Enseñanza de la Estadística usando el software libre R

Dr. Víctor H. Soberanis Cruz

Área de matemáticas, Departamento de Ciencias,
DCI- UQROO

El objetivo de este trabajo mostrar una secuencia didáctica para la presentación del tema Introducción al Análisis Exploratorio de Datos (EDA) usando el software libre R.

ANTECEDENTES

- Los avances del Cálculo de Probabilidades llevaron a la creación de la Estadística teórica, que en cierto modo, se alejo de las ideas estadísticas primitivas centradas en la recogida y análisis de los datos (Cabriá, 1994).

- Recordamos que a partir de la década de los años 60’s, la mayor parte de los textos, como el de Mood y Graybill, el de Wilks, el de Lindgreen, el de Ross etc., se ocupaban especialmente de los modelos clásicos o bayesianos con respecto a conjuntos simples de datos y hubo una matematización de la Estadística, junto con un descuido en la enseñanza de los aspectos prácticos del análisis de datos (Batanero, 2001).

- Como consecuencia de esto y el espectacular desarrollo de la informática en la segunda mitad del siglo XX, en las últimas décadas se han desarrollado tipos de análisis de datos que se sitúan entre la Estadística Descriptiva y la Inferencia Estadística. Entre estos se encuentra el EDA desarrollado por Tukey entre 1960 y 1980.

Análisis exploratorio de datos

•Esencialmente la filosofía del EDA parte del hecho de que un conjunto de datos está constituido por dos partes: la "tendencia" y las "desviaciones".

• Por tendencia entendemos una estructura simplificada de las observaciones (en puntos con tendencia decreciente, por ejemplo, una distribución exponencial ajustada a los datos).

• Las diferencias de los datos con respecto a esta estructura son las desviaciones de los datos, que no necesariamente presentan alguna estructura determinada.

- Tradicionalmente el estudio se ha concentrado en la búsqueda de un modelo que explique la tendencia de los datos.

- ***El EDA por otra parte, desglosa las observaciones en las dos partes antes mencionadas y en lugar de imponer una hipótesis, es decir, imponer un modelo a las observaciones, se genera dicho modelo a partir de los datos mismos.***

- Así pues, el Análisis exploratorio de datos construye a partir de los datos mismos un modelo que en efecto generen esas observaciones.
- De este modo el EDA no solamente le da su debido lugar al cálculo de estadísticos como las medias, varianzas, coeficientes de correlación, etc., ***sino que ha de aumentar la importancia visual de las representaciones de los datos y del desarrollo del razonamiento estadístico.***

BASES TEÓRICAS

- Esta propuesta didáctica, para efectos también de un desarrollo del razonamiento estadístico, se sustenta en el modelo “Estructura del proceso de Aprendizaje Observado” (SOLO, por sus siglas en Inglés) propuesto por Biggs y Collins en 1982.

- SOLO nos permite caracterizar el razonamiento estadístico en cuatro niveles jerárquicos: preestructural, uniestructural, multiestructural y el relacional.

- ***Para la comprensión de gráficas nos apoyaremos en el modelo de Curcio (1989), el cual consta de tres niveles: leer los datos, leer dentro de los datos y leer más allá de los datos***.

Preguntas para esta secuencia didáctica

- ¿Qué tipos de tareas ayudan a los estudiantes a desarrollar el razonamiento estadístico ?
- ¿Cómo pueden los estudiantes construir conocimiento estadístico?
- ¿A qué nivel el uso de software R favorece las aproximaciones de los estudiantes en la descripción de un conjunto de datos e inferir hacia una población?

El siguiente conjunto de datos corresponde a las calificaciones de la asignatura de Matemáticas I de tres secciones en una Escuela Particular de Nivel Medio Superior . Los datos ya están capturados en R.

```
Seccion1<-
c(82,45,89,82,67,72,64,89,93,78,87,75,115,57,86,73,86,85,
82,90,104,64,83,77,83,78,81,96,62,
            77,53,113,67,103,39)
Seccion2<-
c(99,87,72,81,88,82,66,71,88,58,84,68,86,70,88,91,71,108,
109,73,81,96,60,92,85,104,98,104,57,
            25,96,74,74,72,96,88,84,62)
Seccion3<-
c(58,46,72,84,74,48,116,91,69,53,65,109,91,69,69,86,45,48
,61,70,84,96,63,90)
```

•Los siguientes códigos en R nos permiten obtener el histograma convencional para cada Sección así como el histograma para el conjunto de las tres Secciones considerada como un solo conjunto de datos. En el mismo programa tenemos los códigos para obtener los diagrama de Tallo y Hojas, diseñado por Tukey en su EDA, para cada Sección y para las tres Secciones consideradas como un único conjunto de datos.

•Las salidas de R nos permitirán, entre otras cosas, identificar las características del diagrama de Tallos y Hojas que le deben de dar una potencia exploratoria mayor que el histograma convencional.

• Podemos ver como los tres niveles de comprensión de gráficas del modelo de Curcio, son más fáciles de alcanzar con los diagramas de Tallo y Hojas.

PROGRAMA EN R PARA LA OBTENCION DE LOS HISTOGRAMAS CONVENCIONALES Y LOS DIAGRAMAS DE TALLO Y HOJA EN EL ANALISIS EXPLORATORIO DE DATOS.

```
1. Secciones<-c(Seccion1,Seccion2,Seccion3)
2. par(mfrow=c(2,2))
3. Factor<-
c(rep("1",length(Seccion1)),rep("2",length(Seccion2)),rep(
"3",length(Seccion3)))
4. tapply(Secciones,Factor,hist)
5. hist(Secciones)
6. tapply(Secciones,Factor,stem)
7. stem(Secciones)
8. par(mfrow=c(1,1))
9. boxplot(Seccion1,Seccion2,Seccion3)
10. boxplot(Secciones)
```

LA INTERPRETACIÓN DE LAS INSTRUCCIONES 1-10 EN CODIGO R SE LE HACE A LOS ESTUDIANTES PREVIAMENTE. LOS SIGUIENTES HISTOGRAMAS DE IZQUIERDA A DERECHA Y DE ARRIBA HACIA ABAJO CORRESPONDEL A LAS SECCIONES 1,2 Y 3, Y EL DE LA ESQUINA INFERIOR DERECHA A LAS TRES SECCIONES COMO UN SOLO GRUPO.

R

TAREA1. A partir de los histogramas

a) Identifique el intervalo de clase donde hay mayor frecuencia de calificaciones

b) ¿Cuál es la calificación que divide a las calificaciones de la Sección en 50%-50%?

c) ¿Cuál es la calificación promedio de la Sección?

d) ¿ Cual son las calificaciones que parten a las calificaciones de la Sección 0%-25%, 25%-50%, 75%-100%?

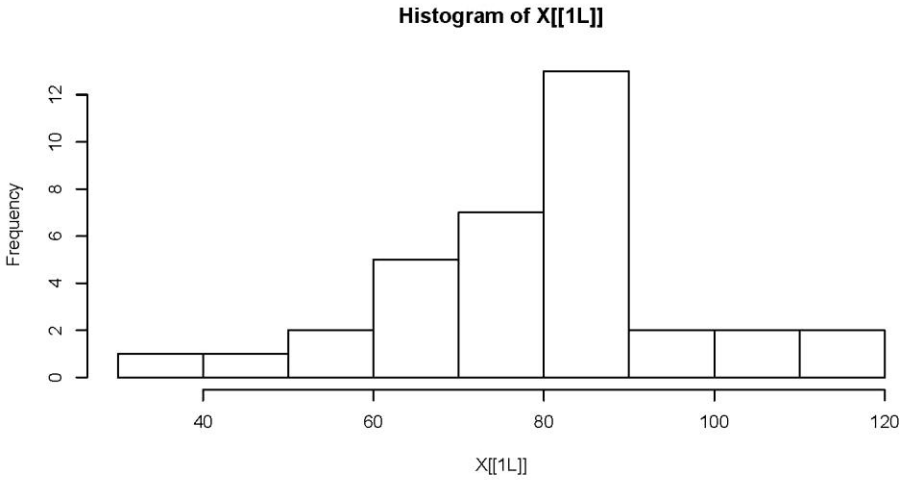
e) ¿Detecta alguna calificación “fuera de lo común”(outliers)?,¿Qué pudo haber ocurrido?

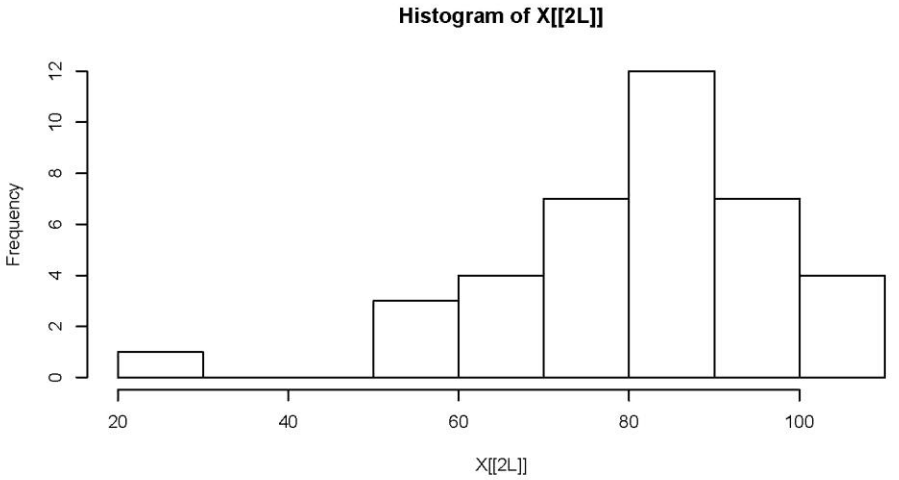
f) ¿ En cual Sección los estudiantes son más “parejos”?

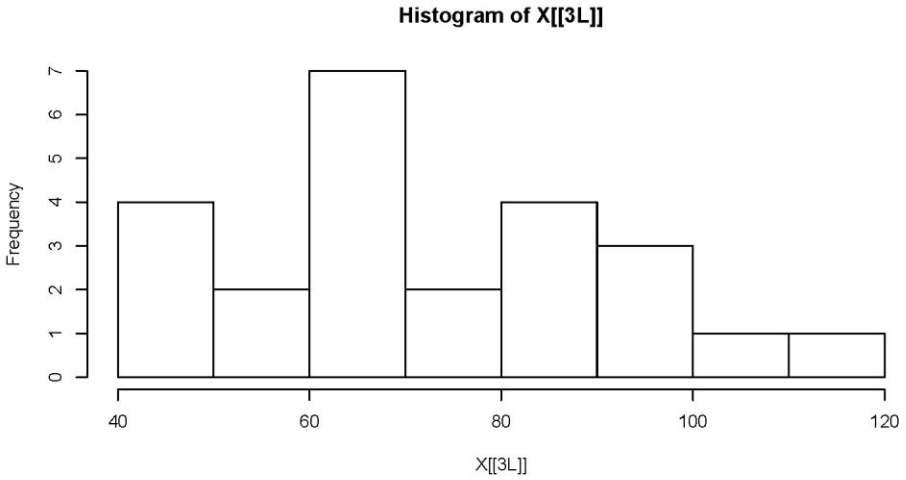
g) ¿Cuál Sección es la más aplicada?

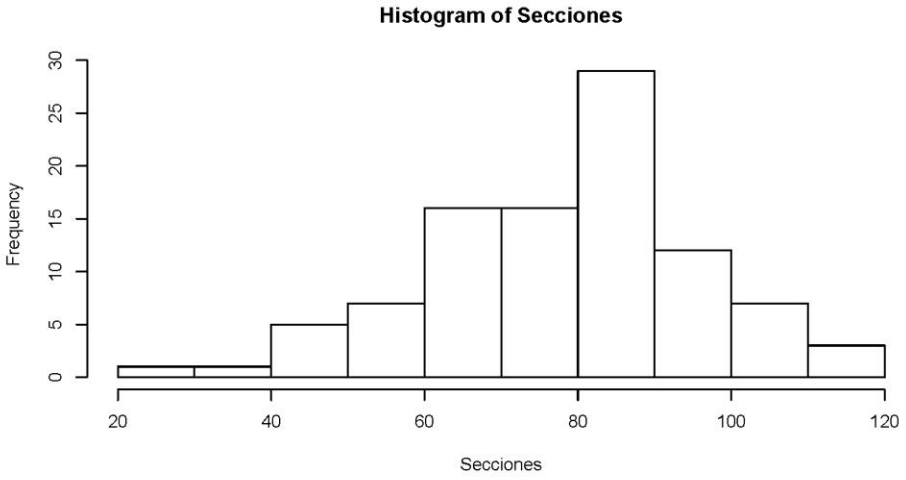
h) ¿ Por que crees que es la más aplicada?

i) ¿Qué propondrías para decir que las comparaciones entre las Secciones son justas?


Histogram of X[[1L]]
Frequency
0
2
4
6
8
10
12
40
60
80
100
120
X[[1L]]


Histogram of X[[2L]]
Frequency
0
2
4
6
8
10
12
20
40
60
80
100
X[[2L]]


Histogram of X[[3L]]
Frequency
0
1
2
3
4
5
6
7
40
60
80
100
120
X[[3L]]


Histogram of Secciones
Frequency
0
5
10
15
20
25
30
20
40
60
80
100
120
Secciones

TAREA2. A partir de los diagramas de Tallo y Hoja (siguiente diapositiva)

a) Identifique el intervalo de clase donde hay mayor frecuencia de calificaciones

b) ¿Cuál es la calificación que divide a las calificaciones de la Sección en 50%-50%?

c) ¿Cuál es la calificación promedio de la Sección?

d) ¿ Cual son las calificaciones que parten a las calificaciones de la Sección en 0%-25%, 25%-50%, 75%- 100%?

e) ¿Detecta alguna calificación “fuera de lo común”(outliers)?,¿Qué pudo haber ocurrido?

f) ¿ En cual Sección los estudiantes son más “parejos”?

g) ¿Cuál Sección es la más aplicada?

h) ¿ Por que crees que es la más aplicada?

i) ¿Qué propondrías para decir que las comparaciones entre las Secciones son justas?

3 | 9
4 | 5
5 | 37
6 | 24477
7 | 2357788
8 | 122233566799
9 | 036
10 | 34
11 | 35

Sección1

2 | 5
3 |
4 |
5 | 78
6 | 0268
7 | 01122344
8 | 112445678888
9 | 1266689
10 | 4489

Sección2

2 |
3 |
4 | 5688
5 | 38
6 | 135999
7 | 024
8 | 4460
9 | 116
10 | 9
11| 6

Sección3

2 | 5
3 | 9
4 | 55688
5 | 337788
6 | 012234456778999
7 | 00112222344457788
8 | 11122223344445566677888899
9 | 00111236666689
10 | 3444899
11 | 356

Las tres Secciones

TAREA3. A partir de los diagramas de Caja (siguiente diapositiva)

a) ¿Cuál es la calificación que divide a las calificaciones de la Sección en 50%-50%?

b) ¿Cuál es la calificación promedio de la Sección?

c) ¿ Cual son las calificaciones que parten a las calificaciones de la Sección en 0%-25%, 25%-50%, 75%- 100%?

d) ¿Detecta alguna calificación “fuera de lo común”(outliers)?,¿Qué pudo haber ocurrido?

e) ¿ En cual Sección los estudiantes son más “parejos”?

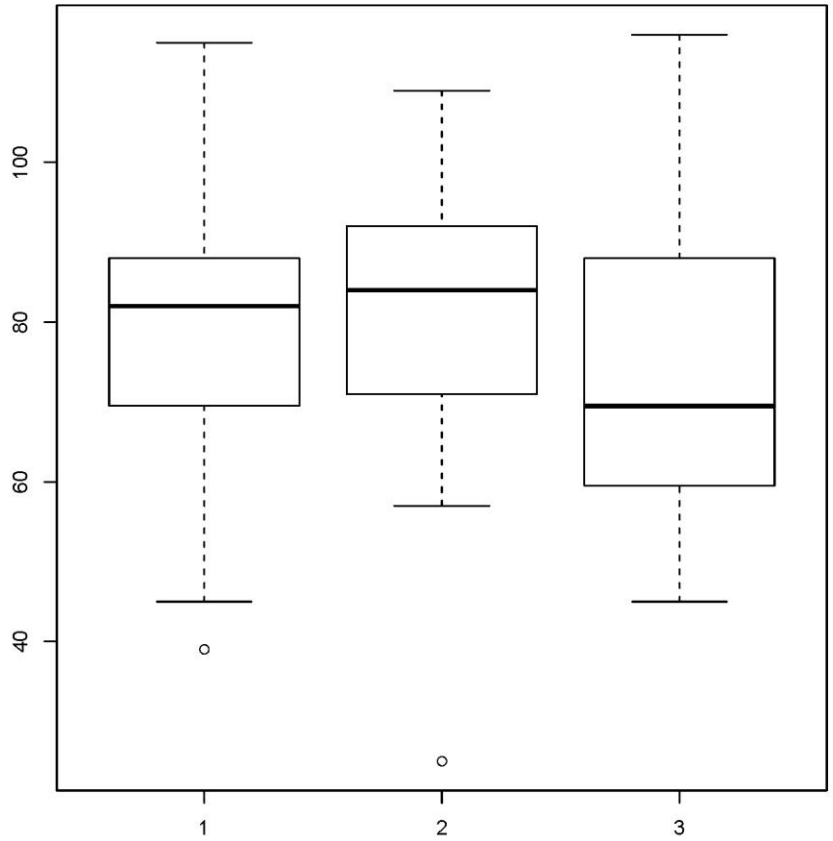
f) ¿Cuál Sección es la más aplicada?

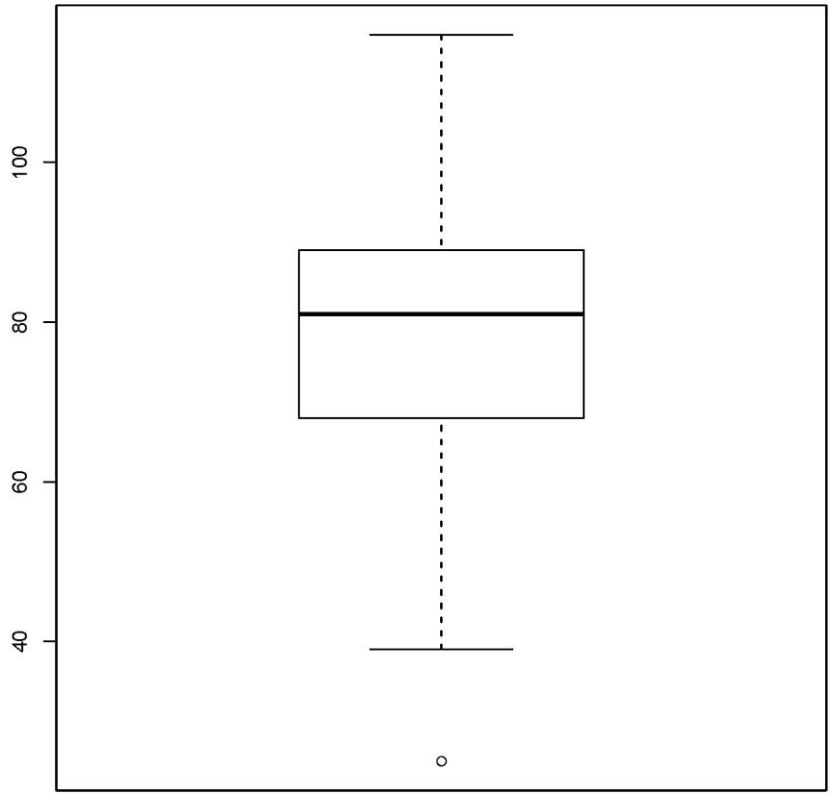
g) ¿ Por que crees que es la más aplicada?

h) Si únicamente tuviésemos el diagrama de caja de la derecha, ¿ Podría decir algo sobra cada Sección o de alguna Sección o comparar dos Secciones?

i) Si tuviésemos las tres cajas de la izquierda y no la de la derecha, ¿ Podríamos decir algo de las tres Secciones como un solo grupo de datos?

El diagrama de caja como una herramienta para medir el razonamiento estadístico en sus cuatro niveles jerarquicos: preestructural, uniestrucural, multiestructural y relacional





Bibliografía:

Batanero, C. (2001). *Didáctica de la Estadística.* Grupo de Investigación en Educación Estadística. Departamento de Didáctica de la Matemática. Universidad de Granada.

Biggs, J.B. y Collins, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Cursio, F.R. (1989). *Developing graph comprehention.* Reston, VA: NCTM