THE UNIVERSITY OF SYDNEY

Business School

Statistics

Mathematics Help Sheet

The University of Sydney Business School

Representation of Data

Frequency Distributions

A frequency distribution is a representation (in tabular or graphical form) of data, illustrating the number of instances in which a variable takes on each of its possible values.

An example of a frequency distribution is as follows

Variable	Frequency
A	12
B	5
C	15
D	7
E	9
F	21

A frequency distribution with class intervals is usually used when there is a large data set involving many observations.

For example,

Age	Frequency
0-9	18
10-19	5
20-29	10
30-39	4
40-49	7
50-59	12

In the above example, the width of each interval is 10 years

Relative Frequency

Relative frequency is found by dividing the frequency of a particular observation (or class interval) by the total number of observations,

$$Relative\ frequency = \frac{f}{n}$$

Frequency Density

Frequency density is found by dividing the frequency of a class interval by the length of that class interval,

$$Frequency\ density = \frac{f}{l}$$

Cumulative Frequency Distribution

A cumulative frequency distribution shows how many observations lie below each value and is obtained by adding each frequency in a frequency distribution to the sum of the preceding cumulated frequency.

For example,

Age	Frequency	Cumulative Frequency
0-9	18	18
10-19	5	23
20-29	10	33
30-39	4	37
40-49	7	44
50-59	12	66

Percentiles and Quantiles

A percentile is a score below which a percentage of observations lie, and a quantile is its equivalent as a decimal. For example, the **75th percentile** is denoted as P_{75} and represents the value under which 75% of all observations lie.

The 75th percentile corresponds to the 0.75 quantile which similarly represents the value under which 75% of observations lie.

Quartiles and Interquartile Range

Quartiles divide data into 4 equal parts, with P_{25} forming the first quartile, P_{50} forming the second, and P_{75} forming the third. These quartiles are denoted as Q_1, Q_2 and Q_3 respectively. Note that Q_3 is also the median of the data set.

The interquartile range (IQR) is the difference between the third and first quartile and represents the distance between the 'middle' 50% of the data. That is, it is $Q_3 - Q_1$.

For example, given the data points (1, 3, 5, 7, 9, 11, 13, 15, 17, 19)

$$Q_1 = 5$$

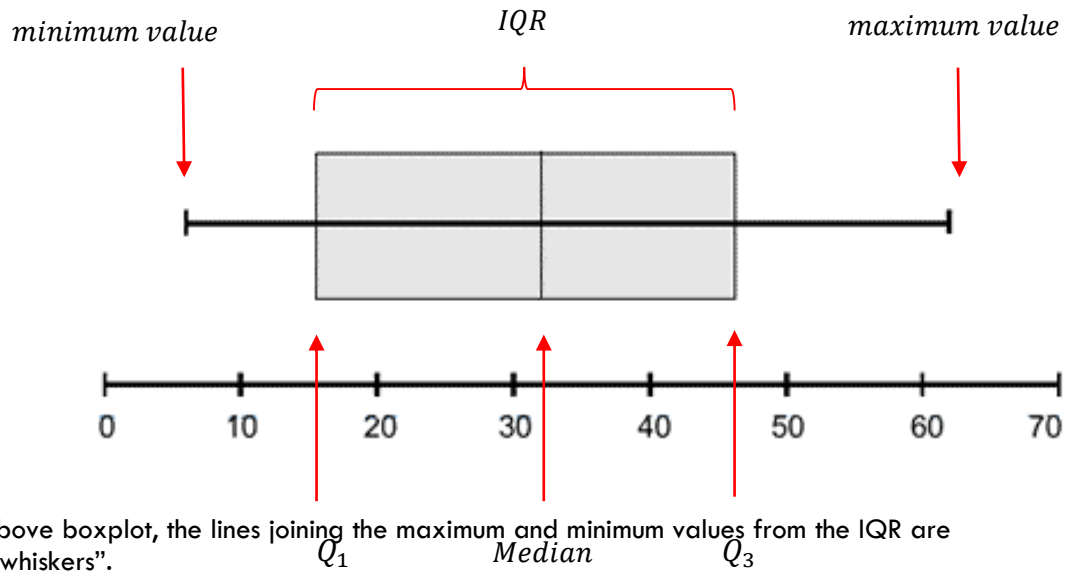
$$Q_3 = 15$$

$$IQR = 10$$

Boxplots

A boxplot is used to graphically illustrate the range, median, quartiles and interquartile range of data. Hence, in order to construct a boxplot, you need to first find the range, the median and the interquartile range of the dataset.

Graphically,



In the above boxplot, the lines joining the maximum and minimum values from the IQR are called "whiskers".

Outliers

A boxplot can also be useful for identifying whether or not there are any **outliers**. An outlier is an observation in the data which is very distant from the other data points and hence has the effect of skewing measures of the data.

The general rule of determining that a data point is an outlier is if the whisker of a box plot is more than **1.5 times** the length of the interquartile range. Hence, in the above boxplot, there are no outliers.

Measures of Central Tendency

Mean

The mean is a useful measure of data as every set of data has a single, **unique mean.** It can be used to compare different sets of data, and being a measure of the "average", it is a concept familiar to most people.

However, the mean can be **affected by outliers** not representative of the majority of data, and where a dataset has extreme outliers, the means ceases to become a good measure of the data.

To find the mean of a sample of data, divide the sum of all observations by the total number of observations,

$$Mean = \frac{sum\ of\ all\ observations}{number\ of\ observations}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

In order to find the mean of a frequency distribution,

$$\bar{x} = \frac{\sum f_i x_i}{n}$$

That is, in the numerator, you must find the sum of all observations multiplied by its frequency.

Median

The median is simply the middle value of a data set where the data has been arranged in either ascending or descending order of magnitude. Where the "middle" is between two values, take the average of those two values. Where there are extreme outliers in the dataset, the median may be a better measure of the data than the mean.

For example, for a data set (1, 3, 7, 7, 9, 9, 12, 15, 180),

The mean of the data is found by,

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{x} = \frac{243}{9}$$

$$\bar{x} = 27$$

The median of the data is found by,

$$(1, \quad 3, \quad 7, \quad 7, \quad 9, \quad \mathbf{9}, \quad 12, \quad 15, \quad 180)$$

Thus the median of the data is 9, which is a much better representation of the data than the mean is in this case, because of the outlier of 180.

Mode

The mode of the data is simply the value that occurs the most often. While both the mean and the median require values which are numerical, the mode does not. A dataset can have more than one mode if the values appear an equal number of times.

For example,

Consider the dataset (1, 5, 6, 6, 8, 8, 12): the modes of the dataset are **6 and 8**.

Consider the dataset (red, blue, green, blue): the mode of the dataset is **blue**.

Measures of Spread

Range and Interquartile Range

The **range** of a dataset is a good measure of its spread, and is given by the maximum value subtracted by the minimum value.

Where there may be extreme outliers in a dataset, the **interquartile range** may be a better measure of spread, more representative of the data, and is given by $Q_3 - Q_1$.

Standard Deviation and Variance

The variance and standard deviation of a dataset measures the variation or dispersion of the data.

Standard deviation is denoted as σ (sigma) and is widely used in statistics.
Variance is simply the standard deviation squared and is hence denoted as σ^2.

The standard deviation of a set of data (or a sample) is found by,

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

That is, the standard deviation is the square root of the sum of all deviations from the mean squared, divided by the total number of observations minus 1.

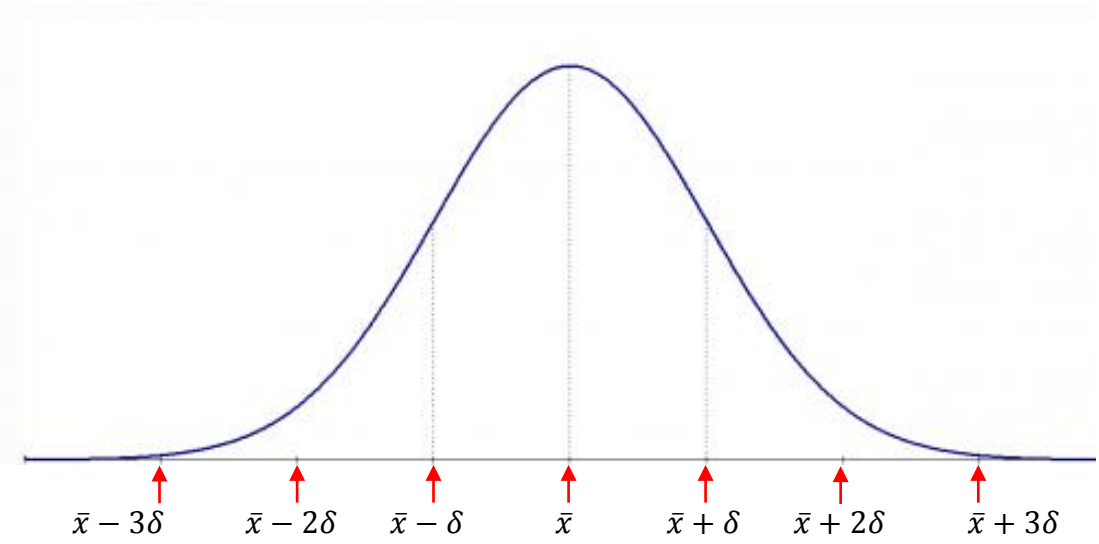
The variance of a set of data is found by,

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

If we assume that a dataset has a normal distribution then,

- 68% of data falls within one standard deviation of the mean
- 95% of the data falls within two standard deviations of the mean
- 99.7% of the data falls within 3 standard deviation of the mean

Graphically,



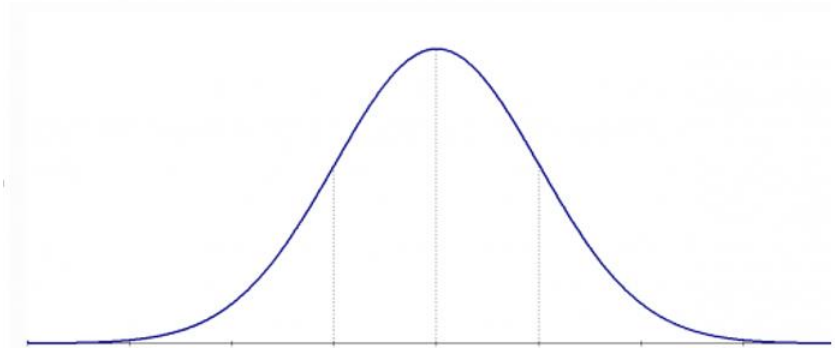
Skewness and Kurtosis

Both skewness and kurtosis measure the shape of a distribution (often a probability distribution). While skewness is a measure of symmetry, kurtosis measures how "peaked" or "flat" a distribution is, relative to the normal distribution.

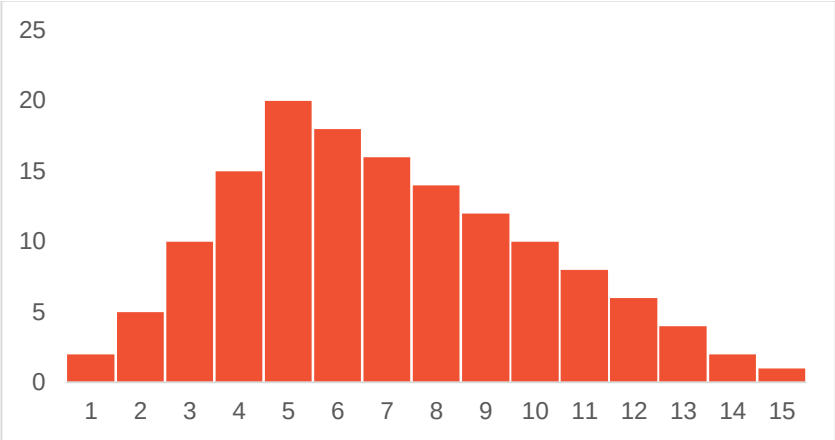
Skewness

Skewness is a measure of symmetry, and the more skewed data is, the less symmetrical it is. A symmetrical distribution of data looks the same from the left as from the right of the centre. That is, its mirror image should be identical to it.

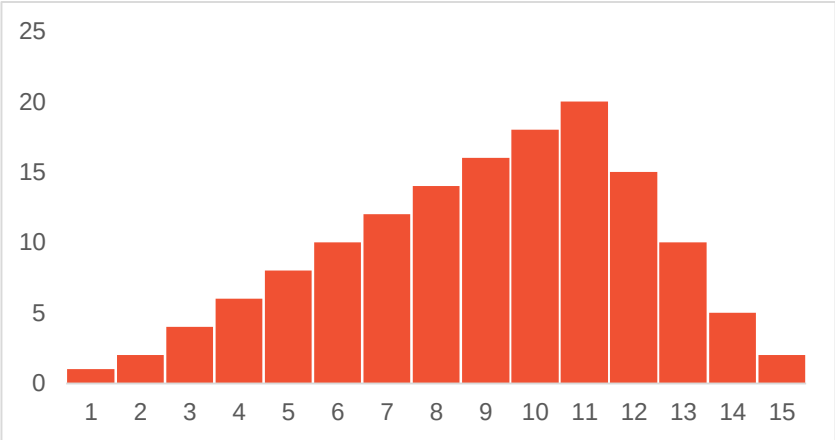
For example, the normal distribution is perfectly symmetrical and is hence **not** skewed,



An example of a positively skewed distribution is,

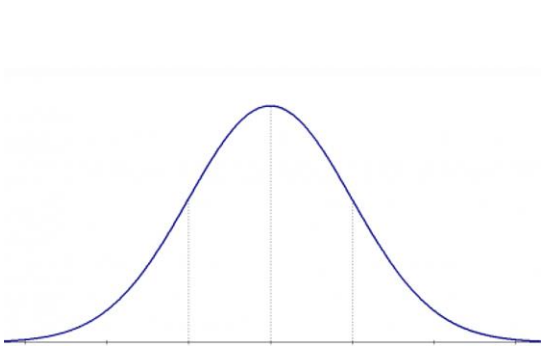


An example of a negatively skewed distribution is,

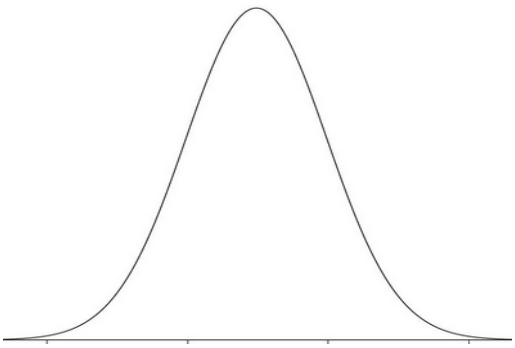


Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. For example, the following is an example of leptokurtic distribution which has a higher kurtosis than the normal distribution,



Normal distribution

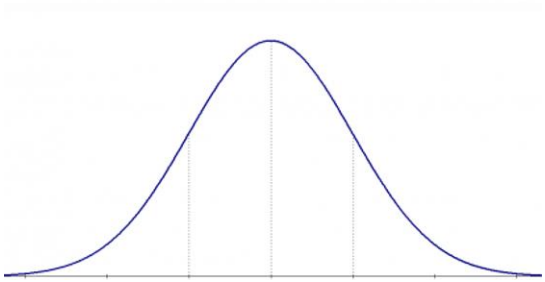


Leptokurtic distribution

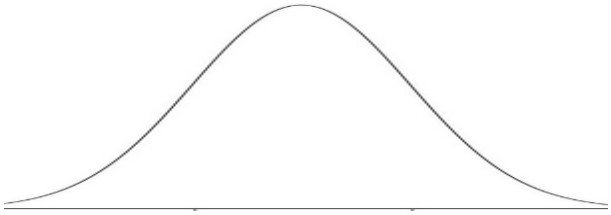
A **leptokurtic distribution** has a higher peak and fatter tails relative to the normal distribution, and hence also has a small variance and standard deviation since values cluster more around the middle.

A **mesokurtic distribution** is one which has the same kurtosis as a normal distribution.

A **platykurtic distribution** is one which is flatter than the normal distribution. It has a higher standard deviation and variance relative to the normal distribution as its values are more spread out.



Normal distribution



Platykurtic distribution