CHAPTER 1

Descriptive Statistics

1.1 Introduction

In this chapter we study how to display data so that we can ask important questions about the population from which the data have been drawn. For example a histogram or stem-and-leaf plot might tell us whether the assumption of a Normal population is likely to be reasonable for a single sample. On the other hand a table of the means and standard deviations of several samples may prompt us to ask whether there is evidence that the corresponding *population* means are really different or whether the observed differences in the *sample* means can be ascribed to random sample variation. Formal answers to these questions can be given by the use of appropriate significance tests but there is no substitute for an initial *descriptive* examination of the data before *any* statistical analysis is undertaken.

To do the basic statistical computing Minitab Handbook, Chapters 1 to 5 with the following omissions is useful. .

- Trimmed means and standard error of means (DESCRIBE command)
- Multiple plots (MPLOT) and Time series plots (TSPLOT)
- Advanced features of the TABLE command
- STACKing and UNSTACKing data

Some basic ideas about MINITAB also provided in the Chapter 9.

First we give here some basic definitions.. You are not expected to learn these by rote but you will be expected to explain underlying concepts of practical statistics in short paragraphs.

DEFINITION 1 A *population* is a collection of units or individuals under investigation. It may be finite or infinite, real or hypothetical. Remark An infinite population is often *modeled* by a *distribution:* see Probability lectures for details.

DEFINITION 2 A variate or *variable* is any measurable quantity or attribute whose value may vary from one unit to another.

DEFINITION 3 A *sample* is a subset of the population. It is selected to give the sample *data,* which are the sampled units together with the values of the variables measured on them.

DEFINITION 4 A *random sample* of size *n* is a sample chosen so that every subset (of the

population) of size *n* is equally likely to be selected. A necessary but not sufficient condition is that each population unit is equally likely to be included in the sample.

DEFINITION 5 An *observation* is the value taken by the variate or variable for a sampled unit. The observations for a single variable X in a sample size *n* are written as

$$x_1, x_2, ...x_n$$

where x_i is the observation for the unit labeled *i* in the sample *(i* =1,2,. . . , *n).*

DEFINITION 6 If there are, say, three variables to be measured, then the data *matrix* is a *n x* 3 array whose i^{th} row is the observations x_i , y_i , z_i on the three variables X, *Y, Z (i* = 1,2,.. . , *n).* Each of the three columns comprises all the sample observations on a variable.

In MINITAB these columns are denoted *C1, C2, C3* and the data matrix together with names of the variables and sums and sums of squares for columns (and rows) form the *worksheet.* This is usually saved so it can be retrieved later without having to read the raw data into the computer all over again.

DEFINITION 7 A *quantitative variable* takes purely numerical values. If any value can be taken in a given range it is called a *continuous* variable. If the values change by steps (often equal) it is called a *discrete* variable.

DEFINITION 8 A *qualitative variable* (or *attribute)* does not normally take numerical values. If there are only two possible values (e.g. True or False, Yes or No, On or Off) it is called a *factor* at *two levels* or *zero/one variable* (because by convention we denote the two values 0 and 1). Similarly if there are *k* possible values we can speak of a *factor at k levels.* If it is meaningful to regard the levels on a scale (e.g. Poor/Average/Good in an opinion of performance) then the variable is an *ordinal variable.* If, however, there is no implied ordering (e.g. Place of Residence) then the variable is a *categorical variable*

1.2 Presentation of Data

The most suitable way to present data depends on the size of the sample and the number and nature or type of variables. Here is a summary of methods of presentation, by hand or using MINITAB, with which you are expected to be familiar.

1. Single Variable

   - Continuous:

     - STEM-AND-LEAF plots
     - HISTOGRAMs
     - DOTPLOTs

- Discrete:

  - frequency table (TALLY) or bar diagram
  - DOTPLOT

- Categorical or Ordinal: frequency table only
- Note that we often *transform* data to try and achieve symmetry before formal significance testing.
  If the distribution is positively skew, try taking
  — square roots (SQRT)
  — natural logarithms (LOGE) or
  — negative reciprocals (—1/C)
  or, if negatively skew, raising to a power > 1. A further objective of transformations is to achieve a distributional shape close to that of a Normal distribution ('transform to Normality') and this can be checked using a Normal probability plot. MINITAB can achieve this simply using the commands NSCORES and PLOT.

2. Two Variables

- Both Continuous: scatter diagram or plot (PLOT)
- Both Categorical: two-way Contingency Table (TABLE)

- One Continuous, one Categorical: STEM-AND-LEAF plot, HISTOGRAM, or DOTPLOT, one for each level of the categorical variable (MINITAB subcommand BY)

3. Three Variables

(a) Plots with Symbols: one continuous variable plotted against another, with symbols denoting the levels of a third, categorical variable (LPLOT)
(b) Tables with Statistics: two categorical variables forming a two-way table with cells containing summary statistics (e.g. mean) of a third, quantitative variable, as well as cell frequencies (TABLE with subcommand STATS, MEAN etc.)

Examples of all these visual presentations are given in the MINITAB handbook Chapters 2 to 4. You should acquire some practice with the different data sets available with MINITAB before attempting question 3 of practical 1.

Example 1

(Data and description are taken from *Statistics for Technology*, by C. Chat-field.) Twenty refrigerator motors were run to destruction under advanced stress conditions and the times to failure (in hours) were recorded as follows.

104.3	158.7	193.7	201.3	206.2
227.8	249.1	307.8	311.5	329.6
358.5	364.3	370.4	380.5	394.6

426.2 434.1 552.6 594.0 691.5

We cannot predict exactly how long an individual motor will last, but if possible, we would like to predict the pattern of behavior of a batch of motors. For example we might want to know the overall proportion of motors which last longer than one week (168 hours).

A DOTPLOT of the data by hand should include a scaled horizontal axis labeled with the name of the (continuous) variable 'Failure Time' and its units of measurement(hours). The plot reveals that the centre of the data is around 350 hours and there are two 'clumps' with a long upper tail to the distribution representing a small but not insignificant proportion of extremely reliable motors.

A STEM-AND-LEAF plot by hand should also be labeled appropriately and the 'stem' chosen to give a good impression of the distribution of data. It is important that the 'leaves' are equally spaced but with this proviso it is not necessary that they consist of only a single digit (as in MINITAB). This allows us to use the main advantage that these plots possess over histograms; namely that the full raw data is preserved in the plot. Another important feature of a *hand* plot is that the stem frequencies can be given **(** on the left hand side). These are more useful than the cumulative type offered by MINITAB. Note that the median class is marked with parentheses.

Stem-and-Leaf plot of Refrigerator Motor Failure Times (hours)

3	1	04.3	58.7	93.7
4	2	01.3	06.2	22.8 49.1
(8)	3	07.8	11.5	29.6 58.5 64.3 70.4 80.5 94.6
2	4	26.2	34.1	
2	5	52.6	94.0	
1	6	91.5		

Any plot should always carry a comment. Here we see a positively skew distribution as in the dotplot. We can check that no data has been missed by totalling the stem frequencies to equal the sample size.

Example 2 (again from Chatfield)
The numbers of cosmic particles striking an apparatus in forty consecutive periods of one minute were recorded as follows.

3	0	0	1	0	2	1	0	1	1
0	3	4	1	2	0	2	0	3	1
1	0	1	2	0	2	1	0	1	2
3	1	0	0	2	1	0	3	1	2

As in Example 1 we cannot exactly predict what the next observation will be because of the experimental uncertainty.
A dotplot can be used also for this discrete variable but just as informative (though less pictorial) is the *frequency distribution*

Frequency distribution of number of cosmic particles

No. of particles	0	1	2	3	4	5+	Total
Frequency	13	13	8	5	1	0	40

Note that the uppermost class should always have a ~ frequency. This is important when we look at goodness-of-fit tests and when using MINITAB output produced by TALLY. Other equivalent ways of displaying these data are stem-and-leaf plots and frequency bar diagrams.

1.3 Measures of Location and Spread

These are used as single variable statistics primarily to compare different samples or sub-samples. We calculate them before asking and answering more formal questions about the population using Significance Tests (Handout 3).

DEFINITION 9 The *sample* mean $\bar{x}$ of *n* observations on a variable x is the *arithmetic mean or average*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The *geometric* mean is the n^{th} root of the product of the observations

$$\sqrt[n]{\prod_{i=1}^{n} x_i}$$

The *harmonic mean* is the reciprocal of the arithmetic mean of the reciprocal observations

$$\frac{n}{\sum_{i=1}^{n} 1/x_i}$$

Remark The last two means are defined only for positive observations (x_i, $i = 1,2,\ldots, n$) and in this case we have the inequality

$$A.M. \geq G.M. \geq H.M.$$

Example 1 revisited $\sum_{i=1}^{n} x_i = 6856.7 \Rightarrow 342.835$ without rounding, but for the final answer only we should round to the same accuracy as the raw data (continuous variables only). Thus $\bar{x} = 342.8$.

The sample mean can be unduly influenced by 'outliers' so should always be accompanied by a visual presentation of the data.

DEFINITION 10 The *sample median m* is the 'middle' observation if the observations are arranged in rank order (SORT), that is the $\left(\frac{n+1}{2}\right)^{th}$ if *n* is odd , and the average of the two middle observations (the $(n/2)^{th}$ and the $(n/2 + 1)^{th}$ if *n* is even.

Example I revisited $n = 20$ is even so $m = \frac{329.6 + 358.5}{2} = 344.05$ and the final answer should be quoted as 344.0 as we round to an even digit by convention with a single final 5.
Remark In general $\bar{x} > m$ for positively skew distributions and $m > \bar{x}$ for negatively skew with $m = \bar{x}$ for symmetry.

DEFINITION 11 The *lower quartile* q_1 is the $\left(\frac{n+1}{4}\right)^{th}$ observation if $\frac{n+1}{4}$ is an integer, otherwise it is determined by interpolation. The *upper quartile q3* is the $\left(\frac{3(n+1)}{4}\right)^{th}$ observation if $\frac{3(n+1)}{4}$ is an integer, otherwise it is determined by interpolation (see the example below).

Example 1 revisited . $\frac{n+1}{4} = 5.25$ so $q_{.1}$ lies 1/4 of the way from the 5^{th} to the 6^{th} observation; that is

$q_1 = 206.2 + 1/4(227.8 - 206.2) = 211.6.$

Similarly
$q_3 = 418.3.$

We could interpret these statistics by expecting that 75% of refrigerator motors would have a lifetime exceeding 211.6 hours, but only 25% exceeding 418.3 hours.

DEFINITION 12 (Measures of Spread or Variability)
The *sample* range *r* is a very crude measure of variability and is simply the largest minus the smallest observation (MAX — MIN). It is clearly prone to 'outliers'.
The *semi-interquartile range* is $\frac{1}{2}(q_3 - q_1)$
The most common measure is the *sample standard deviation s* (STDEV), where

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

is the *sample variance.*

Remark Some authors define the sample variance with a divisor of *n* rather than n – 1. Make sure you know which one your calculator is using. MINITAB uses the divisor $n - 1$ in the function STDEV and does not provide a function for the variance.

By hand we calculate s^2 almost always using the fundamental identity for the corrected sum of squares

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

Note that the semi inter-quartile range, unlike the sample variance, is not so prone to 'outliers'.
Example 1 revisited r = 691.5 - 104.3 = 587.2, $\frac{1}{2}(q_3 - q_1) = 103.4$, $\sum_{i=1}^{n} x_i.^2 = 2776363.47$,

$\sum_{i=1}^{n} x_i = 6856.7$, n = 20

$s^2 = 22402.45924$. Maintaining the same accuracy in intermediate calculations, we obtain $s = 149.674511$ so that the sample standard deviation is 149.7 hours.
The MINITAB command DESCRIBE gives the value of all these summary statistics and can be used for several data sets simultaneously to aid numerical descriptive comparisons.

EXERCISE

1. In a study to investigate the incidence of heart disease among British males aged over 65, three men were selected at random from state pension records: Mr. Brown, who was 1.8 metres tall and weighed 82 kilograms, had not suffered a heart attack recently and classified his diet as 'low-fat'; Mr. Smith (1.85m., 89kg.) also had not suffered but classified his diet as 'normal'; Mr. Jones was 1.75m. tall but weighed 95kg. and had unfortunately had a heart attack recently. Mr. Jones admitted to a 'fatty' diet.

   Identify

   (a) the population

   (b) the variables and whether they are quantitative or qualitative, continuous or discrete, ordinal or categorical

   (c) whether the sample of size 3 can be regarded as random

   (d) the observations on the variable 'Height'

   (e) the data matrix or MINITAB worksheet

   (f) y_3, z_1, w_2 where X, *Y, Z, W* are the letters we use for the variables in order of mention.

2. Show that if we change the 'location' and 'scale' of a sample

$$x_1, x_2, \ldots, x_n$$

to a 'new' sample

$$z_1, z_2, \ldots z_n$$

using the transformation

$$z = kx + c,$$

where *k* and *c* are constants, then the sample mean of the 'new' sample is given by

$$\bar{z} = k\,\bar{x} + c$$

and the sample variance is

$$s^2{}_z = k^2 s^2{}_x$$

Hence by suitable choice of k and c find the sample mean and variance of

37.01, 37.03, 37.02, 37.04, 37.10, 37.03, 37.05.

3. The following data are measurements of the thrust of a rocket engine. Give a stem-and-leaf plot by hand recording the frequencies for each stem. Compare this plot to one produced by MINITAB, identifying the frequencies given there, and comment on the shape of distribution shown by your two plots. (Data taken from *Statistics For Technology* p. 15, by C. Chatfield)

999.1	1003.2	1002.1	999.2	989.7	1006.7
1012.3	996.4	1000.2	995.3	1008.7	993.4
998.1	997.9	1003.1	1002.6	1001.8	996.5
992.8	1006.5	1004.5	1000.3	1014.5	998.6
989.4	1002.9	999.3	994.7	1007.6	1000.9

Note: this data set is also available from the supplied disk, filename ROCKET.DAT. Compute by hand the mean, variance, standard deviation, median, lower and upper quartiles of the sample data and compare these with output from DESCRIBE using MINITAB.

4. Perform a *descriptive* analysis of the PULSE data using the saved MINI-TAB worksheet PULSE.MTW, answering the following questions (refer to the MINITAB handbook p. 319 for the meanings of codes in the data set and for further details):

(a) What types of variable are measured in this survey?
(b) What are the proportions of the three different levels of physical activity?
(c) What is the distribution of the *increase* in pulse rate following the experiment? (You will need to create another variable using the LET command. NAME the new variable appropriately.)
(d) How does this distribution vary according to
  (i) whether the student 'ran in place' or not?
  (ii) whether the student smokes regularly?
  (iii) the sex of the student?
(e) What is the relationship between
  (i) smoking habit and physical activity?
  (ii) height and weight?
  (iii) weight and increase in pulse rate?
(f) How does the mean increase in pulse rate vary according to smoking habit *and* physical activity?
(g) How does the relationship between height and weight vary according to sex?
(h) Ask yourself another question about these data of interest to you and present a descriptive answer.