284

SAMPLING THE REFERENCE SET

By SIR RONALD A. FISHER

University of Adelaide

1. THE REFERENCE SET OF A PROBABILITY STATEMENT

Every test of significance involves probability statements, conditioned on the truth of the hypothesis. Conceptually, these can be verified by sampling the reference set which is their mathematical basis. The misapprehension has, however, been widely promulgated that such sampling involves no more than a repetition of the process by which the data to be tested came into existence. Easy examples, (Fisher, 1956-59) such as the test of significance of a linear regression, or the test of proportionality in a two by two table, have frequently been cited to show that such a method of "repeated sampling from the same population" is erroneous, and irrelevant to the test of significance which it is proposed to verify. The recognition of the appropriate reference set is an essential first step to understanding a test of significance, and, therefore, to setting up an appropriate process of possible verification. We may exemplify such a process using the long-disputed test of significance for the difference between the means of two hypothetical populations, the variances of which are not in any known ratio, when a random sample of each is available.

2. INFERENCES FROM TWO NORMAL SAMPLES

The means of the two samples will be represented by $\bar{x}_1$ and $\bar{x}_2$, where

$$(n_1+1)\bar{x}_1 = S(x_1)$$
$$(n_2+1)\bar{x}_2 = S(x_2)$$

the estimated variances of the means by s_1^2 and s_2^2, where

$$n_1(n_1+1)s_1^2 = S(x_1-\bar{x}_1)^2 = S_1,$$
$$n_2(n_2+1)s_2^2 = S(x_2-\bar{x}_2)^2 = S_2,$$

these estimates being based on n_1 and n_2 degrees of freedom respectively.

The test of significance must involve as parameters the known values n_1, n_2 and s_1/s_2; these are, therefore, known characteristics of the reference set. If a more comprehensive set were first considered, the facts that s_1/s_2 is known to the observer, and is not irrelevant to the probability statements to be made, and that the set in which s_1/s_2 is constant contains no further relevant sub-set, may be taken as defining the set which is to be sampled. If we are to erect a sampling process capable of testing any one of the values tabulated by Behrens or Sukhatme, or others, for Behrens' test, the first step is to obtain random samples having the correct values for this ratio.

3. Verification of Behrens' test

Now if σ_1^2 and σ_2^2 are the true variances of the populations sampled, the distribution of

$$2z = \log (n_1+1)s_1^2\sigma_2^2/(n_2+1)\, s_2^2\sigma_1^2$$

is known in terms of n_1 and n_2. Let z_P stand for that value for which

$$\Pr (z < z_P) = P,$$

and let us take for P a representative series of fractions, such as

$$P_i = (2i-1)/20{,}000$$

where i takes integral values from 1 to 10,000.

For each of these 10,000 values

$$\sigma_2^2/\sigma_1^2 = (n_2+1)s_2^2\, e^{2z}/(n_1+1)s_1^2,$$

and is, therefore, known.

We may now take two Normal populations having equal means, and variances in the ratio required.

Successive samples of the same sizes as those constituting the data may now be taken, but all will be rejected in which the experimental ratio of s_1 to s_2 does not agree, within a specified tolerance, with that originally observed. For each value of i the first which satisfies this condition may be taken as a random representative of the appropriate reference set.

If d_p is the tabular value to be verified, such random pairs of samples may be classified as satisfying one of the three inequalities,

$$(\bar{x}_1-\bar{x}_2) < -d_p \sqrt{s_1^2+s_2^2},$$

$$-d_p \sqrt{s_1^2+s_2^2} < (\bar{x}_1-\bar{x}_2) < d_p \sqrt{s_1^2+s_2^2},$$

$$d_p \sqrt{s_1^2+s_2^2} < (\bar{x}_1-\bar{x}_2),$$

in which the values inserted for $\bar{x}_1$, $\bar{x}_2$, s_1^2 and s_2^2 are derived from the experimental sampling. For tabular values of the 5% point, the expected numbers in these three classes are

$$250,\ 9500,\ 250,$$

while, for values tabulated at the 1% point, the expectations are

$$50,\ 9900,\ 50$$

respectively.

As in other cases the verification need not be carried out; it is sufficient that it is a precisely defined procedure which could be carried out with any degree of precision, and with calculable consequences.

4. The weighted mean

The method set out above of generating a sample of the only possible reference set appropriate to the data may be used to illustrate the solution of some other related problems. For these we should suppose that a sample of a million pairs of samples have been obtained having the correct ratio s_1/s_2. These pairs will differ from each other in the relative difference $d \left(= (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2 + s_2^2} \right)$ observed, but, as in the case of the ratio s_1/s_2, a selection can be made agreeing sufficiently with the original observations in this particular. This doubly screened sample will still show variation in the position of the true common mean relative to the two means observed, and will supply the appropriate verification of the distribution of the true mean, which as L. C. Payne has observed, depends on the four parameters $n_1, n_2, s_1/s_2$ and d and would, therefore, be exceedingly troublesome to tabulate, though capable of an effective asymptotic representation.

For this, from the primary equations

$$\mu - \bar{x}_1 = s_1 t_1,$$

$$\mu - \bar{x}_2 = s_2 t_2,$$

we may derive

$$\mu - \bar{x} = Su,$$

where

$$\bar{x} = \frac{s_2^2 \bar{x}_1 + s_1^2 \bar{x}_2}{s_1^2 + s_2^2}, \quad S = \frac{s_1 s_2}{\sqrt{s_1^2 + s_2^2}}$$

so that

$$\frac{1}{S^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2},$$

and u_P is defined by the relation

$$\Pr(u < u_P) = P.$$

Then if

$$P = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}v^2} dv,$$

u_P may be evaluated by an expansion

$$u_P = x + \sum_{r,s} u_{rs}\, n_1^{-r}\, n_2^{-s},$$

where u_{rs} is a polynomial in x, d and $\cos\theta$, $\sin\theta$, such as

$$\frac{u_{10}}{n_1} = \frac{1}{4n_1}\{(x^3+x)c^4 + 4d(x^2+1)c^3s + (6d^2-2)xc^2s^2 + 4(d^3-d)cs^3\}$$

$$\frac{u_{01}}{n_2} = \frac{1}{4n_2}\{-4(d^3-d)c^3s + (6d^2-2)xc^2s^2 - 4d(x^2+1)cs^3 + (x^3+x)s^4\}$$

in which c and s stand for the cosine and sine of θ, and u_{rs} may be obtained from its conjugate u_{sr} by interchanging c and s and reversing the sign of d.

The expansion is similar to, though owing to the additional parameter, somewhat more complex than the expansion for d (Table 2 of my 1941 paper). Unlike the distribution of d, that of the weighted mean is not symmetrical.

5. The unknown variances

A third problem has been regarded as quite insoluble, namely to determine the simultaneous distribution of σ_1 and σ_2, when, with such data, it is also given that the true means are equal. It may be shown without difficulty that, while the three statistics, s_1, s_2, d are jointly exhaustive, yet no two functions of these three provide simultaneous exhaustive estimates of σ_1 and σ_2. Nor has any function of these quantities a sampling distribution independent simultaneously of σ_1 and σ_2, as would be the case of an Ancillary Statistic.

Now if σ_1' and σ_2' stand for the standard deviations chosen for the two populations to be sampled experimentally, and s_1' and s_2' for the estimates of the standard deviations obtained for the means, because these have been selected so that,

$$\frac{s_1'}{s_1} = \frac{s_2'}{s_2}$$

where s_1 and s_2 are the estimates from the original data, we may obtain σ_1 and σ_2 so that

$$\frac{\sigma_1'}{\sigma_1} \text{ and } \frac{\sigma_2'}{\sigma_2}$$

are also in the same ratio.

Within the selection that has been made to give the correct values of the statistics s_1, s_2 and d, we now have a distribution of pairs of values of σ_1 and σ_2, so that a statement of the form

$$\Pr\,(\sigma_1 < \alpha,\ \sigma_2 < \beta) = F(\alpha, \beta, s_1, s_2, d)$$

can be established.

The function on the right can be expressed as the ratio

$$\int_0^{\alpha} d\sigma_1 \int_0^{\beta} d\sigma_2 \,.\, W \div \int_0^{\infty} d\sigma_1 \int_0^{\infty} d\sigma_2 \,.\, W,$$

where W takes the form

$$* \quad \frac{1}{\sigma_1^{n_1+1}\,\sigma_2^{n_2+1}\sqrt{\sigma_1^2+\sigma_2^2}}\ \exp -\tfrac{1}{2}\left\{ \frac{S_1}{\sigma_1^2}+\frac{S_2}{\sigma_2^2}+\frac{(\bar{x}_1-\bar{x}_2)^2}{\sigma_1^2+\sigma_2^2} \right\}.$$

The three statistics in the index of the exponential term are jointly exhaustive. Hence there is no other observable which could be used to define a recognizable and relevant sub-set. Without the third statistic the observations would merely justify the distribution appropriate to the equivalence,

$$\frac{S_1}{\sigma_1^2} = \chi^2_{n_1},\ \frac{S_2}{\sigma_2^2} = \chi^2_{n_2},$$

for two independent random variables χ^2.

The additional observation

$$\bar{x}_1 - \bar{x}_2$$

introduces the factor

$$\frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}}\ \exp -\tfrac{1}{2}\left\{ \frac{(\bar{x}_1-\bar{x}_2)^2}{\sigma_1^2+\sigma_2^2} \right\}.$$

For variation of $\sigma_1^2+\sigma_2^2$ this is stationary when

$$\frac{-1}{\sigma_1^2+\sigma_2^2}+\frac{(\bar{x}_1-\bar{x}_2)^2}{(\sigma_1^2+\sigma_2^2)^2} = 0\,.$$

The factor increases with $\sigma_1^2+\sigma_2^2$ up to a maximum when

$$\sigma_1^2+\sigma_2^2 = (\bar{x}_1-\bar{x}_2)^2,$$

thereafter decreasing.

* Here and in what follows, replace $\sigma_1^2 + \sigma_2^2$ by $\frac{\sigma_1^2}{n_1+1}+\frac{\sigma_2^2}{n_2+1}$ - R.A.F.

The example is interesting as showing that the possibility of making simultaneous probability statements about parameters is not limited to cases of simultaneous exhaustive estimation, in the strict sense, though a jointly exhaustive set of statistics is still utilized.

References

FISHER, R. A. (1935): The fiducial argument in statistical inference. *Ann. Eug.*, **6**, 391-398.

——— (1941): The asymptotic approach to Behrens' integral with further tables for the *d* test of significance. *Ann. Eug.*, **2**, 142-172.

——— (1956-59): *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Paper received : August, 1960.