STATISTICS IN MEDICINE, VOL. 7, 1031–1043 (1988)

IDENTIFYING IMPORTANT RESULTS FROM MULTIPLE STATISTICAL TESTS

R. A. PARKER

*Division of Biostatistics, Department of Preventive Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232, U.S.A.**

AND

R. B. ROTHENBERG

Division of Chronic Disease Control, Centers for Disease Control, Atlanta, GA 30333, U.S.A.

SUMMARY

When many statistical tests are performed simultaneously, the overall chance of a type I error (incorrect rejection of a true null hypothesis) can substantially exceed the nominal error rate used in each individual test. Numerous techniques exist to adjust results of individual tests to control this problem. In general, these techniques apply a more stringent criterion of statistical significance (a smaller P-value) to each individual test than normally needed to maintain the experimentwise type I error. With an analysis that seeks to identify results for further research, however, such a conservative technique may not be appropriate. We present a new approach that uses a mixture of several distributions to model the set of P-values or of test statistics. One component models the results consistent with a failure to reject the null hypothesis, while the other distribution(s) in the mixture represent results inconsistent with the null hypothesis. These latter results may not achieve statistical significance based on a conventional P-value. We illustrate the use of the method on national mortality data and on several data sets analysed previously.

KEY WORDS Beta distributions Mixture models Multiple comparisons P-plots

1. INTRODUCTION

The decision to reject or not reject a null hypothesis depends on whether the P-value is less than a specified critical value, α, conventionally taken as 0·05. Of necessity, one makes this decision without knowledge of the actual truth of the null hypothesis. The decision may be incorrect in one of two ways: reject a 'true null hypothesis' (type I error, equivalent to the critical value used, α) or fail to reject a 'false null hypothesis' (type II error, β).

When making multiple statistical tests, however, the probability of at least one type I error among all the tests is considerably higher than the nominal critical value used on each test. For example, if $\alpha = 0{\cdot}05$, then the probability of making at least one type I error from ten independent tests is about 0·40, while the probability of making at least one error from 100 independent tests is greater than 0·99. This has led to various multiple comparison procedures, reviewed by Miller.[1] Generally these procedures intend to maintain the overall or experimentwise type I error at a

* Address for correspondence.

0277–6715/88/101031–13$06.50

Received September 1987
Revised May 1988

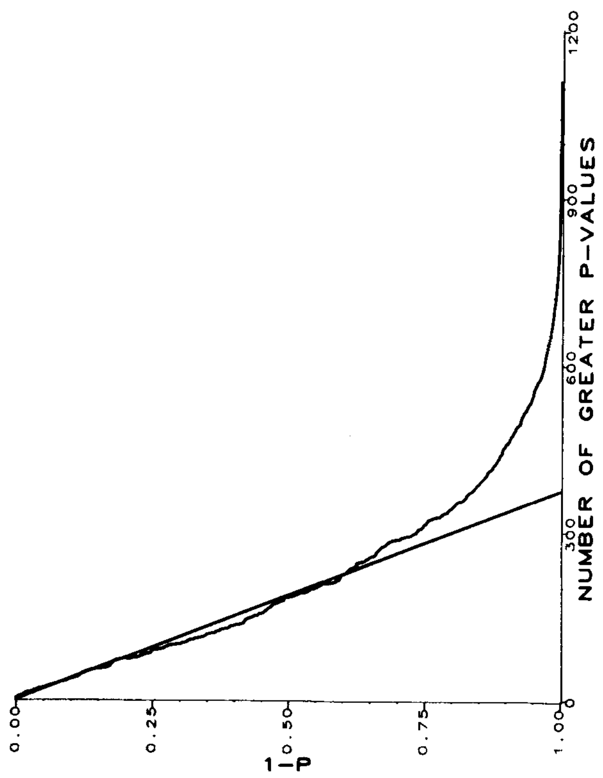

Figure 1. P-value plot of multiple cause of death data.
The straight line, fitted to the plot between $1-P=0{\cdot}0$ and $1-P=0{\cdot}6$, provides an estimate of the number of true null-hypotheses being tested. For details of the data plotted, see Section 4

specified level, normally by making the individual test criterion more stringent. This usually reduces the power of a test to detect individual results as significant.

One may not always find this approach desirable, particularly for an exploratory analysis of a data set. Such an analysis often involves hundreds or even thousands of statistical tests, even in a planned experiment with only one *a priori* hypothesis. Standard multiple comparison procedures apply an extremely stringent significance criterion for this large number of tests. One will likely conduct an exploratory analysis, however, to identify results of interest rather than to test specific null hypotheses. Findings may have interest even if not statistically significant at a nominal significance level. Thus, we need techniques to identify interesting results, whether statistically significant or not.

Recently, Schweder and Spjøtvoll[2] suggested a simple graphical approach to the multiple comparison problem. Although they intended their technique for the standard multiple comparison situation, the method can identify results that arise from false null hypotheses even if the P-value is greater than a nominal 0·05 significance level. Their approach is based on the fact that the distribution of P-values from a statistical test performed on a true null hypothesis is uniform between 0 and 1. For a statistical test of a false null hypothesis, however, the P-value would tend to be near 0, since we should reject a false null hypothesis more often than a true null hypothesis. Thus, Schweder and Spjøtvoll suggest plotting the value $(1-P)$ on the X-axis against N_p, the number of larger P-values, on the Y-axis (Figure 1). If all the significance tests concerned true null hypotheses, then the P-value plot would approximate a straight line from 0 to 1. With contamination from a number of tests of false null hypotheses, however, the plot would show a cluster of results near $1-P=1$. In the latter situation, one fits a straight line to the linear portion of the plot starting from values of $(1-P)$ near 0; Schweder and Spjøtvoll draw this line by visual fit. The value of this straight line at $(1-P)=1$ is an estimate of the total number of true null hypotheses under test. The number of tests remaining, therefore, estimates the number of false null hypotheses under test.

We can extend and formalize their approach by applying a mathematical model to the data set consisting of values of $(1-P)$ and N_p. The model includes one component for the results arising from true null hypotheses; for P-values, this term is a uniform distribution. If this single term adequately fits the data, then there is no evidence that any of the null hypotheses are false. If there is a bulge near $P=0$, however, the model based on a single term will not provide an adequate fit to the data, so we would need other components in the model. These other components represent results from false null hypotheses, the interesting results that we shall attempt to identify. We will refer to results likely to come from these components of the model as 'significant' to indicate that they arise from false null hypotheses, even though they may not be 'statistically significant'.

There are several reasons to develop such an approach. First, use of a model eliminates some of the subjectivity involved in estimation of the number of true null hypotheses by fitting a straight line to part of the plot. The model provides an objective estimate of the number of true null hypotheses in the data set and we can obtain the accuracy of this estimate from the model. After we fit the model, moreover, we can estimate the probability that a particular test result comes from a particular component. In addition to estimation of the probability that a result represents a true null hypothesis, we can use the model to examine whether results cluster in meaningful ways, helpful in deciding which results merit further investigation.

We use standard results for mixture distributions to model the data.[3] Section 2 presents the notation and model. In Section 3, we discuss how to determine the appropriate number of distributions to include in the model, and we illustrate the method in Section 4. We compare our approach to other methods in Section 5, and we briefly discuss advantages and disadvantages in the final section.

2. NOTATION AND MODEL

Assume a set of N null hypotheses, each tested by a statistical test with test statistic y_i. We do not assume independence of these hypotheses or that we use the same statistical test for each null hypothesis. In fact, we could test the same hypothesis with a number of different statistical techniques, in which case we would expect the probabilities from these tests to cluster together. For each null hypothesis, i, $i=1, \ldots, N$, let x_i $(0 \leq x_i \leq 1)$ denote the probability that the test statistic would have a value of y_i, or a more extreme value, if the null hypothesis were true; small values of x_i provide strong evidence that the ith null hypothesis is not true. We wish to estimate T_0, the number of true null hypotheses. In addition, we would like to estimate the probability that any particular test result (y_i) arises from a true null hypothesis.

Let $\beta(r, s)(\ )$ represent the beta distribution with parameters r and s, $0<r, s$, that is, for any $x \subset [0, 1]$, $\beta(r,s)(x)=\Gamma(r+s)x^{r-1}(1-x)^{s-1}/\Gamma(r)\Gamma(s)$. Recall that we can express the uniform distribution as a beta distribution with parameters $r=s=1$.

We can model any distribution of data on the interval $[0,1]$ as a finite mixture of beta distributions.[4] Thus, we can always use a mixture of beta distributions as a generic model for $\{x_i\}$, independent of the specific statistical test(s) used to obtain $\{x_i\}$. For example, the generic approach can be used when statistical significance is determined using a variety of different statistical tests within the same experiment, such as chi-square tests on contingency tables, t-tests for differences between groups for continuous variables and standard normal deviates for correlation of variables.

Let K, a non-negative integer, index a hierarchical family of distributions, with the Kth member of the family given by

$$g_K(x_i)=p_0\beta(1, 1)(x_i)+\sum_{j=1}^{K} p_j\beta(r_j, s_j)(x_i) \tag{1}$$

where p_j denotes the relative proportion of the set $\{x_i\}$ in the jth distribution, $j=0, \ldots, K$, with $0<p_j\leq 1$ and $\sum_{j=0}^{K} p_j=1$. We will refer to this as a K-distribution model to indicate that there are K distributions inconsistent with the null hypothesis included in the model.

We have deliberately expressed the density function (1) as the sum of two terms. All x_i that represent true null hypotheses, no matter what the value of x_i, arise from the first term in the density, $p_0\beta(1, 1)(\;)$. The remaining x_i, that represent false null hypotheses, come from the remaining term in (1). Thus, if all test results arise from true null hypotheses, then $K=0$ and $p_0=1$.

Assume that we know the value of K, that is, that a mixture of K non-null distributions plus 1 null distribution adequately fit $\{x_i\}$. The likelihood of $\{x_i\}$ is

$$L_1=\prod_{i=1}^{N} g_K(x_i), \tag{2}$$

since we do not know *a priori* which component of (1) gives rise to each x_i.

We require an iterative procedure to obtain the maximum likelihood estimators of (2). Titterington *et al.*[3] give an EM algorithm (Section 4.3.2) for mixtures of exponential family densities, including the beta model in (1). We can modify this approach, as needed, to fit other distributions, or we can use a general maximization routine (for example, the Nelder–Mead algorithm[5]). Other methods of estimating the parameters (for example, method of moments) are also available.[3]

Since the distribution $j=0$ is the null distribution that represents those test results that arise from true null hypotheses, $\hat{p}_0$ is an estimate of the proportion of results that come from the null hypothesis. Thus, $N\hat{p}_0$ is an estimate of T_0, the number of true null hypotheses. We can obtain an obvious approximation to the variance of T_0 by treating $\hat{p}_0$ as a proportion, that is,

$$\operatorname{var}(T_0)\doteq[N\hat{p}_0(1-\hat{p}_0)]$$

or we could use the information matrix based on the likelihood formula (2).

Given a value for K and the associated parameters for the mixture, one can directly estimate the probability that a P-value, x, comes from the null distribution. We can do this for any value of x, whether observed or not, so that we could identify a critical region based on the probability that a result represents a true null hypothesis. From (1), the probability that x comes from the null distribution is

$$d_0(x)=p_0B(1, 1)(x)/g_K(x)=p_0/g_K(x). \tag{3}$$

The probability (3) is the probability density that a specific value of x is the result of a test of a true null hypothesis, divided by the total probability density that we observed the exact value x. This probability is not a tail-probability in a conventional significance test; rather, it should be interpreted in the frequentist sense. If we have 100 significance tests that give a P-value of x, then we would expect $100d_0(x)$ of these tests to represent true null hypotheses and the remainder to arise from the non-null distributions. Equation (3) leads to sensible results, with small values of x having a small chance of coming from the null distribution. We can immediately generalize equation (3) to give the probability that the value x comes from any of the component distributions, with use of the density of the particular component distribution as the numerator of (3) instead of p_0.

If all x_i come from the same underlying distribution, we can use an alternative model based on the specific statistical test used. Let $f_\lambda(\ )$ stand for the underlying probability distribution, with parameter λ possibly a vector; symbolize the null distribution by $\lambda = 0$. The parameter, λ, measures displacement from the null hypothesis; it directly relates to the statistical test used to analyse the data initially. Thus, it can be of help to the investigator in interpreting the importance of the false null hypotheses. The beta model is likely to have a less immediate interpretation for the investigator for two reasons. First, a beta distribution is usually unfamiliar to most investigators, since it is not used in routine statistical tests. Second, the individual parameters are not directly interpretable, although the ratio $r/(r+s)$ is the mean of the probability distribution.

To avoid difficulties when $\{x_i\}$ comes from a two-tailed significance test, transform each x_i to an equivalent t_i, based on $f_\lambda(\ )$, such that

$$1 - x_i = \int_{-\infty}^{t_i} f_\lambda(x)\,\mathrm{d}x. \tag{4}$$

Large values of t_i imply strong evidence against the null hypothesis. If x_i results from a one-sided significance test, then $t_i = y_i$. Similar to (1), let

$$g_K(t_i) = p_0 f_0(t_i) + \sum_{j=1}^{K} p_j f_{\lambda_j}(t_i) \tag{5}$$

and thus, assuming K is known, the likelihood is

$$L_2 = \prod_{i=1}^{N} g_K(t_i). \tag{6}$$

The estimation of the parameters in (6) is identical to those of (2). The interpretation of the parameters p_j is the same as for (2), and we can interpret the displacement parameters, λ_j, as a distance from the null component to the non-null components. Similarly, we could use the density function (5) instead of (1) in equation (3) to obtain the probability that a specific value of t comes from a particular component in the model.

One can convert any P-value to an equivalent test statistic for another test, so that it would be possible to transform a variety of test results to standard normal deviates, for example, in which case it would be possible to use a specific approach based on a standard normal deviate to model a variety of test statistics. We feel, however, that this would be inappropriate since the appeal of the specific approach is that it relates to the actual statistical test used to analyse the data.

3. DETERMINING THE NUMBER OF COMPONENT DISTRIBUTIONS

The parameters estimated for the model change as the number of components varies. As we add terms to the model, the estimated value of $\hat{p}_0$ tends to decrease, since we will tend to model more marginal results as false null hypotheses in the density function with more terms. Thus, to have an appropriate estimate of the number of true null hypotheses, it is essential that the model we use has the appropriate number of components. Moreover, if the number of components in the model is incorrect, then the probability that a specific result comes from a specific component is likely to be inaccurate. For example, if we include too few components in the model, we may overlook potentially interesting findings as they may be consistent with the null hypothesis based on (3). However, if we include too many components, then we will misclassify true null hypotheses as potentially interesting findings.

Unfortunately, determination of the appropriate number of components is difficult since we cannot use standard asymptotic results for several reasons. First, the null hypothesis for the more parsimonious model may not be unique. For example, suppose we are testing the K-distribution model against the $(K-1)$-distribution model. The null hypothesis could be that $p_K = 0$, or alternatively, that the parameters λ_K are equal to the parameters of λ_{K^*}, $0 \leq K^* < K$. If there is more than one parameter involved in λ_K, then it is unclear whether we should compare the likelihood ratio test with a chi-square distribution on 1 degree of freedom, corresponding to $p_K = 0$, or to a chi-square distribution based on the number of parameters in λ_K.

Furthermore, in either case we violate the regularity conditions used to develop the asymptotic theory of the likelihood ratio test.[6] These conditions include requirements that the parameters are interior to the parameter space (violated when $p_K = 0$) and that the probability distributions defined by any two sets of parameters differ (violated when $\lambda_K = \lambda_{K^*}$ while holding $p_K + p_{K^*}$ constant). Since we have violated the asymptotic theory underlying the likelihood ratio test, we have no formal method to determine the appropriate value of K. This is a well known problem in the mixture model literature.

We determined K by assessing the goodness-of-fit of each model, and stopped at the smallest K that 'adequately' fit the data. Starting with $K = 0$, we calculated the Cramer–von Mises (CVM) statistic to assess whether $g_K(\ )$ provides an adequate fit to the data. The CVM is a global assessment of the fit of a hypothesized distribution to an observed empirical distribution function. Since calculation of the published critical values for this statistic assumes that one knows the hypothesized distribution, $g_K(\ )$, independent of the data, these critical values are appropriate only for $K = 0$. A simple way to decide on an adequate fit is to compare the CVM statistics for the K-distribution and the $K+1$-distribution models. With little or no improvement between these two statistics, we can select the K-distribution model to provide an adequate fit to the data.

A more rigorous method, used in this paper, is to assess the adequacy of fit by simulation of the empirical distribution function of the CVM statistic itself, when one calculates the statistic from a distribution fit to the data. To do this, we first fit the K-distribution model to the observed data set (the 'observed model'). For each iteration of the simulation, we generate N values at random from the observed model (the 'simulated data'), fit a new K-distribution model to the simulated data (the 'simulated model'), and finally calculate the CVM statistic to assess the adequacy of fit of the simulated model to the simulated data. Repetition of this process provides an empirical distribution function for a CVM statistic calculated for a distribution with a specified number of components fitted to data. With a comparison of the observed CVM from our observed model to the empirical distribution function of the CVM, we can decide formally whether the model provides a reasonable fit to the data. Even with this approach, however, selection of a model that 'adequately' fits the data is, as always, partly subjective.

4. EXAMPLE: MULTIPLE CAUSE OF DEATH DATA

The National Center for Health Statistics has collected multiple cause of death (MCD) data for the United States annually since 1968. These computer tapes contain demographic information from each death certificate and all diseases mentioned on the death certificate, with the exception of certain states in particular years when only a 50 per cent sample of death certificates were coded. The diseases listed include the underlying cause of death, the causal pathway that led to death, and all other conditions that the physician lists on the death certificate.

To examine potential time trends in chronic diseases, we calculated the 'reported prevalence at death' (the proportion of death certificates mentioning a disease) for 16 broad disease groups for

Table I. Results for multiple cause of death trends: t-distribution

Non-null*	Log-likelihood	Proportion†	Displacement‡	CVM§
0	−4162·141	1·000	0·000	131·023 ($P \ll 0·01$)
1	−2517·189	0·457	0·000	2·332 ($P \ll 0·01$)
		0·543	3·362	
2	−2402·093	0·303	0·000	·056 ($P = 0·03$)
		0·451	1·955	
		0·246	4·568	
3	−2389·001	0·278	0·000	0·015 ($P = 0·85$)
		0·415	1·698	
		0·273	4·039	
		0·034	6·511	
4	−2386·676	0·272	0·000	0·013
		0·406	1·647	
		0·270	3·898	
		0·049	5·933	
		0·003	9·333	

* Number of non-null distributions included in model (K)

† Proportion of all test results that arise from distribution (p_j)

‡ Displacement from null hypothesis (λ_j)

§ For 0- and 1-distribution models, P-value of CVM result from standard Cramer–von Mises tables.[7] For models with 2 or 3 non-null distributions, P-values result from simulations (simulations not run for 4-distribution model)

each year from 1968 to 1982 for each of 72 age/race/sex combinations (= 18 age groups × 2 races × 2 sexes). We conducted separate analyses on each age/race/sex strata so that differences in trends in different sub-groups would be readily apparent. As a screening tool, we used simple linear regression of prevalence against calendar year as evidence for time trends in the data. We based each of these regressions on 15 data points, so we calculated P-values from a t-test with 13 degrees of freedom. We performed only 1113 significance tests on the estimated slopes, since there were no deaths for 39 of the 1152 (= 16 × 72) possible combinations of disease and strata. With use of the standard Bonferroni method to adjust for 1113 significance tests, the required significance criterion is $\alpha^* = 0·0000449$ for a nominal $\alpha = 0·05$ test; there are 54 results significant at this level.

Figure 1 is the P-value plot of these results. Depending on the portion of the plot regarded as straight, we estimated a range of 330 to 375 true null hypotheses, which suggests that over 700 results are significant and represent real trends in disease prevalence over time. Some of these significant results must have P-values on the order of $P = 0·1$ or $P = 0·2$, since only 567 tests have a P-value less than the conventional significance level, $P = 0·05$.

4.1. Specific approach: *t*-distribution on 13 d.f.

The results of analysis of the data as a mixture of t-distributions, together with the related CVM statistics, appear in Table I. Neither the model $K = 0$ or $K = 1$ fit the data adequately since the CVM statistic is highly significant compared to published critical values. Thus, we should include at least two non-null distributions in the model.

The difference in likelihood between $K = 2$ and $K = 3$ is highly significant ($P < 0·00001$), although, as pointed out in Section 3, this test is formally invalid. The additional term detected when $K = 3$ implies that 3·4 per cent of the test results come from a very significant population, displaced 6·5 t-units from the null hypothesis, that is, we can think of the statistics as coming from a t-distribution with a mean of 6·5. When going from $K=3$ to $K=4$, the change in likelihood is not statistically significant ($P=0{\cdot}09$), and the new term in the model (0·3 per cent of the data come from a distribution displaced by a mean of 9·3 units) seems to have no practical importance. Thus, it appears that we need to examine the adequacy of the fit to the data only for the models with two or three displaced distributions.

For the two-distribution model, the CVM statistic is 0·056, a very good fit when judged by published tables.[7] Simulations of the actual null distribution for the CVM statistic with the model based on the observed data, however, indicate that only about 5 per cent of the null distribution would be above 0·035. Thus, the observed value 0·056 implies that the fitted model is inadequate ($P=0{\cdot}03$). This illustrates the effect of estimation of the model from the data on the null distribution of the CVM statistic. The CVM statistic for the three-distribution model is 0·015, which provides a very good fit to the data based on simulations ($P=0{\cdot}85$). Thus, we conclude that we need three non-null t-distributions to model the MCD data. We could have reached a similar conclusion without simulation of the empirical distribution function of the CVM, since there is a substantial drop in CVM between the two-distribution model and the three-distribution model, but virtually no drop between the three-distribution and four-distribution models.

The model $K=3$ suggests that 27·8 per cent of the test results come from true null hypotheses, equivalent to 309 true null results. This is somewhat less than the range of 330 to 375 estimated graphically. The location parameters in the non-null distributions give some information about the make-up of the set of results. For example, the first non-null distribution, which includes 41·5 per cent of all the test results, accounts for approximately three-fifths of all the significant results. This group represents relatively marginal age/sex/race disease trends in the data, since the displacement of the distribution is only 1·7 t-units, and implies a conventional one-sided significance level of about $P=0{\cdot}055$ from a t_{13}-distribution. The other two non-null distributions, however, indicate very strong trends.

From (3) one can calculate the probability that a particular t-value came from the null distribution or from one of the displaced distributions, indicating a real effect. For example, the t-value 0·607 (which would have a conventional one-sided significance level of $P=0{\cdot}28$ from a t_{13}-distribution) actually has a 50 per cent chance of coming from the null distribution, with the assumption that the model $K=3$ is true. Similarly, we would expect a t-value of 2·418 (conventional significance level $P=0{\cdot}016$) to come from the null distribution only 5 per cent of the time. As in the selection of the number of distributions to include in the model, the choice of the specific t-value to use to select results of interest is a subjective decision.

By use of the extension of (3), we can determine the most probable distribution that gives rise to each individual significance test. For example, a t-value of 6·00 likely comes from either the second or the third displaced distribution. The probability density for $t=6{\cdot}00$ is 0·0174 for the second distribution and 0·115 for the third distribution, so it more likely comes from the third non-null distribution. Although this approach will involve some misclassification, the results support the idea that, at least for some diseases, the data clump from one or two distributions. For example, over 90 per cent of all the trends for diseases such as cancer, multiple sclerosis, and osteoporosis, come from either the null distribution or the distribution of marginally significant results. Similarly, for diseases such as chronic obstructive pulmonary disease, ischemic heart disease, and stroke, more than half of the trends come from one of the two more displaced distributions. With individual results cross-classified by disease and probable component, there is strong evidence of heterogeneity in trends in diseases ($\chi^2_{45}=337, P\sim 10^{-27}$). This is very important in identification of those results to investigate further. In our case, we concentrated our efforts on investigation of diseases that appeared to change for a number of age/race/sex groups, rather than investigation of diseases that did not show a significant clustering.

Table II. Results for multiple cause of death trends: beta distribution

Non-null*	Log-likelihood	Proportion†	Displacement‡	CVM§
0	0·000	1·000	1·000 1·000	131·023 ($P \ll 0{\cdot}01$)
1	1736·956	0·345	1·000 1·000	0·522 ($P<0{\cdot}05$)
		0·655	0·297 6·956	
2	1768·261	0·327	1·000 1·000	0·045 ($P=0{\cdot}065$)
		0·491	0·401 5·606	
		0·182	0·543 634·117	
3	1768·844	0·337	1·000 1·000	0·047
		0·439	0·470 6·787	
		0·144	0·958 633·761	
		0·080	0·850 10255·852	

* Number of non-null distributions included in model (K)

† Proportion of all test results that arise from distribution (p_j)

‡ Displacement from null hypothesis (r_j, s_j)

§ For 0- and 1-distribution models, P-value of CVM result from standard Cramer–von Mises tables.[7] For 2-distribution model, P-values result from simulations (simulations not run for 3-distribution model)

4.2. Generic approach: beta distributions

It is always possible to fit a model based on beta distributions to any collection of P-values. Table II shows the results of this approach to the multiple cause of death data. The model with a single non-null beta distribution does not provide an adequate fit to the data, since the CVM statistic is significant ($P<0{\cdot}05$), even ignoring the fact that we estimated the model from the data. Since there is no improvement in log-likelihood between the models with two and three non-null beta distributions, and the third distribution models only a very extreme distribution with mean probability, $r_3/(r_3+s_3)$, of 0·00008, it would seem that we would use the model with two non-null beta distributions to model the data. This model provides a marginal fit to the data, based on the simulated CVM statistic ($P=0{\cdot}06$). In practice, we would choose to model the distribution of results as a mixture of a null plus three non-null t-distributions, since the CVM for the t-distribution model is 0·015, considerably less than the CVM of 0·045 for the beta model, and the two models involve fitting the same number of parameters.

Interpretation of the beta model follows the same approach as the specific model. Approximately 364 tests (32·7 per cent) come from a true null hypotheses, compared with 309 tests (27·8 per cent) for the specific model. Similarly, we can calculate the probability that a specific test result comes from the null distribution. For example, the beta model implies that we would expect a test giving a P-value of 0·0109 to come from the null distribution only 5 per cent of the time. This value is smaller than the P-value of 0·0155, equivalent to the t-value of 2·418 obtained from the t-distribution model.

Table III. Results for Duncan[8] data

Non-null*	Log-likelihood	Proportion†	Displacement‡	CVM§
0	−2219·459	1·000	0·000	28·842 ($P \ll 0{\cdot}01$)
1	−516·659	0·421	0·000	1·676 ($P \ll 0{\cdot}01$)
		0·579	5·933	
2	−383·165	0·238	0·000	0·183 ($P \ll 0{\cdot}01$)
		0·510	3·544	
		0·252	8·363	
3	−348·673	0·202	0·000	0·044 ($P = 0{\cdot}10$)
		0·458	3·018	
		0·250	6·563	
		0·090	10·711	
4	−344·711	0·130	0·000	0·014 ($P = 0{\cdot}84$)
		0·286	1·837	
		0·294	3·984	
		0·207	6·967	
		0·083	10·868	
5	−341·319	0·104	0·000	0·013 ($P = 0{\cdot}88$)
		0·251	1·459	
		0·332	3·654	
		0·211	6·639	
		0·078	9·798	
		0·024	12·700	

* Number of non-null distributions included in model (K)

† Proportion of all test results that arise from distribution (p_j)

‡ Displacement from null hypothesis (λ_j)

§ For 0- and 1-distribution models, P-value of CVM result from standard Cramer–von Mises tables.[7] For models with 2 or more non-null distributions, P-values result from simulations

5. COMPARISON WITH OTHER METHODS

We can compare our approach with other methods by use of previously published data sets. The first data set, originally published by Duncan,[8] consists of the means of 17 groups, with an estimated residual mean square based on 64 degrees of freedom. These 17 groups lead to 136 comparisons between pairs of means. Schweder and Spjøtvoll[2] estimate that there are approximately 25 true null hypotheses, while Duncan[8] estimated a range of 34 to 69, depending on the technique one uses to identify significant differences.

Since all test statistics are based on a t-test with 64 degrees of freedom, we fit models with an increasing number of non-null t_{64}-distributions, until we obtained an adequate fit. The fit of models with 0 to 5 non-null distributions appears in Table III with the associated CVM statistic. The models with 0 and 1 non-null distributions grossly fail to fit the data and the two-distribution model fails to fit the data based on the simulated CVM statistic. The fit of the three-distribution model is not unacceptable ($P = 0{\cdot}10$), but addition of a fourth non-null distribution substantially improves the fit ($P = 0{\cdot}84$), and suggests that one should prefer the four-distribution model. This is reflected in the value of the CVM statistics, which drops from 0·044 for the three-distribution model to 0·014 for the four-distribution model. The four-distribution model separates the principal displaced distribution in the three-distribution model (45·8 per cent of the data with a displacement of 3·018 t-units) into two components, one with 28·6 per cent of the tests displaced by

Table IV. Results for Hill[9] data

Non-null*	Log-likelihood	Proportion†	Displacement‡	CVM§
0	−469·918	1·000	0·000	1·259 ($P \ll 0{\cdot}01$)
1	−195·602	0·872	0·000	0·232 ($P = 0{\cdot}14$)
		0·128	7·812	
2	−162·675	0·836	0·000	0·114 ($P = 0{\cdot}35$)
		0·138	5·711	
		0·026	13·054	
3	−154·640	0·812	0·000	0·064 ($P = 0{\cdot}57$)
		0·084	3·535	
		0·078	7·037	
		0·026	13·053	

* Number of non-null distributions included in model (K)

† Proportion of all test results that arise from distribution (p_j)

‡ Displacement from null hypothesis (λ_j)

§ For 0-distribution model, P-value of CVM result from standard Cramer–von Mises tables.[7] For models with 1 or more non-null distributions, P-values result from simulations

1·837 and another with 29·4 per cent of the data with a displacement of 3·984. Predictably, the additional component reduces the estimated proportion of true null hypotheses from 20·2 per cent to 13·0 per cent. The four-distribution model also would lead to a much lower cutoff to identify a false null hypothesis. With four non-null components, there is a 50 per cent chance that a t-value of 0·485 (equivalent to a one-sided significance level of $P = 0{\cdot}31$) indicates a false null hypothesis. The three-distribution model, on the other hand, places the 50 per cent cutoff at a t-value of 1·233 ($P = 0{\cdot}11$), which seems a more reasonable value. The decision between the two models is a tradeoff between model fit and cutoff point, so the final model is, in part, a matter of choice.

If we select the model with three non-null distributions, we would estimate 27·5 true null hypotheses, similar to Schweder and Spjøtvoll's[2] estimate of 25, but smaller than Duncan's[8] minimum estimate of 34.

The second data set consists of 78 correlation coefficients, based on a sample size of 45, originally published by Hill.[9] We used the standard Z-transformation to determine statistical significance. Since this produces an approximate normal variate, we modelled the distribution of the results as a mixture of normal deviates. We adjusted correlation coefficients reported as 0·00 by Hill to 0·005 to avoid P-values of 1·0, which would lead to one-sided normal deviate statistics of $-\infty$. Results appear in Table IV.

The model with one displaced normal distribution identifies a small population with highly significant results. Although it appears to provide an adequate fit to the data, we have a substantial improvement in the log-likelihood when we add an additional non-null distribution to the model. In the two-distribution model, we are identifying one very small, extremely significant population (two tests with displacement of over 10) and another population, less extreme but still highly significant, accounting for approximately 11 tests. This would imply that there are 65 null tests, compared with the estimates of 62 by Hill[9] and 64 by Schweder and Spjøtvoll.[2] If we add a third non-null distribution to the model, the estimated number of true null hypotheses drops to 63. Although the addition of the third distribution reduces the point where there is a 50 per cent chance that a test result represents a true null hypothesis from a normal deviate of 3·17 to 2·41, this would not rule out the three-distribution model. The choice between two or three non-null components in the model depends largely on whether one is comfortable with estimation of six parameters from only 78 values.

6. DISCUSSION

Different methods to assess the number of significant findings give a wide range of results. With total disregard of the multiple comparisons issue, the use of a conventional $\alpha = 0{\cdot}05$ leads to 567 significant results in the multiple cause of death data, while the Bonferroni approach finds only 54 results significant. The fit of a straight line to the P-plot leads to substantially higher estimates, that range from about 738 to 783. Our proposed approach leads to similar estimates, 749 significant results for the mixture of beta distributions and 804 significant results for the mixture of t-distributions, which better fits the data since it has a lower CVM statistic.

Both the approach developed here and that of Schweder and Spjøtvoll[2] can lead to a greater number of significant results than found with use of even an unadjusted P-value. Schweder and Spjøtvoll attribute this, in part, to low power of standard multiple comparison techniques (example 1, Section 3.2 of their paper). We believe that this apparently paradoxical result arises for another reason. Both our technique and P-plots estimate the number of true null hypotheses, independent of the statistical significance associated with each individual result. Thus, both methods identify results as significant (results coming from a true alternate hypothesis) even though the P-value may not indicate statistical significance, that is, there is insufficient evidence formally to reject the null hypothesis. As such, it may be appropriate to classify test results into three rather than two groups. Two of these groups would be similar to the conventional classification: an 'accepted null hypothesis' group for tests that both exceed the conventional rejection criterion and likely arise from the distribution of null hypotheses, and a 'rejected null hypothesis' group for tests that we reject with use of the conventional criterion of statistical significance and that likely come from an alternative hypothesis. We would classify the remaining test results into a third group ('undecided') that would include both tests that represent a true alternative hypothesis and those that represent a true null hypothesis, but for which the evidence in either case is less than overwhelming. For example, with application of our three t-distribution model to the MCD data, we could classify those tests with a t-value less than 0·607 as 'accepted null hypotheses' (with at least a 50 per cent chance that the test result arises from a test of a true null hypothesis), those tests with a t-value greater than 2·418 as 'rejected null hypothesis' (with at most a 5 per cent chance that the result arises from a test of a true null hypothesis), and classify the remaining tests as 'undecided'. This would avoid the dichotomy between 'statistically significant' and 'not statistically significant', and might have particular use in exploratory data analysis.

We can interpret the density functions (1) or (5) in an empirical Bayesian framework. In this interpretation, we use the density function as a prior distribution to obtain a posterior distribution function for each individual test result. We would expect this approach to pull test values closer to the mean of the major distribution likely to give rise to the specific test result. In particular, we would generally find extreme values pulled closer to the (smaller) mean of the distribution, which would reduce the tendency to overinterpret extreme findings. This approach has been applied using one null and one non-null distribution to the multiple comparisons generated in an epidemiologic study of multiple diseases and exposures.[10, 11]

The linear trends that we have examined present a complicated correlation pattern. For example, if there were strong secular trends caused by diagnostic fads or change in life style, then trends would likely exist across many age/race/sex combinations for each disease affected. In addition, we may find trends for one disease correlated, positively or negatively, with trends for another disease. Both factors could lead to clusters of similar effects across age/race/sex strata. As mentioned in Section 4.1, results do appear to cluster within individual diseases, demonstrated by heterogeneity across diseases when we classify test results by probable distribution.

To use our approach, one must make a decision as to the number of distributions to include in the mixture. Although the simulation of the empirical distribution of the CVM statistic provides an objective framework for this decision, our examples illustrate that one cannot make the decision automatically. Moreover, the method is computationally intensive. As there is no theoretically correct way to determine the appropriate number of distributions to include in the model, however, it appears that this simulation to assess fit is highly desirable. As observed in our examples, however, one selects similar models if one continues adding components until the calculated CVM statistic that compares the model and the data becomes stable. Although not a formal test of the adequacy of fit, this might well suffice when computer resources to simulate the empirical distribution of the CVM statistic are unavailable.

Our method describes the underlying distribution of P-values (using the generic approach with beta distributions) or of the test statistics (using the specific approach). The generic approach can be applied to any collection of P-values, while the specific approach should only be used when all test results arise from the same underlying distribution. Both models permit an evaluation, not available with other techniques, of the probability that any given result actually represents a true null hypothesis.

ACKNOWLEDGEMENTS

This work was partially supported by NIH grants HHS 2M01 RR00095 and HL 14214.

REFERENCES

1. Miller, R. G. *Simultaneous Statistical Inference*, Springer-Verlag, New York, 1981.
2. Schweder, T. and Spjøtvoll, E. 'Plots of P-values to evaluate many tests simultaneously', *Biometrika*, **69**, 493–502 (1982).
3. Titterington, D. M., Smith, A. F. M. and Makov, U. E. *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester, 1985.
4. Diaconis, P. and Ylvisaker, D. 'Quantifying prior opinion', *in* Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.) *Bayesian Statistics 2*, Elsevier, New York, 1985, pp. 133–156.
5. O'Neill, R. 'Function minimization using a simplex procedure' *in* Griffiths, P. and Hill, I. D. (eds.) *Applied Statistics Algorithms*, Royal Statistical Society, London, 1985.
6. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974, p. 281.
7. Pearson, E. S. and Hartley, H. O. (eds.). *Biometrika Tables for Statisticians: Volume II*, Cambridge University Press, Cambridge, 1972, p. 359.
8. Duncan, D. B. 'A Bayesian approach to multiple comparisons', *Technometrics*, **7**, 171–222 (1965).
9. Hill, M. 'On looking at large correlation matrices', *Biometrika*, **56**, 249–253 (1969).
10. Thomas, D. C. 'The problem of multiple inference in identifying point-source environmental hazards', *Environmental Health Perspectives*, **62**, 407–414 (1985).
11. Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M. and Armstrong, B. G. 'The problem of multiple inference in studies designed to generate hypotheses', *American Journal of Epidemiology*, **122**, 1080–1095 (1985).