Algunas de las pruebas disponibles son:

- Bonferroni
- Hol-Bongerroni
- Duncan
- LSD
- Tukey

La prueba que realizaremos en este curso será la prueba de Tukey.

$$T = q \cdot \sqrt{\frac{MSE}{n}}$$

Donde q esta dado por el valor critico de la tabla q

Llamado por John Tukey compara todos los pares posibles de medias, y se basa en una distribución de rangos estudiados q (esta distribución es similar a la distribución de t de la prueba t. Por lo general el valor q se obtiene de la tabla q y es necesario conocer la constante K, correspondiente al número de grupos y los grados de libertad correspondientes.

📝Actividad. ANOVA en el data set Iris

⏳ 20 minutos.

Utiliza la técnica ANOVA para determinar si la especie varia significativamente la anchura de petalo.

Recuerda validar los supuestos de ANOVA.
Determina normalidad en los datos
Homogeneidad de Varianza
Estadistico F
Prueba de Tukey

❓ Quiz

Preguntas:

1. Si el estadístico F es mayor a 0.05 debemos:
    a. rechazar H0 porque hay pruebas de que todas las medias difieren
    b. rechazar H0 porque hay pruebas de que al menos una de las medias difiere de las demás
    c. no rechazar H0 porque no hay pruebas de una diferencia en las medias
    d. no rechazar H0 porque una media es diferente de las demás

2. La fórmula del estadístico F corresponde a :
    a. MSW/MSA
    b. SSW/SSA
    c. MSA/MSW
    d. SSA/SSW

3. En ANOVA la hipotesis nula corresponde a
    a. Todas las medias de la población son diferentes
    b. Algunas de las medias de la población son diferentes
    c. Algunas de las medias de la población son iguales
    d. Todas las medias de la población son iguales

Considera la siguiente tabla con información sobre un modelo ANOVA:

Fuente de Variación	DF	Suma de Cuadrados	Cuadrados Medios	*F*
Entre Grupos		6,536	1,634.0	12.51
Intra Grupos	95		130.6	
Total	99	18,943		

**DF corresponde a grados de libertad*

4. El valor para grados de libertad entre grupos es:
    a. 3
    b. 4
    c. 12
    d. 16
5. El valor de la suma de cuadrados intra grupos es:
    a. 12407
    b. 95
    c. 130.6
    d. 4
6. Una empresa de alquiler de coches quiere seleccionar un paquete de software para su sistema de reservas. Hay tres paquetes de software (A, B y C) disponibles en el mercado. La empresa de alquiler de coches elegirá el paquete que tenga el menor número medio de arrendatarios para los que no haya un coche disponible en el momento de la recogida. Se establece un experimento en el que cada paquete se utiliza para hacer reservas durante cinco semanas seleccionadas al azar. ¿Cómo deben analizarse los datos?
    a. prueba chi-cuadrada
    b. ANOVA
    c. prueba t para las diferencias de medias