PART I

SINGLE VARIABLES

CHAPTER 1

GENERAL THEORY FOR A SINGLE VARIABLE

In this chapter an account is given of the theoretical concepts which are required for the treatment of observations of a single variable. The discussion is confined to those parts of the theory which can be developed without the assumption of a particular form for the frequency distribution of the observations. The practical methods of estimation based on a small number of observations are discussed, and illustrated by examples.

1.1 PROBABILITY AND FREQUENCY

If a large number N of observations η are made of a quantity, and the number of these observations which have the value y is N_y, the probability that a particular observation will yield the value y is defined as

$$\Pr\{\eta = y\} = \Pr\{y\} = \mathop{\mathrm{Lt}}_{N\to\infty} N_y/N. \tag{1}$$

The probability of obtaining a value y in a single observation is then by definition proportional to the frequency of occurrence of the value y in a long series of observations. Since $\sum_y N_y = N$,

$$\sum_y \Pr\{y\} = 1. \tag{2}$$

For two variables ξ and η, the probability that a pair of observations yields the values x and y is written $\Pr\{xy\}$. This probability may be evaluated by considering a large number N of pairs of observations. The number of pairs for which $\xi = x$ will be denoted by N_x. Of these N_x, the number for which η has also the value y will be denoted by $N_{y|x}$. Then

$$\Pr\{xy\} = \mathop{\mathrm{Lt}}_{N\to\infty} \frac{N_{y|x}}{N} = \mathop{\mathrm{Lt}}_{N\to\infty} \frac{N_x}{N}\frac{N_{y|x}}{N_x},$$

and so

$$\Pr\{xy\} = \Pr\{x\}\Pr\{y\,|\,x\}, \tag{3}$$

where $\Pr\{y\,|\,x\}$ is the probability that $\eta = y$, given that $\xi = x$. Equation (3) is referred to as the product rule for probabilities.

It will often be true that the value obtained for η does not depend on the value x of ξ. If x and y are independent, so that

$$\Pr\{y\,|\,x\} = \Pr\{y\},$$

then the product rule takes the simpler form

$$\Pr\{xy\} = \Pr\{x\}\Pr\{y\}. \tag{4}$$

Since (3) can also be put in the form

$$\Pr\{xy\} = \Pr\{y\}\Pr\{x \mid y\},$$

it follows that

$$\Pr\{x \mid y\} = \Pr\{x\}\Pr\{y \mid x\}/\Pr\{y\}, \tag{5}$$

or, for variations of x with y fixed,

$$\Pr\{x \mid y\} \propto \Pr\{x\}\Pr\{y \mid x\}. \tag{6}$$

The proportional relation (6) is referred to as Bayes' Theorem. The three terms are called the posterior probability (i.e. the probability after y is fixed), the prior probability (before y is fixed), and the likelihood, respectively.

1.1.1 *Notation*

The symbol $f(y)$ will be used for the probability, $\Pr\{\eta = y\}$, that the observation or measurement will yield the value y. Usually the possible values y are not discrete, but form a continuous set. $f(y)$ is then defined so that $f(y)\,dy$ is the probability that the observation lies in a range dy centred at y. Thus $f(y)$ is referred to as the probability density function. It is also called the frequency function, since the probability of an observation lying in a given range is by definition proportional to the frequency of occurrence of values in that range in a long series of observations.

For a discrete distribution, from (1.1,2),

$$\sum_y f(y) = 1, \tag{1}$$

and for a continuous distribution the sum becomes the integral

$$\int f(y)\,dy = 1. \tag{2}$$

The probability integral or the distribution function is the integral of the frequency function. It is often denoted by $F(y)$, but here, following the usage of the *Biometrika Tables for Statisticians*, the symbol

$$P(y) = \int_{-\infty}^{y} f(u)\,du \tag{3}$$

will be employed. The probability integral $P(y)$ gives the probability that an observed value will be less than or equal to y. The differential of $P(y)$ is

$$dP(y) = P(y+dy) - P(y) = f(y)\,dy. \tag{4}$$

The probability that an observed value is greater than or equal to y will be denoted by $Q(y)$. Thus

$$Q(y) = 1 - P(y) = \int_y^\infty f(u)\,du. \tag{5}$$

If x and y are independent, then from (1.1,4) the probability that x and y lie simultaneously in the ranges dx and dy about x and y is

$$f_1(x) f_2(y)\,dx\,dy.$$

If this probability is written $f(x,y)\,dx\,dy$, where $f(x,y)$ is the combined frequency function, then, when x and y are independent,

$$f(x,y) = f_1(x) f_2(y). \tag{6}$$

1.2 EXPECTATION AND VARIANCE

If a very large number of measurements y of a quantity are made, the fraction of the observations lying in the range dy about y is identical with the probability that a single observation lies in that range, both being equal to $f(y)\,dy$. The average Y of the measurements as the number of observations approaches infinity is given by

$$Y = \int y f(y)\,dy,$$

and is often referred to as the expectation of y, written $E(y)$. Thus

$$E(y) = Y = \int y f(y)\,dy. \tag{1}$$

Y is also referred to as the population mean, the 'population' being simply the aggregate of all possible observations y.

If y and z are two variables, not necessarily independent, and their combined probability distribution is described by the frequency function $f(y,z)$, so that $f(y,z)\,dy\,dz$ is the probability that the observations lie simultaneously in the range dy about y and dz about z, then the expectation of the sum of y and z is

$$\begin{aligned} E(y+z) &= \iint (y+z) f(y,z)\,dy\,dz \\ &= \int y \int f(y,z)\,dz\,dy + \int z \int f(y,z)\,dy\,dz. \end{aligned}$$

Now $f(y,z)\,dz$ gives the probability density for y when the second variable lies in the range dz about z, and so the integral of this quantity over z must be $f(y)$. Hence

$$E(y+z) = E(y) + E(z), \tag{2}$$

and the expectation of a sum is the sum of the individual expectations.

For the product, if the two quantities are independent,

$$E(yz) = \iint yzf(y,z)\,dy\,dz = \iint yzf_1(y)f_2(z)\,dy\,dz,$$

and so $$E(yz) = E(y)\,E(z). \tag{3}$$

The expectation of the product of independent variables is the product of their expectations.

The variance of a quantity is defined as the average of the squares of the deviations from the population mean Y. In symbols

$$\operatorname{var} y = E(y-Y)^2.$$

The square root of the variance is called the standard deviation or the standard error, and is denoted by σ. Thus

$$\operatorname{var} y = \sigma^2 = E(y-Y)^2 = \int (y-Y)^2 f(y)\,dy. \tag{4}$$

The variance is clearly a measure of the spread of the observations about the population mean.

Since $E(y) = Y$,

$$E(y-Y)^2 = E(y^2) - 2YE(y) + Y^2 = E(y^2) - Y^2, \tag{5}$$

a result often useful in calculations.

The variance of the sum of two quantities is

$$E(y+z-Y-Z)^2 = E(y-Y)^2 + E(z-Z)^2 + 2E(y-Y)(z-Z).$$

The last term is the quantity defined as the covariance of y and z,

$$\operatorname{cov}(y,z) = E(y-Y)(z-Z). \tag{6}$$

Then $$\operatorname{var}(y+z) = \operatorname{var} y + \operatorname{var} z + 2\operatorname{cov}(y,z). \tag{7}$$

If y and z are independent, it follows from (3) that

$$E(y-Y)(z-Z) = \{E(y-Y)\}\{E(z-Z)\} = 0.$$

Thus when y and z are independent

$$\operatorname{cov}(y,z) = 0, \tag{8}$$

$$\operatorname{var}(y+z) = \operatorname{var} y + \operatorname{var} z. \tag{9}$$

The variance of the product yz can be obtained by expanding

$$E(yz-YZ)^2 = E\{Z(y-Y) + Y(z-Z) + (y-Y)(z-Z)\}^2.$$

If the quantities y and z are independent, then, using (3),

$$E(yz-YZ)^2 = Z^2E(y-Y)^2 + Y^2E(z-Z)^2 + E(y-Y)^2(z-Z)^2,$$

the other terms vanishing since $E(y-Y) = 0 = E(z-Z)$. Thus

$$\operatorname{var} yz = Z^2 \operatorname{var} y + Y^2 \operatorname{var} z + (\operatorname{var} y)(\operatorname{var} z). \qquad (10)$$

Usually the last term is much smaller than the others, and the approximation

$$\operatorname{var} yz = Z^2 \operatorname{var} y + Y^2 \operatorname{var} z \qquad (11)$$

may be used.

If $\phi(y, z, \ldots)$ is an arbitrary function, then

$$\phi(y,z,\ldots) = \phi(Y,Z,\ldots) + \frac{\partial\phi}{\partial Y}(y-Y) + \frac{\partial\phi}{\partial Z}(z-Z) + \ldots. \qquad (12)$$

If the deviations of the observations y from the population means Y are small, the higher order terms in the expansion (12) may be neglected, and so, by (2),

$$E\{\phi(y,z,\ldots)\} = \phi(Y,Z,\ldots). \qquad (13)$$

If the variables $y, z, \ldots$ are independent,

$$E\{\phi(y,z,\ldots)-\phi(Y,Z,\ldots)\}^2 = \left(\frac{\partial\phi}{\partial Y}\right)^2 E(y-Y)^2 + \left(\frac{\partial\phi}{\partial Z}\right)^2 E(z-Z)^2 + \ldots,$$

or

$$\operatorname{var}\phi(y,z,\ldots) = \left(\frac{\partial\phi}{\partial Y}\right)^2 \operatorname{var} y + \left(\frac{\partial\phi}{\partial Z}\right)^2 \operatorname{var} z + \ldots. \qquad (14a)$$

When $\phi(y, z, \ldots)$ is a product of powers $Cy^a z^b \ldots$, then from (14a)

$$\frac{\operatorname{var}\phi}{\phi^2} = \frac{a^2 \operatorname{var} y}{Y^2} + \frac{b^2 \operatorname{var} z}{Z^2} + \ldots, \qquad (14b)$$

or

$$\left(\frac{\text{S.D.}\,\phi}{\phi}\right)^2 = \left(\frac{a\,\text{S.D.}\,y}{Y}\right)^2 + \left(\frac{b\,\text{S.D.}\,z}{Z}\right)^2 + \ldots. \qquad (14c)$$

1.3 TYPES OF OBSERVED QUANTITY

The quantities observed in practical cases would appear to fall into one of two classes. Firstly, there are those quantities which are constant in magnitude, and which will be referred to as 'controlled' quantities. Many physical quantities are of this type—for example, the mass of a body, the velocity of light, etc. Secondly, there are those quantities which are inherently variable, the value of the various members of the population being distributed according to some frequency function $f(y)$.

Such quantities will be referred to as 'uncontrolled' quantities. Many of the quantities occurring in the biological sciences are of this class. Typical examples are the heights of men, the milk yield of cows, etc.

For controlled quantities, the observed value, population mean and true value all coincide, and an error-free observation will yield the true value Y of the quantity as determined by the experiment. There may, of course, still be unallowed-for systematic errors which cause the value Y to differ from that given by other experiments. For uncontrolled quantities an observation will yield a value y whose expectation is the population mean Y.

Either type of quantity may also be subject to experimental errors of observation. These errors are regarded as random quantities, equally likely to be positive or negative for any particular observation, which are produced by slight transient and unaccounted changes in the experimental conditions and apparatus. Thus if y' is the error-free or corrected value corresponding to an observed value y, the error is

$$\delta = y - y', \tag{1}$$

and the assumption of randomness leads to

$$E(\delta) = 0, \quad E(y) = y'. \tag{2}$$

For controlled variables $y' = Y$, and

$$E(y) = E(y') = Y, \tag{3}$$

so the population mean is the true value, while

$$\operatorname{var} y = \operatorname{var} y' + \operatorname{var} \delta = \operatorname{var} \delta \tag{4}$$

and the variance is a measure of the experimental error δ.

For uncontrolled quantities,

$$E(y) = E(y') = Y = Y', \tag{5}$$

and the population mean is unaffected by observational errors. Also

$$\operatorname{var} y = \operatorname{var} y' + \operatorname{var} \delta, \tag{6}$$

and the variance of the observed quantities y is greater than that of the error-free quantities y'.

1.4 ESTIMATION

In most cases the frequency function $f(y)$ is not known in detail, and a very large number of observations would be required to determine it. But the number of observations available, often referred to as the sample, is usually comparatively small. Hence some method of estimating the true value or population mean is required when the number n of observations y_i is small.

An estimate $\hat{Y}$ is said to be unbiased if its expectation is the true value Y of the quantity. Now any linear function

$$\sum_{i=1}^{n} \lambda_i y_i$$

will provide an unbiased estimate of Y, since

$$E\left\{\sum_{i=1}^{n} \lambda_i y_i\right\} = \sum_{i=1}^{n} \lambda_i E(y_i) = Y \sum_{1}^{n} \lambda_i,$$

and so
$$E\{\Sigma\lambda_i y_i/\Sigma\lambda_i\} = Y. \tag{1}$$

Which estimate will be the best depends on what is adopted as the criterion of 'best'. The most common criterion is a least-squares one, in which the estimate $\hat{Y}$ is chosen so that

$$\Sigma v_i^2 = \Sigma(y_i - \hat{Y})^2 \tag{2}$$

should be a minimum. Then differentiation leads to

$$\Sigma(y_i - \hat{Y}) = 0,$$

or
$$\hat{Y} = \Sigma y_i/n = \bar{y}, \tag{3}$$

where $\bar{y}$ is the sample mean. Thus the estimate obtained from the least-squares postulate is just the sample mean. This is the estimate almost always adopted. A discussion of other postulates is given in § 1.6.

The standard deviation σ can be estimated in terms of the deviations from the mean, usually called the residuals. For the ith residual

$$v_i = y_i - \bar{y}, \tag{4}$$

$$\begin{aligned} E(v_i^2) &= E(y_i - n^{-1}\Sigma y_j)^2 \\ &= E\{(y_i - Y) - n^{-1}\Sigma(y_j - Y)\}^2 \\ &= E\{(n-1)\,n^{-1}(y_i - Y) - n^{-1}\sum_{j \neq i}(y_j - Y)\}^2. \end{aligned}$$

If the observations are independent, $E(y_i - Y)(y_j - Y) = 0$, and

$$E(v_i)^2 = (n-1)^2 n^{-2} \operatorname{var} y + n^{-2}(n-1) \operatorname{var} y,$$

or
$$E(v_i^2) = \frac{n-1}{n} \operatorname{var} y = \frac{n-1}{n} \sigma^2. \tag{5}$$

Hence the quantity s^2 defined by the equation

$$s^2 = \Sigma v_i^2/(n-1) \tag{6}$$

will provide an unbiased estimate of the variance σ^2. Of course, s will also provide an estimate of the standard deviation σ. It is interesting to note that s is not an unbiased estimate of σ, though the bias is almost always negligible. This point is discussed in § 2.5.3.

Σv_i^2 may be calculated by evaluating the individual residuals. It may also be calculated from the formula

$$\Sigma v_i^2 = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2. \tag{7}$$

In estimating the standard deviations of combinations of observed quantities, it is usually necessary to replace in (1.2,9), (1.2,11), and (1.2,14a), the population means and variances by their estimates $\bar{y}$ and s^2 respectively.

1.4.1 *The arithmetic mean*

The mean has been adopted as the best estimate of the population mean or true value Y. Since it is a linear function of the observations,

$$\text{var } \bar{y} = n^{-1} \text{var}\, y. \tag{1}$$

Thus, using (1.4,6),

$$s^2(\bar{y}) = \Sigma v_i^2/n(n-1) \tag{2}$$

will provide an unbiased estimate of the variance of the mean $\bar{y}$.

It is perhaps worth while to emphasize the distinction between $\sigma(y)$, the standard deviation of an observation, and $\sigma(\bar{y})$, the standard deviation of the mean. The standard deviation of an observation is a measure of the spread of the individual observations about the true value or population mean, and its magnitude does not decrease as the number of observations is increased. The standard deviation of the mean does decrease as n is increased, being in fact proportional to $n^{-\frac{1}{2}}$. The larger the number of observations, the less the expected deviation of $\bar{y}$ from the true value or population mean.

Hence, for a controlled variable subject to error, increasing the number of observations will increase the accuracy of the estimate. But there is usually a practical limit to the number of observations that it is profitable to make, since there will almost certainly be undetected systematic errors which are not reduced by increasing n.

1.4.2 *Example*

In this example the refractive index μ of a glass prism will be calculated from measurements of the angle A of the prism, and the angle of minimum deviation θ for a ray of light passing through the prism. In any text-book on optics it is shown that

$$\mu = \frac{\sin \frac{1}{2}(A+\theta)}{\sin \frac{1}{2}A}. \tag{1}$$