Journal of Data Analysis & Operations Research
ISSN 2384-4671
J. Data Ana. Oper. Res. 1, No. 1 (Jun. 2014)

Approximating the Distribution of Combined Dependent P-values from Multiple Experiments

Mehdi Razzaghi

Bloomsburg University, USA

Received: 26 Apr. 2013, Revised: 26 Sept. 2013; Accepted: 30 Dec. 2013, Published: Jun. 2014

Abstract: Observed significance levels or what is commonly known as p-values are routinely used to summarize the results of hypotheses tests in statistical inference. In many experiments, however, it is necessary to summarize results of multiple tests and combine several p-values to arrive at a single observed significance level for the entire experiment. When these p-values are assumed independent, then the classical approach to combining p-values is rather straight forward and by applying a simple transformation, the percentiles of chi-square distribution are used to determine the observed significance level for the entire experiment. However, when the p-values from multiple tests are not independent, the problem is more complex and no closed form distribution can be derived. For such cases, one approach is to use a shifted chi-square distribution to approximate the distribution of the combined p-values. Method of moments has traditionally been utilized to estimate the shift parameter as well as the number of degrees of freedom. Here, we propose a method for estimating these parameters that does not only rely on matching the first two moments. Rather, our method is based on minimizing the distance between the moment generating functions for a chosen set of moments. An approximation for the moment generating function of the distribution of the combined p-values is derived and is used to determine the parameter estimates. The properties of the estimates are discussed and an example is used to illustrate the methodology.

Keywords: Combined p-values, Dependent tests, Moment generating function estimator, Scaled λ^2 approximation

1. Introduction

In many biological experiments it is desired to combine the p-values from multiple tests in order to derive an overall observed significance level for the experiment. For example, in generic experiments with microarrays when thousands of genes are compared simultaneously, often a DNA array is designed to compare two treatments. For each experimental unit, the change in expression levels of genes are tested for significance, and in order to derive an overall p-value for the treatment effect, a combined p-value from all experiments has to be calculated. Fisher (1932) was probably the first to consider this problem. He argued that under the null hypothesis, the p-values from the test statistics with continuous nulls have uniform distributions. Therefore, if $P_1, P_2, \ldots, P_k$ are p-values from k experiments, then assuming independence,

$$Q = -\sum_{i=1}^{k} 2 \ln P_i \qquad (1)$$

has a χ^2 distribution with $2k$ degrees of freedom. Littell and Folks (1971, 1973) showed that Fisher's method for combining p-values is asymptotically optimal among all methods of combining independent tests and Pallini (1994) proved that Fisher's test is Bahadur optimal. A closely related method for combining p-values that has also found some popularity was suggested by Liptak (1958). His method is based on using

$$U = \sum_{t=1}^{k} \Phi^{-1}(1 - P_t)$$

which has been shown to have about the same power of the Fisher's test. Following Fisher, many authors considered different variation of the problem, proposing different procedures. Here we refer to Won et al (2009) and the references therein.

A major shortcoming of the above methods, however, is the assumption of independence of $P_1, \ldots P_k$. There are many experiments in which the multiple tests are not jointly independent. Because of

E-mail: mrazzagh@bloomu.edu

the significant application of this problem in genetics, several authors have recently suggested various modifications of the tests used for independent p-values to include the dependent case. For example, Zaykin et al. (2002) propose a method based on using only p-values below a certain threshold and define a transformation that changes the vector of correlated p-values to uncorrelated values. Similarly, Delongchamp et al. (2006) propose an extension of Liptak's method that uses an estimate of the correlation matrix. However, the simplest and probably the most practical extension is given by Brown (1975). Assuming a multivariate normal distribution with a known covariance matrix for the test statistic, Brown (1975) proposed approximating the distribution of Fisher's statistic Q by a scaled χ^2 distribution. Kost et al. (2002) extended this results to the case when the covariance matrix of $P_1, \dots P_n$ is known up to a scalar quantity. To estimate the scale factor and the number of degrees of freedom, these authors suggest the moment estimators. They propose to set the first two moments of Q equal to the first two moments of the scaled χ^2 distribution to derive the parameter estimates. This method of estimation is clearly too simplistic and although it leads to consistent estimates, the resulting estimators may be highly biased and not efficient. Here, we introduce a method for estimating the parameters of the scaled χ^2 distribution that utilizes several moments including fractional moments. Our method is based on minimizing the sum of squares of differences between the moment generating function $M_Q(t)$ of Q and that of a scaled χ^2 distribution for a chosen set of values of t. The moment generating technique for parameter estimation was first suggested by Quandt and Ramsey (1978) where the authors derive estimates for the parameters of a normal mixture by minimizing the sum of squares of differences between the theoretical and sample moment generating functions. In the next section we outline our method of estimation.

2. Moment Generating Function of Q

As described in Brown (1975) and Kost et al. (2002), we let the random vector $(X_1, X_2, \dots, X_k)^T$ have a joint multivariate normal distribution with mean vector $\mu = (\mu_1, \dots, \mu_k)^T$ and a nonsingular covariance matrix. It is desired to test the hypotheses $H_0: \mu_i = \mu_{i_0}$ against the alternatives $H_i: \mu_i < \mu_{i_0}$ for at least one i where $i = 1, \dots, k$. Then the p-value from each test is given by:

$$P_i = \Phi\left(\frac{X_i - \mu_{i_0}}{\sigma_i}\right) \qquad i = 1, \dots, k$$

where $\Phi(.)$ denotes the cdf of a standard normal distribution and σ_i^2 is the ith diagonal element of the covariance matrix. Now if $P_1, P_2, \dots, P_k$ are independent, then the distribution of Q given in (1) is χ^2 with $2k$ degrees of freedom. However, if the p-values from the multiple tests are not independent, then we have:

$$E(Q) = 2k \tag{2}$$

and

$$V(Q) = 4k + \sum_{i<j}\sum cov\,(-2\ln p_i\,, -2\ln p_i) \tag{3}$$

Brown (1975) suggests approximating the distribution of Q by $a\chi_v^2$ and determines estimates of a and v such that the first two moments of the scaled χ^2 distribution match up with the corresponding moments of Q. This procedure clearly works only when the covariance matrix is completely specified, so that (3) is computable. Brown (1975) notes that covariance between $-2\log P_i$ and $-2\log P_j$ only depends on the correlation between the ith and jth variable and derives empirical quadratic equations to calculate approximations of the covariances. When the covariance matrix is known only up to a scalar constant, Kost et al (2002) show that the p-value from each test has a t-distribution. Using this property, the distribution of Q is again approximated by a scaled χ^2 estimating the parameters using the first two moments as in Brown (1975).

Now, note that the moment generating function of Q is given by:

$$\begin{aligned} M_Q(t) = E[e^{Qt}] &= E\left[e^{-\left(\sum_{i=1}^k 2\ln P_i\right)t}\right] \\ &= E\left[e^{\sum_{i=1}^k \ln p_i^{-2t}}\right] \\ &= E[P_1 P_2 \dots P_k\,]^{-2t} \end{aligned} \tag{4}$$

Let $W_k = P_1 \dots P_k$, then Bailey and Gribskov (1998) prove that when $P_1, \dots, P_k$ are independent, then the cumulative distribution of W_k is given by

$$\begin{aligned} F_{W_k}(u) &= P(W_k \le u) \\ &= u\sum_{i=0}^{k-1}\frac{(-\ln u^x)^i}{i!} \end{aligned}$$

and based on this, in a follow up paper, Bailey and Grundy (1998) propose approximating the cumulative distribution of W_k when $P_1, \dots, P_k$ are not independent by

$$F_{W_k}(u) = P(W_k \le u)$$
$$\cong u^x \sum_{i=0}^{[m]-1} \frac{(-\ln u^x)^i}{i!}$$
$$+ u^x(m-[m])\frac{(-\ln u^x)^{[m]}}{[m]!}. \qquad (5)$$

Here $1 \le m \le k$ is what Bailey and Grundy (1999) call the "effective size" of the sequence family of the dependent p-values. It is estimated by minimizing the root mean square error function that gives equal weight to equal proportional errors in the observed and predicted p-values. More formally, m is a value that minimizes

$$E(m) = \left\{\sum_{i=1}^{k}\left[\ln\left(P_i(m)\right) - \ln\left(\frac{i}{k+1}\right)\right]^2\right\}^{1/2} \qquad (6)$$

where $P_i(m)$ is the ith largest p-value among matches between the given family and k random query sequences. Also, in (5), $x = \frac{m}{k}$ and $[.]$ denotes the integer part function. Using the approximation in (5) and taking its derivative, we get an approximation for the density of W_k as

$$f(u) = \frac{d}{dt}F_{w_k}(u)$$
$$= xu^{x-1}\sum_{i=0}^{[m]-1}\frac{(-\ln u^x)^i}{i!}$$
$$- xu^{x-1}\sum_{i=1}^{[m]-2}\frac{(-\ln u^x)^i}{i!}$$
$$+xu^{x-1}(m-[m])\frac{(-\ln u^x)^{[m]}}{[m]!}$$
$$+u^x(m-[m])\frac{\left(-\frac{x}{u}\right)(-\ln[\![u^x)]\!]^{[m]-1}}{([m]-1)!}$$
$$= xu^{x-1}\frac{(-\ln u^x)^{[m]-1}}{([m]-1)!}$$
$$+ xu^{x-1}\frac{(m-[m])(-\ln u^x)^{[m]}}{[m]!}$$
$$- xu^{x-1}\frac{(m-[m])(-\ln u^x)^{[m]-1}}{([m]-1)!}$$
$$= xu^{x-1}\frac{(-\ln u^x)^{[m]-1}}{([m]-1)!}\left[([m+1]-m)\right.$$
$$\left.-\frac{m-[m]}{[m]}(-\ln u^x)\right] \qquad 0 \le t \le 1$$

(7)

Then, the sth moment of w_k is given by

$$\mu_s' = E(w_k^s) = \int_0^1 u^s f(u)du$$
$$= \frac{x([m+1]-m)}{([m]-1)!}\int_0^1 u^{x+s-1}(-\ln u^x)^{[m]-1}\,du$$
$$+\frac{x(m-[m])}{[m]!}\int_0^1 u^{x+s-1}(-\ln u^x)^{[m]}\,du$$

which after successive integration by parts yields

$$\mu_s' = \left(\frac{x}{x+s}\right)^{[m]}([m+1]-m) +$$
$$\left(\frac{x}{x+s}\right)^{[m]+1}(m-[m]) \qquad (8)$$

Note that $\lim_{s\to 0}\mu_s' = 1$. Therefore from (4), the moment generating function of Q is given by

$$M_Q(t) = \left(\frac{x}{x-2t}\right)^{[m]}([m+1]-m) +$$
$$\left(\frac{x}{x-2t}\right)^{[m]+1}(m-[m]) \qquad (9)$$

3. Parameter Estimation

Now, for a scaled χ^2 random variable V with a scale factor a and v degrees of freedom, we have

$$M_V(t) = E[e^{aVt}] = (1-2at)^{-v/2} \qquad (10)$$

Rather than matching the first two moments of the scaled χ^2 distribution with the first two sample moments of Q, our method of estimation of a and v is based on selecting a suitable set of values $t_1, t_2, \dots, t_n$ and minimizing

$$W = \sum_{i=1}^{n}\{M_Q(t_i) - M_V(t_i)\}^2 \tag{11}$$

Substituting from (10) and setting the derivative of W with respect to a and v equal to zero, we get the two normal equations

$$\sum_{i=0}^{n} t_i\ (1-2at_i)^{-\frac{v+2}{2}}\{M_Q(t_i) - M_V(t_i)\} = 0 \tag{12}$$

And

$$\sum_{i=1}^{n} \ln(1-2at_i)\ M_V(t_i)\{M_Q(t_i) - M_V(t_i)\} = 0 \tag{13}$$

whose roots constitute the estimates of a and v.

4. Properties of the Estimates

We first note that as $m \to k$, then $x \to 1$ and from (8) we have

$$\mu_s' = \frac{1}{(1+s)^k}$$

which is exactly the sth moment of W_k when $P_1, P_2, \dots, P_k$ are independent. On the other extreme, as $m \to 1$, then $x \to \frac{1}{k}$ and we have

$$\mu_s' = \frac{1}{1+ks}$$

which precisely is the expression for the sth moment of W_k when all of the $P_1, P_2, \dots, P_k$ are identical. This shows that (8) provides a reasonable approximation for the moments of w_k. Now, from the general properties of least square estimates, we deduce that $\hat{\theta}_1$and $\hat{\theta}_2$are unbiased with standard expressions for the variance of estimates. This in turn means that estimates of a and v are asymptotically unbiased and consistent.

5. Example

To illustrate the methodology described in this paper, here we provide an example. For the purpose of comparison, we use the same example described in Brown (1975). The data set due to Lifshitz et al (1971) pertains to a survey of constituents of citrus fruits and contains 58 samples of pure lemon juice. The samples are examined for a variety of ingredients. From the sample data, Brown (1975) uses the correlation matrix from four highly correlated ingredients Aspartic Acid, Glutamic Acid, Glycine and Alanine to calculate the variance of Q. Applying the method of moments, the parameter estimates for the approximating shifted χ^2 distribution are given as 0.87 and 4.6 respectively for the shift and the number of degrees of freedom. A hypothetical new sample is used to derive p-values for one-sided t tests for each amino acid separately. Now, here, in order to estimate the value of the effective size m, we utilized the algorithm and the database developed by Grundy (1998) based on a large number of different family sizes. Using an interval $[-B, B]$ around the origin 0, with step sizes of .01for the t-values, we found the estimate of the effective size m when the family size is 4, i.e. four tests are being applied for several values of B. For each estimated value of m, we used (13) to derive estimates for α and ν the shifting parameter and the number of degrees of freedom respectively. Table1 displays the estimates of these parameters with their respective standard deviations for some selected values of B. From Table 1 we see that there is some variation in the parameter estimates depending on the selected values of t. However, this variation does not appear to be drastic and in practice any of the estimates can provide a reasonable result. For example, if we take the estimates derived based on B=4, then the distribution of Q is approximated by a $3\chi^2$ variable with 5.09 degrees of freedom. Hence, similar to Brown (1975), we find that the hypothesis of no adulteration is rejected.

Table1: Estimates of the effect size, shift parameter and number of degrees of freedom for some selected B.

B	m $means \pm sd$	α $means \pm sd$	ν $means \pm sd$
2	3.09 ± 0.71	3.92 ± 0.41	6.30 ± 0.75
3	3.18 ± 0.66	3.33 ± 0.58	5.53 ± 0.58
4	3.09 ± 0.70	3.00 ± 0.25	5.09 ± 0.50

6. The Choice of Moments

The challenging question that remains to be answered is how one should choose the set of moments to be utilized for the parameter estimation i.e. how to select $t_1, t_2, \dots, t_n$. Clearly these values should be chosen in such a way that the corresponding normal equations i.e. (12) and (13) are nonsingular.

Consequently, values that are very close to zero or very large making $M_V(t_i)$ computationally intractable should be avoided. The second more difficult question is whether we can determine a set of values of the argument of the moment generating function $t_1, t_2, \dots, t_n$ that leads to the most efficient estimates. In our computation, we found that although no set rules can be derived for choosing an optimal set of t-values in (13), it is believed that there is little gain when more than 20 to 30 values are used. Also, for values higher than 4 or lower than -4, the computation can become unstable. The choice of values of t is a critical one and further research is needed to determine if an optimal set can be obtained.

7. Discussion

The problem of combining the results of multiple tests has, in recent years, attracted the attention of many researchers. This is partly due to the growth of microarray technology in which gene expressions are examined simultaneously. A primary goal of microarray experiments is to identify genes whose expression levels differ between various classes of samples. Although several methods have been introduced for combining p-values from multiple experiments (see for example Liptak, 1958 and George and Mudholkav, 1983) Fisher's method is still considered simple and very practical. Recently, Won et al (2009) used simulation to make a comparison between different methodologies. In Fisher's derivation, however, it is assumed that the p-values from multiple experiments are independent, an assumption that in many cases may not be true. Brown (1975) introduced a method for approximating the distribution of the Fisher's statistic. His method uses the first two moments to estimate the parameters of the approximating distribution. Here, we have presented an approach for improving this approximation. Our methodology is based on estimating the parameters of the approximate scaled χ^2 distribution in such a way that the distance between the moment generating function of the combined p-values $M_Q(t)$ and the moment generating function of the scaled χ^2 distribution is minimized for a chosen set of values of the associated arguments. This in turn means that a chosen set of moments including fractional moments is used for estimating the parameters. Clearly, by applying several moments, more efficient parameter estimates can be expected. The quality of estimates may improve if a weighted sum of squares with weights inversely proportional to $Var\left(M_V(t_i), M_V(t_j)\right)$ in (11) is used. But, that problem requires further investigation.

Acknowledgement

I am grateful to Dr. Dong Zhang for his computational assistance.

References

Bailey, T. L., & Gribskov, M. (1998). Combining evidence using p-values: applicationto sequence homology searches. *Bioinformatics14*, 48 – 54.

Bailey, T.L., & Grunley, W.N. (1999). *Classifying proteins by family using the product of correlated p-values.* Procedures of the Third International Conference on Computational Molecular Biology, Lyon, France.

Brown, M.B. (1975). A method for combining non-independent one-sided tests of significance.Biometrics 31, 987- 992.

Delongchamp, R., Lee, T. , & Velasco, C. (2006). A method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics, 7 (Suppl. 2), 1 – 9.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*. Fourth ed., Oliver and Boyd, London.

George, E.O. and Mudholkar, G.S. (1983). On the convolution of logistic random variables. *Metrika, 30*, 1-14.

Grundy, W. N. (1998). Homology detection via family pairwise search. *Journal of Computational Biology, 5*, 479 – 491.

Kost, J. T., & McDermott, M. P. (2002). Combining dependent p-values. *Statistics and Probability Letters 60*, 183-190.

Lifshitz, A. and Stepak, Y. (1971). Detection of adulteration of fruit juice. *Journal of the Association of Official Analytical Chemists 54*, 1262 – 1265.

Liptak, T. (1958). On the combination of independent tests. *Maggar Tudom Ãanyos Akad ÃemiaMatematikai Kutat ÃO Intezetenek Kozlemeneyei 3*, 171-197.

Littel, R.C., & Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combiningindependent tests. *Journal of the American Statistical Association 66*, 802-806.

Little, R.C., & Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association 68*, 193-194.

Pallini, A. (1994). Bahadur exact slopes for a class of combinations of dependent tests. *Metron 52*, 53 – 65.

Won, S., Morris, N., Lu, Q. and Elston, R. C. (2009). Choosing an optimal method to combine p-values. *Statistics in Medicine* 28, 1537-1553.

Zaykin, D. V., Zhivotovsky, L.A., Westfall, P.H. , & Weir, B.S. (2002). *Truncated product method for combining p-values*.