Tarea 4 - Titanik

Sebastián Martínez, Juan Darracq, Diego Handalian, Francisco Cabarcos

PROBABILIDAD Y ESTADÍSTICA APLICADA

07/07/2024

Introducción

El objetivo de este trabajo es procesar y analizar datos recreados de los pasajeros del Titanik para encontrar evidencia estadística relevante sobre las personas que estuvieron a bordo durante el desastre. Esto incluye tanto a aquellos que sobrevivieron como a los que no tuvieron la misma suerte. Los datos serán proporcionados en un archivo titanik.csv, el cual deberá ser procesado utilizando Python.

En la primera parte, se obtendrán los datos del archivo titanik.csv y se cargarán en un dataframe en Python. Se corregirán los valores vacíos de la columna "age" utilizando la media de las edades según el género de los pasajeros, justificando por qué esta es una estrategia razonable para la corrección de datos faltantes. Se calcularán la media, mediana, moda, rango, varianza y desviación estándar de las edades. También se determinará la tasa de supervivencia general y la tasa de supervivencia por género. Se creará un histograma de las edades de los pasajeros por clase (primera, segunda y tercera) y se propondrá un modelo para la distribución de la variable edad en el barco. Además, se realizarán diagramas de cajas para las edades de los supervivientes y los no supervivientes.

En la segunda parte, se construirá un intervalo de confianza del 95 % para la edad promedio de las personas en el barco. Se determinará si, con una certeza del 95 %, es posible afirmar que el promedio de edad de las mujeres interesadas en abordar el Titanik es mayor a 56 años, y se realizará el mismo análisis para los hombres. Con una certeza del 99 %, se evaluará si existe una diferencia significativa en la tasa de supervivencia entre hombres y mujeres, así como entre las distintas clases. Finalmente, con una certeza del 95 %, se analizará si en promedio las mujeres eran más jóvenes que los hombres en el barco.

Marco Teórico

Población: La población es el conjunto completo de individuos, elementos, o datos que poseen una característica común que se está estudiando. Por ejemplo, si estamos interesados en la edad promedio de los estudiantes de una materia, la población sería todos los estudiantes de esa materia.

Muestra: Una muestra es un subconjunto de la población que se selecciona para realizar el estudio. Las muestras deben ser representativas de la población para que las inferencias hechas a partir de la muestra sean válidas para toda la población.

Media Muestral :

La media muestral es la suma de todas las observaciones dividida por el número total de observaciones en la muestra.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Varianza Muestral: La varianza muestral es una medida de la dispersión de los datos alrededor de la media muestral.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

ESTIMACIÓN POR INTERVALO: Un estimador por intervalo de un parámetro poblacional es un intervalo aleatorio para el que se predice contendrá al verdadero valor del parámetro con una determinada probabilidad.

Estimación por intervalo para la media poblacional μ *de una población normal con varianza conocida* : $\overline{X} \pm Z_{1-\frac{\sigma}{2}} \frac{\sigma}{\sqrt{n}}$

INTERVALO DE CONFIANZA: Un intervalo de confianza para un parámetro poblacional resulta de evaluar el estimador por intervalo de dicho parámetro en la muestra observada. La confianza que se da al intervalo es la probabilidad de que el estimador por intervalo contenga el verdadero valor del parámetro.

ttest_ind

El método ttest_ind calcula la prueba T para las medias de dos muestras independientes de puntajes.

Desarrollo

Parte 1: Estadística Descriptiva

1) Para obtener los datos del CSV utilizamos la librería pandas de Python y su método read_csv.

```
datos = pd.read_csv('titanik.csv')
datos.head()
```

De esa manera podemos operar sobre los distintos datos del archivo.

2) Para corregir los valores vacíos de la columna “age” primero realizamos la media según el género utilizando la función mean().

```
media_edad_femenina = datos[datos['gender'] == 'female']['age'].mean()
media_edad_masculina = datos[datos['gender'] == 'male']['age'].mean()
datos.loc[(datos['age'].isnull())  &  (datos['gender']  ==  'female'),
'age'] = media_edad_femenina
datos.loc[(datos['age'].isnull()) & (datos['gender'] == 'male'), 'age']
= media_edad_masculina
```

Esto es una estrategia razonable porque la media representa el valor promedio de las edades para cada género, habiendo una estimación balanceada para los datos faltantes.

3) Para calcular la media, mediana, moda, rango, varianza y desviación estándar de las edades utilizamos las funciones mean(),sum(), median(), mode()m, el valor máximo sobre el mínimo para el rango, var() para la varianza y std() para la desviación estándar. Los valores resultantes fueron los siguientes:
Promedio de edad: 57.529133940454855
Mediana de edad: 57.75422001515064
Moda de edad: 58.454742975932184
Rango de edad: 98.86854607597618
Varianza de edad: 580.9367918400833
Desviación estándar de edad: 24.10263039255432.

4) la tasa de supervivencia la calculamos realizando el promedio de la columna 'survived', que indica si un pasajero sobrevivió (1) o no (0)

```
tasa_supervivencia_general = datos['survived'].mean()
```

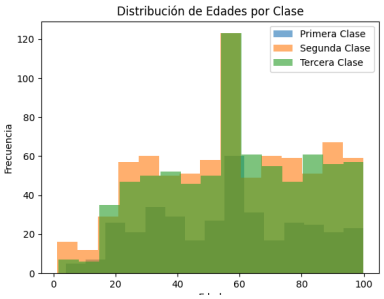
El resultado obtenido fue: 0.40093603744149764.

5) La tasa de supervivencia por género la calculamos realizando el promedio de si sobrevivió o no tal; que su género sea femenino o masculino. Los valores obtenidos fueron los siguientes:

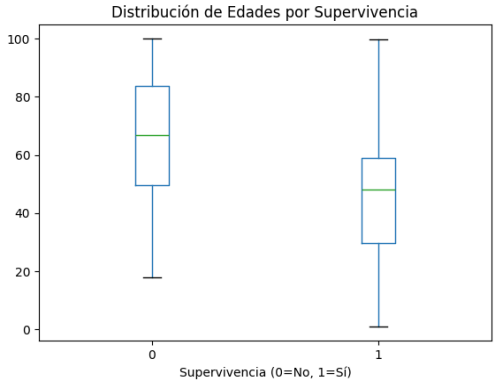
Tasa supervivencia femenina: 0.47789473684210526

Tasa supervivencia masculina: 0.32579650565262075

6) Para visualizar la distribución de edades de los pasajeros por clase (primera, segunda y tercera) en Python realizamos el histograma:



7) Utilizando python realizamos un diagrama de cajas para las edades de los supervivientes, y otro para las edades de los no supervivientes. Del diagrama podemos concluir que la mediana de edad de los pasajeros que no sobrevivieron es mayor que la mediana de edad de los pasajeros que sobrevivieron. Por lo que, en promedio, los pasajeros que no sobrevivieron eran mayores que los que sobrevivieron.



Parte 2: Inferencia Estadística

1)

Para calcular el intervalo de confianza, con confianza 95%, utilizaremos la fórmula $\overline{X} \pm Z_{1-\frac{\sigma}{2}} \frac{\sigma}{\sqrt{n}}$. Así se ve la fórmula en Python para el intervalo de confianza mayor y menor

```
intervalo_confianza_mayor = (
    promedio_edad + (1.96 * (desviacion_edad / raiz_cantidad_personas)))

intervalo_confianza_menor = (
    promedio_edad - (1.96 * (desviacion_edad / raiz_cantidad_personas)))
```

el resultado obtenido es el siguiente: {56.451848301856174, 58.60641957905354}

2)

A partir de los datos de la muestra, con una certeza del 95%.Para este muestra utilizamos una media muestral de 56.5811 para las mujeres , una desviación estándar de 0.7857.

Para las mujeres tenemos un intervalo de 55 y 58 . Entonces con estos datos podemos afirmar con una certeza del 95% que el promedio de edad de las mujeres interesadas en abordar el Titanik es mayor a 56 años.

También, para los hombres tenemos una media muestral igual a 58.4547 años. Entonces con estos datos podemos afirmar con una certeza del 95% que el promedio de edad de las hombres interesadas en abordar el Titanik es mayor a 56 años.

3). A partir de los datos de la muestra, con una certeza del 99 %. Podemos decir que existe una diferencia significativa en la tasa de supervivencia entre hombres y mujeres.

```
t_statistic, valor = ttest_ind(datos[datos['gender'] == 'male']['survived'], datos[datos['gender'] == 'female']['survived'])

if valor < 0.01:
    print("Existe una diferencia significativa en la tasa de supervivencia entre hombres y mujeres.")
else:
    print("No existe una diferencia significativa en la tasa de supervivencia entre hombres y mujeres.")
```

También , a partir de los datos de la muestra, con una certeza del 99 % podemos observar que no existe una diferencia significativa en la tasa de supervivencia en las distintas clases.

4) Mediante las pruebas realizadas en Python se obtuvo un valor p igual a 0.08832796719448419 por lo que no se puede afirmar con una certeza del 95% que en promedio las mujeres eran más jóvenes que los hombres en el barco.

```
resultado_ttest_edad_genero = stats.ttest_ind(datos[datos['gender'] ==
'female']['age'], datos[datos['gender'] == 'male']['age'])
p_valor = resultado_ttest_edad_genero.pvalue
```

Bibliografía

Variables Aleatorias Discretas 1-Facultad de Ingeniería Universidad Católica del Uruguay

Variables Aleatorias Discretas 2-Facultad de Ingeniería Universidad Católica del Uruguay

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html