On the Analysis of Large Numbers of \( p \)-values

David R. Cox\(^1\) and Christiana Kartsonaki\(^2\)

\(^1\)Nuffield College, University of Oxford, Oxford OX1 1NF, UK
\(^2\)MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
E-mail: david.cox@nuffield.ox.ac.uk

Summary

Methods for the analysis of large numbers of \( p \)-values, observed levels of significance, are reviewed with some emphasis placed on the Rényi decomposition hinging on the relation with independent exponentially distributed random variables. Some extensions are described. An empirical example is used in illustration, and simulation results examining potential complications are outlined.

Key words: Exponential distribution; false discovery rate; graphical methods; Rényi decomposition; significance test; spacings.

1 Introduction

Even the role of single significance tests has evoked controversy, largely or perhaps totally resolved by distinguishing at least four possible objectives. These are decision-making, assessment of consistency with a precisely specified null hypothesis, assessment of the direction of an effect and checks for misspecification of the primary model. See, for example, Cox and Donnelly (2011, Section 8.4). When in effect a number or indeed many hypotheses are involved together, several further issues arise. First, the primary question may concern whether there is evidence for rejecting the global null hypothesis that all subhypotheses hold simultaneously. This is addressed in the concept of experiment-wise error rate (Benjamini & Hochberg, 1995) assessing the possibility that the global null hypothesis is falsely rejected when true and, when rejection is appropriate, on specifying the collection of component hypotheses to be rejected. The levels of evidence provided against them individually may be very different, but such a level of detail is often not specified. A second formulation, emphasised in the present paper, puts more emphasis on assessing individually the levels of evidence against each rejected component hypothesis.

That is, there may be overwhelming evidence against some components and rather marginal evidence against others. This can be expressed in various essentially equivalent ways which we now outline.

The starting point of much formal discussion is that if \( U \) is a continuous random variable with a uniform distribution on \((0, 1)\), then \( Z = -\log U \) has an exponential distribution of unit mean. Equivalently, \( 2Z \) has a chi-squared distribution with two degrees of freedom. Here, \( U \) refers typically to the tail area or \( p \)-value of a single test defined through the probability under
the null hypothesis of a deviation as or more extreme than that observed. There is a formal difficulty in applications when discreteness of the distribution under the null hypothesis may need consideration. The discreteness could be resolved by an element of imposed randomisation, although this is typically not to be recommended in applications.

For formal considerations in the nonnull case, it is simplest to assume that $Z$ has an exponential distribution with mean $1/\gamma$ which is equivalent to supposing that $U$ has density $\gamma u^{\gamma - 1}$.

The present paper concentrates on issues where, $n$, possibly large, independent tests are to be assessed simultaneously. To test the corresponding global null hypothesis, the combined statistic $-2\Sigma \log p_j$ may be used, having under the global null hypothesis a chi-squared distribution with $2n$ degrees of freedom (Fisher, 1924). This will, however, be insensitive if $n$ is large and only a few of the component null hypotheses do not hold.

More detailed analysis of sets of $p$-values was considered by Schweder and Spjøtvoll (1982) who examined the shape of the empirical distribution of $p$ near the origin. Interpretation, formal or informal, especially graphical interpretation, is likely to be easier if the region of interest is not tightly concentrated at one end. We therefore consider two different although arithmetically equivalent methods as a basis for interpretation. One is a normal probability plot, relating the ordered values of $\Phi^{-1}(1 - p_j)$ to the expected normal order statistics in samples of size $n$; for the $r$th largest a close approximation to that expected value is (Blom, 1958)

$$\Phi^{-1}\left(\frac{r - 3/8}{n + 1/4}\right).$$

The task is then to interpret correctly any clear departures from a straight line through the origin and of slope one. For example, if the points lie close to a line through the origin of slope greater than one, there may be many real departures present or possibly the sampling error of the test statistics has been systematically underestimated. Formal tests essentially based on this graphical procedure do not seem to have been explored in detail but would be obtained from the joint distribution of a number of relatively extreme order statistics.

In most cases, the emphasis is on the identification and interpretation of those individual $p$-values that signal a real effect by being extreme points in the plot. The formal properties of the sequence of $p$-values are, however, often best assessed from the Rényi decomposition of order statistics (Rényi, 1953). This stems from the notion that the order statistics of a random sample taken in sequence form a Markov process. The present discussion concentrates largely on the use, formal and informal, of the Rényi decomposition as an aid, often somewhat informal, to interpretation. Such use is to be contrasted with the application of formal decision rules to achieve prespecified properties.

2 The Rényi Decomposition

For a more formal probabilistic discussion, it is convenient to change the notation. If $U$, representing a $p$-value, is uniformly distributed on $(0, 1)$, then $Z = -\log U$ has an exponential distribution of mean one. Denote the ordered values from a set of $n$ independent copies of $Z$ by

$$Z_1 \leq Z_2 \leq \ldots \leq Z_{n-1} \leq Z_n.$$

Then, when all $p$-values are independently uniformly distributed on $(0, 1)$,

$$Z_1 = V_1/n, \quad Z_2 = V_1/n + V_2/(n - 1), \quad \ldots, \quad Z_{n-r} = V_1/n + \ldots + V_{n-r}/(r + 1), \ldots, \quad Z_n = V_1/n + V_2/(n - 1) + \ldots + V_{n-1}/2 + V_n,$$
where the normalised spacings $V$ are independently exponentially distributed with mean one. Note that $V_{n-r} = (r + 1)(Z_{n-r} - Z_{n-r-1})$.

This suggests plotting the ordered values of $Z$ against the corresponding expected values, namely,

$$1/n, 1/n + 1/(n-1), \ldots, 1/n + 1/(n-1) + \ldots + 1/2,$$

and finally the value for comparison with max $(Z_j)$, namely,

$$1/n + 1/(n-1) + \ldots + 1/2 + 1.$$

A close approximation to this for large $n$ is $\log n + \gamma$, where Euler’s constant, $\gamma$, is 0.5772.

Again, the simplest possibility is that the points lie on a straight line of slope one. A single suspiciously large observation would correspond to a large value, $z_{\text{max}}$ of $Z_n$, the smallest $p$-value, and

$$P(Z_n \geq z_{\text{max}}) = 1 - (1 - e^{-z_{\text{max}}})^n = 1 - (1 - u_{\text{min}})^n,$$

a standard allowance for selecting the most significant effect out of $n$ independent such tests.

A number of additional aspects have to be considered, especially when $n$ is large. We now consider some possibilities.

### 3 Some Detailed Development

One relatively simple possibility is to examine the $p$-values in order starting with the smallest and stopping when a value is encountered not significant at some assigned level. At each point, we use the test for the most extreme $p$-value, in principle progressively reducing the value of $n$. The resulting confidence set of effects has the formal property that in the context where there is a set of clearly established effects followed by only null population values, the last value reported as nonnull will be wrongly included by chance only with the specified probability.

A rather different possibility is to consider collectively the group of smallest $p$-values. For this, we work with the Rényi representation. That is, we have for the $r$th largest order statistic of the random variables $Z = -\log p$ the representation $V_{n-r}/(r + 1) + \ldots + V_1/n$, where the $V_j$ are in the idealised null case independently exponentially distributed with mean one. Thus, we may test the overall null hypothesis using

$$T_q = Z_n + \ldots + Z_{n-q},$$

where $q$ is a small integer chosen to capture any appreciable effects potentially present. By expressing this in terms of the random variables $V$, the distribution under the global null hypothesis can be determined, for example, by first calculating its moment generating function. A simple approximation may be based on

$$E(T_q) = q + 1 + (q + 1)\{1/(q + 2) + \ldots + 1/n\},$$

$$\text{var}(T_q) = q + 1 + (q + 1)^2\{1/(q + 2)^2 + \ldots + 1/n^2\},$$

treating $T_q$ as distributed proportionally to a chi-squared variable with constant of proportionality and degrees of freedom chosen to match the mean and variance. That is, if we suppose that approximately $T_q = a \chi^2_b$ where $a$ and the degrees of freedom $b$ are to be chosen to agree with the above values of mean and variance, then $a = \text{var}(T_q)/\{2E(T_q)\}$ and
If, as is likely to be the case, \( q \) is modest and \( n \) large, we may use the following approximations:

\[
E(T_q) = (q + 1) \gamma + 1 + \log n - s_{1,q+1},
\]

\[
\text{var}(T_q) = (q + 1) \{1 + (q + 1)(\pi^2/6 - s_{2,q+1})\}.
\]

Here, again, \( \gamma \) is Euler’s constant and for \( r = 1, 2 \)

\[
s_{r,q+1} = 1 + 1/2^r + \ldots + 1/(q + 1)^r.
\]

A formally simpler approach is to base the test not on the directly observed values of \( Z \), that is, on the \( p \)-values, but on the derived variables \( V_j \); this might have some statistical advantage in that any anomalous behaviour in the centre of the distribution of \( p \)-values would be nullified. For this,

\[
M_q = (Z_n + \ldots + Z_{n-q}) - (q + 1)Z_{n-q-1} = V_n + \ldots + V_{n-q},
\]

where when the global null hypothesis holds, \( 2M_q \) has a chi-squared distribution with \( 2q + 2 \) degrees of freedom. A suitable choice of \( q \) has to be made on subject-matter grounds and in practice usually a sensitivity analysis concerning the choice of \( q \) will be needed.

In interpreting this and similar statements, it is important that the same ordered value, \( Z_n \), corresponds in general in hypothetical repetitions to differing contrasts in the underlying population. Any subject-matter interpretation of anomalous \( p \)-values would, however, apply to the specific contrasts involved.

### 4 Departure from Standard Conditions

When \( n \) is large, so that even under the null hypothesis the \( p \)-values of interest are in the extreme tails of the distribution, it is unlikely that probabilities computed by standard methods under the null hypothesis really have a reliable direct interpretation. One way of addressing this issue is to assume that under the null hypothesis a smooth function of the observed \( p \) has a uniform distribution or equivalently that a smooth function of the observed \( Z \) has the Rényi structure. If further we make the rather strong assumption that the latter function can be taken as locally linear in the upper tail we may write the observed values of \( Z = -\log p \) in the upper tail, \( r < r_0 \), for a suitable value of \( r_0 \)

\[
Z_{n-r} = \eta_0 + \eta_1 \{\Omega_1/n + \ldots + \Omega_{n-r}/(r + 1)\},
\]

where under the null hypothesis the \( V \), which were observed via the \( Z \), are replaced by the \( \Omega \), which are not observed, but are again independently exponentially distributed with unit mean. Here, the \( \eta \)'s are unknown parameters capturing smooth departure from the standard assumptions. If we define, as before, \( V_{n-r} = (r + 1)(Z_{n-r} - Z_{n-r-1}) \), then \( V_{(n-r)} = \eta_1 \Omega_{n-r} \) and so has an exponential distribution now of unknown mean \( \eta_1 \). To apply this result, we need first to choose \( q \) in order to define the set of possibly extreme values through

\[
\tilde{M}_q = V_n + \ldots + V_{n-q+1},
\]

and then a suitable base level for comparison

\[
\tilde{D}_{r,r_0} = V_{n-r_0} + \ldots + V_{n-r_0-r}.
\]
Then, under an overall null hypothesis,

\[(\bar{M}_q / q)(\tilde{D}_{r,r_0} / r)^{-1}\]

has the variance ratio distribution with degrees of freedom \((2q, 2r)\). Testing simply the most extreme value corresponds to taking \(q = 1\) when \(\bar{M}_1 / (\tilde{D}_{r,r_0} / r)^{-1}\) has the degrees of freedom \((2, 2r)\). The procedure requires the somewhat arbitrary specification of \((r, r_0)\) but does provide some security against the quite realistic possibility that there is nonstandard behaviour of the extreme tail of the distribution under the null hypothesis. Essentially, the denominator suggests an empirical recalling of the values of \(\log p\).

Note that

\[
\bar{M}_q = Z_n + \ldots + Z_{n-q+1} - qZ_{n-q-2}, \\
\tilde{D}_{r,r_0} = (r_0 + 1)Z_{n-r_0} + Z_{n-r_0-1} + \ldots + Z_{n-r_0+r+1} - (r_0 + r - 1)Z_{n-r_0-r}.
\]

5 Some Further Nonstandard Conditions

A further possibility, especially when the tests are based on standard normal-theory tests using an estimated variance, is that, even though the formal number of degrees of freedom for error may be large, the error variance is in fact wrongly estimated, often, but not necessarily, underestimated. Then, the standard normal probability plot will be linear but not with unit slope.

Local essentially positive serial correlation in the statistics generating the \(p\)-values decreases the sample variance and hence is likely to reduce the effective sample size appropriate for interpreting the set of values. That is, use of the procedures for independent observations is likely to overestimate the significance of extreme values.

Sometimes the test statistics are based on samples of different sizes. For modest amounts of data, no special issue is involved. When, however, very extreme values of \(p\) are involved, the objective may really be more one of scoring the magnitude of an underlying effect in a way to help comparability between analyses with different sample sizes, \(n_0\) and \(n_1\), say. Then, if the underlying estimate of the notional effect remains the same, we have for some value of \(t\) that

\[p_i = \Phi(-t / \sqrt{n_i}).\]

That is, if we wish to convert \(p_0\) calculated from a sample of size \(n_0\) to a notional \(p\)-value at sample size \(n_1\), we write

\[p_1 = \Phi\left\{\frac{\sqrt{n_0}}{\sqrt{n_1}} \Phi^{-1}(p_0)\right\}.
\]

It would, however, be better in principle to reformulate the analysis directly in terms of estimation of a well-defined parameter.

6 Some Developments

The development above can be modified in various ways. In particular,

- it is important to distinguish a number of different objectives

  1 assessing the possibility that all null hypotheses are simultaneously adequate, that is that there are no ‘signals’ of real effects in the whole data
identifying a quite small number of appreciable real effects in a situation in which it is clear that some real effects are present. Individual assessment of the uncertainties of the identified effects may be required.

3 establishing the presence of an appreciable number of individually small effects whose combined impact is appreciable but whose individual impact may be small.

- the impact of the specific assumptions involved in calculating $p$-values needs to be assessed.

1 calculation of extreme values of $p$, say less than $10^{-3}$, typically depends on unrealistic assumptions, although the answer may be relatively more secure if based on a permutation distribution underpinned by randomisation in design.

2 the effect of statistical misspecification on the $p$-values may need study, in particular poor estimation of error variances or the distributional form of test statistics, and the possible effect of internal correlations between different $p$-values.

3 if, as sometimes happens, different $p$-values correspond to seriously differing sample sizes, some allowance for that is needed.

Throughout, we suppose that the emphasis is on the identification and interpretation of individual effects, not on empirical prediction of the response to be anticipated on a new individual.

7 An Empirical Example

To illustrate these methods, we use the obstructive sleep apnea (OSA) data of Gharib et al. (2013). OSA is associated with metabolic dysregulation and systemic inflammation. Gene expression was measured in visceral fat samples from individuals with OSA and healthy controls, aiming to provide insight into the effect of OSA on visceral fat at the molecular level.

The dataset consists of 33,297 gene expression levels from visceral fat samples of 8 OSA patients and 10 controls. The aim is to identify a subset of genes associated with OSA. The data

![Normal probability plot for obstructive sleep apnea data.](wileyonlinelibrary.com)
were first preprocessed by standardising each gene expression variable to have unit variance. A two-sample $t$-test for each gene provided $p$-values to be used in normal probability (Figure 1) and Rényi (Figure 2) plots.

The top 9 genes listed in Table 1 stand out for study and discussion; we will not attempt to interpret them in detail. Their isolation is clear in the normal probability plot and even more so in the Rényi plot. In the latter, there is also a suggestion of a further group of roughly 40 genes above the expected unit line. This might represent a further group of genes with modest effect but also the possibility of slight departure from standard normal-theory assumptions in the tail of the sampling distribution.

We can assess these possibilities rather more formally as follows. For the simplest primary analysis, we take the statistic $T_9$, a monotone function of the product of the nine smallest $p$-values. This gives the value $T_9 = 39.66$, as compared with an expectation and standard error of 90.6 and 4.42. A more cautious approach allows for the possibility that null behaviour in the tails departs from standard normal-theory form, and for this, we compare the most extreme points with those to be expected from the local behaviour in the slightly less extreme tail by using $(\hat{M}_q/q)(\hat{D}_{r,r_0}/r)^{-1}$ with $q = 9$, $r_0 = 10$, $r = 20$, that is comparing the slope with that of a slightly earlier part of the plot. The resulting ratio is 2.26, to be compared with the variance

![Figure 2. Rényi plot for obstructive sleep apnea data. [Colour figure can be viewed at wileyonlinelibrary.com]](image)

| Gene name | Description | $p$       |
|-----------|-------------|-----------|
| MFAP3L    | microfibrillar associated protein 3 like | $5.5 \times 10^{-6}$ |
| TRIP6     | thyroid hormone receptor interactor 6 | $2.2 \times 10^{-5}$ |
| HPF1      | histone PARylation factor 1 | $2.3 \times 10^{-5}$ |
| chr4:170712371-170712853 |          | $2.4 \times 10^{-5}$ |
| ZAN       | zonadhesin (gene/pseudogene) | $3.5 \times 10^{-5}$ |
| SLC12A9   | solute carrier family 12 member | $4.2 \times 10^{-5}$ |
| EPO       | erythropoietin | $5.1 \times 10^{-5}$ |
| POP7      | POP7 homolog, ribonuclease P/MRP subunit | $5.3 \times 10^{-5}$ |
| GNB2      | G protein subunit beta 2 | $5.4 \times 10^{-5}$ |
ratio $F$ statistic with degrees of freedom (18, 40) thus significant at about the 2% level. An alternative interpretation is that the anomalous values represent real effects established with much less security. Study of this would involve repeating the whole analysis with the initial nine values removed.

8 Some Simulation Studies

A range of simulation studies was used to investigate the effect on the Rényi and normal plots of departures from the the standard conditions. The numbers of test statistics considered were 100 and 1000.

The first step was to confirm the expected behaviour when the standard conditions were satisfied. Multivariate normal test statistics with non-zero correlations affected the slope and curvature of the relations.

Independent Student $t$ statistics with small numbers of degrees of freedom produced plots hard to distinguish from those to be expected if a few statistics represent departures from the standard conditions. That is, a few points near the extreme deviate from the anticipated linear behaviour. Test statistics that were equal within pairs or small sets made the deviations from the line even more marked. Autoregressive and moving average correlations between statistics changed the slope of the relations and made a few points at the right top of the plots deviate slightly from the linear relation.

9 Discussion

A normal probability plot and a Rényi plot contain exactly equivalent numerical information but emphasise rather different aspects of the data. It may often be helpful to use both. The Rényi plot has the simplest explicit probability structure, at least under idealised conditions, and puts most emphasis on the upper tail of values. The normal probability plot is more directly related to the estimation of meaningful contrasts and some kinds of departure from it are more directly interpretable. For example, if the plot is a straight line with non-unit slope, incorrect estimation of the error variance underpinning the statistics is suggested. Also, the normal plot may be used in two-sided form if appropriate.

A different approach aims to assess the posterior probability that a particular value represents a ‘real’ departure from the null hypothesis. Cox and Wong (2004, 2007) suggested a simple formulation in which each test statistic either had a standard normal distribution or was generated from a somewhat arbitrarily specified alternative normal distribution with a prespecified prior probability. Efron (2010) and in a wide-ranging discussion Efron and Hastie (2016) used an empirical Bayesian approach in which the underlying distributions under both null and alternative hypotheses are estimated from the data, which needs to be extensive.

Acknowledgements

We are grateful for helpful comments from the referee.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57(1), 289–300.
Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: Wiley.

Cox, D. R. & Wong, M. Y. (2004). A simple procedure for the selection of significant effects. *J. R. Stat. Soc. B*, 66(2), 395–400.

Cox, D. R. & Donnelly, C. A. (2011). *Principles of applied statistics*. Cambridge: Cambridge University Press.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Am. Stat. Assoc.*, 105(491), 1042–1055.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.

Fisher, R. A. (1924). *Statistical Methods for Research Workers*, 1st and 14th editions. Edinburgh: Oliver and Boyd. 1970.

Gharib, S. A., Hayes, A. L., Rosen, M. J. & Patel, S. R. (2013). A pathway-based analysis on the effects of obstructive sleep apnea in modulating visceral fat transcriptome. *Sleep*, 36(1), 23–30.

Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungaricae*, 4, 191–231.

Schweder, T. & Spjøtvoll, E. (1982). Plot of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3), 493–502.

Wong, M. Y. & Cox, D. R. (2007). On the screening of large numbers of significance tests. *J. Appl. Stat.*, 34(7), 779–783.

[Received January 2018, accepted March 2019]