Universidad Simón Bolívar
Departamento de Cómputo Científico y Estadística
CO3321 Estadística para Ingenieros

Laboratorio 3: Intervalos de Confianza

Hecho por:
Jorge M., María Victoria
11-10495

Sartenejas, agosto de 2015

Análisis de la distribución de los datos

Para el análisis de la distribución de las variables utilizadas se utilizaron tres tipos de gráficos: histogramas, diagramas de caja y qqnorm, para este último también se utilizó qqline para agregar una línea teórica con distribución normal a la gráfica y ver qué tan agrupados alrededor de ella se encuentran los datos.

En la Figura 1 podemos ver los gráficos asociados a la variable BPIV. Como se mencionó en el laboratorio 1, en el histograma y en el diagrama de caja ya se puede observar que estos datos no siguen una distribución normal, debido a que no son gráficos simétricos. Este hecho se ratifica con la gráfica del Q-Q Norm, donde a partir de la nota 3 los datos se encuentran por encima de la línea normal y en la nota 0 existe una gran cantidad de datos lejos de esta línea.

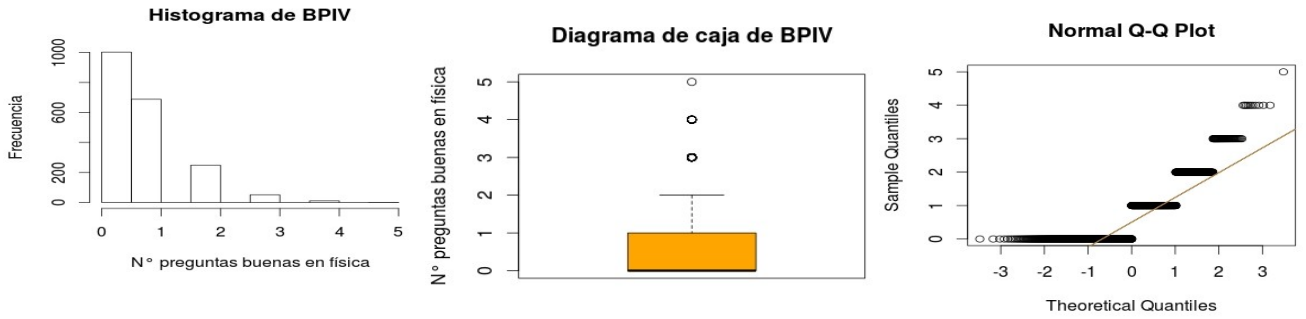

Figura 1

Por otro lado, el número de preguntas malas en la sección de física, representado por la variable MPIV, parece estar más cerca de una distribución normal. A diferencia de las preguntas buenas en esa sección, el histograma de la Figura 2 ya no es una función decreciente y el diagrama de caja tiene bigote inferior, aunque un poco más alejado del primer cuartil que el bigote superior del tercero. Sin embargo, en el Q-Q Norm los datos no parecen seguir completamente la forma de la línea recta, existiendo datos en la parte superior de la gráfica bastante dispersos. Estos datos parecen acercarse más a una distribución normal que los de las preguntas correctas, pero no forman parte de esta distribución.

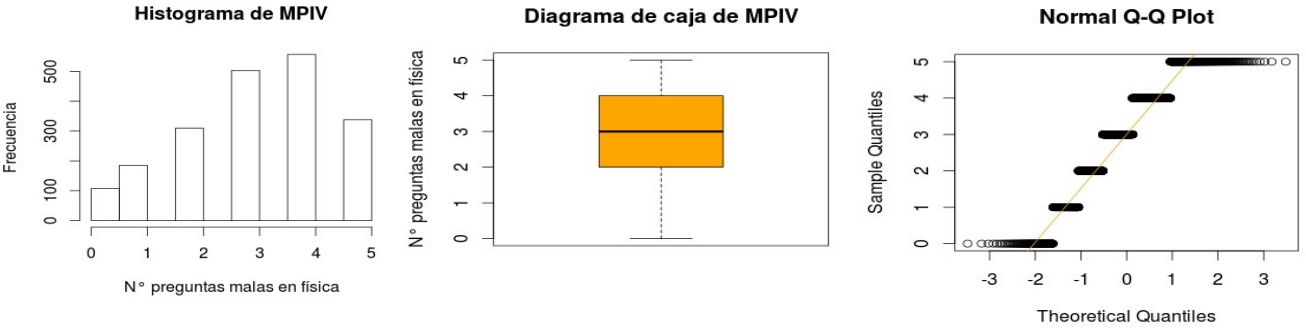

Figura 2

En la Figura 3 se encuentran los gráficos correspondientes al número de preguntas buenas en la sección de química. Al igual que en los datos de BPIV, estos claramente no son normales. El histograma no presenta una forma de campana, como lo haría una normal, sino que es una gráfica decreciente, al diagrama de caja le falta el bigote inferior y se ve una gran cantidad de datos en el área inferior del Q-Q Norm que se encuentran muy alejados de la línea recta.

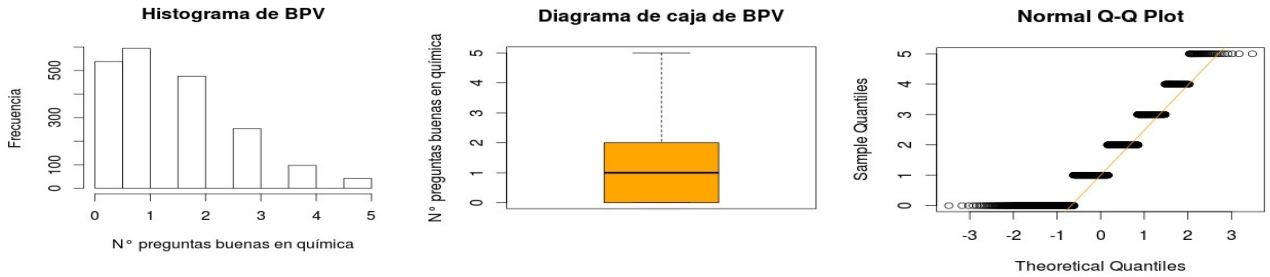

Figura 3

Al igual que en el caso de las variables BPIV y BPV, los datos del número de preguntas malas en la sección de química no siguen una distribución normal. El gráfico del histograma es decreciente, el diagrama de caja no es simétrico, le falta un bigote y la distancia de la mediana respecto a los dos cuartiles es diferente. Además en el Q-Q Norm al igual que en los casos anteriores se ve que la distribución de los datos no sigue la forma de la línea recta.

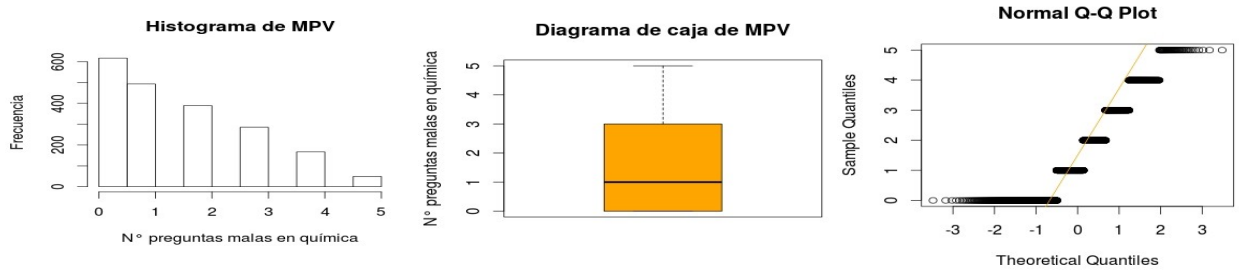

Figura 4

Finalmente, en los datos correspondientes a la nota del examen, representados por la variable NT_EX, se presenta un comportamiento parecido al de todos los datos estudiados antes. El histograma es decreciente, el diagrama de caja tiene ambos bigotes pero uno más alejado que el otro de la caja, además de presentar datos atípicos, y en el Q-Q Norm se observa que en el medio los datos parecen seguir la línea recta pero en los extremos se alejan. Por lo tanto, no siguen una distribución normal, aunque parecen ser los que están más cerca de ser normales entre todos los datos estudiados.

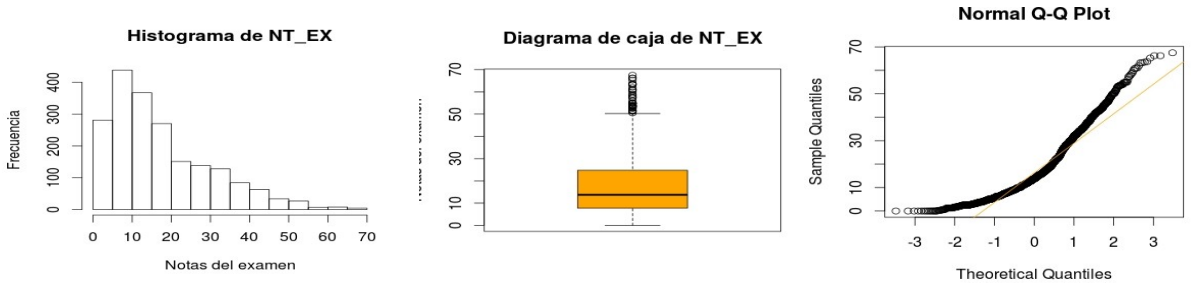

Figura 5

Intervalos de Confianza

Para cada variable se calculó un intervalo de confianza del 95%, para esto se tuvieron dos consideraciones: el tamaño de la muestra y la varianza de los datos. Como se están estudiando datos sobre 2000 estudiantes se considera que la muestra es grande, y aunque como vimos en la sección anterior los datos no

siguen una distribución normal, al ser el tamaño de la muestra grande se puede hacer una aproximación a una distribución normal. Por otro lado, la varianza es desconocida, pero con los datos puede calcularse la varianza muestral y así poder obtener los intervalos de confianza para la media de cada una de las variables por medio de la ecuación correspondiente.

Los intervalos obtenidos para cada variable son los siguientes:

- BPIV: [0.6547311, 0.7272689]
- MPIV: [3.054823, 3.177177]
- BPV: [1.396369, 1.506631]
- MPV: [1.458338, 1.580662]
- NT_EX: [16.92244, 18.05431]

De estos resultados podemos sacar varias conclusiones. Primero, para la sección de física (variables BPIV y MPIV) podemos observar que en promedio se obtuvieron más respuestas incorrectas, esto debido a que los intervalos no son solapados y el rango de MPIV contiene valores más grandes que el de BPIV. Por otro lado, en la sección de química esto no ocurre, ambos rangos están solapados, por lo que no se puede concluir si en promedio se tienen más respuestas de un tipo que de otro en química. De hecho, en promedio parece haber la misma cantidad de preguntas buenas que malas en esa sección. Además, el hecho de que los estudiantes salieron mejor en el área de química que en física mencionado en el Laboratorio 1 se evidencia también en estos intervalos de confianza. El intervalo de preguntas buenas en química no se solapa con el de física y sus valores son mayores, y en el caso de las variables que representan las preguntas incorrectas en ambas secciones pasa lo contrario, el intervalo en física es mayor que el de química, por lo que en promedio se obtuvieron peores resultados en física.

A continuación se presenta el script de R utilizado para la realización de las gráficas presentadas en este informe y del cálculo de los intervalos de confianza para cada variable estudiada.

```
# Laboratorio 3 de Estadística
# Hecho por: María Victoria Jorge 11-10495

# Variables utilizadas: BPIV (Cantidad de preguntas buenas en física 0-5),
#                       BPV (Cantidad de preguntas buenas en química 0-5),
#                       NT_EX (Nota del examen 0-90),
#                       MPIV (Cantidad de preguntas malas en física 0-5) y
#                       MPV (Cantidad de preguntas malas en química 0-5)

datos = read.table("CLargas.txt", header=T)
attach(datos)
names(datos)

# Evaluamos si BPIV se distribuye normal
par(mfrow=c(1,1))
hist(datos$BPIV,main="Histograma de BPIV",xlab = "N° preguntas buenas en física",ylab = "Frecuencia")
boxplot(datos$BPIV,main="Diagrama de caja de BPIV",col = "orange",ylab="N° preguntas buenas en física")
qqnorm(datos$BPIV)
qqline(datos$BPIV, col="orange")

# Evaluamos si MPIV se distribuye normal
par(mfrow=c(1,1))
hist(datos$MPIV,main="Histograma de MPIV",xlab = "N° preguntas malas en física",ylab = "Frecuencia")
boxplot(datos$MPIV,main="Diagrama de caja de MPIV",col = "orange",ylab="N° preguntas malas en física")
qqnorm(datos$MPIV)
```

```
qqline(datos$MPIV, col="orange")

# Evaluamos si BPV se distribuye normal
par(mfrow=c(1,1))
hist(datos$BPV,main="Histograma de BPV",xlab = "N° preguntas buenas en química",ylab = "Frecuencia")
boxplot(datos$BPV,main="Diagrama de caja de BPV",col = "orange",ylab="N° preguntas buenas en química")
qqnorm(datos$BPV)
qqline(datos$BPV, col="orange")

# Evaluamos si MPV se distribuye normal
par(mfrow=c(1,1))
hist(datos$MPV,main="Histograma de MPV",xlab = "N° preguntas malas en química",ylab = "Frecuencia")
boxplot(datos$MPV,main="Diagrama de caja de MPV",col = "orange",ylab="N° preguntas malas en química")
qqnorm(datos$MPV)
qqline(datos$MPV, col="orange")

# Evaluamos si NT_EX se distribuye normal
par(mfrow=c(1,1))
hist(datos$NT_EX ,main="Histograma de NT_EX",xlab = "Notas del examen",ylab = "Frecuencia")
boxplot(datos$NT_EX,main="Diagrama de caja de NT_EX",col = "orange",ylab="Notas del examen")
qqnorm(datos$NT_EX)
qqline(datos$NT_EX, col="orange")


# Intervalos de Confianza
# Función que calcula el IDC del 100(1-alfa)% para la media de una muestra x
intervalo.med = function(x,alfa){
  n = length(x)
  z = qnorm(alfa/2,lower.tail = F)
  limS = mean(x) + z*sqrt(var(x)/n) # Límite superior del IDC
  limI = mean(x) - z*sqrt(var(x)/n) # Límite inferior del IDC
  return (c(limI,limS))
}

# IDC del 95% para BPIV
intervalo.med(datos$BPIV,0.05)

# IDC del 95% para MPIV
intervalo.med(datos$MPIV,0.05)

# IDC del 95% para BPV
intervalo.med(datos$BPV,0.05)

# IDC del 95% para MPV
intervalo.med(datos$MPV,0.05)

# IDC del 95% para NT_EX
intervalo.med(datos$NT_EX,0.05)
```