Universidad de Costa Rica

Escuela de Ciencias de la Computación e Informática

Sigla y Curso	**Probabilidad y Estadística CI-0115**	**Grupo**	**2**
Título del Trabajo	**Laboratorio #3**		
Nombre del Equipo	**Los Compas**		
Fecha de Realización	**22/11/2022**	**Fecha Entrega**	**25/11/2022**
Profesor(a)	**Luis Diego Mora Jiménez**	**Asistente**	**Nayeri Azofeifa Porras**
		Tardía:	(_____)
Estudiante:	**Marcelo Delgado Mora**	**Nota:**	_____ **/100**
Estudiante:	**Jose Antonio Mora Monge**		
Estudiante:	**Jean Paul Chacón Gonzáles**		
Estudiante:			

P1: Mencione brevemente ¿cuál es una de las desventajas de la media como estadística descriptiva?

R/ La desventaja principal de la media es que esta es afectada de manera desproporcionada por sus extremos.

P2: Interprete: ¿qué representa la media de la tarifa (*Fare*) del Titanic?

R/ La media de la tarifa, representa lo que se paga aproximadamente por pasajero lo cual sería $32.20

P3: Brevemente, comente algunas características de la moda.

R/ Sus valores pueden ser de tipo continuo o discreto, y cualitativos o cuantitativos, además puede que no exista o puede haber múltiples modas en el caso en que varios datos cuenten con la misma frecuencia.

P4: Interprete los resultados de la moda en el contexto de la edad de pasajeros del Titanic.

R/ La edad más común en el barco era de 24 años por lo que se observa que en el barco se contaba con grupos de pasajeros de edad muy joven

P5: Calcule la mediana de *Fare*.

R/ El cálculo de la mediana nos lo da summary y será 14.45

P6: Interprete: ¿qué significa el valor de la mediana para el valor de tarifa o *Fare*?

R/ El valor de la mediana indica que en general que la gente en el barco pago ya sea menos de 14 o más de esto siendo el centro de dividiendo las tarifas en dos mitades.

P7: Calcule el rango de los valores de la tarifa (*Fare*) para el conjunto Titanic

R/ El rango del conjunto de datos nos lo da la diferencia entre el min y el max, en el caso de la tarifa se trata de 512.32 - 0 lo que nos indica que el rango es 512.32

P8: Interprete los resultados del *boxplot*, tome en cuenta que debe referirse a lo que le indica la mediana de cada grupo que comparan los *boxplots*. ¿Qué información le están brindando? Además, ¿qué representan los puntos afuera de las barras de error?

R/ En el boxplot, a la hora de analizar las medianas de cada grupo, se puede notar cómo entre más alta sea la clase,(clase 1 siendo la mayor), más alta es la mediana de edades del grupo, además de que, la mediana corresponde aproximadamente al centro de las cajas. Por otro lado, los puntos fuera de las barras de error representan los valores que se alejan del rango normal llamados valores atípicos.

P9: Calcule la varianza y la desviación estándar de *Fare* (tarifa) para el conjunto Titanic.

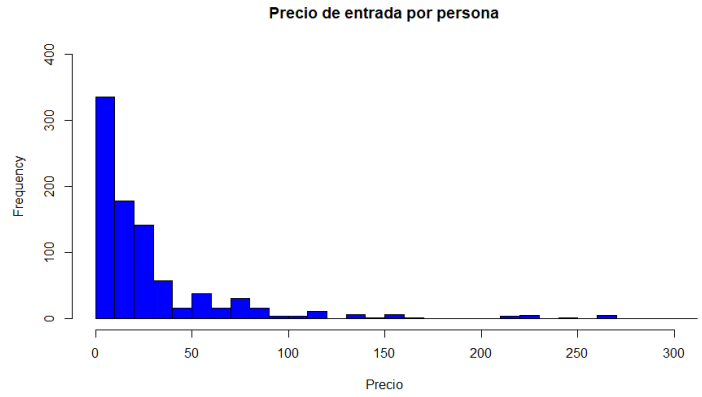
R/ Utilizando las funciones dadas por R básico, podemos obtener la varianza utilizando la función var() la cual nos da la varianza simple, lo que nos da 2469.44 , por otro lado la desviación estándar la obtenemos utilizando la función sd() la cual nos da 49.69.

P10: Calcule la mediana, media, varianza y la desviación estándar de *Age* (edad) para el conjunto Titanic. (NOTA: *Age* tiene valores faltantes).

R/ Podemos utilizar las funciones de R básico mean(), median(), var(), sd() para obtener la media, mediana, varianza y desviación estándar respectivamente de la edad de los pasajeros del Titanic. De esto obtenemos que la mediana sería 28, la media 29.70, la varianza 211.02 y la desviación estándar 14.53.

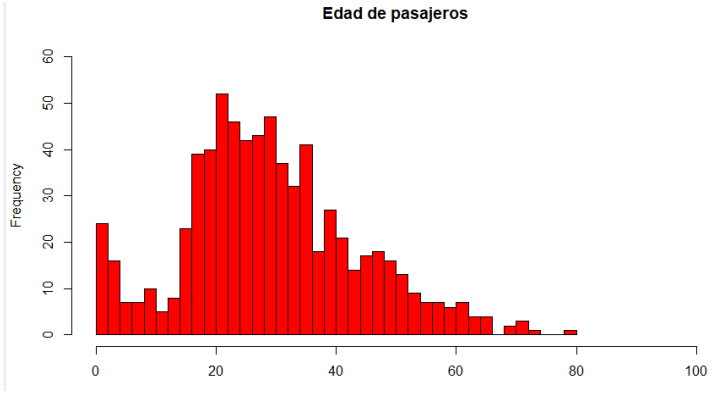
P11: Realice un análisis de distribución. Con base en su conocimiento realice un análisis gráfico que le permita presentar la distribución de los datos de al menos 2 variables del conjunto Titanic (usted elige) y otro análisis gráfico que le facilite la identificación de la distribución normal de los datos que usted visualizó.

Figura 1. Distribución de Precios de entrada por pasajero en la población total del Titanic



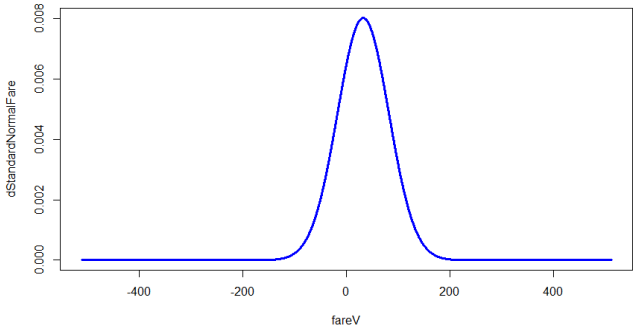
Fuente: Producción Propia

Figura 2. Distribución de edad de pasajeros en la población total del Titanic.



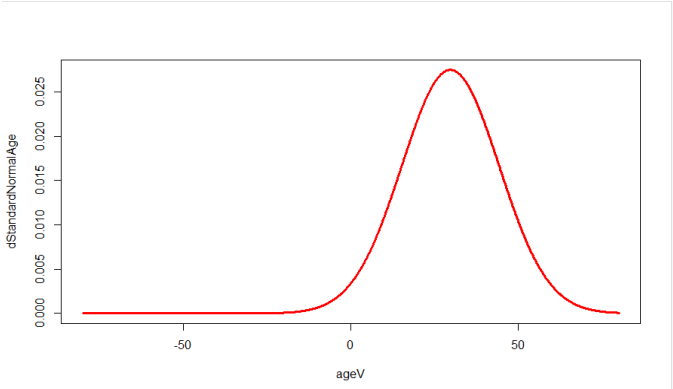
Fuente: Producción Propia

Figura 3. Distribución normal de los datos correspondientes al Precio de entrada por pasajero en la población total del Titanic



Fuente: Producción Propia

Figura 4. Distribución normal de los datos correspondientes a la edad de pasajeros en la población total del Titanic



Fuente: Producción Propia

P12: Modifique la función o construya su propia función para que pueda reportar el valor *p* (*p-value*) al ejecutar una prueba de hipótesis de una o dos colas. Recuerde que es recomendable que incluya una condición que no permita el cálculo de una prueba para más de dos colas, dado que esto no es factible.

R/ Función que calcula p-value dependiendo del valor de Z, que incluye como 4to parámetro una variable para indicar si se quiere la prueba de 1 cola o de 2 colas. Contiene un condicional que si se pone en la variable de la cola un valor mayor a 2, se va a hacer una prueba de 2 colas igualmente. En el caso del ejemplo realizado, el p-value en la prueba de 1 cola a la derecha da 5.689134e-14, y en la prueba de 2 colas da 1.137827e-13, ambos valores muy por debajo del nivel del nivel de significancia del 95%. (ver código).

P13: Una vez realizada la pregunta anterior, interprete los resultados obtenidos, en términos del contexto del análisis que se está realizando. Recuerde que los valores son solamente números a los que debe darle contexto o interpretación con base en una hipótesis.

R/ Debido a que se presenta un valor alto en Z, esto significa que el p-value será muy bajo, y este al ser muy bajo (menor a $\alpha = 0.05$ si se hace un estudio con base al 95% de confianza, el más común o estándar), se rechaza la hipótesis nula y por lo tanto, contextualizando con los datos del Titanic, se puede argumentar que sí parece haber una relación significativa entre las posibilidades de supervivencia y la clase a la que pertenecen los pasajeros.

P14: Interprete los resultados del Chi-cuadrado y explique en el contexto del conjunto de datos del Titanic ¿qué significan estos resultados con respecto a la relación estadística que existe entre las variables?

R/ En ambos casos, al analizar el p-value que devuelven ambas funciones, se puede notar que ambos son menores a 0.05 por mucho, ya que en el caso del sexo, el p-value < 2.2e-16, y en el caso de la clase, el p-value = 4.549e-23, por lo que se puede inferir que la relación entre las variables con la supervivencia es significativa con un 95% de confianza, o relacionándolo con el conuunto de datos, se puede observar que existe una relación significativa entre la superviviencia y el sexo, y la supervivencia y la clase a la que pertenecía el pasajero.

P15: A partir del diagrama de dispersión, ¿se puede observar algún patrón de asociación de las variables?

R/ Debido a que el gráfico no presenta una relación lineal o no forma una línea o patrón evidente positivo o negativo entre sus variables, con la información presentada en el gráfico no parece que exista una correlación entre las variables de precio de entrada y edad.

P16: Interprete: Con base en los resultados de este análisis de correlación, interprete el valor de correlación, y lo que significa en el contexto de los datos del Titanic.

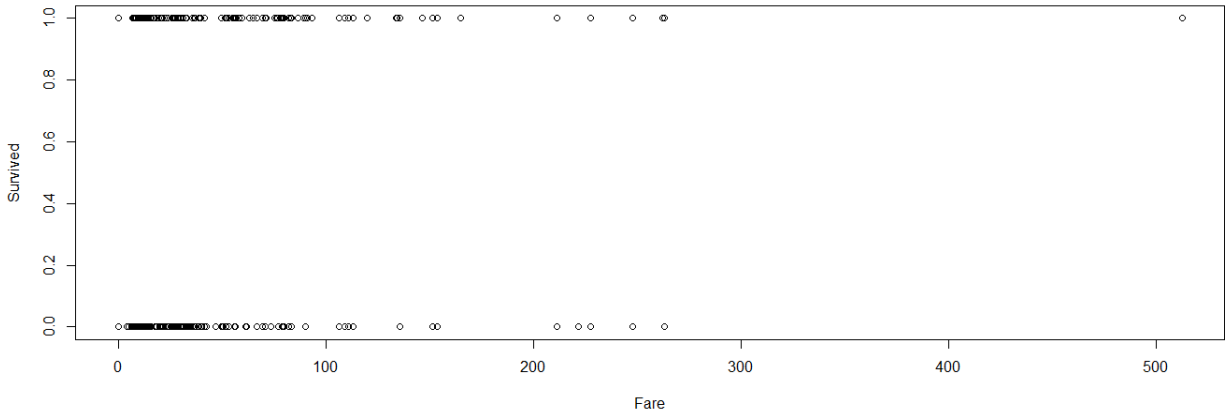
R/ El cálculo de correlación utilizando el método de Pearson confirma las inferencias realizadas con el análisis visual gráfico hecho en la pregunta anterior, ya que la función demuestra que la correlación positiva entre las variables es muy baja, de un 9.6% aproximadamente. Contextualizando estos datos al conjunto de datos del Titanic, se puede inferir que no existe una asociación o relación significativa entre la edad y el precio pagado por tiquete por los pasajeros.

P17: Experimente: Realice al menos 2 pruebas de correlación entre variables del conjunto Titanic e interprete los valores de correlación en el contexto del conjunto de datos del Titanic.

R/ Realizando un cálculo de correlación utilizando el método de Pearson entre las variables supervivencia y precio pagado por tiquete, se puede observar que existe una correlación positiva débil entre las variables, de 26% aproximadamente. Contextualizando los datos a los datos del Titanic, se puede inferir que sí existe una asociación o relación entre el precio pagado por el tiquete y la supervivencia de los

pasajeros, aunque esta es débil y dependiendo del contexto puede resultar no lo suficientemente significativa.

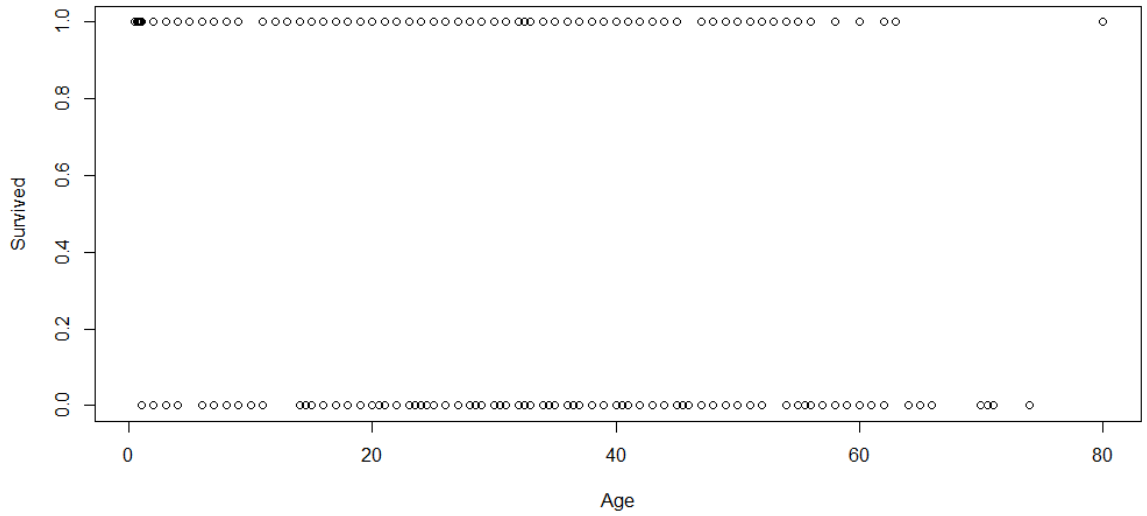
Figura 5. Correlación entre precio de tiquete y supervivencia de los pasajeros del Titanic



Fuente: Producción Propia

Nuevamente, al realizar un cálculo de correlación utilizando el método de Pearson entre la supervivencia y la edad de los pasajeros, se puede notar que existe una correlación negativa muy baja entre las variables, de 7.7% aproximadamente. Contextualizando los datos con el estudio del Titanic, se puede inferir que la asociación o relación entre la edad y la supervivencia de los pasajeros es casi inexistente, o muy poco significativa.

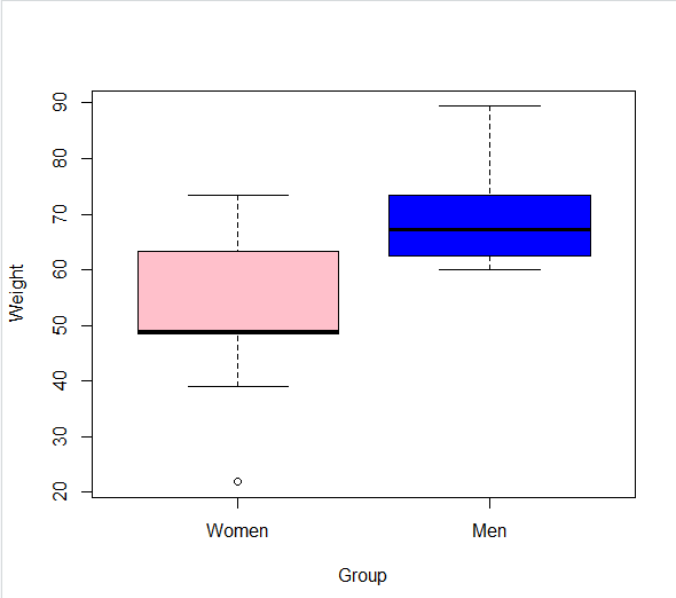
Figura 6. Correlación entre edad y supervivencia de los pasajeros del Titanic



Fuente: Producción Propia

P18: Grafique los datos de peso por grupo (hombre/mujer) en un *boxplot*. Puede utilizar R básico o puede utilizar *ggplot2*. Recuerde que es preferible que coloree por grupo para poder identificar cada uno.

Figura 7. Pesos de hombre y mujer de los pasajeros del Titanic



Fuente: Producción Propia

P19: Identifique la mediana de cada grupo (hombre y mujer).

R/ Mediante la función mean() que puede ser utilizada a través del lenguaje R se calcula que la mediana del grupo de los hombres contiene un valor de 67.3, por otro lado se concluye de la misma manera que la mediana del grupo de las mujeres es de 48.8.

P20: Interprete los resultados de la prueba *F*. Con base en estos, ¿qué se puede inferir de las varianzas? ¿Qué implicaciones tiene esto con la prueba de comparación de medias a realizar?

R/ El valor resultante de p al ejecutar la prueba F propuesta en el enunciado es de 0.17, mediante ese valor se puede deducir que el valor de p es mayor al valor de la significancia, lo que nos ayuda a concluir que las varianzas no tienen diferencia entre sí.

P21: Interprete los resultados: De acuerdo con el valor de *t* y el valor *p* obtenido, ¿qué conclusiones puede extraer usted de la prueba de dos colas realizada a los datos?

R/ Al ejecutar ambos métodos se puede evidenciar que ambos métodos producen el mismo resultado, el resultado del valor p que producen ambos métodos es 0.01327, que es menor al valor de significancia, por lo tanto se rechaza H0 y se puede concluir que el peso promedio de las mujeres es diferente al de los hombres.

P22: ¿Puede usted concluir que alguno de los grupos es mayor o menor? Justifique su respuesta.

R/ La media del grupo de los hombres tiene un valor de 68.9, mientras que la media del grupo de las mujeres tiene un valor de 52.1, por lo tanto el grupo de los hombres es mayor.

P23: Calcule los resultados de la prueba de una cola usando el parámetro "*greater*" para el argumento *alternative*. Interprete los resultados de:

R/

a. **La prueba de una cola one.tail.test con el parámetro "less".**
   El valor p resultante de utilizar el parámetro "less" es 0.99, mucho mayor que 0.05, esto significa que se puede decir que el peso de los hombres no es menor que el de las mujeres.
b. **La prueba de una cola one.tail.test con el parámetro "greater".**
   El valor p es 0.006, lo cual es menor que 0.05, por lo tanto se rechaza la hipótesis nula y si se puede decir que el promedio de peso de los hombres es mayor que el promedio de peso de las mujeres.