2.1 Boxplots (P.15-16)

Recall that observations outside the interval (LT,UT) are called outliers or abnormal observations, where

Lower threshold value (LT) = lower quartile - 1.5 × IQR

Upper threshold value (UT) = upper quartile + 1.5 × IQR.

A popular (box type) graphical representation of the following information from a data set is known as a boxplot:

- Quartiles Q_1, Q_2 and Q_3, (draw a rectangular box from the quartiles Q_1 to Q_3 and mark Q_2 within this box)
- Smallest and largest observations within (LT,UT),
- Outliers, if exist.

Diagram: Suppose that a data set contains three values below the LT (left outliers) and two values above the UT (right outliers). Now we show these information in the diagram below:

Boxplots show the shape of the distribution of data very clearly and are helpful in representing any outlying (or extreme) values of a data set.

THE UNIVERSITY OF SYDNEY
MATH1015 Biostatistics Week 2/3

Example: Consider the following data set of 13 observations x_i from the previous example:

4 6 6 7 7 9 10 11 13 15 22 24 30

1. Find LT and UT for this sample.
2. Identify any outliers if they exist.
3. Draw a boxplot for this sample following the steps:
   (a) Draw a rectangle (horizontal or vertical) of arbitrary width from Q_1 to Q_3.
   (b) Draw a dotted-line across the rectangle at Q_2.
   (c) Draw two lines (called, Whiskers) to and from the observations within (LT,UT) from the above rectangle.
   (d) Mark any identified outliers by ∘

Solution:

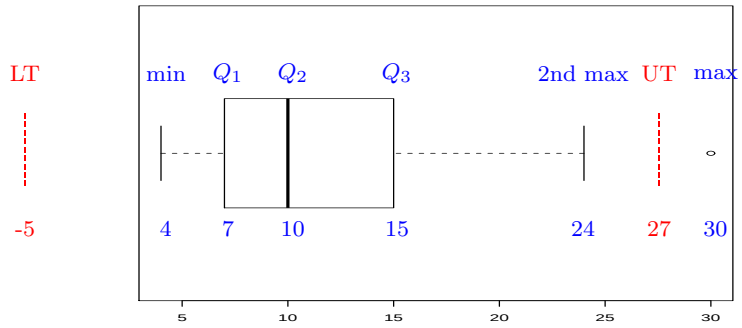
1. From the previous example, we have calculated:
   Median, $Q_2 = 10$; Lower quartile, $Q_1 = 7$. Upper quartile, $Q_3 = 15$. Hence
   IQR = 8;
   LT = -5;
   UT = 27.
2. All observations in the interval (-5,27) are considered "legitimate". Clearly, there is only one data point outside this interval. Therefore, the last observation 30 is considered as abnormally high. This is an outlier.

3. The following *boxplot* summarises the above information as a graph indicating the outlier by o :

Boxplot in R:

R can be used to draw a boxplot.
Let x contains the data.

```
> x=c(4,6,6,7,7,9,10,11,13,15,22,24,30)
> boxplot(x)
```



Notes:

- Boxplots are useful to compare a continuous variable (e.g. length, weight etc) with a nominal variable (e.g. treatment).
- Length of whisker in R is by default chosen to be $1.5 \times$IQR,
- Boxplots give a simple visual display and hence a quick impression of the shape of the data set:
  - Symmetrical: left and right tails are similar
  - left skewed: boxplot is stretched to the left.
  - Right skewed: boxplot is stretched to the right.

Now we look at a number of additional summaries from a data set.

2.2 Measures of Location and Spread (P.9-11)

Measures of Location

We have seen that median is a measure of the center of a data set. Another popular measure of the center of a data set is known as the *mean*. Recall from your high school work that the mean of (4,7, 9, 5, 3) is $\frac{4+7+9+5+3}{5} = 5.6$. Use your calculator to check this answer. Now we develop this concept to handle common problems instatistics. we use the following notation:

A Notation

Suppose that we have n observations from an experiment. This collection (or set) of n values is called a *sample*. Let x_1 be the first sample point or observation; x_2 be the second sample point or observation etc and x_n be the n^{th} sample point or observation.

Example: Suppose that we have a sample of five observations $\{4, 7, 9, 5, 3\}$.
For this sample, the first observed values is 4 and therefore we write $x_1 = 4$ to identify it. Similarly, $x_2 = 7,\ x_3 = 9,\ x_4 = 5,\ x_5 = 3$.

Summation Notation: For simplicity, the sum of these n values $x_1, x_2, \cdots, x_n$ is abbreviated by the sigma notation as follows:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Note: Many calculators use this notation. Please check your calculator now.

Example: Consider the sample: $x_1 = 4, x_2 = 7,\ x_3 = 9,\ x_4 = 5,\ x_5 = 3$. Write down $\sum_{i=1}^{5} x_i$ and evaluate it.

Solution:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = \underline{\qquad\qquad}$$

Example: Evaluate the following summation expressions for the values (3, 4, 5, 1):

$$\sum_{i=1}^{4} x_i,\ \sum_{i=2}^{3} x_i,\ \sum_{i=1}^{4}(2x_i + 3) \text{ and } \sum_{i=1}^{4} x_i^2.$$

Solution:

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = \underline{\qquad\qquad}$$

$$\sum_{i=2}^{3} x_i = x_2 + x_3 = \underline{\qquad\qquad}$$

$$\sum_{i=1}^{4}(2x_i + 3) = (2x_1 + 3) + (2x_2 + 3) + (2x_3 + 3) + (2x_4 + 3)$$
$$= \underline{\qquad\qquad}$$
$$= \underline{\qquad\qquad}.$$

$$\sum_{i=1}^{4} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = \underline{\qquad\qquad}$$

2.2.1 The Sample Mean, p9

The sample mean is the simple arithmetic mean or the average of observations. For n observations $x_1, x_2, \ldots, x_n$, this is denoted by $\bar{x}$ (called x bar) and is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example:

The mean of the sample of 4 values from a previous example is

$$\bar{x} = \underline{\qquad\qquad}.$$

Exercise: Look at your calculator now. Change the mode of your calculator to 'stat' or 'sd' or as per calculator instructions. Chack the above answer using your calculator.

Note: The mean is very sensitive to large or small outliers in the sample. In such cases it is better to use the median as a measure of the "centre" of the data.

Use of R

R can be used to find the mean of a sample. Practice this example.

```
> x=c(3,4,5,1)
> mean(x)
>3.25
```

Exercise: Find the median, mean and mode for the data set: 13.3, 10.7, 11.0, 11.1, 12.9, 11.8, 11.9, 12.2, 10.8, 12.2, 11.6, 11.8

Solution: Order the data x_i to find the median:

10.7, 10.8, 11.0, 11.1, 11.6, 11.8, 11.8, 11.9, 12.2, 12.2, 12.9, 13.3

Ans: mean= 11.775; median = 11.8; mode=11.8 and 12.2
In this case, the mode is not unique. Such datasets are also called *bimodal*.

Exercise: Check the mean of this sample using **your calculator** (now) changing the mode to *stat*.

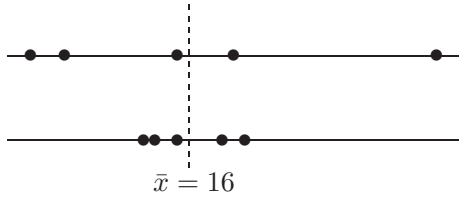
Exercise: Check the answers using R.

2.2.2 Sample Variance and Standard Deviation, p12

In order to motivate this topic, consider the following two sets of observations:

2, 5, 15, 20, 38

12, 13, 15, 19, 21



It is easy to verify that both sets have the same centre or the mean at $\bar{x} = 16$.

However, the two samples visually appear radically different. This difference lies in the greater *spread* or *variability*, or *dispersion* in the first dataset than the second. Therefore, we need a universal measure to find an indication of the amount of variation that a data set exhibits.
We will now describe the most popular measure of spread used in practice known as the sample variance based on n observations.

The Sample Variance

The difference between an observation and the sample mean is known as the 'deviation of the observation' from the sample mean. For example, in sample 1 the deviations from the mean are: $2 - 16 = -14$, $5 - 16 = -11$, $15 - 16 = -1$, $20 - 16 = 4$, $38 - 16 = 22$.
The sum of squared deviations divided by 4 is considered as a good measure of the spread and known as the *sample variance.*
For the above sample 1:
the variance= $\frac{(-14)^2+(-11)^2+(-1)^2+4^2+22^2}{4} = \frac{818}{4} = 204.5$.
Similarly, for the sample 2, the variance is 15. As seen from the data, the sample 1 has more variablity than the sample 2.

Calculation of the Sample Variance

For a set of n observations $x_1, x_2, \ldots, x_n$, the sample variance s^2 is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Note: It is easier to use the following calculation formula in practice. It can be shown after expanding the square term $(x_i - \bar{x})^2$ and re-arranging the terms that the above is equivalent to:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \text{ or } = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right].$$

Note: You do not need to memorize this formula as it is provided on a formula sheet available from the course web site.

Note: The above value is in squared units

Example: Find the mean and variance of the sample:

$$55, 48, 59, 64, 65, 57, 58, 41, 57, 59, 64, 62$$

Solution: $n =$ __. First calculate

$$\sum_{i=1}^{12} x_i = \underline{\qquad\qquad}$$

$$\sum_{i=1}^{12} x_i^2 = \underline{\qquad\qquad}$$

- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{12} x_i =$ _____

- Variance:

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] = \underline{\qquad\qquad}$$

Standard Deviation of a Sample

It is clear that the sample variance has squared units. Therefore, its square root will provide value in original units. This square root is known as the *sample standard deviation*.

Example: Find the standard deviation of the above sample.

Solution: Simply take the square root of the variance. Thus, the Standard Deviation is:

$$s = \underline{\qquad\qquad}$$

Notes:

- Many scientific calculators and computer packages (including `R`) can be used to find the standard deviation of a given dataset.

- Look at your calculator now:

  - Change the mode of your calculator to STAT (or similar depending on your calculator).

– Look for buttons $\bar{x}$, s^2 or σ^2. Many calculators have s^2_{n-1} or σ^2_{n-1} button for the sample sd. Check with the user manual for details.

- It can be proved that after a change in origin of a data set, the variance and standard deviation remain the same. If the sample points change in scale by a factor c, then the variance changes by a factor of c^2 and the sd changes by a factor of c.

Exercise: Consider the data set
110, 96, 118, 128, 130, 114, 116, 82, 114, 118, 128, 124. Show that the mean, variance and sd respectively are (approx) 114.84, 194.52, 13.95.

Note: the second data set is twice the first and hence the second mean is twice the first mean; second variance is four times the first variance and second sd is twice the first sd.

2.2.3 The Coefficient of Variation

The *coefficient of variation*, denoted CV, is the ratio of the standard deviation to the mean.

For a dataset with $\bar{x} \neq 0$, we define

$$\text{CV} = \frac{s}{\bar{x}}$$

This ratio of the standard deviation to the mean is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

Example: The CV for the previous dataset is

$$\text{CV} = \frac{s}{\bar{x}} = \underline{\qquad\qquad}$$

or the s.d. accounts for 12% of the mean.

Note: It is clear that the CV is *dimensionless* as it is a proportion. For example, it is not affected by multiplicative changes of scale. Therefore, the CV is a useful measure for comparing the dispersions of two or more variables that are measured on different scales.

The next section considers the corresponding results for grouped data.

2.3 Grouped Data (P.16-17)

Recall that large datasets can be summarised with a suitable frequency distribution table with k groups or intervals or bins like this:

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

Now we look the problem of calculating the mean and variance from such a frequenccy table.

2.3.1 The mean of Grouped Data

Suppose that we only have the information provided by a grouped frequency table for a data set. That is, we only have access to the published report and not the original data set. Let k be the number of bins (groups or intervals) and $u_1, u_2, \ldots, u_k$ be the *centres of each interval* with corresponding *frequencies* $f_1, f_2, \ldots, f_k$. Then an approximate sample mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i u_i.$$

Example: Consider the data on weight in pounds (recorded to the nearest pound) of 35 female students from week 1.

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

We have the frequency distribution from last week:

CLASS INTERVAL	CLASS CENTER	FREQUENCY
94-104	99	2
104-114	109	5
114-124	119	11
124-134	129	10
134-144	139	3
144-154	149	4
TOTAL		35

Find the grouped mean.

Solution: $n = 35$ (the number of values)

$$\sum_{i=1}^{6} f_i\, u_i \;=\; \underline{\qquad\qquad}$$

$$\Rightarrow \bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{6} f_i\, u_i = \underline{\qquad\qquad}$$

Exercise: Find the exact mean of the data and compare it to the above approximation.

Answer: Using the complete data, check with your calculator and R , sum of all 35 values=4333 and hence the exact mean, $\bar{x} =$ ____.

Note: The grouped mean and the exact mean are close to each other.

2.3.2 The Variance of Grouped Data

For data from a frequency table, the grouped sample variance is:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j(u_j - \bar{x})^2$$

or equivalently

$$s^2 = \frac{1}{n-1}\left[\sum_{j=1}^{k} f_j u_j^2 - \frac{1}{n}(\sum_{j=1}^{k} f_j u_j)^2\right] \overset{\text{or}}{=} \frac{1}{n-1}\left[\sum_{j=1}^{k} f_j u_j^2 - n(\bar{x}^2)\right].$$

Example:

Find the sample variance from the previous frequency distribution table of 35 female students.

Solution:

$$\sum_{i=1}^{6} f_i\, u_i^2 \;=\; \underline{\qquad\qquad\qquad\qquad}$$

$$\Rightarrow s^2 \;=\; \underline{\qquad\qquad\qquad\qquad}$$

$$\Rightarrow s \;=\; \underline{\qquad\qquad\qquad\qquad}$$

Example: Find the exact sample sd and compare with the grouped sd=13.35581.

solution: Check with your calculator and R the following:

$\sum x = 4333;\ \sum x^2 = 542505.$

Thus $s^2 = \frac{\sum x^2 - (\sum x)^2/35}{34} = \frac{542505 - 4333^2/35)}{34} = 178.8118$ and sd=13.37205.

Notice that these two values are also close to each other.

Exercise: Using the following frequecy table for 57 male students from week1 (p14), compute the grouped mean and sd using your calculator and R. Compare them with exact values.

CLASS INTERVAL	CLASS CENTER	FREQUENCY
122-136	129	6
136-150	143	17
150-164	157	17
164-178	171	7
178-192	185	8
192-206	199	1
206-220	213	1
TOTAL		57

Answer: Grouped mean=157.2456 and

grouped variance=367.4431. sd=19.16881.

Exact mean=9021/57 = 158.2632 and

Exact variance=$(1447141 - 9021^2/57)/56 = 347.3045$. sd=18.63611.

Additional worked example:

Consider the two samples:

Sample 1, x: 1.76, 1.45, 1.03, 1.53, 2.34, 1.96, 1.79, 1.21
Sample 2, y: 0.49, 0.85, 1.00, 1.54, 1.01, 0.75, 2.11, 0.92

For each of the two samples,

1. calculate the mean and the standard deviation,
2. find $Q_1,\ Q_2,\ Q_3,\ LT$ and UT,
3. find CV,
4. draw both boxplots on the same page.

Solution: In ascending order:

1. We have $n = 8$ is even and
$\sum_{i=1}^{8} x_i = 13.07,\ \sum_{i=1}^{8} x_i^2 = 22.5873,\ \sum_{i=1}^{8} y_i = 8.67,\ \sum_{i=1}^{8} y_i^2 = 11.2153$

Sample 1:

$$\text{The mean } \bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{13.07}{8} = 1.63$$

$$\text{The sd } s_x = \sqrt{\frac{1}{8-1}\left[\sum_{i=1}^{8} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{8} x_i\right)^2\right]}$$

$$= \sqrt{\frac{1}{7}\left[22.5873 - \frac{13.07^2}{8}\right]} = 0.42$$

Sample 2 :

$$\text{The mean } \bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{8.67}{8} = 1.08$$

$$\text{The sd } s_y = \sqrt{\frac{1}{8-1}\left[\sum_{i=1}^{8} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{8} y_i\right)^2\right]}$$

$$= \sqrt{\frac{1}{7}\left[11.2153 - \frac{8.67^2}{8}\right]} = 0.51$$

2. Sample 1 x_i: 1.03, 1.21, 1.45, 1.53, 1.76, 1.79, 1.96, 2.34
Sample 2 y_i: 0.49, 0.75, 0.85, 0.92, 1.00, 1.01, 1.54, 2.11

Sample 1: $Q_1 = 1.330;\ Q_2 = 1.645;\ Q_3 = 1.875$);
IQR $= Q_3 - Q_1 = 1.875 - 1.330 = 0.545$;
LT $= Q_1 - 1.5 \times IQR = 1.330 - 1.5(0.545) = 0.5125$
UT $= Q_3 + 1.5 \times IQR = 1.875 + 1.5(0.545) = 2.6925$
There is no outlier.

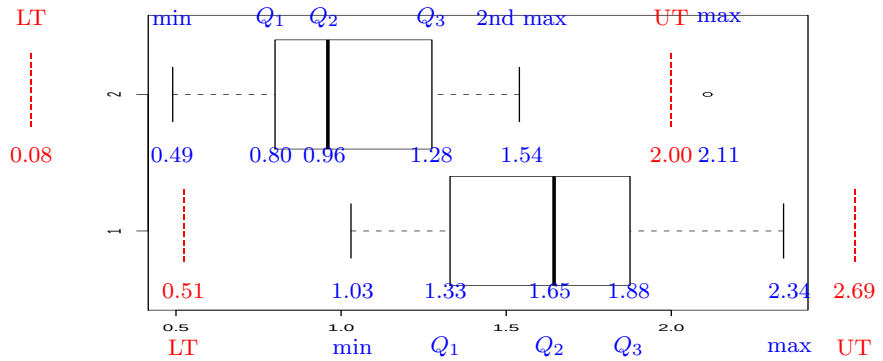
Sample 2: $Q_1 = 0.80;\ Q_2 = 0.96;\ Q_3 = 1.28.$

IQR $= Q_3 - Q_1 = 1.28 - 0.80 = 0.48$;

LT $= Q_1 - 1.5 \times IQR = 0.80 - 1.5(0.48) = 0.08$;

UT $= Q_3 + 1.5 \times IQR = 1.28 + 1.5(0.48) = 2.00$

Since the max = 2.11 lies outside (LT,UT) = (0.08,2.00). 3. CVs are 0.258 and 0.472 respectively. 4.



R commands:

```
mean(x)
sd(x)
sort(x)
median(x)
sd(x)/mean(x) #cv
fivenum(x)
boxplot(x,y) #2 boxplots side by side
```

where `x` and `y` are vectors of measurements.

In order to develop further concepts and applications of biostatistics, it is convenient to understand the basic theory of probability. Now we look at this topic.

3 An Introduction to Probability Theory and Applications, P29

This chapter considers the following topics:

- Basic terminology,
- Theory of sets and Venn diagrams,
- Probability axioms and counting methods,
- Conditional probability and independence.

Preliminaries

- The word *fair* or *unbiased* is regularly used in many life science situations. This means that all possible outcomes of an experiment have the same chance to occur.
- Any experiment to collect information is called a *random experiment*, if we are not certain or cannot predict of its outcome(s).

It is clear that in a random experiment, one cannot state (before the experiment) what a particular outcome will be.

Note: On contrary, a *deterministic* experiment yields known or predictable outcomes when repeated under the same conditions.

For example, consider the following experiments:

1. Toss a fair six-sided die once and observe the number that shows on top.
2. Take a marble from a bag containing 2 red, 1 black and 1 white balls and observe its colour.

It is clear that in these random experiments, one cannot state (before the experiment) what a particular outcome will be at each throw. However, we can make a list of all possible outcomes.

For example:

1. In 1, we observe one of: 1 or 2 or 3 or 4 or 5 or 6.
2. In 2, we observe one colour from: red or black or white.

Now we provide the following definition for later reference:

Definition: The collection (or the set) of all possible outcomes of a random experiment is called the *sample space*. This is denoted by S or Ω and be written as $S = \{\cdots\}$.

For example,

1. in experiment 1 above, $S =$ _____.
2. in experiment 2 above, $S =$ _____.

The following terminology will be useful in many applications:

Definition: An *event* of a random experiment is a collection of outcomes with specified or interested features.

Example: List the event A of observing a number less than 3 in experiment 1 above.

Ans: $A =$ _____.

Example: A card is selected at random from a box containg 10 cards with numbers 1 to 10. List the events: A of observing even numbers and B of observing numbers divisible by 4.

Ans: $A =$ _____; $B =$ _____.

3.1 Probability of equally likely outcomes/events

First consider the concept of equally likely outcomes.

Equally Likely Outcomes: The outcomes of a random experiment (or in a sample space) are called equally likely if all of them have the same chance of occurrence.

In a historical note, the probability was considered as the chance of an event to occur which expresses the strength of one's belief. Therefore, this was known as *subjective probability*. However, this was later developed with a number of common concepts including equally likely outcomes. Therefore, we have the following definion:

Definition: The *probability* of an event A is the relative frequency of its set of outcomes over an indefinitely large number of repeated trials under identical conditions. This is denoted by $P(A)$.

Calculating Probabilities

Suppose we have a random experiment, which has exactly n total possible *equally likely* outcomes. Let A be an event of interest within this sample space containing m number of simple outcomes. Then the probability assigned to A, $P(A)$ is given by:

$$P(A) = \frac{m}{n}.$$

Examples:

1. Throw a fair six-sided die. There are 6 equally likely possible outcomes. The sample space, S of this experiment is

$$S = \_\_\_\_\_.$$

If A denotes the event of observing an even number, then

_____.

$$\text{Prob(an even number)} = P(A) = \_\_\_\_\_.$$

2. Toss a fair coin 3 times. There are 8 possible equally likely outcome and the sample space is

$S =$ _____.

- Let A be the event of observing exactly two heads in this experiment. Then $A =$ _____ and the probability of observing exactly two heads is

$$P(A) = \text{.....}$$

- Let B be the event of observing at least one head. Then the event is $B =$ _____.
  Hence, the probability of observing at least one head is

$$P(B) = \underline{\ldots\ldots}$$

3.2 Probability using tree diagrams, p33

Probability Trees or Tree Diagrams can be used to visualize the events and to calculate simple probabilities.

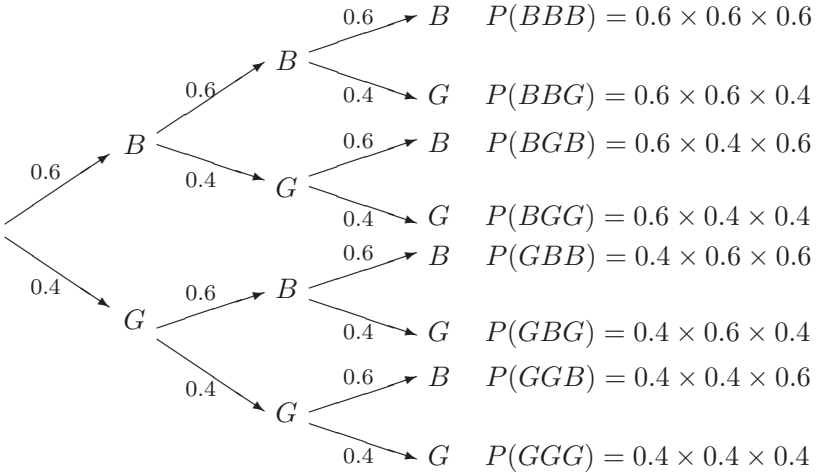
Example: Draw a suitable tree diagram for the experiment of tossing a fair coin two times. Hence list the sample space.

Exercise: Draw a tree diagram for the experiment of tossing a fair coin three times.

Example: A certain country reports that it has a higher rate of male births with probability of a boy is 0.6. Assuming the births are random, (i) draw a tree diagram to repersent the distribution of children in families with three children; (ii) find the probability that there are (a) at most one boy and (b) at least one boy in a family of three children.

Solution (i):

Tree diagram for the distribution of gender of three children



Solution (ii):

(a) $P(\text{at most 1 boy}) =$ _____
$= \underline{\ldots\ldots}$

(b) $P(\text{at least 1 boy}) =$ _____
$= \underline{\ldots\ldots}$