A2 - Estadística inferencial: tests de hipótesis de una y dos muestras

Enunciado

Semestre 2020.1

Índex

Introducción

Los datos que se utilizarán en esta acitividad se corresponden con las ventas de sillitas infantiles para el coche en diferentes tiendas. Estos datos se preprocesaron en la actividad anterior.

Recordemos que las variables del conjunto de datos son:

- Sales: ventas unitarias, en miles, en cada ubicación.
- CompPrice: precio que cobra la competencia en cada ubicación.
- Income: nivel de ingresos comunitarios, en miles de dólares.
- Advertising: presupuesto de publicidad local de la empresa en cada ubicación, en miles de dólares.
- Population: tamaño de la población en la región, en miles.
- Price: precio de las sillitas de coche en cada ubicación.
- ShelveLoc: variable categórica con los niveles Bad, Good y Medium, que indica la calidad de la ubicación de las sillas de coche en la tienda.
- Age: edad media de la población local.
- Education: valor numérico que indica la media del nivel de educación (años de educación) de la población.
- Urban: variable binaria con los niveles Yes y No para indicar si la tienda se encuentra en una ubicación urbana o rural.
- US: variable binaria con los niveles Yes y No para indicar si la tienda se encuentra en EUA o no.

Los datos del estudio están en el archivo ChildCarSeats_clean.csv. Os proporcionamos el archivo resultante de la fase de preprocesado para que lo uséis independientemente del que hayáis obtenido en vuestra actividad anterior. De esta manera, todos usáis el mismo conjunto de datos.

En esta actividad se realizará una investigación para estudiar el comportamiento de las ventas en diferentes tiendas y los factores clave que influyen en las ventas. Usando estadística descriptiva e inferencial, se trata de responder a las cuestiones siguientes:

1. ¿Cuál es el valor promedio de las ventas en las tiendas?
2. ¿Venden más las tiendas de USA que las de fuera de USA?
3. ¿Son diferentes las ventas de zonas urbanas en comparación con las ventas de zonas rurales?
4. ¿Más de la mitad de las tiendas tienen un precio inferior al de la competencia?
5. ¿Hay una estrategia de precios diferente en las tiendas de USA en relación a las de fuera de USA?

Aspectos importantes 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.
- El nivel de confianza por defecto es del 95%, a no ser que se especifique un valor diferente.
- 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.
- El resumen ejecutivo que se pide en este documento es una buena forma de adquirir competencias de comunicación en el ámbito de la estadística basadas en la precisión y concisión.

1 Estadística descriptiva y visualización

En primer lugar, leed el fichero de datos y verificad que los tipos de datos se interpretan correctamente. Si fuera necesario, haced las oportunas conversiones de tipos.

A continuación, realizad una visualización gráfica de los datos del conjunto de datos. Para evitar un exceso de páginas en el documento, podéis agrupar gráficos en un mismo panel (layout). Podéis consultar información en las fuentes siguientes:

1. https://www.statmethods.net/advgraphs/layout.html
2. En el caso de usar la librería ggplot2: https://cran.r-project.org/web/packages/egg/vignettes/Ecosystem.html

Explicad brevemente los gráficos y lo que se puede observar a partir de ellos.

2 Intervalo de confianza de la media poblacional de las ventas

Calculad el intervalo de confianza de la media poblacional de la variable Sales.

2.1 Cálculo

Debéis realizar los cálculos manualmente. No se pueden usar funciones como **t.test** o similares que realicen el cálculo del intervalo de confianza. Si acaso, solamente se pueden usar estas funciones para validar los resultados de vuestros cálculos. Sí que podéis usar funciones básicas como *mean*, *sd*, *qnorm*, *pnorm*.

2.2 Interpretación

Explicad el concepto de intervalo de confianza de la media poblacional a partir del resultado obtenido en la variable Sales.

2.3 Intervalo de confianza de la media poblacional de Sales en USA y fuera de USA

Calculad el intervalo de confianza de la media poblacional de Sales en las tiendas de USA y en las tiendas de fuera de USA respectivamente. Comparad los resultados. ¿Podemos concluir que las dos variables tienen medias poblacionales iguales o diferentes? Explicar.

Nota para la mejor calidad del código: si en el apartado anterior definís una función propia para el cálculo del intervalo de confianza, aquí podéis reaprovecharla sin tener que escribir de nuevo todos los cálculos.

3 Ventas del producto en USA y fuera de USA

Para evaluar si las ventas del producto son superiores en las tiendas de USA que fuera de USA, podemos aplicar un test de hipótesis de dos muestras. Seguid los pasos que se indican a continuación.

3.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

3.2 Test a aplicar

Revisad si se cumple el supuesto de normalidad para la variable Sales y a partir de este resultado, explicad qué test aplicaréis para el test de hipótesis de dos muestras. Podéis usar los tests y las funciones de R que consideréis necesarios para validar el cumplimiento de condiciones.

3.3 Cálculos

Calculad el test de hipótesis. Debéis desarrollar todos los cálculos con vuestras propias instrucciones. Calculad el valor observado, el valor crítico y el valor p. Al igual que antes, no podéis usar fuciones de R como **t.test**, pero sí podéis usar funciones como *qnorm*, *pnorm*, *qt*, *pt* para consultar los valores de las distribuciones normal y t-Student.

3.4 Conclusión

Concluid si se puede afirmar que las ventas en USA son superiores a las ventas fuera de USA con un 95% de nivel de confianza. A continuación, comparad esta conclusión con la de la sección 2.3.

4 Ventas en zonas urbanas y rurales

Nos preguntamos ahora si las ventas en zonas urbanas son diferentes de las ventas en zonas rurales. Realizad un test de hipótesis de dos muestras para responder esta pregunta. Seguid los mismos pasos que en la sección anterior.

4.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

4.2 Test a aplicar

Indicad qué test aplicaréis.

4.3 Cálculos

Calculad el test de hipótesis. Debéis desarrollar todos los cálculos con vuestras propias instrucciones. Calculad el valor observado, el valor crítico y el valor p. Al igual que antes, no podéis usar fuciones de R como **t.test**, pero sí podéis usar funciones como *qnorm*, *pnorm*, *qt*, *pt* para consultar los valores de las distribuciones normal y t-Student.

Si acaso, podéis aprovechar funcionalidades que hayáis desarrollado anteriormente.

4.4 Conclusión

Responded la pregunta formulada en este apartado.

5 Estrategia de precios

Nos preguntamos si la proporción de tiendas que venden el producto por debajo del precio de la competencia es mayor que las que venden por encima del precio de la competencia.

Para responder esta pregunta, se recomienda plantear un test sobre la proporción. Seguid los pasos que se indican a continuación.

5.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa teniendo en cuenta la pregunta formulada.

5.2 Tipo de test

Indicad qué tipo de test aplicaréis y por qué.

5.3 Cálculos

Realizad todos los cálculos con instrucciones propias. Calculad el valor observado, el valor crítico y el valor p. Mostrad los resultados. No se pueden usar fuciones de R que resuelvan el cálculo (en todo caso, solo se pueden usar para validar vuestros resultados). Sí podéis usar funciones como *qnorm*, *pnorm*, *qt*, *pt* para consultar los valores de las distribuciones normal y t-Student.

5.4 Conclusión

A partir de los valores obtenidos, responded la pregunta formulada.

6 Diferencias en la estrategia de precios

Se cree que las tiendas de USA usan una estrategia de precios más agresiva en relación con las tiendas de fuera de USA. Para investigar esta hipótesis, calculamos en cuantas ocasiones el precio de la tienda es inferior al precio de la competencia. A partir de este cálculo, nos preguntamos: **¿la proporción de casos en los que el precio de la tienda es más bajo que la competencia (estrategia de precios bajos) es diferente en las tiendas de USA que en las tiendas fuera de USA?**

Para responder esta pregunta, se recomienda plantear un test sobre la proporción.

Seguid los pasos que se indican a continuación.

6.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa teniendo en cuenta la pregunta formulada.

6.2 Tipo de test

Indicad qué tipo de test aplicaréis y justificadlo.

6.3 Cálculos

Realizad todos los cálculos con instrucciones propias. Calculad el valor observado, el valor crítico y el valor p. Mostrad los resultados. No se pueden usar fuciones de R que resuelvan el cálculo (en todo caso, solo se pueden usar para validar vuestros resultados). Sí podéis usar funciones como *qnorm*, *pnorm*, *qt*, *pt* para consultar los valores de las distribuciones normal y t-Student.

6.4 Conclusión

A partir de los valores obtenidos, responded la pregunta formulada.

7 Resumen ejecutivo

Realizad un resumen ejecutivo de las conclusiones observadas en este análisis. Para contestar este apartado, podéis responder las preguntas formuladas de manera concisa y bien argumentada. Por ejemplo, si afirmáis que las ventas en USA son inferiores a las ventas fuera de USA, es bueno acompañar esta afirmación con el nivel de confianza y/o el valor p.

8 Puntuación de la actividad

- Apartado 1 (10%)
- Apartado 2 (10%)
- Apartado 3 (20%)
- Apartado 4 (10%)
- Apartado 5 (10%)
- Apartado 6 (20%)
- Apartado 7 (10%)
- Calidad del informe dinámico (calidad del código, formato y estructura del documento, concisión y precisión en las respuestas) (10%)