What is Data

Tatjana Kecojević

2016-11-23

Intorduction

There are many situations in modern business and science where data is collected and analysed. The key ideas of data analysis are important in understanding the information provided by such data. In this section we will look into a set of methods to enable data to be explored with the objective of summarising and understanding the main features of the variables contained within the data.

We will start by defining the population. The **population** is the set of all people/objects of interest in the study being undertaken. Usually populations are very large, and in some cases may be conceptual in the sense that they cannot be completely enumerated physically. The majority of data analysis is carried out on a **sample** drawn from the population, and the fundamental problem is to use sample data to draw inferences about the population.

In statistical terms the whole data set is called the population. This represents *perfect information* however in practice it is often impossible to enumerate the whole population. The analyst therefore takes a sample drawn from the population and uses this information to make judgements (inferences) about the population.

Clearly if the results of any analysis are based on a sample drawn from the population, then if the sample is going to have any validity, then the sample should be chosen in a way that is fair and reflects the structure of the population. The process of sampling to obtain a representative sample is a large area of statistical study. The simplest model of a representative sample is a **random sample**, a sample chosen in such a way that each item in the population has an equal chance of being included in the sample. As soon as sample data is used, the information contained within the sample is *imperfect* and depends on the particular sample chosen. The key problem is to use this sample data to draw valid conclusions about the population with the knowledge of and taking into account the *error due to sampling.*

Usually the data will have been collected in response to some design problem, in the hope of being able to glean some pointers from this data that will be helpful in the analysis of the problem. Data is commonly presented to the data analyst in this way with a request to analyse the data.

Before attempting to analyse any data, the analyst should:

- Make sure that the problem under investigation is clearly understood, and that the objectives of the investigation have been clearly specified. The only way to obtain this information is to ask questions, and keep asking questions until satisfactory answers have been obtained.
- Before any analysis is considered the analyst should make sure that the individual variables making up the data set are clearly understood.

A starting point is to examine the characteristics of each individual variable in the data set. The way to proceed depends upon the type of variable being examined.

The variables can be one of two broad types:

1) **Attribute variable**: has its outcomes described in terms of its characteristics or attributes;
2) **Measured variable**: has the resulting outcome expressed in numerical terms.

Statistical Distribution

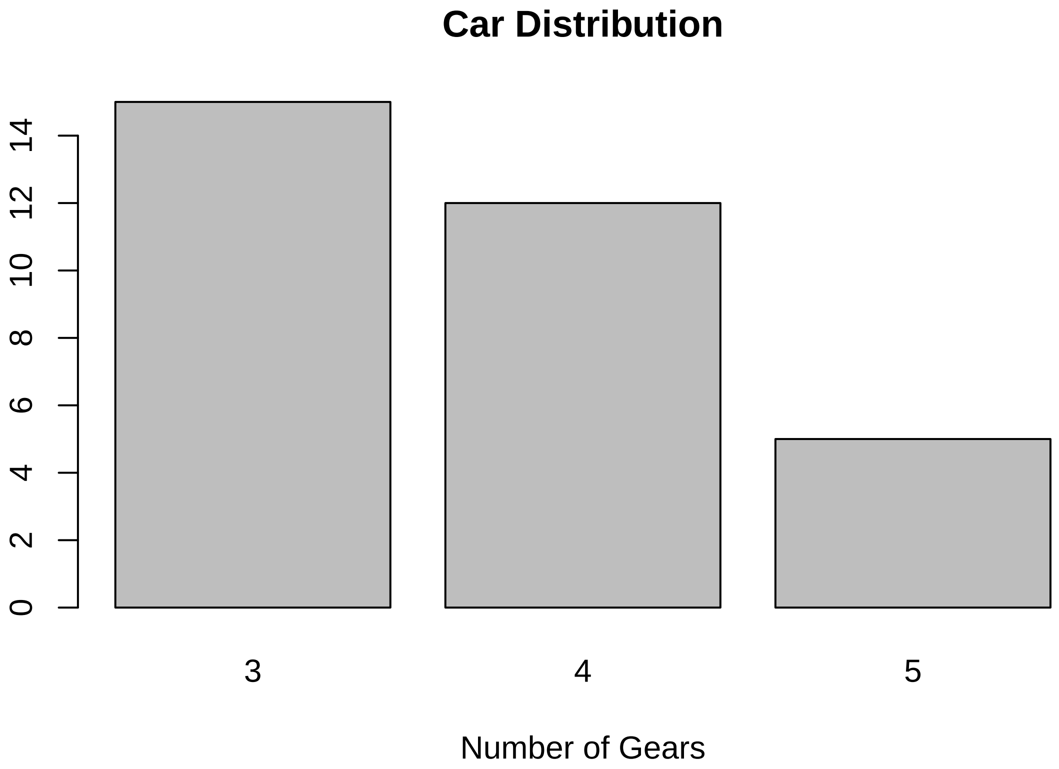
The concept of the **statistical distribution** is central to statistical analysis. This concept relates to the population and conceptually assumes that we have perfect information, the exact composition of the population is known. However, as you are most likely to deal with the sample data you will be looking at sample distribution, based on which you will be drawing conclusions about the population.

If you want to look at the distribution of an attribute variable, you will look at the frequency of occurrence of each level using a bar chart. Let us look at *mtcars* data and the distribution of attribute variable *gear*:

```r
summary(mtcars)
```

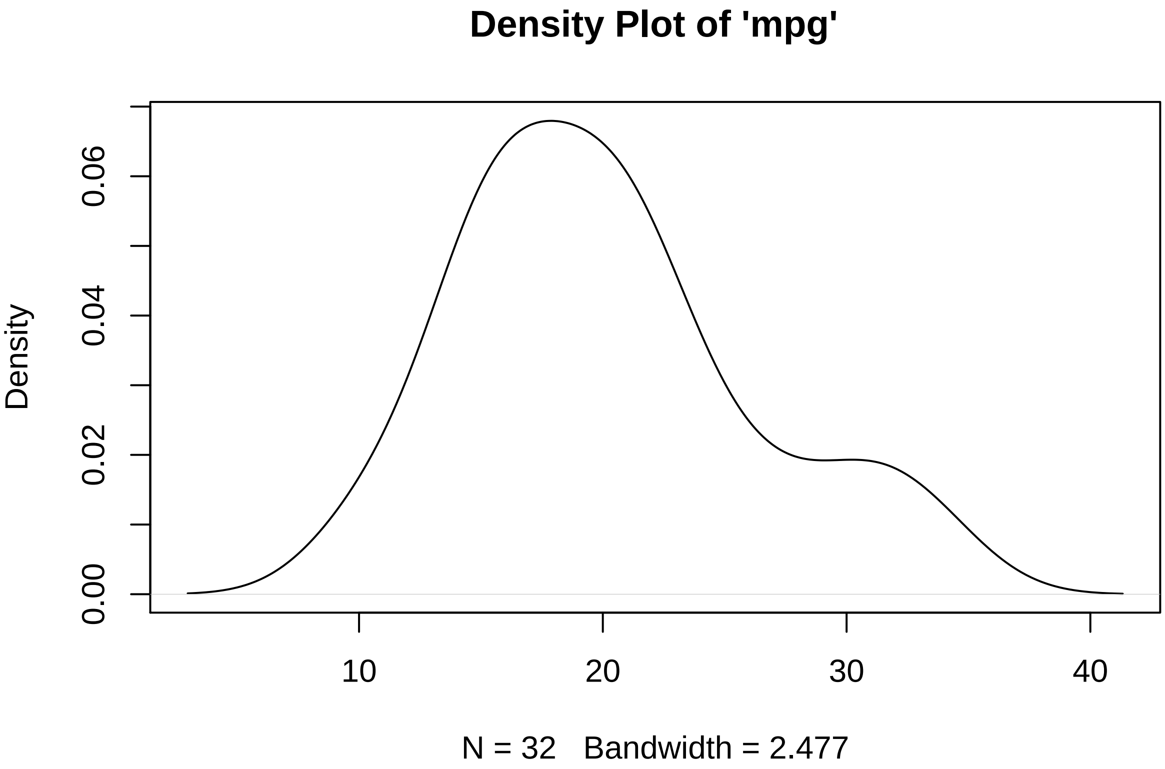
```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt              qsec              vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear             carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```r
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution", xlab="Number of Gears")
```



If the variable under discussion is a measured type then distribution of this variable across its possible range of values may look like in Figure below, which illustrates density plot of the *mpg* measured type variable from *mtcars* data.

```
d <- density(mtcars$mpg)
plot(d, main="Density Plot of 'mpg'"  )
```



where the area under the curve from one height value to another measures the relative proportion of the observations having *mpg* in that range. In other words, the density plot indicates how the range of *mpg* is distributed over the possible range of values.

Let us look at the following study of company share price given in *SHARE_PRICE.csv' spreadsheet file:

```
#==========================================================
#
# A business analyst is studying share prices of companies
# from three different business sectors. As part of the
# study a random sample (n=60) of companies was  selected
# and the following data was collected:
#
#- Share_Price: The market value of a company share
#- Profit: The company annual profit
#- RD: Company annual spending on research and development
#- Turnover: Company annual total revenue
#- Competition: A variable coded:
#  0 if the company operates in a very competitive market
#  1 if the company has a great deal of monopoly power
#- Sector: A variable coded:
#  1 if the company operates in the IT business sector;
#  2 if the company operates in the Finance business sector;
#  3 if the company operates in the Pharmaceutical business
#    sector.
#- Type: A variable coded:
#  0 if the company does business mainly in Europe;
#  1 if the company trades globally.
#
#==========================================================
```

```r
companyd <- read.csv("SHARE_PRICE.csv", header=T)
# ----------------------------------------------
# To check how big the data set is, we can use function dim()
dim(companyd)   # dimension of company data
```

```
## [1] 60  7
```

```r
# Data set has 60 observations and 7 variables
#=========================================================
summary(companyd) # Get the key summary statistics for each variable
```

```
##   Share_Price        Profit             RD            Turnover
##  Min.   :101.0   Min.   :  2.90   Min.   : 39.20   Min.   : 30.3
##  1st Qu.:501.2   1st Qu.: 59.73   1st Qu.: 75.78   1st Qu.:112.3
##  Median :598.5   Median : 88.85   Median : 90.60   Median :173.5
##  Mean   :602.8   Mean   : 84.76   Mean   : 89.64   Mean   :170.2
##  3rd Qu.:739.8   3rd Qu.:106.62   3rd Qu.:104.15   3rd Qu.:216.6
##  Max.   :880.0   Max.   :170.50   Max.   :152.60   Max.   :323.3
##   Competition       Sector       Type
##  Min.   :0.0   Min.   :1   Min.   :0.0
##  1st Qu.:0.0   1st Qu.:1   1st Qu.:0.0
##  Median :0.5   Median :2   Median :0.5
##  Mean   :0.5   Mean   :2   Mean   :0.5
##  3rd Qu.:1.0   3rd Qu.:3   3rd Qu.:1.0
##  Max.   :1.0   Max.   :3   Max.   :1.0
```

```r
# BUT!!! Variables: 'Comparison', 'Sector' and 'Type' are
# attribute variables?! We need to let R know this!
# To encode a measured variable as an attribute variable
# we can use function factor(variable_name).
companyd[, 5] <- factor(companyd[, 5])
companyd[, 6] <- factor(companyd[, 6])
companyd[, 7] <- factor(companyd[, 7])
summary(companyd)
```

```
##   Share_Price        Profit             RD            Turnover
##  Min.   :101.0   Min.   :  2.90   Min.   : 39.20   Min.   : 30.3
##  1st Qu.:501.2   1st Qu.: 59.73   1st Qu.: 75.78   1st Qu.:112.3
##  Median :598.5   Median : 88.85   Median : 90.60   Median :173.5
##  Mean   :602.8   Mean   : 84.76   Mean   : 89.64   Mean   :170.2
##  3rd Qu.:739.8   3rd Qu.:106.62   3rd Qu.:104.15   3rd Qu.:216.6
##  Max.   :880.0   Max.   :170.50   Max.   :152.60   Max.   :323.3
##  Competition Sector Type
##  0:30        1:20   0:30
##  1:30        2:20   1:30
##              3:20
##
##
##
```

```r
# ----------------------------------------------
# Alternatively, to get an individual summary for measured variable
# type:
sapply(companyd[,1:4], summary)
```

```
##         Share_Price Profit     RD Turnover
## Min.          101.0   2.90  39.20     30.3
## 1st Qu.       501.2  59.73  75.78    112.4
## Median        598.5  88.85  90.60    173.5
## Mean          602.8  84.76  89.64    170.2
## 3rd Qu.       739.8 106.60 104.20    216.6
## Max.          880.0 170.50 152.60    323.3
```

```
# ----------------------------------------------
# To focus on the centre of the distributions for the measured
# variables you can ask only for the rows showing mean and
# median to be displayed.
sapply(companyd[,1:4], summary)[3:4, ]
```

```
##         Share_Price Profit     RD Turnover
## Median        598.5  88.85 90.60    173.5
## Mean          602.8  84.76 89.64    170.2
```

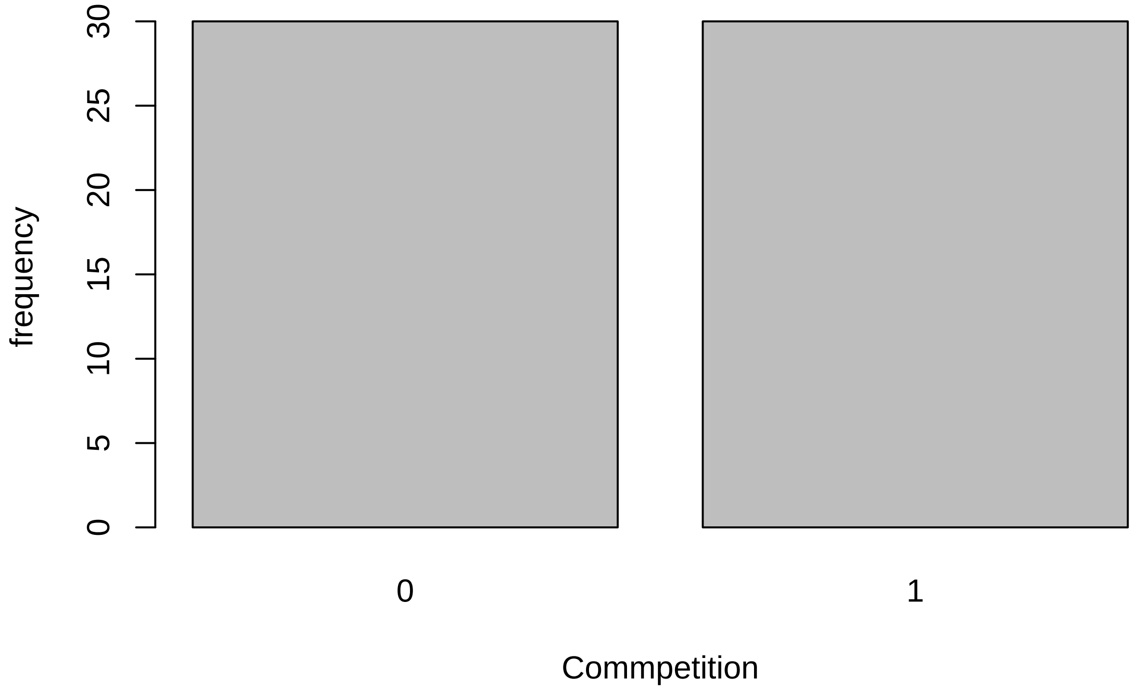
```
# To observe spread of the data we can use standard deviation
# and/or Inter Quartile Range
sapply(companyd[,1:4], sd)
```

```
## Share_Price      Profit          RD    Turnover
##   177.28461    37.76443    24.13231    75.72712
```

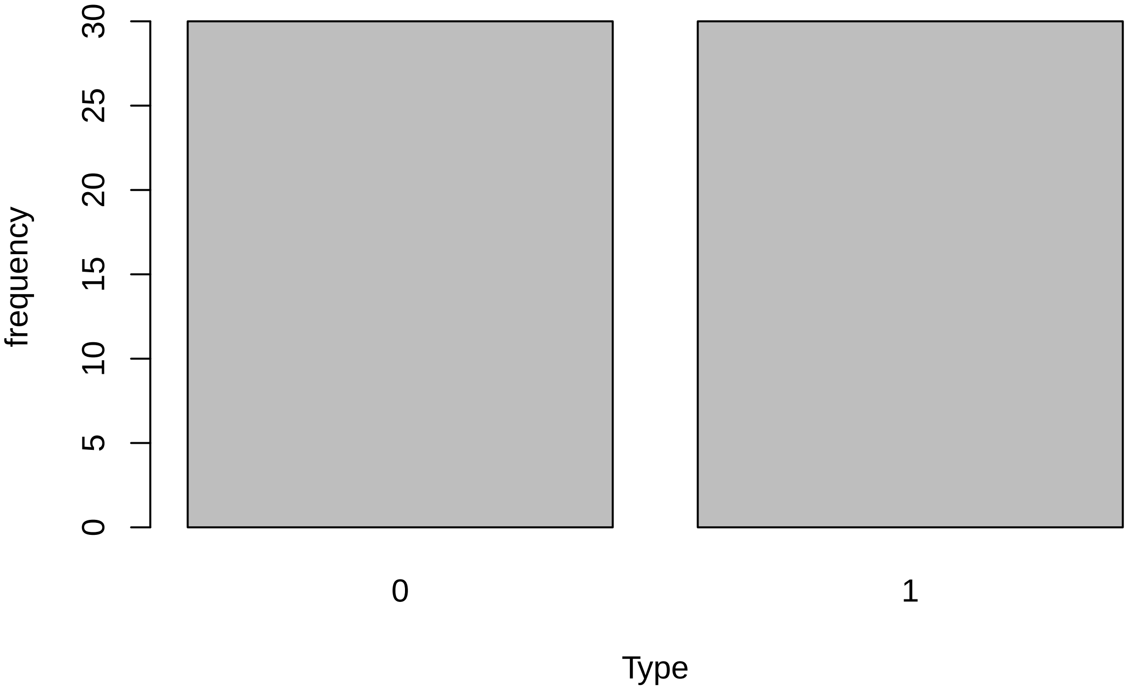
```
sapply(companyd[,1:4], IQR)
```

```
## Share_Price      Profit          RD    Turnover
##     238.500      46.900      28.375     104.250
```

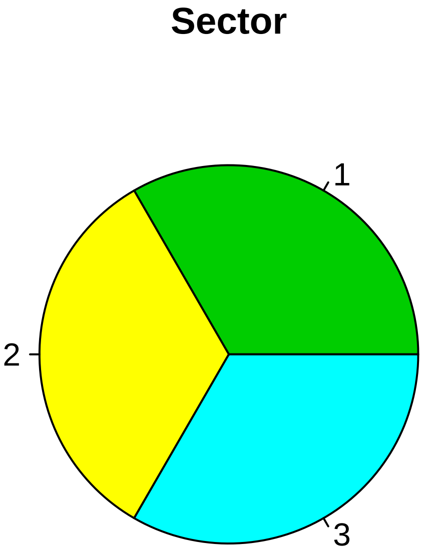
```
#=========================================================
# To explore the distributions of the variables visually
# you should use the appropriate graphs.
# Usually you use a pie chart or a bar plot if you want to
# visualise an attribute variable.
barplot(table(companyd[, 5]), xlab="Commpetition", ylab="frequency")
```



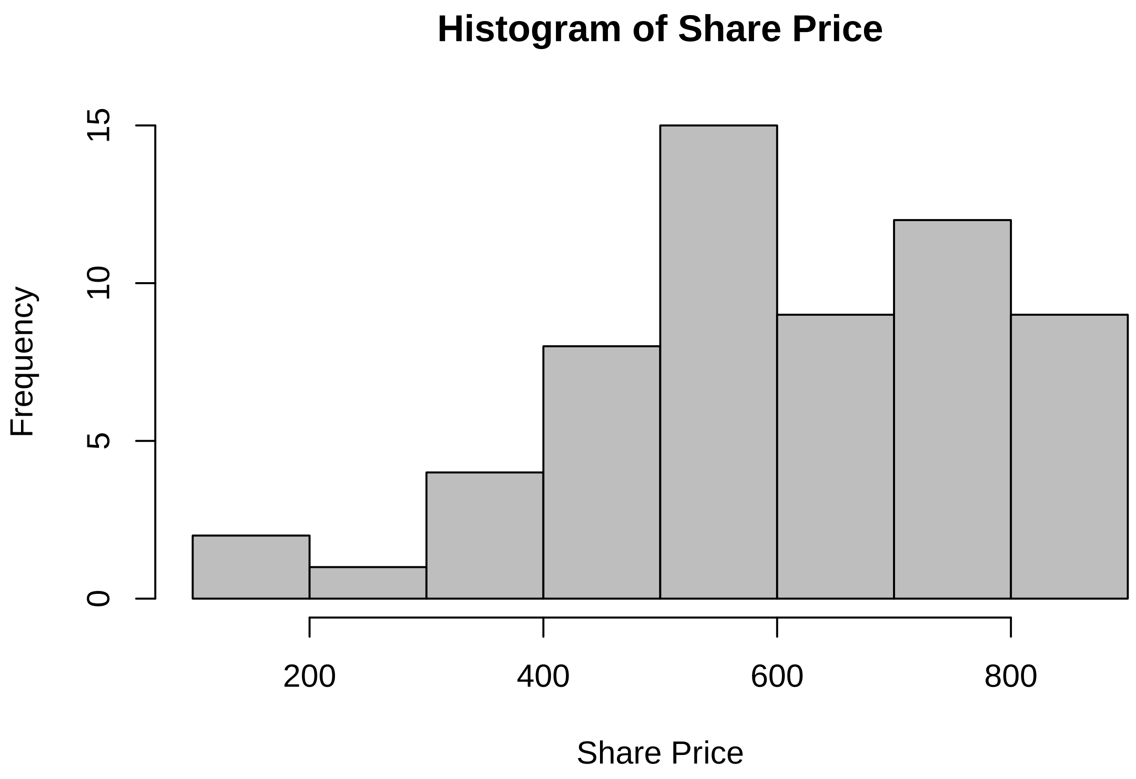
```
barplot(table(companyd[, 7]), xlab="Type", ylab="frequency")
```



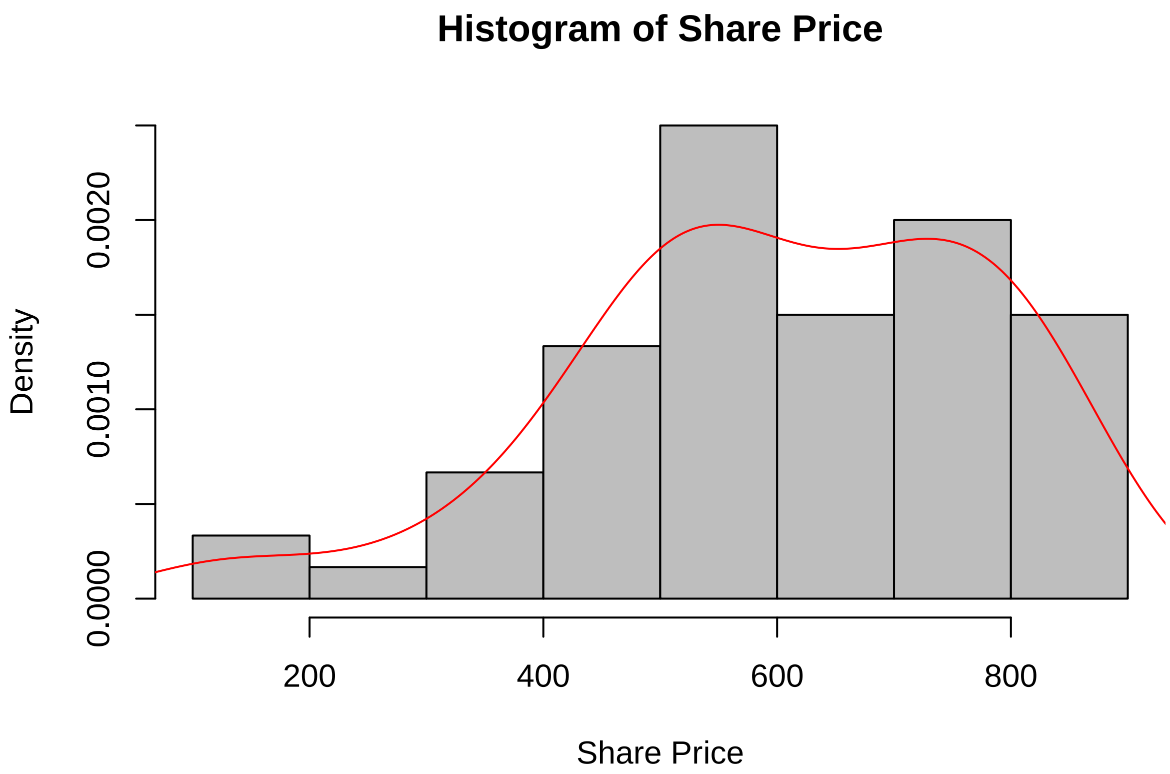
```
pie(table(companyd[, 6]), labels=names(companyd$Sector), col=c(3, 7, 5), main="Sector")
```



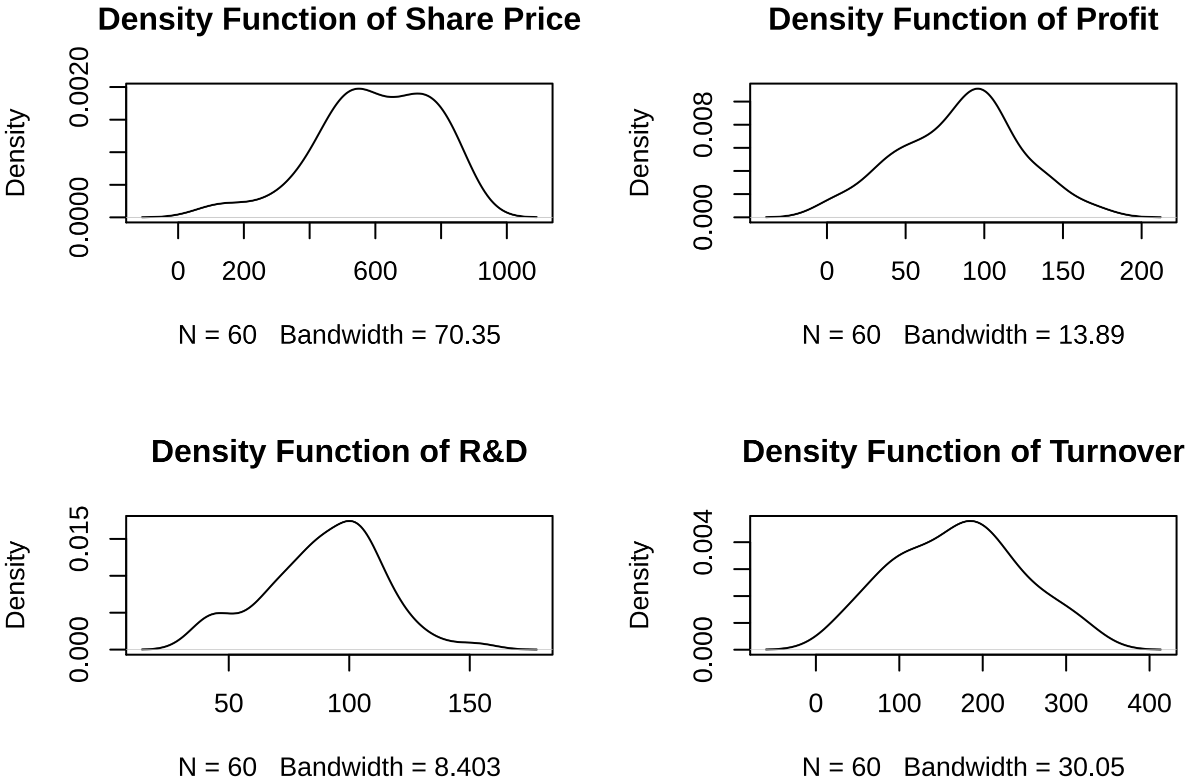
```
# Histogram is appropriate when you have a measured variable
# to graphically explore.
hist(companyd[, 1], xlab="Share Price", main="Histogram of Share Price", col="gray")
```



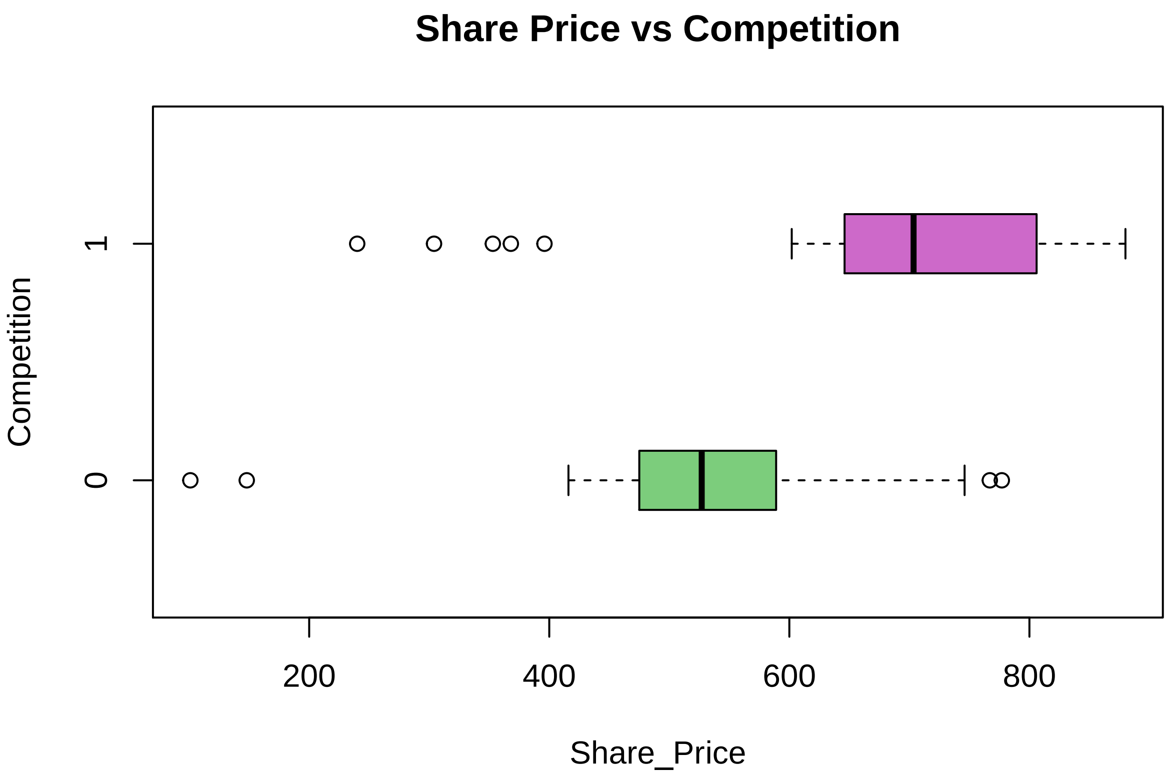
```
# If you would like to see the density smoothing of the
# histogram, on your histogram you will plot the
# probability density rather than the frequency of the
# measured variable, over which you can superimpose
# a kernel density smoothing line.
hist(companyd[, 1], xlab="Share Price", main="Histogram of Share Price", col="gray", prob=T)
lines(density(companyd[, 1]), col="red")
```



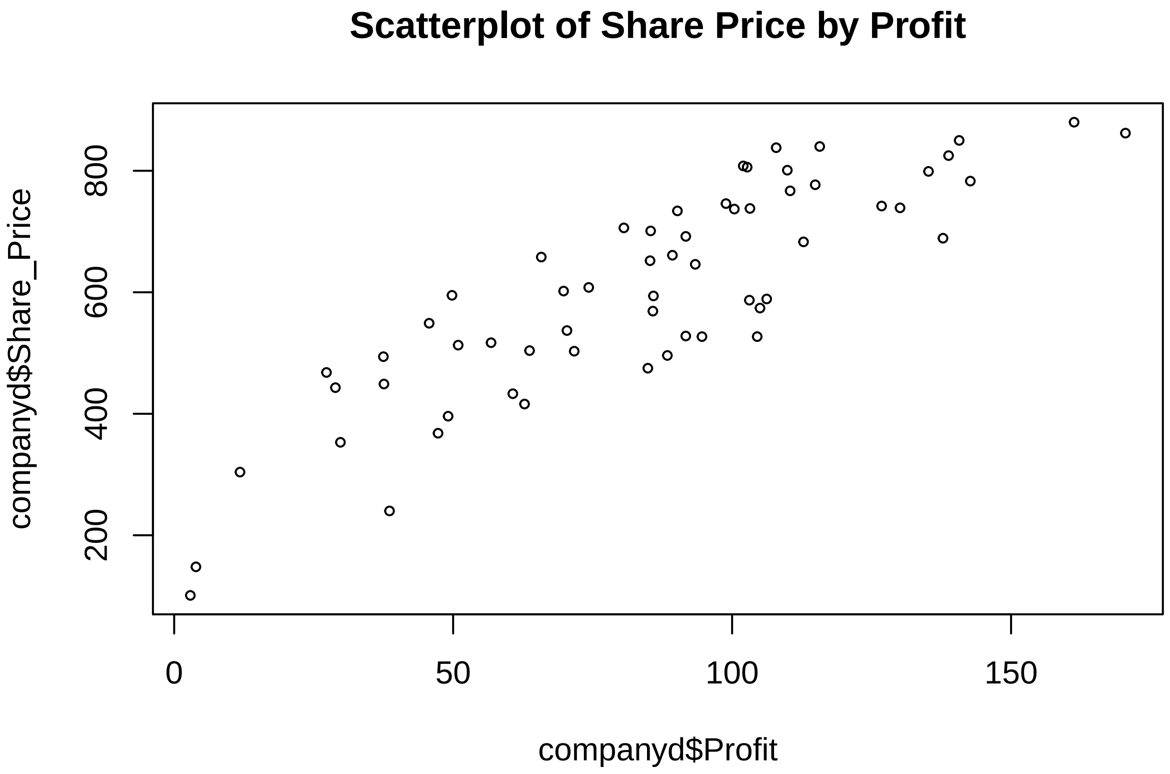
```
# Or you can have a kernel density smoothing as an
# individual plot.
par(mfrow=c(2, 2))  # splits the graph window into 2 rows and 2 columns
plot(density(companyd[,1]), main="Density Function of Share Price")
plot(density(companyd[,2]), main="Density Function of Profit")
plot(density(companyd[,3]), main="Density Function of R&D")
plot(density(companyd[,4]), main="Density Function of Turnover")
```



```
par(mfrow=c(1, 1))      # puts the graph window back onto a single plot
#=========================================================
# To investigate the possible relationship between attribute
# and measured variables you can use a box plot:
boxplot(Share_Price ~ Competition, data = companyd, boxwex = 0.25, main="Share Price vs Competition", xl
```



```
# If you are interested in analysing a potential
# relationship between measured variables use a scatter
# plot:
plot(companyd$Profit, companyd$Share_Price, cex=.6, main="Scatterplot of Share Price by Profit")
```



```
# Note: the plot() function gives a scatterplot of two
# numerical variables. The first variable listed will be
# plotted on the horizontal axis and the second on the
# vertical axis, ie. you 'feed' as the arguments first
# variable representing X and then variable
# representing Y.
# ---------------------------------------------
```

Your Turn

Use *birthwt* data from *MASS* package in *R*.

i. What type of information variables are providing. Provide key information about each of the variables and use the appropriate plot to illustrate your findings.
ii. How two variables are related: boxplot or scatterplot? Illustrate the potential relationships between the two variables using appropriate graphs.
iii. Write down questions that you could answer with this data.

Further Reading

One of the great things about R that makes it so powerful to use, is the freely available excellent documentation that you can access not just from CRAN, but from other websites created by a vast community of R enthusiast. Here are a couple of them that you can explore for yourself:

- Good for quick and useful tips http://www.ats.ucla.edu/stat/r/
- More comprehensive - for those of you already familiar with R http://www.statmethods.net/

Acknoledgment

I would like to thank Ian McGowan and George Rawlings for letting me adapt their teaching material.