LARGE SAMPLE TESTS OF STATISTICAL HYPOTHESES CONCERNING SEVERAL PARAMETERS WITH APPLICATIONS TO PROBLEMS OF ESTIMATION

By C. RADHAKRISHNA RAO

Communicated by M. S. Bartlett

Received 18 April 1947

1. Introduction

If the probability differential of a set of stochastic variates contains k unknown parameters, the statistical hypotheses concerning them may be simple or composite. The hypothesis leading to a complete specification of the values of the k parameters is called a simple hypothesis, and the one leading to a collection of admissible sets a composite hypothesis. In this paper we shall be concerned with the testing of these two types of hypotheses on the basis of a large number of observations from any probability distribution satisfying some mild restrictions and their use in problems of estimation.

If we have a number of samples whose probability densities involve a set of parameters, we may have to test whether a single set is relevant to all the samples before combining them to arrive at the best estimates. This test, which may be called the test of homogeneity of parallel samples, involves a composite hypothesis. A general test of homogeneity different from the χ^2 test of independence of samples each arranged in some categories has been proposed and applied to test for agreement in gene frequencies between two samples giving the distribution in O, A, B and AB blood-group classes.

Another important group of problems is the estimation of parameters subject to restrictions which are sometimes derived from empirical considerations. The validity of these restrictions may be formally tested before giving final estimates. The use of such empirical relations among the parameters to be estimated, when known, enhances the precision of the estimates, although they may not be strictly accurate. A slightly inaccurate relationship may introduce bias in the estimates, but such estimates are more useful than the less efficient estimates so long as the bias, in any case, is small in comparison with its standard error. This, in some way, is secured when the test for a hypothesis specifying some restrictions indicates close agreement with the observations. It is also of importance to satisfy oneself that the increase in efficiency is of such a magnitude as to justify the use of some restrictions, although they may introduce errors which are smaller in comparison with standard errors of estimates. An instance to this point is the empirical formula $y_{12} = (y_1+y_2)/(1+4y_1y_2)$ suggested by Kosambi (1944) giving the relation connecting the recombination fractions y_1 and y_2 for two successive segments of a chromosome with y_{12} that for the combined segment. The use

of this has been found to enhance considerably the precision of the estimates of the recombination fractions.

Methods for determining the confidence regions in the case of several parameters have also been discussed in the light of the new tests proposed above.

2. The problem of distribution

There are two problems of distribution which are useful in deriving tests of significance for simple and composite hypotheses. Let

$$x_1, \ldots, x_p, \quad y_1, \ldots, y_q, \quad \ldots,$$

be independent sets of observations from probability laws with densities represented by $f_1(x \mid \theta), f_2(y \mid \theta), \ldots$, such that each function contains at least one of the unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. The likelihood of the parameters which is the same as the probability density at the observed sets of data is given by

$$L = f_1(x \mid \theta) \times f_2(y \mid \theta) \ldots.$$

We define, following Fisher (1935), the quantities

$$\phi_i = \frac{\partial \log L}{\partial \theta_i} \quad (i = 1, 2, \ldots, k)$$

as efficient scores. The mean values of these scores are zero. Their covariance matrix is represented by (α_{ij}) and its reciprocal by (α^{ij}). We shall assume that there exist positive quantities η such that

$$E\left(\frac{1}{f_i}\frac{\partial f_i}{\partial \theta_j}\right)^{2+\eta} \tag{2·1}$$

are finite. Under these conditions, if the non-vanishing terms in the sequence $\partial \log f_i/\partial \theta_j$ $(i = 1, 2, \ldots)$ for any j form a sufficiently large set, it follows from general limit theorems that the limiting distribution of $\phi_1, \ldots, \phi_k$ at the true values $\theta_1, \ldots, \theta_k$ tends to the multivariate normal form with zero mean and covariance matrix (α_{ij}). From this it follows that the statistic

$$\chi^2 = \Sigma\Sigma\alpha^{ij}\phi_i\phi_j$$

is distributed, in large samples, as χ^2 with k degrees of freedom when the true values of the parameters are $\theta_1, \theta_2, \ldots, \theta_k$.

In the case where the probability densities $f_1, f_2, \ldots$ are the same, it is enough for the limiting properties to hold that

$$E\left(\frac{1}{f}\frac{\partial f}{\partial \theta_j}\right)^2$$

is finite for every j, which is less restrictive than the condition (2·1). I am grateful to Mr Bartlett for drawing my attention to this.

Suppose that the θ's are subject to s restrictions defined by s independent relations

$$\psi_i(\theta_1, \ldots, \theta_k) = 0 \quad (i = 1, 2, \ldots, s). \tag{2·2}$$

The maximum likelihood estimates are given by

$$\left.\begin{aligned} \phi_i + \sum_j \lambda_j \frac{\partial \psi_j}{\partial \theta_i} = 0 \quad (i = 1, 2, \ldots, k) \\ \psi_i = 0 \quad (i = 1, 2, \ldots, s) \end{aligned}\right\} \tag{2·3}$$

where λ's are Lagrangian multipliers. Let $\dot{\theta}_1, \ldots, \dot{\theta}_k$ be the maximum likelihood

estimates. Since the set of equations (2·3) involves $(k-s)$ linear restrictions on $\phi_i(\dot{\theta})$, it is expected that the statistic

$$\chi^2 = \Sigma\Sigma\alpha^{ij}(\dot{\theta})\phi_i(\dot{\theta})\phi_j(\dot{\theta})$$

is distributed as χ^2 with s degrees of freedom which is $k-s$ less than the degrees of freedom for true values $\theta_1, \ldots, \theta_k$.

This can be demonstrated if we assume that the restrictions (2·2) specify s of the parameters which may be taken as $\theta_{k-s+1}, \ldots, \theta_k$ as functions of the $k-s$ free parameters $\theta_1, \ldots, \theta_{k-s}$, so that the likelihood is an explicit function of these parameters only, and further that the joint distribution of $\dot{\theta}_1, \ldots, \dot{\theta}_{k-s}$ tends to the multivariate normal form in large samples with variances and covariances of $O(n^{-1})$. It is known that the latter assumption is true provided the probability laws satisfy the condition (2·1), and further that the maximum likelihood estimates are *uniformly consistent* (Wald, 1943; Doob, 1934). I have omitted this latter condition by mistake in an earlier paper (Rao, 1947) in establishing some optimum properties of the maximum likelihood estimates. This does not seem to be a necessary condition, and the approach to normality is probably true under less stringent conditions.

Let us take the case of two parameters and one restriction which may be taken as $\theta_2 = w(\theta_1)$. The differential coefficient $d\theta_2/d\theta_1$ is denoted by $\lambda(\theta_1)$. The maximum likelihood estimates satisfy

$$\phi_1(\dot{\theta}) + \lambda(\dot{\theta}_1)\phi_2(\dot{\theta}) = 0, \quad \dot{\theta}_2 - w(\dot{\theta}_1) = 0. \tag{2·4}$$

If the given relation is true, then the statistic

$$\chi_0^2 = \Sigma\Sigma\alpha^{ij}(\theta)\phi_i(\theta)\phi_j(\theta) \tag{2·5}$$

depends only on θ_1, and is distributed as χ^2 with 2 degrees of freedom at the true value of θ_1. The expression (2·5) treated as a function of θ_1 may be expanded in the neighbourhood of $\dot{\theta}_1$. The first term is

$$\chi_1^2 = \Sigma\Sigma\alpha^{ij}(\dot{\theta})\phi_i(\dot{\theta})\phi_j(\dot{\theta}). \tag{2·6}$$

The second term is

$$2(\theta_1 - \dot{\theta}_1)[\phi_1(\theta)\{\alpha^{11}(\alpha_{11} + \lambda\alpha_{12}) + \alpha^{12}(\alpha_{12} + \lambda\alpha_{22})\} + \phi_2(\theta)\{\alpha^{22}(\alpha_{12} + \lambda\alpha_{22}) + \alpha^{12}(\alpha_{11} + \lambda\alpha_{12})\}]$$
$$= 2(\theta_1 - \dot{\theta}_1)[\phi_1(\dot{\theta}) + \lambda\phi_2(\dot{\theta})] = 0, \tag{2·7}$$

in virtue of (2·4). In the expression (2·7) terms of the order (n^0) only have been retained, $\partial\phi_i/\partial\theta_j$ being replaced by α_{ij} and terms of the type

$$\frac{\partial\alpha_{ij}}{\partial\theta_i}(\theta_1 - \dot{\theta}_1)\phi_i\phi_j$$

being omitted as they are of $O(n^{-\frac{1}{2}})$.

The third term can be easily shown to be

$$\chi_2^2 = (\theta_1 - \dot{\theta}_1)^2[\alpha_{11}(\dot{\theta}) + 2\lambda\alpha_{12}(\dot{\theta}) + \lambda^2\alpha_{22}(\dot{\theta})].$$

Neglecting terms of higher order of smallness we get

$$\chi_0^2 = \chi_1^2 + \chi_2^2.$$

Since
$$1/V(\theta_1 - \dot{\theta}_1) \sim \alpha_{11}(\dot{\theta}) + 2\lambda\alpha_{12}(\dot{\theta}) + \lambda^2\alpha_{22}(\dot{\theta}),$$

it follows that χ_2^2 is distributed in large samples as χ^2 with 1 degree of freedom.

It can be demonstrated by expanding $\phi_i(\dot{\theta})$ in powers of $(\theta_1 - \dot{\theta}_1)$ that $(\theta_1 - \dot{\theta}_1)$ and $\phi_i(\dot{\theta})$ tend to be uncorrelated in large samples, so that χ_1^2 and χ_2^2 are independently distributed in the limiting case.

Since χ_0^2 is distributed as χ^2 with 2 degrees of freedom and χ_2^2 with 1 degree of freedom, it follows that the residual part χ_1^2 is distributed as χ^2 with 1 degree of freedom.

In the case of s relations and k ($\geqslant s$) parameters χ_0^2 can be expressed as a function of $(k-s)$ parameters and split into two portions, one of which is a χ_2^2 with $(k-s)$ degrees of freedom measuring the discrepancy of the $(k-s)$ estimated parameters from their true values, and another a χ_1^2 with s degrees of freedom measuring the departures from the assigned relationships.

3. Derivation of statistics for simple and composite hypotheses

In the case of a single parameter, to test the simple hypothesis $\theta = \theta^0$, the statistic $\phi_1^2(\theta^0)/\alpha_{11}(\theta^0)$ is used as χ^2 with 1 degree of freedom. The quantity $\phi_1(\theta^0)$ has been called by Fisher (1935, 1946) the efficient score at the assigned value, and its use leads to elegant analysis in statistical tests. The optimum properties of this test have been discussed by Wald (1941) and Rao and Poti (1946).

In the multiparameter case let us consider the set of values $\theta_1^0 + h_1, \ldots, \theta_k^0 + h_k$, where $h_1, \ldots, h_k$ are small as alternatives to $\theta_1^0, \ldots, \theta_k^0$. The proportionate increase in the likelihood is given by

$$h_1\phi_1 + \ldots + h_k\phi_k. \tag{3·1}$$

The best test of the hypothesis in the sense that it affords the maximum discrimination when the alternatives differ from the assigned values by small quantities is provided by the statistic

$$w = h_1\phi_1 + \ldots + h_k\phi_k, \tag{3·2}$$

which leads to the use of the statistic

$$\chi^2 = w^2/\Sigma\Sigma h_i h_j \alpha_{ij}(\theta^0) \tag{3·3}$$

as χ^2 with 1 degree of freedom.

If the ratios of $h_1, \ldots, h_k$ can be assigned from *a priori* considerations, which is sometimes possible, the test can be carried out with exactitude. On the other hand, we may have to determine $h_1, \ldots, h_k$ from the departures of the assigned values $\theta_1^0, \ldots, \theta_k^0$ from those values indicated by the data and introduce suitable changes in judging the significance of the derived statistic. This may be done by finding the ratios of $h_1, \ldots, h_k$ such that χ^2 of (3·3) is maximum. The maximum value comes out as

$$\chi^2 = \Sigma\Sigma\alpha^{ij}(\theta^0)\,\phi_i(\theta^0)\,\phi_j(\theta^0). \tag{3·4}$$

In large samples this can be used, as shown in the previous section, as χ^2 with k degrees of freedom to test the hypothesis that the values of $\theta_1, \ldots, \theta_k$ are $\theta_1^0, \ldots, \theta_k^0$ respectively. This differs from the statistic proposed by Wald (1943), wherein he uses

$$\chi^2 = \Sigma\Sigma\alpha_{ij}(\dot{\theta})\,(\dot{\theta}_i - \theta_i^0)\,(\dot{\theta}_j - \theta_j^0)$$

to test the above hypothesis, where the $\dot{\theta}$'s are the maximum likelihood estimates. The test associated with (3·4) besides being simpler than Wald's has some theoretical advantages as shown in § 5.

A composite hypothesis specifies that the admissible sets of values lie on the intersections of surfaces

$$\psi_j(\theta_1, \ldots, \theta_k) = 0 \quad (j = 1, 2, \ldots). \tag{3·5}$$

If $s \leqslant k$ of these functions are independent the composite hypothesis is said to have $(k-s)$ degrees of freedom. Since a single set is responsible for the observed sample we may find their best estimates subject to the restrictions (3·5) and change the problem to that of testing a simple hypothesis whether these estimates agree with the data.

If the best estimates under the above restrictions are $\dot{\theta}_1, \ldots, \dot{\theta}_k$, the statistic

$$\chi^2 = \Sigma\Sigma\alpha^{ij}(\dot{\theta})\phi_i(\dot{\theta})\phi_j(\dot{\theta}) \tag{3·6}$$

can be used as χ^2 as shown in (2·4) with s degrees of freedom to test the composite hypothesis that the parameters satisfy s conditions. The $(k-s)$ degrees of freedom have been lost in constructing a suitable simple hypothesis from the composite hypothesis. As a general rule we may say that the degrees of freedom of χ^2 for testing a composite hypothesis is k, the number of parameters $-f$, the degrees of freedom of the hypothesis, which is the same as $(k-f)$ the number of restrictions they obey.

4. A GENERAL TEST OF HOMOGENEITY OF PARALLEL SAMPLES

The test of agreement of parallel samples, where each sample consists of observations arranged in mutually exclusive classes, can be treated as a test of independence in a contingency table if nothing is specified about the nature of the distribution in the various classes. Thus if we have r samples each arranged in p classes, the χ^2 test of independence has $(r-1)(p-1)$ degrees of freedom. If the distribution in the p classes can be specified by a probability law involving $k \leqslant (p-1)$ parameters, then the test of agreement in parallel samples is equivalent to a test of a composite hypothesis which specifies $k(r-1)$ relations among the rk parameters, and these are exactly the degrees of freedom of the χ^2 test of composite hypothesis. The disagreement in parallel samples is specified by $k(r-1)$ degrees of freedom, and a test for their significance need only be carried out. The exact expression for the χ^2 statistic is

$$\sum_{s=1}^{r} \sum_{i,j=1}^{k} \alpha_s^{ij}(\dot{\theta})\phi_i^s(\dot{\theta})\phi_j^s(\dot{\theta}), \tag{4·1}$$

where $\dot{\theta}$'s are obtained from the equations $\sum_s \phi_i^s(\theta) = 0$ $(i = 1, 2, \ldots, k)$ and (α_s^{ij}) is the matrix inverse to the information matrix for the sth sample and ϕ_i^s is the ith efficient score for the sth sample. This test is applicable in all cases whether the variables are continuous or discontinuous provided the sample size is large. The test is illustrated with an example given below, and the calculations are similar in any analogous situation.

The distributions in the four O, A, B and AB blood-group classes of 140 Christians who are army cadets and 295 other Christians are given in Table 1. The problem is to test whether the two samples agree in the gene frequencies.

Table 1. *Blood-group frequencies in two samples of Christians (Indian)*

	O	A	B	AB	Total
Army cadets	56	60	18	6	140
Other Christians	120	122	42	11	295
Total	176	182	60	17	435

If p, q, r are the A, B, O gene frequencies, then the probabilities and their derivatives are:

	Probabilities and derivatives		
	π	$\frac{\partial \pi}{\partial p}$	$\frac{\partial \pi}{\partial q}$
O	r^2	$-2r$	$-2r$
A	$p(p+2r)$	$2r$	$-2p$
B	$q(q+2r)$	$-2q$	$2r$
AB	$2pq$	$2q$	$2p$

On the given hypothesis the maximum likelihood values are to be obtained from the combined sample. Fairly approximate solutions as obtained from Bernstein's (1925) method are

$$p = 0{\cdot}26449, \quad q = 0{\cdot}09317, \quad r = 0{\cdot}64234.$$

The probabilities and coefficients for the calculation of efficient scores are:

	Probability π	Coefficients for scores	
		$\frac{1}{\pi}\frac{\partial \pi}{\partial p}$	$\frac{1}{\pi}\frac{\partial \pi}{\partial q}$
O	0·41260	−3·11362	−3·11362
A	0·40974	3·13543	−1·27104
B	0·12838	−1·45217	10·00685
AB	0·04928	3·75086	10·73307

The information matrix for a single observation is

$$I_{pp} = 9{\cdot}00315, \quad I_{pq} = 2{\cdot}47676,$$
$$I_{pq} = 2{\cdot}47676, \quad I_{qq} = 23{\cdot}21612.$$

The elements of the inverse matrix are

$$I^{pp} = 0{\cdot}114430, \quad I^{pq} = -0{\cdot}012208,$$
$$I^{pq} = -0{\cdot}012208, \quad I^{qq} = 0{\cdot}044376.$$

The efficient scores for each sample are obtained by multiplying the observed frequencies with the coefficients for scores given above and adding up over all the classes:

	ϕ_p	ϕ_q
Sample 1	10·30918	−7·30340
Sample 2	−10·51362	7·21019
Total = Φ	− 0·20444	−0·09321

The small additive corrections to the approximate values p and q are given by

$$dp = \frac{(I^{pp}\Phi_p + I^{pq}\Phi_q)}{N} = -0{\cdot}0000{,}5116,$$

$$dq = \frac{(I^{pq}\Phi_p + I^{qq}\Phi_q)}{N} = -0{\cdot}0000{,}0377.$$

The efficient scores and informations matrix at these values are needed for the test. They can be obtained by slight adjustments if the approximations are good to start with. The changes in the elements of the information matrix are negligible. The χ^2 is 0·17258, which is small for 2 degrees of freedom, thus indicating close agreement.

Table 2. *Adjusted efficient scores and* χ^2

Sample	n	$\phi_1' = \phi_1 - \frac{n}{N}\Phi_1$	$\phi_2' = \phi_2 - \frac{n}{N}\Phi_2$	$\chi^2 = \frac{1}{n}\Sigma\Sigma I^{ij}\phi_i'\phi_j'$
1	140	10·37497	−7·27341	0·11704
2	295	−10·37497	7·27341	0·05554
Total	435	0	0	0·17258
				2 D.F.

We can in such cases give the best estimates of gene frequencies as derived from the combined sample:

$$\dot{p} = 0{\cdot}26444, \quad V(\dot{p}) = \frac{I^{pp}}{N} = 0{\cdot}00026305,$$

$$\dot{q} = 0{\cdot}09317, \quad V(\dot{q}) = \frac{I^{qq}}{N} = 0{\cdot}00010202,$$

$$\dot{r} = 0{\cdot}64239, \quad V(\dot{r}) = \frac{I^{pp} + 2I^{pq} + I^{qq}}{N} = 0{\cdot}00030893.$$

The general formula for χ^2 in the case of two samples can be written as

$$\chi^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\Sigma I^{ij}\phi_i\phi_j,$$

where I^{ij} are elements of the matrix inverse to the information matrix for a single observation and n_1 and n_2 are sample sizes, and ϕ's are efficient scores at the estimated values for one of the samples. The degrees of freedom in this case are equal to the number of parameters under consideration.

5. CONFIDENCE REGIONS AND INTERVALS

It has been shown in § 2 that the statistic

$$\chi^2 = \Sigma\Sigma\alpha^{ij}(\theta)\phi_i(\theta)\phi_j(\theta)$$

considered as a function of the observations and the unknown parameters $\theta_1, \ldots, \theta_k$ is distributed, in the limit, independently of the parameters. When such pivotal quantities as defined by Fisher (1945) exist it is possible to divide the set of parameters into two groups S_1 and S_2 such that any hypothesis assigning a set of parameters belonging to only one of the groups S_1 (say) is rejected by the observed data on a desired probability level α %. The groups S_1 and S_2 are defined by the inequalities

$$\Sigma\Sigma\alpha^{ij}(\theta)\phi_i(\theta)\phi_j(\theta) \geq \alpha\ \%\ \text{value of } \chi^2 \text{ with } k \text{ degrees of freedom},$$
$$< \alpha\ \%\ \text{value of } \chi^2 \text{ with } k \text{ degrees of freedom},$$

respectively. The region defined by the group S_2 in a space of k dimensions in which the sets of parameters may be represented, is called the confidence region. The regions so constructed from the observations satisfy the property that in repeated samples they exclude the true set of parameters only α % of times. Some optimum properties of these regions are mentioned in an abstract of a paper by Wilks (1939).

The confidence region constructed above is useful only when all the parameters are considered simultaneously. If the confidence interval for a single parameter (say) θ_1 irrespective of the others is required then the following procedure is necessary. If θ_1 is considered known the maximum likelihood estimates $\dot{\theta}_2, \ldots, \dot{\theta}_k$ can be determined as functions of θ_1 and the observations. This amounts to estimating the parameters with the restriction that the value of θ_1 is given. Under such circumstances it has been shown in § 2 that the statistic

$$\chi^2 = \Sigma\Sigma\alpha^{ij}(\theta_1, \dot{\theta})\,\phi_i(\theta_1, \dot{\theta})\,\phi_j(\theta_1, \dot{\theta})$$

is distributed as χ^2 with $k-(k-1)$ degrees of freedom. The α % confidence interval for θ_1 is defined by the inequality

$$\Sigma\Sigma\alpha^{ij}(\theta_1, \dot{\theta})\,\phi_i(\theta_1, \dot{\theta})\,\phi_j(\theta_1, \dot{\theta}) < \alpha\,\%\text{ value of } \chi^2 \text{ with 1 degree of freedom.}$$

Similarly confidence regions for any subset of s parameters can be determined. The χ^2 to be used in this case has $k-(k-s)$ degrees of freedom.

REFERENCES

BERNSTEIN, F. Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Z. indukt. Abstamm.- u. VererbLehre*, 37 (1925), 237–70.

DOOB, J. L. Probability and statistics. *Trans. Amer. Math. Soc.* 36 (1934), 759–75.

FISHER, R. A. The detection of linkage with dominant abnormalities. *Ann. Eugen., London*, 6 (1935), 187–201.

FISHER, R. A. The logical inversion of the notion of the random variable. *Sankhyā*, 7 (1945), 130–3.

FISHER, R. A. A system of scoring linkage data with special reference to pied factors in mice. *Amer. Nat.* 80 (1946), 568–78.

KOSAMBI, D. D. The estimation of map distance from recombination values. *Ann. Eugen. London*, 12 (1944), 172–6.

RAO, C. R. and POTI, S. J. On locally most powerful tests when alternatives are one sided. *Sankhyā*, 7 (1946), 439.

RAO, C. R. Minimum variance and the estimation of several parameters. *Proc. Cambridge Phil. Soc.* 43 (1947), 280–3.

WALD, A. Some examples of asymptotically most powerful tests. *Ann. Math. Stat.* 12 (1941), 396–408.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54 (1943), 426–82.

WILKS, S. S. Optimum fiducial regions for simultaneous estimation of several population parameters from large samples. *Ann. Math. Statist.* 10 (1939), 85.

KING'S COLLEGE
CAMBRIDGE