10 JOINT DISTRIBUTIONS OF ATTRIBUTES

10.1 Data on two or more attributes

In some investigations, it may be appropriate to collect data, for a given set of individuals, on more than one character at the same time. The object here would be to look for any relationship that may obtain among the characters. In this chapter we shall be concerned with data on several *attributes*. The case of several *variables* will be taken up in Chapters 11 and 12. (There may, of course, be a third case, where some of the characters are attributes and the others are variables. But this case can be dealt with by suitably adapting the methods appropriate for the first two cases and hence will not be discussed separately.)

Thus the data may relate to the proficiency in English and the proficiency in mathematics of a group of high-school students (for each attribute there being, say, five classes : very good/good/mediocre/bad/very bad) ; or to the subject-matter (fiction/non-fiction) and the readability (easy-reading/difficult-reading) of a number of books ; or to the sex, the economic status (rich/middle-class/poor) and the level of education(illiterate/primary/high-school/university) of a group of adults.

In Table 10.1 data on two attributes are presented in a summary form. The figure in each cell stands for the number of individuals (i.e. the frequency) corresponding to a pair of forms of the two attributes. Thus, e.g., 19 is the number of adults attacked with fever among those administered quinine during the period, 193 the number of adults attacked with fever among those not administered quinine, and so on. The cell-frequencies, together with their grand total, give the *joint (frequency) distribution* of the attributes, because they show how the two attributes vary jointly in the given group of individuals. From the joint distribution, we also obtain two other types of distribution. Thus the row-totals (marginal frequencies), together with the grand total, give the distribution of the attribute

'precautionary measure', to be called the *marginal frequency distribution* of precautionary measure in the present context. On the other hand, the column-totals, together with the grand total, give the marginal distribution of 'outcome'. The other type of distribution is given by each column or each row of frequencies of the table, together with the corresponding column or row-total. Take, e.g., the frequencies in the first row, together with the row-total 606. For these frequencies the form of the first attribute, 'precautionary measure', is the same but the form of the second varies. As such it is said to give a *conditional frequency distribution*—the conditional distribution of 'outcome' for the form 'quinine used' of precautionary measure. Similarly, the second row gives the conditional distribution of outcome for the form 'no quinine used' of precautionary measure. The frequencies in the first column and those in the second, in the same way, give the conditional distributions of the attribute 'precautionary measure' for the forms 'attacked with malaria' and 'not attacked with malaria', respectively, of the attribute 'outcome'.

TABLE 10.1

DATA ON THE USE OF QUININE AND INCIDENCE OF MALARIA COLLECTED IN AN INVESTIGATION IN A STATE OF INDIA

(Each figure relates to the number of adults in each category among a total of 3,540 adults)

Precautionary measure (B)	Outcome (A): Attacked with malaria (A)	Outcome (A): Not attacked with malaria (α)	Total
Quinine used (B)	19 (f_{AB})	587 ($f_{\alpha B}$)	606 (f_B)
No quinine used (β)	193 ($f_{A\beta}$)	2,741 ($f_{\alpha\beta}$)	2,934 (f_β)
Total	212 (f_A)	3,328 (f_α)	3,540 (n)

Clearly, the attributes need not have just two forms each; i.e., the table need not be a 2×2 table. Thus in Table 10.2 we have data on two attributes each of which has three forms.

In each case, we might consider the relative frequencies, instead of the frequencies, which would also give the distributions—joint, marginal or conditional—of the attributes, although in a different form.

10.2 Independence and association*

Consider again two attributes, A and B. In the 2×2 case, the two forms of A may be denoted by A (the 'positive' form, indicating the *presence* of the character A) and α (the 'negative' form, indicating the *absence* of the character A) and, similarly, the two forms of B may be denoted by B and β. The four cell-frequencies may be denoted by f_{AB}, $f_{A\beta}$, $f_{\alpha B}$ and $f_{\alpha\beta}$ and the total by n. Also, the (marginal) frequencies for the A-classes may be denoted by f_A and f_α, and the (marginal) frequencies for the B-classes by f_B and f_β. Thus

$$\left.\begin{aligned} f_A=f_{AB}+f_{A\beta},\ f_\alpha=f_{\alpha B}+f_{\alpha\beta},\\ f_B=f_{AB}+f_{\alpha B},\ f_\beta=f_{A\beta}+f_{\alpha\beta}, \end{aligned}\right\} \quad \ldots \quad (10\,1)$$

and

$$n=f_{AB}+f_{A\beta}+f_{\alpha B}+f_{\alpha\beta} \quad \ldots \quad (10.2a)$$

$$=f_A+f_\alpha \quad \ldots \quad (10.2b)$$

$$=f_B+f_\beta. \quad \ldots \quad (10.2c)$$

Suppose the individuals under consideration constitute the population itself and not just a sample from the population. Also suppose that none of the maginal frequencies is zero. Then the ratios f_{AB}/f_A and $f_{\alpha B}/f_\alpha$ give, respectively, the proportions of members of the population having B, among those having A and among those having α. If these proportions be equal, we may say that the presence or absence of the character A in an individual does not in any way determine whether B will be present. A and B may then be called *statistically* unrelated or *independent*. As opposed to the notion of independence, there is the notion of *association*. Thus A and B are said to be associated if they are not independent.

We have seen that, for A and B to be independent, we must have

$$\frac{f_{AB}}{f_A}=\frac{f_{\alpha B}}{f_\alpha}. \quad \ldots \quad (10.3)$$

This implies

$$\frac{f_{AB}}{f_A}=\frac{f_{AB}+f_{\alpha B}}{f_A+f_\alpha}=\frac{f_B}{n}$$

or

$$f_{AB}=\frac{f_Af_B}{n}. \quad \ldots \quad (10.4a)$$

*The ideas in this section are comparable to those in Section 3.5.

Actually, (10.3) also implies

$$f_{\alpha B}=\frac{f_\alpha f_B}{n}, \qquad \cdots (10.4b)$$

$$f_{A\beta}=\frac{f_A f_\beta}{n} \qquad \cdots (10.4c)$$

and

$$f_{\alpha\beta}=\frac{f_\alpha f_\beta}{n}, \qquad \cdots (10.4d)$$

Since equation (10.4a) itself leads to (10.4b), (10.4c), (10.4d) and to (10.3), it is taken as the defining equation for the independence of A and B. This is done irrespective of whether f_A and/or f_B is zero.

Suppose A and B are not independent, i.e. are associated. We may distinguish two cases. (i) If

$$f_{AB}>\frac{f_A f_B}{n}, \qquad \cdots (10.5)$$

A and B occur together more frequently than they would have if they had been independent. Hence in this case the attributes are said to be *positively associated* (or, simply, associated). (ii) On the other hand, if

$$f_{AB}<\frac{f_A f_B}{n}, \qquad \cdots (10.6)$$

i.e. if A and B occur together less frequently than they would have if they had been independent, then they are said to be *negatively associated* (or *disassociated*).

As regards the definition of *perfect association*, we may adopt one of two alternatives. (1) Thus we may say that there is perfect positive association between A and B if all A's are B's *and/or* all B's are A's, i.e. if $f_{A\beta}=0$ *and/or* $f_{\alpha B}=0$. Likewise, there may be said to be perfect negative association if no A's are B's *and/or* no α's are β's, i.e. if $f_{AB}=0$ *and/or* $f_{\alpha\beta}=0$. (2) Alternatively, we may say that there is perfect positive association if all A's are B's *and* all B's are α's, i.e. if $f_{A\beta}=0$ *and* $f_{\alpha B}=0$; and that there is perfect negative association if no A's are B's *and* no α's are β's, i.e. if $f_{AB}=0$ *and* $f_{\alpha\beta}=0$.

To keep these two cases distinct, the association will be said to be *complete* (positive or negative) in the first case and to be *absolute* in the second.

10.3 Measures of association for the 2×2 case

We shall consider measures of the extent to which A and B, each of which occurs in two possible forms, may be said to be associated. Clearly, there are certain desiderata that such a measure should fulfil. For one thing, it should be independent of the total frequency n, just as, say, the mean or the moments are, and should thus depend on the relative frequencies in the cells rather than on their frequencies. Secondly, it should be zero in the case of independence, negative in the case of negative association and positive in the case of positive association. Thirdly, it should increase from its lowest possible value through zero to its highest possible value as we proceed from perfect negative association through independence to perfect positive association. Lastly, it should preferably vary between two definite limits, like -1 and $+1$.

Obviously, the difference

$$\delta_{AB}=f_{AB}-\frac{f_A f_B}{n}, \qquad \ldots \quad (10.7)$$

between the actual frequency for the cell (A, B) and the value that it should assume if A and B are independent, may serve as the basis for such a measure. Keeping all the desiderata in mind, one may use

$$Q_{AB}=\frac{n\delta_{AB}}{f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}} \qquad \ldots \quad (10.8)$$

$$=\frac{f_{AB}f_{\alpha\beta}-f_{A\beta}f_{\alpha B}}{f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}} \qquad \ldots \quad (10.9)$$

as a measure of association. It has been called the *coefficient of association* between A and B and is due to Yule. It may be seen that Q statifies all the desiderata stated above. In particular, $Q_{AB}=0$ if and only if $\delta_{AB}=0$, i.e. if and only if A and B are independent. Its lowest possible value (-1) occurs when $f_{AB}f_{\alpha\beta}=0$, i.e. when $f_{AB}=0$ and/or $f_{\alpha\beta}=0$, i.e. when there is *complete* negative association between A and B. Likewise, its highest possible value $(+1)$ occurs when there is *complete* positive association between A and B.

A measure with the same general properties as those of Q_{AB} is the *coefficient of colligation* Y_{AB}, also due to Yule and defined by

$$Y_{AB}=\frac{\sqrt{f_{AB}f_{\alpha\beta}}-\sqrt{f_{A\beta}f_{\alpha B}}}{\sqrt{f_{AB}f_{\alpha\beta}}+\sqrt{f_{A\beta}f_{\alpha B}}}. \qquad \ldots \quad (10.10)$$

There is yet a third measure, viz.

$$V_{AB}=\frac{n\delta_{AB}}{\sqrt{f_A f_\alpha f_B f_\beta}}=\frac{f_{AB}f_{\alpha\beta}-f_{A\beta}f_{\alpha B}}{\sqrt{f_A f_\alpha f_B f_\beta}}. \quad \cdots \quad (10.11)$$

This has properties similar to those of Q and Y, but unlike Q and Y, $V=\mp 1$ when and only when there is *absolute association* between the two characters.

To prove this result, let us use the symbols a, b, c and d for f_{AB}, $f_{A\beta}$, $f_{\alpha B}$ and $f_{\alpha\beta}$, respectively. Then

$$V_{AB}=\frac{ad-bc}{\{(a+b)(c+d)(a+c)(b+d)\}^{1/2}},$$

and this equals ∓ 1 if and only if

$$(ad-bc)^2=(a+b)(c+d)(a+c)(b+d),$$

i.e. if and only if

$$a^2(bc+bd+cd)+b^2(ac+ad+cd)+c^2(ab+ad+bd)$$
$$+d^2(ac+ab+bc)+4abcd=0. \quad \cdots \quad (10.12)$$

But this expression can vanish only if at least two of the non-negative quantities, a, b, c and d, vanish. We assumed, however, that the marginal frequencies are all non-zero, precluding the cases $a=b=0$, $c=d=0$, $a=c=0$, and $b=d=0$. Hence $V_{AB}=\pm 1$ if and only if $b=c=0$ or $a=d=0$. In the former case there is absolute positive association between A and B and $V_{AB}=+1$, while in the latter there is absolute negative association and $V_{AB}=-1$.

Ex. 10·1 For the data of Table 10·1, let us denote by A the attribute 'outcome' and by B the attribute 'precautionary measure'. We then have for the data

$$Q_{AB}=\frac{19\times 2741-193\times 587}{19\times 2741+193\times 587}=\frac{52079-113291}{52079+113291}$$
$$=-61212/165370=-0\cdot 37015,$$

$$Y_{AB}=\frac{\sqrt{52079}-\sqrt{113291}}{\sqrt{52079}+\sqrt{113291}}=\frac{228\cdot 208-336\cdot 587}{228\cdot 208+336\cdot 587}$$
$$=-108\cdot 379/564\cdot 795=-0\cdot 19189,$$

while

$$V_{AB}=\frac{19\times 2741-193\times 587}{\sqrt{212\times 3328\times 606\times 2934}}=\frac{52079-113291}{\sqrt{125444583\times 10^4}}$$
$$=-61212/(11200\times 10^2)=-0\cdot 05465.$$

Each of the measures indicates only a slight negative association between the two attributes. In other words, there is only slight evidence in support of the belief that use of quinine is generally followed by exemption from attack of malaria.

One important point is to be noted in this connection. The notion of independence is, by its very nature, related to a population and so is the notion of association. However, it is perfectly legitimate to study the presence or absence of association in the population from sample data. We may thus compute a measure of association according to one of the formulæ given above, where n is now to be regarded as the sample size and the frequencies as the sample frequencies for the cells or the margins. As in many other cases, the sample measure is to be taken, at least for large n, as a good approximation to the corresponding population value. Indeed, the data of Table 10.1 are, more appropriately, to be considered to be sample data for a random sample of size 3,540 taken from the population of all adults in the given Indian State.

10.4 Manifold two-way ($k \times l$) classification

... two attributes,