Contrastes de hipótesis

Clase del 3/12/2024

Contenidos

Objetivos

El **análisis de la varianza (ANOVA)** es una técnica estadística que permite determinar si existen diferencias significativas entre las medias de dos o más grupos o poblaciones. Su objetivo principal es evaluar si las variaciones observadas en los datos se deben a factores específicos (efecto de tratamientos o grupos) o a la variabilidad aleatoria inherente a la muestra.

Contraste

- **Hipótesis nula (H_0)**: Todas las medias poblacionales son iguales ($\mu_1 = \mu_2 = \mu_3 = \ldots$).
- **Hipótesis alternativa (H_a)**: Al menos una media poblacional es diferente.

Idea

- **Variación entre grupos**: Diferencias en las medias de los grupos (atribuidas al factor de interés).
- **Variación dentro de los grupos**: Variabilidad natural o aleatoria entre los datos de cada grupo.
- **Cálculo principal**: El ANOVA descompone la variabilidad total en dos componentes:

$$\text{Suma de cuadrados total (SST)} = \text{Suma de cuadrados entre grupos (}$$

Luego se compara la proporción de estas variaciones utilizando la estadística F:

$$F = \frac{\text{Varianza entre grupos}}{\text{Varianza dentro de los grupos}}.$$

Fuentes de variación

1. $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$
2. $SSB = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$
3. $SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$

$$SST = SSB + SSW$$

Fuentes de variación

- **Suma de cuadrados total (SST)** Es la variación total de los datos con respecto a la media global ($\bar{X}$).
- **Suma de cuadrados entre grupos (SSB)** Calcula la variación debido a las diferencias entre las medias de los grupos ($\bar{X}_i$) y la media global ($\bar{X}$):
- **Suma de cuadrados dentro de los grupos (SSW)** Calcula la variación dentro de cada grupo alrededor de su propia media ($\bar{X}_i$):

- X_{ij}: Observación j del grupo i.
- $\bar{X}$: Media global de todas las observaciones.
- k: Número de grupos; n_i: Tamaño del grupo i.
- $\bar{X}_i$: Media del grupo i.

Tabla ANOVA

Cuadro: *Tabla ANOVA*

Fuente de variación	Gr. de libertad	Suma cuadr.	Cuadrados medios	F (k-1,n-k)
SCE	$k-1$	SSB	$MS_{\text{entre}} = \frac{SS_{\text{entre}}}{k-1}$	$F = \frac{MS_{\text{entre}}}{MS_{\text{dentro}}}$
SCD	$N-k$	SSW	$MS_{\text{dentro}} = \frac{SS_{\text{dentro}}}{N-k}$	
SCT	$N-1$	SST		

Estadístico

La estadística F se calcula como la proporción de la varianza entre grupos a la varianza dentro de los grupos:

$$F = \frac{\text{MSB}}{\text{MSW}}$$

donde:

- $\text{MSB} = \frac{SSB}{k-1}$: Media de los cuadrados entre grupos.
- $\text{MSW} = \frac{SSW}{N-k}$: Media de los cuadrados dentro de los grupos.

Aquí, N es el número total de observaciones y k el número de grupos.

Interpretación

Si el valor de F es grande y significativo (según un nivel de significancia, como $\alpha = 0.05$), se rechaza H_0 y se concluye que hay diferencias significativas entre al menos dos medias.

Supuestos

- Los datos dentro de cada grupo son aproximadamente normales.
- Las varianzas son homogéneas (igualdad de varianzas entre los grupos).
- Las observaciones son independientes entre sí.

Tipos

- **ANOVA de un factor**: Un solo factor o variable independiente (por ejemplo, analizar el efecto de diferentes fertilizantes sobre el crecimiento de plantas).
- **ANOVA de dos factores**: Dos factores o variables independientes (por ejemplo, estudiar el efecto de fertilizantes y tipo de suelo).
- **ANOVA de medidas repetidas**: Mismo grupo de sujetos evaluado en diferentes condiciones.

Campo Médico

Se desea analizar la eficacia de tres tratamientos médicos (**A**, **B** y **C**) para reducir la presión arterial. Se miden los cambios en la presión arterial (en mmHg) después de aplicar los tratamientos en 30 pacientes por grupo.
Los datos generados son:

Paciente	**Tratamiento A**	**Tratamiento B**	**Tratamiento C**
1	10.99	13.00	7.18
2	9.52	11.45	8.55
3	10.65	12.82	10.59
...	...	...	...
30	9.46	12.40	7.99

Cuadro: *Muestra de los datos recogidos.*

- Hipótesis nula (H_0): Las medias de reducción de presión arterial son iguales para todos los tratamientos ($\mu_A = \mu_B = \mu_C$).
- Hipótesis alternativa (H_1): Al menos una media de reducción es diferente.

El ANOVA descompone la variabilidad total en dos componentes:

Suma de cuadrados total (SST) = Suma de cuadrados entre grupos (SSB

Se evalúa la relación entre estas componentes utilizando la estadística F:

$$F = \frac{\text{Varianza entre grupos}}{\text{Varianza dentro de los grupos}}.$$

Los resultados del ANOVA son:

- F-Estadístico: 29.64
- Valor-p: 1.52×10^{-10}

Dado que el valor-p es mucho menor que 0.05, se rechaza la hipótesis nula (H_0). Esto implica que hay diferencias significativas en las medias de reducción de presión arterial entre los tres tratamientos.

Tratamiento	**Media (mmHg)**	**Desviación Estándar (mmHg)**
A	10.06	2.04
B	12.02	1.86
C	8.04	1.81

Cuadro: *Resumen estadístico por tratamiento.*

Dado que el valor-p es mucho menor que 0.05, se rechaza la hipótesis nula (H_0). Esto implica que hay diferencias significativas en las medias de reducción de presión arterial entre los tres tratamientos.

Homogeneidad de varianzas

- H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
- H_1: Al menos una varianza es significativamente diferente de las demás.

Cochran: $n_1 = n_2 = \ldots = n_k$)

Se utiliza el estadístico de Cochran que se define como:

$$C = \frac{\max(S_i^2)}{\sum_{i=1}^{k} S_i^2}$$

Donde:

- S_i^2: Varianza de la i-ésima muestra.
- k: Número de grupos.

El valor de C representa la proporción de la mayor varianza respecto a la suma total de varianzas.

Cochran: Criterio de decisión

1. Se compara el valor calculado de C con un valor crítico de la **tabla de Cochran**, que depende del nivel de significancia (α), el número de grupos (k) y el tamaño de muestra de cada grupo (n).
2. **Rechazar** H_0 si $C > C_{\text{crítico}}$, lo que indica que las varianzas no son homogéneas.
3. **No rechazar** H_0 si $C \leq C_{\text{crítico}}$, lo que sugiere que no hay evidencia para afirmar que las varianzas son heterogéneas.

Cochran:Aplicaciones

- Es adecuado cuando se tienen múltiples grupos y se desea detectar varianzas atípicas.
- Es simple y efectivo en situaciones donde los datos cumplen con los supuestos de normalidad.

1. **Diseños experimentales**: Para comprobar la homogeneidad de varianzas antes de aplicar un ANOVA.
2. **Análisis en calidad**: Evaluación de consistencia en procesos industriales.
3. **Investigaciones biomédicas**: Comparación de variabilidad entre tratamientos o condiciones.

Barlett: No es necesario que $n_1 = n_2 = \ldots = n_k$)

Estadístico de Barlett:

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)}$$

- $N = \sum_{i=1}^{k} n_i$: Tamaño total de la muestra.
- k: Número de grupos.
- n_i: Tamaño de la muestra del i-ésimo grupo.
- S_i^2: Varianza del i-ésimo grupo.
- $S_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{N-k}$: Varianza ponderada.

Barlett: Criterio de decisión

1. Comparar el valor calculado de χ^2 con el valor crítico de la distribución chi-cuadrado para un nivel de significancia (α) y $k-1$ grados de libertad.
2. **Rechazar** H_0 si $\chi^2_{\text{calculado}} > \chi^2_{\text{crítico}}$, indicando que las varianzas no son homogéneas.
3. **No rechazar** H_0 si $\chi^2_{\text{calculado}} \leq \chi^2_{\text{crítico}}$, sugiriendo que las varianzas son homogéneas.

Método de Bonferroni: Detección de diferencias

En el ANOVA, el resultado significativo indica que al menos una media es diferente, pero no especifica qué pares de medias son diferentes. Para identificar las diferencias específicas, se realizan comparaciones múltiples.

- H_0: $\mu_i = \mu_j$
- H_1: $\mu_i \neq \mu_j$

Bonferroni: Explicación de parámetros

Para cada par de grupos (i, j), NO se rechaza H_0 si:

$$\frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MSE}{n_i} + \frac{MSE}{n_j}}} < t_{n-k, \frac{\alpha}{K}}$$

- $\bar{X}_i, \bar{X}_j$: Medias de los grupos i y j.
- MSE: Error cuadrático medio (MSW) del ANOVA.
- n_i, n_j: Tamaños de las muestras de los grupos i y j.
- N. de comparaciones: $K = \binom{k}{2}$

Ejemplo práctico

Supongamos que se realiza un ANOVA con 3 grupos (A, B, C), y se obtiene

1. $SSW = 2.5$
2. $n = 10$ por grupo.
3. $\bar{X}_A = 15, \bar{X}_B = 12, \bar{X}_C = 18$
4. $K = \binom{3}{2} = 3.$
5. $\alpha_{\text{ajustado}} = \frac{0.05}{3} \approx 0.0167$

Valor crítico: Tenemos $df = 27$ y $\alpha_{\text{ajustado}} = 0.0167$, $t_{27,0.0167} \approx 2.77.$

Resultados

- $A - B$:

$$t = \frac{|15 - 12|}{\sqrt{\frac{2.5}{10} + \frac{2.5}{10}}} = \frac{3}{\sqrt{0.5}} = 4.24 > 2.77$$

- $A - C$:

$$t = \frac{|15 - 18|}{\sqrt{\frac{2.5}{10} + \frac{2.5}{10}}} = \frac{3}{\sqrt{0.5}} = 4.24 > 2.77$$

- $B - C$:

$$t = \frac{|12 - 18|}{\sqrt{\frac{2.5}{10} + \frac{2.5}{10}}} = \frac{6}{\sqrt{0.5}} = 8.49 > 2.77$$

Se concluye que hay diferencias significativas entre cada par de grupos.

Ejemplo 1 IA

Sobre N conjuntos de datos, aplicacamos tres algoritmos distintos (SVM, Random Forest y KNN) y queremos evaluar si hay un rendimiento significativamente mejor que los demás en algún algoritmo. Para ello se realiza una ANOVA para analizar si existen diferencias significativas en las precisiones promedio de los algoritmos evaluados en varios conjuntos de datos.

Ejemplos 2 IA

Para analizar el rendimiento de una red neuronal con una ANOVA de dos factores para analizar la interacción entre la tasa de aprendizaje (0.01, 0.1, 0.5) y el número de neuronas por capa (16, 32, 64) en términos de la pérdida promedio en el conjunto de validación.

Ejemplos 2 IA

Existen distintas técnicas de preprocesamiento de datos (normalización,estandarización, imputación de valores faltantes), lo que podría afectar al rendimiento de un modelo de IA. Con un ANOVA de un factor podríamos comparar la precisión promedio de un modelo entrenado con tres técnicas distintas de preprocesamiento de datos, identificando qué técnica produce mejores resultados.