Estadística Avanzada - Actividad 4

Enunciado

Semestre 2022.1

Índice

Introducción

En una investigación médica se estudió la capacidad pulmonar de los fumadores y no fumadores. Se recogieron datos de una muestra de la población fumadora, no fumadora y fumadores pasivos. A cada persona se realizó un test de capacidad pulmonar consistente en evaluar la cantidad de aire expulsado (AE). La muestra de n individuos se categorizó en 6 tipos:

- No fumadores (NF)
- Fumadores pasivos (FP)
- Fumadores que no inhalan (NI): personas que fuman pero no inhalan el humo.
- Fumadores ligeros (FL): personas que fuman e inhalan de uno a 10 cigarrillos al día durante 20 años o más.
- Fumadores moderados (FM): personas que fuman e inhalan entre 11 y 39 cigarrillos por día durante 20 años o más.
- Fumadores intensivos (FI): personas que fuman e inhalan 40 cigarrillos o más durante 20 años o más.

En esta actividad se analizará si la capacidad pulmonar está influida por el tipo de fumador. Para ello, se aplicaran distintos tipos de análisis, revisando los contrastes de hipótesis de dos muestras, vistos en la actividad A2, y luego realizando análisis más complejos como ANOVA.

Notas importantes a tener en cuenta para la entrega de la actividad:

- Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida debe incluir el código y el resultado de su ejecución (paso a paso). Se debe incluir un índice o tabla de contenidos. Y se debe respetar la numeración de los apartados del enunciado.
- No realizar listados de los conjuntos de datos, puesto que estos pueden ocupar varias páginas. Si queréis comprobar el efecto de una instrucción sobre un conjunto de datos podéis usar la función **head** o **tail** que muestran las primeras o últimas filas del conjunto de datos.

1 Preprocesado

Cargar el fichero de datos "Fumadores.csv". Consultar los tipos de datos de las variables y si es necesario, aplicar las transformaciones apropiadas. Averiguar posibles inconsistencias en los valores de Tipo, AE, género y edad. En caso de que existan inconsitencias, corregidlas.

2 Análisis descriptivo de la muestra

2.1 Capacidad pulmonar y género

Mostrar la capacidad pulmonar en relación al género. ¿Se observan diferencias?

2.2 Capacidad pulmonar y edad

Mostrar la relación entre capacidad pulmonar y edad usando un gráfico de dispersión. Interpretar.

2.3 Tipos de fumadores y capacidad pulmonar

Mostrar el número de personas en cada tipo de fumador y la media de AE de cada tipo de fumador. Mostrad un gráfico que visualice esta media. Se recomienda que el gráfico esté ordenado de menos a más AE.

Luego, se debe representar un boxplot donde se muestre la distribución de AE por cada tipo de fumador. Interpretar los resultados.

Nota: Para calcular la media o otras variables para cada tipo de fumador, podéis usar las funciones **summarize** y **group_by** de la libreria **dplyr** que os serán de gran utilidad. Para realizar la visualización de los datos, podéis usar la función **ggplot** de la librería **ggplot2**.

3 Intervalo de confianza de la capacidad pulmonar

Calcular el intervalo de confianza al 95% de la capacidad pulmonar de las mujeres y hombres por separado. Antes de aplicar el cálculo, revisar si se cumplen las asunciones de aplicación del intervalo de confianza. Interpretar los resultados. A partir de estos cálculos, ¿se observan diferencias significativas en la capacidad pulmonar de mujeres y hombres?

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

4 Diferencias en capacidad pulmonar entre mujeres y hombres

Aplicar un contraste de hipótesis para evaluar si existen diferencias significativas entre la capacidad pulmonar de mujeres y hombres. Seguid los pasos que se indican a continuación.

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

4.1 Hipótesis

Escribir la hipótesis nula y alternativa.

4.2 Contraste

Explicad qué tipo de contraste aplicaréis y por qué. Si es necesario, validad las asunciones del test.

4.3 Cálculos

Aplicad los cálculos del contraste. Mostrar el valor observado, el valor de contraste y el valor p.

4.4 Interpretación

Interpretad los resultados y comparad las conclusiones con los intervalos de confianza calculados anteriormente.

5 Diferencias en la capacidad pulmonar entre Fumadores y No Fumadores

¿Podemos afirmar que la capacidad pulmonar de los fumadores es inferior a la de no fumadores? Incluid dentro de la categoría de no fumadores los fumadores pasivos. Seguid los pasos que se indican a continuación.

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

5.1 Hipótesis

Escribir la hipótesis nula y alternativa.

5.2 Contraste

Explicad qué tipo de contraste aplicaréis y por qué. Si es necesario, validad las asunciones del test.

5.3 Preparación de los datos

Preparad las muestras. Una de ellas contiene los valores de AE de los fumadores y la otra, los valores de AE de los no fumadores y fumadores pasivos.

5.4 Cálculos

Aplicad los cálculos del contraste. Mostrar el valor observado, el valor de contraste y el valor p.

5.5 Interpretación

Interpretar el resultado del contraste

6 Análisis de regresión lineal

Realizamos un análisis de regresión lineal para investigar la relación entre la variable capacidad pulmonar (AE) y el resto de variables (tipo, edad y género). Construid e interpretad el modelo, siguiendo los pasos que se especifican a continuación.

6.1 Cálculo

Calculad el modelo de regresión lineal. Podéis usar la función *lm*.

6.2 Interpretación

Interpretad el modelo y la contribución de cada variable explicativa sobre la variable AE.

6.3 Bondad de ajuste

Evaluad la calidad del modelo.

6.4 Predicción

Realizad una predicción de la capacidad pulmonar para cada tipo de fumador desde los 30 años de edad hasta los 80 años de edad (podéis asumir género hombre). Mostrad una tabla con los resultados. Mostrad también visualmente la simulación.

7 ANOVA unifactorial

A continuación se realizará un análisis de varianza, donde se desea comparar la capacidad pulmonar entre los seis tipos de fumadores/no fumadores clasificados previamente. El análisis de varianza consiste en evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias variables independientes, denominadas factores. En el caso que nos ocupa, nos interesa evaluar si la variabilidad de la variable AE puede explicarse por el factor tipo de fumador. Hay dos preguntas básicas a responder:

- ¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?
- Si existen diferencias, ¿entre qué grupos están estas diferencias?

7.1 Normalidad

Evaluar si el conjunto de datos cumple las condiciones de aplicación de ANOVA. Seguid los pasos que se indican a continuación. Mostrad visualmente si existe normalidad en los datos y también aplicar un test de normalidad.

Nota: podéis usar el gráfico normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de los residuos.

7.2 Homoscedasticidad: Homogeneidad de varianzas

Otra de las condiciones de aplicación de ANOVA es la igualdad de varianzas (homoscedasticidad). Aplicar un test para validar si los grupos presentan igual varianza. Aplicad el test adecuado e interpretar el resultado.

Nota: podéis usar tests como el de Levene o Bartlett test.

7.3 Hipótesis nula y alternativa

Independientemente de los resultados sobre la normalidad e homoscedasticidad de los datos, proseguiremos con la aplicación del análisis de varianza. Concretamente, se aplicará ANOVA de un factor (one-way ANOVA o independent samples ANOVA) para investigar si existen diferencias en el nivel de aire expulsado (AE) entre los distintos tipos de fumadores. Escribid la hipótesis nula y alternativa.

7.4 Cálculo ANOVA

Podéis usar la función **aov**.

7.5 Interpretación

Interpretad los resultados de la prueba ANOVA y relacionarlos con el resultado gráfico del boxplot mostrado en el apartado 2.3.

7.6 Profundizando en ANOVA

A partir de los resultados del modelo devuelto por **aov**, identificar las variables SST (Total Sum of Squares), SSW (Within Sum of Squares), SSB (Between Sum of Squares) y los grados de libertad. A partir de estos valores, calcular manualmente el valor F, el valor crítico (a un nivel de confianza del 95%), y el valor p. Interpretar los resultados y explicar el significado de las variables SST, SSW y SSB.

7.7 Fuerza de la relación

Calcular la fuerza de la relación e interpretar el resultado.

8 Comparaciones múltiples

Independientemente del resultado obtenido en el apartado anterior, realizamos un test de comparación múltiple entre los grupos. Este test se aplica cuando el test ANOVA devuelve rechazar la hipótesis nula de igualdad de medias. Por tanto, procederemos como si el test ANOVA hubiera dado como resultado el rechazo de la hipótesis nula.

8.1 Test pairwise

Calcular las comparaciones entre grupos sin ningún tipo de corrección. Podéis usar la función **pairwise.t.test**. Interpretar los resultados.

8.2 Corrección de Bonferroni

Aplicar la corrección de Bonferroni en la comparación múltiple. Interpretar el resultado y contrastar el resultado con el obtenido en el test de comparaciones múltiples sin corrección.

9 ANOVA multifactorial

En una segunda fase de la investigación se evalua el efecto del género como variable independiente, además del efecto del tipo de fumador, sobre la variable AE.

9.1 Análisis visual

Se realizará un primer estudio visual para determinar si existen efectos principales o hay efectos de interacción entre género y tipo de fumador. Para ello, seguir los pasos que se indican a continuación:

1. Agrupar el conjunto de datos por tipo de fumador y género y calcular la media de AE en cada grupo. Podéis usar las instrucciones **group_by** y **summarise** de la librería **dplyr** para realizar este proceso. Mostrar el conjunto de datos en forma de tabla, donde se muestre la media de cada grupo según el género y tipo de fumador.
2. Mostrar en un gráfico el valor de AE medio para cada tipo de fumador y género. Podéis realizar este tipo de gráfico usando la función **ggplot** de la librería **ggplot2**.
3. Interpretar el resultado sobre si existen sólo efectos principales o existe interacción. Si existe interacción, explicar cómo se observa y qué efectos produce esta interacción.

9.2 ANOVA multifactorial

Calcular ANOVA multifactorial para evaluar si la variable dependiente AE se puede explicar a partir de las variables independientes género y tipo de fumador. Incluid el efecto de la interacción.

9.3 Interpretación

Interpretad el resultado.

10 Resumen técnico

Realizad una tabla con el resumen técnico de las preguntas de investigación planteadas a lo largo de esta actividad.

N	Pregunta	Resultado (valor observado, crítico, valor p...)	Conclusión
P1	IC AE mujeres 95%	(inf,sup)	El IC es...
P2	texto	valores	texto
P3	texto	valores	texto
P4	texto	valores	texto
P5	texto	valores	texto
...	...	...	...

11 Resumen ejecutivo

Escribid un resumen ejecutivo como si tuvieráis que comunicar a una audiencia no técnica. Por ejemplo, podría ser un equipo de gestores o decisores, a los cuales se les debe informar sobre las consecuencias de fumar sobre la capacidad pulmonar, para que puedan tomar las decisiones necesarias.

Puntuación de la actividad

- Pregunta 1: 10%
- Pregunta 2: 10%
- Pregunta 3: 10%
- Preguntas 4,5: 10%
- Pregunta 6: 10%
- Pregunta 7: 10%
- Pregunta 8: 10%
- Pregunta 9: 10%
- Pregunta 10: 10%
- Pregunta 11: 10%