CI-0115 Probabilidad y Estadística
Prof. Sebastián Ruiz Blais
I-2024
Grupos 2 y 3

Lab 4: Análisis de varianza de varios grupos

En este laboratorio continuaremos estudiando factores que contribuyen o no a la esperanza de vida de una localidad (país o ciudad). Para preguntas teóricas responda como un comentario en el jupyter notebook. Utilice la mayor cantidad de datos para cada pregunta. Para la pregunta 5, puede utilizar un subconjunto de los datos, con las diferentes alturas.

1. (10%) Trabajando con la base de datos de *life-expectancy-who*, Investigue a qué corresponden las columnas de 'hepatitis B', 'measles' (sarampión), ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditure', 'Diphtheria ', ' HIV/AIDS', 'thinness 1-19 years', ' thinness 5-9 years', 'Income composition of resources' y para cada una explique cuál es el propósito de tenerlas.
2. (10%) Utilice **diagramas de cajas** para visualizar la distribución de las variables "esperanza de vida", "total expenditure" y "schooling". Investigue en Google cómo hacer esto (p.ej., https://en.wikipedia.org/wiki/Box_plot).
3. (10%) Utilice **diagramas de dispersión** para visualizar las relaciones entre estas variables. En total, son 3 diagramas para visualizar las relaciones. Investigue en Google (p.ej., https://en.wikipedia.org/wiki/Scatter_plot).
4. Para nuestro análisis, dividimos el planeta en 3 zonas: la zona tropical, ubicada entre el Trópico de Capricornio y el Trópico de Cáncer; el Sur, ubicado por debajo del Trópico de Capricornio y el Norte, ubicado por encima del Trópico de Cáncer.
    a. (10%) Prepare los datos y utilice diagramas de cajas para estudiar las distribuciones de los países para **esperanza de vida**, **total expenditure** y **schooling** en cada una de las zonas planetarias. Debe haber una figura por cada variable y cada una debe tener las 3 cajas correspondientes a la zona (ver https://www.reneshbedre.com/blog/anova.html).
    b. (20%) Compare las tres variables estudiadas entre estas regiones utilizando **ANOVA** (https://en.wikipedia.org/wiki/Analysis_of_variance). Siga el enlace de la pregunta anterior para investigar cómo realizar el análisis de varianza en Python.
    c. (10%) Dé una **interpretación** a estos resultados, con base en lo que investigó previamente en los enlaces proporcionados. Explique cuáles son las hipótesis del ANOVA y cómo se interpretan los resultados. ¿Hay diferencias en esperanza de vida, gasto total y escolaridad entre zonas del mundo?
5. Para esta pregunta, trabaje con el subconjunto de datos generado en el lab 3 que contiene información de la altura, correspondiente a entre 1000 y 2000 datos (o más si así lo desea).

a. (5%) Grafique la **esperanza de vida** en función de la **altitud** usando un diagrama de dispersión. Comente lo que observa.
b. (10%) Para analizar si la altitud influye en la esperanza de vida, estudiamos las distribuciones de la esperanza de vida correspondiente a las ciudades, en las categorías de altitud 0-500, 500-1500, 1500-2300 y 2300+. Realice un **diagrama de cajas** de la esperanza de vida en las cuatro categorías de altitud (4 grupos).
c. (15%) Utilice **ANOVA** para comparar las cuatro poblaciones. Explique los resultados obtenidos. ¿Hay diferencia entre los grupos?