A2 - Análitica descriptiva e inferencial

Enunciado

Semestre 2021.2

Índice

Introducción

En esta actividad nos introducimos en la inferencia estadística. Para ello, usaremos el conjunto de datos `CensusIncome_clean.csv` que se ha preprocesado en la actividad anterior. Este conjunto de datos surge de un censo, donde se registra la información demográfica, personal y laboral de una muestra de la población. El conjunto de datos original se ha obtenido de la base de datos de la web Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

Las variables del conjunto de datos son:

- *CS_ID:* Identificador del individuo.
- *age:* Edad del individuo.
- *workclass*: Categorización del individuo en base al perfil laboral.
- *fnlwgt*: Número de unidades de la población objetivo que representa la unidad respondiente.
- *education_num*: Número de años de formación educativa del individuo.
- *marital_status*: Estado civil del individuo.
- *occupation*: Categorización del individuo en base a la tipología del trabajo.
- *relationship*: Parentesco de la unidad que responde de la familia.
- *race*: Grupo racial al que pertenece el individuo.
- *gender*: Género del individuo.
- *hours_per_week*: Horas por semana trabajadas por el individuo.
- *income*: Salario (anual) del individuo.
- *education_cat*: Nivel de educación.

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

---

1. Lectura del fichero y preparación de los datos

Leed el fichero `CensusIncome_clean.csv` y guardad los datos en un objeto con identificador denominado *censo*. A continuación, verificad que los datos se han cargado correctamente.

---

2. Edad

Para empezar el análisis, nos interesa conocer el valor medio de la edad del censo, a partir de los datos de la muestra. Para ello, calculad el intervalo de confianza de la media edad. Seguid los pasos que se especifican a continuación.

2.1. Distribución de edades

Visualizad gráficamente la distribución de la edad. Escoged el gráfico que sea más apropiado, considerando que se quiere conocer la distribución de la variable y si ésta sigue una distribución normal.

2.2. Normalidad

¿Podemos asumir normalidad para el cálculo del intervalo de confianza de la media de edad? Argumentad la respuesta.

2.3. Intervalo de confianza

Calculad manualmente el intervalo de confianza de la media de la variable age. Para ello, definid una función IC que reciba la variable, la confianza, y que devuelva un vector con los valores del intervalo de confianza.

La cabecera de la función es:

```
IC <- function( x, NC ){}
```

Nota: No se pueden usar funciones como t.test para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

2.4. Cálculos

Calculad el intervalo de confianza al 90 % y 95 %. Comparad los resultados.

2.5. Interpretación

Explicad cómo se interpreta el intervalo de confianza a partir de los resultados obtenidos.

---

3. Salario

Vamos a investigar ahora el salario de la población. En particular, nos preguntamos si en media, el salario de las personas Self-Employed es inferior al del resto de modalidades. Seguid los pasos que se especifican a continuación.

3.1. Pregunta de investigación

Formulad la pregunta de investigación.

3.2. Hipótesis

Escribid las hipótesis (hipótesis nula e hipótesis alternativa).

3.3. Test a aplicar

Explicad qué tipo de test podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

Nota: Podéis usar las funciones de R que consideréis necesarias para responder esta pregunta.

3.4. Cálculo

Calculad el test usando una función propia. Implementad una función que realice el cálculo del test y que podáis usar con distintos valores de nivel de confianza.

Calculad el contraste para un nivel de confianza del 95 % y del 90 %. Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No se pueden usar funciones como *t.test* para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

3.5. Conclusión

A partir de los resultados obtenidos, dad respuesta a la pregunta de investigación.

---

4. Proporción de Self-Employed

Nos preguntamos si el porcentaje de Self-Employed en la población es superior al 10 %. Aplicad el test necesario para dar respuesta a esta pregunta. Seguid los pasos que se indican a continuación.

4.1. Pregunta

Formulad la pregunta de investigación que se plantea en esta sección.

4.2. Hipótesis

Escribid la hipótesis nula y la hipótesis alternativa.

4.3. Análisis visual

Representad de forma gráfica la proporción de Self-Employed en la muestra.

4.4. Contraste

Explicad qué tipo de contraste podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

4.5. Cálculo

Calculad el test usando una función propia. Podéis crear una función que reciba los parámetros necesarios y el nivel de confianza. Luego, calculad el contraste, llamando esta función, con nivel de confianza del 95 %. Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No podéis usar *prop.test* o funciones ya implementadas en R. Sí podéis usar *qnorm*, *qt*, etcétera.

4.6. Conclusión

A partir de los resultados obtenidos, dad respuesta a la pregunta de investigación.

---

5. Proporción de Self-Employed en mujeres y hombres

Nos preguntamos si la proporción de Self-Employed es menor entre las mujeres que entre los hombres en la población. Para dar respuesta a esta pregunta, seguid los pasos que se indican a continuación.

5.1. Pregunta de investigación

Formulad la pregunta de investigación que se plantea en esta sección.

5.2. Análisis visual

Representad de forma gráfica la proporción de Self-Employed en la muestra de hombres y mujeres respectivamente.

5.3. Hipótesis

Escribid la hipótesis nula y la hipótesis alternativa.

5.4. Test

Explicad qué tipo de test podéis aplicar dada la pregunta de investigación planteada y las características de la muestra. Justificad vuestra elección.

5.5. Cálculo

Calculad el test usando una función propia. Al igual que en apartados anteriores, se recomienda definir una función que realice el cálculo y que reciba los parámetros necesarios.

Calculad el contraste para un nivel de confianza del 97 %. Mostrad los resultados (valor observado, crítico y valor p) en una tabla.

Nota: No podéis usar funciones como *prop.test* o funciones ya implementadas en R para el contraste. Sí podéis usar funciones básicas como *qnorm*, *qt*, etcétera.

5.6. Conclusión

A partir de los resultados obtenidos, proporcionad una respuesta a la pregunta de investigación.

---

6. Dependencia Género - Self-Employed

Otra forma de abordar si existen diferencias en la proporción de Self-Employed según el género es realizando un test de independencia de dos variables cualitativas. Concretamente, nos preguntamos si el género y ser Self-Employed están relacionadas o se pueden considerar variables independientes. Las variables serían independientes si el género no influye en la proporción de Self-Employed, es decir, si no hay diferencias en las proporciones de Self-Employed según el género.

En esta sección se pide aplicar el test de independencia Chi cuadrado para evaluar si las variables género y Self-Employed son independientes. Seguid los pasos que se indican a continuación.

6.1. Pregunta de investigación

Formulad la pregunta de investigación.

6.2. Hipótesis

Escribid la hipótesis nula y alternativa.

6.3. Test

Describid brevemente en qué consiste el test chi cuadrado. Calculad la matriz de contingencia y mostrad sus valores.

6.4. Cálculos

Realizad los cálculos del test Chi cuadrado, implementando una función propia. Calculad el contraste para un nivel de confianza de 97 %.

Nota: No podéis usar la función *chisq.test* de R. Sí podéis usar *pchisq* para consultar los valores de la distribución.

6.5. Conclusión

Responded la pregunta de investigación planteada en este apartado. Relacionad el resultado con la aproximación de la sección anterior, donde se realiza un test sobre las proporciones.

---

7. Resumen y conclusiones

Presentad una tabla con los resultados principales de cada sección: la pregunta de investigación planteada, los valores obtenidos del contraste y la conclusión obtenida en cada apartado. La tabla puede tener un formato como el que se muestra a continuación (se aporta un ejemplo para la primera fila de datos).

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
2	Intervalo de confianza de la media de edad al 95 %	(25.22,26.24)	El intervalo de confianza al 95 % es...
3	texto	valores	texto
4	texto	valores	texto
5	texto	valores	texto
6	texto	valores	texto

---

8. Puntuación de la actividad

- Apartado 1 (10 %)
- Apartado 2 (10 %)
- Apartado 3 (15 %)
- Apartado 4 (10 %)
- Apartado 5 (20 %)
- Apartado 6 (15 %)

- Apartado 7 (10 %)
- Calidad del informe dinámico (10 %)