MISCELLANEOUS

SAMPLING EXPERIMENTS ON THE COMBINATION OF INDEPENDENT χ^2-TESTS

By NIKHILESH BHATTACHARYA
Indian Statistical Institute

SUMMARY. The present communication reports on some model sampling work on the relative powers of three methods of combining independent χ^2-tests. One is based on the straightforward addition of χ^2's and of the corresponding degrees of freedom; the second is the P_{λ_n} technique using upper tail probabilities associated with the χ^2's; and the third is Tippett's test (Birnbaum, 1954) based on the minimum upper tail probability. Although the experiments are in no way exhaustive, they indicate that the first two methods are almost equally powerful, (which has interesting implications,) and that the third method is usually inferior to the other two.

1. Outline of the experiment

1.1. Let k denote the d.f. and λ the parameter of non-centrality for a non-central χ^2. For each of a number of combinations of k and λ, one series of independent non-central χ^2's was built up by using Wold's table of normal deviates (1948). The three methods of combination were then applied* to mutually exclusive sets of n non-central χ^2's of each series, n being, in turn, 2, 6, 12 or 24. Table 1 summarises the results.

1.2. Another type of experiments was carried out for χ^2's with single d.f. Two series of single d.f. χ^2's were taken, the two differing in respect of λ, and a 'mixed' series was built up by picking up alternate elements from the two 'pure' series. The three methods of combination were then applied to mutually exclusive sets of n non-central χ^2's of the 'mixed' series, where $n = 2$, 6, 12 or 24. Results for such 'mixed' series are shown in Table 2.

1.3. Power figures given in the tables are all 'estimates' based on the model sampling work, although 'true' values were also calculated for the first and the third methods using Patnaik's approximate rules (1949). These 'true' values agreed with the corresponding estimates to within limits of sampling error. The 'estimates' are, however, presented here instead of 'true' values, in the interest of making the power comparisons more sensitive.

1.4. To save space, estimates of power are given for two particular levels of significance. Results for the other levels were very similar. Also, lines for $n = 12$ or 24 are omitted in case the number of experiments fall below 50.

2. Results

2.1. As regards the relative powers of $\Sigma\chi^2$ and P_{λ_n} tests, Table 1 shows that these are almost equally efficient when the χ^2's combined have the same k and λ, where $k = 1$, 6, 12 or 24, and λ assumes the common range of values. Table 2 indicates that some variation in λ does not alter the situation if $k = 1$. For $k = 2$, it may be recalled, the two methods are strictly equivalent, whether the λ's are different or not. From all these, it becomes probable that the two methods are almost equally powerful even in the general case of combining χ^2's with varying k and λ.

*Upper tail probabilities (q) were, of course, easily found for χ^2's with single degrees of freedom. For higher d.f., formula (22) given by Pearson and Hartley (1958, Introduction, pp.13-14) was used when $q > 0.001$; when $q < 0.001$, Tables of the Incomplete Gamma Function (K. Pearson, 1946) were used.

TABLE 1. RELATIVE POWERS OF THREE METHODS OF COMBINING n INDEPENDENT χ^2-TESTS WHEN k AND λ ARE EQUAL FOR ALL THE χ^2's

parameters of the individual non-central χ^2's combined		number of χ^2's combined (n)	number of model experiments	estimated powers (%)					
				at 5% level			at 0.1% level		
k	λ			$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test	$\Sigma\chi^2$ test	P_{λ_n} test	Tippett's test
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0.25	2	600	7.3	7.7	6.7	0.7	0.7	0.3
		6	200	12.0	11.5	9.0	—	—	—
		12	100	14.0	17.0	11.0	3.0	3.0	—
		24	50	18.0	20.0	10.0	4.0	4.0	—
1	1.0	2	600	21.2	21.7	20.0	1.3	1.7	1.0
		6	200	39.0	37.5	24.0	7.0	7.5	1.5
		12	100	51.0	54.0	23.0	17.0	17.0	1.0
		24	50	82.0	84.0	34.0	28.0	30.0	—
1	2.25	2	600	46.2	46.7	40.5	6.8	7.7	3.8
		6	200	82.0	82.0	54.5	31.5	34.0	6.5
		12	100	97.0	100.0	64.0	66.0	69.0	6.0
		24	50	100.0	100.0	68.0	100.0	100.0	8.0
1	4.0	2	600	72.3	73.2	64.8	22.2	22.8	12.3
		6	200	98.0	98.5	84.0	78.0	78.5	15.5
		12	100	100.0	100.0	95.0	100.0	100.0	23.0
		24	50	100.0	100.0	100.0	100.0	100.0	32.0
1	6.25	2	600	90.5	91.2	86.2	48.3	45.3	28.2
		6	200	100.0	100.0	97.0	97.0	98.0	46.5
		12	100	100.0	100.0	100.0	100.0	100.0	55.0
		24	50	100.0	100.0	100.0	100.0	100.0	70.0
6	1.5	2	600	17.3	17.3	13.5	1.7	1.5	0.8
		6	200	24.5	23.5	17.5	3.5	4.0	1.0
		12	100	39.0	38.0	17.0	8.0	8.0	2.0
		24	50	68.0	68.0	22.0	22.0	22.0	2.0
6	6.0	2	600	59.7	58.3	51.5	17.2	15.8	7.7
		6	200	95.0	91.5	73.0	60.0	58.5	11.5
		12	100	100.0	100.0	82.0	95.0	93.0	15.0
		24	50	100.0	100.0	90.0	100.0	100.0	24.0
12	3.0	2	300	21.7	22.0	18.3	3.0	2.7	3.3
		6	100	39.0	36.0	20.0	8.0	9.0	1.0
		12	50	68.0	68.0	32.0	22.0	22.0	—
12	12.0	2	300	82.7	80.3	77.7	38.3	37.3	23.0
		6	100	100.0	100.0	91.0	95.0	93.0	37.0
		12	50	100.0	100.0	98.0	100.0	100.0	52.0
24	6.0	2	150	34.0	33.3	29.3	4.7	4.7	4.7
		6	50	68.0	70.0	44.0	22.0	24.0	8.0
24	24.0	2	150	98.3	98.0	96.0	76.0	78.0	57.3
		6	50	100.0	100.0	100.0	100.0	100.0	82.0

TABLE 2. RELATIVE POWERS OF THREE METHODS OF COMBINING n INDEPENDENT SINGLE D.F. χ^2-TESTS, WHEN THE λ's ARE NOT EQUAL FOR ALL THE χ^2's

values of λ for the χ^2's combined: for one set of $\frac{n}{2}$ χ^2's	values of λ for the χ^2's combined: for the other set of $\frac{n}{2}$ χ^2's	number of χ^2's combined (n)	number of model experiments	estimated powers (%) at 5% level: $\Sigma\chi^2$ test	at 5% level: P_{λ_n} test	at 5% level: Tippett's test	at 0.1% level: $\Sigma\chi^2$ test	at 0.1% level: P_{λ_n} test	at 0.1% level: Tippett's test
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	1	2	600	13.5	13.8	14.2	1.0	0.8	0.5
		6	200	22.0	20.5	17.0	3.0	2.0	1.0
		12	100	25.0	26.0	17.0	4.0	4.0	1.0
		24	50	36.0	38.0	24.0	10.0	10.0	2.0
0.25	4	2	600	43.7	42.8	42.2	6.8	6.0	7.3
		6	200	79.0	81.0	81.5	28.5	27.0	8.5
		12	100	90.0	99.0	77.0	60.0	81.0	15.0
		24	50	100.0	100.0	86.0	98.0	98.0	22.0
0	6.25	2	600	61.8	61.2	63.2	15.0	14.2	16.0
		6	200	95.5	95.0	84.0	54.0	51.0	28.0
		12	100	100.0	100.0	96.0	98.0	97.0	36.0
		24	50	100.0	100.0	100.0	100.0	100.0	42.0
1	2.25	2	600	33.0	33.2	30.5	3.8	3.8	3.0
		6	200	61.5	62.5	38.5	16.0	17.5	5.0
		12	100	93.0	93.0	45.0	34.0	36.0	4.0
		24	50	100.0	100.0	50.0	74.0	84.0	4.0
1	4	2	600	52.3	52.8	48.3	9.5	9.7	6.3
		6	200	87.0	86.5	65.0	35.0	37.5	8.5
		12	100	98.0	97.0	78.0	80.0	81.0	9.0
		24	50	100.0	100.0	88.0	98.0	98.0	14.0
2.25	6.25	2	600	74.7	74.8	69.5	24.2	24.5	18.2
		6	200	97.5	97.5	90.0	82.0	82.5	28.0
		12	100	100.0	100.0	96.0	98.0	99.0	35.0
		24	50	100.0	100.0	98.0	100.0	100.0	50.0

2.2. Tippett's method seems to be less efficient than the other two in most cases, although for $n = 2$, the difference is small or sometimes even zero. The difference increases with n and is more marked for the 0.1% level. It is, however, possible that when one or a few of the λ's is sufficiently higher than the rest, the method may even be superior to the other two.

2.3. That the addition method is more efficient than the (min q) method is indirectly seen from Table 1 : the (min q) method applied to a set of non-central χ^2's becomes more efficient when applied to sub-totals of the same χ^2's.

3. Further observations

3.1. Combination of χ^2's may become necessary when one has carried out a series of goodness of fit tests or tests on a number of contingency tables, and where just one test on the pooled data may not be meaningful or adequate. In some cases there may be need of giving unequal weights to the tests being combined, (Yates. 1955a, 1955b; Zelen. 1957), but this point has been ignored in the present study.

3.2. The result for single d.f. χ^2's seems to have interesting implications, and in what follows, only single d.f. χ^2's are considered. Let $x_1, x_2, \ldots, x_n$ be independent normally distributed variates, each having unit s.d., but with $E(x_i) = \mu_i$; and let us suppose that the μ_i's are free to have any sign and magnitude. Let y_i be the incomplete probability integral corresponding to x_i, calculated with reference to the standard normal distribution. Then to test the hypothesis $H(\mu_1=\mu_2= \ldots = \mu_n=0)$, one can use either $\sum\limits_{i=1}^{n} x_i^2$ or the Sukhatme form of $P_{\lambda_n} = \prod\limits_{i=1}^{n} z_i$, where $z_i = 1-2\,|\,y_i-\frac{1}{2}\,|$. Since z_i is the upper tail probability corresponding to x_i^2, which is a non-central χ^2 with 1 d.f., Tables 1 and 2 imply that these two tests are of nearly equal power, although Σx_i^2 is generally believed to have some optimum properties for this well-known model which has direct bearing on the combination of independent two-sided tests, and hence to tests of homogeneity.

3.3. The two criteria are, however, closely similar. Whereas Σx_i^2 is a sum of single d.f. χ^2's, $-2\log_e(P_{\lambda_n})$ is the sum $\Sigma\,(-2\log_e z_i)$, and $-2\log_e z_i$ is that value of χ^2 with 2 d.f. which corresponds to x_i^2 in having the same incomplete probability integral y_i.

3.4 Birnbaum (1954) considered this model for the simple case $n = 2$, and found that the critical regions defined by the two criteria are very similar.* Earlier, Lancaster (1949) had studied the problem of combining two-sided tests on 2×2 tables or on binomial data, for the case where small frequencies are involved; and his work seemed to suggest that for combining the single d.f. χ^2's, the summation method and the P_{λ_n} technique (using upper tail areas) would be about equally powerful.

3.5. Earlier still, E. S. Pearson (1938) had showed that the critical region given by low values of $P_{\lambda_n} = \prod\limits_{i=1}^{n} [1-2\,|\,y_i-\frac{1}{2}\,|\,]$ is optimum for testing whether a sample of x-values $(x_1, x_2, \ldots, x_n)$ has probably arisen from a population $N(0, 1)$, where the alternative hypothesis states that x is $N(0, \sigma)$, with $\sigma > 1$, y_i being the incomplete probability integral of x_i under

*As regards Tippett's test Birnbaum's (1954) recommendations were largely influenced by his consideration for the heterogeneous case. For (more or less) homogeneous cases even criteria leading to non-convex acceptance regions may be definitely superior to Tippett's test. Even for ordinary heterogeneous cases, Tippett's test will become comparatively inefficient as n increases beyond 2. This is obviously because, unlike the P_{λ_n}-test or the $\Sigma\chi^2$-test, Tippett's test is unduly dependent on one extreme observation.

the null hypothesis. This result was not entirely correct, but it suggested that the Sukhatme form of the P_{λ_n} test is almost as efficient as the UMP test based on Σx_i^2. It is interesting to note that this model is a first approximation to that mentioned in para 3.2.

3.6. Yates (1955b) remarked that the problems considered by Lancaster (1949) were unrealistic. It seems, however, that the problem of combining two-sided tests may arise with binomial data.

3.7. Suppose one is given some (large-sample) binomial data, arranged group-wise, and wants to test for a preassigned proportion p_0 for all the groups, where the group proportions can individually exceed or fall short of p_0. Mather's monograph (1951, pp. 15-20) shows the application of χ^2-tests to such problems. The P_{λ_n} test could equally be used in such cases, and might even be adapted to give approximate analysis of the total divergence into components for 'deviation' and 'heterogeneity'.

4. Homogeneity of correlation coefficients

4.1. One may next consider a situation where one has k sample correlation coefficients $r_1, r_2, \ldots, r_k$, based on independent random samples from k bivariate normal populations. Let the respective sample sizes be $n_1, n_2, \ldots, n_k$, and the population correlation coefficients be $\rho_1, \rho_2, \ldots, \rho_k$; and suppose it is desired to test the hypothesis $H(\rho_1 = \rho_2 = \ldots = \rho_k = \rho_0)$, ρ_0 being a preassigned value, where the ρ_i's may either exceed or fall below ρ_0 individually.

4.2. The Fisherian test based on the z-transformation is well-known. K. Pearson (1933) suggested an alternative method based on probability integrals, but this was not properly oriented, and David (1938, pp. xxii-xxviii) rightly modified the Pearson test. Let $p_i = \int_{-1}^{r_i} P(r/\rho_0, n_i)dr$, where $P(r/\rho_0, n_i)$ is the frequency function of r_i under the null hypothesis. Then the Pearson-David criterion is $\prod_{i=1}^{n} [1-2|p_i-\frac{1}{2}|]$, small values of the product being significant.

4.3. Consider the case where the n_i's are so large that $y_i = \sqrt{n_i-3}\,(z_i-\xi_0)$ can be regarded as standard normal deviates under the null hypothesis, where $z_i = \tanh^{-1} r_i$, and $\xi_0 = \tanh^{-1}\rho_0$. If now one notes that p_i is the probability integral of y_i also, the problem is seen to be equivalent to that considered in para 3.2, Fisher's criterion $\Sigma\, y_i^2$ corresponding to that based on $\Sigma\, x_i^2$, and the Pearson-David criterion to P_{λ_n} of that para. Theoretical considerations suggest that the Fisherian test would have some optimum properties; but it involves approximations, while the other test is exact, and as far as the present investigation can show, the differences in power are almost negligible in most cases.

4.4. There could be many other instances where the Sukhatme form of P_{λ_n} can be applied to test whether a number of unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$ are simultaneously equal to a preassigned value θ_0. For a strict test of homogeneity, however, θ_0 should be left unspecified. In such cases, the parameter θ_0 has to be estimated from sample data before carrying out homogeneity tests and the exact distribution of P_{λ_n} becomes unknown. It is customary to still regard $-2\log_e(P_{\lambda_n})$ as a χ^2 with $2k$ d.f., but this number $2k$ is obviously too high.

ACKNOWLEDGEMENTS

The author is glad to acknowledge the help received from Shri Rabindranath Mukherjee and Shri Bimalendu Mahalanobis who organised the model sampling experiments.

REFERENCES

BIRNBAUM, ALLAN (1954): Combining independent tests of significance. *J. Amer. Stat. Ass.*, **49**, 559-574.

DAVID, F. N. (1938): *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, First Edition, Cambridge University Press.

LANCASTER, H. O. (1949): The combination of probabilities arising from data in discrete distributions. *Biometrika*, **36**, 370-382.

MATHER, K. (1951): *The Measurement of Linkage in Heredity*, Second Edition. Methuen & Co. Ltd., London.

PATNAIK, P. B. (1949): The non-central χ^2 and F-distributions and their applications. *Biometrika*, **36**, 202-232.

PEARSON, E. S. (1938): The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, **30**, 134-148.

——— and HARTLEY, H. O. (1958): *Biometrika Tables for Statisticians*, Vol. I, Second Edition, Cambridge University Press.

PEARSON, KARL (1933): On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**, 379-410.

———— (1946): *Tables of the Incomplete* Γ*-function*. Second Re-issue, Cambridge University Press.

WOLD, H. O. A. (1948): *Random Normal Deviates : Tracts for Computers*, No. XXV, Cambridge University Press.

YATES, FRANK (1955a): The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika*, **42**, 382-403.

———— (1955b): A note on the application of the combination of probabilities test to a set of 2×2 tables. *Biometrika*, **42**, 404-411.

ZELEN, MARVIN (1957): The analysis of incomplete block designs. *J. Amer. Stat. Ass.*, **52**, 204-217.

Paper received : February, 1960.