CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

Chapter I

INTRODUCTION

1.1. Description of certain statistical terms. The statistical methods described in this book are all concerned with the treatment of *variables*. By a variable is meant a quantity which assumes different values that may be measured in some appropriate unit. Height, weight, test scores, readings on a thermometer, etc., are examples of variables, or variates, as they are sometimes called. Variables are usually denoted by X or Y in this book. The number of times a particular value of a variable occurs in a set of observations is called the *frequency* of occurrence of that value, and a table showing the frequency of occurrence of all the values of a variable in a set of observations is named a *frequency distribution table*.

A series of observations may be represented by one value which is called an *average*, and the way in which the different values of the variable lie about this average is described as the *scatter* or *dispersion* of the observations. Measures of averages and scatter are descriptive statistics, since they yield in a condensed form a description of a whole series of observations. This is the first function of statistical method: the other chief function is the examination of various hypotheses which are made about observational data.

It is usually impossible to measure all the values of any variable, so that the data from a single experiment are only a *sample* drawn from the *total population* of possible observations. For example, if the variable is human height, then the total population of that variable would be the height of every man, woman and child ever on earth; it is manifestly impossible to measure all these values and in practice we have to content ourselves with measuring the heights of a sample of some convenient size. The distribution of the total population can usually be expressed in a mathematical form by using a small number of constants or *parameters*. Obviously we can

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

never know the exact values of these parameters, since we cannot measure the whole population, but we can make estimates of them by measuring *random samples*, that is, samples drawn purely at random from the population. These estimates are known as *statistics*, and their accuracy as estimates depends on the size of the sample and the type of distribution of the variable.

In calculating some statistics it is essential to know the number of *degrees of freedom* available for the calculation. The conception of 'degrees of freedom' is not an easy one for the beginner, so that whenever the term is used in this book categorical rules for determining the number of degrees of freedom are supplied. Consideration of the following example may give some idea of the meaning of the term. Suppose 100 shillings are to be shared amongst 10 boys. We may give as many as we like (up to a total of 99) to each of 9 of the boys, but we are bound to give the tenth boy what is left over, i.e. we have only 9 degrees of freedom in sharing out the shillings. If we were told further that 60 shillings were to be shared amongst the oldest 5 boys and the remainder amongst the youngest 5, we should only have 8 degrees of freedom for doing this, since the fifth boy in each group would have to have what was left over —we should not be free to vary his share.

Two variables which are related together, so that a knowledge of the values of one variable indicates likely values of the other, are said to be *associated* or *correlated*. If the variables are unrelated, they are said to be *independent*.

Other terms, applicable to particular methods, will be described in their appropriate places.

1.ii. Notation. A certain amount of symbolism is essential in the description of statistical methods. Unfortunately there is a lack of agreement amongst different authors, which is apt to be confusing to the beginner. For this reason an attempt at a consistent method of notation is made in this book. Being based on first principles it is hoped that it will be readily understood by the learner and will enable him to follow the notation used in the standard books on statistical

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

methods. Symbols in general established use are taken over unchanged. Here again there is confusion, since owing to the multiplicity of statistics the same symbol may have to stand for quite different quantities. Thus the symbol z is used for three different quantities and particular care is needed on the part of the student to avoid misconception in such cases.

Certain symbols have the same significance throughout the whole of this book. For example, N always stands for the total number of observations in a sample and S always signifies 'the sum of'. The other notation is made as unambiguous as possible. Certain Greek letters are used as symbols: a list of these with their pronunciation is given below:

β (beta)	π (pi)
γ (gamma)	ρ (rho)
δ (delta)	σ (sigma)
ζ (zeta)	τ (tau)
η (eta)	ϕ (phi)
μ (mu)	χ (ki)
ν (nu)	

Σ (capital sigma) is used in some places to indicate 'the sum of'. As a rule S is used for summation of sample values and Σ for derived quantities.

Care must be taken by the student to avoid confusing *suffixes* and *indices*. Suffixes are small numbers or letters written after a symbol at the foot, e.g. x_1, σ_x, etc.; these are merely descriptive and confine the use of the symbol to a particular purpose. Indices are small numbers written after and above symbols and have their usual algebraical significance; for example, x^2 (x squared) means x multiplied by x, y^3 (y cubed) means y multiplied by y multiplied by y, and so on.

The usual arithmetical symbols, $+$, $-$, $\times$ and $\div$, have their accustomed significance. There are three other symbols with which the non-mathematical student may not be familiar. Vertical lines drawn on each side of a quantity mean 'the positive numerical value of ', e.g. $|a-b|$ means 'the positive numerical value of the difference between a and b'. Using this

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

notation, therefore, it does not matter whether we write $|a-b|$ or $|b-a|$. Secondly, there is the factorial sign, '!'. This latter is best explained by examples, e.g. 4! stands for $4\times3\times2\times1$, 6! for $6\times5\times4\times3\times2\times1$, and so on. Thirdly, $\binom{n}{p}$ means the number of combinations of n things taken p at a time. Expanded algebraically, $\binom{n}{p}=\frac{n!}{p!(n-p)!}$.

Since a good deal of arithmetical work is involved in certain of the statistical methods described in the following chapters, it is an advantage for the student to be familiar with the use of logarithms (unless he has a calculating machine available).

1.iii. References. Reference is made in the following pages to two invaluable books on statistics and to certain books of statistical tables. The references are made numerically to the following works:

(1) *An Introduction to the Theory of Statistics.* G. Udny Yule and M. G. Kendall. 11th edition. Charles Griffin and Co. Ltd. 1937.
(2) *Statistical Methods for Research Workers.* R. A. Fisher. 9th edition. Oliver and Boyd. 1944.
(3) *Tables for Statisticians and Biometricians*, Part I. Edited by Karl Pearson. Biometrika Office, University College, London.
(4) *Barlow's Tables of Squares, Cubes, Square-roots, Cube-roots and Reciprocals of all Integral Numbers up to* 12,500. E. and F. N. Spon. 4th edition. 1941.
(5) *Statistical Tables for Biological Agricultural and Medical Research.* Fisher and Yates. Oliver and Boyd. 2nd edition. 1942.
(6) *The Advanced Theory of Statistics.* M. G. Kendall. Charles Griffin and Co. Ltd. Vol. I, 1945; vol. II, 1946.

Since no attempt is made in this book to prove or justify the various methods and formulae used, the student wishing to go into such matters is referred to the first two and the last of the foregoing works.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

1.iv. Use and abuse of statistical methods. The student who works conscientiously through the following chapters should learn how to make use of the commoner methods of statistics. He should never forget, however, that statistical methods are merely tools for a research worker. They enable him to describe, relate and assess the value of his observations. They cannot make amends for incorrect observation nor can they of themselves provide a single fact of psychology, biology or any other subject of research. Statistical methods are to the research worker what tools are to a carpenter. The latter has first to learn how to use his tools and he may then by employing them reveal the useful and beautiful purposes to which his material may be put. But the tools themselves must be used for their correct functions. The craftsman will not, for instance, use a mallet and chisel or a fretsaw to plane a plank of wood, nor will he use a hammer to drive in a screw. In the same way statistical methods must only be used by the research worker for the purposes for which they have been devised.

Further, a carpenter's tools cannot tell him directly anything about the materials he is using. They cannot by themselves distinguish between mahogany and deal nor prove that oak is more durable than white wood. No carpenter's tools have ever yet made a piece of wood; similarly no statistical method has ever yet produced a biological fact.

The student is advised, therefore, to try to acquire an understanding of the specific purpose of each statistical method he learns to use, to appreciate the scope of and the assumptions underlying the use of each formula, and to realise that the outcome of each calculation is a statistical statement which has to be interpreted in terms of the particular branch of science from which the data for examination are drawn.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

Chapter II

AVERAGES

2.i. The arithmetic mean. The best known and most useful form of average is the *arithmetic mean*, usually referred to as the 'mean' or the 'average'. It is easily calculated by adding together all the observations to be averaged and dividing the sum or total by the number of observations.

Example 1. Find the mean of the following observations: 22, 24, 20, 23, 21, 19, 23, 22, 20, 22, 20, 22, 23, 25, 21, 21, 22, 24, 23, 22, 23, 21, 22, 21, 23.

Add together all the observations.

$$\begin{aligned} &\text{The sum} && = 549,\\ &\text{The number of observations} && = 25,\\ &\text{The arithmetic mean} && = \frac{\text{sum of observations}}{\text{no. of observations}}\\ & && = \frac{549}{25}\\ & && = 21{\cdot}96. \end{aligned}$$

This procedure may be generalised to cover all cases. If X is a variable which has different values X_1, X_2, X_3, etc., then the arithmetic mean of a number N of such values is the sum of the various values of X, which we denote by $S(X)$, divided by N, the number of them. In general, therefore,

$$m_x = \bar{X} = \frac{S(X)}{N}. \qquad (1)$$

Here m_x and $\bar{X}$ (called X-bar) are different ways of denoting 'the mean of X'.

2.ii. If N is large and no adding machine is available, the process of addition may be very laborious. It may, however, be made easier by the construction of a *frequency distribution table*. This is a table showing how often each value of the variable occurs in the sample under consideration. In Example 1, the values taken by X all lie between 19 and 25

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

inclusive. If we count how many times each different value of X occurs and write the totals in tabular form, we obtain the frequency distribution table given below.

TABLE I

X	f
19	1
20	3
21	5
22	7
23	6
24	2
25	1
	25

In this table the first column, headed X, shows the different values assumed by the variable X in the sample, and the second column, headed f, gives the number of times, or frequency, of occurrence of each. The total of the f column is, of course, the total number of observations we are averaging, i.e. $S(f) = N$. The next step is to write down a third column, headed fX, which is produced by multiplying together the corresponding pairs of numbers in the X and f columns. We then sum the fX column, giving us $\Sigma(fX)$, and the arithmetic mean is then obtained by dividing this sum by N as before; i.e.

$$m_x = \bar{X} = \frac{\Sigma(fX)}{N}. \qquad (2)$$

Example 2. Calculate the mean of the observations in Example 1 by constructing a frequency of distribution table.

X	f	fX
19	1	19
20	3	60
21	5	105
22	7	154
23	6	138
24	2	48
25	1	25
	25	549

$$\Sigma(fX) = 549,$$
$$N = 25,$$
$$\bar{X} = \frac{549}{25}$$
$$= 21{\cdot}96.$$

It will be noted that this result is identical with that obtained in Example 1.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

2.iii. The method of Section 2.ii is useful when the range of the X values is small, but if there are many different values of X the method again becomes laborious. Suppose the observations in Example 1 were the lengths of 25 sticks measured in centimetres, each one being measured to the nearest centimetre. Now let us suppose we had a large number of such sticks and measured them in millimetres. The range of the measurements might now be from 190 to 252 mm., so that if we constructed a frequency distribution table of these we should have 63 different values of X to tabulate. This would

TABLE II

X	f
190–193	2
194–197	4
198–201	7
202–205	12
206–209	19
210–213	24
214–217	27
218–221	35
222–225	26
226–229	21
230–233	18
234–237	13
238–241	6
242–245	5
246–249	2
250–253	1
	222

be tedious, and the calculation may be shortened, with some small sacrifice of accuracy, by subdividing the range of the X's into a convenient number of groups. In practice, a number of groups between 12 and 20 should be chosen, and the best unit for grouping may be found by dividing the range first by 12 and then by 20, and taking a convenient unit in between these two quotients. For instance, if the range is 63, then the results of dividing 63 first by 12 and then by 20 are 5·25 and 3·15. Hence a convenient working unit for grouping would be 4 in this case. This means that we should group the values of

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

X together in 4's, so that the first group would comprise 190, 191, 192 and 193, the second 194, 195, 196 and 197, and so on, the last group being 250, 251, 252 and 253. We could now construct a frequency distribution table of these groups. Such a table might be, for example, as Table II. This gives the frequency distribution of the lengths of 222 sticks measured in 4 mm. groups.

Now the method of calculating the mean length of the sticks from such a table depends on the assumption that the average length of the sticks in each group is equal to the mean value of X for that group. For instance, there are 12 sticks in the 202–205 group and we shall assume that the average length of those 12 sticks is equal to the average of 202, 203, 204 and 205, i.e. 203·5. This is, of course, an assumption, but the larger the number of readings in each group the nearer it becomes to being true.

Care should be taken in arranging the grouping that this assumption should be as nearly as possible true for the end groups, leaving the middle ones, which usually have more readings in them, to look after themselves. For instance, if the two readings in the 190–193 group were both 190, then their average would be 190 instead of 191·5, which is assumed by the grouping. In such a case it would have been better to have started the grouping from 189, which would assume an average for the group of 190·5.

In order to calculate the mean of the observations in a grouped frequency distribution table, such as Table II, we take an *arbitrary origin*, or starting point, and then calculate the discrepancy between this point and the true mean. Let us take our arbitrary origin near the middle of the range, since this simplifies the arithmetic. It is convenient to have it at the centre of a group so we will choose the 218–221 group, so that our arbitrary origin will be 219·5, the centre of the group. This group we number 0. The next group in the table, 222–225, has an average of 223·5, which is one group unit, or working unit, above the arbitrary origin, and we therefore number this group 1. In a similar manner the 226–229 group is numbered 2, and so on.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-11620-6 - Statistical Calculation for Beginners
E. G. Chambers
Excerpt
More information

The 214–217 group averages 215·5, which is one working unit less than the arbitrary origin, so this group is numbered −1. Similarly the 210–213 group becomes −2, and so on.

Example 3. Calculate the mean of the data in Table II.

X	f	x	fx
190–193	2	−7	−14
194–197	4	−6	−24
198–201	7	−5	−35
202–205	12	−4	−48
206–209	19	−3	−57
210–213	24	−2	−48
214–217	27	−1	−27
218–221	35	0	−253*
222–225	26	1	26
226–229	21	2	42
230–233	18	3	54
234–237	13	4	52
238–241	6	5	30
242–245	5	6	30
246–249	2	7	14
250–253	1	8	8
	222		256
			−253
			3

$\Sigma(fx) = 3,$

$N = 222,$

$D = \frac{3}{222} = 0{\cdot}0135,$

$w = 4,$

$m_a = 219{\cdot}5,$

$m_x = 219{\cdot}5 + 0{\cdot}0135 \times 4$

$= 219{\cdot}5 + 0{\cdot}054$

$= 219{\cdot}554.$

* Since there will be no entry in the fx column corresponding to $x = 0$, this is a convenient place to add the negative entries in the fx column.

We can now replace the X column by another column, which we will head 'x', which indicates the number of working units that each X-group lies away from the arbitrary origin. The X column is now neglected and a fourth column, headed fx, is written down. This is obtained by multiplying corresponding entries in the f and x columns. By adding this column we get $\Sigma(fx)$, and the discrepancy, D, between the true mean and the arbitrary origin, *in working units*, is given by

$$D = \frac{\Sigma(fx)}{N}. \qquad (3)$$

This quantity D tells us how many working units the true

© Cambridge University Press www.cambridge.org