Chapter 2 - Descriptive Statistics

Our ultimate aim in this class is to study Inferential statistics but before we can start making decisions about data we need to know how to describe this data in a useful manner. So in Chapter 2 we are going to concentrate on learning some Descriptive statistics. This should take us around 5 or 6 weeks and will provide us with most of the tools we will need for the main part of this course ie: Inferential Statistics.

Some of the terminology in this chapter may be quite familiar but there will be very specific definitions here which may be different from what you have seen before.
So the next 5 or 6 weeks will look like this:

2.1 Describing Categorical Data -
Bar Graphs, Pie Charts etc

2.2 Describing Numerical Data-
Dot Plots, Histograms, Stem and Leaf Diagrams

2.3 Summation Notation

2.4 Measuring the Centre
Mean, Median, Mode

2.5 Measuring Variability
Range, Variance, Standard Deviation

2.6 What is Standard Deviation?
Chebyshev and the Empirical Rule

2.7 Relative Standing (How do you compare?)
Percentile, Z-Score

2.8 Quartiles and Box Plots
Quartiles, IQR, Box-Plots

2.9 Lying with statistics
Be careful when looking at graphs
SECTION 2.1 Describing Categorical Data - Bar Graphs, Pie Charts etc

2.1.1 Definitions

1. A Class is one of the categories into which categorical data can be classified.
2. Class Frequency is the number of observations which fall into a particular class.
3. Class Relative Frequency is the proportion of observations in a particular class.

2.1.2 Example

Premiership results weekend Sept 11th
Arsenal  3   Villa   1
Chelsea  1   Newcastle  0
Coventry 3   Leeds   4
Liverpool 2   Man U  3
Middlesb 3   Southam  2
Shef Wed 0   Everton  2
Sunderland 2   Leicester  0
West Ham 1   Watford  0
Wimbledon 2   Derby  2
Bradford 1   Spurs  1

Classes: Win, Lose, Draw
<table>
<thead>
<tr>
<th>Class</th>
<th>Class Frequency</th>
<th>Class Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIN</td>
<td>8</td>
<td>8/20 = 0.4</td>
</tr>
<tr>
<td>LOSE</td>
<td>8</td>
<td>8/20 = 0.4</td>
</tr>
<tr>
<td>DRAW</td>
<td>4</td>
<td>4/20 = 0.2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

### 2.1.3 Example
Students in this class belong to 3 different groups: 2nd Science, 1st Arts & 1st Actuary
Create a Frequency/Relative Frequency table for the each row and for the class

### 2.1.4 Bar Graphs/Histograms for Categorical Data
Bar graphs are easy to draw it just takes 4 steps:

1. Draw Horizontal and Vertical axes intersecting at the bottom left corner of the Graph
2. Write the category names on the horizontal axis
3. Mark a scale on the vertical axis appropriate to either the Class Frequencies or Class Relative Frequencies
4. Above each category label construct a rectangle whose height is the corresponding Frequency or Relative Frequency

**NOTE:** Each rectangle should have the same base width.
### 2.1.5 Example Rainfall in Ireland 1996

<table>
<thead>
<tr>
<th>Location</th>
<th>Rainfall (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belmullet</td>
<td>1,136</td>
</tr>
<tr>
<td>Birr</td>
<td>837</td>
</tr>
<tr>
<td>Cork</td>
<td>1,433</td>
</tr>
<tr>
<td>Dublin</td>
<td>787</td>
</tr>
<tr>
<td>Shannon</td>
<td>886</td>
</tr>
<tr>
<td>Malin</td>
<td>926</td>
</tr>
</tbody>
</table>

![Bar chart showing rainfall frequency and amounts for different locations in Ireland in 1996.](chart.png)
2.1.6 Example
Histogram showing frequencies of Students in your row belonging to each of the 3 different faculties

2.1.7 Example
Histogram showing Relative Frequencies for Favourite subject among students in your row

2.1.8 Pie Charts
An alternative to the Histogram is the Pie Chart, here the 360 degrees of a disk are divided up in arcs representing the relative frequencies of the different categories in the dataset.
The angle subtended at the centre of the disk is calculated by multiplying each relative frequency by 360 degrees.

*** 2.1.9 Example
200 First year Commerce students were asked which of 3 subjects they preferred
130 said Accounting, 20 Economics and 50 Management

<table>
<thead>
<tr>
<th>Subject</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>130</td>
<td>0.65</td>
<td>360*.65 = 234</td>
</tr>
<tr>
<td>Economics</td>
<td>20</td>
<td>0.10</td>
<td>360*.10 = 36</td>
</tr>
<tr>
<td>Management</td>
<td>50</td>
<td>0.25</td>
<td>360*.25 = 90</td>
</tr>
</tbody>
</table>
2.1.10 Example
Draw a Pie Chart showing Relative Frequencies from Example 2.1.6 showing the distribution of students in your row among the 3 different faculties.
Section 2.2 Describing Numerical Data-
Dot Plots, Histograms, Stem and Leaf Diagrams

2.2.0 Dot Plots
A long time ago before computers could draw decent graphs they used to draw Dot Plots as a way of describing numerical data. The range of values in the dataset are marked on an X-axis and then a dot is placed above the relevant point on the axis for each value in the dataset. If two or more observations have the same value then dots are stacked on top of each other. You should use minitab to generate some Dot Plots.

*** 2.2.1 Example
Draw a Dot Plot for the following dataset

50 35 70 55 50 30 40 65 50 75 60 45 35 75 60 55 55 50 40 55 50

* * *
* * *
* * *
* * *
* * *
* * *

30 35 40 45 50 55 60 65 70 75
2.2.3 Example
Construct a dot plot using the ages of the 6 people sitting closest to you now as the data set.

2.2.4 Stem and Leaf Diagrams
Stem and Leaf Diagrams are graphical ways to display a group of integers in a dataset.

Steps for Constructing a Stem and Leaf Diagram
1. Select one or more of the leading digits to be the Stem values, the remaining digits become the Leaves.
2. List Possible Stem values in a column
3. Record the Leaf for every observation beside the corresponding Stem value.
4. Indicate on the display what units are used for the Stems and Leaves.

*** 2.2.5 Example
The following are a selection of exam marks
71 52 52 75 64 60 48 56
67 29 11 53 25 46 58 46
49 62 66 40 19 54 57 54
60 19 59 43 51 40 21 45
46 62 73 59 36 45 55 46
45 32 55 46 51 46 65 49 61 40
A Stem And Leaf Diagram will look like this:

1 | 1 9 9
2 | 1 5 9
3 | 2 6
4 | 0 0 0 3 5 5 5 6 6 6 6 6 6 6 8 9 9
5 | 1 1 2 2 3 4 4 5 5 6 7 8 9 9
6 | 0 0 1 2 2 4 5 6 7
7 | 1 3 5

STEM UNIT = TENS
LEAF UNIT = ONES

2.2.6 OSCAR Winners 1928 -1988
Construct a stem and leaf diagram for the ages of the Best Actress Oscar winners between 1928 and 1988.

Comparing the Stem and Leaf diagrams for Best Actress and Best Actor Winners, can we draw any conclusions?

INSERT
2.2.7 Histogram for Discrete Numerical Data

1. Draw a horizontal X-axis and on it mark the possible values taken by the observations.

2. Draw a vertical Y-axis marked with either relative frequencies or frequencies.

3. Above each possible value on the X-axis draw a rectangle centred on the value with width 1 and height equal to the relative frequency or frequency of that value.
### 2.2.8 Example

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>40</td>
<td>150</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
</tr>
</tbody>
</table>

![Histogram showing frequency distribution]
2.2.9 Frequency Distributions for Continuous Data
Creating a Frequency distribution for discrete data is as we have seen very simple. However continuous data presents us with some problems. Given a set of REAL numbers as opposed to INTEGERS you may find that the dataset does not contain 2 numbers which are the same. If we then create a frequency distribution as before we will find that each observation value has frequency 1 which is meaningless.

2.2.9A Example
A meaningless frequency distribution:

Dataset: 12.0 12.3 13.1 14.2 11.5 12.7

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.5</td>
<td>1</td>
</tr>
<tr>
<td>12.0</td>
<td>1</td>
</tr>
<tr>
<td>12.3</td>
<td>1</td>
</tr>
<tr>
<td>12.7</td>
<td>1</td>
</tr>
<tr>
<td>13.1</td>
<td>1</td>
</tr>
<tr>
<td>14.2</td>
<td>1</td>
</tr>
</tbody>
</table>

It makes much more sense when dealing with continuous numerical data to define **Class Intervals** on the REAL line which may contain several observations which are close together if not exactly the same.
2.2.9B Example
A more Meaningful Frequency Distribution

Using the same dataset as before:
**Dataset: 12.0 12.3 13.1 14.2 11.5 12.7**

This time we split up the observations into 4 intervals instead of looking at individual values.

These Class Intervals are:

\([11.0, 12.0)\)   \([12.0, 13.0)\)   \([13.0, 14.0)\)   \([14.0, 15.0)\]

which we could also write as

\(11.0-<12.0\)   \(12.0-<13.0\)   \(13.0-<14.0\)   \(14.0-<15.0\)

So the Frequency Distribution looks like this:

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.0-&lt;12.0</td>
<td>1</td>
</tr>
<tr>
<td>12.0-&lt;13.0</td>
<td>3</td>
</tr>
<tr>
<td>13.0-&lt;14.0</td>
<td>1</td>
</tr>
<tr>
<td>14.0-&lt;15.0</td>
<td>1</td>
</tr>
</tbody>
</table>

**NOTE:**
The value 12.0 belongs to the interval 12.0-<13.0 and not to the interval 11.0-<12.0.
2.2.10 Relative Frequency Distributions for Continuous Data

As with Discrete or Categorical data we can define the relative frequency of a particular Class Interval to be the Frequency of that Interval divided by the Total number of observations in the dataset.

*** 2.2.10A Example

The following dataset represents Zinc intake in mg for a sample of 40 patients with Rheumatoid Arthritis:

8.0 12.9 13.0 8.9 10.1 7.3 11.1 10.9 6.2 8.1 8.8 10.4
15.7 13.6 19.3 9.9 8.5 11.1 10.7 8.8 10.7 6.8 7.4 4.8
11.8 13.0 9.5 8.1 6.9 11.5 11.2 13.6 4.9 18.8 15.7 10.8
10.7 11.5 16.1 9.9

We will define class intervals as:
3-<6  6-<9  9-<12  12-<15  15-<18  18-<21

This gives us the following Frequency/Relative Frequency Distribution:

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-&lt;6</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>6-&lt;9</td>
<td>12</td>
<td>0.3</td>
</tr>
<tr>
<td>9-&lt;12</td>
<td>16</td>
<td>0.4</td>
</tr>
<tr>
<td>12-&lt;15</td>
<td>5</td>
<td>0.125</td>
</tr>
<tr>
<td>15-&lt;18</td>
<td>3</td>
<td>0.075</td>
</tr>
<tr>
<td>18-&lt;21</td>
<td>2</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Looking at this table we can now answer some questions, such as:

A. How many individuals have a Zinc intake of less than 12 mg?
B. What proportion of individuals have a Zinc intake between 9 and 15 mg?

Answers:
A. $2 + 12 + 16 = 30$
B. $0.4 + 0.125 = 0.525$

2.2.11 Cumulative Relative Frequencies

Cumulative Relative Frequencies measure the proportion of observations falling below a specified value. The Cumulative Relative Frequency for a particular Class Interval is calculated by summing up the Relative Frequencies for that Class Interval together with the Relative Frequencies for all previous Class Intervals
**2.2.11A Example**
The following shows CRFs for the Zinc intake example

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Relative Freq</th>
<th>Cum Rel Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-&lt;6</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>6-&lt;9</td>
<td>0.3</td>
<td>0.05+0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0.35</td>
</tr>
<tr>
<td>9-&lt;12</td>
<td>0.4</td>
<td>0.05+0.3+0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0.35 + 0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0.75</td>
</tr>
<tr>
<td>12-&lt;15</td>
<td>0.125</td>
<td>0.75 + 0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0.875</td>
</tr>
<tr>
<td>15-&lt;18</td>
<td>0.075</td>
<td>0.875 + 0.075</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 0.95</td>
</tr>
<tr>
<td>18-&lt;21</td>
<td>0.05</td>
<td>0.95 + 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 1.0</td>
</tr>
</tbody>
</table>

**2.2.12 Example**
Below is a dataset containing the current earnings of some Statistics graduates. Construct a Cumulative Relative Frequency Distribution for this dataset and identify what proportion of graduates earn less than £20,000 and what proportion earn more than £30,000. Use Class Intervals of width £5,000
### Data in IR£000’s:

14.1  13.8  32.7  25.1  38.9  17.8  16.5  31.2  24.6  
27.5  19.4  16.9  18.4  21.6  23.8  15.6  13.7  14.9  
16.4  42.7  36.8  28.1  19.8  29.8  30.3  22.5  23.2

<table>
<thead>
<tr>
<th>C Interval</th>
<th>Freq</th>
<th>Rel Freq</th>
<th>CR Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-&lt;15</td>
<td>4</td>
<td>.148</td>
<td>.148</td>
</tr>
<tr>
<td>15-&lt;20</td>
<td>8</td>
<td>.296</td>
<td>.444</td>
</tr>
<tr>
<td>20-&lt;25</td>
<td>5</td>
<td>.185</td>
<td>.629</td>
</tr>
<tr>
<td>25-&lt;30</td>
<td>4</td>
<td>.148</td>
<td>.777</td>
</tr>
<tr>
<td>30-&lt;35</td>
<td>3</td>
<td>.111</td>
<td>.888</td>
</tr>
<tr>
<td>35-&lt;40</td>
<td>2</td>
<td>.074</td>
<td>.962</td>
</tr>
<tr>
<td>40-&lt;45</td>
<td>1</td>
<td>.037</td>
<td>.999 ~ 1</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.2.13 Histograms for Continuous Data when Class Intervals have Equal Width

When we constructed Frequency/Relative Frequency Distributions for Continuous Data so far the Class Intervals we used all had the same widths. When this is the case it is easy to go one step further and draw Histograms for such data. These Histograms are drawn in almost exactly the same way as all the previous ones.

1. Mark the ends of each Class Interval on the horizontal X-axis
2. Mark either the Frequencies or Rel Freq.s on the vertical Y-axis
3. Draw a rectangle for each class directly above the corresponding interval so that the sides of the rectangles are at the ends of the Class Intervals which you have marked on the X-Axis.

*** 2.2.13A Example

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Freq</th>
<th>Rel Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-&lt;10</td>
<td>5</td>
<td>0.125</td>
</tr>
<tr>
<td>10-&lt;20</td>
<td>11</td>
<td>0.275</td>
</tr>
<tr>
<td>20-&lt;30</td>
<td>10</td>
<td>0.250</td>
</tr>
<tr>
<td>30-&lt;40</td>
<td>9</td>
<td>0.225</td>
</tr>
<tr>
<td>40-&lt;50</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>50-&lt;60</td>
<td>1</td>
<td>0.025</td>
</tr>
<tr>
<td>60-&lt;70</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>1.0</td>
</tr>
</tbody>
</table>
2.2.14 Example
Construct a Frequency Histogram for the salaries of Statistics graduates in Example 2.2.12

2.2.15 Histograms of Continuous Data when the Class Interval Widths are Unequal
In all of the previous examples of Histograms the height of each rectangle represented the Frequency or Relative Frequency. Now the area of each rectangle is just the base multiplied by the height. Since all rectangles had the same base we find that the area of each rectangle was proportional to the Frequency or Relative Frequency.
It is not always necessary to have the bases of the rectangles be the same width, in many cases it makes more sense to have the bases of some rectangles be wider than others. If we do this however we must change the way that we draw Histograms.

**From now on we are going to draw Histograms so that the AREA of each rectangle represents the Relative Frequency.**

If the area is the Relative frequency we get the following relationship:

Relative Frequency = Area = Base * Height

So we find

Height= (Relative Frequency)/Base

This new Height measurement we will call Density and since the Base of each rectangle is actually the width of a Class Interval we end up with a final equation which we will use from now on.

Rectangle Height = Density =
(Relative Frequency)/(Class Interval Width)
### 2.2.16 Example Misrepresentation of Grades

Many times when people are asked what their results were on exams the truth gets distorted. People who have not done very well tend to add a few percent to their exam results “just to make them respectable”. And similarly some people who do extremely well on exams tend to underreport their exams “for fear of being seen as Brainiacs”. The following table shows the difference in some peoples reported marks from their actual, ie:

\[
\text{the observation} = \text{reported score} - \text{actual score}
\]

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Relative Frequency</th>
<th>CI Width</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>-50% -&lt; -10%</td>
<td>0.023</td>
<td>40</td>
<td>0.000575</td>
</tr>
<tr>
<td>-10% -&lt; -5%</td>
<td>0.055</td>
<td>5</td>
<td>0.011</td>
</tr>
<tr>
<td>-5% -&lt; -2.5%</td>
<td>0.097</td>
<td>2.5</td>
<td>0.0388</td>
</tr>
<tr>
<td>-2.5% -&lt; 0%</td>
<td>0.21</td>
<td>2.5</td>
<td>0.084</td>
</tr>
<tr>
<td>0 -&lt; 2.5%</td>
<td>0.189</td>
<td>2.5</td>
<td>0.0756</td>
</tr>
<tr>
<td>2.5% -&lt; 5%</td>
<td>0.139</td>
<td>2.5</td>
<td>0.0556</td>
</tr>
<tr>
<td>5% -&lt; 10%</td>
<td>0.116</td>
<td>5</td>
<td>0.0232</td>
</tr>
<tr>
<td>10% -&lt; 50%</td>
<td>0.171</td>
<td>40</td>
<td>0.004275</td>
</tr>
</tbody>
</table>

INSERT EXAMPLES of Correct/Incorrect Histogram for this data
*** Example 2.2.17 Marijuana Usage
A telephone survey was conducted on marijuana usage. The frequency distribution gives the amount of marijuana in grams smoked per week for those drug fiends (respondents) who indicated they used the drug.

<table>
<thead>
<tr>
<th>Amt Smoked</th>
<th>Frequency</th>
<th>Rel Freq</th>
<th>Cum Rel Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 -&lt; 3</td>
<td>94</td>
<td>0.188</td>
<td>0.188</td>
</tr>
<tr>
<td>3 -&lt; 11</td>
<td>269</td>
<td>0.538</td>
<td>0.726</td>
</tr>
<tr>
<td>11 -&lt; 18</td>
<td>70</td>
<td>0.14</td>
<td>0.866</td>
</tr>
<tr>
<td>18 -&lt; 25</td>
<td>48</td>
<td>0.096</td>
<td>0.962</td>
</tr>
<tr>
<td>25 -&lt; 46</td>
<td>10</td>
<td>0.02</td>
<td>0.982</td>
</tr>
<tr>
<td>46 -&lt; 60</td>
<td>7</td>
<td>0.014</td>
<td>0.996</td>
</tr>
<tr>
<td>60 -&lt; 74</td>
<td>2</td>
<td>0.004</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Calculate the Relative Frequencies & CRF for this dataset.

Draw a Histogram of this data.

What proportion of respondents smoked less than 25g per week?

Approximately what proportion of respondents smoked more than 53g per week?
Answer:

<table>
<thead>
<tr>
<th>Amt Smoked</th>
<th>Rel Freq</th>
<th>CI width</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 -&lt; 3</td>
<td>0.188</td>
<td>3</td>
<td>0.06267</td>
</tr>
<tr>
<td>3 -&lt; 11</td>
<td>0.538</td>
<td>8</td>
<td>0.06725</td>
</tr>
<tr>
<td>11 -&lt; 18</td>
<td>0.14</td>
<td>7</td>
<td>0.02</td>
</tr>
<tr>
<td>18 -&lt; 25</td>
<td>0.096</td>
<td>7</td>
<td>0.0137</td>
</tr>
<tr>
<td>25 -&lt; 46</td>
<td>0.02</td>
<td>21</td>
<td>0.000952</td>
</tr>
<tr>
<td>46 -&lt; 60</td>
<td>0.014</td>
<td>14</td>
<td>0.001</td>
</tr>
<tr>
<td>60 -&lt; 74</td>
<td>0.004</td>
<td>14</td>
<td>0.000286</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 2.2.18 A final word on Histograms

The general shape of a histogram is important. The number of peaks in the histogram determines whether a distribution is classed as Unimodal, Bimodal or Multimodal.
In addition to this classification we can further classify UniModal distributions as to whether they are symmetric or not.

A unimodal distribution is defined to be Symmetric if there is a vertical line of symmetry through the middle of the distribution such that the distribution to the left of this line is the mirror image of the distribution to the right of this line.

These 3 distributions are symmetric:

![Symmetric Distributions](image)

The right part of a unimodal distribution is called the Upper Tail of the distribution while the left part is called the lower tail:

![Upper and Lower Tails](image)
A Unimodal distribution which is not symmetric is called skewed, there are two types of skewness.

**Positive Skew:** If the upper tail of the distribution stretches out more than the lower tail then the distribution is said to be positively skewed.

![Positive Skew Diagram]

**Negative Skew:** If the Lower tail of the distribution stretches out more than the upper tail then the distribution is said to be negatively skewed.

![Negative Skew Diagram]
OK that’s it for histograms! 2.3 is a short section:

Section 2.3 Summation Notation

Sometimes we want to refer to the numbers in a general numerical dataset. We could talk about “the first number/observation in the dataset” and “the second observation in the dataset” but this is very longwinded and after a while we’d get fed up writing all this down. So instead we will use a shorthand notation:

\[ X_1 = \text{The 1}\text{st Observation in the dataset} \]
\[ X_2 = \text{The 2}\text{nd Observation in the dataset} \]
\[ X_3 = \text{The 3}\text{rd Observation in the dataset} \]
\[ \ldots \]
\[ X_n = \text{The n}\text{th or Last Observation in the dataset} \]

ExampleA:

Dataset: 23, 43, 56, 98, 190, 3, 11, 21, 56

\[ X_1 = 23 \]
\[ X_3 = 56 \]
\[ X_n = X_9 = 56 \quad \text{as n}=9 \]
Sometimes we want to find the sum of all the observations (numbers) in the dataset.

\[ X_1 + X_2 + X_3 + X_4 + X_5 + \ldots + X_n \]

We can also use a shorthand notation for writing down this sum:

\[ \sum_{i=1}^{n} X_i \]

\[ = X_1 + X_2 + X_3 + X_4 + X_5 + \ldots + X_n \]

**Example B** using the same dataset as in example A

**calculate** \[ \sum_{i=1}^{n} X_i \]

**Dataset:** 23, 43, 56, 98, 190, 3, 11, 21, 56

\[ \sum_{i=1}^{n} X_i = 23 + 43 + 56 + 98 + 190 + 3 + 11 + 21 + 56 \]

\[ = 501 \]
Another calculation that we want to do quite often is to square each number in the dataset and then add up the answers. The shorthand way to write that down is:

\[ \sum_{i=1}^{n} X_i^2 \]

\[ = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 + \ldots + X_n^2 \]

**Example C**
For the following dataset we get

Dataset: 2, 3, 7, 4, 6, 5

\[ \sum_{i=1}^{n} X_i^2 \]

\[ = 2^2 + 3^2 + 7^2 + 4^2 + 6^2 + 5^2 \]
\[ = 4 + 9 + 49 + 16 + 36 + 25 \]
\[ = 139 \]

Note: A final shorthand is sometimes seen, this is really just laziness or sloppiness but sometimes the top and bottom of the sigma are left out so you get:

\[ \sum X \] instead of \[ \sum_{i=1}^{n} X_i \]
2.3.1 Example
Dataset: 5, 1, 2, 2, 1
Find
\[ \sum X \quad \sum X^2 \quad \sum (X - 1) \]
\[ \sum (X - 1)^2 \quad (\sum X)^2 \]

Answers:
12, 40, 7, 21, 144

MORE EXAMPLES to be done:
2.26, 2.27, 2.28 on page 43 of the Textbook
Section 2.3A Sampling a preview

In Chapter 6, which is covered in the other part of this course the theory behind Sampling distributions will be covered in detail. But since inferential statistics is based on sampling, and since it will be a while before you reach Chapter 6 perhaps a preview is a good idea. We have seen some of the basic concepts in Chapter 1. Remember the definitions of a Population and a Sample.

Our aim in Inferential statistics is to make a measurement about some Characteristic (property) of the Population but usually the Population is too large for us to perform the required measurement.

So instead we take a Representative Sample from the Population and measure the value taken by the Sample Statistic which corresponds to Population Characteristic we are interested in.

So we have two groups of Objects: The Population and The Sample.

And we have two corresponding measures associated with these: The Population Characteristic and The Sample Statistic
An example of this would be
Population = Entire population of Ireland

Sample = A selection of 1000 Irish people chosen at random.

Population Characteristic: We are interested in measuring the Mean (average) Age of the Population of Ireland. We use the Greek letter $\mu$ (pronounced mu) to represent this Population Mean.

As already mentioned it would take too much time and effort for us to measure $\mu$.

Sample Statistic: So Instead we measure the corresponding Mean age for the sample of 1000 people. We use $\bar{x}$ (x bar) to represent this Sample Mean.

In this example and in general both $\mu$ and $\bar{x}$ represent the same concept the only difference being that $\mu$ refers to the population and $\bar{x}$ to the sample.

The next two sections will examine which Sample Statistics best measure the Centre of a Population and the Variability of a Population.