Chapter 1 Descriptive Statistics

Outcomes: You must be able to

* draw rod diagrams and histograms for sets of raw data
* calculate the arithmetic mean, median, mode(s), Pearson's coefficient of skewness for a set of raw data
* calculate various quantiles.

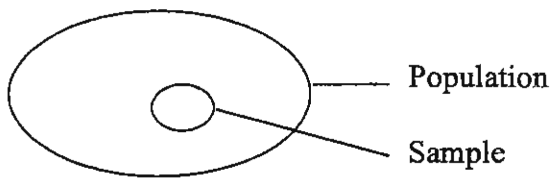
1.1 Basic Concepts

1.1.1 Definition

The basic concern of statistics is to find objective scientific procedures by which to both describe and make inferences from data in which *variability* and *uncertainty* are important factors.

1.1.2 Populations and Samples

Invariably in statistical experimental situations one is confronted with one or more relatively small subsets of data, called *samples*, drawn from generally large, inaccessible sets of data, called *populations* and one wishes to gain information about the populations by examination of the samples.



NB: Venn diagrams, as used above, are very convenient to represent populations and samples. However, it should be noted that the elements comprising a sample should be thought of as being *randomly distributed* throughout the population and *not* (as might be inferred from the diagram) a sub-group of elements in close proximity to each other!

1.1.3 Specific Objectives

The essential problem in statistics is to find quantitative procedures for *describing* and *interpreting* sets of data. There are two aspects to this problem:

1. The *description* of a set of data (*sample / population*) in terms of a small set of descriptive quantities (called *statistics / parameters* respectively). This aspect is called *descriptive statistics.*

2. Drawing *inferences* about population parameters by examining sample statistics. This aspect is called *inferential statistics.*

In this Chapter we focus on Descriptive Statistics. Inferential Statistics will be dealt with from Chapter 7 onwards.

1.1.4 The Statistical Experiment

Five distinct phases can be identified in any statistical experiment:

- ***Collection of data***
- ***Organisation of data***
- ***Mathematical description of data*** — ***Descriptive Statistics***
- ***Analysis of data***
- ***Interpretation of data*** — ***Inferential Statistics***

1.1.5 Notation

A subtle, but important, distinction is made between the upper case variable, X, and the lower case variable x. We are familiar with the fact that, in mathematics, the (algebraic) variable, x, is simply a symbol that can assume any *particular* value of a specified set of numbers. And depending on how the set of numbers is specified, x can be either a *continuous variable* (as, for example, the set of real numbers, R) or a *discrete variable* (as, for example, the set of all integers, Z). Suppose x represents the set of all real numbers. The statement $x = x_1$ is interpreted as referring to a *unique* element of R that is distinct from, say, $x = x_2$. The numbers x_1 and x_2 are also seen as two distinct points on the real number line, R. In contrast, however, the variable X is used to represent all possible *data* elements that can (theoretically) be selected from a numeric population. Here, $X_1, X_2, X_3 \ldots, X_n, \ldots$, are seen as a sample of n unique data elements drawn from a numeric population. However, if $X = X_1 = x_1$ and $X = X_2 = x_1$ then X_1 and X_2 are still seen as two *unique* data elements in the sample even though $X_1 = X_2 = x_1$. The number, x_1, although unique on the real number line R, could be associated with numerous elements in the sample.

Thus the X_i are the actual elements of the data set, while the x_i are the numbers that can be found in the data set.

Example: Consider this data set: 1, 2, 2, 4, 5, 5, 5, 6, 7

The X_i are $X_1 = 1, X_2 = 2, X_3 = 2, X_4 = 4, X_5 = 5, X_6 = 5, X_7 = 5, X_8 = 6, X_9 = 7$.
The x_i are $x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 5, x_5 = 6, x_6 = 7$.

1.2 Data

1.2.1 Categorisation of Data

Data can be classified according to the *kind* of measuring scale to differentiate between particular data elements. The following table summarises the types of data that are generally encountered.

<table>
<tr><th>Data</th><th>Measuring Scale</th><th>Type of Statistics</th><th colspan="2">Type of Data</th></tr>
<tr><td rowspan="4">Data Elements
X</td><td>Ratio Scale
[eg mass, length, count]</td><td rowspan="2">Parametric Statistics</td><td rowspan="3">Numeric data</td><td>Continuous
[eg mass, length, time, temperature]</td></tr>
<tr><td>Interval Scale
[eg temperature, time]</td><td>Discrete
[eg count, order, position]</td></tr>
<tr><td>Ordinal Scale
[eg position, order]</td><td rowspan="2">Non-parametric Statistics</td><td></td></tr>
<tr><td>Nominal Scale
[eg gender, blood-group]</td><td>Categorised data</td><td>Discrete
[eg blood-group, gender]</td></tr>
</table>

Note: Numeric data can always be categorized. For example, consider a set of marks expressed as percentages:

If mark $\geq$ 50%, category is 'pass'; If mark < 50%, category is 'fail'

Conversely, categorized data can sometimes be quantified. For example, consider the subjective assessment of an essay according to the following rule:

Excellent $\Rightarrow$ 4; Good $\Rightarrow$ 3; Satisfactory $\Rightarrow$ 2; Poor $\Rightarrow$ 1; Very poor $\Rightarrow$ 0.

1.2.2 Collection of Data

Often, what passes as a cogent experimental result, falls apart simply as a result of the *manner* in which the data for the experiment had originally been obtained. Paradoxically, this very crucial aspect of a statistical experiment can only be fully appreciated once the essentials of inferential statistics have been dealt with. It should be noted, however, that central to the *data collection* aspect of a statistical experiment is the extent to which *randomness* was observed during this phase of the experiment. We shall deal more extensively with the concept of randomness in Chapter 4.

Data that is collected from an experiment is initially usually numerically unordered. It is called *raw data* at this stage. When the data has been arranged in ascending or descending order of magnitude we say it is an *array*. The difference between the largest and smallest numbers is called the *range* of the data.

1.3 Organisation of Data: Frequency Distributions.

When summarising large amounts of data it is often useful to group the data into classes (usually of equal sizes) and to determine the number of data elements in each class, called the *class frequency*. The table used to represent the classes and their frequencies is called a *frequency distribution* or a *frequency table*. Grouping data destroys much of the original detail of the data but it can make the overall picture of a large amount of data easier to see.

Many types of diagrams are used to gain a pictorial representation for a set of data. In this course we shall restrict ourselves to *rod-diagrams* (also called *bar-diagrams*) and *histograms* only.

1.3.1 Rod-Diagrams

Consider a set of n numeric data elements, X_1, $X_2, X_3, \ldots X_n$ of a *discrete* variable X. We can represent its distribution by a *rod-diagram*. Rod diagrams are usually used for small amounts of data. The raw data is arranged in numerically ascending order and the frequency of each number is noted. The results are plotted on a frequency distribution graph.

Example 1.1

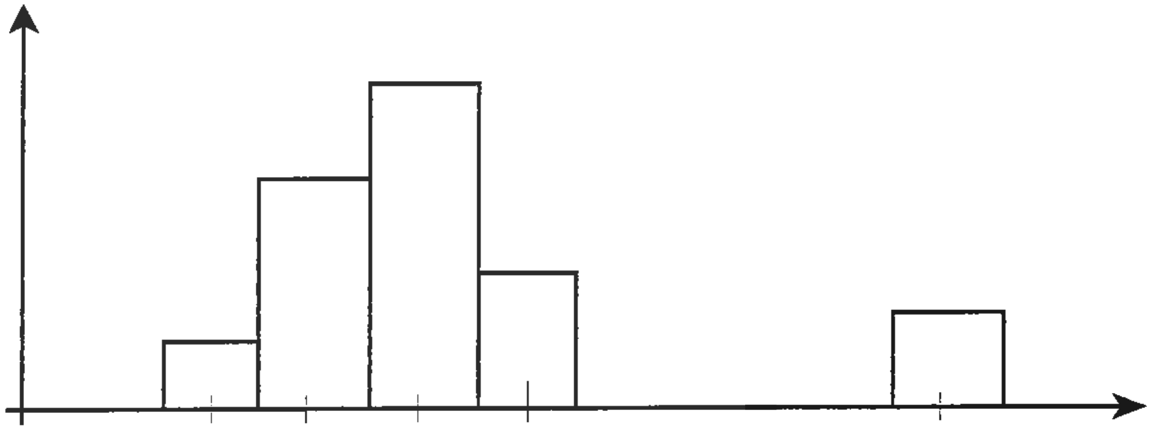
Represent the following data by means of a rod-diagram:

2, 3, 5, 4, 4, 1, 6, 3, 3, 3, 4, 5, 5, 7, 2, 3, 4, 2, 3, 6

1.3.2 Histograms

Definition of Terms

In the histogram below the data variable X can either be *continuous* or *discrete.*



- ***Class midpoints*** $= x_1, x_2, x_3, \ldots x_k$
- ***Number of Classes*** $= k$
- ***Frequency*** $= f_1, f_2, f_3, \ldots f_k$
- ***Class boundaries*** $= c_0, c_1, c_2, c_3, \ldots c_{k-1}, c_k$
- ***Class length*** = L

Note: If n is the number of data elements then $n = f_1 + f_2 + f_3 + f_4 + \ldots\ldots\ldots\ldots + f_k = \sum_{i=1}^{k} f_i$

NB: Gaps should not be inserted between the "bins" of the histogram either when the data variable X is discrete, or when it is continuous and rounded. The reasons for this will become clear when *rounding correction* and *continuity correction* are dealt with in Chapter 4.

Continuous Variables

Consider a set of n numeric data elements, X_1, $X_2, X_3, \ldots X_n$ of a *continuous* variable X.

Example 1.2

Represent the data below by means of a *histogram.* Assume that the data elements in this theoretical situation have been determined *exactly* to a very large (infinite) number of decimal places; i.e. ignore the practical question of *rounding* of the data elements collected.

Class Limits of X	f
$0 < X \leq 10$	3
$10 < X \leq 20$	5
$20 < X \leq 30$	14
$30 < X \leq 40$	28
$40 < X \leq 50$	2

In the above table it would make no difference whatsoever whether the inequality '$\leq$' or '<' were used since X is a *continuous* data variable that can *theoretically* be expressed to an infinite number of decimal places. In *practical* situations, however, data elements selected from a *continuous variable* will have been *rounded off* to a given level of accuracy. In such a situation one would have to be careful not to be contradictory when using inequalities (as would seem to be have been the case above).

Example 1.3

In the following table, X represents the *mass* of a subject in a group of university male students in which individual masses have been measured in kilograms rounded to the first decimal place:

Class Limits of X	f	Class Boundaries
$40 \leq X \leq 49{,}9$	5	
$50 \leq X \leq 59{,}9$	20	
$60 \leq X \leq 69{,}9$	17	
$70 \leq X \leq 79{,}9$	11	
$80 \leq X \leq 89{,}9$	3	

The boundaries for the bins are found by halving the gap between the last number in a class and the first number in the next class, eg $50 - 49{,}9 = 0{,}1$,

$\therefore$ boundary is at $49{,}9 + 0{,}05 = 49{,}95$.

Discrete Variables

Histograms can also be used to represent a distribution of a *discrete* variable.

Example 1.4

In the following table, X represents the number of students attending lectures at a particular time in each of the small lecture theatres at the University:

Class Limits of X	f	Class Boundaries
0 – 9	6	
10 – 19	11	
20 – 29	27	
30 – 39	43	
40 – 49	56	
50 – 59	35	
60 – 69	20	

1.4 Profiles of Frequency Distributions.

In practical situations, the profile of frequency distributions (discrete and continuous) tend to take on certain characteristics shapes:

Positively skewed profiles | Bell-shaped profiles | Negatively skewed profiles

All of the above characteristic shapes are referred to as *unimodal* frequency distributions: i.e. there exists only *one* clear peak in the distribution. However, frequency distributions do occasionally occurs in practice where there exists more than one local peak. These distributions are referred to as *multimodal* distributions: or more specifically, as *bimodal, trimodal, ... distributions*. When such distributions do occur, it often masks the existence of two or more distinct sub-distributions. For example, consider the following bimodal distribution:

1.5 Mathematical Description of Data

The following mathematical descriptions are frequently used. Those in *italics* are particularly pertinent to this course.

- **Measures of central tendency:** *median; mode; arithmetic mean*; geometric mean; harmonic mean; quadratic mean.
- **Measures of spread:** *range;* mean deviation; quartile deviation; 10-90 percentile range; *standard deviation; variance.*
- **Measures of deviation of a distribution from a symmetrical profile:** skewness; *Pearson's coefficient of skewness;* quartile coefficient of skewness; 10-90 percentile coefficient of skewness.
- **Other descriptive measures:** *quantiles* (quartiles, deciles, percentiles); kurtosis.

1.5.1 Mean, Standard Deviation and Moment Coefficient of Skewness

1. **Raw Data :** Let the variable, X, represent a numeric population of N data elements (discrete or continuous). Let $X_1, X_2, X_3, \ldots, X_n$ denote the elements in *sample* of n elements [or let $X_1, X_2, X_3, \ldots, X_N$ denote the elements of the whole *population*]. The *arithmetic mean, standard deviation* and *moment coefficient of skewness* of the set of elements is defined by the following:

	Sample	Population (Finite)
Arithmetic Mean:	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$
Standard deviation:	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2}$	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$
Pearson's Coefficient of Skewness:	$\alpha_3 = \frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{k}(X_i - \bar{x})^3$	$A_3 = \frac{1}{N\sigma^3}\sum_{i=1}^{k}(X_i - \mu)^3$

Note: For the sake of brevity, we shall from now on refer to the *arithmetic mean*, $\bar{x}$ or μ and the *Pearson's coefficient of skewness*, α_3, or A_3 , simply as the *mean* and *skewness* respectively.

The mean and standard deviation can be calculated using a calculator. For a finite population, the standard deviation σ must be calculated using n data elements. On a Sharp calculator the σ_x button is used and on the Casio check that you use the correct version. For a sample, the standard deviation s must be calculated using n-1 elements. On a Sharp calculator the s_x button is used and on the Casio check that you use the correct version.

Example 1.5

Evaluate the *mean, standard deviation* and *skewness* of the following sample of data elements:

7, 2, –8 and 7 expressed in some unit of measurement.

2. **Arrays:** Let the variable, X, represent a numeric population of N data elements (discrete or continuous). Let x_1; x_2; .... x_k, occurring with respective frequencies f_1; f_2;.........f_k, represent the values of the data elements in a *sample* of n elements [or the values of the data elements of the whole *population* of N elements]. The *arithmetic mean, standard deviation and moment coefficient of skewness* of the set of elements are defined by:

	Sample	**Population (Finite)**
Mean:	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$	$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i f_i$
Standard deviation:	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}$	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 f_i}$
Skewness:	$\alpha_3 = \frac{n\sum_{i=1}^{k}(x_i - \bar{x})^3 f_i}{(n-1)(n-2)s^3}$	$\alpha_3 = \frac{1}{N\sigma^3}\sum_{i=1}^{k}(x_i - \mu)^3 f_i$
Note:	$n = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$	$N = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$

Example 1.6 Evaluate the arithmetic mean, standard deviation and Pearson's coefficient of skewness of the following sample of data elements:

0, 1, 1, 0, 0, 1, 2, 0, 3, 3, 2, 0, 3, 0, 4, 5, 0, 4, 2, 1

3. Grouped data: Let the variable, X, represent a numeric population of N data elements (discrete or continuous). Consider a *sample* of n elements [or a *population* of N elements] which have been grouped into k classes with class midpoints x_1; x_2; ...x_k with respective class frequencies f_1; f_2; ....f_k. [Such data is referred to as *grouped data.*] If the original data is not available, then the *arithmetic mean, standard deviation* and *moment coefficient of skewness* of the set of elements is fairly closely estimated by:

	Sample	Population (Finite)
Mean:	$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} x_i f_i$	$\mu \approx \frac{1}{N}\sum_{i=1}^{k} x_i f_i$
Standard deviation:	$s \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}$	$\sigma \approx \sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 f_i}$
Skewness:	$a_3 \approx \frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{k}(x_i - \bar{x})^3 f_i$	$\alpha_3 \approx \frac{1}{N\sigma^3}\sum_{i=1}^{k}(x_i - \mu)^3 f_i$
Note:	$n = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$	$N = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$

Example 1.7 Evaluate the *mean, standard deviation* and *skewness* of the following sample of data elements:

x	f
10 – 11	2
12 – 13	3
14 – 15	6
16 – 17	8
18 – 19	12
20 – 21	6

Exercise 1.8 Evaluate the *mean, standard deviation* and *skewness* of each of the samples of data elements given in examples 1.1; 1.2; 1.3 and 1.4.

1.5.2 Other Descriptive Measures

1. **Variance:** The variance of a set of data elements is defined as the square of the standard deviation of the data elements
i.e. *variance* = σ^2.

2. **Mode:** The mode of a set of *discrete* data elements is the value of that data element which occurs *most often* in the set. It is thus the best representative of the typical item. It is this form of average that is implied by such expressions as "the average student takes four subjects in first year". This statement implies that there are more students taking four subjects, compared to those taking less than four or more than four. It is not an arithmetic average, and is cause for confusion when one is expecting accurate statistical terminology.

The mode of a set of *continuous* data elements is the value of that data element at which the *probability density* of the set of elements is greatest. [The concept of *probability density* will be addressed when the *probability distribution of a continuous variable* is dealt with.]

The mode of a set of grouped data is the class which has the greatest frequency.

3. **Range:** The *range* of a set of data elements is the difference between the data element with the greatest value and the data element with the least value.

4. **Median:** The *median* of a set of data elements is that value below which 50% of the data elements lie.

The median of a set of data with an odd number of elements is thus the middle element. The median of a set of data with an even number of elements can be one of two options: if the middle two elements are the same, then that is the median, OR if the middle two elements have different values, then the median is the average of the two numbers.

5. **Quartiles:** The *first quartile* (Q_1) of a set of data elements is that value below which 25% of the data elements lie; the *second quartile* (Q_2) of a set of data elements is that value below which 50% of the data elements lie; the *third quartile* (Q_3) of a set of data elements is that value below which 75% of the data elements lie.

6. **Deciles:** The *first decile* (D_1) of a set of data elements is that value below which 10% of the data elements lie; the *second decile* (D_2) of a set of data elements is that value below which 20% of the data elements lie; and so on.

7. **Percentiles:** The *first percentile* (P_1) of a set of data elements is that value below which 1% of the data elements lie; the *second percentile* (P_2) of a set of data elements is that value below which 2% of the data elements lie; and so on.

Note: $Q_2 = D_5 = P_{50} = Median;\ Q_1 = P_{25};\ Q_3 = P_{25};\ D_1 = P_{10};\ D_2 = P_{20};\ldots\ldots\ldots;\ D_9 = P_{90}$

Tutorial 1 Descriptive Statistics

Important: Use the statistics functions on your calculator to find the mean and standard deviation. The skewness must be found using the formula or by using Excel.

1. Determine the arithmetic mean, standard deviation and Pearson's coefficient of skewness of the following sample of data: 2, 41, 19, 10

2. Draw a rod diagram representing the frequency distribution of the following sample of discrete data:

1	0	4	0	1	0	2	1	2	0	0	2	0	4
3	4	3	1	4	0	3	0	1	0	3	1	3	3
0	1	5	1	0	2	0	2	1	3	0	3	1	0

Determine the arithmetic mean, standard deviation and skewness of the sample.

3. The frequency distribution of a sample of 100 data elements of a continuous variable x, rounded off to the nearest one hundredth of a unit, is given by the following table:

x	f
0,00 - 0,09	10
0,10 - 0,19	12
0,20 - 0,29	23
0,30 - 0,39	50
0,40 - 0,49	5

Draw the corresponding histogram. Estimate the arithmetic mean, standard deviation and skewness of the frequency distribution. Is it possible to find the exact mean, standard deviation and skewness? Justify your answer.

4. A sample of 50 steel girders were weighed and the weights were rounded off to the nearest kilogram. The following results were obtained:

345	311	322	356	310	374	363	252	305	323
358	388	307	350	342	309	358	332	387	329
340	240	379	470	247	323	355	403	349	327
329	260	319	362	329	288	361	277	303	311
309	288	369	288	358	301	265	208	293	356

a) Determine the arithmetic mean and standard deviation of the above sample.

b) Arrange the given data in ascending order. Categorise the above data into between 5 and 10 equal sized classes and draw a histogram. Use the classes you have created to estimate the arithmetic mean and standard deviation of the data. Compare the results you get with the answers for (a). Comment on why they are different.

c) Is weight a continuous variable or a discrete variable?

5. The following data represents the distance workers for a construction company travel when they go home for the December break. The distances have been rounded off to the nearest km.

80	3	90	272	80	8	485	176	10	72
294	22	144	160	50	64	224	480	56	141
259	96	104	90	72	37	208	40	120	400
208	160	410	90	80	278	240	192	35	45

a) Arrange the given data in ascending order and draw up a table showing data values, frequency and cumulative frequency.

b) Calculate the following statistics: mean, standard deviation, skewness.

c) i) Find the 90^{th} percentile.
ii) Calculate the median.
iii) X% of the distances are less than 200km. Calculate X.
iv) Y% of the distances are greater than 300km. Calculate Y.

d) Using class lengths of 50km, draw up a frequency distribution table and draw the associated histogram.

e) Use the frequency table to find the class in which the median and mode lie.

Answers

1) 18; 16,83; 1,056

2) 1,55; 1,48; 0,56

3) 0,273; 0,107; -0,833

4) 326,56; 46,295; answers depend on how many bins you use; continuous

5) 151,875; 127,57;1.186; 294; 100; 70%; 10%; 51-100; 51-100.