A2 - Analítica Descriptiva e Inferencial

Enunciado

Semestre 2019.1

Índice

Introducción

En esta actividad se realizará un análisis estadístico descriptivo e inferencial de los datos procesados en la actividad 1. Recordamos que el conjunto de datos usado en la actividad previa consistía en datos recolectados en una investigación sobre mujeres premenopáusicas, a las que se medía el nivel de estradiol. El objetivo

del estudio era determinar posibles relaciones entre el nivel de estradiol y la obesidad, así como con otros parámetros como la edad de la mujer o haber tenido hijos.

Concretamente, los atributos del fichero son:

- Id: identificador.
- Estrad (serum estradiol): medida del estradiol (analítica hormonal).
- Ethnic (ethnicity): African American o Caucasian.
- Entage: edad de la persona.
- NumChild: número de hijos.
- Agefbo: edad a la que la persona ha tenido el primer hijo.
- Anykids: "Yes" si ha tenido hijos, "No" si no ha tenido.
- Agemenar: edad de la menarquia (primera menstruación).
- BMI: medida de la adiposidad general. Corresponde al ratio $peso(kg)/Altura^2(m)$.
- WHR: medida de la adiposidad abdominal. Corresponde al ratio $cintura/cadera$.
- Area: "Rural" o "Urban".

Puesto que el resultado del preprocesado de los datos puede ser ligeramente distinto entre las distintas soluciones que habéis aportado, os suministramos el fichero preprocesado. Esta actividad se realizará con el fichero que os suministramos, independientemente del proceso de preprocesado que hayáis realizado en la actividad anterior. El nombre del fichero es **ESTRADL_cleean.csv**.

Nota importante a tener en cuenta para entregar 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 la ejecución del mismo (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.

1. Analítica descriptiva

1.1. Lectura del fichero

Leer el fichero **ESTRADL_clean.csv**. Validar que los datos leídos son correctos. Si no es así, realizar las conversiones oportunas.

1.2. Análisis descriptivo visual

Representar de forma visual las variables del conjunto de datos y las distribuciones de sus valores. Escoged la representación más apropiada en cada caso.

2. Edad de menarquia media de la población

A partir de los datos de la muestra, se desea estimar la edad de menarquia media de las mujeres. En concreto, se desea estudiar si la edad de menarquia media de la población es de 14 años, o bien es inferior a 14 años. Para ello, realizar un contraste estadístico con un nivel de confianza del 98 %. Seguid los pasos que se indican a continuación.

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el intervalo de confianza. En cambio, sí se pueden usar funciones como qnorm, pnorm, qt y pt.

2.1. Hipótesis nula y alternativa

Escribid la hipótesis nula y la hipótesis alternativa.

2.2. Método

Indicad cuál es el método más apropiado para realizar este contraste, en función de las características de la muestra.

2.3. Cálculos

Calcular el estadístico de contraste, el valor crítico y el valor p.

2.4. Interpretación

A partir de los resultados obtenidos, realizar la interpretación de los mismos, especialmente en relación a la hipótesis que se formula al inicio de esta sección.

3. Intervalo de confianza de Estradiol

3.1. Calcular el intervalo de confianza del 95 % de la variable Estradiol

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el intervalo de confianza. En cambio, sí se pueden usar funciones como qnorm, pnorm, qt y pt. Escribid las instrucciones paso a paso e imprimid los resultados de los pasos más relevantes.

3.2. Interpretar el resultado

A partir del resultado obtenido del intervalo de confianza, explicar la interpretación del mismo en cuanto al valor de estradiol en las mujeres.

3.3. Comparar intervalos

Si calculáramos el intervalo de confianza del 97 %, ¿cómo sería el intervalo de confianza en relación al calculado previamente? Justificar la respuesta. No es necesario realizar los cálculos.

4. Diferencias en el nivel de estradiol según etnia

En la investigación médica, se han planteado un conjunto de hipótesis que se desean validar estadísticamente a partir de análisis inferencial. La primera hipótesis es que los niveles de estradiol en mujeres caucásicas es diferente del de las mujeres negras. ¿Qué dicen los datos al respecto? Seguid los pasos que se indican a continuación.

4.1. Escribir la hipótesis nula y alternativa

4.2. Método

En función de las características de la muestra, decidir qué método aplicar para validar la hipótesis planteada. Para ello, debéis especificar como mínimo: a) si es un contraste de una muestra o de dos muestras (en caso de dos muestras, si éstas son independientes o están relacionadas), b) si podéis asumir normalidad y por qué, c) si el test es paramétrico o no paramétrico, d) si el test es bilateral o unilateral.

4.3. Cálculos

Realizar los cálculos para validar o rechazar la hipótesis de la investigación, con un nivel de confianza del 95 %. Calcular: el estadístico de contraste, el valor crítico y el valor p.

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el contraste de hipótesis. En cambio, sí se pueden usar funciones como qnorm, pnorm, qt y pt. Se aconseja escribir una función propia en R que realice los cálculos necesarios para que se pueda aprovechar posteriormente, si es necesario.

4.4. Interpretar

Interpretar los resultados y concluir si se puede afirmar que existen diferencias significativas en el nivel de estradiol según la etnia.

5. Nivel de estradiol según los hijos

A continuación, se desea evaluar si existen diferencias en el nivel de estradiol de las mujeres según si han tenido hijos o no. Es decir, se podría afirmar que el nivel de estradiol es inferior en las mujeres que han tenido hijos, con un nivel de confianza del 95 %? ¿Y con un nivel de confianza del 90 %?

Seguir los pasos que se indican a continuación para dar respuesta a esta hipótesis (que son análogos a la pregunta anteior). Recordad que se deben realizar los cálculos manualmente. No se pueden usar funciones de R que calculen directamente el contraste de hipótesis. En cambio, sí se pueden usar funciones como qnorm, pnorm, qt y pt.

Especificad todos los pasos detalladamente e imprimid los resultados de las variables relevantes de este contraste, tal como se requiere en la sección anterior.

5.1. Escribir la hipótesis nula y alternativa

5.2. Método

5.3. Cálculos

5.4. Interpretación

6. Estudio longitudinal: ¿estradiol aumenta con los años?

Los investigadores del estudio encontraron que existía una posible correlación entre la edad de las mujeres y el nivel de estradiol. Por ello, realizaron un estudio longitudinal con una muestra reducida de mujeres. En un

grupo de 10 mujeres voluntarias de la muestra original, se midió los niveles de estradiol al cabo de 7 años. El fichero **ESTRAD7.csv** recoge esta medida. Concretamente, el fichero contiene el identificador de la mujer, el nivel de estradiol original y su nivel de estradiol medido al cabo de 7 años del estudio original. La hipótesis de la investigación es que estradiol aumenta con la edad. ¿Qué dicen los datos en relación a esta hipótesis? ¿Podemos afirmar que aumenta el nivel de estradiol con un nivel de confianza del 97 %?

Seguid los pasos que se indican a continuación.

6.1. Escribir la hipótesis nula y alternativa

6.2. Asunción de normalidad

Para determinar el tipo de prueba a aplicar, se comprueba primero si se cumple la asunción de normalidad de los datos. Para ello, podemos examinar si se puede aplicar el teorema del límite central. Además, se puede realizar una visualización gráfica con las curvas Q-Q y aplicar el test de Shapiro-Wilk. Podéis consultar la información siguiente:

https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/qqnorm (curvas Q-Q en R),

https://data.library.virginia.edu/understanding-q-q-plots/ (interpretación de las curvas Q-Q)

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html (función shapiro.test en R)

En este apartado debéis realizar estas comprobaciones y determinar si se puede asumir normalidad. Justificar vuestra conclusión en base a los resultados obtenidos.

6.3. Método

Independientemente de la conclusión obtenida en el apartado anterior, aplicar un test no paramétrico a la muestra. Existen dos tipos de tests no paramétricos para el contraste de dos muestras: 1) el test de suma de rangos (también conocido como test U de Mann-Whitney): https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/ y 2) el test de rangos y signos de Wilcoxon: https://www.r-bloggers.com/wilcoxon-signed-rank-test/

Decidid cuál es el contraste que debe aplicarse en este caso. Debéis justificar vuestra elección.

6.4. Cálculo e interpretación

Aplicad el contraste e interpretar el resultado. Podéis usar funciones R (no desarrolléis los cálculos en este apartado).

6.5. Explicar el test escogido

Explicar brevemente cómo se calcula el test que habéis escogido en el apartado anterior. La explicación no debe ser en base a un código, sino en vuestras propias palabras. Tratad de ser claros y concisos en la explicación

7. Conclusiones

En esta sección, debéis realizar un resumen de los aprendizajes adquiridos al realizar la actividad:

- En base a los conceptos aprendidos al realizar esta actividad y al consultar la documentación relacionada, escribid el proceso para realizar un contraste de hipótesis de dos muestras. Es decir, explicar brevemente qué tipos de contrastes podemos aplicar y cuáles son las condiciones de aplicación de los mismos. Considerad todas las "configuraciones" posibles: tests paramétricos-no paramétricos, pareados-independientes, bilaterales-unilaterales...

8. Comentarios importantes sobre la actividad

1. **No se puede inspeccionar ni corregir de manera manual** el fichero de datos. Por ejemplo, **no** se pueden realizar instrucciones de este tipo:

   ```
   data[1,5] <- 32.5
   ```

   Este tipo de transformaciones se deben hacer con funcionalidades de búsqueda (buscar los registros que tienen errores o inconsistencias) y luego hacer las correcciones oportunas con funcionalidades de R. Así el procedimiento de limpieza es útil, independientemente del fichero de datos y de la posición y valores concretos del archivo.

2. **No se pueden hacer listados completos de los datos del fichero a pantalla**, porque generan archivos de salida excesivamente grandes. Si se desea validar el resultado de una instrucción sobre los datos, se puede usar la función **head** que muestra las primeras filas de la tabla de datos o **tail** que muestra las últimas.

Puntuación de la actividad

- Apartado 1 (10 %)
- Apartado 2 (10 %)
- Apartado 3 (10 %)
- Apartado 4 (20 %)
- Apartado 5 (10 %)
- Apartado 6 (20 %)
- Apartado 7 (10 %)
- Calidad del informe dinámico (10 %)