Proyecto 3. Inferencia Estadística y Modelación Estadística

1. Objetivo

En este proyecto, se aplicarán los conceptos de estadística inferencial, análisis de varianza (ANOVA) y regresión lineal para investigar un conjunto de datos real. El objetivo es evaluar las relaciones entre variables, identificar diferencias significativas entre grupos y modelar la influencia de factores independientes en las variables dependientes a través del análisis de datos y pruebas estadísticas.

2. Conjunto de Datos

Cada estudiante es responsable de:

1. Seleccionar una muestra aleatoria (n) de 500 estudiantes (de high school de ambos sexos, todos los años) de un estado en el siguiente enlace:

   https://ww2.amstat.org/censusatschool/RandomSampleForm.cfm

   Nota: No escoger alguno de los siguientes estados: Arkansas, Canal Zone, District of Columbia, Guam, Hawaii, Louisiana, Maine, Mississippi, Montana, Nevada, North and South Dakota, Puerto Rico, Virgin Islands, Wyoming.

2. Descargar el cuestionario en el siguiente enlace:

   http://ww2.amstat.org/censusatschool/pdfs/C@SQuestionnaire.pdf

   A partir del cuestionario revisa los siguientes conceptos donde se explica el origen y relevancia de los datos para el análisis:

   a) Unidad de estudio.

   b) Población objeto de estudio.

   c) Muestra de estudio.

   d) Variable de análisis.

   e) Tipo de variable y su escala de medición.

3. Depurar la base de datos, esto puede incluir la eliminación de valores anómalos (error de entrada de datos), la imputación de valores perdidos y la transformación de variables.

3. Pasos del Proyecto

Conformar grupos de 3 estudiantes y unir los datos de los 3 estados. A partir de esta base de datos con 1500 muestras, realizar lo siguiente

3.1. Primer paso: Preparación y Exploración de Datos

- Carga el conjunto de datos en Python o R.
- Inspecciona las primeras filas, dimensiones y tipos de variables.
- Identifica variables clave como dependientes, independientes, continuas, etc.

3.2. Segundo paso: problemas a resolver

1. **Problema:** Seleccionar una variable categórica como factor y una variable continua como variable dependiente. Suponga que están interesados en investigar si hay diferencias significativas en la variable dependiente entre al menos tres grupos que ustedes escojan de la base de datos del "Census at School".

   Pasos:

   a) Análisis de Varianza (ANOVA): Primero, realiza un ANOVA para determinar si hay diferencias significativas en la variable dependiente entre los diferentes grupos. Deben proporcionar el código a mano de cómo llegaron a las tablas de resúmen y ANOVA (ver ejemplo en clase).

   b) Post-ANOVA: Si el resultado del ANOVA es significativo, entonces deben realizar un análisis post-hoc para determinar qué grupos específicos difieren en la variable dependiente. Usar la prueba LSD de Fisher.

   c) Intervalos de Confianza: Calcular los intervalos de confianza para las diferencias en la variable dependiente entre los grupos. Esto les dará una idea de la incertidumbre asociada a las estimaciones de las diferencias.

   d) Gráficas: Finalmente, visualizar los resultados utilizando un boxplot de la variable dependiente entre los diferentes grupos para visualizar las diferencias. También crear un gráfico de los intervalos de confianza para visualizar la incertidumbre de las estimaciones.

2. **Problema:** Leonardo da Vinci (1452-1519) dibujó la figura de un hombre, indicando que la distancia entre los brazos extendidos de una persona (midiendo por la espalda con los brazos extendidos para formar una "T") es casi igual a la estatura de una persona. Para probar lo dicho por él, usen la base de datos de los 1500 estudiantes del *Census at school*.

a) Trace una gráfica de dispersión para la distancia entre los brazos extendidos y estatura. Use la misma escala en los ejes horizontal y vertical. Describa la relación entre las dos variables.

b) Si da Vinci estaba en lo correcto y la distancia entre los brazos extendidos de una persona es casi igual a la estatura de esa persona, ¿cuál debe ser la pendiente de la recta de regresión?

c) Calcule la recta de regresión para predecir la estatura con base en la distancia entre los brazos extendidos de una persona. ¿El valor de la pendiente confirma las conclusiones de ustedes del inciso anterior?

d) Si una persona tiene una distancia de 67 pulgadas entre los brazos extendidos, ¿cuál sería el pronóstico de ustedes respecto a la estatura de la persona?

e) ¿Los datos dan suficiente evidencia para indicar que hay una relación lineal entre distancia y estatura? Pruebe al nivel de significancia de 5 %.

f) Construya un intervalo de confianza de 95 % para la pendiente de la recta de medias.

g) Si Leonardo da Vinci tenía razón y la distancia entre los brazos extendidos de una persona es casi igual a la estatura de esa persona, ¿el intervalo de confianza construido en el inciso anterior confirma esta suposición? Explique.

3. **Problema:** Imaginen que son ingenieros de datos en una empresa de investigación de mercado. Han recopilado información sobre la relación entre la inversión publicitaria (x) y las ventas de un producto (y). Sin embargo, saben que en el mundo real, las relaciones entre variables a menudo están sujetas a variaciones y ruido. Para tener en cuenta esta variabilidad, necesitan generar un conjunto de datos sintéticos más grande con ruido para simular un escenario más realista.

   a) Generación de Datos:

   - Utilicen una función para generar 500 valores de X de manera uniforme (todos los eventos posibles son igualmente probables) entre 50 y 100.
   - Utilicen la relación lineal $y = 1{,}4x + \epsilon$, donde ϵ es un término de error que representa el ruido. Agreguen este término de error para cada valor de X para introducir variabilidad en la relación.

   b) Exploración de Datos:

   - Visualicen los datos generados en un gráfico de dispersión (scatter plot) para comprender la relación entre x y y.

   c) Aplicación de Regresión Lineal:

- Utilicen un el método de los mínimos cuadrados para la regresión lineal simple que ajuste un modelo a estos datos. Grafíque la recta junto con el diagrama de dispersión.
- Calcule los coeficientes de correlación y de determinación para evaluar la calidad del modelo. Interprete los resultados. ¿Cómo describe la fuerza de la relación entre y y x?
- Construya una tabla ANOVA para la regresión lineal.

d) Intervalos de Confianza y Pruebas de Hipótesis:

- Construya intervalos de confianza de 95 % para los coeficientes de la regresión (pendiente e intercepto).
- Interpretar los intervalos de confianza. ¿Qué información proporcionan sobre la estimación de los coeficientes?
- Realice pruebas de hipótesis para verificar si los coeficientes son estadísticamente significativos.

3.3. Tercer paso: Entrega del Proyecto

- Elaborar (y compartir) en, por ejemplo "Colab", el código que les permitió generar los datos, figuras y cálculo de indicadores.
- Elaborar (y compartir) en un documento "PDF" que incluya las respuestas a los problemas anteriores.