Trabajo Práctico

Estadística

Docentes: Rodrigo Del Rosso - Ezequiel Nuske - Gustavo Levinis

Deadline: Lunes 15 de Julio de 2024

Consideraciones

El presente trabajo práctico constituye el único de la materia **Estadística** de la Maestría en Ciencia de Datos de la **Universidad Austral** de la modalidad Online.

La finalidad es articular los conceptos teóricos estudiados en las clases con una aplicación práctica mediante la utilización de R.

Ahora bien, podrán utilizar cualquier paquete o diseñar cualquier función adicional que consideren necesaria, siempre indicando el uso de los mismos. Pueden emplear cualquier lenguaje de programación, no se limiten únicamente a R.

- Algunas cuestiones prácticas:
    1. La fecha de entrega es inclusive. Tienen hasta las 23:59 de ese día para entregar el trabajo.
    2. Cualquier entrega tardía será penalizada, descontando un 20% de la nota obtenida.
    3. El trabajo es grupal. La cantidad máxima de integrantes es de 3 (tres). No se admiten modificaciones de integrantes.
    4. El siguiente **formulario** deberá ser completado con los datos requeridos de cada grupo.
    5. Cada integrante del grupo deberá subir como máximo 3 (tres) archivos: un PDF con el informe, un archivo del script (.R, .py, etc) y de corresponder, la base de datos utilizada.
    6. Cabe destacar que el lenguaje empleado en el informe deberá ser de índole académico. Por ejemplo, una buena escritura académica aconseja no emplear gerundios (Ejemplo: "Planteando", "Analizando", etc.), y por convención, los textos académicos escapan el uso de la primera persona (especialmente del singular) por considerar que tiñe de informalidad, de subjetividad o de falta de rigor la comunicación científica. Asimismo, es útil emplear paráfrasis y referencias a distintos autores. Es importante mencionar que se deberá utilizar las normas de estilo APA [1]. Se recomienda consultar el libro de *Manuel Scarano* ante cualquier inquietud.

Los siguientes ejercicios deberán resolverlos en forma grupal mediante la utilización de R. El puntaje mínimo para aprobar es de **60 puntos** completos y correctos. Cada ejercicio cuenta con el puntaje que otorga su resolución en forma completa y correcta.

Ejercicio N° 1 - Estadística Descriptiva (20 puntos)

Seleccione una base de datos pública de su interés y seleccione una muestra de n observaciones aleatorias al azar. Alternativamente, simule una muestra aleatoria simple de al menos 1000 registros.

Para la entrega de este examen deberá adjuntar la muestra seleccionada y el procedimiento (ya sea en R o Excel) que permitir obtener los registros muestrales. A partir de la misma se solicita lo siguiente,

[1] Acrónimo en inglés de *American Psychological Asociation*

- Describir para este caso la población, la muestra tomada, el experimento que estará usando, las variables bajo análisis y cualquier otra característica relevante para el procedimiento. Si la base es simulada, exponga los motivos por los cuales el procedimiento es viable.
- Generar un set de estadística descriptiva sobre la misma que le permita resumir la información obtenida, explicando para cada una de ellas su significado.
- Generar un histograma para una de las variables cuantitativas, utilizando la forma que considere más correcta para agrupar las categorías. Elija la variable que mayor simetría consiga en el histograma resultante.

Ejercicio N° 2 - Probabilidad (20 puntos)

Ejercicio 1

1. Si lanzamos una moneda, ¿Cuál es la probabilidad esperada de obtener una cara?. Si lanzamos una moneda 10 veces, ¿Cuál es la cantidad esperada de caras?
2. Lanzar una moneda 10 veces y contar el número de caras. Repetirlo 8 veces y almacenar el número de caras para cada una.
3. Lanzar una moneda 10 veces, contar el número de caras, almacenar el resultado y repetirlo 1000 veces.
4. ¿Cómo difieren los resultados del experimento en (2) de los resultados en el experimento (3)? Justificar

Ejercicio 2

Una persona te propone jugar un juego con dados, el cual te solicita tirar 2 dados.

- Si sale un 7, te pagarán $ 3
- Si sale un 11, te pagarán $ 5.
- Si sale cualquier otra combinación, deberás pagar $ 0.70

1. ¿Cuál es la probabilidad de sacar un siete?
2. ¿Cuál es la probabilidad de sacar un once?
3. ¿Cuál es la probabilidad de sacar un siete o un once?
4. Simular tirar 2 dados mediante la función Roll1Dice(). Simular tirar 2 dados 100 veces y almacenar los resultados. Calcular los puntos anteriores (1, 2 y 3) a partir de los datos.
5. Suponga que jugó 10 veces y obtuvo una ganancia de $ 30. ¡Qué fácil parece ser el juego! ¡Debería seguir jugando! ¿Es correcta la suposición? Demostrar con una simulación.
6. Ahora dicha persona te ofrece disminuir el monto a pagar a $ 0.68. ¿Deberías aceptarlo?

Ejercicio N° 3 - Variables Aleatorias (20 puntos)

1. Simular la suma de dos variables normales mediante la función rnorm(x,mu,sigma) en R con media igual a 0 y desvío igual a 1. Probar con el tamaño de muestra n = 10 y n = 100. ¿Qué observa si grafica ambos objetos?
2. Simular la suma de dos variables normales mediante la función rnorm(x,mu,sigma) en R con media igual a 0 y desvío igual a 1. Probar con distintos valores de tamaño de muestra. Podría probar con n = 100, n = 1000, n = 10000, n = 100000. Gráficar para estos distintos valores de muestra. Comprobar que la media (valor esperado) es igual a la suma de sus valores esperados y la varianza es igual a la suma de varianzas individuales.

3. Simular la suma de diez variables normales mediante la función rnorm en R con media igual a 0 y desvío igual a 1. Probar con distintos valores de tamaño de muestra. Podría probar con n = 100, n = 1000, n = 10000, n = 100000. Gráficar para estos distintos valores de muestra. Comprobar que la media (valor esperado) es igual a la suma de sus valores esperados y la varianza es igual a la suma de varianzas individuales.

4. Extra: Probar lo anterior pero sumando normales con distintas medias y desvíos.

Hint: Crear vectores vacíos y utilicé una estructura repetitiva para rellenar con el resultado de cada simulación.

Ejercicio N° 4 - Teorema Central del Límite (20 puntos)

En clase hemos visto que la media de variables Normales es una Normal. Ahora bien, ¿Ocurrirá lo mismo si las variables que se promedian no son normales?.

Se plantea el siguiente ejercicio para que intenten resolver, de forma tal que descubran al Teorema Central del Límite.

Repetir el proceso visto en clase mediante R cuando la variable aleatoria original se distribuye de la forma siguiente,

1. Poisson de parámetro $\lambda = 1.3$
2. Exponencial de parámetro $\mu = 1.5$
3. Uniforme en el intervalo [5,10]
4. Weibull de parámetros shape = 1.2 y scale = 0.5

Realizar un Gráfico de Histograma y plotear la densidad de una variable normal para cada caso.

Ejercicio N° 5 - IC y Prueba de Hipótesis (20 puntos)

Proceda a cargar el siguiente dataset de un repositorio en *github*,

```
install.packages("repmis")
url = "https://github.com/hllinas/DatosPublicos/blob/main/Estudiantes.Rdata?raw=false"
repmis::source_data(url)
datos <- Estudiantes
```

Realizar los siguientes ejercicios. Interprete todas sus respuestas.

a) Considerar solamente las observaciones que van desde la 2 hasta la 35 y definir el data frame “datos2a35”. Verificar su tamaño, variables y estructura.

Todas los puntos siguientes resolverlo con el dataset “datos2a35”,

b) Definir el objeto “Sexo” (género de los estudiantes). Conviértalo en factor y diga cuáles son sus respectivos niveles.

c) Construir una tabla de frecuencias para la variable Sexo y el diagrama de barras correspondiente.

d) Determinar la proporción de mujeres.

e) Mediante el método de la región crítica: Al nivel del 5%, determine si el porcentaje poblacional de mujeres es menor o igual que el 30%. Escribir un resumen del enunciado del problema, verificar los supuestos, concluya, diga cuál es la fórmula, el valor de prueba, el valor crítico, la región crítica e interprete.

f) Mediante el método del P-valor: Determine si el porcentaje poblacional de mujeres es menor o igual que el 30%. Halle el P-valor, interprete y compare su decisión con el inciso (e).

g) Realizar la misma prueba del inciso (h) con la función prop.test y compare los resultados obtenidos.

h) Construir un intervalo del 95% de confianza para la proporción poblacional de mujeres y compare los resultados obtenidos en los incisos anteriores.

i) Construya intervalos de Confianza del 95% para la media mediante bootstrap con 10000 repeticiones.

Ejercicio N° 6 - Naive Bayes (Bonus)

Para responder los siguientes puntos, se le requiere cargar el dataset "infert" de la librería datasets y utilizar una semilla igual a 123 para que los resultados sean reproducibles.

```
data(infert, package ="datasets")
```

a) Convertir las variables predictoras que sean categóricas en factores.
b) Entrenar el algoritmo de Naive Bayes con el 70% de los datos y el resto utilizarlos para testear.
c) Calcular la matriz de confusión sobre las predicciones realizadas sobre los datos de testing.
d) Calcular las métricas de accuracy, especificidad y sensibilidad.