Facultad de Ingenierías
Inferencia Estadística
Maestría en Inteligencia Artificial y Ciencia de Datos
Prof. Cristian E Garcia.

Desafío 2

Condiciones:

- Subir la tarea en formato pdf en la plataforma UAO-Virtual.
- Es necesario incluir el código de R en formato R. Mostrar los resultados a partir de tablas, gráficos o indicadores que les permita dar respuesta a los planteamientos.
- Deben interpretar los resultados obtenidos en cada situación de acuerdo al contexto.
- Realizar la actividad en grupos máximo de 4 personas.

Situación 1

Un experimento utilizó una muestra de estudiantes universitarios para investigar si el uso del teléfono celular afecta los tiempos de reacción de los conductores. En una máquina que simulaba situaciones de conducción, se encendía de manera irregular una luz roja o verde. Se les indicó a los participantes que presionaran un botón de freno tan pronto como detectaran una luz roja. Bajo la condición de uso del teléfono celular, cada estudiante tenía una conversación con alguien en otra habitación. En la condición de control, los mismos estudiantes escuchaban una transmisión de radio. El archivo de datos **CellPhone** registra los tiempos de respuesta promedio de los estudiantes (en milisegundos) en varias pruebas para cada condición: $\{y_{i1}\}$ para la condición del teléfono celular y $\{y_{i2}\}$ para la condición de control.

[(a)] Las comparaciones de medias o proporciones suponen muestras independientes para los dos grupos. Explica por qué las muestras para estas dos condiciones son **dependientes** en lugar de independientes.

[(b)] Para comparar μ_1 y μ_2, puedes usar $\{d_i = y_{i1} - y_{i2},\ i = 1,\ldots,n\}$, aquí con $n = 8$. Especifica el parámetro μ_d y la hipótesis nula H_0 para hacer esto, y explica por qué$\mu_d = \mu_1 - \mu_2$.

[(c)] Indica las suposiciones y la estadística de prueba, y explica por qué sigue una distribución t con$df = n - 1$. Reporta el valor P con una hipótesis alternativa bilateral H_a, e interpreta el resultado. También es posible realizar análisis de pares relacionados

usando intervalos de confianza, al comparar pesos de niñas con anorexia antes y después de un tratamiento analizando la diferencia media de pesos).

Situación 2

Una empresa farmacéutica produce píldoras donde el ingrediente activo debe ser exactamente de **10 mg**. Una desviación, incluso pequeña, puede tener consecuencias. El equipo de control de calidad toma una muestra de $n = 40$ píldoras para decidir si el lote de producción debe ser rechazado. La hipótesis nula es que la media del lote es de 10 mg.

El desafío es que no se sabe con certeza la distribución exacta del contenido del ingrediente. Puede ser perfectamente normal o puede estar "contaminada" con valores atípicos debido a fallos esporádicos en la maquinaria.

1. Planteamiento de la Prueba de Hipótesis:
   - **Hipótesis Nula (H0):** La media del contenido del ingrediente activo en el lote es de 10 mg ($\mu = 10$).
   - Hipótesis Alternativa (H1): La media no es de 10 mg ($\mu \neq 10$).
   - Nivel de Significancia: $\alpha = 0.05$.
2. **Escenarios de Simulación:** Los estudiantes deben simular datos bajo dos escenarios distintos:
   - **Escenario 1 (Normal):** Los datos provienen de una distribución `Normal(μ, σ²)`, con `σ = 0.5`.
   - **Escenario 2 (Contaminado):** Los datos provienen de una **distribución mixta** para simular outliers. El 95% de los datos viene de `Normal(μ, σ²)` y el 5% viene de una distribución con mayor varianza, `Normal(μ, (3σ)²)`.

3. **Cálculo de la Función de Potencia por Simulación:** Para cada uno de los dos escenarios, los estudiantes deben escribir un programa que calcule la potencia de las siguientes tres pruebas sobre un rango de valores verdaderos de μ (por ejemplo, de 9.5 a 10.5 en pasos de 0.05):
   - Z-test para una muestra: Asumiendo que `σ = 0.5` es conocida.
   - t-test para una muestra: Sin asumir que σ es conocida.
   - **Prueba de rangos con signo de Wilcoxon:** Una prueba no paramétrica que no asume normalidad.
4. El algoritmo para calcular la potencia para **un valor específico de μ** es: a. Inicializar un contador de rechazos a cero. b. Realizar un bucle de `B = 5,000` iteraciones. c. En cada iteración, generar una muestra de `n = 40` datos del escenario de distribución correspondiente (Normal o Contaminado) con esa media μ. d. Aplicar la prueba de hipótesis (Z, t, o Wilcoxon) a la muestra generada. e. Si el p-valor es menor que α, incrementar el contador de rechazos. f. La **potencia estimada** para ese μ es `(contador de rechazos) / B`.

5. **Análisis y Visualización:**

- Generar dos gráficos, uno para cada escenario (Normal y Contaminado).
- En cada gráfico, dibujar las tres curvas de la función de potencia (una para cada prueba), con μ en el eje X y la Potencia en el eje Y.
- **Verificar que para $\mu = 10$ (bajo H0), la potencia es igual a $\alpha = 0.05$ para todas las pruebas.** Esto confirma que el nivel de significancia está bien controlado.

Discusión:

- **Escenario Normal:** ¿Qué prueba es la más potente? ¿Era esto lo esperado? ¿Por qué?
- **Escenario Contaminado:** ¿Cómo cambia la potencia de cada prueba en presencia de outliers? ¿Sigue siendo la misma prueba la más potente?
- **Robustez:** Basado en los gráficos, ¿qué prueba considerarían más "robusta"? Es decir, ¿cuál mantiene un buen rendimiento en ambos escenarios?
- ¿Qué sucede con la potencia a medida que el valor verdadero de μ se aleja de 10? ¿Tiene sentido este comportamiento?

Situación 3

Una firma decide estudiar una muestra aleatoria de 20 proyectos que envió para ser evaluados, tanto a consultores externos, como a su propio departamento de proyectos. Las variables medidas fueron

X : n^ode días que demoro la evaluación

Y : n^o de variables consideradas en la evaluación.

Z: Consultor al que se le envió el proyecto

$$Z = \begin{cases} -1 & ; \text{Depto. de Evaluación} \\ 0 & ; \text{Robani Consultores} \\ 1 & ; \text{Tanaka Ltda.} \end{cases}$$

W: Costo de la evaluación (en U.F.)

Los resultados de este muestreo son:

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	4	2	8	10	1	3	8	3	2	2	4	4	5	6	7	2	1	3	4	9
Y	3	1	6	8	3	2	6	2	1	1	4	4	4	7	10	3	2	4	5	10
Z	-1	-1	0	0	0	0	1	0	0	1	-1	-1	0	1	1	-1	-1	0	1	-1
w	40	30.5	80.3	68.5	24.7	40.5	90.5	38.5	50.4	50.2	60.1	60.8	70.9	80	90	30	27	40	50	40

(a) Estime con un 90 % de confianza el costo medio de los proyectos.

(b) Estime con un 90 % de confianza la proporción de proyectos cuyo costo fue inferior a 50 U.F. dado que no involucraron más de 6 variables y que fueron resueltos en un tiempo superior a 2 días.

(c) El Depto. de control afirma que el costo medio de enviar los proyectos a asesores externos es significativamente mayor que el de evaluarlos allí mismo. ¿Qué concluye usando$\alpha = 0{,}05$?. **Nota**: Compruebe primero si las varianzas son iguales o diferentes para poder decidir que test utilizar para la diferencia de medias. **Hint:** Use la distribución F.

(d) Tanaka Ltda. Afirma que la proporción de proyectos que ellos evalúan, que toman más tiempo de más de 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo de más 4 días, no es superior a la proporción de proyectos que evalúa Robani Consultores, que toman un tiempo más de 4 días. Concluya si la afirmación de Tanaka Ltda. es correcta. (Use$\alpha = 0{,}01$).

Situación 4 “Investigar un poco”

La siguiente tabla contiene 40 recuentos anuales del número de reclutas y reproductores en una población de salmones. Las unidades están en miles de peces.

R	S	R	S	R	S	R	S
68	56	222	351	311	412	244	265
77	62	205	282	166	176	222	301
299	445	233	310	248	313	195	234
220	279	228	266	161	162	203	229
142	138	188	256	226	368	210	270
287	428	132	144	67	54	275	478
276	319	285	447	201	214	286	419
115	102	188	186	267	429	275	490
64	51	224	389	121	115	304	430
206	289	121	113	301	407	214	235

- **Reclutas (R)**: peces que ingresan a la población capturable.

- **Reproductores (S)**: peces que están poniendo huevos. Los reproductores mueren después de poner huevos.

El modelo clásico de Beverton-Holt para la relación entre reproductores y reclutas es:

$$R = \frac{1}{\beta_1 + \beta_2 / S}, \quad \beta_1 \geq 0, \quad \beta_2 \geq 0,$$

donde R y S son los números de reclutas y reproductores, respectivamente. Este modelo puede ajustarse mediante regresión lineal con las variables transformadas $1/R$ y $1/S$.

Para mantener una pesca sostenible, la población total solo se estabilizará si $R = S$. La población total disminuirá si se producen menos reclutas de los reproductores que murieron generándolos. Si se producen demasiados reclutas, la población también disminuirá debido a la competencia por los recursos. Por lo tanto, hay un nivel intermedio de reclutas que se puede mantener indefinidamente en una población estable. Este nivel estable es el punto donde la línea de 45° intercepta la curva que relaciona R y S.

Instrucciones

(a) Ajustar el modelo de Beverton-Holt y encontrar una estimación puntual para el nivel estable de la población donde $R = S$. Usar bootstrap para obtener un intervalo de confianza del 95% y un error estándar, utilizando dos métodos: remuestreo de los residuales y remuestreo de los casos. Representar histogramas para cada distribución bootstrap y comentar sobre las diferencias en los resultados.
(b) Proporcionar una estimación corregida por sesgo y un error estándar correspondiente para el estimador corregido.
(c) Usar bootstrap anidado con pivoteo para encontrar un intervalo de confianza del 95% para el punto de estabilización.

Situación 5 "Investigar un poco más"

Un equipo de MLOps ha desarrollado un nuevo algoritmo (B) para procesar transacciones en tiempo real y necesita determinar si es significativamente más rápido que el algoritmo actual (A). La métrica clave de rendimiento es el **tiempo de procesamiento**, pero se sospecha que esta métrica no se distribuye de forma normal; de hecho, puede tener una cola larga debido a picos de carga ocasionales. El objetivo es comparar el "peor caso" del rendimiento, que se definirá como el **percentil 95** del tiempo de procesamiento.

Generación de Datos Simulados:

Simule los tiempos de procesamiento para ambos algoritmos, creando un escenario realista.

1. Dimensiones de las Muestras:
   - Algoritmo A (control): $n_A = 1500$ observaciones.
   - Algoritmo B (tratamiento): $n_B = 1400$ observaciones.
2. Modelos de Generación (Distribuciones No Normales):
   - **Tiempos del Algoritmo A:** Generar los datos de una **distribución Gamma**, que a menudo modela bien los tiempos de espera. Por ejemplo, `Gamma(forma=3, escala=10)`.
   - **Tiempos del Algoritmo B:** Generar los datos de una **distribución Lognormal** para simular una mejora. Esta distribución también es asimétrica y de cola derecha. Por ejemplo, `Lognormal(media_log=2.5, sd_log=0.5)`. El objetivo es que la media de B sea menor que la de A, pero la comparación debe hacerse sobre el percentil 95.

Desafío a evaluar

1. Validación de Supuestos (Prueba de Normalidad):
   - Para cada conjunto de datos (A y B), aplicar una prueba de normalidad formal, como la de **Shapiro-Wilk**, vista en la Sesión 5.
   - Interpretar el p-valor para concluir (y demostrar) que los datos no son normales, justificando así la necesidad de usar métodos no paramétricos como el bootstrap.
2. Estimación por Intervalos de Confianza mediante Bootstrap:
   - El estadístico de interés es la **diferencia entre los percentiles 95** de ambos algoritmos: $\Delta = P95(A) - P95(B)$.
   - Implementar el procedimiento de **bootstrap no paramétrico** (visto en la Sesión 4) para estimar la distribución muestral de Δ.
     - Generar `B = 10,000` réplicas bootstrap.
     - En cada réplica, remuestrear con reemplazo de las muestras originales A y B para crear `muestra_A*` y `muestra_B*`.

        - Calcular el estadístico de interés para la réplica: $\Delta^* =$ `P95(muestra_A*) - P95(muestra_B*)`.
    - Construir un intervalo de confianza del 95% para Δ utilizando el método de percentiles sobre las `B` réplicas de Δ^*.
3. Prueba de Hipótesis Basada en Bootstrap:
    - Formular una prueba de hipótesis para determinar si el Algoritmo B es significativamente más rápido en el percentil 95.
        - **Hipótesis Nula (`H0`):** No hay diferencia en el percentil 95 del tiempo de procesamiento. `P95(A) = P95(B)` o $\Delta = 0$.
        - **Hipótesis Alternativa (`H1`):** El percentil 95 del tiempo de procesamiento del Algoritmo B es menor que el de A. `P95(B) < P95(A)` o $\Delta > 0$.
    - Calcular el **p-valor** a partir de la distribución bootstrap de Δ^*. El p-valor será la proporción de réplicas Δ^* que son menores o iguales a cero.
    - Tomar una decisión: Con un nivel de significancia de $\alpha = 0.05$, ¿se rechaza la hipótesis nula?