Descriptive Statistics

September 5, 2013

Outline

- **Qualitative Data**
- **Quantitative Data**
- **Updating R Graphs**

R Graphics- Qualitative

The **table** command allows us to look at tables. Its simplest usage looks like **table(x)** where x is a categorical variable.
Example: A survey asks people if they smoke or not.
Yes, No, No, Yes, Yes
We can enter this into R with the c() command, and summarize with the table command as follows

```
> x=c("Yes","No","No","Yes","Yes")
> table(x)
```

Note: The **table** command simply adds up the frequency of each unique value of the data.

Example

The Final grades of 25 students in MA 345 is given as below:
C,D,D,D,C,D,C,C, A,C,B, A,B, A,B,C,B,C, A, A,C, A,D,C,A
We will create a table as

```
> x<-c("C","D","D","D","C",'D','C','C', "A","C","B", "A",
"B", "A","B","C","B","C", "A", "A","C", "A","D","C","A")
> table(x)
x
A B C D
7 4 9 5
```

Data frame called **hsb2** has categorical variable race, which has four levels (1 = Hispanic, 2 = Asian, 3 = African American and 4 = Caucasian). Similarly, gender is coded 1 for female and 0 for male. We can access the data and create a table.

```
example=read.table('http://www.ats.ucla.edu/stat/data/hsb2.csv', header=T, sep=",")
attach(example)

tab = table(example$race)
```

factor

Categorical data is often used to classify data into various levels or factors. For example, the smoking data could be part of a broader survey on student health issues. R has a special class for working with factors which is occasionally important to know as R will automatically adapt itself when it knows it has a factor. To make a factor is easy with the command **factor** or **as.factor.**

```
>x=c("Yes","No","No","Yes","Yes")
>factor(x)
```

Bar Graph

A bar chart draws a bar with height proportional to the count in the table. The height could be given by the frequency, or the proportion. A bar graph can be drawn for frequency data or relative frequency data

```
> grade<-c(12,20,10,3,5)
> names(grade)=c("A", "B", "C", "D", "F")
>barplot(grade,col=c("1","2","3","4","5"),
main="Grade Distribution")
```

```
> grade<-c(12,20,10,3,5)
> names(grade)=c("A", "B", "C", "D", "F")
> barplot(grade/(sum(grade)),col=c("1","2","3","4","5"),
main="Grade Distribution",xlab="Grade", ylab="Relative frequency
```

Pie Chart

Pie charts A circle divided into sectors that represent the percentages of a population or a sample that belongs to different categories is called a pie chart. To construct a pie chart, we multiply by 360 the relative frequency for each category to obtain the degree measure or size of the angle for the corresponding category.

```
>students<-c(120, 200, 350, 230,125)
>names(students)=c("Maths","Science","Engineering",
"Education","Technology")
>pie(students,col=c("1","2","3","4","5"),
main="Student Enrollment")
```

Histogram

A histogram is a graphic that gives an idea of the "shape" of a sample, indicating regions where sample observations are concentrated and regions where they are sparse. A histogram is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. The number of classes suggested by Sturge's formula is $c = 1 + \log_2(n)$ where n is the number of observations in the data.

```
> x<-c (25,37,20,31,31,21,12,25,36,27,38,16,40,32,33,24,39,26,27,19)
> hist(x)

> x<-c (25,37,20,31,31,21,12,25,36,27,38,16,40,32,33,24,39,26,27,19)
>hist(x,main="Histogram of Chicago Temp.",xlab="Temperature", ylab="# of Days", c=3)

>hist(x, breaks=15) it suggests 10 breaks
>hist(x, breaks = c(10, 20, 30, 40))# uses these breaks
>hist(x, breaks="scott") Use "Scott" algorithm
```

Plotting Density Curve

The problems associated with drawing histograms and density functions of continuous variables are much more challenging. The subject of density estimation is an important issue for statisticians. You can get a feel for what is involved by browsing the **?density** help window. The algorithm used in **density.default** disperses the mass of the empirical distribution function over a regular grid. The choice of bandwidth is a compromise between smoothing enough to rub out insignificant bumps and smoothing too much so that real peaks are eliminated. The bandwidth is given as

$$b = \frac{max(x) - min(x)}{2(1 + \log_2 n)}$$

Example

We will access faithful data from the data set and plot density curve for eruptions.

```
>attach(faithful)
>(max(eruptions)-min(eruptions))/(2*(1+log(length(eruptions),base=2)))

> par(mfrow=c(1,2))
> hist(eruptions,15, freq=FALSE,main="Density Curve",col=2)
> lines(density(eruptions,width=0.6,n=200))
> truehist(eruptions,nbins=15, main="Density Curve",col=3)
> lines(density(eruptions,n=200))
```

Note that **truehist** is in the package **MASS**

Graphic Parameters, *par()*

The function **par()** is used to set or get graphical parameters. **par** allows us to plot multiple (x, y)'s in a single graphic. This is accomplished by selecting par(new=T) following each call to plot.

```
x<-seq(-5,5,0.1)
y1<-dnorm(x)
y2<-dcauchy(x)
y3<-0.5*dexp(abs(x))
yrange<-range(y1,y2,y3)
plot(x,y1,xlab="x",ylab="f(x)",lty=1, type="l",xlim=c(-5,5),ylim=yrange,col=1)
par(new=TRUE)
plot(x,y2,xlab="",ylab="",lty=3,type="l",xlim=c(-5,5),ylim=yrange,col=2)
par(new=TRUE)
plot(x,y3,xlab="",ylab="",lty=2,type="l",xlim=c(-5,5),ylim=yrange,col=4)
legend(1,.5,legend=c("N(0,1)","C(0,1)","L(0,1)"),lty=c(1,3,2),col=c(1,2,4))
title(cex=1,"probability density functions of standard Normal, standard Cauchy and
\n standard Laplace distributions")
```

Stem and Leaf Plot

Stem-and-leaf plot is a simple way of summarizing quantitative data. In a stem-and-leaf plot each data value is split into a "stem" and a "leaf". The "leaf" is usually the last digit of the number and the other digits to the left of the "leaf" form the "stem". Usually there is no need to sort the leaves, although computer packages typically do.

```
> x<-c(78,74,82,66,94,71,64,88,55,80,91,74,82,75,96,
78,84,79,71,83)
> stem(x)

The decimal point is 1 digit(s) to the right of the |
5 | 5
6 | 46
7 | 11445889
8 | 022348
9 | 146
```

Boxplot

A boxplot is a way of summarizing a set of data measured on an interval scale. It is often used in exploratory data analysis. It is a type of graph which is used to show the shape of the distribution, its central value, and variability. The picture produced consist five number summaries. The median for each dataset is indicated by center line, and the first and third quartiles are the edges of the box. The extreme values (within 1.5 times the inter-quartile range from the upper or lower quartile) are the ends of the lines extending from the IQR. Points at a greater distance from the median than 1.5 times the IQR are plotted individually as asterisks. These points represent potential outliers.

```
>x=c(24,58,61,67,71,73,76,79,82,83,85,87,88,88,92,93,94,97)
>boxplot(x, main="Boxplot of test scores", col=2)
> arrows(1,24,1.2,30)
> text(1.4,31,"This is an Outlier")
```

Example:

Link below provides the number of Atlantic hurricane from 1870 to 2010 http://biostatistics.it/Didattica/Dati/SilwoodWeather.txt
We will import the subject data and plot a boxplot for monthly data

```
>temperature="http://biostatistics.it/Didattica/Dati/SilwoodWeat
>weather=read.table(temperature,header=T)
>attach(weather)
> names(weather)
[1] "upper" "lower" "rain" "month" "yr"
```

Before we can plot the data we need to declare month to be a factor. At the moment, R just thinks it is a number.

```
>month<-factor(month)
>plot(month,upper)
```

Saving Graphs

Since R runs on so many different operating systems, and supports so many different graphics formats, it's not surprising that there are a variety of ways of saving your plots, depending on what operating system you are using, what you plan to do with the graph, and whether you're connecting locally or remotely.

Format	Driver
JPG	jpeg
PNG	png
WMF	win.metafile
PDF	pdf
Postscript	postscript