1 Introduction

1.1 What is Statistics?

Statistics is a *scientific study* of *numerical data* based on *natural phenomena*.

It is also the *science* of *collecting*, *organising*, *interpreting* and *reporting data*.

There are four phases of an experiment, survey or study.

1. Planning
Advise on the best way to collect data, what readings should be taken, how bias can be reduced or eliminated. This stage is guided by the questions the study wants to address, time and costs.

2. Data Analysis
Summarize the data using graphs and numerical measures to get an impression of the *variability* and *shape* of the data.

3. Model Building
Develop a mathematical model based on probability theory to explain the patterns observed.

4. Inference
Draw conclusion about the *population* (the entire collection of units from which observations can be made) based on a model and observations from a *sample* (a subset of units from the population).

Some applied statistical problems

1. A Quality Control (QC) Problem.
   In many manufacturing companies, a QC team regularly sample output and test for defects. They usually set prior limits to accept or reject the production. For example, the process is deemed to be okay if the proportion of defective items (overall) is at most 5%.

   Suppose that a company inspects a sample of 60 items and finds 4 defectives. Should they reject the whole batch? 5% of 60 is 3. If the true defect rate was 5% can the observed '4' be explained by chance?

2. In vitro fertilization (IVF) data
   Suppose that 6 of the first 7 births on the IVF program in Australia were female. Does this provide evidence of sex bias in IVF babies?

In order to make a sound statistical argument to answer such questions, we need to study a number of statistical concepts and methods.

1.2 Graphical Methods

A graphical display helps to visualize the data and obtain a preliminary understanding of the data distribution. Stem-and-Leaf plots, bar graphs, histograms, and box plots are popular graphical methods.

1.2.1 Stem-and-Leaf plots (P.5-6)

This is a simple graphical representation of a data set. It shows all data information and their distribution. Hence is suitable for relatively smaller data sets. This is prepared in following steps:

1. Separate each data point into a stem component and a leaf component.
2. Write all stem digits left of a vertical line.
3. Write all leaf digits right of the vertical bar.
4. Re-arrange each leaf digit.

Example: No. of persons killed in road accidents in 13 major cities during the Easter holidays in 2005 are given below:

6 24 7 9 15 10 31 6 23 7 13 11 4

Draw a stem-and-leaf diagram for this data.

Solution: Step 1: Separate the data into stem (10 digit) and leaf (unit digit) components.

Step 2 & 3: Write stem digits to the left and leaf digits to the right.

Step 4: Order the leaf digits.

1.2.2 Use of software - R

In this unit of study, we will use a statistical language called `R`. It provides a wide variety of statistical and graphical techniques and is available as *free software* from the CRAN (Comprehensive R Archive Network) website: `http://www.r-project.org/`

`R` is an *expression language*, which means it is not menu-driven. All commands are case sensitive and are written and executed in the console window at the prompt. However, there are certain tasks which can be implemented through the menus (like installing new packages, for example).

Data in `R` are organised as named structures. In this course we will mainly deal with the simplest such structures: vectors and matrices. They can be numerical *vectors* (like height, weight,

etc) or categorical *factors* (like gender, type of diet, etc). `R` treats factors and vectors differently, as will be seen later. From time to time we will use other types of data structures, like the *data frame*, which can combine both factors and vectors.

Statistical output is generated in the same window (the `R` console), where the commands are executed. Results can be copied and pasted to other programs (like a word processor), or can be saved as a text file, using the "File" menu. Graphs are produced in a separate graphics window and can be saved or printed by right-clicking on them and choosing the appropriate command.

Most of the time we will read data with the `scan` and `read.table` and `write.table` commands. Help for any built-in function can be obtain by typing `?name_of_the_function`. For example, `?scan`. Alternatively, you can type `help(scan)`.

For a more detailed introduction to R and the way to login the computer system in the School is given in the tutorial sheet and the website. The pdf manuals are available through the "Help" menu.

For the stem-and-leaf plot in the example, the R codes are

```
x=c(6,24,7,9,15,10,31,6,23,7,13,11,4)
stem(x)
```

```
  The decimal point is 1 digit(s) to the right of the |
  0 | 466779
  1 | 0135
  2 | 34
  3 | 1
```

Example: Consider the following data on weight in pounds (recorded to the nearest pound) of 92 students (available on website). Prepare a stem-and-leaf plot for the female students.

Males:
140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175
175 170 180 135 170 157 130 185 190 155 170 155 215 150 145 155
155 150 155 150 180 160 135 160 130 155 150 148 155 150 140 180
190 145 150 164 140 142 136 123 155

Females:
140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120
118 125 135 125 118 122 115 102 115 150 110 116 108 95 125 133
110 150 108

Solution: In R,

```
x=c(140,120,130,138,121,125,116,145,150,112,125,130,
  120,130,131,120,118,125,135,125,118,122,115,102,115,
  150,110,116,108,95,125,133,110,150,108)
stem(x)
```

```
  The decimal point is 1 digit(s) to the right of the |

   9 | 5
  10 | 288
  11 | 002556688
  12 | 0001255555
  13 | 0001358
  14 | 05
  15 | 000
```

You may try the stem-and-leaf plot for male students.

```
12 | 3
13 | 005568
14 | 000255558
15 | 000000355555555557
16 | 000045
17 | 000055
18 | 0005
19 | 00005
20 |
21 | 5
```

Remark:

1. The stems may have many digits, but each leaf contains only the final digit of the observation.
2. The stem-and-leaf plot is not appropriate for some versatile datasets.

For example, if the data are:

$$0.01 \quad 4.37 \quad 48.25 \quad 167.389$$

then they can not be presented in a meaningful way on a stem-and-leaf plot.

```
x=c(0.01,4.37,48.25,167.389)
stem(x)
```

```
  The decimal point is 2 digit(s) to the right of the |
```

```
0 | 00
0 | 5
1 |
1 | 7
```

Note that the stem is 100 digit whereas the leaf is 10 digit. Because of the large data range, there is a loss of accuracy in the data presentation.

1.2.3 Frequency distribution (P.7)

Sometimes the sample size is too large to display all the data. In such a situation, we group data using a frequency table before a graph is drawn. Suitable intervals are chosen and we record the number of values in each interval, called its *frequency*.

A *frequency distribution table* is a table consisting of ordered intervals with counts of the number of values in each interval or frequencies. In drawing the table, one needs to choose the *number* of bins/classes, the *widths* of the bins, and what to do with observations which equals exactly to the *endpoints* of the bins.

We adopt the following guidelines:

1. The number of bins should be between 5 and 15, but can be lower for small datasets, or larger for big datasets. A suggestion from Sturges (Sturges rule) is:

No. of observations	**Appropriate no. of classes**
10 to 100	4 to 8
100 to 1000	8 to 11
1000 to 10000	11 to 14

or The no. of classes $k = 1 + 3.322 \times \log_{10}(n)$.

2. All bins intervals should have the same length, equal to the range of the data divided by the number of bins, and rounded to the next integer.

3. The intervals will be *right-closed*; that is, an observation equal to the boundary between two bins should be included in the left bin.

The general layout of grouped data with k groups:

Group/Class interval	Class center	Frequency	Relative frequency
$y_1 < x \leq y_2$	$u_1 = (y_1 + y_2)/2$	f_1	f_1/n
$y_2 < x \leq y_3$	$u_2 = (y_2 + y_3)/2$	f_2	f_2/n
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_k < x \leq y_{k+1}$	$u_k = (y_k + y_{k+1})/2$	f_k	f_k/n
TOTAL		n	1.000

There is no definite rule in choosing y_1 but the first interval must contain the minimum of the data.

The ratio of $\dfrac{\text{frequency}}{\text{total sample size}}$ in each interval is called the *relative frequency*. This gives the percentage of values falling in a particular interval.

Example: Consider the data on weight in pounds (recorded to the nearest pound) of 92 students again. Prepare a suitable frequency table for the female data.

Solution: We have $n = 35$.

1. The number of bins chosen is

   _____.

2. The range is _____. Round up to _____.

3. The classes are

   _____.

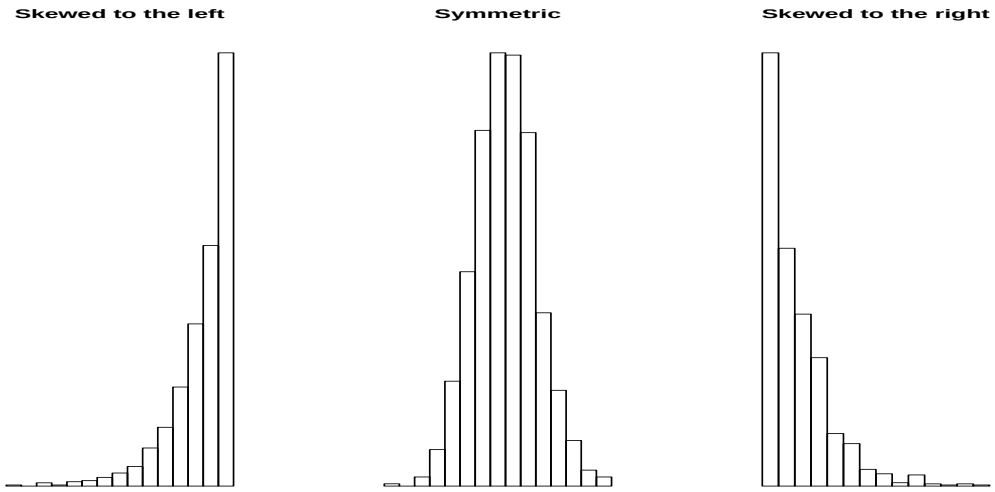
In R,

```
xc=cut(x,breaks=c(94,104,114,124,134,144,154))
table(xc)
xc
 (94,104] (104,114] (114,124] (124,134] (134,144] (144,154]
        2         5        11        10         3         4
```

1.2.4 Histogram (P.8)

Frequency tabulations can be represented as a graph of frequency against the measurement variable. Histogram is one kind of such graph consisting of rectangles. The area of each rectangle is proportional to the number of data values that occur within the interval. The width of each rectangle is the width of its class.

The histogram describes the shape of the distribution like the overall pattern (symmetric or skewed), area of concentration, and the presence/absence of outliers. Note that the shape of the histogram changes with the classes chosen.

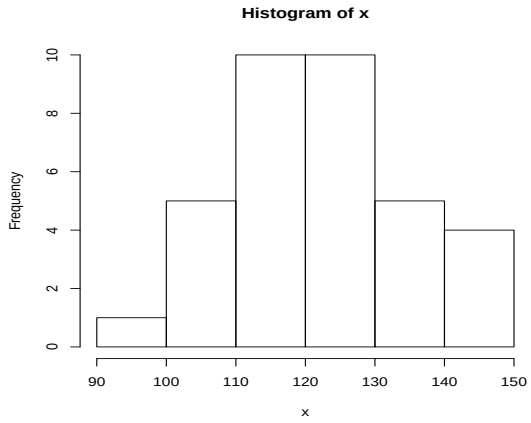


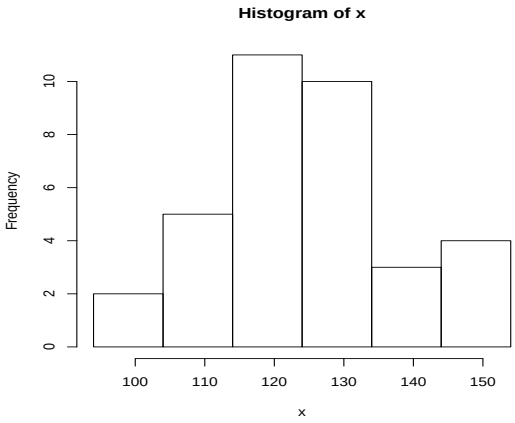
Example: Consider the data on weight in pounds (recorded to the nearest pound) of 92 students again. Draw a histogram for the female data and describe the distribution.

Solution: In R,

```
> x=c(140,120,130,138,121,125,116,145,150,112,125,130,
+ 120,130,131,120,118,125,135,125,118,122,115,102,115,
+ 150,110,116,108,95,125,133,110,150,108)
```

```
> hist(x)
> hist(x,breaks=c(94,104,114,124,134,144,154))
```



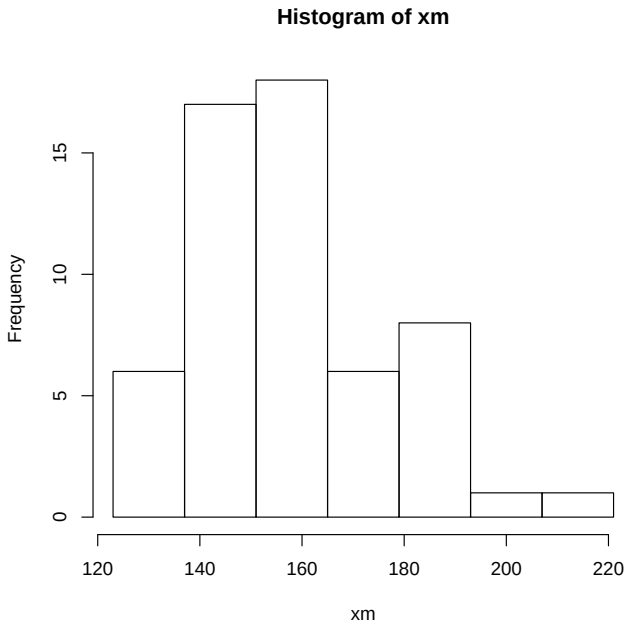


The first histogram does not specify the breaks and the default is (90,100], ..., (140,150].

The second histogram specifies the breaks to be (94,104], ..., (144,154].

The shape of the histogram changes with the classes chosen. The distribution of the weight of female student is _____.

You may try the histogram plot for male students.



The distribution of the weight of male student is _____.

Useful R commands:

```
x=c(1,2,5,7,9,13,11,15) #for example
stem(x)
hist(x,breaks=c(0,5,10,15))
xc=cut(x,breaks=c(0,5,10,15))
table(xc) #give frequencies
```

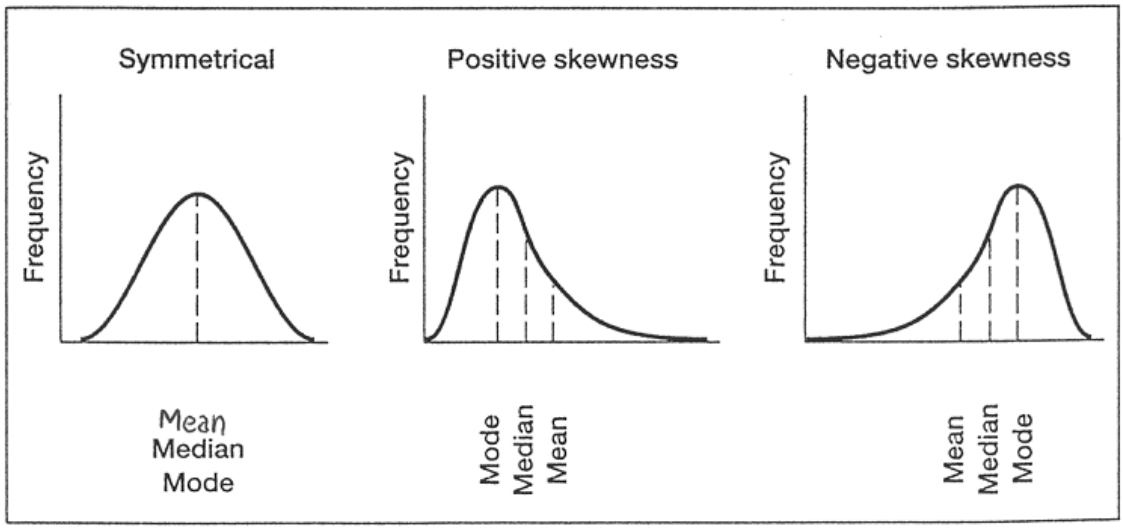
1.3 Shapes of Distributions (P.9)

In many statistical analysis problems, the analyst need to know the shape of the distribution of data to make inferences. Histograms and boxplots can be used to get an impression of the shape of the data set. There are three main shapes:

1. *Symmetric*

2. *Right skewed (positive skewness):* the boxplot is stretched to the right.

3. *Left skewed (negative skewness):* the boxplot is stretched to the left.



How to judge the symmetry of a distribution?

- If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median.

- If the upper quartile is further from the median than the lower quartile, then the distribution is positively (or right) skewed.

- If the lower quartile is further from the median than the upper quartile, then the distribution is negatively (or left) skewed.