Trabajo Teórico Estadística
Curso 2022/2023
Estudio sobre NFL

Grupo

G11 - 2ºB

Participantes

Georgi Angelov Chervenyashki

Índice

1. INTRODUCCION

Hemos elegido una base de datos sobre la **NFL** (National Football League) o Liga Nacional de Fútbol Americano. Más concretamente datos registrados durante la **temporada 2016** o la 97.ª edición. Estos datos están limitados a los **32 equipos** que participan en la competición registrando así los datos más importantes de cada partido disputado durante las 17 semanas de duración de la competición.

2. Presentación datos de entrada y sus tipos

Hemos recogido estos datos en un **dataframe** con las siguientes variables:

- **Team**: Nombre del equipo
- **Wins**: Número de victorias durante la temporada
- **Losses**: Número de derrotas durante la temporada
- **Ties**: Número de empates durante la temporada
- **WinPct**: Porciento de victorias
- **PointsFor**: Puntos a favor
- **PointsAgainst**: Puntos en contra
- **NetPts**: Diferencia de puntos
- **YardsFor**: Yardas a favor
- **YardsAgaint**: Yardas en contra
- **TDs**: Número de Touchdowns
- **Division**: División del equipo (Este, Oeste, Norte, Sur)
- **Conference**: Tipo de federación (AFC o NFC)

2.1. Variables Cualitativas

Uno de los aspectos que podemos analizar con las variables cualitativas de nuestra base de datos es la organización de los 32 equipos y la relación que se da entre el número de equipos de distintas división o federación.

```
tabla <- table(NFL$Division, NFL$Conference)
tabla
```

```
      AFC NFC
East    4   4
North   4   4
South   4   4
West    4   4
```

```
prop.table(tabla)
```

```
       AFC   NFC
East  0.125 0.125
North 0.125 0.125
South 0.125 0.125
West  0.125 0.125
```

La tabla nos permite ver de un vistazo que existe una paridad o distribución uniforme entre el número de equipos de cada división y federación.

También podemos ver las frecuencias relativas y absolutas en términos de proporciones

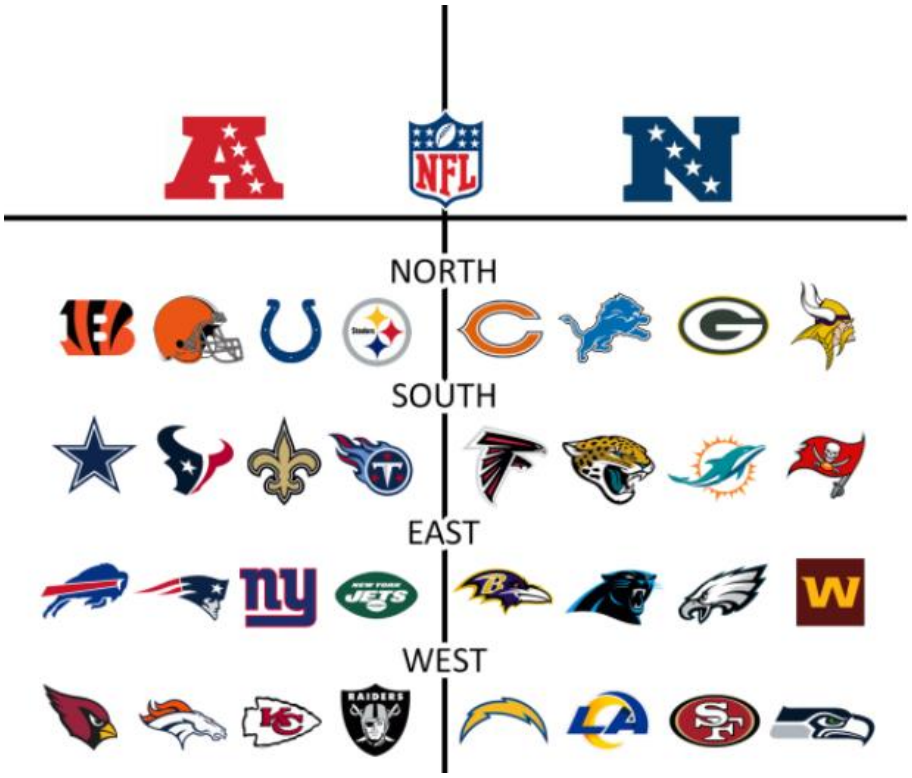
El resultado es otro objeto de la clase table al que se le han añadido una o varias filas o columnas, que contienen las frecuencias marginales, tanto absolutas como relativas.

```
addmargins(tabla)
```

```
      AFC NFC Sum
East    4   4   8
North   4   4   8
South   4   4   8
West    4   4   8
Sum    16  16  32
```

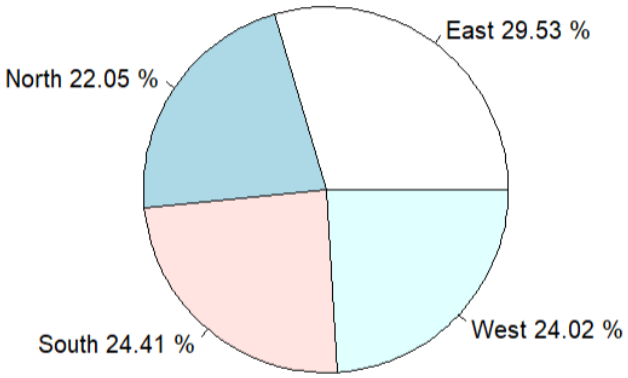
```
addmargins(prop.table(tabla))
```

```
       AFC   NFC   Sum
East  0.125 0.125 0.250
North 0.125 0.125 0.250
South 0.125 0.125 0.250
West  0.125 0.125 0.250
Sum   0.500 0.500 1.000
```

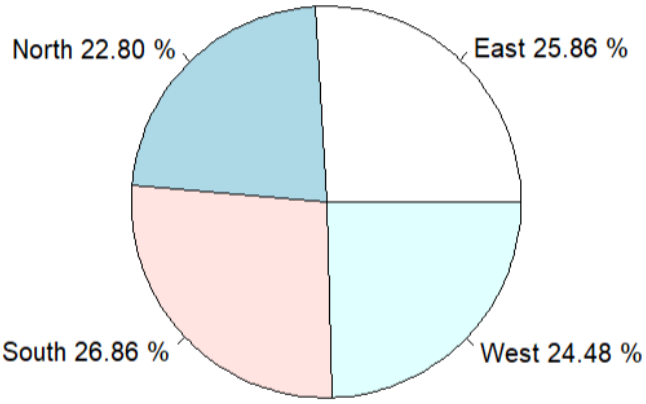


2.2. Variables Cuantitativas

Porciento de victorias por división

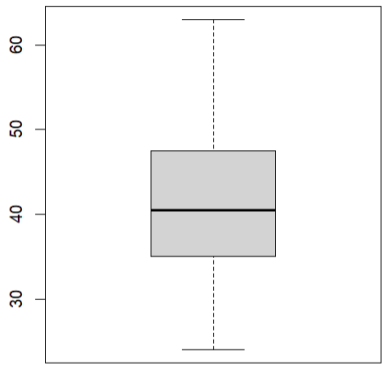


Porciento de touchdowns por división

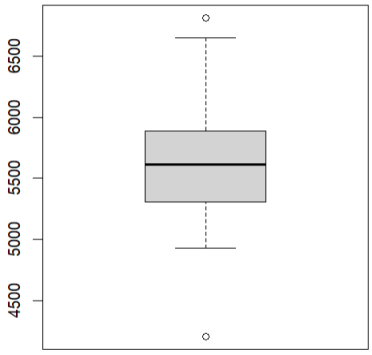


Vamos a identificar cual es la división más competitiva calculando las medias de Touchdowns, yardas a favor y puntos a favor de cada una de ellas. Después de identificar las variables que vamos a utilizar comprobamos si existen outliers, en el caso de la variable yardas a favor existen 2 outliers uno inferior y otro superior, procedemos a identificarlos, ver a que división pertenecen y eliminar para calcular la media.

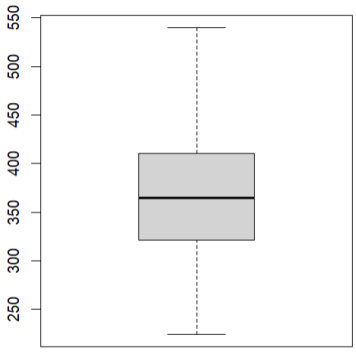
Boxplot de Touchdowns



Boxplot de yardas a favor



Boxplot de puntos a favor



Media de Touchdowns

```
East  North  South   West
42.250 37.250 43.875 40.000
```

Media de puntos a favor

```
East    North    South    West
371.500 339.375 390.125 356.625
```

Media de yardas a favor

```
East     North    South     West
5700.875 5534.875 5807.625 5382.125
```

Media yardas a favor sin outliers

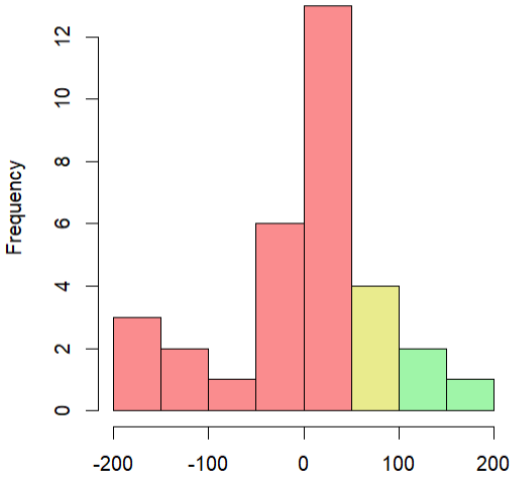
```
East     North    South    West
5700.875 5534.875 4955.625 4856.750
```

Podemos ver que los outliers tenían una influencia significativa en la media, si se tienen en cuenta la división del Sur es la más competitiva en todos los aspectos.

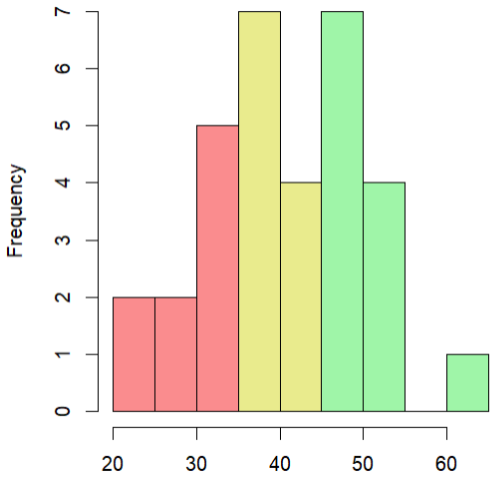
Con las variables cuantitativas de nuestra base de datos hemos medido el rendimiento de cada equipo, los hemos categorizado en 3 grupos según los Touchdowns y la diferencia de puntos obtenidos durante la temporada, donde podemos encontrar los siguientes grupos según el rendimiento de cada equipo:

- Good: NetPts <= 10 y TDs <= 30
- Decent: NetPts <= 100 y TDs 30 - 45
- Excelent: NetPts >= 100 y TDs >= 45

Distribución de frecuencias variable NetPts



Distribución de frecuencias variable TDs



Hemos obtenido los siguientes resultados:

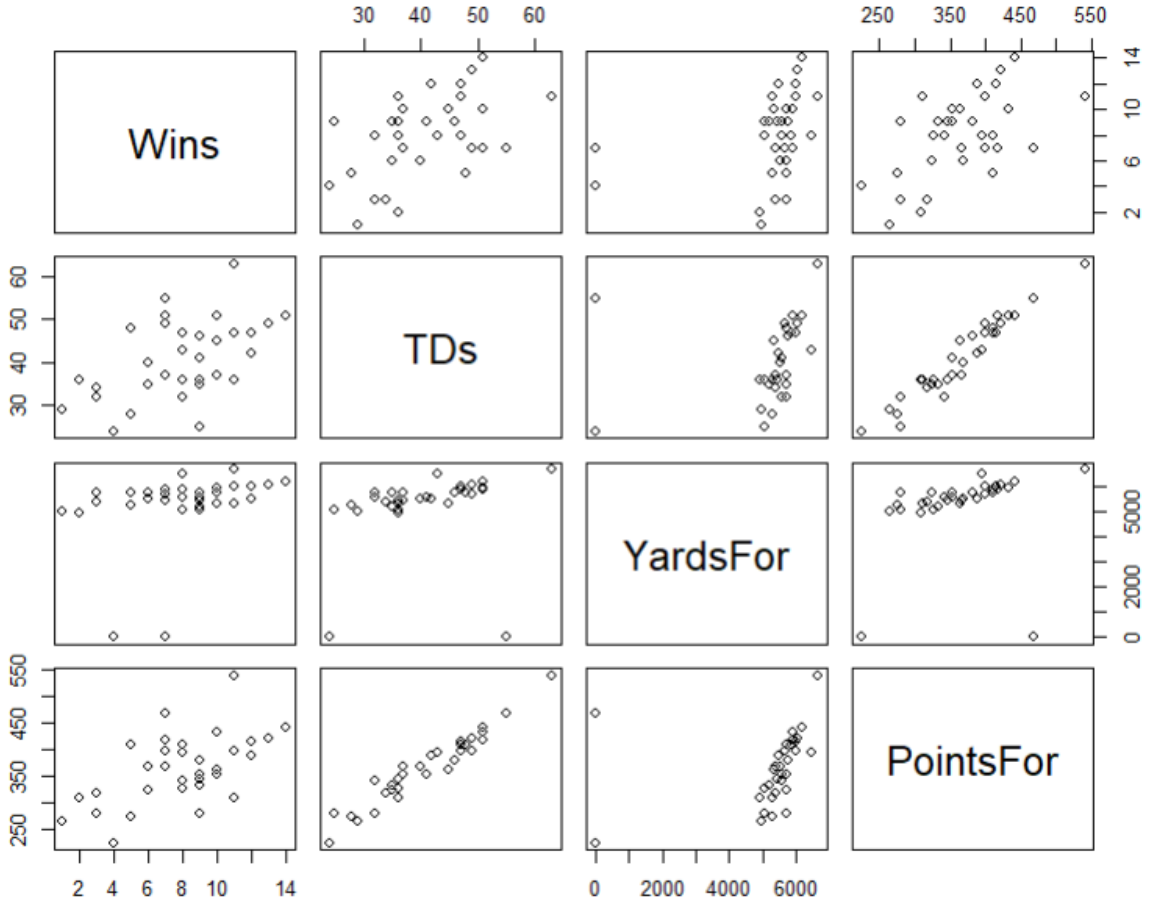
- Good: 14 equipos
- Decent: 15 equipos
- Excelent: 3 equipos

Podemos ver que los 3 equipos que han tenido una temporada excelente según nuestro criterio están muy bien posicionados en la calificación.

X	Team	Wins	Losses	Ties	WinPct	PointsFor	PointsAgainst	NetPts	YardsFor	YardsAgainst	TDs	Division	Conference
1	New England Patriots	14	2	0	0.875	441	250	191	6179	5222	51	East	AFC
2	Dallas Cowboys	13	3	0	0.813	421	306	115	6027	5502	49	East	NFC
3	Kansas City Chiefs	12	4	0	0.750	389	311	78	5488	5896	42	West	AFC
4	Oakland Raiders	12	4	0	0.750	416	385	31	5973	6002	47	West	AFC
5	Atlanta Falcons	11	5	0	0.688	540	406	134	6653	5939	63	South	NFC

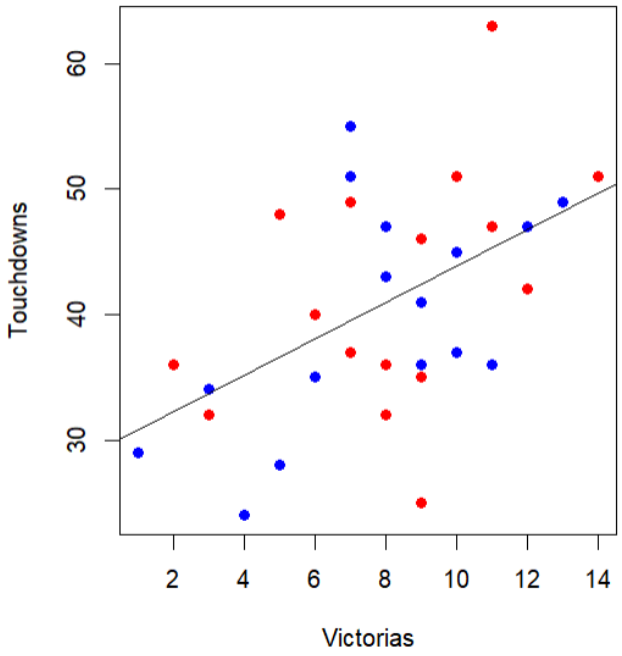
3. Análisis de relaciones entre variables

Observamos las relaciones entre las variables de nuestro dataset en la matriz de gráficos y tabla con los coeficientes de correlación.

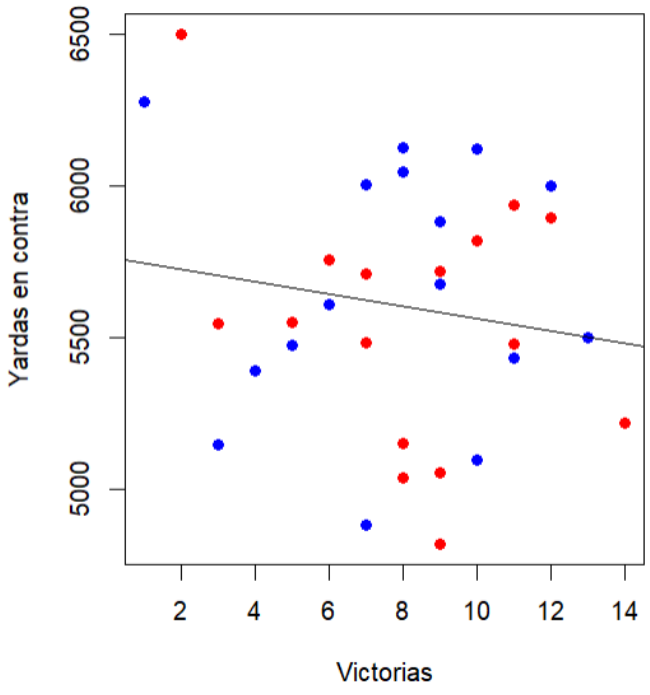


	Wins	TDs	YardsFor	PointsFor
Wins	1.0000000	0.5082924	0.3228361	0.5737728
TDs	0.5082924	1.0000000	0.2317405	0.9637721
YardsFor	0.3228361	0.2317405	1.0000000	0.2702086
PointsFor	0.5737728	0.9637721	0.2702086	1.0000000

Hemos optado por dos relaciones. La primera es la relación entre victorias y Touchdowns destacar aquí que un Touchdown equivale a 6 puntos a favor. Está relación nos da un coeficiente de Pearson de 0.5082924 lo que quiere decir que existe una relación positiva moderada entre las dos variables, cuantos más touchdowns una mayor posibilidad de ganar. Está relación también se muestra en la siguiente figura:



Otra relación podría darse entre las victorias y las yardas en contra en esta relación obtenemos un coeficiente de Pearson de -0.1561484 lo que nos indica que existe cierta relación inversa débil.

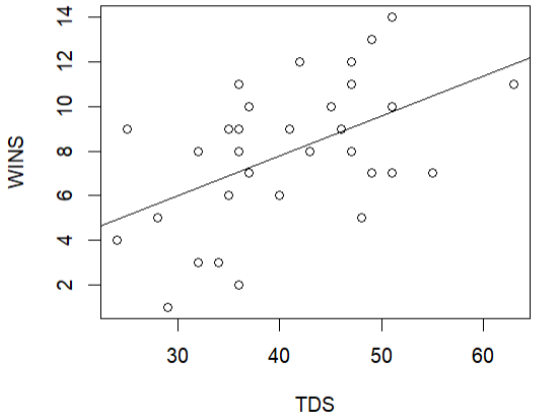


4. Modelo de regresión lineal

El modelo de regresión muestra las victorias y los touchdowns con un coeficiente de Pearson de 0.5082924 y una bondad de ajuste es de 0,2584.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.65630    2.30616   0.285  0.77792
TDS          0.17827    0.05514   3.233  0.00298 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.803 on 30 degrees of freedom
Multiple R-squared:  0.2584,    Adjusted R-squared:  0.2336
F-statistic: 10.45 on 1 and 30 DF,  p-value: 0.002976
```



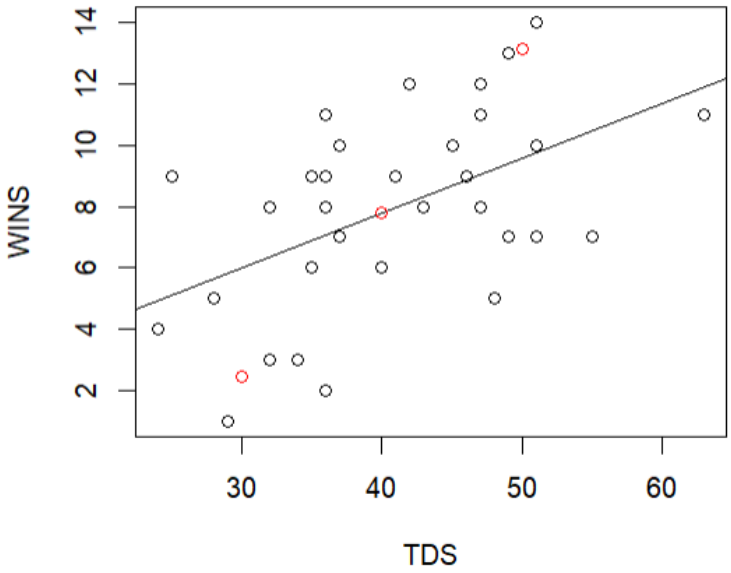
Según el modelo de regresión construido, la relación entre el numero de victorias y el número de touchdowns viene definida por la siguiente función Según el modelo de regresión construido, la relación entre el número de victorias y el número de touchdowns viene definida por la siguiente función:

$$y = 0.65630 + 0.17827\ x$$

En este modelo se toma como variable dependiente (X) el número de victorias y como variable independiente (Y) el número de touchdowns, debido a que el número de touchdowns tiene una mayor variabilidad que el número de victorias.

Según la predicción

- Un equipo con 10 touchdowns tiene entre 2 y 3 victorias (2.438999)
- Un equipo con 30 touchdowns tiene entre 7 y 8 victorias (7.787085)
- Un equipo con 70 touchdowns tiene entre 13 y 14 victorias (13.13517)



5.Contrastes de hipótesis

Se quiere comparar si la media de yardas a favor de la división South (Conferencias NFC y AFC) es igual a la media de yardas a favor de la división North. Para ello se realiza un test de hipótesis con un nivel de confianza del 97%.

- H0: Media yardas a favor South = Media yardas a favor North
- H1: Media yardas a favor South ≠ Media yardas a favor North

Es un contraste bilateral para la media

```
t.test(SouthYardsFor, NorthYardsFor, paired = TRUE, alternative = "two.side", conf.level = 0,97)
```

```
data:  SouthYardsFor and NorthYardsFor
t = 0.79947, df = 7, p-value = 0.4503
alternative hypothesis: true mean difference is not equal to 97
0 percent confidence interval:
 272.75 272.75
sample estimates:
mean difference
         272.75
```

El p-value obtenido es 0,4552, mayor que el nivel de significación α = 0,03. Podemos aceptar la hipótesis H0 y concluir con una confianza del 97% que la media de puntos de la división del norte es igual a la de la división del sur.

Realizaremos un contraste de hipótesis para contrastar con un nivel de confianza del 95% que la media de touchdowns de la federación NFC es mayor a 40.

- H0: Media >= 40
- H1: Media < 40

```
t.test(NFCTDs, mu=40, alt="greater", conf.level = 0.95)
```

```
data:  NFCTDs
t = 0.6881, df = 15, p-value = 0.2509
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 37.3883     Inf
sample estimates:
mean of x
  41.6875
```

Como podemos observar el p-value es mayor que nuestro nivel de significación 0,05 por tanto, no se descarta H0 y podemos asegurar al 95% que la media de la NFC se encuentra por encima o igual a 40 touchdowns.

Realizaremos un contraste de hipótesis para contrastar con un nivel de confianza del 99% que la media de la NetPts de la división West es menor a 0.

- H0: Media NetPts West <= 0
- H1: Media NetPts West > 0

```
t.test(WestNetPts, mu=0, alt="less", conf.level = 0.99)

    data:  WestNetPts
    t = -0.31593, df = 7, p-value = 0.3806
    alternative hypothesis: true mean is less than 0
    99 percent confidence interval:
         -Inf 96.56603
    sample estimates:
    mean of x
      -11.375
```

Como el p-value es mayor que el nivel de significación mayor que el nivel de significación aceptamos H0.