

# 1. Statistics for Engineers

## Recommended Books

1. Miller & Freund's Probability & Statistics for Engineers

R.A. Johnson (*7<sup>th</sup>* Ed. Pearson, 2005)

2. Statistics for Business & Economics

Anderson, Sweeney, Williams, Freeman, Shoesmith  
(*9<sup>th</sup>* Ed. Thomson, 2007)

+ many Stats texts in Library + Chapters in Stroud etc.

## Websites

<http://www.maths.manchester.ac.uk/service/>

Online course materials (Statistics)

<http://onlinestatbook.com/rvls/>

(Rice University Multimedia Resource)

Additional texts:

mentioned on (Maths service) course website:

**<http://www.maths.manchester.ac.uk/service/>**

*E Kreyszig*, Advanced Engineering Mathematics, Wiley

*RE Walpole and RHE Myers*, Probability and Statistics for Engineers and Scientists - 5th Edition, Prentice-Hall (1993)

*G James et. al.*, Modern Engineering Mathematics, Pearson

*K Weltner et al.*, Probability and Statistics for Maths for Scientists and Engineers, Stanley Thornes

# 1. Describing Data

## 1.1 Introduction

The aim of this section is to present certain standard ways of exploring, summarising and presenting data, such as:

- the mean, median and mode;
- the range and inter-quartile range;
- variance and standard deviation.

We shall use the following example for illustrative purposes throughout this section.

### *Example 1.*

Recorded below are the total number of alignment errors per aeroplane from a sample of 50 planes. This data set

was obtained from the Time Series Data Library:  
<http://www-personal.buseco.monash.edu.au/hyndman/TSDL/>

Data.

7 6 6 7 4 7 8 12 9 9 8 5 5 9 8 15 6 4 13 7 8 15 6 6 10 7  
13 4 5 9 3 4 6 7 14 18 11 11 11 8 10 8 7 16 13 12 9 11  
11 8

It will be convenient to have the data ordered in ascending order. Therefore the ordered data are:

3 4 4 4 4 5 5 5 6 6 6 6 6 6 7 7 7 7 7 7 8 8 8 8 8 8 8  
9 9 9 9 9 10 10 11 11 11 11 11 11 12 12 13 13 13 14 15 15  
16 18

A plot helps us understand the data better. One of the

most useful graphical methods is a histogram (bar chart).

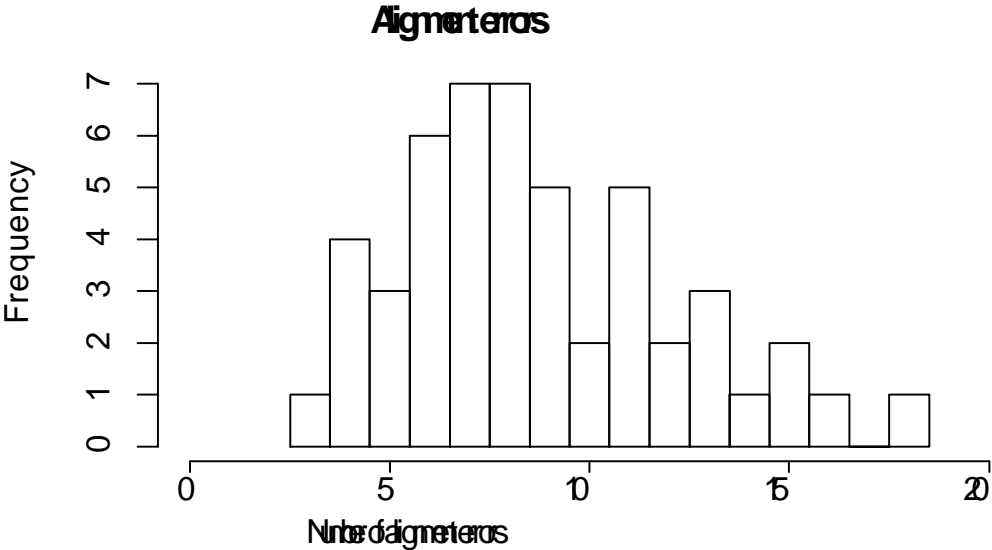


Figure 1.1 Histogram of the number of alignment errors.

### 1.2 Measure of 'central tendency'

The first question asked about data is often: 'What is a typical value?' There are three common measures or "summary statistics":

- Mean - 'average'. (Most commonly used.)

- Median - middle value.
- Mode - most frequent value.

*Mean.*

The mean is defined as follows:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of observations}}.$$

In the above example, the sum of all the values = 436 and the number of observations = 50. Therefore the mean number of alignment errors is  $\frac{436}{50} = 8.72$  per aeroplane.

Convention:

$\bar{x}$  represents the *sample* mean.

$\mu$  represents the (*whole*) *population* mean.

Suppose that the sample comprises  $N$  observations

$\{x_1, x_2, \dots, x_N\}$ . Then

$$\bar{x} = \frac{1}{N} \sum_{j=1}^N x_j.$$

We are usually interested in estimating the *population* mean,  $\mu$ . (see later on in the course). The most usual estimate for the population mean,  $\mu$ , is simply the sample mean,  $\bar{x}$ .

*Median.*

The median is that value which divides the data in half in the sense that 50% of the values are less than or equal to the median and 50% are greater.

Let  $N$  denote the number of observations.

If  $N$  is odd, that is,  $N = 2m + 1$ , where  $m$  is a whole number, then the median value is  $(m + 1)^{st}$  observation in the sorted list.

If  $N$  is even, that is,  $N = 2m$ , where  $m$  is a whole number, then the median value is the *average* of the  $m^{th}$  and  $(m + 1)^{st}$  observations in the sorted list.

Example: 50 observations. The median is the average of the 25<sup>th</sup> and 26<sup>th</sup> observations in the sorted list.

25<sup>th</sup> observation: 8

26<sup>th</sup> observation: 8

Therefore median =  $\frac{8+8}{2} = 8$ .

*Mode.*

The value that occurs the most frequently. If the data has more than one mode we say that the data is multimodal.

Example: This is easily obtained in this example from the histogram. Clearly, the modes are: 7 and 8.

### **1.3 Range and variation**

The above measures of 'central tendency' are informative about the data but don't tell the whole story. To illustrate this consider the following two data sets consisting of 10 observations:

Data set 1: 4, 5, 6, 5, 5, 4, 5, 5, 6, 5

Data set 2: 8, 3, 6, 9, 4, 1, 5, 2, 5, 7

Both data sets have mean=5, median=5 and mode=5, but the two data sets are really rather different.

Therefore the second question which is asked of the data is: 'How much variation (spread) is there in the data?' There are a number of common approaches taken to answer this question:

*Range.*

The difference between the minimum and maximum of the data.

Example. Range=  $18 - 3 = 15$ .

*Inter-quartile range (IQR).*

The lower quartile is the observation which 25% of the values are less than or equal to. The upper quartile is the observation which 25% of the values are greater than or equal to.

Example: 50 observations

Lower quartile: 6 (13<sup>th</sup> observation)

Upper quartile: 11 (38<sup>th</sup> observation)

*Variance.*

The sum of the squares (SS) of all the deviations of the values from their mean divided by (Number of values -1)

$$\begin{aligned}s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N - 1} \\ &= \frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2.\end{aligned}$$

This expression can be rewritten as

$$\begin{aligned}s^2 &= \frac{1}{N - 1} \left( \sum_{i=1}^N x_i^2 - N\bar{x}^2 \right) \\ &= \frac{1}{N - 1} \left( \sum_{i=1}^N x_i^2 - \frac{(\sum_{i=1}^N x_i)^2}{N} \right)\end{aligned}$$

The latter expression involves only the sum of  $x$  – values and the SS of  $x$  – values and is a convenient formula for calculating the variance.

The expression above is appropriate for analysing a sample (from some larger population). If we were considering the whole of some finite population we may use a divisor of  $N$  and the Greek  $\sigma$  for standard deviation (the square root of the variance):

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$$

Summary: to obtain the variance divide by  $N - 1$  if estimating the variance based on a sample and divide by  $N$ , if calculating the variance based on the whole population. Note that for large  $N$  there is very little difference between the two calculations.

Example.

$$\sum_{i=1}^N x_i^2 = 4386$$

Therefore variance,  $s^2$ , is:

$$s^2 = \frac{1}{49}(4368 - 50 \times (8.72)^2) = 11.92.$$

*Standard deviation.*

The square root of the variance, *i.e.*  $s$ .

Example.  $s = 3.452$ .

The main reason to take square roots is to recover a measure of variability that has the same units as the data.

## **1.4 Concluding comments**

The mean, median and mode are all natural measures of *central tendency* and are useful in describing data that cluster around some particular value. The range, interquartile range, variance and standard deviation are measures of the way data are scattered around the central

tendency. The most commonly used measures are: mean, variance and standard deviation.

### Extra Optometry Data

Optometrists IOP/ <i>mmHg</i>		Engineers IOP/ <i>mmHg</i>	
14	15	16	15
15	14	17	19
15	15	15	17
15	14	13	13
15	16	13	23
15	15	10	13
16	15	20	18
15	15	16	12
15	16	10	17
15	15	8	15

Analyse the data.

For both data sets (the number of observations)  $n=20$ .

## *Optometrists*

Mean:  $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{1}{20}(300) = 15$

Median: 15 ( $10^{th} obs. = 15, 11^{th} obs. = 15$ )

Mode: 15 (occurs 14 times)

Range:  $16-14 = 2$

Variance:  $s^2 = \frac{1}{19} \left( \sum_{i=1}^{20} x_i^2 - 20\bar{x}^2 \right) = 0.3158$

Standard deviation:  $s = 0.5619$

## *Engineers*

Mean:  $\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{1}{20}(300) = 15$

Median: 15 ( $10^{th} obs. = 15, 11^{th} obs. = 15$ )

Mode: 13 (occurs 4 times; 15 occurs 3 times)

Range:  $23-8 = 15$

Variance:  $s^2 = \frac{1}{19} \left( \sum_{i=1}^{20} x_i^2 - 20\bar{x}^2 \right) = 13.263$

Standard deviation:  $s = 3.642$

<http://onlinestatbook.com/rvls/>