Trabajo Práctico III: Estudio de simulación sobre los supuestos de Anova

---

Aclaraciones:

- El TP se puede hacer en grupos de **tres o cuatro personas**.
- La resolución del mismo debe incluir los razonamientos, gráficos y código necesario. Se debe presentar en formato Notebook. Todo el código debe correr sin errores y dar los mismos resultados.
- La fecha límite de entrega es el **Jueves 30 de Junio**.

En este trabajo de deben analizar la importancia de los supuestos de Anova en el caso de muestras balanceadas. Consideramos una escenario con tres grupos donde se quiere testear si la medias de los mismos son iguales o si alguna de ellas es distinta. Para índices $1 \leq i \leq 3$ e $1 \leq j \leq J$ las variables que componen la muestra son

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Aqui μ_i es la media del grupo i y ε_{ij} un componente de error aleatorio para cada observación.

Los supuestos necesarios para ANOVA son:

- Normalidad de los errores:
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{ij}).$$
- Homocedasticidad: Los terminos de error tiene la misma varianza.
$$\sigma^2_{ij} = \sigma^2$$
- Independencia de los errores: Los errores son todos independientes.

1. Describa la hipótesis nula y la hipótesis alternativa del test.
2. Utilizando las función `escenarioI(medias, J)` ya definida en el script de R genere una muestra con $I = 3$ grupos y de $J = 20$ observaciones por grupo. Utilice los valores de media $\mu_1 = \mu_2 = \mu_3 = 0$. Esta función genera muestras con varianza por grupo $\sigma^2 = 3$.

   Aclaración: En el resto del trabajo utilice la misma estructura para generar muestras. Es decir, una matriz con $I = 3$ columnas y J filas. Donde cada columna es la muestra obtenida para cada grupo.
3. Con la muestra obtenida corrobore que se satisface la identidad

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \overline{Y}_{..}\right)^2}_{S_{tot}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \overline{Y}_{i.}\right)^2}_{S_w} + \underbrace{J\sum_{i=1}^{I}\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2}_{S_b}.$$

4. Defina una función `rechazaH0(muestra)` que tome una muestra como argumento y determina si rechaza H_0 al nivel $\alpha = 0.1$. Recuerde que se rechazará H_0 si y solo si $F_{obs} > F_{I-1,I(J-1),1-\alpha}$, con el estadístico

$$F = \frac{SS_b}{I-1}\frac{I(J-1)}{SS_w}.$$

Supuesto de normalidad

5. Corrobore que el nivel del test es exacto mediante una simulación que satisfaga todos los supuestos con $I = 3$, $J = 20$, $\sigma^2 = 1$ y $\mu_1 = \mu_2 = \mu_3 = 0$. Para esto

   (a) Genere una muestra con la función `escenario1()`.

   (b) Decidir si se rechaza o no H_0.

   (c) Repetir (a) y (b) un total de $N = 10000$ veces.

   (d) Calcule la proporción de veces que se rechaza H_0.

6. Definir una función `escenario2(medias, J)` que reciba como argumento un vector de medias de grupo y un tamaño de muestra J y devuelvan una muestra cuyos errores que siguen una distribución $\mathcal{U}(-3, 3)$

7. Definir una función `escenario3(medias, J)` que reciba como argumento un vector de medias de grupo y un tamaño de muestra J y devuelva una muestra cuyos errores siguan una distribución t de student con 3 grados de libertad.

8. Para cada uno de los tres escenarios, defina una función que reciba un vector de medias y un J y que estime mediante una simulación la proporción de veces que se rechaza la hipótesis nula.

9. Para cada escenario estime el nivel del test cuando las medias son $\mu_1 = \mu_2 = \mu_3 = 0$ y la cantidad de observaciones por grupo son $J = 3, 6, 9, 12, 15, 18, 21, 24, 27, 30$.

10. Para un tamaño de muestra $J = 20$ estime la potencia del test en cada uno de los tres escenarios. Para esto tome la media del primer y del segundo grupo $\mu_1 = \mu_2 = 0$ y la media del tercer grupo tomando valores en una grilla entre 0 y 2. Presente los resultados en un gráfico apropiado.

11. ¿Que conclusiones saca? ¿Que puede decir de la importancia del supuesto de la normalidad de los errores?

Supuesto de homocedasticidad

Vamos a estudiar el nivel de nuestro test si todos los supuestos se cumplen salvo el de homocedasticidad.

12. Defina una función `escenario4(medias, desvios)` que tome como argumento un vector de medias de grupo, un vector de desvíos de grupo y devuelva una muestra de tres grupos con $J = 20$ observaciones y errores normales e independientes con los desvíos indicados para cada grupo.

Sugerencia: No programe más de lo necesario. Adapte el código y las funciones que ya tiene definidas.

13. Defina una función que reciba un vector de medias y un vector de desvíos y que estime mediante una simulación la proporción de veces que se rechaza la hipótesis nula dentro del escenario 4.

14. Dentro del escenario 4, estimar el nivel del test si la medias de los grupos son $\mu_1 = \mu_2 = \mu_3 = 0$, el desvío de los primeros dos grupos es $\sigma_1 = \sigma_2 = 1$ y el desvío del tercer grupo es $\sigma_3 = 1.5$.

15. Aproxime el nivel del test cuando las medias son $\mu_1 = \mu_2 = \mu_3 = 0$ y los desvíos de los primeros dos grupos son $\sigma_1 = \sigma_2 = 1$ y el del tercer grupo toma valores entre 1 y 3. Presente los resultados en un gráfico apropiado.

16. Estime la potencia del test en el siguiente escenario: Una cantidad de observaciones por grupo $J = 20$, la media y desvío del primer y segundo grupo son $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, el desvío del segundo grupo es $\sigma_1 = 1.5$ y la media del tercer grupo toma valores en una grilla entre 0 y 2. Presente los resultados en un gráfico apropiado.

17. ¿Que conclusiones saca? ¿Que puede decir de la importancia del supuesto de homocedasticidad?

Estudio de nivel - Supuesto de independencia

Vamos a estudiar el nivel del test si todos los supuestos se cumplen salvo el de independencia de los errores. Este supuesto se puede quebrar de muchas maneras distintas. En este caso vamos a ver que ocurre si las muestras vienen apareadas. Es decir los $\varepsilon_{1,j}$, $\varepsilon_{2,j}$, $\varepsilon_{3,j}$ son dependientes.

Vamos a utilizar la función `escenario5(medias, dep)` definida en el script de `R` que toma como argumento un vector de medias y un número $0 \leq dep \leq 1$ y devuelve una muestra con $J = 20$ observaciones por grupo con errores correlacionados. El valor de dep mide la fuerza de la dependencia. Si $dep = 0$ los errores son independientes. Si $dep = 1$ los errores están perfectamente correlacionados.

18. Viendo el código de la función `escenario5()` explique como funciona y por qué da muestras que no son independientes.

19. Defina una función que reciba un vector de medias y un valor de dep y que estime mediante una simulación la proporción de veces que se rechaza la hipótesis nula dentro del escenario 5.

20. Para una grilla de valores de dep entre 0 y 1 estime el nivel del test con medias $\mu_1 = \mu_2 = \mu_3 = 0$ y $J = 20$. Muestre estos valores en un gráfico apropiado.

21. Estime la potencia del test en el siguiente escenario: Una cantidad de observaciones por grupo $J = 20$, la media y desvío del primer y segundo grupo son $\mu_1 = \mu_2 = 0$, el valor de $dep = 0.5$ y la media del tercer grupo toma valores en una grilla entre 0 y 2. Presente los resultados en un gráfico apropiado.

22. ¿Que conclusiones saca? ¿Que puede decir de la importancia del supuesto de indendencia de los errores?