THE UNIVERSITY OF MELBOURNE

FACULTY OF BUSINESS & ECONOMICS

Helpsheet

Giblin Eunson Library

STATISTICS

Use this sheet to help you:

- Understand frequency distributions
- Calculate the five points needed to sketch a box plot of a given set of data
- Calculate the three measures of central tendency of a given data set
- Calculate the variance and standard deviation of a set of data

Author: Carter, D.
Design and layout: Pesina, J.

Representation of data

Frequency distributions

A distribution is an arrangement of observations in increasing or decreasing order of magnitude.

The **frequency** of a particular observation is the number of times the score appears in the data.

A f**requency distribution** is a tabular representation of the observations in ascending order of magnitude with their corresponding frequencies.

Frequency distributions with class intervals

A frequency distribution for a data set with a large number of observations is constructed as follows

1. determine the range of the data set
2. decide the width of the class intervals
3. divide the range by the chosen width of the class interval to determine the number of intervals.

Standardised frequencies

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

where f is the frequency of a particular observation (or class interval) and n is the number of observations

2 Percentage frequency $= \frac{f \times 100}{n}$

where f is the frequency of a particular observation (or class interval) and n is the number of observations.

Frequency density

fr. density = $\frac{\text{frequency of class interval}}{\text{length of class interval}}$

Cumulative frequency distribution

A cumulative frequency distribution is a tabular display of data showing how many observations lie below certain values. It is obtained from the frequency distribution by adding each frequency to the sum of the preceding frequencies.

Cumulative frequency curve

The cumulative frequency curve is the graph of the cumulative frequency distribution, and it is drawn by joining the point plotted at the upper ends of the intervals.

Most cumulative frequency curves have a stretched S-shape and they are used to determine how many observations lie below (or above) a particular score and to read off directly the percentage of observations less (or more) than any specified value.

Percentiles and quantiles

A percentile is a score below which lie a percentage of cases; and a quantile is its decimal equivalent.

e.g. the 25^{th} percentile is denoted by P_{25}. This is known as the 0.25 quantile.
The percentiles and quantiles can be obtained from the cumulative frequency curve and are useful to describe the behaviour of the data set.

Quartiles and the interquartile range

Quartiles divide the set of measurements into four equal parts
P_{25} is known as the first or **lower quartile**, and is denoted by Q_1 (or Q_L)
P_{75} is known as the third or **upper quartile**, and is denoted by Q_3 (or Q_U)
The interquartile range (IQR) is the difference between Q_3 and Q_1.
i.e. $IQR = Q_3 - Q_1$.
P_{50} is is known as the median.

Boxplots

A boxplot is a device used to illustrate the range, median quartiles and interquartile range of data.

The box displays the interquartile range, the line inside the box represents the median and the whiskers are the lines joining the quartiles to the extreme values.

Outliers

An outlier is an observation away from the main body of the data. The boxplot may be used to identify outliers. If there are outliers then the whiskers can be very long and this can misrepresent the data. The rule for determining an outlier is that the whisker must be no longer 1.5 times IQR (i.e. the length of the box). Therefore an outlier is an observation which lies outside the interval.

$$Q_1 - 1.5 \times IQR \leq x \leq Q_3 - 1.5 \times IQR.$$

Outliers are individual observations which do not fit the overall pattern. An outlier tells us something about the observation. An unusually high score in an aptitude test, for example, may inform us that the student is a genius or that there is a design fault in the design of the test.

An outlier is often the product of a typing error or an equipment malfunction. You have to determine the cause of the outlier(s) and then decide how important a part of the data set it might be.

Measures of central tendency

The mean

$$\text{Mean} = \frac{\text{sum of all measurements}}{\text{number of measurements}}$$

symbolically, $\bar{x} = \frac{\sum x}{n}$

The mean of grouped data

The mean of a frequency distribution is

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

where is the mean value of the sample x is the value of each measurement f is the frequency of each measurement $\sum f x$ is the sum of all the measurements n is the number of measurements.

The mean of a class interval frequency distribution is

$$\bar{x} = \frac{\sum f xm}{n}$$

where $\bar{x}$ is the mean value of the sample
m is the midpoint of each interval
f is the frequency of each interval
n is the number of measurements

Advantages of the mean

- Every set of data has one and only one mean
- The mean is useful for comparing sets of data
- The concept of the mean is familiar and clear to most people

Disadvantages of the mean

- The mean is reliable and it reflects all measurements in the data set, but it is affected by extreme measurements that are not truly representative of the data.

The median

If the measurements of a variables are arranged in ascending (or descending) order of magnitude, then the **median** is the middle score.

If the number of measurements is even, then the median is the average of the two middle measurements.

Advantages of the median

- The median is easy to comprehend
- The median can be read directly from the cumulative frequency curve
- Extreme values in the data set will not affect the median.

The mode

The **mode** is the score which occurs most often

Measures of spread

The range

The **range** is the difference between the largest value and the smallest value of a given set of data

Interquartile range

$IQR = Q_3 - Q_1$.

The **interquartile range** describes the spread of the middle half of the observations.

The deviation

The **deviation** from the mean of a particular score, x, in a sample is the difference between that score, x, and the mean, x, of the sample.

The mean deviation

The **mean deviation** for a sample is the average of the positive values of the deviations from the mean.

$$\text{Mean dev} = \frac{\text{sum of the positive deviations}}{\text{the number of scores}}$$

Practical applications of the mean deviation: the mean deviation enables us to make a comparison of scores, say of the same marks obtained by the same student for different subjects.

Variance and standard deviation

In practice the mean deviation and the range are rarely used to find the spread of a sample. Usually we use the **standard deviation** to find the spread of the sample since this measure has wider applications in applied statistics.

Sample standard deviation for ungrouped data.

The variance of a sample of size n is defined by

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

The standard deviation is the positive square root of the variance and is therefore defined as:

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

The standard deviation is the unit of measurement for the spread of a distribution.

For any set of observations, about 95% of the observations lie in the interval

$$\bar{x} - 2s \le x \le \bar{x} + 2s$$

For any set of observations,

- about 66% of the observations lie in the interval

$$\bar{x} - s \le x \le \bar{x} + s$$

- about 99% of the observations lie in the interval

$$\bar{x} - 3s \le x \le \bar{x} + 3s$$

Sample standard deviation for grouped data

Sample variance $s^2 = \frac{\sum (x-\bar{x})^2 f}{n-1}$

where n = Σf
and the standard deviation is $s = \sqrt{\frac{\sum (x-\bar{x})^2 f}{n-1}}$