A2 - Analítica descriptiva e inferencial

Enunciado

Semestre 2021.1

Índex

1 Lectura del fichero y preparación de los datos **4**

2 Coste de los siniestros **4**
2.1 Análisis visual . . . 4
2.2 Comprobación de normalidad . . . 4
2.3 Intervalo de confianza de la media poblacional de la variable `UltCost` . . . 4

3 Coste inicial y final de los siniestros **4**
3.1 Justificación del test a aplicar . . . 5
3.2 Escribid la hipótesis nula y la alternativa . . . 5
3.3 Cálculos . . . 5
3.4 Conclusión . . . 5
3.5 Comprobación . . . 5

4 Diferencia de salario según género **5**
4.1 Análisis visual . . . 5
4.2 Interpretación . . . 5
4.3 Escribid la hipótesis nula y la alternativa . . . 5
4.4 Justificación del test a aplicar . . . 6
4.5 Cálculos . . . 6
4.6 Conclusión . . . 6
4.7 Comprobación . . . 6

5 Salario semanal (II) **6**
5.1 Escribid la hipótesis nula y la alternativa . . . 6
5.2 Justificación del test a aplicar . . . 6
5.3 Cálculos . . . 6
5.4 Conclusión . . . 6
5.5 Comprobación . . . 6

6 Diferencia de jornada según género **7**
6.1 Análisis visual . . . 7
6.2 Interpretación . . . 7
6.3 Hipótesis nula y alternativa . . . 7
6.4 Tipo de test . . . 7
6.5 Cálculos . . . 7
6.6 Conclusión . . . 7
6.7 Comprobación . . . 7

7 Salario por hora **7**
7.1 Hipótesis nula y alternativa . . . 7

7.2 Tipo de test 8
7.3 Cálculos 8
7.4 Conclusión 8
7.5 Comprobación 8

8 Resumen ejecutivo **8**

Introducción

El conjunto de datos train_clean2.csv se inspira (ha sido modificado por motivos académicos) en la base de datos disponible en la plataforma Kaggle: https://www.kaggle.com/c/actuarial-loss-estimation.

Este conjunto de datos contiene información de una muestra de indemnizaciones otorgadas por una compañía de seguros por el tiempo que ha estado de baja laboral el trabajador.

Las principales variables que se usarán en esta actividad son:

- ClaimNumber: Identificador de la póliza.
- DateTimeOfAccident: Fecha del accidente.
- DateReported: Fecha que se comunica a la compañía y ésta abre un expediente del siniestro.
- Age: Edad del trabajador.
- Gender: Sexo.
- MaritalStatus: Estado civil, (M)arried, (S)ingle, (U)nknown.
- DependentChildren: Número de hijos dependientes.
- DependentsOther: Número de dependientes excluyendo hijos
- WeeklyWages: Salario semanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWeek: Número horas por semana.
- DaysWeek: Número de días por semana.
- ClaimDescription: Descripción siniestros.
- IniCost: Estimación inicial del coste realizado por la compañía.
- UltCost: Coste total pagado por siniestro.
- Time: Tiempo desde que se apertura a cierra el siniestro.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, análisis descriptivo e inferencia estadística.

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 `train_clean2.csv` y guardad los datos en un objeto con identificador denominado *claim*. A continuación, verificad que los datos se han cargado correctamente.

---

2 Coste de los siniestros

La compañía de seguros está interesada en investigar los valores que toma la variable coste de los siniestros en la población. Para ello, realizad un primer análisis visual de esta variable (UltCost) a partir de la muestra. Posteriormente, realizad un análisis de normalidad y calculad el intervalo de confianza de la variable UltCost de los siniestros. Seguid los pasos que se indican a continuación.

2.1 Análisis visual

1. Mostrad con un diagrama de caja la distribución de la variable 'UltCost'.
2. Transformad la variable 'UltCost' a escala logarítmica y mostrad el diagrama de caja.
3. Interpretad los gráficos brevemente.

2.2 Comprobación de normalidad

¿Podemos asumir que la variable `UltCost` tiene una distribución normal? Debéis justificar la respuesta a partir de métodos visuales y contrastes.

- Realizad inspección visual de normalidad en base a los gráficos que consideréis oportunos.
- Realizad contraste de normalidad de Lilliefors (p.ej. con función lillie.test de la libreria nortest).
- Realizad inspección visual y contraste de normalidad a la variable UltCost en escala logarítmica.

2.3 Intervalo de confianza de la media poblacional de la variable `UltCost`

- Calculad manualmente el intervalo de confianza al 95% de la media poblacional de la variable 'UltCost' en escala original (No se pueden utilizar funciones como t.test o z.test para el cálculo). Sí se pueden usar funciones como 'mean', 'sd', 'qnorm', 'pnorm', 'qt' y 'pt'.
- ¿Podemos asumir la hipótesis de normalidad para el cálculo del intervalo de confianza sobre la media muestral del coste en escala original? Argumentar la respuesta.
- A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

---

3 Coste inicial y final de los siniestros

La compañía de seguros está interesada en investigar si la estimación inicial del coste que hace de los siniestros (IniCost) en promedio es suficiente para cubrir el coste total pagado (UltCost). Para eso, nos plantean la pregunta siguiente:

> ¿Podemos aceptar que no hay diferencias entre IniCost y UltCost?

Responded a la pregunta utilizando un nivel de confianza del 95%.

Seguid los pasos que se detallan a continuación.

3.1 Justificación del test a aplicar

Explicad qué tipo de contraste se puede aplicar en este caso. Es decir, explicad si se trata de un contraste de una muestra o dos muestras, sobre la media/varianza/proporción, si es bilateral o unilateral, etcétera.

3.2 Escribid la hipótesis nula y la alternativa

3.3 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Estos cálculos deben ser acordes con el método (contraste) elegido.

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 pueden usar funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`.

3.4 Conclusión

A partir de los valores obtenidos, explicad si podemos aceptar o rechazar la hipótesis planteada. También debéis responder la pregunta de investigación formulada.

3.5 Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R t.test.

---

4 Diferencia de salario según género

Existe una opinión generalizada que las mujeres cobran menos que los hombres. Vamos a comprobar qué dicen los datos al respecto. Nos preguntamos si las mujeres reciben un menor salario (**WeeklyWages**) que los hombres.

Para ello, debéis obtener dos muestras. La primera muestra contiene todas las mujeres (**Gender** igual a F). La segunda muestra contiene todos los hombres (**Gender** igual a M). Usad un nivel de confianza del 95%.

Seguid los pasos que se indican a continuación.

4.1 Análisis visual

En primer lugar, mostrad en un diagrama de caja la distribución de la variable `WeeklyWages` (en logaritmos) según el género.

4.2 Interpretación

Interpretad cualitativamente el gráfico mostrado en el apartado anterior, explicando si se pueden observar diferencias (visualmente) entre el salario de mujeres y hombres.

4.3 Escribid la hipótesis nula y la alternativa

Escribid las hipótesis nula y alternativa para la pregunta de investigación siguiente:

> ¿Podemos aceptar que los hombres cobran más que las mujeres en promedio a la semana?

Responded a la pregunta utilizando un nivel de confianza del 95%, usando la variable `WeeklyWages` en sus unidades originales (sin usar logaritmo).

Seguid los pasos que se detallan a continuación.

4.4 Justificación del test a aplicar

Explicad qué tipo de contraste se puede aplicar en este caso.

4.5 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%.

Nota: se deben realizar los cálculos manualmente. Para el cálculo del contraste, implementad una función que os permita utilizarla en los siguientes apartados.

4.6 Conclusión

A partir de los valores obtenidos, debéis concluir si podemos aceptar o rechazar la hipótesis. Asimismo, responded la pregunta formulada.

4.7 Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R t.test.

---

5 Salario semanal (II)

En este apartado, seguimos interesados en investigar si los hombres tienen un salario mayor a las mujeres, y si éste es mayor en una cantidad de 50 euros como mínimo. Por tanto, la pregunta que realizamos es:

> ¿Podemos aceptar que los hombres cobran al menos 50 euros más que las mujeres en promedio a la semana?

Seguid los pasos que se indican a continuación.

5.1 Escribid la hipótesis nula y la alternativa

5.2 Justificación del test a aplicar

Justificad qué tipo de contraste podemos aplicar en este caso.

5.3 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Se recomienda usar las funciones desarrolladas en apartados anteriores, si éstas son útiles para este contraste.

5.4 Conclusión

A partir de los valores obtenidos, responded si podemos aceptar o rechazar la hipótesis y responded la pregunta formulada.

5.5 Comprobación

Comprobar si los valores obtenidos coincicen con los de la función de R t.test.

---

6 Diferencia de jornada según género

Existe una opinión generalizada que las mujeres tienden a utilizar más la jornada a tiempo parcial. Vamos a comprobar qué dicen los datos al respecto.

Nos preguntamos si las mujeres realizan más frecuentemente una jornada a tiempo parcial **PartTimeFullTime** que los hombres.

6.1 Análisis visual

Mostrad un un diagrama de barras que muestre los porcentajes de cada categoría de la variable `PartTimeFullTime` según el género.

6.2 Interpretación

Interpretad los resultados y dad respuesta a la pregunta planteada.

6.3 Hipótesis nula y alternativa

La pregunta que realizamos sobre los datos es:

> ¿La proporción de personas que trabajan a tiempo completo es diferente para hombres que para mujeres?

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

6.4 Tipo de test

Indicad qué tipo de test aplicaréis y justificadlo.

6.5 Cálculos

Realizad todos los cálculos con instrucciones propias. Calculad el valor observado, el valor crítico y el valor p. Mostrad los resultados.

6.6 Conclusión

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

6.7 Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R prop.test.

---

7 Salario por hora

Anteriormente hemos comparado el salario semanal entre hombres y mujeres. Ahora bien, la pregunta que nos hacemos ahora es:

> ¿Podemos afirmar que los hombres cobran más que las mujeres por hora trabajada?

7.1 Hipótesis nula y alternativa

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

7.2 Tipo de test

Indicad qué tipo de test aplicaréis y justificadlo.

7.3 Cálculos

Calculad el estadístico de contraste, el valor crítico y el valor p con un nivel de confianza del 95%. Para realizar estos cálculos, usad la función que habéis implementado previamente.

7.4 Conclusión

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

7.5 Comprobación

Comprobar si los valores obtenidos coincicen con los de la función de R t.test.

---

8 Resumen ejecutivo

Resumid las conclusiones principales del análisis. Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Puntuación de la actividad

- Apartados 1 y 2 (10%)
- Apartado 3 (10%)
- Apartado 4 (20%)
- Apartado 5 (10%)
- Apartado 6 (20%)
- Apartado 7 (10%)
- Apartado 8 (10%)
- Calidad del informe dinámico (10%)