Trabajo Final Inferencia Estadística

Autor: Joan Benlloch Garcia

Profesora: Mar Angulo Martinez

24-05-2024

1. Introducción/Resumen

Este trabajo consiste en el análisis estadístico de una base de datos a libre elección. Durante este proyecto usaremos el Dataset que nos han facilitado en clase, **titanic3**, sobre los datos de los pasajeros del Titanic. Este conjunto de datos describe el estado de supervivencia de **1045** de los **pasajeros** del Titanic, sin contar miembros de la tripulación.

Este conjunto de datos tiene las siguientes variables, las cuales estaremos analizando durante el proyecto:

Pclass: Representa la clase en la que viajaba el pasajero. Toma valores 1, 2 o 3 representado respectivamente primera clase, segunda clase y tercera clase.

Survival: toma valores de 0 si el pasajero no sobrevivió, 1 si lo hizo.

Name: Nombre completo del pasajero

Sex: Sexo del pasajero

Age: Edad del pasajero

Sibsp: Número de hermanos o pareja del psajero a bordo

Parch: Número de padres o hijos del pasajero a bordo

Ticket: Número del ticket del pasajero

Fare: Tarifa del pasajero, en libras británicas de antes del 1970

Cabin: Cabina del pasajero

Embarked: Puerto en el que el pasajero embarcó. Toma valores de “C” para Charbourg, “Q” para Queenstown y “S” para Southampton.

Boat: Número del bote salvavidas

Body: Número identificador del cuerpo

Home.dest: Población de residencia y destino al que se dirigían.

Este conjunto de datos está preparado para hacer regresiones logísticas, pero nosotros lo usaremos para todo el proceso de análisis estadístico que nos han pedido, extrayendo conclusiones en la medida de lo posible y trabajando con los datos de forma adecuada a los problemas que plantearemos y resolveremos.

Como cabe esperar, las variables más interesantes y que más juego dan para este trabajo son las numéricas. Especialmente la edad del pasajero y la clase en la que viajaba. Además del sexo ya que, como veremos en el modelo de regresión, tuvo una gran influencia en la supervivencia de los pasajeros. Muchos de los estudios serán, por supuesto, para determinar las variables influyentes en la supervivencia de lo pasajeros, aunque también incluimos estudios sobre el análisis de la población y otros estudios vistos durante el curso.

El análisis estadístico consiste en **tres partes**. En la primera parte nos centraremos en la Inferencia Paramétrica. Concretamente realizaremos diversos contrastes de hipótesis estudiando la muestra de pasajeros en cuanto a su edad y las proporciones de supervivientes. Admás incluimos en la primera parte diversos estudios sobre intervalos de confianza para las medias de edades y proporciones de pasajeros en las distintas clases, entre otros.

En la segunda parte estudiaremos los métodos no paramétricos, esta parte

2. Desarrollo

La parte del desarrollo consiste en el análisis en R y Python de el dataset titanic3. Hemos usado R para la inferencia paramétrica y no paramétrica y Python para los modelos regresivos.

Por supuesto no nos adentraremos en la preparación de los datos a no ser que sea necesario tener en cuenta alguna preparación especifica para poder realizar alguno de los estudios, en cuyo caso se mencionaría antes de hacer el análisis.

Dado que este dataset es una muestra de la población real del titanic sacaremos conclusiones sobre la población real desconocida a partir de los datos con los que contamos, teniendo en cuenta que algunos valores no están completos para todos los pasajeros y por tanto el tamaño de la muestra varía según el análisis que estamos realizando.

2.1. Inferéncia paramétrica

Empezamos la parte de desarrollo con el estudio de inferencia paramétrica.

Siguiendo el orden del temario visto en clase, el primer apartado del trabajo va dirigido al estudio de Intervalos de confianza en las variables de nuestros datos.

Como se ha comentado en la introducción trataremos en la mayoría del trabajo las variables de la edad y las clases a las que pertenecían los pasajeros.

El primer intervalo de confianza lo vamos a hacer sobre la media de los pasajeros. Queremos encontrar un intervalo de confianza del 95% para el valor medio de la edad de los pasajeros.

Encontramos el intervalo siguiente:

```
> IC <- c(media_edad-margen_error,media_edad+margen_error)
> IC
[1] 29.00767 30.75629
```

Con un 95% de seguridad podemos decir que la edad media se encuentra entre 29 y 30,75 años. Pero más interesante es estudiar la media de edades por clase ya que el tamaño de las muestras varía considerablemente. Esto es de esperar ya que en primera clase sule haber mucha menos gente y en tercera la mayoría. En nuestro caso, tenemos 284 pasajeros en primera, 261 en segunda y n total de 500 en tercera, las tres muestras grandes:

Primera:

```
> IC_primera
[1] 37.46794 40.85189
```

Segunda:

```
> IC_segunda
[1] 27.85208 31.16133
```

Tercera:

```
> IC_tercera
[1] 23.75892 25.85708
```

Estos datos son muy curiosos, ya que vemos un claro envejecimiento en primera clase respecto a tercera clase. Si lo pensamos en el contexto histórico es entendible, ya que para pertenecer a primera clase solo familias o personas adineradas se lo podían permitir y eran pasajeros que viajaban por motivos muy diversos, mientras que en tercera clase viajaban personas muy pobres, pagando lo mínimo posible que buscaban empezar una nueva vida en Estado Unidos desde jóvenes e iban a buscar trabajo.

Voy a hacer algunos intervalos de confianza para proporciones. Por ejemplo, con una confianza del 90% ¿Qué proporción de pasajeros podemos decir que es menor de edad? ¿Y con un 99%?

Para un 90%:

```
> IC
[1] 0.1029200 0.1321945
```

La proporción de menores de edad está entre el 10,29% y el 13,21%.

Para un 99% de confianza:

```
> IC
[1] 0.09463542 0.14047909
```

Conun 99% de confianza podemos decir que la verdadesra proporción de menores de edad entre los pasajeros del titánic era de entre el 9,46% y el 14,04% de los pasajeros.

A continuación, hacemos un intervalo de confianza del 95% para la proporción de pasajeros que sobrevivieron.

```
> IC <- c(p-margen_error,p+margen_error)
> IC
[1] 0.3788079 0.4384169
```

Este intervalo de confianza nos dice que al 95% de confianza podemos decir que sobrevivieron entre un 37,88% y un 43,84% de los pasajeros del titánic.

Realizamos ahora el intervalo para un 99% de confianza:

```
> IC <- c(p-margen_error,p+margen_error)
> IC
[1] 0.3694427 0.4477822
```

Para un 99% de confianza nos sale que el intervalo está entre el 36,94% y el 44,77%.

El siguiente intervalo de confianza realizado será para estimar con un 95% de seguridad la proporción de pasajeros que viajaban en primera clase. Pero esta vez vamos a considerar los datos de nuestra base de datos como si fuesen la población y extraer una

muestra de tan solo 100 pasajeros. Para hacer esto usamos las herramientas vistas en clase en R.

```
pasajeros.elegidos <- sample(1:n,100)
```
Run Chunk | Run Above
```{r}
#Veamos cuales son los pasajeros elegidos en la muestra
muestra.pasajeros = titanic_df[pasajeros.elegidos,]
```

Y a continuación realizamos tres intervalos de confianza distintos con la librería vista en clase, epitools:

```
> #Y vamos ya a calcular el intervalo de confianza exacto con "epitools"
> binom.exact(número.pasajeros.primera,100,conf.level=0.95)
   x   n proportion     lower     upper conf.level
1 30 100        0.3 0.2124064 0.3998147       0.95
> ###Intervalo de Confianza por el método de Wilson###
> binom.wilson(número.pasajeros.primera,100,conf.level=0.95)
   x   n proportion     lower     upper conf.level
1 30 100        0.3 0.2189489 0.3958485       0.95
> #Comprobamos cómo quedaría el IC utilizando el método de Laplace
> binom.approx(número.pasajeros.primera,100,conf.level=0.95)
   x   n proportion     lower     upper conf.level
1 30 100        0.3 0.2101832 0.3898168       0.95
```

Entramos ahora en la parte de contrastes de hipótesis.

Como inciso para este apartado diremos que la mayoría de los contrastes de hipótesis vistos en clase son para poblaciones normales. El problema que tenemos en este dataset es que ninguna de sus variables sigue una distribución normal. Esto lo veremos en más profundidad en el primer apartado de la sección de métodos no paramétricos en la que buscaremos algunas distribuciones que se ajustes a las distribuciones de nuestras variables en la muestra.

Por ejemplo mi primer contraste será para contrastar la edad media de pasajeros en primera y tercera clase. Esto lo puedo hacer ya que mi muestra es grande, más de 40 y por el TCL la diferencia de medias podemos decir que sigue una distribución normal.

Este contraste de hipótesis tiene como hipótesis nula decir que ambas medias son iguales. Este estudio tiene valor, o lo tendría en caso de estar haciendo inferencia en datos no estudiados antes, porque la edad media en las distintas clases habla sobre la sociedad del momento en Inglaterra.

```
> p.valor <- 2*pt(-abs(t),df=n.primera+n.tercera-2)
> p.valor
[1] 1.545265e-40
```

Dado que el p-valor es del orden de 10^-40, es decir, extremadamente pequeño, podemos afirmar que la diferencia de edad es muy significativa, la edad media en primera clase, es significativamente mayor.

Vemos cuánto.

Realizo un contraste de hipótesis diciendo que la diferencia de edad media entre primera y tercera clase es de 10 años.

```
> p.valor <- 2*pt(-abs(t),df=n.primera+n.tercera-2)
> p.valor
[1] 2.060033e-05
```

Vemos que el p-valor es muy pequeño, por tanto rechazamos H0 y no podemos decir que la diferencia de edad sea de 10 años entre primera y tercera clase.

Vemos si son 15 años.

```
> #Veamos si podemos aceptar que sean 15 años mayores en media la pblación de
> t_0 <- 15
> t <- (media.primera-media.tercera-t_0)/sqrt(std.primera^2/n.primera+std.terc
> t
[1] -0.6380378
> p.valor <- 2*pt(-abs(t),df=n.primera+n.tercera-2)
> p.valor
[1] 0.5236358
> #p-valor = 0.5236358 => aceptamos la hipótesis nula, la diferencia de edades
```

Ahora sí, obtenemos un p-valor de 0,5236 un valor que acepta H0, es decir podemos afirmar que la diferencia entre primera y tercera clase es de 15 años.

Otro contraste de hipótesis interesante es comparar si la clase influye en la probabilidad de sobrevivir. Esto lo podemos hacer haciendo un contraste de proporciones con las muestras de primera clase y tercera clase y ver si podemos decir que la proporción de supervivientes en primera es similar a la de los supervivientes en las otras clases.

Una vez hechos los contrastes vemos que tenemos que rechazar todas las hipótesis nulas. La clase a la que se pertenece influye significativamente en la proporción de supervivientes. Obtenemos p-valores de:

Contraste de primera con tercera:

```
> p.valor
[1] 2.428093e-27
```

Primera con segunda:

```
> p.valor
[1] 2.714983e-06
```

Y de segunda con tercera:

```
> p.valor
[1] 9.795336e-07
```

Para terminar el apartado de inferencia paramétrica vamos a plantear las siguientes preguntas.

¿Podemos decir que la probabilidad de sobrevivir fuese del 50%? ¿Y del 30%? ¿Y del 40%?

Para contestar estas preguntas realizamos un contrate de proporciones para nuestra muestra sobre los pasajeros que sobrevivieron.

Usando la función pbinom obtenemos los siguientes p-valores

Para un 50% de probabilidad de sobrevivir:

```
> p.valor
[1] 1.891052e-09
```

Rechazamos H0

Para un 30% se sobrevivir:

```
> p.valor
[1] 3.608225e-14
```

Todavía menor, rechazamos que la probabilidad de sobrevivir sea del 30%

Para el 40%

```
> p.valor
[1] 1.452287
```

Hay evidencia fuerte de que la probabilidad de sobrevivir esté en el 40%

2.2.Métodos no paramétricos

Antes de empezar el desarrollo de los métodos no paramétricos hacemos un inciso en los motivos por los que estamos haciendo un estudio sobre los datos de los pasajeros del titánic. La idea de hacer inferencia sobre un crucero de 1912 no es extraer conclusiones sobre todos los barcos o cruceros que hacen viajes, pero tal vez si que se puede hablar a grandes rasgos de los barcos de la misma compañía o de los barcos que viajaban de Inglaterra a Estados Unidos en 1912, lo que nos aporta datos de relevancia histórica según las características y motivos de los viajeros del momento.

Dicho esto me apreció interesante poder ver si alguna de las variables de la muestra de pasajeros de titanic3, nuestra base de datos, seguía una distribución conocida. Como ya hemos adelantado, ninguna de las variables sigue una distribución normal, pero a continuación exponemos el contraste de bondad de ajuste chi cuadrado para la muestra de edades de los pasajeros.

Para hacer este contraste primeo debemos ordenar las edades, luego crear unos intervlos para las edades, yo he creado intervalos de edades de 5 en 5 años. Es decir empezando por el 0 hasta los 80 años que es la persona más longeva del dataset, tenemos 15 intervalos. A continuación, generamos la tabla de frecuencias. Y como podemos ver en el

gráfico de abajo, todas las frecuencias suman un mínimo de 5 individuos, requisito necesario para realizar un contraste de bondad de ajuste chi cuadrado.

```
> print(tabla_frecuencia)
intervalos_edades
   [0,5)   [5,10)  [10,15)  [15,20)  [20,25)  [25,30)  [30,35)  [35,40)
      51       31       27      116      184      160      132      100
 [40,45)  [45,50)  [50,55)  [55,60)  [60,65)  [65,70) [70,Inf)
      69       66       43       27       27        5        8
```

Según lo visto en clase generamos la lista de frecuencias teóricas:

```
> round(frecuencias.teóricas,2)
 [1]  24.60  44.51  71.48 101.91 128.96 144.87 144.46 127.87 100.47  70.08
[11]  43.39  23.85  11.63   5.04   2.87
```

Y para terminar hacemos el test de chi cuadrado para ver si nuestra distribución de frecuencias de edades sigue una distribución normal.

```
> chisq.test(frecuencias.empíricas,p=probabilidades.teóricas)

        Chi-squared test for given probabilities

data:  frecuencias.empíricas
X-squared = 134.27, df = 14, p-value < 2.2e-16
```

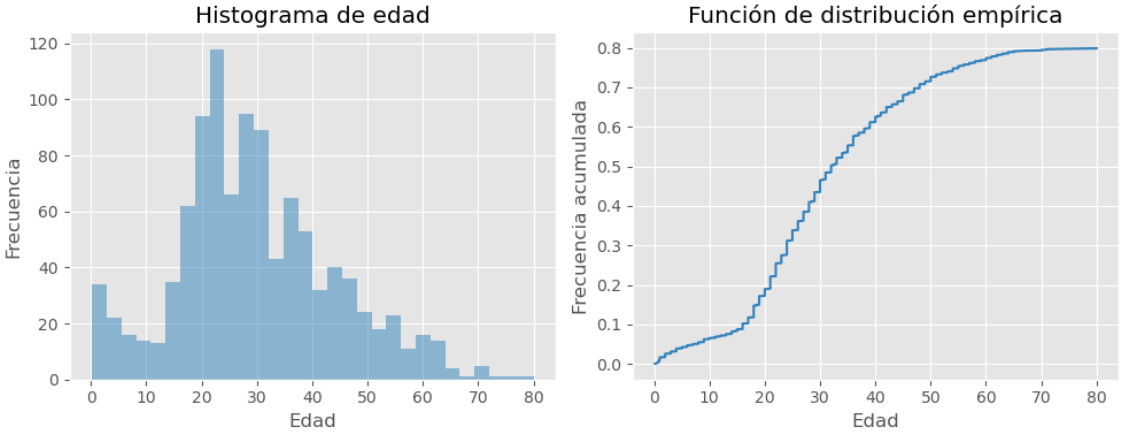
Como podemos ver el p-valor nos sale del orden de 10^(-10), un valor muy pequeño, por tanto rechazamos que nuestra distribución siga una distribución normal.

Dado que hemos visto que nuestra distribución no sigue una normal vamos a buscar a qué distribución se le parece. Para esto vamos a aplicar métodos gráficos vistos en clase. El código de esta sección está en el notebook de Python CódigoPracticaFinal.ipynb.

Antes que nada vamos a dejar una visualización del método describe de los dataframes de Python que aporta mucha información sobre cada una de las variables. Aunque no entraremos a describir cada punto porque no es el objetivo del apartado creo que es interesante verlo.

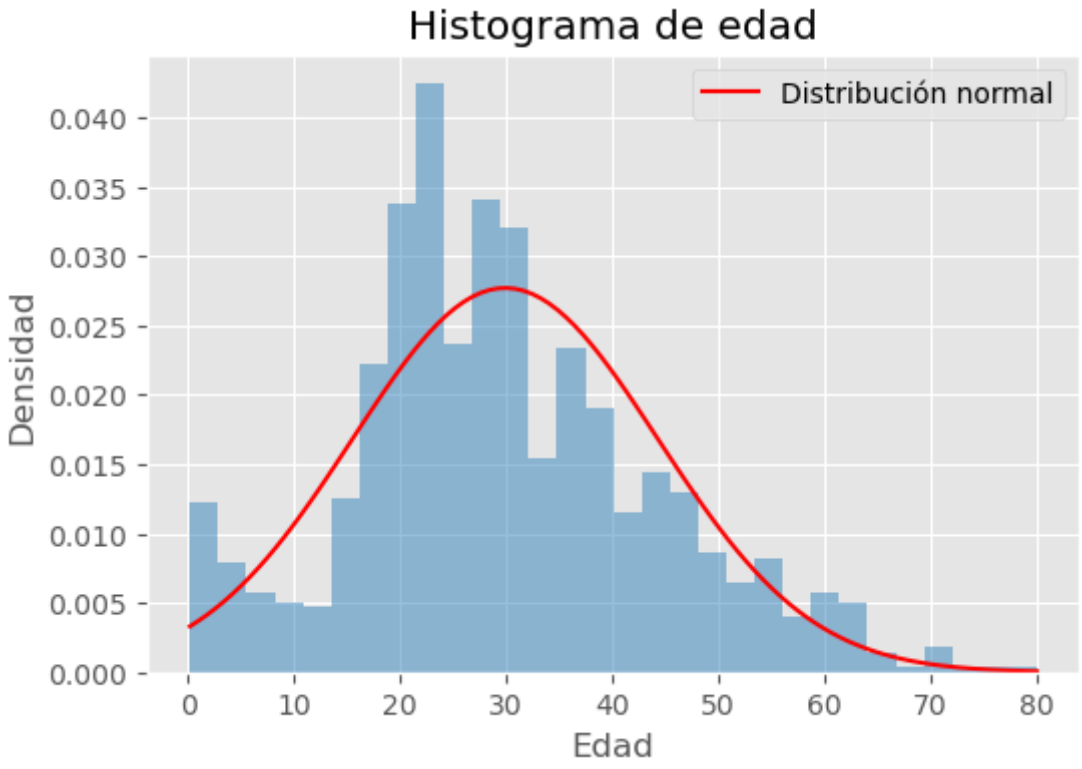
	pclass	survived	age	sibsp	parch	fare	body
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	2.294882	0.381971	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.837836	0.486055	14.413500	1.041658	0.865560	51.758668	97.696922
min	1.000000	0.000000	0.166700	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	21.000000	0.000000	0.000000	7.895800	72.000000
50%	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	155.000000
75%	3.000000	1.000000	39.000000	1.000000	0.000000	31.275000	256.000000
max	3.000000	1.000000	80.000000	8.000000	9.000000	512.329200	328.000000

El estudio gráfico es el mismo visto en clase. Primero visualizamos las frecuencias y la función de distribución empírica:



A continuación, probamos a hacer un ajuste por una distribución Normal.

Primero definimos la distribución a la que pretendemos comprobar si se ajustan los datos, en este caso una normal. Después con el método fit obtenemos los parámetrso y por último graficamos el resultado:



Mostramos los resultados del ajuste:

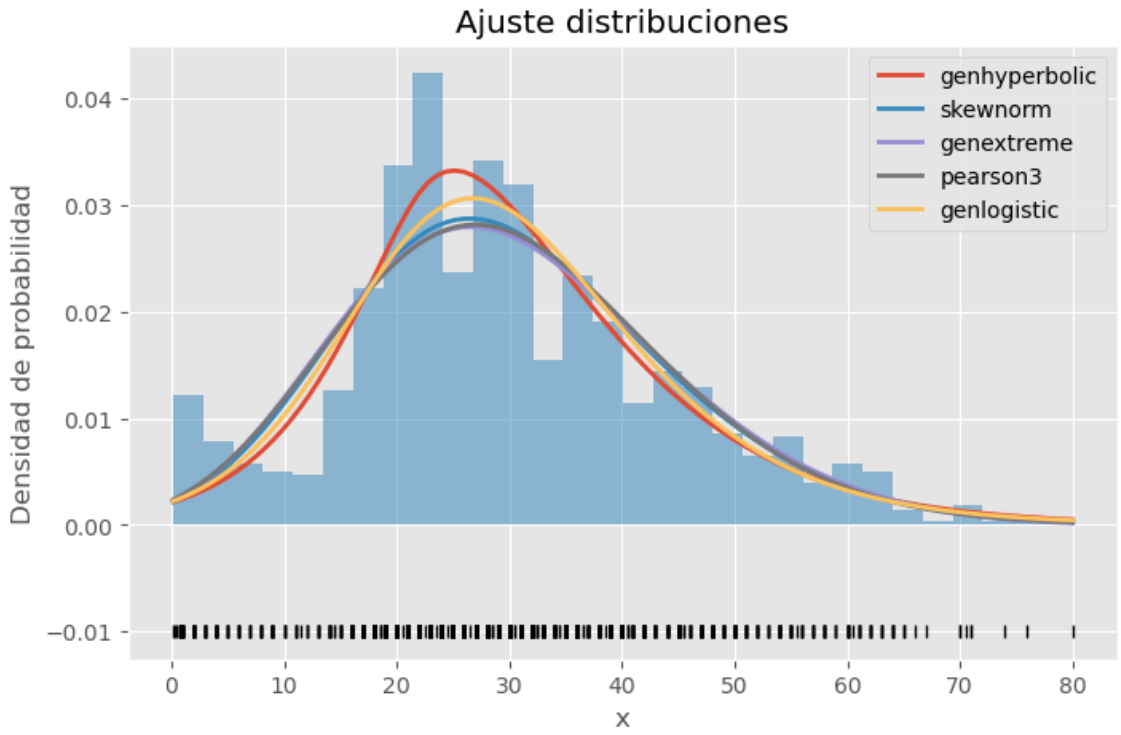
```
--------------------
Resultados del ajuste
--------------------
Distribución:   norm
Dominio:        [-inf, inf]
Parámetros:     {'loc': 29.8811345124283, 'scale': 14.40660823421771}
Log likelihood: -4274.610314500958
AIC:            8553.220629001917
BIC:            8563.126086291166
```

Como podemos ver el log_likelihood my negativo nos indica que el modelo normal no describe la distribución de nuestras edades.

Ajuste simultaneo de varias distribuciones:

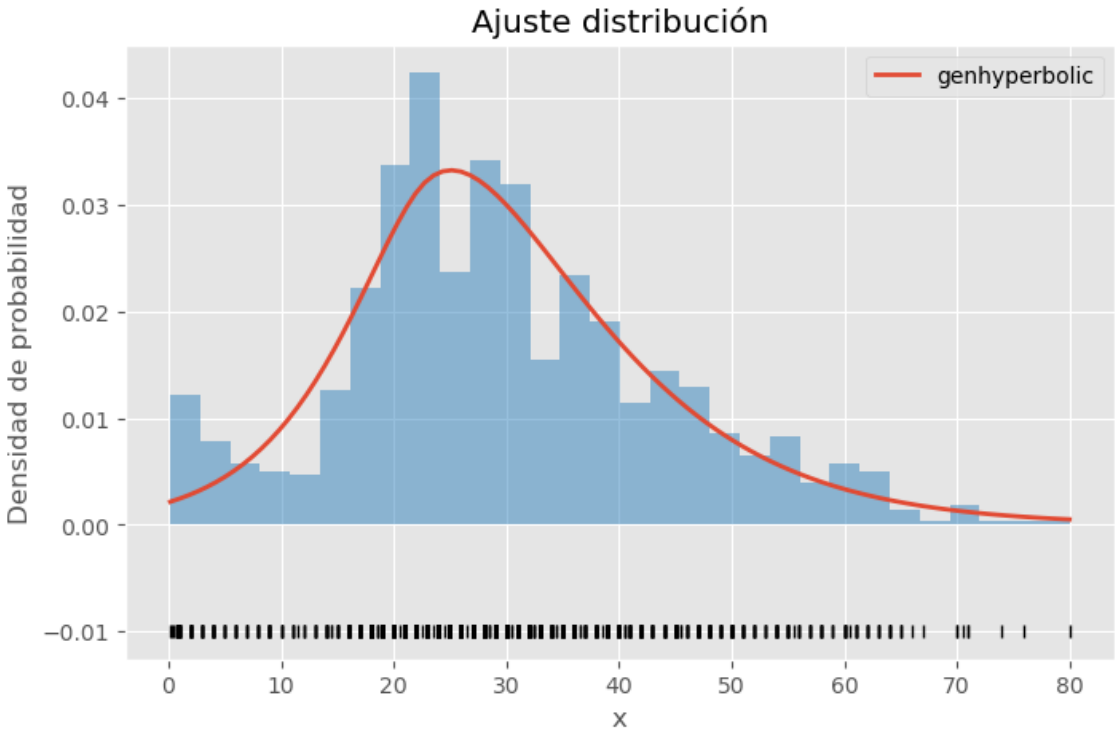
Como hemos visto que nuestra distribución no se ajusta bien por una normal vamos a seguir un método grafico que consiste en intentar ajustar al mimo tiempo con diversas funciones de distribuciones, a ver en cual "encaja" mejor nuestra función de distribución de la edad.

Mostramos a continuación los resultados.



El mejor de los ajustes resulta ser el genhyperbolic.

```
--------------------
Resultados del ajuste
--------------------
Distribución:   genhyperbolic
Dominio:        [-inf, inf]
Parámetros:     {'p': 2.178894844386954,
Log likelihood: -4255.499527850544
AIC:            8520.999055701088
BIC:            8545.762698924213
```



En caso de que tuviésemos que hacer estudios relativos a la edad procuraríamos usar esta distribución. ¡Esto es fundamental tenerlo en cuenta ya que es el motivo por el que estamos intentando ajustar nuestra variable a una distribución conocida!

Contrastes de homogeneidad de Willcoxon y Mann-Whitney

Vamos a estudiar si hay homogeneidad ente hombres y mujeres y lo que pagaron en el billete. Para esto primero vamos a hacer un test de homogeneidad de Willcoxon.

Recurrimos a este test de inferencia no paramétrica porque no se cumple la normalidad y por tanto no podemos aplicar técnicas de inferencia paramétrica.

```
stat, p = mannwhitneyu(f_fare.fare, m_fare.fare, alternative='less')
print('p-value:', p)
```

```
p-value: 1.0
```

Este test, con el hyperparámtro les nos indica que no hay homogeneidad entre lo pagado por hombres y mujeres y además dado eu el p-valor es 1 nos dice que las mujeres pagaron más.

Repito el proceso pero con un test de Willcoxon para ver si el resultado es muy distinto o nos sigue saliendo lo mismo.

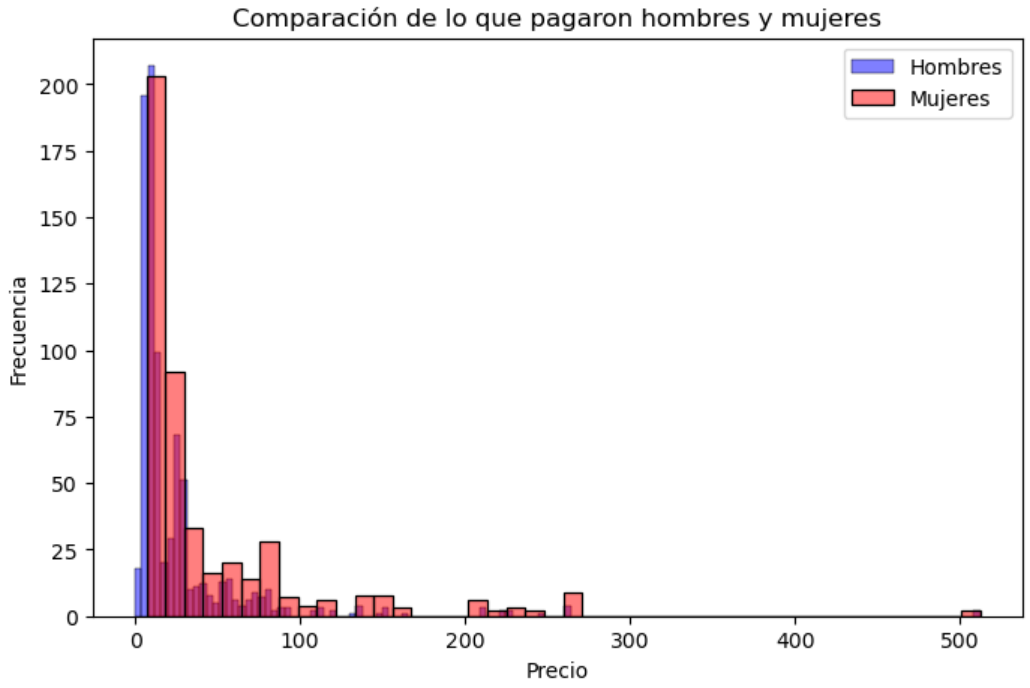
En este test la hipótesis nula sería que no hay diferencias significativas entre el sexo de los pasajeros y lo que pagaron por su billete y la hipótesis alternativa sería que si que hay diferencias significativas entre sexos y lo que pagaron respectivamente.

```
Statistics=32521.500, p=0.041
Rechazamos H0: Hay diferencias significativas
```

Una vez aplicado el test, obtenemos un p-valor de 0,041. Al 95% de certeza rechazamos H0, pero al 99% no podríamos rechazarlo.

Podemos decir después de los dos tests que hay diferencias significativas entre lo que pagaron hombres y mujeres, y además podemos decir que las mujeres pagaron más

Para visualizar esto hago una grafica con la comparación de lo que pagaron hombres y mujeres.



Contraste de homogeneidad de más de dos muestras.

El contraste kruskal sirve para ver la homogeneidad cuando queremos comparar más de dos muestras. Hasta ahora comparábamos los sexos con lo que pagaban, vamos a comparar ahora las tres clases de los pasajeros con sus respectivas edades. Decidí comparar con las edades y no con el precio del tique porque la distinción entre clases es precisamente lo que pagaron por su billete y por tanto no van a ser homogéneas las distribuciones.

Aplico el test de Kruskal:

```
stat, p = kruskal(class1.age, class2.age, class3.age)
print(stat,p)
```

```
173.07483768708096 2.613825063985474e-38
```

La interpretación de este resultado es que el p-valor es bajísimo, del orden de 10^(-38), es decir rechazamos H0 con seguridad. Lo que significa que las edades en las tres clases difieren significativamente.

Lo que pasa es que este test nos dice únicamente que hay al menos dos variables que no son homogéneas entre si, pero para saber cuáles de las tres posibles parejas no son homogéneas hay que hacer un contraste de willcoxon individual para cada una de las parejas.

Aplicamos la corrección de Bonferroni según el número de variables que estamos observando:

```
#Aplicamos la corrección de Bonferroni
# Bonferroni: 3 comparaciones
alpha_2=0.05/3
alpha_2
```

```
0.016666666666666666
```

Y seguimos el siguiente esquema para comparar de dos en dos las clases.

```
stat, p = mannwhitneyu(class1.age, class2.age,alternative='two-sided')
print('p-value:', p)
if p > alpha_2:
    print('No rechazamos H0: No hay diferencias significativas entre los grupos 1 y 2.')
else:
    print('Rechazamos H0: Hay diferencias significativas entre los grupos 1 y 2.')
```

```
p-value: 8.078340942588266e-14
Rechazamos H0: Hay diferencias significativas entre los grupos 1 y 2.
```

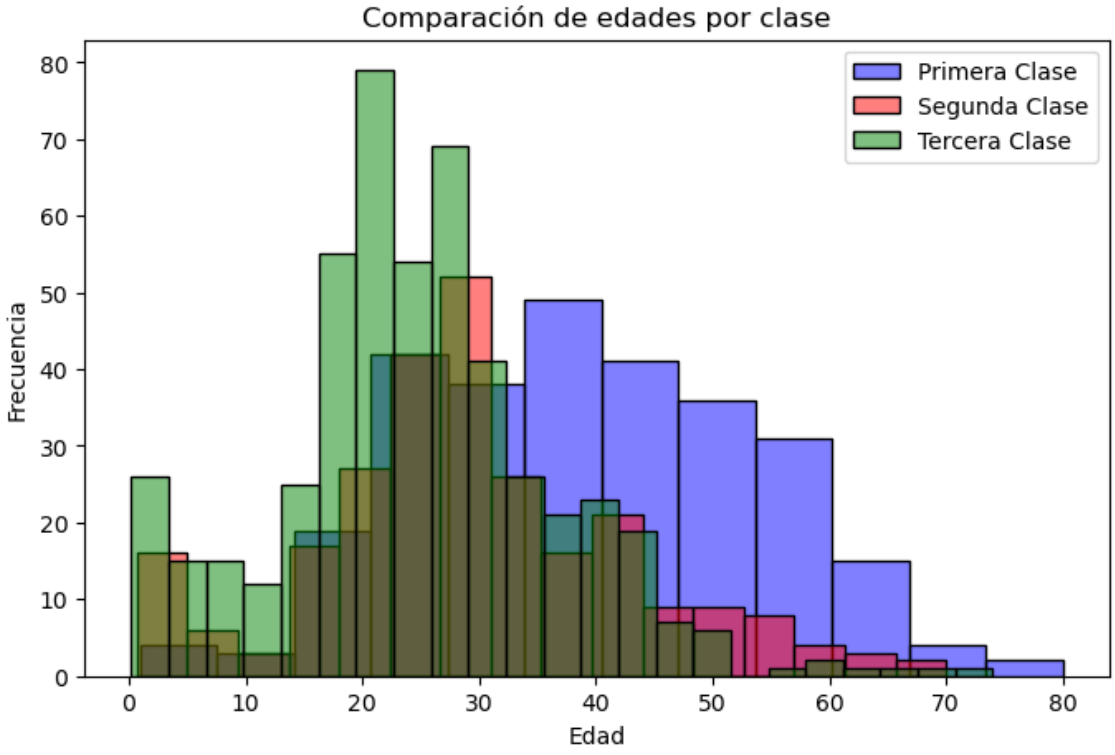
Obteniendo también los resultados:

```
p-value: 4.7417379898130856e-07
Rechazamos H0: Hay diferencias significativas entre los grupos 2 y 3.
```

```
p-value: 2.7732044525827l9e-38
Rechazamos H0: Hay diferencias significativas entre los grupos 1 y 3.
```

Vemos que las diferecias son significativas en edad entre las tres clases con el test de Kruskal-Wallis. Y analizando las clases dos a dos con el test de Mann-Whitney vemos que todas ellas son diferentes.

Para hacer una visualización de todo esto muestro de forma gráfica las distribuciones de edades de las tres clases. Aunque matemáticamente menos preciso, nos ayuda a ver que realmente hay claras diferencias entre las distribuciones de edades de las tres clases.



Podemos hacer el test de Kruskal Wallis para comparar el precio pagado entre los pasajero de distintas clases, pero claramente nos va a dar distinto, ya que ese el motivo principal de diferencia entre clases.

```
602.4393587194828 1.5203914315752115e-131
```

El p-valor es de 10^(-131)!!

Volvemos a R para concluir con los métodos paramétricos.

Antes de terminar la sección de métodos no paramétricos contraste de independencia. Este contraste sirve para ver si hay dos variables dependientes, cuyos valores están relacionados o no.

Por ejemplo, voy a ver si el hecho de ser mujer es independiente de haber sobrevivido al naufragio.

Obtengo primero la tabla de contingencia:

```
        0   1
female 127 339
male   682 161
```

Esta tabla lo que nos dice es las proporciones de hombres y mujeres, que sobrevivieron [1] o fallecieron [0].

Obtenemos ahora la tabla de frecuencias esperadas en caso de que no fuesen independientes, es decir si fuese arbitrario sobrevivir y el sexo del pasajero.

```
             0        1
[1,] 288.0015 177.9985
[2,] 520.9985 322.0015
```

La primera fila y primera columna son el número de mujeres que debería haber fallecido en caso de ser variables independientes, la primera fila y la segunda columna son el número de mujeres que deberían haber sobrevivido en caso de ser variables independientes, la supervivencia del sexo.

Aplicamos el test:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  tabla.contingencia
X-squared = 363.62, df = 1, p-value < 2.2e-16
```

De nuevo el p-valor nos sale muy pequeño, por tanto, las variables no son independientes.

Miramos ahora la clase a la que pertenecen los pasajeros si es independiente.

Obtenemos su tabla de contingencia y su tabla de frecuencias esperadas:

```
> (tabla.contingencia = table(titanic_df$pclass,titanic_df$survived))

     0   1
  1 123 200
  2 158 119
  3 528 181
> (tabla.frec.esperadas =rowSums(tabla.contingencia)%*%t(colSums(tabla.conting$
+  /sum(tabla.contingencia))
           0       1
[1,] 199.6234 123.3766
[2,] 171.1940 105.8060
[3,] 438.1826 270.8174
```

Y aplicamos el test para ver si son independientes:

```
        Pearson's Chi-squared test

data:  tabla.contingencia
X-squared = 127.86, df = 2, p-value < 2.2e-16
```

Una vez más obtenemos que la clase en la que viajaban los pasajeros no es independiente de haber sobrevivido o no a accidente del titánic.

Hasta aquí el estudio de métodos no paramétricos.

2.3.Modelos y series temporales

En este apartado me centrar en el desarrollo de un modelo Logit.

El objetivo de hacer un modelo Logit es poder predecir si un pasajero va a sobrevivir o no según el valor de sus variables. A cada variable se le asignan unos “pesos” que son indicativo de cuánto influye la variable en la probabilidad de sobrevivir.

Por ejemplo, en nuestro primer modelo de Regresión Logistica obtenemos los siguientes pesos:

```
pclass:          Weight=-1.0688115380285343,     Standard deviation=-0.9007670517833876
age:             Weight=-0.037684503934559965,   Standard deviation=-0.5486050734455753
sex_num:         Weight=-2.4680958516026656,     Standard deviation=-1.1889248177566638
sibsp:           Weight=-0.3443743827383492,     Standard deviation=-0.3197994246523243
parch:           Weight=0.05816679674481951,     Standard deviation=0.05044978701123106
fare:            Weight=0.0017909418131150149,   Standard deviation=0.09816583298257188
```

Estas son las variables que vamos a tener en cuenta a la hora de crear nuestro modelo Logit. Por ejemplo, que pclass tenga asociado un valor de -1 aproximadamente significa

que por cada unidad que aumentamos la clase disminuye en una unidad la probabilidad de sobrevivir.

A primera vista podemos intuir que las variables que más influencia tienen en si un pasajero se salvaba o no son el sexo, la clase y si tenía familiares a bordo.

Calculamos su z-valor para ver cuales son más influyentes en la función objetivo, que es predecir si el pasajero va a sobrevivir o no.

```
pclass:          z_value=1.1865570969902186
age:             z_value=0.068691497324073 65
sex_num:         z_value=2.075905738312049
sibsp:           z_value=1.0768449102519242
parch:           z_value=1.1529641687460543
fare:            z_value=0.018244044375734826
```

Y su p-valor:

```
pclass:          p_value=0.23540236725959796
age:             p_value=0.9452351863706787
sex_num:         p_value=0.03790266989587021
sibsp:           p_value=0.2815495578428957
parch:           p_value=0.24892509182199252
fare:            p_value=0.9854441661436075
```

sabiendo ahora el p-valor, podemos decir que la variable que más influye en si un pasajero se salvaba o no es el sexo, seguido de la clase y si tenía familiares a bordo.

Pero la única relevante con un p-valor menor a 0.05 es el sexo. Y las que se podrían descartar son la edad y lo que pagó por el tiquete.

De momento hacemos algunas predicciones con este modelo:

Por ejmplo la probabilidad de sobrevivir de un hombre de 30 años, de segunda clase, sin hemanos ni padres a bordo que pagó 100 libras:

```
4  probabilidad = model.predict_proba([[1, 30, 2, 0, 0, 100]])[0][1]
5
6  print(f"La probabilidad de que sobreviva es de {probabilidad * 100:.2f}%")
✓ 0.0s
La probabilidad de que sobreviva es de 8.92%
```

Si ahora hacemos que sea de primera clase:

```
La probabilidad de que sobreviva es de 53.61%
```

¿Y si además fuese menor? Edad 16 años:

```
La probabilidad de que sobreviva es de 66.20%
```

¿Y si fuese una mujer?

```
La probabilidad de que sobreviva es de 85.08%
```

Y si mantemos todas las características, es decir una mujer de 16 años, sin familiares a bordo, pero fueses en tercera clase.

```
La probabilidad de que sobreviva es de 3.93%
```

Inmediatamente la probabilidad se supervivencia cae en picado. Lo que nos da información muy interesante, y es que ¡la clase a la que se pertenecía es mucho más influyente que el sexo!

Ahora realizamos un segundo modelo eliminado las características con un p-valor despreciable, es decir nos quedamos solo con las variables de la clase, sexo y familiares a bordo y recalculamos los pesos:

```
pclass:         Weight=-0.8355968636008496,     Standard deviation=-0.7028531701172088
sex_num:        Weight=-2.475069597885323,      Standard deviation=-1.1958306286249374
sibsp:          Weight=-0.2208515736115964,     Standard deviation=-0.20142425874299036
parch:          Weight=0.10456600841570661,     Standard deviation=0.08779879975353133
```

Y sacamos los nuevos p-valores:

```
pclass:         p_value=0.234493170332250l4
sex_num:        p_value=0.038475828092960684
sibsp:          p_value=0.2728820122477069
parch:          p_value=0.23366414662817192
```

Calculamos ahora la probabilidad de sobrevivir de un hombre de segunda clase sin familiares a bordo.

```
La probabilidad de que sobreviva es de 5.76%
```

Personalmente creo que el primer modelo hace mejores predicciones. Porque, aunque las características del tiquete y la edad no fuesen muy influyentes, al final se ajusta mejor y retorna predicciones que tienen más sentido. Ya que creo que decir que tener hijos, pareja o familiares en general no afecta mucho como ocurre en el segundo modelo, en el que ocurre que estar casado es negativo es claramente erróneo.

3. Conclusiones

El análisis estadístico realizado sobre el conjunto de datos del Titanic proporciona una visión detallada del accidente del Titanic, dándonos información relevante sobre las variables que influyeron significativamente en la supervivencia de los pasajeros. Usando métodos paramétricos y no paramétricos hemos obtenido conclusiones interesantes tanto para un análisis de datos exclusivamente como para un análisis desde un punto de vista del estudio histórico del acontecimiento.

Me gustaría hablar en las conclusiones de importancia de la clase social y el sexo a la hora de sobrevivir en el accidente. Estas variables han sido las que hemos estado trabajando durante todo el trabajo porque ciertamente son las más relevantes y las que más información aportan. A nivel de estudio estadístico, pero también si quisiésemos hacer inferencia sobre el contexto social de la época y el comportamiento de las personas en 1912. Los resultados subrayan las desigualdades sociales que había. Los pasajeros de primera clase tenían una probabilidad mucho mayor de sobrevivir en comparación con los de tercera clase. Tristemente, este análisis puede servir como un recordatorio de cómo las estructuras sociales pueden influir en los resultados en situaciones de emergencia.

Pero al mismo tiempo se ve reflejado que la sociedad tenía una claro lo que podríamos llamar un sentido de nobleza, al priorizar la supervivencia de mujeres y niños en caso de tragedia.

Comentar también las limitaciones del análisis realizado, ya que la muestra con la que hemos trabajado tenía valores nulos en diversas columnas y además no tiene en cuenta la tripulación del Titanic en los datos. Esto último es importante, porque teniendo en cuenta que la mayoría de la tripulación eran hombres que fallecieron en el accidente los resultados hubiesen sido todavía más negativos para el sexo masculino. Además una característica que no existe en la base de datos es la nacionalidad o raza de los pasajeros. Que hubiese sido muy interesante estudiar como influyó o si fue influyente en la supervivencia o no de los pasajeros.

En resumen, este análisis estadístico del dataset del Titanic ha permitido no solo identificar las variables más influyentes en la supervivencia de los pasajeros, sino también proporcionar un contexto histórico y social que facilita nuestra comprensión de la tragedia.