Plots of P-Values to Evaluate Many Tests Simultaneously

T. Schweder; E. Spjotvoll

Biometrika, Vol. 69, No. 3. (Dec., 1982), pp. 493-502.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198212%2969%3A3%3C493%3APOTEMT%3E2.0.CO%3B2-%23

Biometrika is currently published by Biometrika Trust.

Biometrika (1982), **69**, 3, *pp.* 493–502
Printed in Great Britain

Plots of *P*-values to evaluate many tests simultaneously

By T. SCHWEDER

Institute of Mathematical and Physical Sciences, University of Tromsø, Norway

and E. SPJØTVOLL

Department of Mathematics and Statistics, Agricultural University of Norway, Aas, Norway

Summary

When a large number of tests are made, possibly on the same data, it is proposed to base a simultaneous evaluation of all the tests on a plot of cumulative P-values using the observed significance probabilities. The points corresponding to true null hypotheses should form a straight line, while those for false null hypotheses should deviate from this line. The line may be used to estimate the number of true hypotheses. The properties of the method are studied in some detail for the problems of comparing all pairs of means in a one-way layout, testing all correlation coefficients in a large correlation matrix, and the evaluation of all 2×2 subtables in a contingency table. The plot is illustrated on real data.

Some key words: Simultaneous tests; Multiple comparison; P-value plot; One-way layout; Correlation matrix; Contingency table.

1. Introduction

We consider a situation in which a large number of tests are made on the same data or are related to the same problem. A classical example is the one-way layout in the analysis of variance when all pairs of means are compared. With 10 means there are 45 pairwise comparisons. For this example there exist simultaneous tests such as the S-method, the T-method and others. These methods, however, are not very powerful when a large number of comparisons is involved. Thus, if few real differences are detected, it may be due to lack of sensitivity of the tests. In this paper we present a simple graphical procedure, called a P-value plot, which gives an overall view of the set of tests. In particular, from the graph it is possible to estimate the number of hypotheses that ought to be 'rejected'. The technique is primarily intended for informal inference, and it is difficult to make exact probability statements. The method is general and can be used in many situations where other simultaneous methods are inapplicable.

The P-value plot is closely related to the half-normal plot of Daniel (1959) in the case of a known error variance, or its application to correlation coefficients (Hills, 1969), or 2^m contingency tables (Cox & Lauh, 1967). By applying a normal scores transform to both axes a P-value plot is converted to a normal plot. The inverse transformation may also be carried out. In §4 we argue slightly in favour of the P-value plot.

2. Problem and method

Consider a situation where we have T null hypotheses H_t ($t = 1, \ldots, T$). Suppose that H_t is rejected when a statistic Z_t is large. Let F_t be the cumulative distribution function of Z_t

under H_t. The P-value, i.e. significance probability, for the hypothesis H_t is

$$P_t = 1 - F_t(Z_t),$$

with possible correction if the distribution is discrete (Cox, 1977). We assume that the distribution of Z_t is completely known when H_t is true, so that P_t does not depend upon unknown parameters. We shall base our procedure upon the observed significance probabilities $P_1, \ldots, P_T$. If H_t is true, the significance probability P_t is uniformly distributed on the interval $(0, 1)$. If H_t is not true, P_t will tend to have small values.

Let T_0 be the unknown number of true null hypotheses, and let N_p be the number of P-values greater than p. Since the P-value should be small for a false null hypothesis, we have

$$E(N_p) \simeq T_0(1-p) \tag{1}$$

when p is not too small. A plot of N_p against $1-p$ should therefore for large p indicate a straight line with slope T_0.

For small values of p, we have $E(N_p) > T_0(1-p)$ since false hypotheses are then also counted in N_p. Hence for small p, N_p will lie above the line indicated by the N_p for large p.

The above analysis suggests the use of a cumulative plot of $Q_t = 1 - P_t$ against t $(t = 1, \ldots, T)$. This is of course a probability plot versus the uniform distribution. With absolute frequency along the vertical axes the plot may be considered a plot of the observed values of $(1-p, N_p)$. The left-hand part of the plot should be approximately linear. According to (1) the slope of that straight line is an estimate of T_0, the number of true null hypotheses. One should reject the null hypotheses corresponding to the points deviating from the straight line on the right-hand part of the plot.

Often, the plot will not show a clearcut break but rather a gradual bend. This indicates the presence of some 'almost true' null hypotheses, and the interpretation of the plot is less clear.

The number of P-values or null hypotheses needed to get sensible results with our methods will depend on several things. The number of true null hypotheses should be sufficiently large to permit a straight line to be fitted. The uncertainty in the plot is greater the more positive correlations there are between the P-values. This will vary with the type of application, as seen from the variance calculations we present below. These or other similar calculations will be of help in assessing the precision with which T_0 may be estimated. For example, Quesenberry & Hales (1980) have studied the variability of plots of the cumulative distribution function of a random sample from a uniform distribution. But also, in cases where no quantified estimate of T_0 is aimed at, one may obtain a valuable overview of the situation by a P-value plot, even based on as few as 15 P-values.

The results of Silverman (1976) may perhaps be used to study the asymptotic behaviour of the plot.

3. ONE-WAY LAYOUT

3·1. *The problem*

As an application of the above method, consider the one-way layout in the analysis of variance

$$X_{ij} = \mu_i + e_{ij} \quad (i = 1, \ldots, a;\ j = 1, \ldots, n_i),$$

where the errors $\{e_{ij}\}$ are assumed to be independent and identically normally distributed. The $\frac{1}{2}a(a-1)$ null hypotheses are H_{ij}: $\mu_i = \mu_j$ $(i < j)$. As test statistics we may use

$$T_{ij} = |\bar{X}_{i.} - \bar{X}_{j.}|\,(1/n_i + 1/n_j)^{-\frac{1}{2}} s^{-1},$$

where $\bar{X}_{i.}$ is the average of the ith group and s is an independent standard deviation estimate with ν degrees of freedom.

The observed significance probabilities are

$$P_{ij} = 2\{1 - H(T_{ij})\},$$

where H is the cumulative t distribution with ν degrees of freedom.

To base a simultaneous analysis on the set of pairwise comparisons is of course not a new idea (Lehmann, 1975, § 5·6). But utilizing the P-values in the way we propose, is, to our knowledge and surprise, novel; and it may be done whether t tests, the Wilcoxon method or any other appropriate test method is used.

3·2. *An example*

To demonstrate the method we analyse the same data as Duncan (1965). There are $a = 17$ observed group means: 654, 729, 755, 801, 828, 829, 846, 853, 861, 903, 908, 922, 933, 951, 977, 987, 1030. Each mean is the average of 5 observations. The residual mean square is 1713 on 64 degrees of freedom. The experiment is actually a two-factor experiment without interaction and hence with 64 and not 68 degrees of freedom, but this is of no importance in our context.

The plot of the P-values of the $T = 136$ tests comparing two means is given in Fig. 1. It seems that the left-hand part of the plot lies close to a straight line. The line given in the figure is drawn by visual fit. The slope of the line, which is an estimate of T_0, is approximately 25.

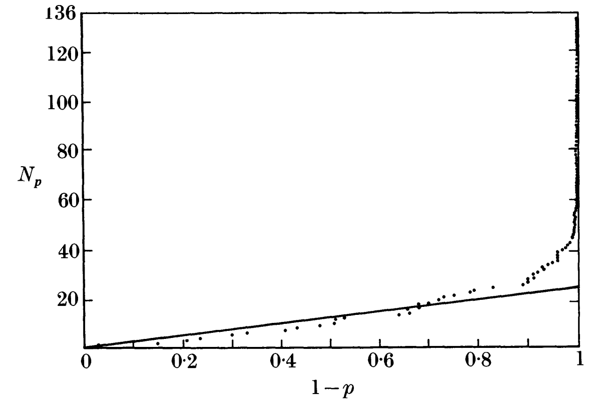

Fig. 1. Plot of P-values for Duncan's data involving 17 means

Duncan (1965, Table 5) gives the number of hypotheses not rejected by various multiple comparisons methods. When using level 0·05 it ranges from 34 for the least significant difference method to 69 for the S-method. Since the P-value plot indicates that there are approximately only 25 true null hypotheses, one conclusion is that we are in a situation where the standard multiple comparison methods do not detect differences due to low power. There are 34 hypotheses with P-values larger than 0·05. Most likely, many of these are due to false null hypotheses since the estimate of the number of true hypotheses is 25. Furthermore, one would expect that at most 1 or 2 of the P-values less than 0·05 correspond to true hypotheses. In other words, if one used the least significant difference at level 0·05, one would include 1 or 2 false rejections.

An approximate formal way of identifying null hypotheses that clearly should be rejected is the following. Since there are about 25 true null hypotheses we should, when aiming at an overall level α for at least one false rejection, use level $\alpha/25$ for the individual tests. This is an improvement over the same Bonferroni argument applied to all tests which would have given level $\alpha/136$ for the individual tests.

The P-value plot is primarily of help in deciding on the number of null hypotheses to be rejected, or alternatively on the significance level to be used. When the set of rejected hypotheses is settled, the further analysis is done in a standard way.

3·3. *The plot variance*

The P-value plot in the present situation, i.e. the null case, has a larger sampling variation than a P-value plot based on T independent uniformly distributed variates. This is due to the positive correlation between the P-values. Some care must therefore be taken in order not to over-interpret the P-value plot.

To get some idea of the amount of chance variation involved in the plots we shall calculate the variance of N_p in the null case where there are no differences in the μ's. Let D_{ij} be the indicator of $P_{ij} \geqslant p$ such that

$$N_p = \sum_{i<j} D_{ij}.$$

Due to the common standard deviation s, all the terms D_{ij} are slightly positively correlated. We shall neglect this correlation, that is we take $\nu = \infty$, and only take into account the correlation due to a common index in pairs of index pairs (i, j). Let X and Y be standardized binormal variates with correlation ρ and let z_p be the upper $\frac{1}{2}p$ fractile of the univariate normal distribution. Let

$$B(p, \rho) = \mathrm{pr}\,[\{|X| \leqslant z_p\} \cap \{|Y| \leqslant z_p\}] \tag{2}$$

denote the centred quadrat probability of the bivariate normal distribution. By a simple argument it is seen that

$$\mathrm{cov}\,(D_{12}, D_{13}) = B[p, \{(1+n_1/n_2)(1+n_1/n_3)\}^{-\frac{1}{2}}] - (1-p)^2,$$

when $\nu = \infty$ and under the null case. Therefore

$$\mathrm{var}\,(N_p) = \tfrac{1}{2}a(a-1)\,p(1-p) + \sum_{i \neq j \neq k} (B[p, \{(1+n_i/n_j)(1+n_i/n_k)\}^{-\frac{1}{2}}] - (1-p)^2). \tag{3}$$

With a balanced one-way layout, (3) simplifies to

$$\mathrm{var}\,(N_p) = \tfrac{1}{2}a(a-1)\,p(1-p) + a(a-1)(a-2)\,\{B(p, \tfrac{1}{2}) - (1-p)^2\}. \tag{4}$$

All the calculations we have done indicate that $B(p, \rho)$ is convex in ρ. Since the correlations occurring in (3) scatter around $\frac{1}{2}$, we may take (4) as a lower bound for (3). And in turn, (3) is a lower bound for $\mathrm{var}\,(N_p)$ when ν is finite.

To compute $\mathrm{var}\,(N_p)$, Table 1 or the table by Krishnaiah & Armitage (1965) may be used.

Table 1. *Binormal quadrat probability* (2) *for different values of the correlation coefficient* ρ *and different values of* p.

				$1-p$			
ρ	0·5	0·6	0·7	0·8	0·85	0·9	0·95
0·10	0·251	0·361	0·491	0·641	0·723	0·811	0·903
0·20	0·254	0·365	0·495	0·644	0·726	0·812	0·904
0·25	0·256	0·367	0·498	0·647	0·728	0·814	0·904
0·30	0·259	0·370	0·501	0·649	0·730	0·815	0·905
0·40	0·266	0·379	0·510	0·657	0·736	0·819	0·907
0·50	0·277	0·391	0·522	0·666	0·744	0·824	0·909
0·60	0·291	0·407	0·537	0·679	0·753	0·831	0·912
0·70	0·312	0·429	0·557	0·694	0·765	0·839	0·917

3·4. *The number of true hypotheses*

In the above example, we estimated the number T_0 of true null hypotheses by fitting visually a straight line through the points in the first part of the plot. The slope is used to estimate T_0. This technique could be formalized using some form of least squares fit. Since the points are not statistically independent, it would, however, be difficult to evaluate the properties of the estimation procedure.

To get some idea of the uncertainties involved, we shall consider a simple, but perhaps not very efficient estimator of T_0. Take a fixed value of p, find the corresponding N_p, and use the estimate $\hat{T}_0 = N_p/(1-p)$. By (1) this should be an approximately unbiased estimator for p not too small. We have

$$\operatorname{var}(\hat{T}_0) = \operatorname{var}(N_p)/(1-p)^2.$$

To have a small variance a small p should be used. But a small p would increase the bias of the estimator since many of the nonzero differences would be included in N_p. Therefore, a moderate value of p belonging to the part of the plot where there is an approximate straight line relationship should be used.

When we calculate $\operatorname{var}(\hat{T}_0)$ in an actual case, it is important to adjust for the fact that the number of true null hypotheses is less than T. A reasonable estimate of the variance is obtained by using for a in (4) the value $\hat{a}$ found from $\hat{T}_0 = \frac{1}{2}\hat{a}(\hat{a}-1)$. Let us illustrate this on the data in the example. With $p = 0{\cdot}3$ we find $N_p = 18$ and $\hat{T}_0 \bumpeq 26$. This corresponds to $\hat{a} \bumpeq 7$ and an adjusted $\operatorname{var}(\hat{T}_0) \bumpeq 8{\cdot}9$. Adding two standard deviations, we get 32 which may be considered as a certain approximate upper confidence limit for the number of true null hypotheses. The use of $T_0 = 32$ instead of $T_0 = 25$ will not greatly modify the conclusions drawn in § 3·2.

4. Correlation matrices

The $T = \frac{1}{2}a(a-1)$ empirical correlation coefficients based on an a-variate sample of size n may be plotted by their normal scores (Hills, 1969), or they may be plotted by their P-values relative to the hypotheses of no pairwise correlation. As an example we use data taken from Hills (1969). There are altogether $a = 13$ variables with sample size $n = 45$. Thus there are $T = 78$ different correlation coefficients. The plot of the significance probabilities is given in Fig. 2. The points in the main part of the plot seem to oscillate around a straight line. Such oscillations are natural when one considers the correlations among order statistics. The estimate of the number of true null hypotheses as read from the visually fitted line is $\hat{T}_0 = 64$.

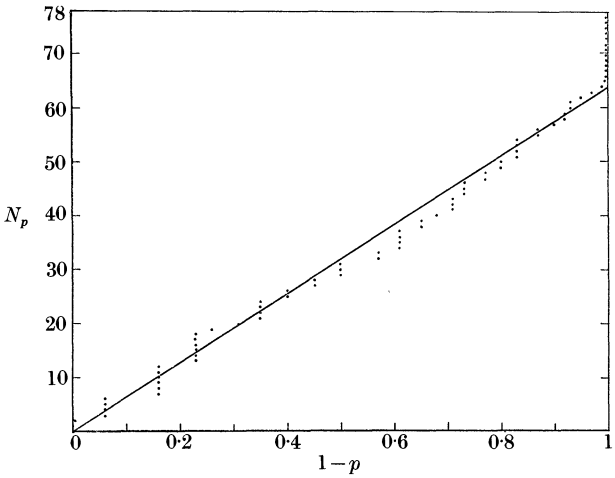

Fig. 2. Plot of P-values for 78 correlation coefficients

The formula for the variance of N_p, i.e. the number of significance probabilities greater than p, is very simple in the null case when all correlations are zero. This is because any two empirical correlation coefficients are uncorrelated when the corresponding true correlations are zero. This may be seen by utilizing results of Anderson (1958, p. 53) and by conditioning. We thus have

$$\operatorname{var}(N_p) = \tfrac{1}{2}a(a-1)\,p(1-p).$$

In the example with $p = 0{\cdot}1$ and $\hat{T}_0 = 64$ we find that $\operatorname{var}(\hat{T}_0)$ is estimated to be 7·1 when we adjust as in § 3·4. The plot, however, indicates that the precision of the estimate of T_0 is greater than indicated by this variance. This corresponds to the impression that a smaller p, say $p = 0{\cdot}05$, could have been used in estimating T_0, which would have reduced $\operatorname{var}(\hat{T}_0)$ considerably.

The half-normal plot of Hills (1969) is closely related to the P-value plot. The plots may be obtained from each other by a simple transformation. Hills presented his plot only in the present context of correlation coefficients, but it may be constructed whenever a P-value plot may be. We tend to prefer the P-value plot over the half-normal plot. First, the uniform probability transform, or the P-value, is widely used when testing hypotheses. Furthermore, it may be slightly easier to evaluate the variance of a P-value plot than of a half-normal plot. Finally, when we compare Hills's (1969) Fig. 1 with our Fig. 2, it seems somewhat easier to fit the line in the P-value plot, and the reason is that the tails of both the axes are too stretched out in the half-normal plot.

If we return to the example, Hills's estimate of T_0 as read from the normal plot is 62, since the ordered observations number 63 and 64 seem to deviate a good deal from the preceding observations (Hills, 1969, p. 250). In the P-value plot, however, this does not seem to be the case. The significance probabilities 0·03 and 0·01 associated with these observations are not suspiciously small, considering that they are the smallest among more than 60 significance probabilities.

5. Subtables of a contingency table

5·1. *The problem*

As a final illustration of the applicability of P-value plots we shall consider a two-way contingency table with r rows and c columns. Sometimes it is of interest to investigate subtables of such a table. We consider the set of 2 × 2 subtables. These may for example be of help in identifying one or more deviant cells relative to the assumption of independence between the two classifications. Many patterns may be conceived in which such a set of outlying cells is hard to identify by traditional methods based on residuals. The 2×2 subtables may also be of interest for other reasons, for example in testing for independence in a table with missing cells.

Also P-value plots other than those based on 2×2 subtables may be of interest in connexion with contingency tables. One may test for independence between pairs of rows or columns. Or, for higher dimensional contingency tables, one may plot all the P-values obtained by testing the interactions in say a log linear model. Cox & Lauh proposed a half-normal plot for the interactions in a 2^m-table: by use of the P-values this may be extended to any higher-dimensional table, although the P-values will not generally be uncorrelated as is the case in the simple 2^m case.

5·2. *Two examples*

We give two examples of plots for contingency tables. The first is taken from Kendall & Stuart (1967, p. 558), and concerns classification of students according to level of intelligence and standard of clothing. The data are given in Table 2. The P-value plot based on the chi-squared test of independence in the 90 subtables of size 2×2 is given in Fig. 3(a). The first part of the plot may be approximated reasonably well by a straight

Table 2. *Distribution of* 1725 *school children according to their standard of clothing and their intelligence: Kendall & Stuart* (1967, p. 558) *after Gilby.*

	Intelligence class						
Standard of clothing	A, B	C	D	E	F	G	Total
Very well clad	33	48	113	209	194	39	636
Well clad	41	100	202	255	138	15	751
Poor but passable	39	58	70	61	33	4	265
Very badly clad	17	13	22	10	10	1	73
Total	130	219	407	535	375	59	1725

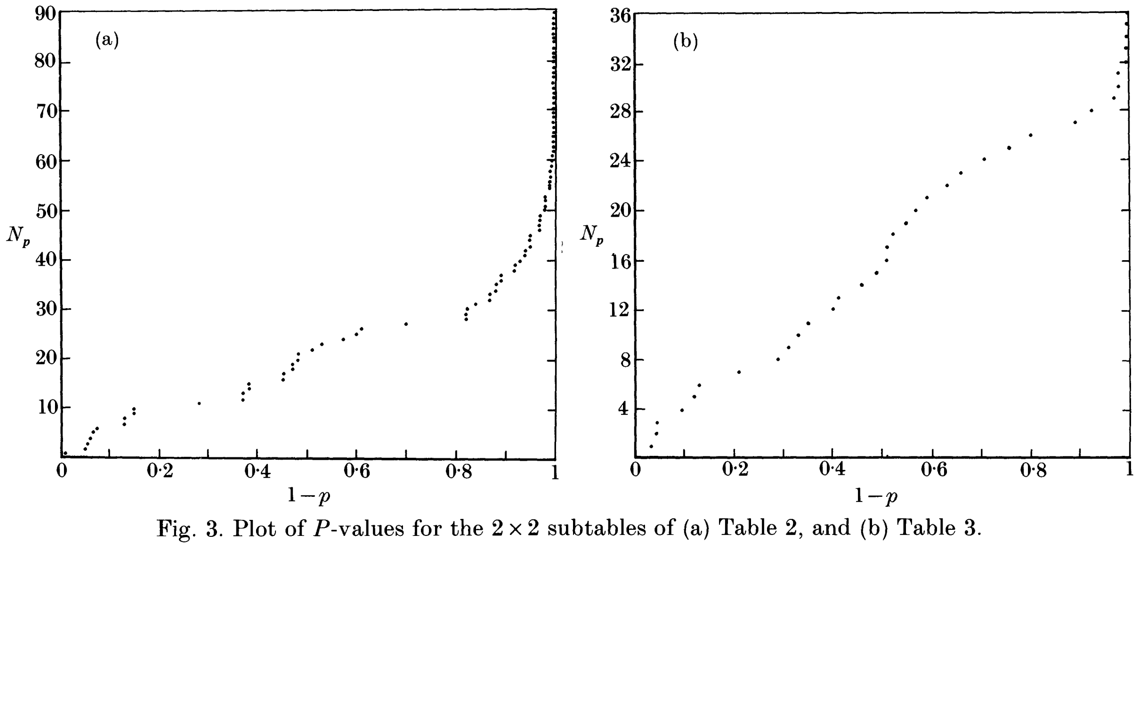

Fig. 3. Plot of P-values for the 2×2 subtables of (a) Table 2, and (b) Table 3.

line. There is, however, a good deal of uncertainty in the choice of the position of the line. Lines drawn through the main parts of the left-hand points will give an estimate of T_0 in the neighbourhood of 40.

When one looks more carefully at which hypotheses are not rejected, i.e. the ones with the 40 largest P-values, Table 2 may be simplified. The 'poor but passable' and the 'very badly clad' classes may be put together into one. The same is true for the F and G intelligence classes.

The second example is taken from Fisher (1946, p. 88), and concerns an experiment with crosses in mice. The data showing the results of 1204 crosses are given in Table 3. This is actually a 2^4 table. The table does not allow a simple interpretation in terms of a log linear model. Let us therefore look at the table in search of deviant cells relative to a simple structure. A reasonable approach would be to test all the 24 subtables of size 2×2 obtained by keeping the levels of 2 of the 4 factors fixed at a time. Instead of looking at a P-value plot based on these subtables, we have chosen to consider the P-value plot shown in Fig. 3(b), which is based on testing for independence in all the 36 subtables of size 2×2 obtained from Table 3 regarded as a two-way table. From the P-value plot it is seen that only the 5 or 6 smallest P-values seem to deviate from a straight line behaviour. When one looks at the 2×2 tables giving these P-values, it is seen that they all include the observation 130 for F_1 males in the repulsion phase for the cross Brown Piebald. This is obviously an outlying observation. The rest of the table does not show any deviation from homogeneity.

Table 3. *Distribution of* 1204 *crosses in mice according as the male or female parents were heterozygous, F_1, in the two factors, and according to whether the two dominant genes were received both from one, i.e. coupling, or one from each parent, i.e. repulsion; Fisher* (1946, p. 88) *after Wachter.*

	Black self	Black Piebald	Brown self	Brown Piebald	Total
Coupling					
F_1 males	88	82	75	60	305
F_1 female	38	34	30	21	123
Repulsion					
F_1 males	115	93	80	130	418
F_1 females	96	88	95	79	354
Total	337	297	280	290	1204

5·3. *The variance of N_p*

In this case it takes more labour to determine the variance of N_p. We assume the multinomial model with $E(X_{ij}) = np_{ij}$, $\Sigma_{ij} p_{ij} = 1$, and n large. To facilitate the notation let $r_i = \Sigma_j p_{ij}$, $c_j = \Sigma_i p_{ij}$ such that the independence assumption, under which we will work, reads $p_{ij} = r_i c_j$.

By linearizing the chi-squared test statistic for a 2×2 table and performing a fairly tedious calculation, the following asymptotic variance is found:

$$\begin{aligned}\operatorname{var}(N_p) = {} & \binom{r}{2}\binom{c}{2}(1-p) - \binom{r}{2}\binom{c}{2}(2r-3)(2c-3)(1-p)^2 \\ & + \Sigma_{ij}\Sigma_{C_{ij}} B[p, \{(1+r_i/r_{i_1})(1+r_i/r_{i_2})(1+c_j/c_{j_1})(1+c_j/c_{j_2})\}^{-\frac{1}{2}}] \\ & + \Sigma_i \Sigma_{D_i} B[p, \{(1+r_i/r_{i_1})(1+r_i/r_{i_2})\}^{-\frac{1}{2}}] \\ & + \Sigma_j \Sigma_{E_j} B[p, \{(1+c_j/c_{j_1})(1+c_j/c_{j_2})\}^{-\frac{1}{2}}], \qquad (5)\end{aligned}$$

where the index set

$$C_{ij} = \{(i_1, i_2, j_1, j_2) \mid i \neq i_1 \neq i_2, j \neq j_1 \neq j_2\}$$

refers to the pairs of 2×2 subtables with the (i, j)th cell in common,

$$D_i = \{(i_1, i_2, j_1, j_2) \mid i \neq i_1 \neq i_2, j_1 \neq j_2\}$$

refers to the pairs of subtables with cells from the ith row in common, and correspondingly

$$E_j = \{(i_1, i_2, j_1, j_2) \mid i_1 \neq i_2, j \neq j_1 \neq j_2\}.$$

The term $B(p, \rho)$ occurring in the variance formula denotes the binormal quadrat probability (2).

When the marginal probabilities are equiprobable $r_i = 1/r$, $c_j = 1/c$ and the variance simplifies to

$$\begin{aligned}\operatorname{var}(N_p) = \binom{r}{2}\binom{c}{2}\{ & (1-p) - (2r-3)(2c-3)(1-p)^2 \\ & + 4(r-2)(c-2)B(p,\tfrac{1}{4}) + (2r+2c-8)B(p,\tfrac{1}{2})\}. \qquad (6)\end{aligned}$$

Since $B(p, \rho)$ for fixed p is convex in ρ, we may take (6) as a lower bound for (5) since for one group of summands in (5) we have B-terms with correlations scattered around $\frac{1}{2}$ and for another group the correlations scatter around $\frac{1}{4}$. From numerical work, however, our impression is that inhomogeneity in the marginals seldom increases $\operatorname{var}(N_p)$ by more than 20% relative to (6).

Returning to the example with intelligence and clothing where $r = 4$ and $c = 6$, we find $\operatorname{var}(N_p) = 74{\cdot}4$ at $1-p = 0{\cdot}7$ in the null case. This, however, is not quite relevant when estimating $\operatorname{var}(\hat{T}_0)$ because the null case does not hold. The estimate of T_0 is about 40, corresponding approximately to the number of 2×2 subtables obtained from a 3×6 table. At $1-p = 0{\cdot}7$ the null variance of N_p from such a table is $29{\cdot}6 = 3 \times 15 \times 0{\cdot}66$. A relatively conservative estimate of the variance of N_p is then

$$\operatorname{var}(N_p) = 40 \times 0{\cdot}66 \times 1{\cdot}2 = 31{\cdot}7$$

and we obtain $\sqrt{\operatorname{var}(\hat{T}_0)} \simeq \sqrt{31{\cdot}7/0{\cdot}7} = 8$.

This high value is due to the strong positive correlation among the P-values. If these correlations erroneously had been disregarded, the estimated standard deviation of $\hat{T}_0$ would have been 4·5 instead of 8. This illustrates the importance of taking the correlation of the P-values into account in order not to draw too strong conclusions on the basis of the P-value plot.

6. Other applications

We have studied in some detail three general situations where P-value plots may be applied. Other cases are easily imagined. For example, in a medical study one may want to examine a large number of binomial or Poisson parameters. For a two-factor analysis of variance experiment one may study interactions by looking at all 2×2 tables. And in a multifactor experiment one may want to test a large number of interactions. Also in cluster analysis the P-value plot based on pairwise comparisons of individuals may be of help. Even in less structured situations where a number of tests have been done, say in a larger social science study, one may benefit from the overall view obtained by the P-value plot.

Let us finish by mentioning a situation where the plot cannot be used. As mentioned in § 4 the P-value plot may in some cases be considered as a probability transform of a half-normal or a normal plot. However, it cannot be used in cases where the half-normal plot or a variant of it is used both for estimating an unknown variance and detecting significant effects (Daniel, 1959; Gnanadesikan & Wilk, 1970; Schweder, 1981).

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Cox, D. R. (1977). The role of significance tests. *Scand. J. Statist.* **4**, 49–70.

Cox, D. R. & Lauh, E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics* **9**, 481–8.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two level experiments. *Technometrics* **1**, 311–41.

Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171–222.

Fisher, R. A. (1946). *Statistical Methods for Research Workers*, 10th edition. Edinburgh: Oliver and Boyd.

Gnanadesikan, R. & Wilk, M. B. (1970). A probability plotting procedure for general analysis of variance. *J. R. Statist. Soc.* B **32**, 88–101.

Hills, M. (1969). On looking at large correlation matrices. *Biometrika* **56**, 249–53.

Kendall, M. G. & Stuart, A. (1967). *The Advanced Theory of Statistics*, **2**, 2nd edition. London: Griffin.

Krishnaiah, P. R. & Armitage, J. V. (1965). *Tables for the Distribution of the Maximum of Correlated Chi-square Variates with One Degree for Freedom*. United States Air Force Wright-Patterson Air Force Base, Ohio: Aerospace Research Laboratories, Office of Aerospace Research.

Lehmann, E. L. (1975). *Nonparametrics, Statistical Methods Based on Ranks*. San Francisco: Holden Day.

Quesenberry, C. P & Hales, C. (1980). Concentration bands for uniformity plots. *J. Statist. Comput. Simul.* **11**, 41–53.

Schweder, T. (1981). A simple test for a set of sums of squares. *Appl. Statist.* **30**, 11–21.

Silverman, B. (1976). Limit theorems for dissociated random variables. *Adv. Appl. Prob.* **8**, 806–19.

[*Received March* 1981. *Revised April* 1982]