

Note on the χ^2 smooth testBy D. A. S. FRASER, *University of Toronto*

1. In his recent contribution to *Biometrika*, Seal (1948) presents a theorem to be used in generalizing the χ^2 test of a theoretical frequency curve or of a graduation of a mortality table. Although incorrectly stated in his paper, the theorem produces a rather striking result for the normal distribution and is then applied as an approximation to the multinomial distribution or to the joint distribution of a series of binomial distributions.

The proof indicated is, however, applicable to the following corrected statement:

THEOREM. If $x_i (i = 1, 2, \dots, n)$ are the residuals of n independent normal variates, with means zero and variances one, after removing the regression on k linearly independent vectors, then the probability distribution of $q^2 = \sum x_i^2$ and the distribution of the signs of the residuals are independent. (The x_i are the residuals of n independent normal variates with means zero and variances one, after fitting k independent homogeneous linear constraints by regression.)

The proof follows by showing that the conditional distribution of the signs, given q^2 , is independent of q^2 . Because the original n normal variates were independent and had unit variances, the distribution in n -space is spherically symmetrical and any orthogonal rotation of the n -space will give independent variates. Hence the probability density function of the residuals is the conditional distribution of the original normal variates in the linear subspace determined by the constraints, and has the following form:

$$c \exp \left[-\frac{1}{2} \sum x_i^2 \right],$$

where c is a constant depending on the number of constraints and the x_i are connected by the k constraints. For each value of q the density is a constant. To find the probability of any particular sign pattern is then to find the proportion of the surface of a hypersphere (in the subspace) which is contained in a particular 'quadrant' of the n -dimensional space. Since the sphere has centre the origin, this is independent of q , which is the radius of the hypersphere.

In the statement of the theorem in §1, the *residuals* were to have means zero and variances one. It is interesting to note that frequently this situation is impossible, depending on the selection of constraints. An example will illustrate this: Let y_1, y_2 be independent, normal, and with means zero and variances one. Let x_1, x_2 be the residuals after fitting the constraint $y_1 = 0$. Then the variance of x_1 is zero. No other selection of a joint distribution for y_1 and y_2 would overcome this.

2. Seal applies this theorem to testing the graduation of a mortality table where more than one constraint is applied. Although there is independence between the usual χ^2 and the sign patterns, the basis for the suggested sign tests no longer exists. When the single constraint $\sum x_i = 0$ is applied, all sign combinations are equally likely except two—all positive and all negative, which cannot occur. With additional linear constraints, the sign patterns will no longer be equally likely, *unless* the constraints are of the form $\sum \pm x_i = 0$ which is not the case for the graduation of the mortality table. It is worth noting that for each additional constraint of the form $\sum \pm x_i = 0$, two sign patterns become impossible, but the remaining sign patterns have equal probability.

3. In her paper in *Biometrika*, David (1948) discusses the effect of applying a sign test to a mortality graduation using the moment type of constraint without taking account of the sign patterns which no longer have equal probability. The conclusion is that, as the number of moment constraints increases, the correlation between adjacent deviations approaches -1 and consequently the sign test as proposed is of doubtful validity.

At the end of §2 of her paper, the relation

$$\theta_{ij} = \cos^{-1} r_{ij}$$

is presented without proof. The proof follows immediately by noting that the rearrangement of the quadratic exponent implies that z_1, \dots, z_n are independent with equal variance, and hence the expression for the correlation between two linear combinations of the z_i will be identical with that for the cosine of the angle between the two planes determined by the linear combinations set equal to zero.

4. Using the theorem in §1, a sign test can be constructed by calculating the probability for each sign pattern—or for enough 'extreme' patterns to form a rejection region of the proper size. This would involve integration to find the proportion of the surface of a hypersphere (in the subspace satisfying the constraints) which is contained in the 'quadrants' of n -space corresponding to the appropriate sign patterns. The work necessary to calculate these probabilities would be prohibitive unless n is small.

If it is desirable to combine the usual χ^2 test with a test based on signs, the sign patterns could be ordered by reference to a 'reasonable' criterion or, perhaps better, by reference to a representative

alternative hypothesis. From the data, the probability of a more extreme pattern could be calculated, transformed to a χ^2 with p degrees of freedom, and combined with the χ^2 on $n-k$ degrees of freedom to form a combined χ^2 on $n-k+p$ degrees of freedom. The choice of p will determine the relative weight to be given to the two tests. It should be made by considering alternative hypotheses, but this seemingly would be a prohibitive task and might better be left to the discretion of the statistician in weighting the tests before observing the data. Seal's choice of $p = 1$ (1948) would appear to the writer to underestimate greatly the sign test except perhaps when $n-k$ is small.

The author wishes to express his appreciation of valuable discussion with Dr Seal.

REFERENCES

- DAVID, F. N. (1948). Correlations between χ^2 cells. *Biometrika*, **35**, 418.
 SEAL, H. L. (1948). A note on the χ^2 smooth test. *Biometrika*, **35**, 202.

An alternative form of χ^2

By F. N. DAVID

1. We consider a population, Π , divided into k groups or strata. The proportion in the i th stratum is p_i ($i = 1, 2, \dots, k$). A sample of N observations is randomly drawn from Π , the number coming from the i th stratum being n_i ($i = 1, 2, \dots, k$). If the functional form of the population is fully specified, the proportions p_i are known, and the quantity commonly calculated in order to test whether the sample is representative of the population Π is

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - Np_i)^2}{Np_i} = \sum_{i=1}^k \frac{n_i^2}{Np_i} - N.$$

This quantity has as its first two moments

$$\begin{aligned} \mathcal{E}(\chi^2) &= k - 1, \\ \sigma_{\chi^2}^2 &= 2(k-1) \left(1 - \frac{1}{N}\right) + \frac{1}{N} \sum_{i=1}^k \frac{1}{p_i} - \frac{k^2}{N}, \end{aligned}$$

and provided N is large, it is known that it may be assumed to be distributed as a Pearson Type III variable with range zero to plus infinity. Recently, and in quite another connexion, Neyman (1949, p. 239) has considered the quantity

$$\chi_1^2 = \sum_{i=1}^k \frac{(n_i - Np_i)^2}{n_i} = \sum_{i=1}^k \frac{N^2 p_i^2}{n_i} - N.$$

This quantity has certain advantages from the computational point of view, and it appears of interest therefore to obtain some idea of how closely the distribution of χ_1^2 approximates to the limiting distribution of χ^2 .

2. We take the expression

$$\chi_1^2 + N = \sum_{i=1}^k \frac{N^2 p_i^2}{n_i} = \sum_{i=1}^k \frac{Np_i}{1 + \frac{n_i - Np_i}{Np_i}},$$

and expand the denominator. This will be legitimate only for

$$|n_i - Np_i| < Np_i,$$

but we may assume that N is sufficiently large for this to be true in all but a negligible proportion of cases. Thus

$$\chi_1^2 = \sum_{i=1}^k \frac{(n_i - Np_i)^2}{Np_i} - \sum_{i=1}^k \frac{(n_i - Np_i)^3}{N^2 p_i^2} + \sum_{i=1}^k \frac{(n_i - Np_i)^4}{N^3 p_i^3} - \dots,$$

a series which is in increasing powers of $1/N$. On taking expectations, we have to order $1/N^2$,

$$\mathcal{E}(\chi_1^2) = k - 1 + \frac{1}{N} \left[\sum_{i=1}^k \frac{2}{p_i} - 3k + 1 \right] + \frac{1}{N^2} \left[\sum_{i=1}^k \frac{6}{p_i^2} - \sum_{i=1}^k \frac{12}{p_i} + 7k - 1 \right].$$