Discovery: A Statistical Perspective

David R. Cox
Nuffield College, Oxford, OX1 1NF, UK

Abstract
A general statistical formulation of discovery problems is sketched and important distinctions drawn. Procedures for checking for the existence of a signal are analyzed from different viewpoints.

1 Introduction

Discovery is taken to mean finding and verifying a rare signal against a noisy background. There are many variants and formulation is critical. Key elements are the following:

- the reference frame in which signals are defined
- the statistical properties of the noise
- the temporal sequence of data collection
- the statistical character of the signal
- the frequency of occurrence of signals
- the multi-stage character of the search process

Thus the reference framework may be a set of bins, the order of which is essentially ignored, an ordered set of histogram bins, or a one or higher dimensional continuum, corresponding to energy, time, spectral frequency or to one or more spatial dimensions. Another possibility is the linked use of two or more reference frames. Thus a weak signal when data are ordered by energy level and a weak signal when ordered by some other feature might, under some circumstances, become a strong signal if it could be shown that the same originating events were involved in each case.

Typical noise processes are either Gaussian processes or Poisson processes and it may be important to allow for errors in estimating their properties. Conventional statistical thinking would, if there is extensive observation of the background, tend to rely on the empirical variability of the background rather than on *a priori* assumptions about its form. This would, for example, cover the possibilities of over- or under-dispersion relative to the Poisson distribution.

Observation may be in one step or by the gradual accretion of frequencies over time.

The signal may be a single blip, or a set of occurrences at nearby points in the reference frame.

A distinction crucial to the following discussion is between two main situations. First there may be no signal present or just one. In the second situation there are likely to be a limited but nonzero number of signals present and the challenge is to find as many as possible of them with few false alarms. The former seems the version more appropriate for current issues in particle physics and therefore we largely concentrate on that.

Examples range from particle physics to genetic epidemiology, especially GEWAS (genome-wide association studies), isolation of faint signals in complex spectra, and aspects of drug development and of plant breeding programmes. All have distinctive features.

2 Simplest Formulation

We start with the simplest formulation. At each of a large number n of sites a one-sided test of significance yields the set of p-values $p_1, \ldots, p_n$. For simplicity we assume the underlying test distributions to be essentially continuous and moreover the different tests to be statistically independent.

This is an explicitly significance test based formulation in which only a null hypothesis is specified formally, together with a criterion judged sensitive to relevant departures. There is no statistically formulated model of the signal, so that no estimation question arises. Of course richer formulations allow richer solutions.

Schweder and Spjotvoll (1982) discussed the analysis and interpretation of the empirical distribution of the p-values. The analysis could be done in terms of an equivalent standard normal variable, $t = \Phi^{-1}(1-p)$, where $\Phi(.)$ denotes the standard normal cumulative distribution function, or in terms of the transform $z = -\log p$. In a sense all are equivalent, but there are substantial advantages to the last because it places the region of interest prominently in large values, rather than crowded near zero for p and in the upper tail of a normal distribution for t. Use of z or a near equivalent appears a favoured method in analysing GEWAS, following Wellcome Trust Case Control Consortium (2007).

For more careful analysis and for dealing with generalizations it is helpful to use the Renyi decomposition. Under the global null hypothesis, that is that there is no signal present, the random variables $(z_1, \ldots, z_n)$ are independently exponentially distributed with unit mean. The Renyi decomposition is that the corresponding ordered values $(z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)})$ can be represented in terms of a another set of independent unit exponential variables $(v_1, v_2, \ldots, v_n)$ in the form

$$z_{(1)} = v_1/n, z_{(2)} = v_1/n + v_2/(n-1), \ldots, z_{(n)} = v_1/n + v_2/(n-1) + \cdots + v_n.) \tag{1}$$

That is, as the ordered values evolve in sequence they form a nonstationary random walk with exponentially distributed steps.

Two immediate consequences are first that the ordered z may be plotted against the expected values of the decomposition, Eq. (1), namely $1/n, 1/n + 1/(n-1), \ldots, 1/n + 1/(n-1) + \ldots + 1$. Interest lies in the upper end of the plot. Secondly the largest value, $z_{\max} = z_{(n)}$ is such that its associated significance level is

$$P(z_{\max} \geq z) = 1 - (1 - e^{-z})^n$$

and this is to a close approximation

$$1 - \exp(-ne^{-z}). \tag{2}$$

This is the standard allowance for selecting the most significant of many test statistics and the second form shows virtual equivalence to using the Gumbel distribution of extreme value theory for $z_{\max}$.

3 Extensions

The most immediate use of the above results is to show the assembled p-values graphically, in particular leading to evidence for an isolated signal if $z_{\max}$ is large. We now sketch a number of extensions and modifications.

Efron in series of papers, synthesized and extended in a monograph (Efron, 2010), has emphasized that the distribution of the test statistic under the real subject-matter null hypothesis may not be that specified by statistical theory. Indeed if interest lies in the 5σ region proposed in the particle physics context it is hardly plausible that the probabilistic interpretation associated with the Gaussian distribution will hold quantitatively. The implication for testing for a single extreme point is that $z_{\max}$ should be compared not with its expected value from the exponential distribution but with an extrapolation from previous values in the plot.

More formally if the p's are independent and identically distributed but not uniformly distributed the Renyi decomposition applies to a nonlinear function of the v's so that the plot mentioned above should be a smooth curve and if the upper section can be treated as locally linear the largest order statistics should have the form

$$\eta_0 + \eta_1\{v_1/n + v_2/(n-1) + \cdots + v_{n-k+1}/(n-k)\},$$

where η_1 is the local slope of the plot at large values.

This can be estimated from

$$\hat{\eta}_1 = \{2z_{n-1} + z_{n-2} + \ldots + z_{n-k} - (k+1)z_{n-k-1}\}/k.$$

Here the effective slope of the plot is estimated bearing in mind the special structure of the order statistics as a random walk. Then $(z_n - z_{n-1})/\hat{\eta}_1$ has under the null hypothesis the standard variance ratio or F distribution with degrees of freedom $(2, 2k)$. Choice of k will in general be based on inspection of the plot.

Next if the z are based on histogram bins, choice of bin size may be problematic. One resolution is to take a number of bin sizes $h, 2h, \ldots$ and to take the largest z, denoted again by $z_{\max}$. This will again have a Gumbel distribution, that is leading to a significance level of

$$1 - \exp(-n^* e^{-z_{\max}}). \tag{3}$$

Here n^* is an effective sample size, intermediate between the number of small bins and the total number of bins, the latter corresponding to the Bonferroni bound. The constant n^* could perhaps be calculated theoretically but would probably best be found by simulation; note that finding a single constant by simulation is much easier than studying the extreme tails of a distribution!

A further possibility is that the test statistics have some correlation structure, for example that of a stationary time series. Then the z-plot may be nonlinear with a nonunit slope and $z_{\max}$ will have a Gumbel distribution, Eq. (3), that is with a modified effective sample size, n^{**}, where in extreme value theory n/n^{**} is termed the extremal index. Such correlation might arise from the noise process. Another way would be if the signal is expected to be spread over a number of bins suggesting the use of a statistic, in Gaussian form, in the one sided-case of

$$S_m = T_m + aT_{m-1} + a^2 T_{m-2} + \ldots.$$

Here T_m is a standard Gaussian test statistic from bin m. After dividing S_m by its standard error it can be converted to z-form and then plotted as before.

Thus in summary the use of $z = -\log p$ provides a graphical analysis and a range of test procedures adaptable to various situations.

Figure 1, very kindly provided by Professor Brad Efron, illustrates such a plot based on a genetical application in which 6033 genes were examined for a possible connection with prostate cancer. A full description is given by Efron (2010, Section 2.1). The upward curvature of the plot shows a very clear departure from the theoretical null hypothesis distribution. A direct interpretation would be that the clear departure from linear form indicates that an appreciable number of genes are overexpressed relative to the null hypothesis. That there is a single fairly smooth curve implies that the data do not point to an isolated single anomalous point and that a possible explanation for all the data is essential agreement with the theoretical global null hypothesis distribution.

4 False Discovery Rate

The previous discussion has concentrated on testing the global null hypothesis, that is on the question of whether there is a signal at all. When it is likely that there is a small number of signals present, Benjamini and Hochberg (1995) suggested controlling the false discovery rate. There are two slightly different definitions of this. Let S signals be declared present of which F are false. Then the false discovery rate is defined as either

$$E(F/S \mid S > 0) \quad \text{or} \quad E(F/S).$$

The justification for the former definition is that if $S = 0$ there can be no false detection! In effect Benjamini and Hochberg's procedure uses cut off level among the $z_{(n-k)}$. Their formal analysis is in the

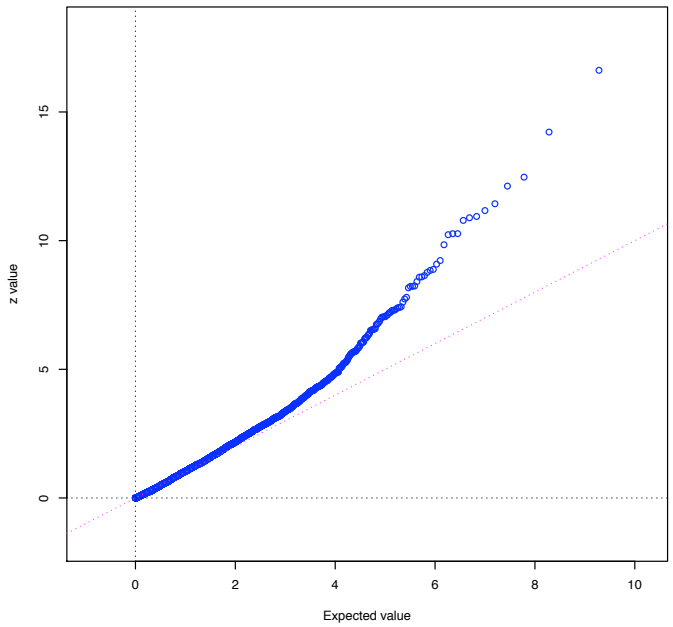

Fig. 1: Diagram provided by Professor Brad Efron arising from Efron (2010, Section 2.1). 6033 genes each tested for possibly being more active in prostate cancer patients. p-values converted to ordered z values where $z = -\log p$. Plotted against expectation assuming universal null hypothesis of no genuine effects. Upward curvature at right suggests some genes significantly overexpressed relative to null hypothesis.

spirit of Neyman-Pearson theory; it does not distinguish between overwhelming signals and those that only just emerge over threshold. This could be remedied to some extent by using several false discovery rates.

5 Bayesian Approach

We now turn briefly to a parallel Bayesian discussion. In the work summarized in his monograph Efron (2010) has developed a fully nonparametric empirical Bayes procedure in terms of two distributions, one for the null hypothesis and one applying when there is a signal, and a prior probability that the null is false. All are estimated empirically and in the monograph a number of examples of effective application are given. These do, however, require substantial amounts of data and a nontrivial number of signals. Cox and Wong (2004) gave a much simpler discussion in which the two associated distributions are normal with unit variance, the mean being zero under the null hypothesis and unknown under the signal. There are thus only two unknown parameters to be estimated; it was shown by simulation that satisfactory false recovery rates can be achieved. In both these procedures each potential signal is assigned an estimated posterior probability of being real.

To apply the procedure of Cox and Wong (2004) to the detection of a single signal we must somehow assign numerical values to the prior probability that there is a signal at a given site and to the mean of the corresponding normal distribution.

Suppose then that with n sites the prior probability of a signal at a particular site is α/n and that the corresponding normal distribution has mean μ_s. The total number of signals present thus has a Poisson distribution with mean α and the prior probability that the global null hypothesis holds is $e^{-\alpha}$. Thus if, for example, $\alpha = \log 2 = 0.69$ the prior probability of the global null hypothesis is one-half.

If the maximum normal-based test statistic is $t_{\max}$ the corresponding posterior log odds that there is a real signal, that is the log of the ratio of the probability of a real signal to the probability of a null

signal, is approximately

$$t_{\max}(\mu_s - t_{\max}/2) + \log(\alpha/n).$$

There is in this context no possibility of estimating μ_s from the data under analysis. The final term will typically be small and negative and, for given $t_{\max}$, the first term has maximum $t_{\max}^2/2$ achieved when $\mu_s = t_{\max}$. If $t_{\max} = \mu_s = 5$, the first term gives an odds ratio of approximately 10^5 compared with roughly $10^{-6.5}$ for the corresponding p-value, illustrating the superficially less extreme answers from the Bayesian formulation.

6 Discussion

A number of important issues are ignored in the preceding discussion. Typically discovery will be a multi-step procedure, starting often with preliminary data processing. Uncertainty at all stages will need consideration, although formal synthesis into a single measure of overall uncertainty may be neither necessary nor feasible. An extreme case of such a discovery process is traditional plant breeding in which a large number of varieties are reduced to a very small number in a series of trials involving progressively fewer and fewer varieties. Here issues of statistical significance at the separate stages are virtually irrelevant.

In many applications data are accrued over time and any signal will gradually emerge from noise. Analysis will proceed over time probably until a notable result can be reported and often data collection will continue past that. In a Bayesian formulation, provided the prior distribution and the general formulation do not change over time, there is no need to account for the repeated analysis.

No appreciable allowance for repeated analysis is required in frequentist theory, provided the formulation is in terms of confidence intervals with a target set for their width. The main formulation in previous sections is in terms of pure significance testing with no probability model of behaviour in the presence of a signal. Then allowance for the effect of repeated testing on the p-value is in general needed, but in the case of very extreme levels such as 5σ this allowance is likely to be negligible.

A more serious difficulty in application is the clash between the desire for the degree of objectivity achieved by precise prior specification of the procedure of analysis and the need to learn from the unexpected. In the present context it may be appropriate to concentrate on the single bin histogram approach possibly with allowance for data-dependent choice of bin width in the way outlined above.

It is a pleasure to thank Brad Efron for Fig. 1 and for helpful comments and Louis Lyons for comments on the paper and for many discussions of these issues.

References

[1] Y. Benjamini, Y. Hochberg, *J. R. Statist. Soc.* B **57**, 289-300 (1995).

[2] D.R. Cox and M.Y. Wong, *J. R. Statist. Soc.* B **66**, 395-400 (2004).

[3] B. Efron, *Large-scale inference.* IMS Monograph (Cambridge University Press, 2010).

[4] T. Schweder and E. Spjotvoll, *Biometrika* **69**, 493-502 (1982).

[5] Wellcome Trust Case Control Consortium, *Nature* **447**, 661-682 (2007).