Análisis de la varianza

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025

Universidad
Rey Juan Carlos

Introducción

- En temas previos hemos comparado las medias de dos poblaciones
  $H_0 : \mu_1 = \mu_2$
  $H_1 : \mu_1 \neq \mu_2$
- ¿Y si queremos comprobar si hay diferencia entre 3 o más medias muestrales? Es decir, comprobar la hipótesis de que todas las medias (3 o más) son iguales –> **ANOVA**

ANOVA

- **Análisis de la varianza**, conocido como **ANOVA** (del inglés *Analysis of Variance*)
- Técnica estadística utilizada para comparar las medias de dos o más grupos y determinar si existen diferencias significativas entre ellos
- Desarrollado por Fisher en las primeras décadas del siglo XX.
- La idea central es analizar la variabilidad de los datos y dividirla en componentes atribuibles a diferentes fuentes de variación

- En su forma más simple, el ANOVA se utiliza para probar hipótesis sobre las diferencias entre las medias de grupos
- Por ejemplo, si quisiéramos comparar el rendimiento de tres tipos diferentes de métodos educativos, podríamos usar ANOVA para determinar si el método educativo tiene un efecto significativo en el rendimiento académico:

Tipo de método	Rendimiento académico
Método 1	85, 78, 90, 82
Método 2	88, 79, 91, 85
Método 3	80, 75, 88, 83

Procedimiento general

1. **Formulación de hipótesis**: Se establece una hipótesis nula que indica que no hay diferencias entre las medias de los grupos y una hipótesis alternativa que sugiere que al menos una media es diferente
   $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
   H_1 : al menos una de las medias es distinta
2. **Cálculo de la varianza**: Se calculan dos tipos de varianza: la varianza dentro de los grupos (variabilidad debido a diferencias dentro de los mismos grupos) y la varianza entre los grupos (variabilidad debido a diferencias entre los grupos)

Procedimiento general

3. **F-test**: Se realiza una prueba F de Fisher para evaluar la relación entre las varianzas. Si la varianza entre los grupos es significativamente mayor que la varianza dentro de los grupos, esto sugiere que hay diferencias significativas entre las medias de los grupos
4. **Análisis de resultados**: Si la prueba F indica que hay diferencias significativas, se pueden realizar pruebas adicionales para identificar entre qué grupos existen estas diferencias.

- Es una herramienta poderosa porque permite comparaciones múltiples mientras controla la tasa de error tipo I
- Es ampliamente utilizado en experimentos donde se comparan tratamientos o condiciones en diferentes grupos o en diferentes momentos

Distribución F (Fisher-Snedecor)

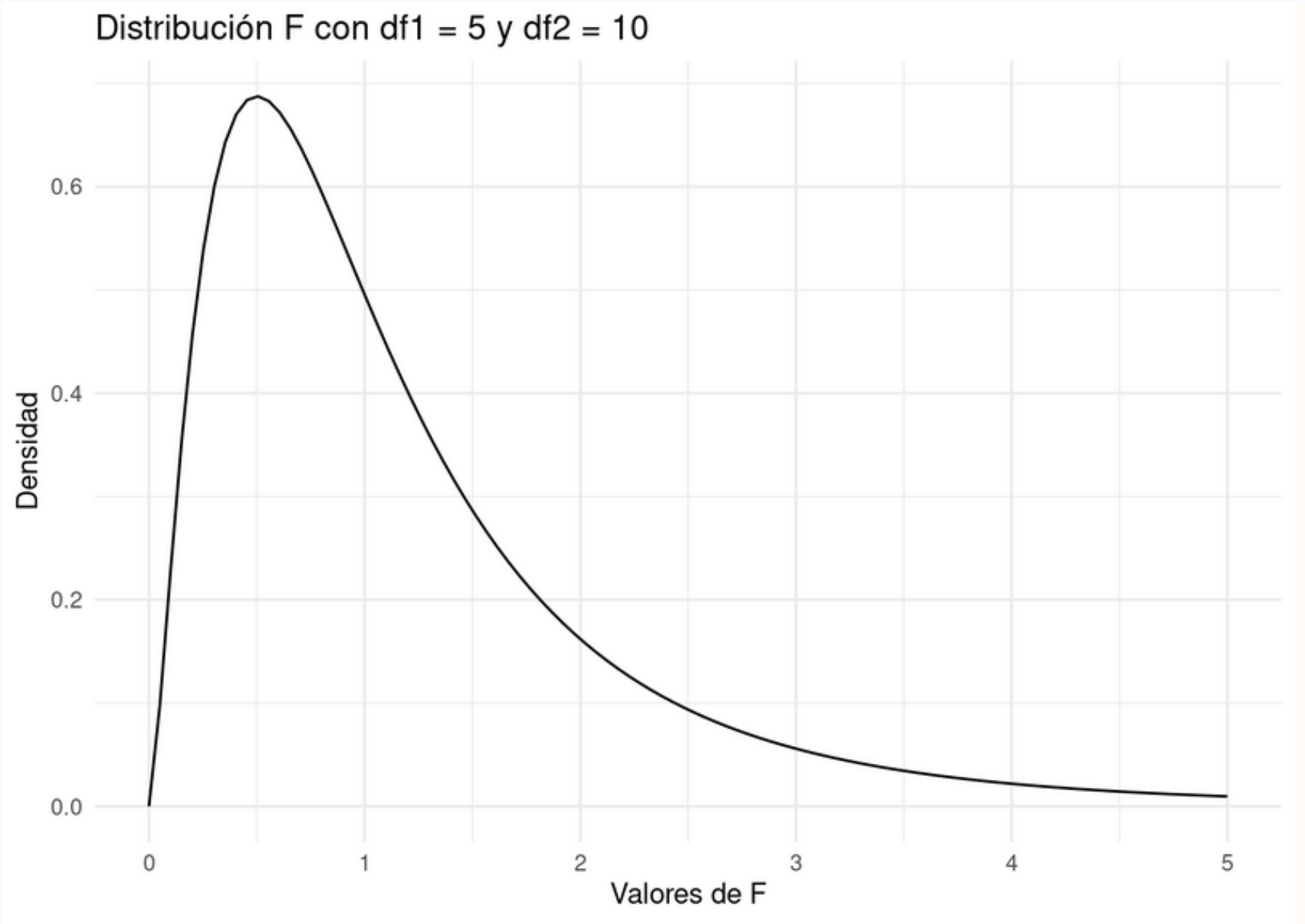
- Necesitamos una distribución muestral para comparar varianzas
- Se toma el estadístico cociente S_X^2/S_Y^2. Cuanto más próximo a 1, más parecidas las varianzas
- La distribución muestral de S_X^2/S_Y^2 es la distribución F
- Dadas dos muestras $\mathbf{X} = (X_1, \ldots, X_{n+1})$, $\mathbf{Y} = (Y_1, \ldots, Y_{m+1})$ de tamaño n y m, procedentes de dos poblaciones Normales con varianzas σ_X^2 y σ_Y^2. Entonces

$$\frac{nS_X^2}{\sigma_X^2} \sim \chi_n^2 \qquad \frac{mS_Y^2}{\sigma_Y^2} \sim \chi_m^2$$

y

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma^2 Y} \sim F_{m,n}$$

Distribución F



Modelo con un factor

- Se utiliza cuando se estudia el efecto de un solo factor (variable independiente) en una variable dependiente continua
- Permite comparar las medias de varios grupos para determinar si existen diferencias significativas entre ellos
- **Hipótesis**:
  - **Hipótesis Nula (H_0)**: Todas las medias de los grupos son iguales ($\mu_1 = \mu_2 = \cdots = \mu_k$).
  - **Hipótesis Alternativa (H_1)**: Al menos una de las medias de los grupos es diferente.

Modelo con un factor

Tenemos una variable aleatoria Y que toma valores reales y una variable cualitativa o factor X con k niveles $1, 2, \ldots, i, \ldots, k$. La variable Y toma valores $Y_{ij}, j = 1, \ldots, n_i$ en el nivel i del factor X, siendo n_i el número de observaciones en el nivel i del factor X

X_1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
$\vdots$	$\vdots$
X_i	$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$
$\vdots$	$\vdots$
X_k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$

Modelo con un factor

Tenemos los siguientes supuestos:

- Normalidad: Las distribuciones de las poblaciones de las que provienen las muestras son normales. Se supone que los errores ϵ_{ij} están distribuidos como una Normal de media 0 y varianza σ^2
- Homogeneidad de varianzas: Las varianzas de las poblaciones son iguales
- Independencia: Las observaciones son independientes entre sí

Modelo con un factor

El modelo teórico es como sigue:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Donde:

- Y_{ij} es la observación j-ésima del grupo i-ésimo
- μ es la media general
- ϵ_{ij} es el término de error aleatorio, $\epsilon_{ij} \sim N(0, \sigma^2)$
- τ_i es el efecto del grupo i-ésimo en la media de la variable respuesta Y. Esto es, cuánto aumenta o disminuye la media de Y por pertenecer la observación a la categoría i. De modo que podemos llamar

$$Y_i = \mu + \tau_i$$

al efecto medio del grupo i-ésimo.

La suma de las diferencias al cuadrado de cada dato respecto a la media general se calcula como sigue:

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

donde $\bar{Y}_{\cdot\cdot}$ es la media general de todas las observaciones.

Teniendo en cuenta que: $Y_{ij} - \bar{Y}_{..} = Y_i + \epsilon_{ij} - \bar{Y}_{..}$

Podemos descomponer la suma de cuadrados, como sigue:

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \underbrace{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y}_{..})^2}_{SSB} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}_{SST}$$

¿Qué son SSB y SST? ¿Qué interpretación le dais?

- SSB: La varianza entre grupos se calcula como la suma de las diferencias al cuadrado de las medias de los grupos respecto a la media general, ponderada por el tamaño de los grupos:

$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y}_{..})^2$$

  donde $\bar{Y}_i$ es la media del grupo i.
- SSW: La varianza dentro de los grupos es la suma de las diferencias al cuadrado de cada dato respecto a la media de su grupo

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Esto es, se descompone la variabilidad total de los datos en dos componentes, SSB que refleja la diferencia de cada grupo respecto a la media global y SSW que refleja la variabilidad intrínseca dentro de cada grupo:

$$\text{SST} = \text{SSB} + \text{SSW}$$

Nota:

- SST: Sum of squares total
- SSB: Sum of squares between
- SSW: Sum of squares within

Cálculo del Estadístico F:

$$F = \frac{\text{Varianza Entre Grupos (MSB)}}{\text{Varianza Dentro de los Grupos (MSW)}}$$

Donde:

- MSB (Mean Square Between): Media cuadrática entre grupos.
- MSW (Mean Square Within): Media cuadrática dentro de los grupos.

Esto es:

$$F = \frac{SSB/df_B}{SSW/df_W}$$

siendo $df_B = k - 1$ son los grados de libertad entre los grupos y $df_W = N - k$ son los grados de libertad dentro de los grupos y N es el número total de observaciones.

Una vez se dispone de toda esta información, es común representarla en forma de tabla, en la llamada *Tabla ANOVA*:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Disferencias entre grupos	SSB	k-1	MSB
Diferencias dentro de los grupos, Residual o Error	SSW	N-k	MSW
Total	SST	N-1	

El estadístico de prueba $F \sim F_{df_B, df_W}$ bajo la hipótesis nula de igualdad de medias.

El $p - valor$ se obtiene a partir de la distribución F, considerando los grados de libertad de los numeradores y denominadores. Esto es:

$$p - valor = P(F_{df_b, df_W} > F_{muestral})$$

Como en otros contrastes, si el $p - valor$ es menor que el nivel de significancia α, se rechaza la hipótesis nula, concluyendo que al menos una de las medias de los grupos es diferente.

Modelo con un factor: Ejemplo

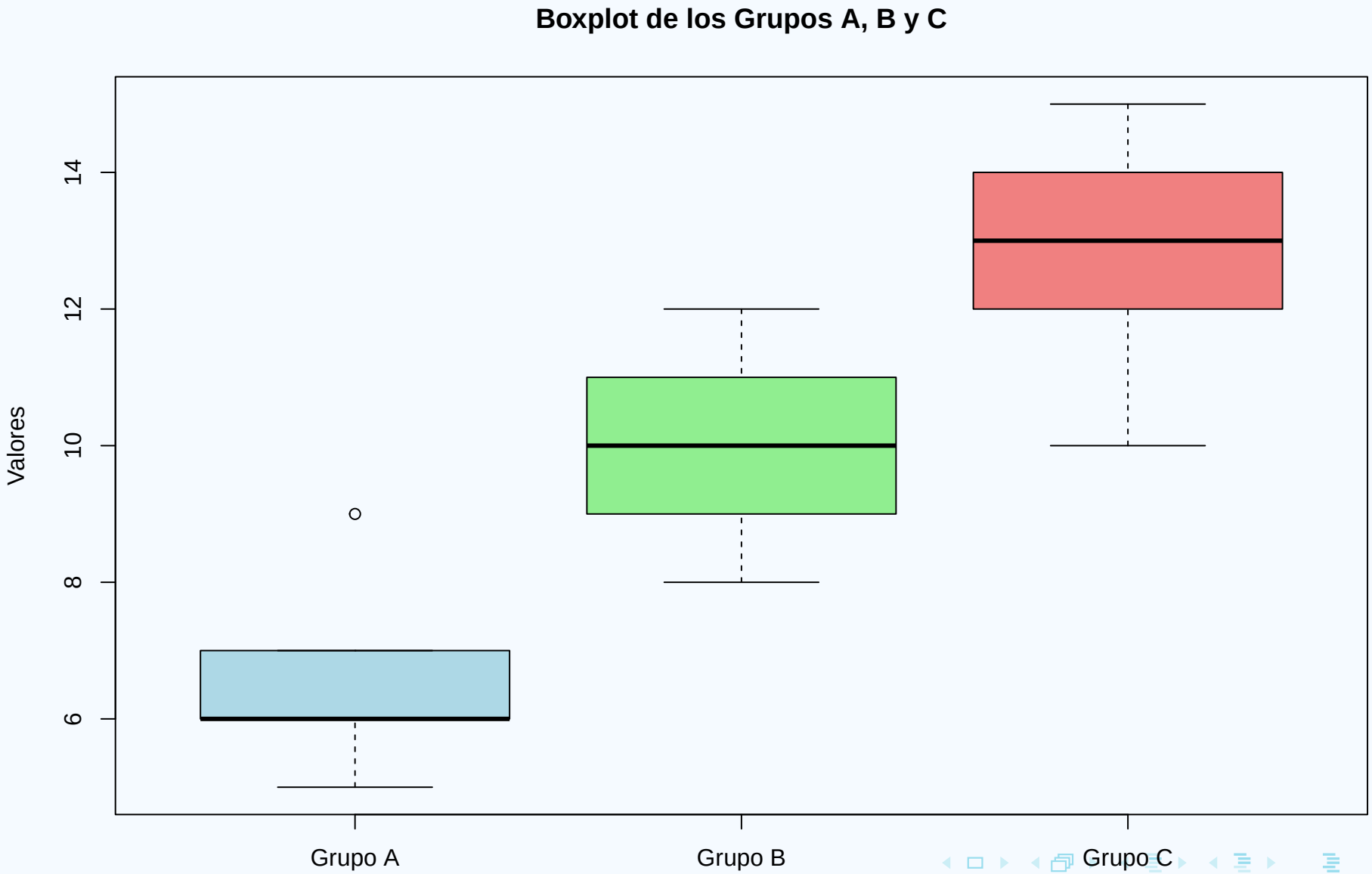
Supongamos que tenemos tres tratamientos (A, B y C) y sus correspondientes muestras de datos son:

- Grupo A: $[5, 7, 6, 9, 6]$
- Grupo B: $[8, 12, 9, 11, 10]$
- Grupo C: $[14, 10, 13, 15, 12]$

Nuestro objetivo es determinar si existe una diferencia significativa entre las medias de estos tres grupos.

Modelo con un factor: Ejemplo

Hagamos un boxplot con los tres tratamientos:



Modelo con un factor: Ejemplo

Comenzamos calculando la media de cada grupo y la media general:

- Media de Grupo A $\bar{Y}_A = \frac{5+7+6+9+6}{5} = \frac{33}{5} = 6.6$
- Media de Grupo B $\bar{Y}_B = \frac{8+12+9+11+10}{5} = \frac{50}{5} = 10$
- Media de Grupo C $\bar{Y}_C = \frac{14+10+13+15+12}{5} = \frac{64}{5} = 12.8$
- Media General $\bar{Y}$:

$$\bar{Y} = \frac{6.6 + 10.0 + 12.8}{3} = \frac{29.4}{3} = 9.8$$

Modelo con un factor: Ejemplo

Calculemos ahora cada uno de los componentes de SST = SSB + SSW

Comencemos por SSB:

$$\text{SSB} = n_A(\bar{Y}_A - \bar{Y})^2 + n_B(\bar{Y}_B - \bar{Y})^2 + n_C(\bar{Y}_C - \bar{Y})^2$$

siendo $n_A = n_B = n_C = 5$ (número de observaciones en cada grupo).

$$\text{SSB} = 5(6.6 - 9.8)^2 + 5(10.0 - 9.8)^2 + 5(12.8 - 9.8)^2 = 96.4$$

Modelo con un factor: Ejemplo

Calculemos a continuación la suma de los cuadrados dentro de los grupos:

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Para cada grupo, calculamos la suma de las diferencias al cuadrado entre cada dato y la media del grupo:

- Grupo A:
  $(5-6.6)^2 + (7-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (6-6.6)^2 = 9.2$
- Grupo B:
  $(8-10)^2 + (12-10)^2 + (9-10)^2 + (11-10)^2 + (10-10)^2 = 10$
- Grupo C:
  $(14-12.8)^2 + (10-12.8)^2 + (13-12.8)^2 + (15-12.8)^2 + (12-12.8)^2 = 14.8$

Con ello,

$$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = 9.2 + 10.0 + 14.8 = 34.0$$

Por tanto que la Suma Total de Cuadrados (SST) es:

$$\text{SST} = \text{SSB} + \text{SSW} = 96.4 + 34.0 = 130.4$$

Modelo con un factor: Ejemplo

Para realizar el contraste, calculamos el estadístico F

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/df_B}{\text{SSW}/df_W}$$

Los grados de libertad son:

- Grados de libertad entre los grupos: $df_B = k - 1 = 3 - 1 = 2$
- Grados de libertad dentro de los grupos: $df_W = N - k = 15 - 3 = 12$

Luego, el valor del estadístico del contraste es

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/df_B}{\text{SSW}/df_W} = \frac{96.4/2}{34/12} = \frac{48.2}{2.83} = 17.01$$

El p-valor se obtiene utilizando la distribución F-Snedecor con $df_B = 2$ y $df_W = 12$.

El comando en R es `pf(17.01, df1 = 2, df2 = 12, lower.tail = FALSE)` que devuelve un p-valor de 0.00031. Como es menor que el grado de significancia $\alpha = 0.05$, indica una diferencia significativa entre los grupos.

Modelo con un factor: Ejemplo

```r
grupo_a <- c(5, 7, 6, 9, 6)
grupo_b <- c(8, 12, 9, 11, 10)
grupo_c <- c(14, 10, 13, 15, 12)

datos <- data.frame(
  valores = c(grupo_a, grupo_b, grupo_c),
  grupo = factor(rep(c("Grupo A", "Grupo B", "Grupo C"), each

anova = aov(datos$valores ~ datos$grupo)
summary(anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
datos$grupo   2   96.4   48.20   17.01 0.000314 ***
Residuals    12   34.0    2.83
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo ANOVA con dos factores

- Se utiliza cuando se estudian dos factores simultáneamente para evaluar su efecto individual y conjunto en una variable dependiente
- Puede verse como una generalización del caso de ANOVA con un único factor
- Este modelo es más complejo y permite entender no solo los efectos principales de cada factor, sino también si hay una interacción entre ellos

Modelo ANOVA con dos factores

Sean A y B dos factores que se desean estudiar, con m_A y m_B niveles. Trabajaremos con las siguientes hipótesis nulas:

Opciones de hipótesis:

- Hipótesis nula para los efectos principales H_0:
  - No hay efecto del primer factor
  - No hay efecto del segundo factor
- Hipótesis nula para la interacción H_0: No hay interacción entre los dos factores

Modelo con dos factores sin interacción

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

Donde:

- Y_{ijk} es la observación k-ésima del nivel j-ésimo del factor B y nivel i-ésimo del factor A
- μ es la media general
- α_i es el efecto del nivel i-ésimo del factor A
- β_j es el efecto del nivel j-ésimo del factor B
- ϵ_{ijk} es el término de error aleatorio, $\epsilon_{ijk} \sim N(0, \sigma^2)$

Modelo con dos factores sin interacción

En este caso, la tabla ANOVA queda como sigue:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Diferencias entre niveles del factor A	SSB_A	$m_A - 1$	MSB_A
Diferencias entre niveles del factor B	SSB_B	$m_B - 1$	MSB_B
Error	SSW	$N - m_A - m_B + 1$	MSW
Total	SST	$N - 1$	

Modelo con dos factores sin interacción

Para estudiar la importancia de cada factor se calcula el estadístico F particular para cada uno de ellos como sigue:

$$F_A = \frac{MSB_A}{MSW} \sim F_{m_A-1, N-m_A-m_B+1}$$

y

$$F_B = \frac{MSB_B}{MSW} \sim F_{m_B-1, N-m_A-m_B+1}$$

A partir de estos estadísticos de prueba podemos contrastar las hipótesis nulas de no existencia de efectos asociados a los factores A y B respectivamente.

Modelo con dos factores con interacción

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Donde $(\alpha\beta)_{ij}$ representa el efecto de interacción entre el nivel i-ésimo del factor A y el nivel j-ésimo del factor B

Modelo con dos factores con interacción

En este caso, la tabla ANOVA añade el factor de interacción:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Diferencias entre niveles del factor A	SSB_A	$m_A - 1$	MSB_A
Diferencias entre niveles del factor B	SSB_B	$m_B - 1$	MSB_B
Diferencias debidas la interacción	SSB_{AB}	$(m_A - 1) * (m_B - 1)$	MSB_{AB}
Error	SSW	$N - m_A * m_B$	MSW
Total	SST	$N - 1$	

Modelo con dos factores con interacción

Para estudiar la importancia de cada de la interacción se calculan el estadístico F correspondiente:

$$F_{AB} = \frac{MSB_{AB}}{MSW} \sim F_{(m_A-1)*(m_B-1),N-m_A*m_B}$$

Con este estadístico, contrastamos la hipótesis nula de no existencia de interacción entre los dos factores A, B:

- Si podemos rechazar esa hipótesis, es decir, si existe interacción entre los factores, entonces hemos terminado. Es decir, no podemos eliminar ningún factor del modelo.
- En cambio, si no rechazamos la hipótesis nula, es decir, si no existe interacción entre los factores, podemos eliminar dicho efecto (la interacción) del modelo y pasar a un modelo sin interacción (como el previo)

Modelo con dos factores con interacción

Se ha realizado un estudio para evaluar el efecto de dos factores sobre la calidad del sueño:

- **Factor A**: Tipo de rutina antes de dormir (con tres niveles: "Leer", "Meditar" y "Uso de pantallas").
- **Factor B**: Ambiente de sueño (con dos niveles: "Silencio" y "Con ruido").

Cada combinación de los factores tiene dos observaciones para medir la calidad del sueño (con puntuaciones del 1 al 10, siendo 10 la mejor calidad). Los datos son los siguientes:

	Silencio	Con ruido
Leer	7, 8	5, 6
Meditar	9, 10	6, 7
Uso de pantallas	4, 5	3, 4

Comparaciones múltiples

- Con ANOVA podemos rechazar la siguiente hipótesis nula: $H_0 : \mu_1 = \mu_2 \ldots = \mu_k$, siendo k el número de niveles en el factor
- ¿En qué niveles del factor se encuentran las principales diferencias? Es decir, qué hipotesis (una o varias) de las siguientes son rechazadas:

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 = \mu_3$$

$$\ldots$$

$$H_0 : \mu_{k-1} = \mu_k$$

Comparaciones múltiples

- En un caso con k niveles en el factor, hay $k * (k-1)/2$ posibles contrastes de igualda de medias
- Si realizamos todos esos contrastes, aumenta la probabilidad de cometer errores de tipo I (rechazar incorrectamente la hipótesis nula)
- Este fenómeno se conoce como el problema de las **comparaciones múltiples**

Comparaciones múltiples

- Cuando realizamos una sola prueba de hipótesis, establecemos un nivel de significancia predeterminado (ejemplo $\alpha = 0.05$)
- Esto es, aceptamos una probabilidad de error de tipo I del 5%, es decir, hay un 5% de probabilidad de rechazar incorrectamente la hipótesis nula cuando es verdadera
- Cuando se realizan múltiples test de hipótesis, la probabilidad acumulada de cometer al menos un error de tipo I aumenta con cada prueba adicional
  - Ejemplo: Con 10 tests de hipótesis independientes, cada uno con un nivel de significancia de $\alpha = 0.05$, la probabilidad de cometer al menos un error de tipo I aumenta a más del 40% $(1 - (1 - 0.05)^{10} \approx 0.40)$

Comparaciones múltiples

- Solución: Aplicar correcciones cuando se realizan comparaciones múltiples para controlar este aumento en el riesgo de error
- Existen varios métodos para controlar el problema de las pruebas múltiples:
  - Bonferroni, Holm-Bonferroni, LSD (Least Significant Differences), Tuley HSD (Honestly-significant-difference), entre otros
- Estos métodos controlan la tasa global de error de tipo I para todas las comparaciones realizadas, manteniendo un nivel de significancia general específico

Método de Bonferroni

- Este método es relativamente simple y conservador
- Idea: ajustar el nivel de significancia individual para cada prueba de hipótesis realizada. En lugar de utilizar un nivel de significancia estándar (por ejemplo, $\alpha = 0.05$), se divide el nivel de significancia global deseado por el número total de pruebas realizadas (m):

$$\alpha' = \frac{\alpha}{m}$$

  Esta división produce un nivel de significancia más estricto (α') para cada prueba individual, lo que ayuda a controlar el riesgo global de error de tipo I
- Se utiliza el nivel de significancia individual ajustado para cada prueba de hipótesis. Si el $p - valor$ de una prueba es menor que el nivel de significancia ajustado, se rechaza la hipótesis nula de la prueba

Método de Bonferroni

- Fácil de entender e implementar
- Proporciona un control conservador sobre el error de tipo I en comparaciones múltiples
- Puede ser un método demasiado conservador en situaciones donde se realizan muchas comparaciones, lo que puede resultar en una pérdida de potencia estadística

Referencias

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

Canavos, G. C., & Medal, E. G. U. (1987). *Probabilidad y estadística* (p. 651). México: McGraw Hill.