Properties of Sufficiency and Statistical Tests

M.S. Bartlett
Imperial Chemical Industries, Ltd.,
United Kingdom

Introduction

1—In a previous paper*, dealing with the importance of properties of sufficiency in the statistical theory of small samples, attention was mainly confined to the theory of estimation. In the present paper the structure of small sample tests, whether these are related to problems of estimation and fiducial distributions, or are of the nature of tests of goodness of fit, is considered further.

The notation $a|b$ implies as before that the variate a is conditioned by† a given value of b. The fixed variate b may be denoted by $|b$, and analogously if b is clear from the context, $a|b$ may be written simply as $a|$. Corresponding to the idea of ancillary information introduced by Fisher for the case of a single unknown θ, where auxiliary statistics control the accuracy of our estimate, I have termed a conditional statistic of the form $T|$, quasi-sufficient, if its distribution satisfies the "sufficiency" property and contains all the information on θ. In the more general case of other unknowns, such a statistic may contain all the *available* information on θ.

* Bartlett (1936*a*): I have noticed an error on p. 128 of this paper which I will take this opportunity to correct. In the example the order of magnitude of the two observations was lost sight of. The information in one observation, *if it be recorded whether it is the greater or smaller*, is found to be $1 \cdot 386$, and is thus more than that in the mean.

† With this notation and phraseology, b is in general a known statistic. Inside a probability bracket, it may sometimes be necessary to stress that the distribution depends on an unknown parameter θ, and the usual notation is then adopted of writing $p(a)$ more fully as $p(a|\theta)$, and $p(a|b)$ as $p(a|b, \theta)$.

S. Kotz et al. (eds.), *Breakthroughs in Statistics*

Sufficient Statistics and Fiducial Distributions

2—It has been noted (Bartlett 1936*a*) that if our information on a population parameter θ can be confined to a single degree of freedom, a fiducial distribution for θ can be expected to follow, and possible sufficiency properties that would achieve this result have been enumerated. A corresponding classification of fiducial distributions is possible.

Since recently Fisher (1935) has put forward the idea of a simultaneous fiducial distribution, it is important to notice that the sufficient set of statistics $\bar{x}$ and s^2 obtained from a sample drawn from a normal population (usual notation) do not at once determine fiducial distributions for the mean m and variance σ^2. That for σ^2 follows at once from the relation

$$p(\bar{x}, s^2|m, \sigma^2) = p(\bar{x}|m, \sigma^2)p(s^2|\sigma^2), \tag{1}$$

but that for $\bar{x}$ depends on the possibility of the alternative division

$$p\{\Sigma(x - m)^2|\sigma^2\}p(t), \tag{2}$$

where t depends only on the unknown quantity m. No justification has yet been given that because the above relations are equivalent respectively to fiducial distributions denoted by $fp(m|\sigma^2)fp(\sigma^2)$ and $fp(\sigma^2|m)fp(m)$, and hence symbolically to $fp(m, \sigma^2)$, that the idea of a simultaneous fiducial distribution, and hence by integration the fiducial distribution of either of the two parameters, is valid when *both* relations of form (1) and (2) do not exist (Bartlett 1936*b*). Moreover, even in the above example, the simultaneous distribution is only to be regarded as a symbolic one, for there is no reason to suppose that from it we may infer the fiducial distribution of, say, $m + \sigma$.

3—In certain cases where a fiducial distribution exists for a population parameter, it will similarly exist for the corresponding statistic in an unknown sample. If, for example, a sufficient statistic T_1 exists for θ in the known sample S_1, we shall have a corresponding unknown statistic T_2 in an unknown sample S_2, and an unknown statistic T for the joint sample S. If we write

$$p(T_1|\theta)p(T_2|\theta) = p(T|\theta)p(T_1, T_2|T), \tag{3}$$

then $p(T_1, T_2|T)$ depends only on T_1 and the unknown T_2 (for which T_1 may be regarded as a sufficient statistic), and will lead to a fiducial distribution for T_2. Alternatively, if the unknown sample S_2 is merely the remainder of a "sample" from which, in order to infer its contents, a subsample S_1 has been drawn, we may obtain the fiducial distribution of T. If T_2 or T is an unbiased estimate of θ, we obtain the fiducial distribution of θ by letting the size of sample S_2 tend to infinity.

No corresponding fiducial distribution for T_2 (or T) exists if these statistics are only quasi-sufficient, since the configuration of the second sample will be unknown. T_2 has not then the same claim to summarize the contents of sample S_2.

For similar inferences on both $\bar{x}$ and s^2 (or $\bar{x}_2$ and s_2^2) in normal theory, the relevant probability distribution will be

$$p(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2 | \bar{x}, s^2), \tag{4}$$

which is necessarily independent of m and σ^2. This distribution can be split up into two factors in three ways, corresponding to the association of s_1^2, s_2^2 or $(n_2 - 1)s_1^2 + (n_2 - 1)s_2^2$ with the t-factor. We have

$$p\left(\frac{\bar{x}_1 - \bar{x}_2}{s_1}\right) p\left(\frac{s_2^2}{s^2}\right) \tag{5}$$

$$= p\left(\frac{\bar{x}_1 - \bar{x}_2}{s_2}\right) p\left(\frac{s_1^2}{s^2}\right) \tag{6}$$

$$= p\left(\frac{\bar{x}_1 - \bar{x}_2}{s}\right) p\left(\frac{s_1^2}{s_2^2}\right). \tag{7}$$

Since (5) is equivalent to $fp(\bar{x}_2)fp(s_2^2|\bar{x}_2)$, and (7) to $fp(\bar{x}_2|s_2^2)fp(s_2^2)$, it is consistent to speak of the simultaneous distribution $fp(\bar{x}_2|s_2^2)$. But while (5) is also equivalent to $fp(\bar{x})fp(s^2|\bar{x})$, $fp(\bar{x}|s^2)$ is obtained from the first factor, and $fp(s^2)$ from the second factor, of (6), so that $fp(\bar{x}, s^2)$ also exists, but by virtue of a *different* factorization (cf. Fisher 1935).

For discontinuous variation, a relation (3) may similarly hold. While a fiducial distribution (in Fisher's sense) will no longer exist, the probability distribution $p(T_1, T_2|T)$ will still be available for inferences on T_2 or T. Thus if S_1 contains r_1 members out of n_1 with an attribute A, etc., we obtain

$$p(r_1, r_2|r) = p(r_1)p(r_2)/p(r)$$

$$= \frac{n_1!n_2!(n-r)!r!}{(n_1 - r_1)!(n_2 - r_2)!r_1!r_2!}, \tag{8}$$

which will determine the chance, say, of obtaining as few as r_1 members in S_1 when S contains r such members, or S_2 at least r_2 such members.*

4—The equivalence of a sufficient statistic (or, when relevant, the fiducial distribution derived from it) to the original data implies that when it exists it should lead to the most efficient test. It does not follow that a uniformly most powerful test, as defined by Neyman and Pearson (1933), will necessarily exist; but if the probability (or fiducial probability) distribution is known, the consequences of any procedure based on it will also be known.

The converse principle, that the existence of a uniformly more powerful test must depend on the necessary sufficiency properties being present, and is, moreover, only possible for the testing of a single unknown parameter, has been denied by Neyman and Pearson (1936*b*); but while agreeing that the examples they give are formal exceptions, I think it is worth while examining

* For approximate methods of using equation (8), see Bartlett (1937).

their examples further, since they could reasonably be regarded as bearing out the principle to which formally they are anomalous. It seems to me more valuable to recognize the generality of the principle that a test of a single parameter should be most sensitive to variations in it than to reject the principle because of apparent exceptions.

In example I of their paper, the distribution

$$p(x) = \beta e^{-\beta(x-\gamma)}, \quad (x \geqslant \gamma), \tag{9}$$

is considered. It is required to test whether $\gamma \leqslant \gamma_0$ and/or $\beta \geqslant \beta_0$. Since if any observation occurs that is less than γ_0 no statistical test is necessary, we are effectively concerned only with samples for which all observations are greater than γ_0. *For such observations*, the distribution law is

$$\begin{aligned} p(x) &= \beta e^{-\beta(x-\gamma)}/e^{-\beta(\gamma_0-\gamma)}, \quad (x \geqslant \gamma_0), \\ &= \beta e^{-\beta(x-\gamma_0)}, \end{aligned} \tag{10}$$

and is independent of γ. The sufficient statistic for β is $\bar{x}$, so that we are merely testing one independent parameter β for which a sufficient statistic $\bar{x}$ exists.

Example II is merely a special case of this with $\beta = 1 + \gamma^2$, ($\gamma_0 = 0$), and again $\bar{x}$ is the sufficient statistic for β and hence for γ, ($\gamma \leqslant 0$).

The above examples remind us, however, that a fiducial distribution is of more limited application than the sufficient statistic from which it is derived, and if restrictions are placed on the possible values of an unknown parameter, may become irrelevant. If the restriction is on an eliminated unknown, it might prove more profitable to use an inequality for a test of significance than an exact test. Thus if in normal theory it were known that $\sigma^2 \leqslant \sigma_0^2$, a test based on $p(\bar{x}|m, \sigma_0^2)$ might be more useful than one based on $p(t)$, though the possibility of using exact information on the range of other parameters in this way is in practice rare.

Conditional Variation and Exact Tests of Significance

5—By exact tests will be meant tests depending on a known probability distribution; that is, independent of irrelevant unknown parameters. It is assumed that no certain information is available on the range of these extra parameters, so that their complete elimination from our distributions is desirable.

In order for the variation in the sample S to be independent of irrelevant unknowns ϕ, a sufficient set of statistics U must exist for ϕ. All exact tests of significance which are to be independent of ϕ must be based on the calculable *conditional variation* of the sample $S|U$. We may in fact state as a general principle that *all exact tests of composite hypotheses are equivalent to tests of simple hypotheses for conditional samples.* For this principle to be general, conditional variation is understood to include theoretical conditional varia-

tion; for we have seen that in certain cases allied to problems in estimation, the set U may be functions of a primary unknown θ.

A useful illustration of the principle is given by the known exact test (Fisher 1934, p. 99) for the 2×2 contingency table (observed frequencies n_{11}, n_{12}, n_{21}, n_{22}). The sufficient statistics U for the unknown probabilities of independent attributes A and B are $n_{1.}|n$ and $n_{.1}|n$, where $n_{1.} = n_{11} + n_{12}$, etc. Hence any exact test of independence *must* be based on the variation $S|U$, which has one degree of freedom, and a distribution

$$p(S|U) = p(S)/p(n_{1.}|n)p(n_{.1}|n)$$
$$= \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n!}. \tag{11}$$

6—It is of some importance to consider the relation of tests dependent on this principle of conditional variation with those obtained by the likelihood criterion introduced by Neyman and Pearson. Suppose there is only one degree of freedom after elimination of irrelevant unknowns, as in quasi-sufficient solutions of estimation problems; and suppose further the relation exists,

$$p(T_1|\theta_1)p(T_2|\theta_2) = p(T_1, T_2|T)p(T|\theta), \tag{12}$$

when $\theta_1 = \theta_2 = \theta$. We have $p(T_1, T_2|T) \equiv p(T_1|T)$ and $T_1|T$ is the statistic for testing discrepancies between θ_1 and θ_2. By the likelihood criterion, however, the appropriate variate will be of the form

$$\lambda = \frac{p(T_1|T)p(T|\hat{\theta})}{p(T_1|\hat{\theta}_1)p(T_2|\hat{\theta}_2)},$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ from the distribution of T, etc., whence

$$\lambda = \frac{f(T_1|T)F(T)}{f_1(T_1)f_2(T_2)}, \tag{13}$$

say, which, for variation of S, will not necessarily be equivalent to $T_1|T$, or independent of θ. The condition for λ to provide the same test as $T_1|T$ appears to be that $T_1|T$ should be equivalent to a function $\psi(T_1, T_2)$ independent of T, and that $F(T)/f_1(T_1)f_2(T_2)$ should be an appropriate function $\phi(\psi)$. This holds when λ provides the proper test in normal theory, but it clearly must fail when only quasi-sufficiency (not convertible into pure sufficiency) properties exist.

A modification of the criterion when statistics U exist is proposed here. For a comprehensive test of "goodness of fit" involving all the remaining degrees of freedom of the sample, the test may be based directly on the *conditional likelihood* of $S|$. For the joint testing of definite unknowns, the conditional likelihood of the relevant statistics $T|$ would be considered. A criterion of this kind, if it differs from λ, is denoted subsequently by μ.

The mathematical definition of likelihood adopted is not separated from the more fundamental conception of the chance $p(S)$ of the observed data. For discontinuous data* the two are identical, so that the logarithm L is $\log p(S)$. For continuous variation it is sometimes convenient to drop the infinitesimal elements dx in $p(S)$, but some caution is necessary, this becoming more apparent when the likelihood of derived statistics is to be considered. Thus for s^2 (n degrees of freedom) in normal theory, $L(s^2)$ must, like $p(s^2)$, be invariant for a simultaneous change in scale in both s^2 and σ^2, and is defined to be

$$C + n \log s - n \log \sigma - \frac{n}{2}\left(\frac{s^2}{\sigma^2} - 1\right), \tag{14}$$

it being permissible to drop the term $d \log s^2$ (but not ds^2).

As an example we shall derive the μ appropriate for testing discrepancies among several variances. It is assumed that means and other regression parameters have already been eliminated, the statistic U being the pooled residual variance s^2, and the T the k individual variances s_r^2. Then

$$L(T) = L(T|U) + L(U),$$

or

$$L(T|U) = C' + \Sigma n_r \log s_r - n \log s.$$

For convenience $L(T|U)$ is measured from its maximum value C', so that

$$-2 \log \mu = n \log s^2 - \Sigma n_r \log s_r^2. \tag{15}$$

This criterion is not identical with that proposed by Neyman and Pearson, the weights being the number of degrees of freedom and not the number of observations. It is, however, put forward as an improvement, and a practicable method of using it derived below. With the original criterion λ it would be possible, if several regression parameters were eliminated from samples of unequal size, for fluctuations of a variance reduced to one or two degrees of freedom to mask real discrepancies in more stable samples; this effect is corrected when the weight for such a variance is reduced correspondingly.

If any likelihood with f degrees of freedom tends to a limiting normal form $C \exp\{-\frac{1}{2}A(x, x)\}$, then $-2 \log \lambda$ will tend to be distributed as χ^2 with f degrees of freedom. This, apart from its practical importance, is a useful reminder of the goodness of fit, or test of homogeneity, character of such tests, and should warn us against pooling together components which there is reason to separate.

To obtain more precisely the value of the χ^2 approximation in the present problem, consider first μ from (14) (that is, $k = 2$, $n_1 = n$, $n_2 = \infty$). From the known form (14) of the distribution of s^2, we readily obtain by integration the characteristic function of $-2 \log \mu$ for this case; the expected value

* Variation in which only discrete values of the variate are possible is specified for convenience by the term "discontinuous variation".

Table I.

	$n_1 = n_2$					$n_2 = \infty$			
n_1	1	2	3	6	12	2	4	9	∞
C	1.5	1.25	1.167	1.083	1.042	1.167	1.083	1.037	1
$P = 0.10$	2.48	2.66	2.69	2.70	2.70	2.68	2.70	2.70	2.706
$P = 0.02$	4.61	5.16	5.31	5.39	5.41	5.28	5.38	5.41	5.412

$$M \equiv E(\mu)^{-2t} = \frac{\Gamma\left\{\frac{n}{2}(1-2t)\right\}\left(\frac{n}{2}\right)^{nt} e^{-nt}}{(1-2t)^{n/2}\Gamma\left(\frac{n}{2}\right)}, \tag{16}$$

$$K \equiv \log M = t\left(1 + \frac{1}{3n} + \cdots\right) + \frac{t^2}{2!}\left(2 + \frac{4}{3n} + \cdots\right) + \frac{t^3}{3!}\left(8 + \frac{24}{3n} + \cdots\right) + \cdots$$

$$= t\left(1 + \frac{1}{3n}\right) + 2\frac{t^2}{2!}\left(1 + \frac{1}{3n}\right)^2 + 8\frac{t^3}{3!}\left(1 + \frac{1}{3n}\right)^3 + \cdots \tag{17}$$

approximately. If we call the exact function $K(n)$, we have for the general equation (15), owing to the equivalence of the statistics $s_r^2 | s^2$ to angle variables independent of s^2,

$$K = \Sigma K(n_r) - K(n), \tag{18}$$

or neglecting the effect of terms of $O\left(\frac{1}{n_r^2}\right)$, as in (17), we write

$$-\frac{2\log\mu}{C} = \chi^2 \tag{19}$$

with $k - 1$ degrees of freedom, where

$$C = 1 + \frac{1}{3(k-1)}\left\{\Sigma\frac{1}{n_r} - \frac{1}{n}\right\}. \tag{20}$$

The practical value of (19) was checked by means of the special case $k = 2$, representative values of (19) being given in Table I* for $P = 0.10$ and $P =$

* The values for $n_1 = n_2$ were obtained by means of Fisher's z table. When $n_2 = \infty$, the values for $n_1 = 2$ and 4 were obtained by an iterative method. It was subsequently noticed that the case $n_1 = 2$ could be checked from Table I of Neyman and Pearson's paper (1936*a*), from which the values for $n_1 = 9$ were added.

0.02, corresponding to the correct values of χ^2 (one degree of freedom) of 2.706 and 5.412 respectively. The use of C tends to over-correct in very small samples the otherwise exaggerated significance levels, but it greatly increases the value of the approximation, and serves also as a gauge of its closeness.

Continuous Variation—Normal Theory

7—For continuous variation, such as in normal theory, exact tests of significance have often been obtained owing to the readiness of the sample to factorize into independent statistics. Thus, for all inferences on the normality of a sample the relevant distribution

$$p(S|\bar{x}, s^2)$$

is expressible as a product of t distributions, and the usual statistics g_1 and g_2 (or β_1 and β_2) for testing for non-normality are independent of $\bar{x}$ and s^2.

The usual χ^2 goodness of fit test is not, but since the expected frequencies corresponding to $p(S|)$ would in any case only be used in a "large sample" test, which is an approximation, the alternative use of the estimated normal distribution, $m = \bar{x}$, $\sigma^2 = s^2$, will be legitimate. We may appeal to Fisher's proof that the distribution of χ^2 when m and σ^2 are efficiently estimated follows the well-known form with $f - 3$ degrees of freedom, where f is the number of cells.

But it is of theoretical interest to note that the *true* expected values for $S|$ could be found. Since the expected frequencies would then have three *fixed* linear conditions, for n, $\bar{x}$ and s^2, the number of degrees of freedom for χ^2 could, from this point of view, never have been questioned. $S|$ implies a sample point distributed over the $n - 2$-dimensional surface of a hypersphere of radius $s\sqrt{(n-1)}$ in the "plane" $\Sigma x = n\bar{x}$, but the *expected* distribution of the n variates is more simply expressed by the distribution (n times that for any one variate)

$$E(S|) = \frac{n\Gamma\left(\frac{n}{2}\right)}{\Gamma(\frac{1}{2})\Gamma\left(\frac{n-1}{2}\right)}\left\{1 - \frac{(x-\bar{x})^2}{(n-1)s^2}\right\}^{(1/2)(n-3)} \frac{dx}{s\sqrt{(n-1)}}. \qquad (21)$$

There is here a distinction between an exact test of goodness of fit for the normal law (which does not imply fitting a *normal* distribution at all), and the estimation of the normal law, which may be taken to be $m = \bar{x}$, $\sigma^2 = s^2$ (or $(n-1)s^2/n$).

Similarly for the exponential distribution with known origin 0 but unknown scale, the sufficient statistic for the scale is $\bar{x}$, the geometrical sample point is distributed at random over the "plane" $\Sigma x = n\bar{x}$, $(x > 0)$, and the expected distribution is

$$E(S|) = \frac{n}{\bar{x}}\left(1 - \frac{x}{n\bar{x}}\right)^{n-2} dx. \tag{22}$$

For the distribution $p = dx/\pi\{1 + (x - m)^2\}$, for which

$$p(S|m) = p(\bar{x}|C, m)p(C) \tag{23}$$

(where C is the configuration), the goodness of fit will be based on C and the estimation problem has entirely disappeared.

8—For two correlated variates x_1 and x_2, no function of the estimated correlation coefficients r and r' from two samples S and S' is a sufficient statistic for the true coefficient ρ. Hence no "best" test for a discrepancy in correlation coefficients is possible.

If, however, the degree of association between x_1 and x_2 were to be compared in the two samples on the basis of two populations of the same variability, an appropriate distribution is

$$p(v_{12}, v'_{12}|V_{11}, V_{12}, V_{22}), \tag{24}$$

where v_{12} is the sample covariance of S, V_{12} that of $S + S'$ (with elimination of both sample means), etc. This distribution, which is necessarily independent of σ_1^2, σ_2^2 and ρ, is thus a valid test of the difference between two covariances, although owing to the conditional nature of the distribution, the test would be rather an impracticable one even if the mathematical form of the distribution were known.

Discontinuous Variation—Poisson and Binomial Theory

9—For discontinuous variation, as for continuous variation which has been grouped, it is expedient for all but very small samples to be treated by approximate tests, but it is still important to consider the exact tests when they exist, not only for use with very small samples, but so that the basis of the approximate tests may be more clearly realized.

Consider first the Poisson distribution. For two samples with observed frequencies r_1 and r_2, the distribution of $(r_1, r_2|r)$, where $r = r_1 + r_2$, is simply a partition or configuration distribution giving the number of ways of dividing r between the two samples, and is

$$p(r_1, r_2|r) = \frac{r!}{2^r r_1! r_2!}, \tag{25}$$

or the terms of the binomial distribution $(\frac{1}{2} + \frac{1}{2})^r$. This will be the distribution for testing a discrepancy between the observed values of two Poisson samples.

For several samples, we have similarly a distribution depending on the multinomial distribution

$$\left(\frac{1}{n}+\frac{1}{n}+\cdots+\frac{1}{n}\right)^r. \tag{26}$$

The χ^2 test for dispersion is

$$\chi^2 = \Sigma(r_i - k_1)^2/k_1 = (n-1)k_2/k_1 \tag{27}$$

(where the usual semi-invariant notation is used, so that $k_1 = r/n$). Before an exact test is possible, a suitable criterion must be adopted. χ^2 and the μ criterion will no longer be completely equivalent in small samples, but since for distributions near to the Poisson the ratio k_2/k_1 may be defined as an index of dispersion, it seems in practice convenient still to consider χ^2, or equivalently (since k_1 is fixed and is the *true* expected value in (27)) the variance k_2.

The moments of $k_2|k_1$ are of some interest; they would be obtained from the identity

$$k_2 = \frac{1}{n-1}\{\Sigma r_i^2 - nk_1^2\}.$$

Thus, after some algebra with factorial moments, it is found that

$$\kappa_1(k_2|k_1) = k_1, \tag{28}$$

$$\kappa_2(k_2|k_1) = \frac{2k_1(nk_1-1)}{n(n-1)}. \tag{29}$$

These results may be compared with

$$\kappa_1(k_2) = m, \tag{30}$$

$$\kappa_2(k_2 - k_1) = \frac{2m^2}{(n-1)}, \tag{31}$$

and the approximate solution from (27),

$$\kappa_2(k_2|k_1) \sim \frac{2k_1^2}{n-1}. \tag{32}$$

When a large number of samples are available, a Poisson distribution may be fitted. Analogously to the goodness of fit tests for the normal and other continuous distributions, the *true* expected frequencies for observed values 0, 1, 2, etc.; could be found from the distribution of $S|$, but as a good enough approximation we fit the estimated Poisson distribution, $m = k_1$. The true expected distribution for $S|$ is the binomial

$$n\left(\frac{n-1}{n}+\frac{1}{n}\right)^r. \tag{33}$$

10—Similar, though somewhat more complicated, properties hold for the binomial distribution. For two samples, the exact distribution for inferring

the contents of a second sample was given by (8), and this distribution may similarly be used for comparing two known samples. The problem is a special case of a 2×2 contingency table where the marginal frequencies one way, corresponding to the sizes of the two samples, are absolutely fixed.

For l samples, we have similarly

$$p(S|) = \frac{(N-R)!R!}{N!} \prod_{i=1}^{l} \frac{n_i!}{(n_i - r_i)!r_i!} \tag{34}$$

(N and R referring to the total sample $S = \Sigma S_i$).

For l samples with more than a two-way classification of the contents, $R_1 \ldots R_m$ being the total observed numbers in the m groups,

$$p(S|) = \frac{R_1! \ldots R_m!}{N!} \prod_{i=1}^{l} \frac{n_i!}{r_{i1}! \ldots r_{im}!} \tag{35}$$

This corresponds also to an $l \times m$ contingency table.

For testing the dispersion among l binomial samples of equal size n, the usual test is

$$\chi^2 = \frac{N\Sigma(r_i - k_1)^2}{k_1(N - lk_1)} = \frac{N(l-1)k_2}{k_1(N - lk_1)}, \tag{36}$$

with $l - 1$ degrees of freedom. The exact distribution of $k_2|k_1$ could always be investigated, if necessary, from (34). The alternative use of the μ criterion is considered in section 12.

11—The moments of $k_2|k_1$ could also be found*, using factorial moments of r_i. For example,

$$\kappa_1(k_2|k_1) = \frac{nR(N-R)}{N(N-1)}, \tag{37}$$

where $k_1 = R/l$ (cf. equation (46), of which this is a special case).

It might be noticed that the factorial moment-generating function for (33) is either the coefficient of x^R in

$$\frac{(N-R)!R!}{N!}(1 + x + xt)^{n_1}(1 + x)^{n_2}, \tag{38}$$

or the coefficient of x^{n_1} in

$$\frac{n_1!n_2!}{N!}(1 + x + xt)^R(1 + x)^{N-R}. \tag{39}$$

The expression (39) is most readily generalized for classification of individuals into more than two *groups*, and becomes

* *Note added in proof*, 23 *March* 1937. Compare Cochran's treatment (1936). The exact value for the variance κ_2 appears somewhat complicated, but for large l it becomes approximately $2n^3R^2(N-R)^2/\{N^4(l-1)(n-1)\}$, which checks with Cochran's result.

$$\frac{n_1!n_2!}{N!}\prod_{j=1}^{m}(1+x+xt_j)^{R_j}, \tag{40}$$

while (38) is most easily generalized for the case of more than two *samples*, and becomes

$$\frac{(N-R)!R!}{N!}\prod_{i=1}^{l}(1+x+xt_i)^{n_i}. \tag{41}$$

The general case (35), corresponding to the $l \times m$ contingency table, is more complicated. Its generating function can be regarded as the coefficient of $x_1^{R_1}x_2^{R_2}\ldots x_m^{R_m}$ in

$$\frac{R_1!R_2!\ldots R_m!}{N!}\prod_{i=1}^{l}\left\{\prod_{j=1}^{m}x_j(1+t_{ij})\right\}^{n_i}. \tag{42}$$

For a large number of equal binomial samples, the expected distribution of $S|$ is the hypergeometric distribution

$$E(S|) = \frac{l\cdot n!(N-n)!R!(N-R)!}{(n-r)!r!(N-n-R+r)!(R-r)!N!}. \tag{43}$$

12—The 2×2 contingency table has been already mentioned. The exact solution for testing interactions in a $2 \times 2 \times 2$ table is also known (Bartlett 1935), but the immediate derivation of the probability of any partition, complete with constant, is no longer possible, owing to the complication which lack of independence introduces. Thus the number of ways of filling up a 2×2 table is the coefficient of $x^{n_{1\cdot}}y^{n_{\cdot 1}}$ in

$$(1+x+y+xy)^n = (1+x)^n(1+y)^n$$

or

$$\frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!n!},$$

but the number of ways of filling up a $2 \times 2 \times 2$ table *when testing the interaction* is the coefficient of

$$x_{11}^{n_{\cdot 11}}x_{12}^{n_{\cdot 12}}x_{21}^{n_{\cdot 21}}x_{22}^{n_{\cdot 22}}y_{11}^{n_{1\cdot 1}}y_{12}^{n_{1\cdot 2}}y_{21}^{n_{2\cdot 1}}y_{22}^{n_{2\cdot 2}}z_{11}^{n_{11\cdot}}z_{12}^{n_{12\cdot}}z_{21}^{n_{21\cdot}}z_{22}^{n_{22\cdot}},$$

in

$$\begin{aligned}(&x_{11}y_{11}z_{11}+x_{12}y_{12}z_{11}+x_{21}y_{11}z_{12}+x_{22}y_{12}z_{12}\\&+x_{11}y_{21}z_{21}+x_{21}y_{22}z_{21}+x_{21}y_{21}z_{22}+x_{22}y_{22}z_{22})^n,\end{aligned} \tag{44}$$

this last expression no longer factorizing.*

* The symbols x_{11}, x_{12}, x_{21} and x_{22} represent four parallel edges of a cube, the y's four other parallel edges, and the z's the remaining four. Each observed frequency n_{ijk}, $(i, j, k = 1, 2)$, corresponds to a corner of the cube, and hence to the three edges which intersect there. The sum of the frequencies at the end of every edge is fixed.

The expected value in the cell of a 2×2 table is the first factorial moment. For example,

$$E(n_{11}) = \frac{n_{1.}n_{.1}}{n}. \tag{45}$$

While this result is evident, it should be noted that the expected values in other χ^2 problems have not remained unaltered when S is modified to $S|$; χ^2 for a contingency table appears to have a slight theoretical advantage here when approximations to small sample theory are being considered.

Since the expected value corresponding to (34) must also be expressible as a rational fraction, the solution in terms of a cubic equation (Bartlett 1935) is an approximation, valid for large sample theory.

For the $l \times m$ table the second factorial moment for any cell is

$$\frac{n_{i.}(n_{i.} - 1)n_{.j}(n_{.j} - 1)}{n(n-1)},$$

whence the expected value of χ^2 itself is readily shown to be

$$E(\chi^2) = \frac{n}{n-1}(l-1)(m-1), \tag{46}$$

so that the bias of χ^2 is small and unimportant in comparison with the more general effects of discontinuity on its distribution (Yates 1934).

Since for an $l \times m$ contingency table to be tested for independence, $p(S|)$ is given by (35), the μ criterion is

$$\mu = \prod_{i=1}^{l} \prod_{j=1}^{m} \frac{n'_{ij}!}{n_{ij}!}, \tag{47}$$

where the n'_{ij} are the values of n_{ij} maximizing $p(S|)$. If this criterion were used, the values n'_{ij} must be found by inspection, though they will be near the expected values of n_{ij}. Equation (47) may be contrasted with the λ criterion given by Wilks (1935).

From (47) a small sample test is always possible. Thus for three binomial samples of 20 (equivalent to a 2×3 contingency table), with numbers

$$2 \quad 0 \quad 7$$

the exact significance level corresponding to μ (and also in this instance to χ^2) is found to be 0.007.

For large samples $-2 \log \mu$ will, like $-2 \log \lambda$, be distributed like χ^2, the three tests becoming equivalent. For medium to small samples, for which an exact test is too laborious, it is doubtful whether the usual χ^2 test can be bettered, for μ is not easier to compute, and its approximate distribution is only known in so far as μ tends to the form $\exp(-\frac{1}{2}\chi^2)$.

13—If a test of significance is to be independent of the particular *kind of population* from which the sample values were obtained, the whole set S of sample values must be regarded as fixed. This might seem to imply that noth-

ing is left to vary; but permutations of order are still possible. The relation of the sample values x with different groups or treatments, or with values of a second variate y, leads to tests for the significance of differences in means, or for the significance of apparent association. Thus in an experiment the assignment of treatments is at our choice; randomization ensures the validity of a test along these lines, this test tending to the usual test for a reasonable number of replications.

Summary

Properties of sufficiency must necessarily be considered for all small sample tests of significance, whether these are related to problems of estimation and fiducial distributions, or are of the nature of tests of goodness of fit.

The idea of "conditional variation" is developed, and its bearing on common tests, depending either on continuous or discontinuous variation, is shown. In particular, the use of χ^2 and other likelihood criteria is re-examined; and a new application of χ^2 proposed for testing the homogeneity of a set of variances.

References

Bartlett 1935 *J.R. Statist. Soc.* (Suppl.), **2**, 248.
——— 1936*a* *Proc. Roy. Soc.* A, **154**, 124.
——— 1936*b* *Proc. Camb. Phil. Soc.* **32**, 560.
——— 1937 *J.R. Statist. Soc.* (Suppl.), **4**.
Cochran 1936 *Ann. Eugen.* **7**, 207.
Fisher, R.A. 1934 "Statistical Methods for Research Workers," 5th ed.
——— 1935 *Ann. Eugen.* **6**, 391.
Neyman and Pearson 1933 *Philos. Trans.* A, **231**, 289.
——— 1936*a* *Statistical Research Memoirs*, **1**, 1.
——— 1936*b* *Statistical Research Memoirs*, **1**, 113.
Wilks 1935 *Ann. Math. Statist.* **6**, 190.
Yates 1934 *J.R. Statist. Soc.* (Suppl.), **1**, 217.