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. Problema 1

Una empresa embotelladora de refrescos ha instalado una nueva máquina llenadora de botellas de 500 ml. De acuerdo con las especificaciones, el volumen de llenado sigue una distribución normal con una media objetivo de 500 ml y una desviación estándar de 5 ml. Para asegurar la calidad del producto, se ha implementado un sistema automatizado de control que toma muestras periódicas del proceso. Cada 10 minutos, el sistema selecciona automáticamente 5 botellas de la línea de producción y mide su contenido. Si el promedio de la muestra es mayor que 502 ml o menor que 498 ml, el sistema detiene automáticamente la línea de producción para su ajuste.

1. Plantear una hipótesis nula y alterna adecuada detrás del proceso de control.
2. Enunciar correctamente todos los elementos necesarios para soportar la prueba detrás del proceso de control.
3. ¿Cuál es la probabilidad de parar el proceso si la media de llenado es correcta ($\mu = 500$ ml)?
4. ¿Cuál es la probabilidad de No parar el proceso si la media real de llenado del proceso es 497 ml?
5. Si se desea reducir el nivel de significancia del proceso de control al 5 %, encuentre el tamaño de muestra necesario para que la potencia de la prueba sea 0.98 cuando la media real es 503 ml. ¿Cuáles serían los nuevos límites de control?

3. Problema 2

La Tabla 1 (presentada más adelante) del artículo Burrows-Wheeler Transform Based Lossless Text Compression Using Keys and Huffman Coding de Md. Atiqur Rahman y Mohamed Hamada compara el rendimiento de distintos algoritmos de compresión sin pérdida, tales como PAQ8n, Deflate, Bzip2, Gzip, LZMA, LZW y Brotli, frente al método propuesto por los autores. En esta tabla, se observa que el enfoque de Rahman y Hamada logra, en promedio, un ratio de compresión de 1.884, superando a los demás algoritmos evaluados en términos de eficiencia de compresión. Brotli, el mejor entre los métodos convencionales comparados, obtiene un ratio de 1.667, mientras que LZW presenta el menor ratio de compresión, con 1.288. Esta diferencia de rendimiento sugiere que el método propuesto por Rahman y Hamada es especialmente efectivo para reducir el tamaño de archivos, posicionándose como una alternativa atractiva para aplicaciones que requieren compresión sin pérdida de alta eficacia.

Cuadro 1: Comparación de ratios de compresión entre diferentes métodos.

Textos	**PAQ8n**	**Deflate**	**Bzip2**	**Gzip**	**LZMA**	**LZW**	**Brotli**	**Propuesto**
1	1.582	1.548	1.335	1.455	1.288	1.313	1.608	1.924
2	1.497	1.427	1.226	1.394	1.214	1.283	1.544	1.935
3	1.745	1.655	1.460	1.574	1.338	1.399	1.692	1.925
4	1.523	1.463	1.261	1.382	1.200	1.268	1.531	1.899
5	1.493	1.408	1.228	1.390	1.195	1.170	1.625	1.949
6	1.242	1.228	1.051	1.199	1.057	1.036	1.250	1.429
7	1.154	1.040	1.026	1.061	1.000	0.946	1.287	1.448
8	1.566	1.430	1.316	1.465	1.298	1.254	1.783	1.893
9	1.295	1.265	1.092	1.219	1.050	1.275	1.380	1.536
10	1.495	1.371	1.307	1.419	1.216	1.174	1.511	1.629
11	1.455	1.309	1.219	1.373	1.168	1.134	1.466	1.632
12	1.497	1.306	1.249	1.370	1.222	1.209	1.580	1.773
13	1.369	1.201	1.126	1.250	1.097	1.092	1.493	1.660
14	1.595	1.407	1.336	1.462	1.321	1.305	1.637	1.773
15	1.559	1.302	1.243	1.380	1.249	1.227	1.492	1.788
16	2.401	2.082	2.214	2.121	1.888	1.559	2.269	2.466
17	1.380	1.211	1.353	1.302	1.113	1.103	1.428	1.903
18	1.755	1.537	1.477	1.585	1.401	1.394	1.782	1.931
19	1.507	1.370	1.261	1.417	1.247	1.234	1.542	1.815
20	2.020	1.744	2.010	1.783	1.596	1.430	1.941	2.033
Promedio	**1.643**	**1.486**	**1.418**	**1.504**	**1.325**	**1.288**	**1.667**	**1.884**

Aplicar un análisis de varianza (ANOVA) de un solo factor en Python, calculando cada elemento de la tabla ANOVA mediante fórmulas manuales. El objetivo es verificar si existen diferencias estadísticamente significativas en el rendimiento de los algoritmos y, en caso de encontrar diferencias, realizar una prueba post-ANOVA utilizando el LSD de Fisher para determinar cuáles grupos difieren entre sí.

Pasos:

1. 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).

2. 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.

3. 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.

4. 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.

4. Problema 3

El famoso polímata italiano Leonardo da Vinci (1452-1519) propuso en sus estudios anatómicos que existe una proporción particular en el cuerpo humano: la distancia entre los brazos extendidos horizontalmente (formando una "T" con el cuerpo) es aproximadamente igual a la estatura de la persona. Para verificar esta hipótesis, utilizaremos datos de 1500 estudiantes del *Census at school* que usaron en el primer proyecto.

a) Realice un diagrama de dispersión (scatter plot) para visualizar la relación entre la distancia entre brazos extendidos (eje x) y la estatura (eje y). Utilice la misma escala en ambos ejes. ¿Qué tipo de relación observa entre estas variables?.

b) Según la hipótesis de da Vinci, ¿qué valor debería tener la pendiente de la recta de regresión? Justifique su respuesta.

c) Determine la ecuación de la recta de regresión para predecir la estatura a partir de la distancia entre brazos. Compare la pendiente obtenida con su respuesta del inciso anterior. ¿Qué puede concluir?

d) Para una persona con una distancia entre brazos de 65 pulgadas: ¿Cuál sería su estatura predicha? ¿Cuál es el error estándar de esta predicción? Construya un intervalo de predicción del 95 % para esta estatura.

e) Realice un análisis de residuos: Grafique los residuos vs valores ajustados ¿Se cumplen los supuestos de linealidad y homocedasticidad? Identifique posibles valores atípicos o influyentes.

f) Pruebe la existencia de una relación lineal entre las variables: Plantee las hipótesis nula y alternativa. Use $\alpha = 0{,}05$. Interprete el resultado en el contexto del problema.

g) Para la pendiente de la recta de regresión: Construya un intervalo de confianza del 95 % ¿Este intervalo contiene el valor teórico según la hipótesis de da Vinci? ¿Qué implica esto sobre la validez de la hipótesis?

h) Calcule e interprete: El coeficiente de correlación r. El coeficiente de determinación R^2. ¿Qué nos dicen estos valores sobre la fortaleza de la relación?

5. Entregas

- 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.