Estadística Computacional

Proyecto grupal Curso 2023/24

Primera entrega. Presentación del estudio

Instrucciones

En esta primera entrega, cada grupo deberá decidir el contenido y los sujetos y variables de estudio sobre las que se realizará el análisis estadístico en el resto de entregas. Para ello, cada grupo podrá seleccionar una base de datos entre las distintas opciones que aparecen en Blackboard, o bien proponer una alternativa distinta (que requerirá de validación).
Será necesario:

- Seleccionar 50 ó más sujetos en el estudio ($n > 50$).

- Seleccionar al menos 4 variables de estudio, de las cuales al menos una deberá ser categórica/cualitativa, y las otras tres tendrán que ser numéricas, dos de ellas continuas y la tercera puede ser continua o discreta.

Procedimiento y requisitos mínimos

Se deberá enviar en la fecha establecida de entrega en formato PDF, con una longitud de entre 1 y 2 páginas:

- Las características de la población de estudio, las variables escogidas (al menos 4, pueden ser más) y su descripción (unidades y tipo).

- Hipótesis y marco de trabajo: deberá argumentarse la elección de variables y datos seleccionados y las hipótesis previas de trabajo: presumibles correlaciones, relaciones y diferencias. Motivo y objetivo del estudio.

La longitud máxima de este informe será de dos páginas con tipo de letra Arial, Verdana, Times New Roman o similar y un tamaño de fuente de 11 puntos.

Fecha límite de entrega: 25/02.

Segunda entrega. Estadística Descriptiva

Instrucciones

En esta segunda entrega, se procederá a realizar un análisis descriptivo básico de los datos, empleando los estadísticos más comunes, de dispersión y de centralidad (media, mediana, desviación típica, asimetría, kurtosis). El análisis se realizará de modo global y de modo específico según los niveles o variable o variables categóricas escogidas. Además, en esta etapa del estudio se deberá representar visualmente los datos.

Se deberá entregar un informe en PDF que recoja los cálculos y análisis estadísticos descriptivos solicitados en la fecha establecida de entrega. La longitud del informe debe ser de entre 5 y 7 páginas y deberá contener los requisitos mínimos explicados a continuación.

Procedimiento y requisitos mínimos

- Estadísticos descriptivos básicos totales: media, mediana, varianza, desviación típica, curtosis, coeficiente de asimetría y cuartiles; y por niveles según una de las variables categóricas escogidas.

- Gráficas que representen e ilustren aspectos significativos del conjunto de datos. Pueden emplearse: histogramas, box-plots, diagramas de barras o de tarta, gráficos Q- Q y/o gráficos de probabilidad para comparar muestras, diagramas de dispersión,... del conjunto total de los datos y de algunos subgrupos o niveles concretos de la población.

- Todos los gráficos deben estar comentados, explicando qué es lo que cuentan sobre los datos (abstenerse de poner teoría generalista sobre los gráficos).

La **discusión y la visualización de los datos** obtenidos es donde recaerá la mayor parte del peso de la evaluación en esta parte.

El documento no debe contener código Matlab ni “pantallazos” tomados directamente de Matlab.

Fecha límite de entrega: 01/04.

Después de la entrega, los estudiantes recibirán feedback por parte del profesor. Deberán corregir o incorporar las sugerencias a su trabajo final.

Tercera Entrega. Inferencia

Instrucciones

La tercera entrega se basa en realizar diversos análisis inferenciales a partir los datos obtenidos en la primera entrega, complementando el análisis exploratorio realizado en la entrega número dos. En concreto:

- **Distribuciones y contrastes de bondad de ajuste**: a la vista de los histogramas calculados en la segunda entrega, se deberá proponer una distribución a la que se ajusten las variables dadas (normal, gamma, exponencial...), estimando los parámetros necesarios (media y Varianza en caso de distribución normal, lambda en caso de exponencial, etc). Además, se debe realizar un contraste sobre la bondad del ajuste mediante la distribución chi-cuadrado, según la teoría vista en Inferencia Estadística. En caso de que el ajuste paramétrico no sea bueno, se deberá recurrir a las técnicas no paramétricas vistas en Estadística computacional.

- **Estimación paramétrica**: se deben calcular el sesgo y la eficiencia para algunos de los estimadores calculados.

- Proponer **modelos de regresión lineal (o no lineal)** para la variable respuesta Y. Se deberá utilizar el método de validación cruzada para calcular el error de predicción en los modelos de regresión lineal propuestos.

- **Intervalos de confianza**: se incluirán intervalos de confianza de los siguientes parámetros muestrales: media y varianza, y/o proporción, en función de las variables escogidas. Los intervalos se calcularán de forma teórica, asumiendo normalidad de los datos, y se compararán con el intervalo Bootstrap estándar, y/o con el intervalo percentil.

- **Contraste de hipótesis**: además del contraste de bondad de ajuste, deberá realizarse algún otro contraste de hipótesis, en función de las hipótesis previas y de las variables escogidas. En ellos podrá por ejemplo analizarse si los parámetros muestrales (media, varianza) se desvían de valores establecidos o conocidos (ej. si la tasa de desempleo en Madrid se desvía significativamente de un valor dado, como el nacional) o analizar diferencias entre los distintos subgrupos o niveles (factores) de la variable cuantitativa/ categórica (p.ej. ¿existen diferencias significativas en la tasa de desempleo entre distritos?).

La longitud de esta parte del informe no debe de sobrepasar las 10 páginas, incluyendo los gráficos necesarios sobre la distribución de las variables, en caso de necesitarlos, los resultados obtenidos y la discusión de los resultados.

Como anexo al final (sin incluir en las 8 páginas) se debe adjuntar el código empleado para la inferencia.

Se valorará en especial que los resultados obtenidos sean correctos y tengan sentido, y sobre todo que las hipótesis planteadas sean coherentes.

Procedimiento y requisitos mínimos

Se deberá incorporar al informe PDF elaborado en la entrega 2 los puntos descritos previamente, completando el análisis descriptivo ya realizado. Deberá contener como mínimo:

- Proponer modelos probabilísticos para al menos 2 de las variables escogidas (si siguen una distribución normal, gamma, binomial, etc). Para dichas variables se deberá realizar un test sobre la bondad del ajuste usando las técnicas vistas en inferencia estadística.
- Para una de las variables numéricas, estimar la función de densidad de probabilidad mediante alguno de los métodos no paramétricos vistos en computacional.
- Proponer al menos un modelo de regresión lineal y emplear Validación cruzada para estimar el error.
- Para el modelo de regresión propuesto, calcular sesgo y error estándar del coeficiente de correlación entre la variable independiente y la (o una de las) dependiente(s) usando la técnica Jackkife.
- Intervalos de confianza para al menos 2 de las variables estudiadas. Al menos dos de los intervalos se calcularán de forma teórica, asumiendo normalidad de los datos, y se compararán con el intervalo Bootstrap estándar, y con el intervalo percentil.
- Contraste de hipótesis: se deberá realizar un constate de hipótesis sobre la media de una de las variables de forma teórica, y mediante simulación MC. Además, se debe hacer una comparación y análisis de diferencias significativas entre medias o proporciones de una de las variables escogidas según los distintos niveles/valores de la variable categórica. Se puede usar la función ttest2 de Matlab, que sirve para ver si dos medias son significativamente distintas o no, y además usar la técnica vista en clase en Inferencia y comparar.

En la discusión y análisis de los datos obtenidos es dónde recaerá la mayor parte del peso de la evaluación. Se valorarán aquellos análisis extra realizados al margen de los requisitos mínimos, siempre sin superar la longitud máxima. Se valorará especialmente la elección más adecuada de las variables, los gráficos y las hipótesis estudiadas para justificar el objetivo del estudio inicialmente propuesto.

El informe total constará de cada una de las 3 partes explicadas, habiendo sido la parte descriptiva corregida o mejorada con el feedback del profeso. La longitud total del informe final no debe superar las 20 páginas incluyendo introducción, resultados, gráficos y discusión de los resultados y conclusiones. El código empleado se anexará a parte.

Además del informe a entregar, se hará una presentación grupal la última semana de clase, durante el horario de Inferencia o de computacional, donde los estudiantes tendrán que contestar a las preguntas de ambas profesoras. El profesor podrá hacer cualquier pregunta a cualquiera de los integrantes del grupo, independientemente de la parte en la que haya trabajado cada alumno.

La nota del trabajo constará de 2 partes: una grupal por el documento y la presentación, y otra individual según la calidad de las respuestas del estudiante a las preguntas de las profesoras.