A2 - Analítica descriptiva e inferencial

Enunciado

Semestre 2022.1

Índice

Introducción

En esta actividad nos introducimos en la inferencia estadística. Para ello, usaremos el conjunto de datos `gpa.csv` que se ha preprocesado en la actividad anterior. Este conjunto de datos contiene la nota media de estudiantes universitarios tras el primer semestre de clases (GPA: grade point average, en inglés), así como información sobre la nota de acceso, la cohorte de graduación en el instituto y algunas características de los estudiantes.

Este conjunto de datos surge de una encuesta realizada a una muestra representativa de estudiantes de una universidad de EEUU (por razones de confidencialidad el conjunto de datos no incluye el nombre de la universidad). Las variables incluidas en el conjunto de datos son:

- sat: nota de acceso (medida en escala de 400 a 1600 puntos)
- tothrs: horas totales cursadas en el semestre
- colgpa: nota media del estudiante al final del primer semestre (medida en escala de 0 a 4 puntos)
- athlete: indicador de si el estudiante practica algún deporte en la universidad
- hsize: número total de estudiantes en la cohorte de graduados del bachillerato (en cientos)
- hsrank: ranking del estudiante, dado por la nota media del bachillerato, en su cohorte de graduados del bachillerato
- hsperc: ranking relativo del estudiante, en porcentaje (hsrank/hsize)
- female: indicador de si el estudiante es mujer
- white: indicador de si el estudiante es de raza blanca o no
- black: indicador de si el estudiante es de raza negra o no
- gpaletter: letra que indica el nivel de la nota colgpa (A,B,C,D)

El objetivo de esta actividad es estudiar la nota de los estudiantes a partir de las variables de interés así como la proporción de atletas entre la población de estudiantes. Para ello, las preguntas que nos planteamos son:

- P1. ¿Cuál es el intervalo de confianza de la nota entre los estudiantes?
- P2. ¿Ser atleta influye en la nota?
- P3. ¿Las mujeres obtienen mejor nota que los hombres?
- P4. ¿Hay diferencias significativas en la nota según la raza?
- P5. ¿La proporción de atletas en la población es inferior al 5%?
- P6. ¿Hay más atletas entre los hombres que entre las mujeres?

Al final del análisis, os pedimos un resumen ejecutivo donde se deben resumir y explicar brevemente las conclusiones del estudio.

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

Leed el fichero `gpa.csv` y guardad los datos en un objeto denominado `gpa`. A continuación, verificad el tipo de cada variable.

---

2 Estadística descriptiva y visualización

2.1 Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas). Mostrad el número de observaciones y el número de variables.

2.2 Visualización

Estudiaremos de forma visual la distribución de las variables `sat` y `colgpa`:

1. Mostrad la distribución de las variables 'sat' y 'colgpa' (por separado).
2. Mostrad la distribución de la variable 'sat' con respecto a la variable género ('female'), la variable atleta ('athlete') y la raza ('white', 'black')
3. Realizad el mismo tipo de visualizaciones con la variable 'colgpa' y las variables 'female', 'athlete' y 'white'/'black'.
4. Interpretad los gráficos brevemente.

---

3 Intervalo de confianza de la media poblacional de la variable `sat` y `colgpa`

3.1 Supuestos

Revisad los supuestos necesarios para el cálculo del intervalo de confianza y si éstos se cumplen para las variables de interés.

3.2 Función de cálculo del intervalo de confianza

Definid una función IC que calcule el intervalo de confianza de una variable a partir del nivel de confianza. La función recibe como parámetro la variable, el nivel de confianza, y devuelve un vector con los valores del intervalo de confianza.

La cabecera de la función es:

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

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

3.3 Intervalo de confianza de la variable `sat`

Calculad el intervalo de confianza al 90% de la media poblacional de la variable `sat` de los estudiantes. Repetid el cálculo para un intervalo de confianza del 95%.

3.4 Intervalo de confianza de la variable `colgpa`.

Calculad el intervalo de confianza de la variable `colgpa` con un nivel de confianza del 90% y del 95%.

3.5 Interpretación

A partir de los resultados obtenidos, explicad cómo se interpreta el intervalo de confianza.

Nota: Se recomienda que guardéis toda esta información en variables para el posterior resumen ejecutivo.

---

4 ¿Ser atleta influye en la nota?

En este apartado queremos analizar si ser atleta influye en la nota `colgpa`. Es decir, si hay diferencias significativas entre atletas y no atletas en esta nota, con un nivel de confianza del 95%.

Como realizaremos preguntas similares en los siguientes apartados, se recomienda implementar una función que permita realizar contrastes de hipótesis para la diferencia de medias. La función debe recibir como parámetro las dos muestras, el nivel de confinza requerido y otros parámetros que se puedan requerir.

Seguid los pasos que se detallan a continuación.

4.1 Análisis visual

Realizad un pequeño análisis visual para orientar esta pregunta.

4.2 Función para el contraste de medias

Implementad una función que calcule el contraste entre medias y que devuelva: el valor del estadístico de contraste, el valor crítico, y el valor p. Esta función se debe implementar después de analizar el apartado 4.5 (pero por motivos de estructura del documento, os pedimos que la desarrolléis en este apartado).

Nota: se deben realizar los cálculos manualmente. No se pueden usar funciones de `R` que calculen directamente el contraste como `t.test` o similar. Sí se puede usar `var.test` y funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`.

4.3 Pregunta de investigación

Formulad la pregunta de investigación que se plantea en este apartado.

4.4 Hipótesis nula y la alternativa

Escribid las hipótesis nula y alternativa.

4.5 Justificación del test a aplicar

Explicad qué test es el más adecuado para dar respuesta a esta pregunta y por qué. Comprobad si se cumplen los supuestos de aplicación del test.

4.6 Cálculo

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Usad la función implementada anteriormente.

4.7 Interpretación del test

Concluid dando respuesta a la pregunta planteada.

---

5 ¿Las mujeres tienen mejor nota que los hombres?

En este apartado queremos analizar si el género influye en la nota `colgpa`. Específicamente, si las mujeres tienen mejor nota que los hombres, con un nivel de confianza del 95%. Comprobad también para un nivel de confianza del 90%. Seguid los mismos pasos que anteriormente.

5.1 Análisis visual

Realizad un análisis visual para orientar la respuesta a esta pregunta de investigación.

5.2 Función

Si podéis usar la misma función anterior, no es necesario implementarla de nuevo. De lo contrario, implementad la función que será apropiada para dar respuesta a esta pregunta. Igual que anteriormente, esta función se debe implementar (si es necesaria) después de analizar el apartado 5.5.

5.3 Pregunta de investigación

Formulad la pregunta de investigación que se plantea en este apartado.

5.4 Hipótesis nula y la alternativa

Escribid las hipótesis nula y alternativa.

5.5 Justificación del test a aplicar

Explicad qué test es el más adecuado para dar respuesta a esta pregunta y por qué. Comprobad si se cumplen los supuestos de aplicación del test.

5.6 Cálculo

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95% (y si es necesario del 90%). Usad la función implementada anteriormente.

5.7 Interpretación del test

Concluid dando respuesta a la pregunta planteada.

---

6 ¿Hay diferencias en la nota según la raza?

Antes de desarrollar este apartado, queremos hacer énfasis que el hecho que existan (si existen) diferencias entre razas (blancos y negros), no necesariamente nos lleva a concluir que la raza influye en la nota (existen

factores socioeconómicos que pueden afectar). Por tanto, se debe ser muy cauteloso a la hora de realizar conclusiones en este sentido. Un tema similar puede suceder con las diferencias en cuanto a género. A pesar de estas puntualizaciones, el estudio es interesante puesto que en caso de detectar diferencias en uno u otro sentido, se pueden analizar los causantes de estas diferencias y realizar intervenciones educativas apropiadas.

Después de este matiz, nos gustaría estudiar si las personas de raza blanca tienen una nota diferente en `colgpa` que las personas de raza negra. Seguid los mismos apartados que anteriormente.

6.1 Análisis visual

6.2 Función

6.3 Pregunta de investigación

6.4 Hipótesis nula y la alternativa

6.5 Justificación del test a aplicar

6.6 Cálculo

6.7 Interpretación del test

---

7 Proporción de atletas

Nos preguntamos si la proporción de atletas en la población es inferior al 5% con un nivel de confianza del 95%. Para ello, seguid los mismos pasos que en los casos anteriores.

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

7.1 Análisis visual

7.2 Pregunta de investigación

7.3 Hipótesis nula y la alternativa

7.4 Justificación del test a aplicar

7.5 Realizad los cálculos del test

7.6 Interpretación del test

---

8 ¿Hay más atletas entre los hombres que entre las mujeres?

Nos preguntamos si la proporción de atletas entre los hombres es superior a la de las mujeres con un nivel de confianza del 95%. Seguid los mismos pasos que anteriormente.

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

8.1 Análisis visual

8.2 Pregunta de investigación

8.3 Hipótesis nula y la alternativa

8.4 Justificación del test a aplicar

8.5 Realizad los cálculos del test

8.6 Interpretación del test

---

9 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). Esta tabla nos ayuda a tener una visión general y técnica de los resultados del estudio.

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
P1	Intervalo de confianza de la media de stat al 95%	(inf,sup)	El intervalo de confianza al 95% es...
P2	texto	valores	texto
P3	texto	valores	texto
P4	texto	valores	texto
P5	texto	valores	texto
...	...	...	...

---

10 Resumen ejecutivo

Resumid las conclusiones del estudio para una audiencia no técnica, indicando las respuestas a las preguntas de investigación planteadas. El resumen no debe ocupar más de media página.

Nota: esta pregunta trabaja la competencia de comunicación que es muy importante en el rol de analista de datos.

Puntuación de la actividad

- Apartados 1,2 (10%)
- Apartado 3 (20%)
- Apartado 4 (10%)
- Apartado 5 (5%)
- Apartado 6 (5%)
- Apartado 7 (10%)
- Apartado 8 (10%)
- Apartados 9, 10 (20%)

- Calidad del informe dinámico (10%)