

LABORATORIO DE ESTADÍSTICA – EXAMEN PL3 (1,2 puntos)

APELLIDOS: Yaccelga Castillo

NOMBRE: Sergio Andrés

DNI: Y0668612Q

GRUPO: B1

INSTRUCCIONES:

一 Se debe descargar el examen del campus virtual, en la sección:

LABORATORIO > Laboratorio > Pruebas de Laboratorio (PL) > PL3

二 En esa carpeta hay un archivo zip que hay que descomprimir, que incluye los archivos:

三 PL3.docx con el formulario del examen

四 datos.csv con datos a usar en el examen

五 Crear un proyecto con RStudio para el examen

六 En las cajas debajo de cada enunciado de pregunta:

七 pegar una copia (texto o imagen) de la consola de RStudio, en la que aparezca la ejecución de los comandos utilizados en los cálculos;

八 y en la segunda caja indicar el resultado pedido en el enunciado.

九 Al finalizar hay que enviar el archivo PL3.docx (o en formato pdf) a través del buzón de entrega PL3, desde la sección: **LABORATORIO > Laboratorio > Pruebas de Laboratorio (PL) > PL3 > Buzón de entrega de la prueba PL3 > PL3**

Realizar las siguientes acciones:

a Descargar el fichero **datos.csv** en un directorio (carpeta) del ordenador que será el directorio de trabajo,

NOTA: **El fichero contiene datos de una muestra de equipos de fútbol extraída de una población con un total de 300 equipos.**

b cargar en una variable de tipo data frame que tenga como nombre **tu primer apellido** el contenido del archivo datos.csv, usando la función `read.csv2()`. Por ejemplo, si el alumno se llama Antonio Pérez, la variable debe ser `perez`:

```
perez = read.csv2 ...
```

a eliminar en esa variable la fila cuyo número de fila coincida con **la última cifra de tu DNI**, la que va antes de la letra. Si la cifra es 0, borrar la fila 10. Por ejemplo, si el DNI del alumno Antonio Pérez fuese 01234567Z, habría que borrar la fila 7, con el comando:

```
perez = perez[-7,]
```

a EN EL RESTO DEL EXAMEN HAY QUE UTILIZAR LA VARIABLE QUE HAYA RESULTADO CON LA FILA BORRADA.

```
Console Terminal x Jobs x
R 4.1.2 · ~/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen/ ↗
> setwd("/home/s3rzh/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen")
> getwd()
[1] "/home/s3rzh/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen"
> yacelga <- na.omit(read.csv2("./datos.csv"))
> yacelga <- yacelga[-2, ]
> head(yacelga)
  ZONA PRESUPUESTO VICTORIAS GOLES
1 Norte      89.87         14    48
3 Norte     113.70         22    26
4 Norte      97.00         22    59
5 Norte      72.90         14    45
6 Norte     148.02          8    51
7 Norte      79.37         12    48
> |
```

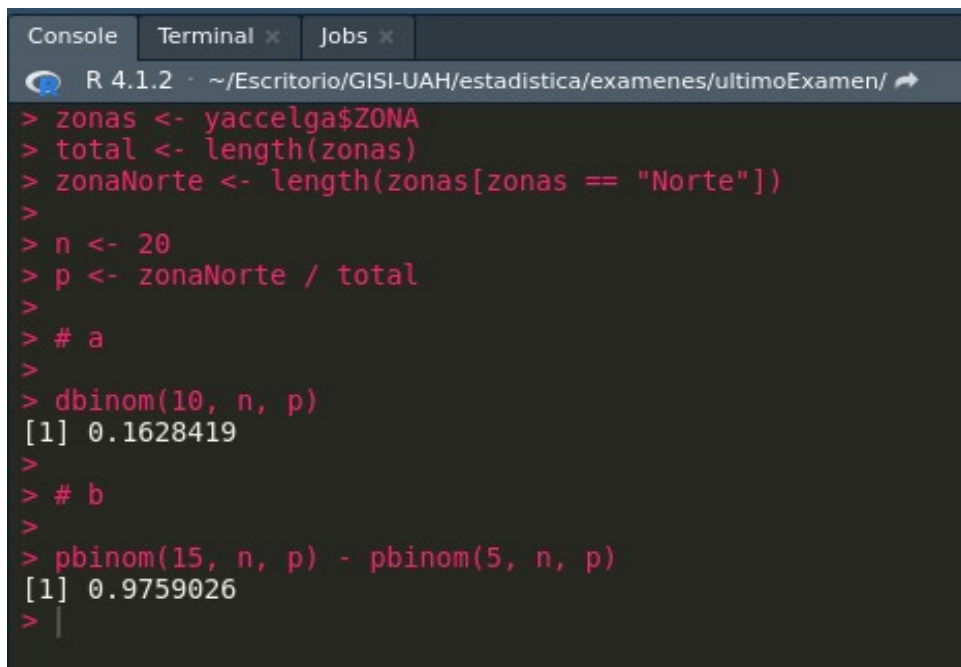
(0,2p) Nos fijamos en La variable estadística ZONA, que representa la ubicación geográfica de la ciudad de cada equipo: Norte o Sur. Si se eligen aleatoriamente 20 equipos, responder a las siguientes preguntas:

a. Calcular la probabilidad de que 10 de los 20 equipos elegidos sean de la zona Norte.

b. Calcular la probabilidad de que entre 5 y 15 (ambos valores incluidos) de los 20 equipos elegidos sean de la zona Norte.

NOTA: Utilizar una variable aleatoria binomial $X \sim B(n, p)$ que represente el “número de equipos que son de la zona Norte en un conjunto de 20 equipos”. Recordar que en distribuciones discretas $P[X < \text{valor}]$ no es lo mismo que $P[X \leq \text{valor}]$.

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:



```
Console Terminal x Jobs x
R 4.1.2 ~/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen/
> zonas <- yaccelga$ZONA
> total <- length(zonas)
> zonaNorte <- length(zonas[zonas == "Norte"])
>
> n <- 20
> p <- zonaNorte / total
>
> # a
>
> dbinom(10, n, p)
[1] 0.1628419
>
> # b
>
> pbinom(15, n, p) - pbinom(5, n, p)
[1] 0.9759026
> |
```

RESULTADOS:

n = 20 equipos

$p = \text{zonaNorte} / \text{total}$

a)

Probabilidad de que 10 de los 20 equipos elegidos sean de la zona Norte = 0.1628419

b)

Probabilidad de que entre 5 y 15 de los 20 equipos elegidos sean de la zona Norte = 0.9759026

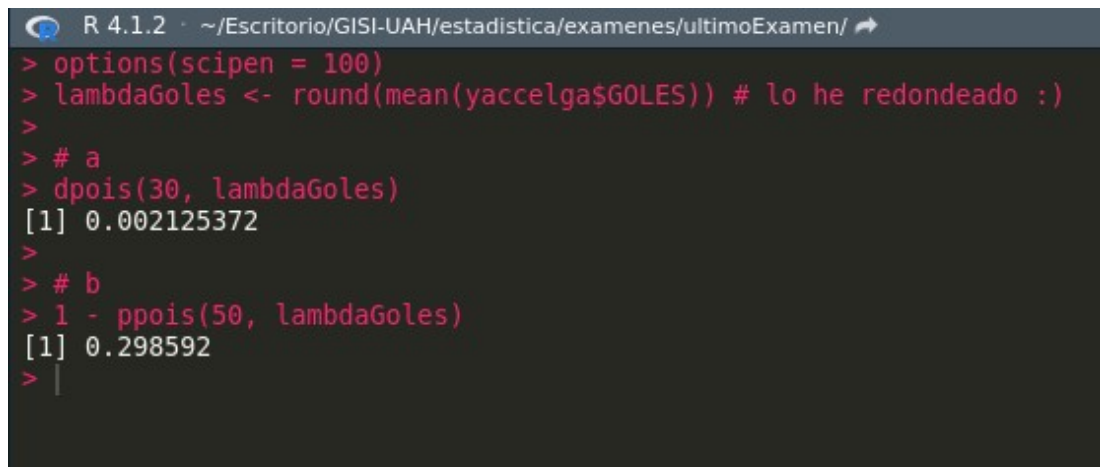
(0,2p) Nos fijamos en La variable estadística GOLES que representa los goles marcados por cada equipo en un año.

a. Calcular la probabilidad de que un equipo marque 30 goles en un año.

b. Calcular la probabilidad de que un equipo marque más de 50 goles en un año.

NOTA: Utilizar una variable aleatoria de Poisson $X \sim P(\lambda)$ que represente los “goles marcados por un equipo en un año, sabiendo que la media de goles marcados por un equipo es de λ goles en un año”. Al calcular λ redondear a un valor entero.

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:



```
R 4.1.2 · ~/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen/ ↵
> options(scipen = 100)
> lambdaGoles <- round(mean(yaccelga$GOLES)) # lo he redondeado :)
>
> # a
> dpois(30, lambdaGoles)
[1] 0.002125372
>
> # b
> 1 - ppois(50, lambdaGoles)
[1] 0.298592
> |
```

RESULTADOS:

$\lambda = 47$ goles de media marcados por equipo en un año.

a)

Probabilidad de que un equipo marque 30 goles en un año = 0.021

b)

Probabilidad de que un equipo marque más de 50 goles en un año = 0.298

(0,2p) Nos fijamos en La variable estadística PRESUPUESTO que representa el presupuesto anual de cada equipo en millones de euros.

a. Calcular la probabilidad de que un equipo elegido al azar tenga un presupuesto de entre 20 y 50 millones de euros.

b. Dibujar en un único diagrama el histograma con los datos reales de la variable estadística PRESUPUESTO junto a la función de densidad de la variable aleatoria, para poder comprobar visualmente si son similares.

NOTA: Utilizar una variable aleatoria Normal $X \sim N(m, d)$ que represente el “presupuesto anual de un equipo”. Al ser una variable continua se acepta que $P[X \leq \text{valor}]$ coincide con $P[X < \text{valor}]$.

```
R 4.1.2 · ~/Escritorio/GIS-UAH/estadistica/examenes/ultimoExamen/ ↗
> presupuestos <- yaccelga$PRESUPUESTO
> mPresupuesto <- mean(presupuestos)
> dPresupuesto <- sd(presupuestos)
>
> # a
>
> probabilidad20y50m <- pnorm(50, mPresupuesto, dPresupuesto) - pnorm(20, mPresupuesto, dPresupuesto)
>
> # b
>
> par(mfrow=c(1,2))
> curve(dnorm(x, mPresupuesto, dPresupuesto), 0, 200, main="Función de densidad")
>
> hist(presupuestos, freq=FALSE)
> curve(dnorm(x, mPresupuesto, dPresupuesto), 0, 200, add = TRUE)
> # Sí son similares.
> par(mfrow=c(1,1))
> |
```

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:

RESULTADOS:

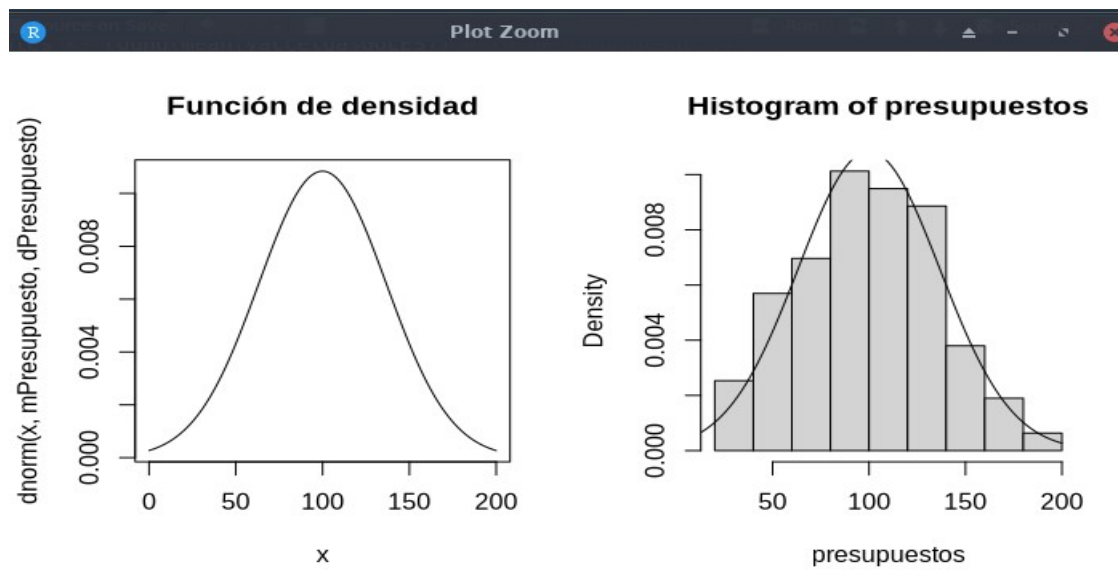
m = 99.94709 millones de euros

d = 36.78711 millones de euros

a)

Probabilidad de que un equipo tenga un presupuesto de entre 20 y 50 millones de euros = 0.07239287

b) (Pegar la imagen del diagrama)

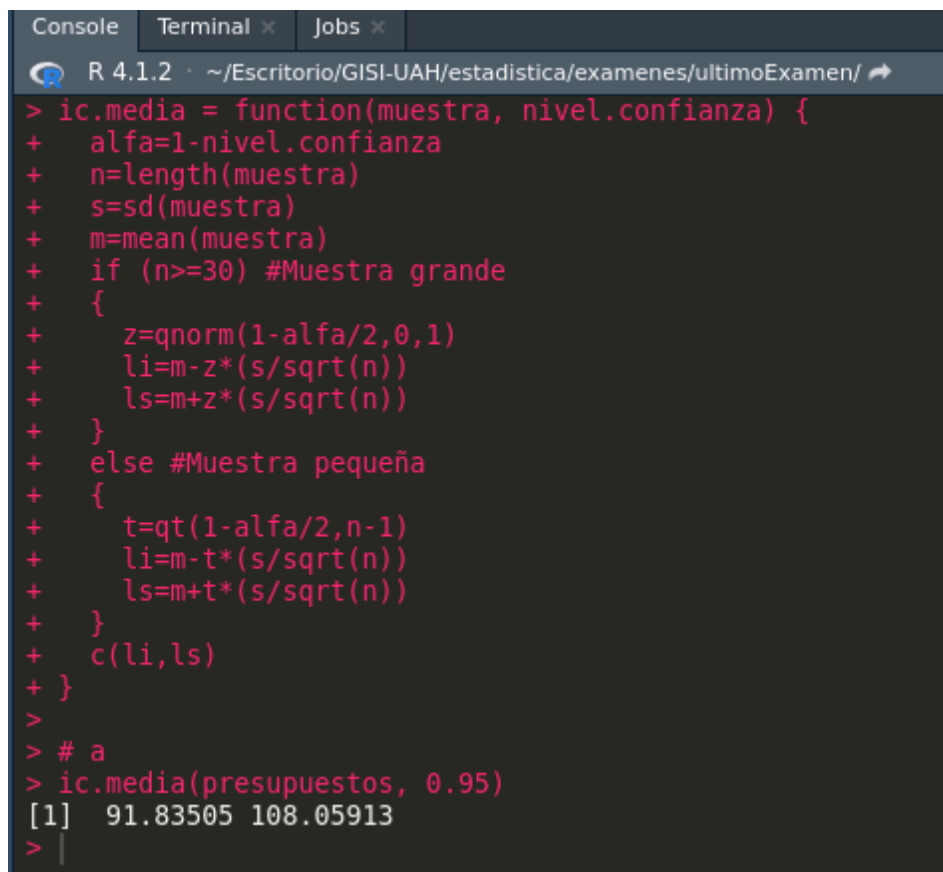


(0,2p) Nos fijamos en la variable estadística PRESUPUESTO. Suponiendo que los datos se refieren a una muestra de una población de equipos de fútbol con distribución Normal:

a) Calcular el intervalo de confianza de la media poblacional del presupuesto de los equipos de fútbol con un nivel de confianza del 95%.

NOTA: No se puede utilizar la función `t.test()` de R. Hay que aplicar la fórmula que corresponda según el tamaño de la muestra, teniendo en cuenta en esta pregunta que una muestra grande es aquella cuyo tamaño es superior o igual a 30 y que una muestra pequeña es aquella cuyo tamaño es inferior a 30. No es obligatorio crear una función para calcular el intervalo, pero si se crea una función su código R debe incluirse en la sección de ejecución de comandos R en la consola.

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:



```
Console Terminal x Jobs x
R 4.1.2 ~/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen/
> ic.media = function(muestra, nivel.confianza) {
+   alfa=1-nivel.confianza
+   n=length(muestra)
+   s=sd(muestra)
+   m=mean(muestra)
+   if (n>=30) #Muestra grande
+   {
+     z=qnorm(1-alfa/2,0,1)
+     li=m-z*(s/sqrt(n))
+     ls=m+z*(s/sqrt(n))
+   }
+   else #Muestra pequeña
+   {
+     t=qt(1-alfa/2,n-1)
+     li=m-t*(s/sqrt(n))
+     ls=m+t*(s/sqrt(n))
+   }
+   c(li,ls)
+ }
>
> # a
> ic.media(presupuestos, 0.95)
[1] 91.83505 108.05913
>
```


RESULTADOS:

$n = 79$ equipos, por lo que es una muestra **Grande** (*Indicar si es grande o pequeña*)

a)

Límite inferior del intervalo = 91.83505 millones de euros

Límite superior del intervalo = 108.05913 millones de euros

(0,2p) Nos fijamos en la variable estadística ZONA, que representa la ubicación geográfica de la ciudad de cada equipo: Norte o Sur.

a) Calcular el intervalo de confianza de la proporción poblacional de equipos que pertenecen a la zona Norte con un nivel de confianza del 95%.

b) Suponiendo que la población es de 300 equipos, inferir el mínimo y máximo número de equipos de la población que son de la zona Norte con un nivel de confianza del 95%. Redondear a valores sin ninguna cifra decimal.

NOTA: No se puede utilizar la función `prop.test()` de R. No es obligatorio crear una función para calcular el intervalo, pero si se crea una función su código R debe incluirse en la sección de ejecución de comandos R en la consola.

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:

```
Console Terminal x Jobs x
R 4.1.2 ~/Escritorio/GISI-UAH/estadistica/examenes/ultimoExamen/ ➔
> ic.proporcion = function(ng, n, nivel.confianza) {
+   alfa=1-nivel.confianza
+   p=ng/n
+   z=qnorm(1-alfa/2,0,1)
+   li=p-z*sqrt(p*(1-p)/n)
+   ls=p+z*sqrt(p*(1-p)/n)
+   c(li,ls)
+ }
> nZonas <- length(zonas)
>
> # a
> ic.proporcion(zonaNorte, nZonas, 0.95)
[1] 0.4344809 0.6541267
>
>
> # b
> n300 <- 300
> 108 * ic.proporcion(zonaNorte, n300, 0.95)
[1] 11.19757 19.76243
> |
```

RESULTADOS:

ng = 43 equipos de la zona Norte

n = 79 equipos totales de la muestra

p = _____ es la proporción de equipos de la zona Norte

a)

Límite inferior del intervalo = 0.4344809

Límite superior del intervalo = 0.6541267

b)

Mínimo (sin decimales) = 11 equipos de la población son de la zona Norte

Máximo (sin decimales) = 20 equipos de la población son de la zona Norte

(0,2p) Nos fijamos en la variable estadística PRESUPUESTO. Suponiendo que los datos se refieren a una muestra de una población de equipos de fútbol con distribución Normal:

a) Calcular el intervalo de confianza de la diferencia de las medias poblacionales de los presupuestos de los equipos de las zonas Norte y Sur con un nivel de confianza del 95%.

b) ¿Existe una diferencia significativa entre las medias de las dos poblaciones? ¿Por qué?

NOTA: No se puede utilizar la función `t.test()` de R. Son dos muestras independientes. Hay que aplicar la fórmula que corresponda según el tamaño de las muestras, teniendo en cuenta en esta pregunta que se aplica la fórmula de muestras grandes si el tamaño de cada muestra es superior o igual a 30 y la fórmula de muestras pequeñas si el tamaño de cada muestra es inferior a 30. No es obligatorio crear una función para calcular el intervalo, pero si se crea una función su código R debe incluirse en la sección de ejecución de comandos R en la consola.

EJECUCIÓN DE COMANDOS R EN LA CONSOLA:

```
> ic.dif.medias = function(muestra1, muestra2, nivel.confianza, pareados=FALSE) {  
+   alfa=1-nivel.confianza  
+   if (pareados==FALSE)  
+   {  
+     n1=length(muestra1)  
+     n2=length(muestra2)  
+     m1=mean(muestra1)  
+     m2=mean(muestra2)  
+     s1=sd(muestra1)  
+     s2=sd(muestra2)  
+     if ((n1>=30)&(n2>=30))  
+     {  
+       z=qnorm(1-alfa/2,0,1)  
+       li=m1-m2-z*sqrt((s1^2/n1)+(s2^2/n2))  
+       ls=m1-m2+ z*sqrt((s1^2/n1)+(s2^2/n2))  
+     }  
+   }
```

```

+ else

+ {

+   #f = aproximación de Welch

+   f=(((((s1^2/n1)+(s2^2/n2))^2)/((s1^2/n1)^2/(n1+1)+(s2^2/n2)^2/(n2+1))))-2

+   t=qt(1-alfa/2,f)

+   li=m1-m2-t*sqrt((s1^2/n1)+(s2^2/n2))

+   ls=m1-m2+ t*sqrt((s1^2/n1)+(s2^2/n2))

+ }

+ }

+ else

+ {

+   d=muestra1-muestra2

+   m=mean(d)

+   s=sd(d)

+   n=length(d)

+   if (n>=30)

+   {

+     z=qnorm(1-alfa/2,0,1)

+     li=m-z*(s/sqrt(n))

+     ls=m+z*(s/sqrt(n))

+   }

+   else

+   {

+     t=qt(1-alfa/2,n-1)

+     li=m-t*(s/sqrt(n))

```

```

+   ls=m+t*(s/sqrt(n))

+   }

+   }

+   c(li,ls)

+   }

>

> presupuestoSur <- yaccelga[yaccelga$ZONA == 'Sur', ]$PRESUPUESTO

> presupuestoNorte <- yaccelga[yaccelga$ZONA == 'Norte', ]$PRESUPUESTO

>

> ic.dif.medias(presupuestoSur, presupuestoNorte, 0.95)

[1] -23.959778  9.202375

>

> # b

>

> ic.1=ic.media(presupuestoSur, 0.95)

> li.1=ic.1[1]

> ls.1=ic.1[2]

> ic.2=ic.media(presupuestoNorte,0.95)

> li.2=ic.2[1]

> ls.2=ic.2[2]

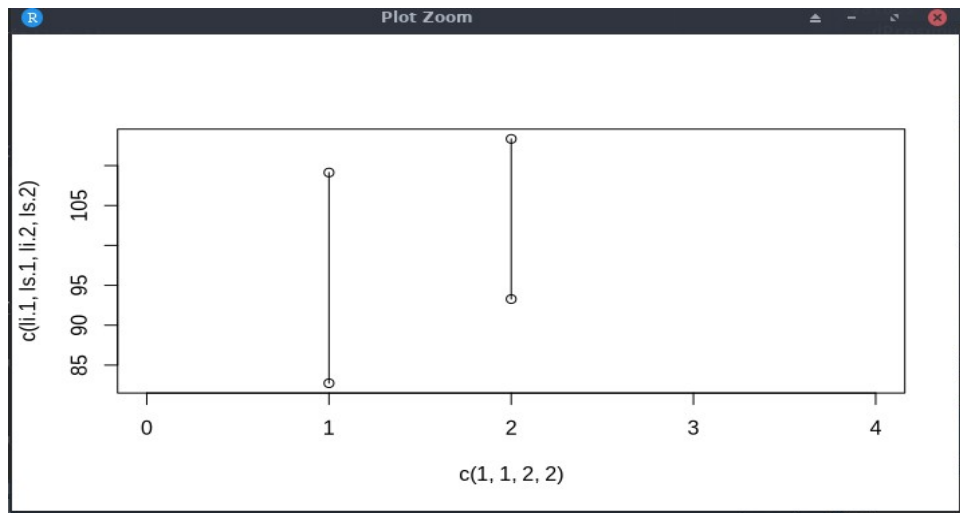
> plot(c(1,1,2,2),c(li.1,ls.1,li.2,ls.2), xlim = c(0,4))

> lines(c(1,1),c(li.1,ls.1))

> lines(c(2,2),c(li.2,ls.2))

> # No, no existe una diferencia significativa porque no se colapsan

```



RESULTADOS:

a)

Límite inferior del intervalo = -23.959778 millones de euros

Límite superior del intervalo = 9.202375 millones de euros

b)

¿Existe una diferencia significativa entre las medias de los presupuestos de los equipos del Norte y del Sur?: No (*indicar Sí o NO*)

¿Por qué?: Porque los dos intervalos no se solapan, es un buen indicio de que no va a haber coincidencia.