Categorical data

We often view categorical data with tables but we may also look at the data graphically with bar graphs or pie charts.

Using tables

The table command allows us to look at tables. Its simplest usage looks like table(x) where x is a categorical
variable.
Example: Smoking survey
A survey asks people if they smoke or not. The data is
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)
x
No Yes
2 3
```

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

The table command will summarize bivariate data in a similar manner as it summarized univariate data.
We can handle this in R by creating two vectors to hold our data, and then using the table command.

```
> smokes = c("Y","N","N","Y","N","Y","Y","Y","N","Y")
> sex = c(□F□,□F□,□M□,□M□,□M□,□M□,□F□,□M□,□M□,□F□)
> table(smokes,sex)

      sex
smokes F M
     N 1 3
     Y 3 3
```

Bar charts

```
> barplot(x) # this isn't correct
> barplot(table(x)) # Yes, call with summarized data
> barplot(table(x)/length(x)) # divide by n for proportion
```

For bivariate

```
> barplot(table(smokes,sex))
> barplot(table(smokes,sex),beside=TRUE)
```

Pie charts

```
>pie(table(x))
```

Mode

No built-in function!!!

```
>which(table(x)==max(table(x)))
>which.max(table(X))
```

Numerical□data:□

Numeric measures of center and spread:

I) Measures of central tendency

Is a value that represents a typical, or central, entry of a data set

Measures of central tendency	Function in statistic	Function in R
Mean	$\bar{x} = \frac{\sum x}{n}$	**mean(x)**
Median	1) if n (odd number) $M = x_{\left(\frac{n+1}{2}\right)}$ 2) if n (odd number) $M = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$	**median(x)**
Mode	The data entry occurs with the greatest frequency	

II) Measures of variation

Measures of Variation	Function in statistic	Function in R
Range	Range = (max data entry) - (min data entry)	**range(x)** return vector of two elements (min(x),max(x)) the actual range **range(x)[2] – range(x)[1]** or **diff(range(x))**
Variance	$\operatorname{var}(x)=\frac{\sum(x-\bar{x})^2}{n}=\frac{\sum x^2}{n}-\bar{x}^2$	var(x)
Standard deviation	$sd(x)=\sqrt{\frac{\sum(x-\bar{x})^2}{n}=\frac{\sum x^2}{n}-\bar{x}^2}$	sd(x)

III) Measures of Position

A) **Quartiles:** the three quartiles Q_1, Q_2, and Q_3 approximately divided an ordered data set into four equal parts

max data entry------ x--------x--------x ------min data entry

Q_3 Q_2 Q_1

Measures of Position	Function in statistic	Function in R
Quartiles	• The first and the third quartiles are the medians of the lower and upper halves of the data set. • The second quartile is the same as the median of the data set.	**quantile(x)** return a vector of three elements
Inter quartile range (IQR)	$IQR = Q_3 - Q_1$	IQR(x)
Semi-inter quartile range (SIQR)	$SIQR = \frac{Q_3 - Q_1}{2}$	

Outliers

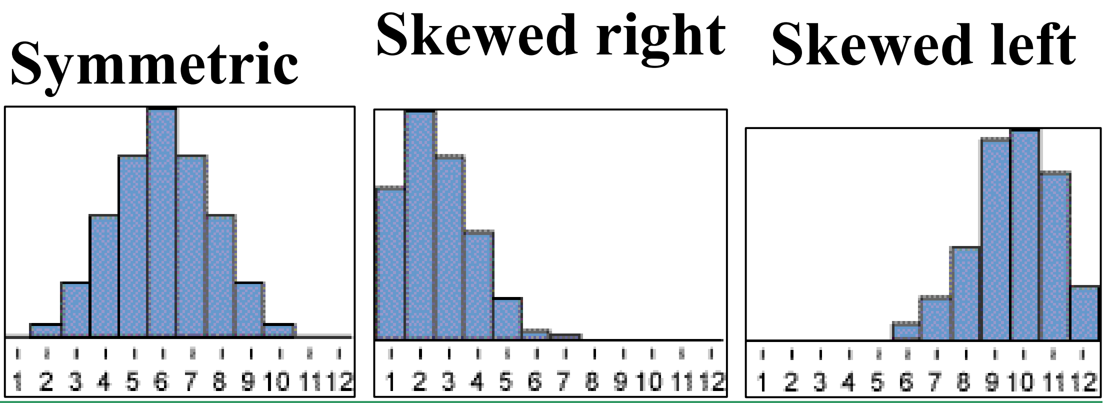
	Measures	Function in R
central tendency	Trimmed Mean	**mean(x,trim=)**
	Median	**median(x)**
variation	IQR	IQR(x)
	Median Average Deviation $(\text{Median}\ \|X_i - \text{median}\| * 1.4826)$	Mad(x) or `median(abs(x - median(x))) * 1.4826`

Shape of a distribution

Histogram

The purpose of a histogram is to graphically summarize the distribution of a univariate data set. The histogram graphically shows the following:

- center (i.e., the location) of the data;
- spread (i.e., the scale) of the data;
- skewness of the data;
- presence of outliers; and
- presence of multiple modes in the data.




```
> hist(x) # frequencies
> hist(x,probability=TRUE) # proportions (or
probabilities)
```

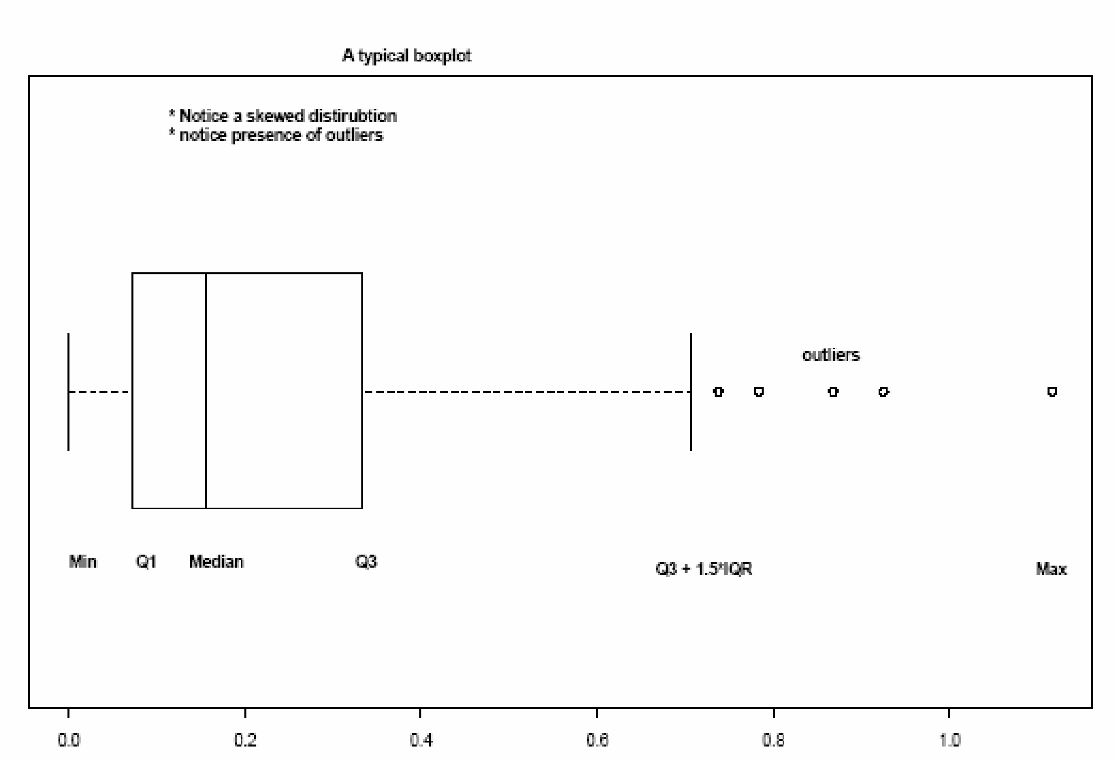
Box Plot

box-and-whisker plot is an exploratory data analysis tool that highlights the important features of a data set.
The **five-number summary** is used to draw the graph.

- The minimum entry
- Q1
- Q2 (median)
- Q3
- The maximum entry

Calculate the following points:

L1 = Q1- 1.5*IQR
L2 = Q1 - 3.0*IQR
U1 = Q3 + 1.5*IQR
U2 = Q3 + 3.0*IQR



->boxplot(x)
For bivariate :boxplot(x,y)

Goodness of fit tests

Chi Square test

The chi-square test is used to test if a sample of data came from a population with a specific distribution.
The test requires that the data first be grouped.
The chi-square goodness-of-fit test can be applied to discrete distributions disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The chi-square test is defined for the hypothesis:

H_0: The data follow a specified distribution.

H_a: The data do not follow the specified distribution.

For the chi-square goodness-of-fit computation, the data are divided into ns and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

Kolomogrov Smirnov test

A goodness-of-fit test for any statistical distribution. The test relies on the fact that the value of the sample cumulative density function is asymptotically normally distributed.

H0 : $F'(x) = F(x)$ for all x

H1 : $F'(x) \neq F(x)$ for at least one value of x

To apply the Kolmogorov-Smirnov test, calculate the cumulative frequency of the observations as a function of class. Then calculate the cumulative frequency for a true distribution (most commonly, the normal distribution). Find the greatest discrepancy between the observed and expected cumulative frequencies, which is called the "D-statistic." Compare this against the critical D-statistic for that sample size. If the calculated D-statistic is greater than the critical one, then reject the null hypothesis that the distribution is of the expected form.

$$D = \max_x \{|F'(x) - F(x)|\}$$

- ks.test(x,"pnorm",mean=…. ,sd=……)

Exploratory Data Analysis (EDA) Functions

```
eda.shape<-function(x)
{par(mfrow =c(2,2))
hist(x)
boxplot(x)
iqd<-summary(x)[5]-summary(x)[2]
plot(density(x,width=2*iqd),xlab="x",ylab="",type="l")
qqnorm(x)
qqline(x)}

eda.ts<-function(x)
{par(mfrow =c(2,2))
ts.plot(x)
acf(x)
invisible()}
```