https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-535

Rational Approaches to Correcting for Multiple Tests

Christopher W. Tyler, Ph.D., D.Sc.

Smith Kettlewell Eye Research Institute
San Francisco

Abstract

The logic of the Bonferroni correction for multiple tests, or family-wise error, is to set the criterion to reduce the expected number of erroneous false positives, or Type I errors, below 1. This is a very stringent criterion for false positives in cases where the test may be applied millions of times, and will necessarily introduce a large proportion of false negatives (missed positives, or Type II errors). A proposed solution to this problem is to adjust the criterion for False Discovery Rate (Benjamini & Hochberg, 1995), which allows the number of false positives to increase proportionally to the number of true positives, though remaining at a small proportion, dramatically reducing the number of false negatives. This approach may be conceptualized as working with a relaxed confidence level that any one test is a true rather than a false positive, bringing the criterion more into line with our societal assessment of the validity of statements in general, and even in science, as having less than 100% certainty. The analytic strategy to the assessment of statistical significance provides a more intuitive approach to the identification of sparse signals in large datasets than the standard Bonferroni approach to correction for multiple tests.

Introduction

The role of statistical testing is to set meaningful limits on what may be regarded as the error range for statistical estimates. However, functional error ranges depend on the number of times a test is applied, which in the era of 'big data' may number in the millions of applications, making a criterion that reduces false alarms to 1 in 20 applications (symbolized as $p < 0.05$) subject to unacceptable numbers of false alarms. It is therefore important to provide an adequate approach to the assessment of large numbers of multiple tests when applying them to large datasets, such as those encountered in product quality control, identification of genetic substrates of diseases or individual characteristics, or high-resolution medical imaging such as brain connectivity analysis.

The Era of Big Data

The 21st century may be characterized as the beginning of the era of Big Data, spearheaded by the Human Genome Project whose initial deadline was the turn of the millennium in the year 2000. The term 'Big Data', which seems to have its origins in computer graphics developed by Silicon Graphics for Hollywood special effects (Mashey, 1998) and econometrics around the same time (Diebold, 2003), span many arenas of human endeavor.

Climate data span the globe, with NASA satellites sending back more than ten million gigbytes of data per year.

The human genome contains 3 billion base pairs, so it was a decade-long effort to sequence it for one individual, but techniques have rapidly improved and now it can be done for any of the 7 billion individuals on the planet in less than an hour.

The internet has a somewhat longer history dating back to the mid-20th century. It has now reached the scope of more than 1 million terabytes of information (Kemp, 2017), although these are exchanged among the relatively modest number of only about 1 billion websites (i.e., less than 1 per every 7 people on the planet). By contrast, the human brain contains as many as 100 billion neurons, so a single brain is still more highly connected than the entire internet by a couple of orders of magnitude. Indeed, the numbers of potential connections in these systems are enormously larger than the numbers of nodes, at about 1 quadrillion for the Internet and 10 quintillion for the human brain.

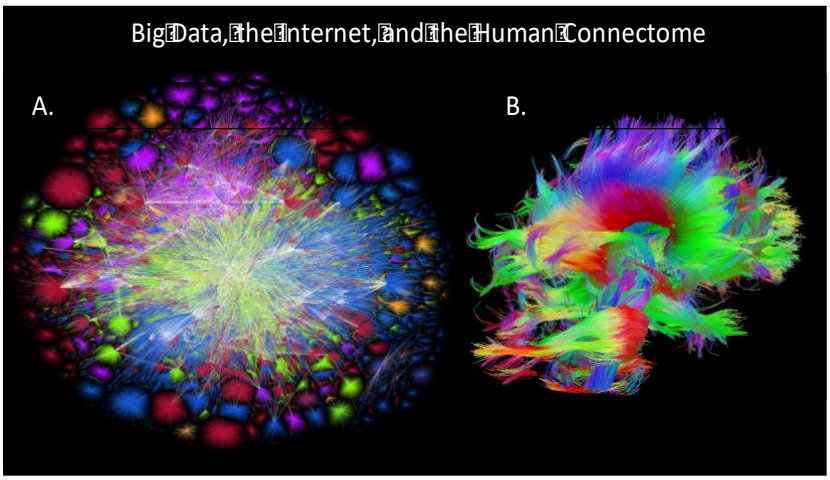

Fig. 1. A. A depiction of the Internet in 2005 from http://www.opte.org/the-internet/. B. The shape and properties of the brain's white-matter structures have been shown to be related to behavior, cognition and neurological diseases. (Courtesy of the USC Laboratory of Neuro Imaging and Martinos Center for Biomedical Imaging, Consortium of the Human Connectome Project).

The Human Connectome Project is the current effort to study the structural, functional and directed (or effective) connectivity of the human brain as a function of its interaction with the environment. Although it reaches very much into the domain of Big Data, the connectome does not reach the scale of the neural connectivity. Structural connectivity is performed at a resolution of about 0.5 mm^3, or about 50,000 voxels of segmented neural information, providing for the order of a billion possible connections. Functional connectivity, whether undirected or directed, is performed at a resolution of about 2 mm^3, or about 1,000 voxels of segmented neural information, providing for the order of only a million possible connections. Nevertheless, determining which of these connections is significantly identifiable or activated in a particular brain at a particular time requires an individual statistical test for every one of the million or billion possible connections, and thus demands a well-considered approach to the correction of multiple tests. (At present, it is not possible to envisage brain connectivity analysis at the full scale of neuronal resolution, but advances will undoubtedly be made toward that ultimate goal.)

Signal Detection Theory and the Student t Test

The core concept for classical statistical testing is based on the assumption that performance is limited by additive Gaussian noise (based on the outcome of the Central Limit Theorem, which is that the combination of various sources of noise is asymptotically Gaussian). Statistical tests for the detection of a given signal in a noisy environment (which, in the limit, is the case for all knowledge acquisition and assessment) are based on the Signal Detection Theory (SDT) formalism (Tanner & Swets, 1954). The signal is drawn from a distribution $f_S(x)$ of values perturbed by the added Gaussian noise, while the non-signal null samples are drawn from a separate distribution $f_N(x)$ with the same standard deviation (Fig. 1). In practice, the distribution

must be assessed by drawing multiple samples in conditions when no signal is known to be present, and fitting a theoretical distribution to the result. For typical multi-source noise, the Central Limit Theorem makes the Gaussian distribution the best candidate for the theoretical distribution, but other distributions may be appropriate for special situations.

To test whether one sample is a signal or a noise sample, the SDT formalism is to set a criterion level on the basis of the noise distribution $f_N(x)$ and treat any sample larger than that value as signal. This approach will allow a known proportion of noise samples to be treated as signal, or false alarms (e.g. 5% for a criterion level of $p < 0.05$), indicated by the blue region in Fig. 1. For a mean signal level $f_S(x)$ matching this criterion, however, 50% of true signals will be rejected (misses), with progressively greater discriminability as the signal level increases.

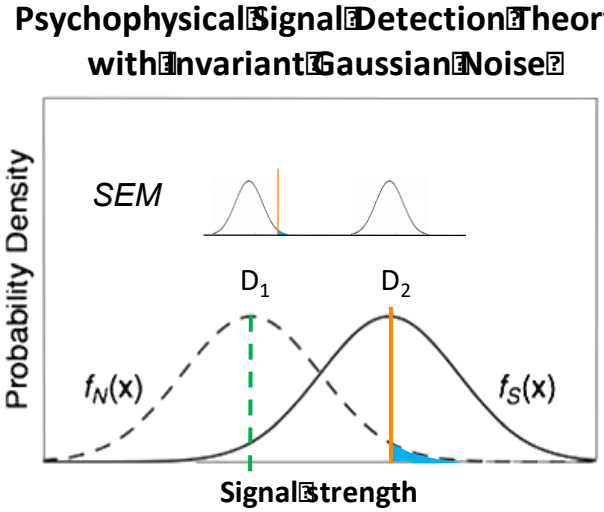

Fig. 1. Classical Signal Detection Theory analysis, which also corresponds to the basis of the *t*-test. The noise ($f_N(X)$) and signal+noise ($f_S(X)$) distributions are shown separated by 1.96 σ between the means D_1 and D_2, leaving a criterion level at 5% of the noise distribution as the cutoff for the signal estimates. Computing the standard errors of the means for 100 samples allows the estimated criterion level to be substantially reduced according to the standard error of the means (SEM).

The next level of assessment is to take multiple samples, or readings, of the same situation, or sampling distribution, in order to improve the accuracy of the statistical assessment. Taking the average of these readings will reduce the spread of the distribution by the square root of the number of samples, *n* (Fig. 1 inset). Applying the same false alarm proportion of $p < 0.05$ to the new multiple-sample noise distribution will result in a proportionately lower criterion level, so that the signal strength will result in a much higher proportion of signal samples being categorized as valid signal.

The applicable statistical test for the multiple sample situation depicted in Fig. 1 (inset) is the "Student" t-test (Gosset, 1908), defined as the ratio of the difference between the means to their standard error (SEM):

$$t = \frac{\text{difference between means}}{\text{difference SEM}} \quad (1)$$

$$= \frac{M_2 - M_1}{\sqrt{\left(\left(\frac{s_1}{n_1}\right)^2 + \left(\frac{s_2}{n_2}\right)^2\right)}}$$

To assess its significance, the value of *t* must be checked against a statistical table of *t* values at the preferred level of significance based on the applicable number of degrees of freedom (*df*), which are calculated by the simple formula:

$$df = n_1 + n_2 - 2 \quad (2)$$

The criterion value of *t* varies from about 10 for $df = 2$ to 1 for $df \sim= \infty$.

Multiple Testing

So far, we are considering a single application of the criterion level for a single statistical test. When the test is applied multiple times, the applicable criterion has to be adjusted, since some proportion of the tests will inevitably be classified as valid results even when they are drawn from purely random effect distributions. The standard approach to setting the criterion for multiple testing or, "family-wise error", was codified by Bonferroni (1936) as the criterion to reduce the probability of an erroneous false positives below 1. This is a very stringent criterion for false positives in cases where the test may be applied tens of thousands or millions of times, as is often the case in large datasets. The equation for the Bonferroni corrected criterion level p_c is simply to divide the desired uncorrected level p_u by the number of tests T to be performed:

$$p_c = p_u / T \quad (3)$$

For example, at $p < 0.05$ the Bonferroni correction for a significant signal in any one of 10,000 voxels in a human brain scan is $p < 0.000005$, which requires a signal z-score level of 5 times the prevailing noise level, a radical increase over the z-score of 1.96 required for a single test.

Benjamini/Hochberg correction for false discovery rate, or confidence level

Another way to characterize this criterion is in terms of *confidence level* in each assessment of significance. Thus, it is generally accepted that the logic of everyday calculations is not absolute, even down to the definition of object concepts In practical application, there is always some fuzzy fringe, or region of uncertainty, surrounding the definitive core of any concept (Zadeh, 1965; Tyler & Likova, 2010). This uncertainty can be expressed as a confidence level in the categorization of any aspect of reality, as for how likely it is to match the ideal definition of that concept. For example, gender recognition by current software has error rates of 0.8% for light-skinned males, 7% for light-skinned females, 12% for dark-skinned males and 35% for dark-skinned females (Buolamwini & Gebru, 2018). Although human recognition capabilities are better than this, they are not absolute, so we cannot be 100% confident in any given object or feature identification. It therefore makes the most sense to apply a specified confidence level (such as 90%) to any given statistical assessment rather than requiring an unrealistic level of absolute certainty.

This situation has been addressed in an interesting form by Benjamini & Hochberg (1995) to take into account the number of tests that pass the significance criterion rather than focusing entirely on eliminating any false positives, or "false discoveries". Thus, rather than reducing the ***number*** of false positives to zero, as was the case for the Bonferroni criterion, the concept is to hold the ***rate*** of false positives to some small fraction of the number of true positives, or valid results. For example, it may be acceptable to allow a 10% rate of false positives among the valid results, giving 90% confidence in any one result.

Holm (1979) had made the proposal to order the significance levels for multiple tests and consider them in order relative to the Bonferroni criterion, stepping down the n in the numerator by one each time that a test passes the its applicable criterion. This allows for more tests to pass at high significance, but it still rejects many valid results that would pass the usual criterion on a single test basis by holding to the concept of absolute validity of all the results. The Benjamini-Hochberg approach based on an acceptable ***rate*** of false positives has been characterized as a step-up

method rather than Holm's step-down method, ending up rejecting fewer true positives.

Applying this concept of a confidence level to the correction for multiple tests is simple in the case of no signal conditions, since the noise distribution can be assumed to be fully Gaussian. The problem arises when the signal is present, since its distribution is arbitrary and unknown. The signal distribution will be assumed to be flat (if necessary) but it may not be necessary. Given the number of tests, the area under the normal distribution is known for any given p value. The excess of positive results over that number therefore represents, on average, the signal results. For example, if 2000 tests are assessed at the $p < 0.05$ criterion level, the number of false positives in each tail should be 50 and the Bonferroni criterion required to eliminate all false positives is $p < 0.000025$ (i.e., 0.05/2000).

Benjamini-Hochberg (B-H) Procedure

Implementing the B-H procedure is a two-step strategy. The first step is to run the analysis at the uncorrected level of significance in order to determine the number of possibly significant results if tested individually. If there is no signal in the dataset, the expected number of false positives in each tail of the distribution is given by

$$E = p_u \, x \, n \, / \, 2 \qquad (4)$$

where p_u is the uncorrected criterion and n is the number of applications of the test (e.g., 500 false positives for 10,000 tests at p_u = 0.05).

The possible number of truly significant results at p_u is then estimated as

$$T = P_u - E \qquad (5)$$

where P_u is the empirical number of positive results at $\underline{p_u}$.

Thus, for the relaxed criterion of the confidence level, C, the new approach defines the corrected significance level p_c as

$$p_c = p_u \, / \, T \, x \, C \qquad (6)$$

For comparison, Bonferroni correction is $p_c = p_u / n$.

For m tests

1. Rank the individual p-values in descending order.
2. Draw a line from 0.05 to 0 over the interval 0 to m.
   [$p_{BH} = (i/m)Q$]
3. Take all tests below this line as significant.

Simulated example

In the above example, depicted in Fig. 4, the blue curve shows the samples from a simulated Gaussian distribution plotted in inverse percentile rank order. The red curve is a similar function for the example of 10% of significant test results in 10,000 tests (at a level of z = 2). In the lower half of the figure (on an expanded ordinate to show the bottom 5% of the scale) are depicted the four significance criteria discussed in this paper.

The upper dashed line shows the standard uncorrected criterion of p = 0.05.

The lower dashed line shows the Bonferroni corrected criterion of p = 0.000005.

The green curves show the rank-dependent criteria levels for the Holm approach for 10, 100, and 1000 tests in descending order in the graphic. Thus, it can be seen that the Holm

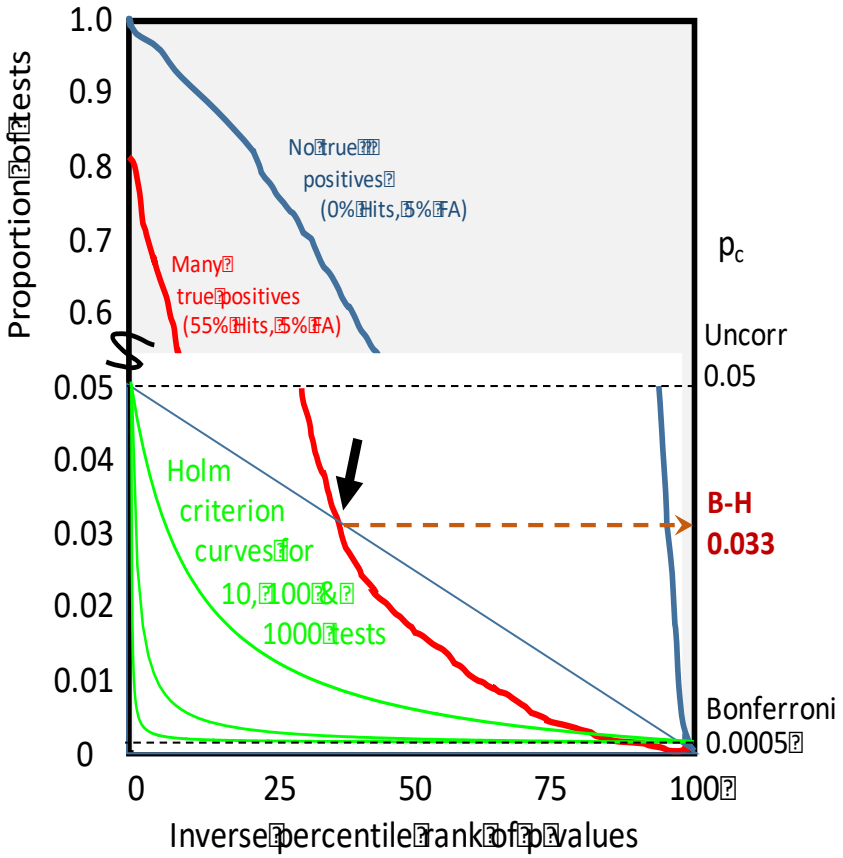

Fig. 2. MonteCarlo simulation comparing the Benjamini-Hochberg (B-H) method with simple Bonferroni correction and the Holm sequential method, for two distributions: a 10% signal distribution (red curve) at the 50% criterion level (z = 1.96) and a pure noise distribution (blue curve). For both the Holm and B-H approaches, *p* values for all individual tests are inversely rank-ordered from high-to-low to define the criterion line between the uncorrected and the Bonferroni corrected criteria. The ordinate is expanded up to the uncorrected criterion value of 0.05 for clarity. In this format, the Holm criteria are shown as the green curves for 10, 100 and 1000 tests, in descending order. The points where the B-H and Holm rank-order curves cross the criterion line define the hit/false-alarm ratios for the two distributions.

approach approximates the low Bonferroni cut progressively more closely as the *n* increases, providing progressively less benefit for larger *n*. By the 10,000 test level, it closely approximates the Bonferroni line along most of its extent. It is therefore ineffective in recovering the 100 true positives in this sample.

The 90% B-H criteria is indicated by the oblique line running from the uncorrected criterion to the Bonferroni criterion lines, with the buff coloration indicating the zone of significant results. It can be seen that, due to the curvature of the inverse ranked curve for the 10% significant results (red curve), a full range of the simulated results are categorized as positive 'hits'. The actual proportion of hits was 55% of the total number of true positives, with a 5% false alarm rate corresponding to 90% of the 'signal' cases (i.e., a 90% confidence level that the signal cases are valid). The lowest applicable significance criterion is lowered to 0.033 in this example, only a little more stringent than the uncorrected criterion.

It should be stressed that, in the absence of any actual positive in the dataset, the B-H procedure will not produce many more false positives than the Bonferroni, but that it will detect a higher proportion of positives if they have higher z-scores (i.e., when the tail of the inverse percentile rank curve extends further to the right).

Thus, it is clear that there is no free lunch, but that the graded criterion approach provides for the detection of the majority of signals for the price of allowing contamination with small proportion of noise elements. In many fields of endeavor, this is well worth the cost for the benefit of recovering potentially a large number of signals that would be excluded by the stringent Bonferroni criterion. Given that there are many uncertainties of control in a particular experimental design, such as contaminated samples, temperature control, treatment compliance, disease diagnosis, etc, etc, it does not seem reasonable to hold the statistical criterion to an absolute standard of perfection.

Popularity of the Benjamini-Hochberg correction for false discovery

As mentioned, the 21st century may be characterized as the beginning of the era of Big Data, so it might be expected that a statistical procedure such as the B-H correction for 'discovery', which is designed for use with large datasets, would achieve widespread popularity in this era. Since 1995 when it was first published, however, the B-H correction has had a limited usage history (see Fig. 3). Barely registering until 2000, it had a burst of popularity for about 5 years. But since 2005 its growth leveled off to a markedly slower rate of about 10% per year, currently reaching a level of a little over 1000 publications per year. This is a tiny proportion of the roughly 2 million scientific publications every year, so it is clear that the technique has by no means overtaken traditional approaches to statistical analysis, despite its obvious advantages in data mining.

Conclusion

There are two basic philosophies for statistical analysis. The traditional approach is to adopt a strategy than ensures that, within the bounds of statistical estimation, no result is considered unless it has an absolute certainty of not being attributable to random processes. This approach works well for individual cases of statistical assessment but founders in the era of Big Data, when an investigation may require thousands or millions of statistical tests to be applied to a dataset.

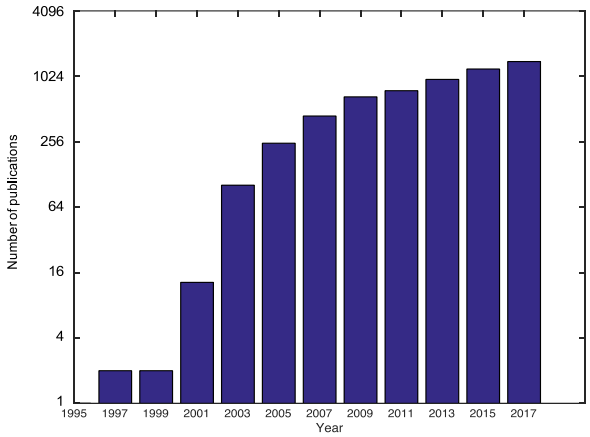

Fig. 3. Bar graph of the number of publications using the term 'false discovery' in two-year time bins since its original publication in 1995. Note slow growth following a rapid increase in the first decade of the 21st century.

The alternative approach is to allow for some probability that any given result may have occurred by chance, scaled to the number of positive results in the dataset under study. The Benjamini-Hochberg procedure provides an implementation of this approach that can radically increase the number of underlying events detected in a data structure at the price of only a small decrease in the level of confidence in their validity.

References

[1] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B, 57: 289–300.

[2] Bonferroni CE (1936) Teoria Statistica delle Classi e Calcolo delle Probabilità. Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

[3] Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification." Proceedings of Machine Learning Research, 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency.

[4] Diebold FX (2003) 'Big Data' dynamic factor models for macroeconomic measurement and forecasting (Discussion of Reichlin and Watson papers), in Dewatripont M, Hansen LP, Turnovsky S (Eds.), Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society. Cambridge University Press: Cambridge, 115-122.

[5] Gosset WS (1908) The probable error of a mean.
Biometrika, 6: 1–25. doi:10.1093/biomet/6.1.1.

[6] Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statistics, 6: 65–70

[7] Kemp S (2017) Digital in 2017. https://www.linkedin.com/pulse/digital-data-trends-every-country-world-simon-kemp/ (accessed 3/14/2018).

[8] Mashey J (1998) Big Data and the Next Wave of Infrastress.https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems (accessed 3/3/2018).

[9] Tanner WP, Swets JA (1954) A decision-making theory of visual detection. Psych Review, 61: 401-409.

[10] Tyler CW, Likova LT (2010) An algebra for the analysis of object encoding. NeuroImage, 50:1243-50.

[11] Zadeh LA (1965) Fuzzy sets. Information and Control, 8: 338–353.