Universidad Simón Bolívar
Dpto. de Cómputo Científico y Estadística
CO-3321 Estadística para Ingeniería
Intensivo Julio-Agosto 2016

Laboratorio 3: Intervalos de Confianza

Estudiantes:
Alessandra Marrero,12-11091
Verónica Mazutiel, 13-10853
Profesor: Pedro Ovalles.

Sartenejas, 3 de agosto de 2016

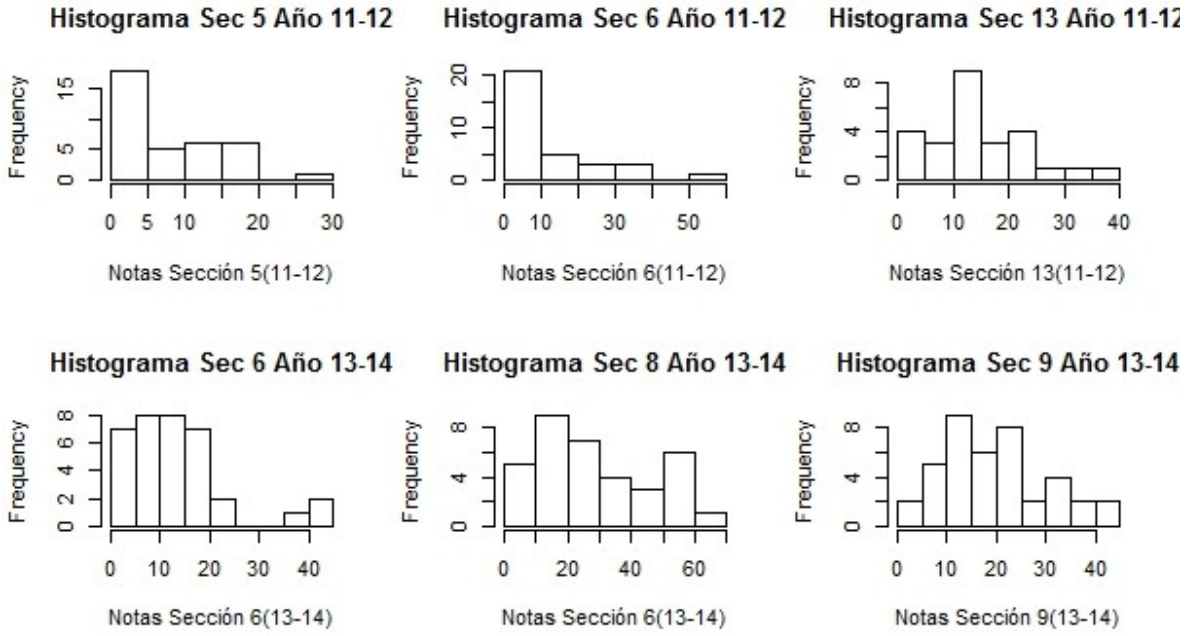
Laboratorio 3: Intervalos de confianza

Se tienen las notas del primer examen de tres secciones de un mismo curso. Los examenes son evaluados sobre 100 puntos.

1. Utilizando las gráficas vistas en clases, ¿cuáles de las secciones tienen notas que se distribuyen normal?

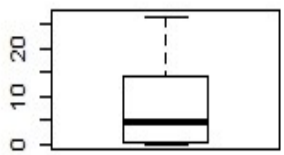
Utilizaremos los datos Notas11-12/Notas13-14

```
> Notas11_12 = read.delim("Notas11-12.txt", header=T)
> Notas12_13 = read.delim("Notas12-13.txt", header=T)
> par(mfrow=c(2,3))
> hist(Notas11_12$S5,main="Histograma Sec 5 Año 11-12",xlab="Notas Sección 5(11-12)")
> hist(Notas11_12$S6,main="Histograma Sec 6 Año 11-12",xlab="Notas Sección 6(11-12)")
> hist(Notas11_12$S13,main="Histograma Sec 13 Año 11-12",xlab="Notas Sección 13(11-12)")
> hist(Notas13_14$S6,main="Histograma Sec 6 Año 13-14",xlab="Notas Sección 6(13-14)")
> hist(Notas13_14$S8,main="Histograma Sec 8 Año 13-14",xlab="Notas Sección 6(13-14)")
> hist(Notas13_14$S9,main="Histograma Sec 9 Año 13-14",xlab="Notas Sección 9(13-14)")
```

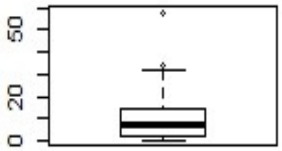


Observando los histogramas, sólo pudiéramos decir que la sección 9 para el año 13-14 tienen una distribución más o menos Normal. Las demás secciones muestran una gráfica más "movida" hacia la izquierda de lo que realmente es la normal.

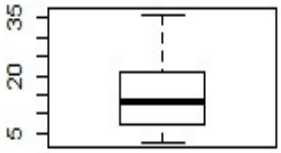
```
> boxplot(Notas11_12$S5,xlab="Notas sección 5(11-12)")
> boxplot(Notas11_12$S6,xlab="Notas sección 6(11-12)")
> boxplot(Notas11_12$S13,xlab="Notas sección 13(11-12)")
> boxplot(Notas13_14$S6,xlab="Notas sección 6(13-14)")
> boxplot(Notas13_14$S8,xlab="Notas sección 8(13-14)")
> boxplot(Notas13_14$S9,xlab="Notas sección 9(13-14)")
```



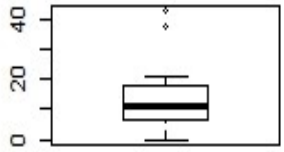
Notas sección 5(11-12)



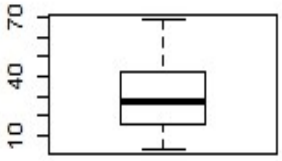
Notas sección 6(11-12)



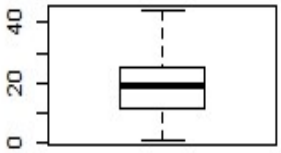
Notas sección 13(11-12)



Notas sección 6(13-14)



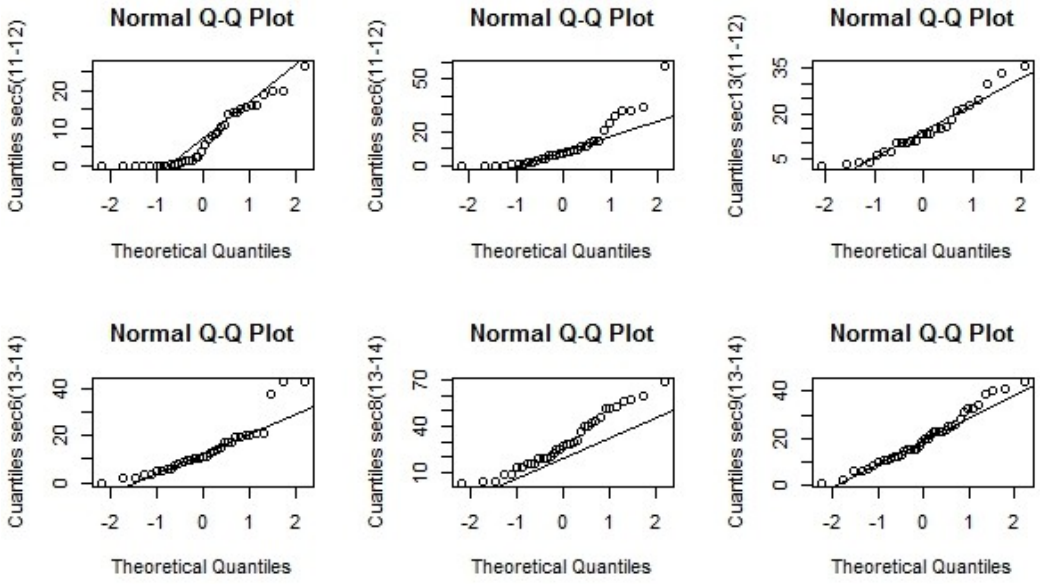
Notas sección 8(13-14)



Notas sección 9(13-14)

De igual forma, con los diagramas de caja, se podría ver que las notas de la sección 9 del año 13-14 tienen una distribución normal dado que la caja está ubicada más o menos en el centro del diagrama. Ahora se aprecia además que las notas de la sección 8 del año 13-14 también pudieran tener una distribución normal, pues la caja se encuentra cercana al centro. De los demás diagramas no se puede afirmar que se tenga una distribución normal de las notas.

```
> qqnorm(Notas11_12$S5,ylab="Cuantiles sec5(11-12)")
> qqline(Notas11_12$S5)
> qqnorm(Notas11_12$S6,ylab="Cuantiles sec6(11-12)")
> qqline(Notas11_12$S6)
> qqnorm(Notas11_12$S13,ylab="Cuantiles sec13(11-12)")
> qqline(Notas11_12$S13)
> qqnorm(Notas13_14$S6,ylab="Cuantiles sec6(13-14)")
> qqline(Notas13_14$S6)
> qqnorm(Notas13_14$S8,ylab="Cuantiles sec8(13-14)")
> qqline(Notas13_14)
> qqnorm(Notas13_14$S9,ylab="Cuantiles sec9(13-14)")
> qqline(Notas13_14$S9)
```



A pesar de que en los diagramas anteriores no se pudiera afirmar que las notas de todas las secciones tienen una distribución normal, con ayuda de qqnorm podemos ver que tanto se acercan los datos a los de una distribución normal.
Vemos que las notas de las secciones 6 y 9 del año 13-14 y las 6 y 13 del año 11-12 se acercan bastante a la línea de la normal, así que podemos asumir que tienen una distribución normal. Por otro lado, casi ningún dato de la sección 5 del año 11-12 está sobre la línea de la norma l, y ninguno de los de la sección 8 del año 13-14 está sobre dicha línea.
Sin embargo, para nuestros cálculos asumiremos que todas las notas se distribuyen normal.

2. Calcule intervalos de confianza del 97 % para la media de la nota de cada sección. ¿Qué consideraciones se deben tomar en cuenta para construir estos intervalos?

```
> intervalo=function(x,alfa) {
+   n=length(x)
+   cuantil= qnorm(1-alfa/2)
+   LS=mean(x)+ cuantil*sqrt(var(x)/n)
+   LI=mean(x)- cuantil*sqrt(var(x)/n)
+   return(c(mean(x),LI,LS))}
>
> intervalo(Notas11_12$S5[!is.na(Notas11_12$S5)],0.03)
[1]  7.361111  4.617464 10.104758
> intervalo(Notas11_12$S6[!is.na(Notas11_12$S6)],0.03)
[1] 11.674242  6.775896 16.572589
> t.test(Notas11_12$S13[!is.na(Notas11_12$S13)])
        One Sample t-test
data:  Notas11_12$S13[!is.na(Notas11_12$S13)]
t = 8.1085, df = 25, p-value = 1.836e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

10.86007 18.25531
sample estimates:
mean of x
14.55769
> intervalo(Notas13_14$S6[!is.na(Notas13_14$S6)],0.03)
[1] 13.614286 9.826452 17.402119
> intervalo(Notas13_14$S8[!is.na(Notas13_14$S8)],0.03)
[1] 29.74286 23.25049 36.23523
> intervalo(Notas13_14$S9[!is.na(Notas13_14$S9)],0.03)
[1] 19.82500 16.06699 23.58301

Solución

Intervalos de confianza para Notas11-12:

- Sección 5: [4.617464, 10.104758]
- Sección 6: [6.775896, 16.572589]
- Sección 13: [10.86007, 18.25531]

Intervalos de confianza para Notas13-14:

- Sección 6: [9.826452, 17.402119]
- Sección 8: [23.25049, 36.23523]
- Sección 9: [16.06699, 23.58301]

Se debe tener en cuenta si los datos tienen una distribución normal, como estos datos no tienen una distribución exactamente normal utilizamos la fórmula de la distribución general. Para usarla n debe ser mayor o igual a 30, lo cual se cumple en todas las secciones menos una. Esta sección tiene datos que se acercan a la distribución normal, por lo tanto se calcula el intervalo usando el comando t.test de R. Además hay que estar pendiente de que hay secciones que tienen en sus datos NA, los cuales deben de quitarse a la hora de calcular el intervalo puesto que cambian el valor de n.

3. ¿Qué conclusiones puede sacar a partir de los intervalos de confianza de sus variables? En particular, ¿cómo compararía los resultados entre secciones?

Observando los intervalos de confianza se puede concluír que en promedio los alumnos salen muy mal en el primer examen, puesto que el intervalo del promedio de notas tiene todos sus valores por debajo de 50 puntos. Esto quiere decir que en promedio los alumnos reprueban el examen.

Para comparar el resultado entre las secciones habria que calcular el intervalo de la diferencia de las medias de las secciones, el cual puede calcularse mediante el comando t.test de R. Si la diferencia no incluye al 0 en su intervalo y es negativa entonces el promedio de notas de los alumnos de la segunda sección es mayor que el de la primera utilizada para calcular el intervalo de la diferencia. Si no incluye al 0 pero es positiva entonces ocurre lo contrario, el promedio de notas de la primera sección es mayor a la segunda. Y si el intervalo incluye al 0 entonces no se puede concluír si hay diferencia entre las medias de las notas de las secciones comparadas o no.

4. Si realizan la comparación de los intervalos de confianza entre años, cómo los compararía? ¿Qué puede concluir al hacer la comparación?

Agrupamos las notas de cada año en un vector para cada año, sin distinción de sección. Buscamos un IDC para la diferencia de las medias entre los años. Para ello primero verificamos si las varianzas son iguales usando el comando var.test de R

```
> (año11_12=c(Notas11_12$S5,Notas11_12$S6,Notas11_12$S13))
  [1]  7.00 16.00  0.50  8.50  5.50  0.00  9.00 16.00 10.50  1.50 14.00  0.00  4.00
 [14]  1.50 20.00  2.50  0.00  0.00 14.00 15.00  0.00 19.75 13.50 26.50 19.00  8.00
 [27]  0.00  0.00  0.75  2.50  1.50  0.75 10.75 15.50  1.00  0.00  4.50 58.00  0.50
 [40]  8.50  0.00  1.50 25.00  3.50  6.00  8.50 11.50 29.00  7.00 34.25  4.00  9.00
 [53]  9.50 13.50  6.75  0.50  6.50  2.00 31.75 11.50  1.00 20.50 14.25  0.50 31.50
 [66]  2.50 14.25  7.00  1.00    NA    NA    NA  3.00 17.75  3.50 10.50 30.00 33.50
 [79] 22.50  3.75 36.00 10.50 13.50  7.50 15.00 11.00 13.00 10.75  6.00 24.25 13.00
 [92] 15.25 21.00 21.50 15.75  7.25  2.50 10.25    NA    NA    NA    NA    NA    NA
[105]    NA    NA    NA    NA
> (año11_12= año11_12[!is.na(año11_12)]) # Eliminamos los NA
  [1]  7.00 16.00  0.50  8.50  5.50  0.00  9.00 16.00 10.50  1.50 14.00  0.00  4.00
 [14]  1.50 20.00  2.50  0.00  0.00 14.00 15.00  0.00 19.75 13.50 26.50 19.00  8.00
 [27]  0.00  0.00  0.75  2.50  1.50  0.75 10.75 15.50  1.00  0.00  4.50 58.00  0.50
 [40]  8.50  0.00  1.50 25.00  3.50  6.00  8.50 11.50 29.00  7.00 34.25  4.00  9.00
 [53]  9.50 13.50  6.75  0.50  6.50  2.00 31.75 11.50  1.00 20.50 14.25  0.50 31.50
 [66]  2.50 14.25  7.00  1.00  3.00 17.75  3.50 10.50 30.00 33.50 22.50  3.75 36.00
 [79] 10.50 13.50  7.50 15.00 11.00 13.00 10.75  6.00 24.25 13.00 15.25 21.00 21.50
 [92] 15.75  7.25  2.50 10.25
> (año13_14=c(Notas13_14$S6,Notas13_14$S8,Notas13_14$S9))
  [1] 19.0 37.5  6.0 17.0 42.5  6.0  3.5 10.5 21.0  0.0 10.0  9.5 17.0  2.0 14.5  9.0
 [17] 20.0 11.0 20.0  5.0 14.0 17.0  7.0  3.5 10.5  2.0 42.5  5.0  9.5  8.0 12.5 11.0
 [33] 21.0 13.0 19.0   NA   NA   NA   NA   NA 46.0 68.5 19.0  9.0 16.0 51.5 40.0 53.0
 [49] 57.0 55.5 31.0 36.0 25.0 59.0 27.5 20.0 42.0  3.5 28.0 30.0 43.0 39.5 19.0 28.5
 [65] 13.0 16.0 13.0 16.0 19.0 25.0  8.5 52.0 22.5  4.5  4.0   NA   NA   NA   NA   NA
 [81] 25.0 15.5 28.0 12.0 31.5  8.0 23.5 18.0  2.0 15.0 15.0 25.0 33.0 12.0 11.5 23.0
 [97] 41.0  6.0  1.0 23.0 21.0 10.5 10.5 13.0 19.5 44.0 16.5 19.5 40.0 15.5 25.5 23.0
[113] 32.5 39.0  6.0 22.5 14.5 34.0 10.0  7.0
> (año13_14= año13_14[!is.na(año13_14)]) #Eliminamos los NA
  [1] 19.0 37.5  6.0 17.0 42.5  6.0  3.5 10.5 21.0  0.0 10.0  9.5 17.0  2.0 14.5  9.0
 [17] 20.0 11.0 20.0  5.0 14.0 17.0  7.0  3.5 10.5  2.0 42.5  5.0  9.5  8.0 12.5 11.0
 [33] 21.0 13.0 19.0 46.0 68.5 19.0  9.0 16.0 51.5 40.0 53.0 57.0 55.5 31.0 36.0 25.0
 [49] 59.0 27.5 20.0 42.0  3.5 28.0 30.0 43.0 39.5 19.0 28.5 13.0 16.0 13.0 16.0 19.0
 [65] 25.0  8.5 52.0 22.5  4.5  4.0 25.0 15.5 28.0 12.0 31.5  8.0 23.5 18.0  2.0 15.0
 [81] 15.0 25.0 33.0 12.0 11.5 23.0 41.0  6.0  1.0 23.0 21.0 10.5 10.5 13.0 19.5 44.0
 [97] 16.5 19.5 40.0 15.5 25.5 23.0 32.5 39.0  6.0 22.5 14.5 34.0 10.0  7.0
> #Comparaciones de las varianzas
> var.test(año11_12,año13_14,conf.level=0.97)

        F test to compare two variances

data:  año11_12 and año13_14
F = 0.50657, num df = 94, denom df = 109, p-value = 0.0008254
alternative hypothesis: true ratio of variances is not equal to 1
97 percent confidence interval:
 0.3290525 0.7850281
sample estimates:
ratio of variances
```

0.5065683

Dado que el intervalo no incluye el 1, entonces las varianzas son distintas, por lo que para hallar el IDC de la diferencia de medias usamos el comando t.test de R, dejando el argumento var.equal = FALSE que toma R por defecto.

```
> # Comando de R
> t.test(año11_12,año13_14, conf.level = 0.97)
        Welch Two Sample t-test
data:  año11_12 and año13_14
t = -5.7588, df = 196.13, p-value = 3.23e-08
alternative hypothesis: true difference in means is not equal to 0
97 percent confidence interval:
 -14.038152  -6.313044
sample estimates:
mean of x mean of y
 10.82895  21.00455
```

Como el intervalo de la diferencia de las medias [-14.038152, -6.313044] no incluye al 0 y además es negativo entonces se puede concluír que la media de las notas del año 13-14 es mayor a la media de las notas del año 11-12; lo que quiere decir que en promedio los alumnos salieron mejor en el primer examen del año 13-14 que el del año 11-12.