Commentary

Inna Gerlovina*, Mark J. van der Laan and Alan Hubbard

Big Data, Small Sample

Edgeworth Expansions Provide a Cautionary Tale

DOI 10.1515/ijb-2017-0012

Abstract: Multiple comparisons and small sample size, common characteristics of many types of “Big Data” including those that are produced by genomic studies, present specific challenges that affect reliability of inference. Use of multiple testing procedures necessitates calculation of very small tail probabilities of a test statistic distribution. Results based on large deviation theory provide a formal condition that is necessary to guarantee error rate control given practical sample sizes, linking the number of tests and the sample size; this condition, however, is rarely satisfied. Using methods that are based on Edgeworth expansions (relying especially on the work of Peter Hall), we explore the impact of departures of sampling distributions from typical assumptions on actual error rates. Our investigation illustrates how far the actual error rates can be from the declared nominal levels, suggesting potentially wide-spread problems with error rate control, specifically excessive false positives. This is an important factor that contributes to “reproducibility crisis”. We also review some other commonly used methods (such as permutation and methods based on finite sampling inequalities) in their application to multiple testing/small sample data. We point out that Edgeworth expansions, providing higher order approximations to the sampling distribution, offer a promising direction for data analysis that could improve reliability of studies relying on large numbers of comparisons with modest sample sizes.

Keywords: finite sample inference, hypothesis testing, multiple comparisons

1 Introduction

Explosive proliferation of recordable information that keeps Big Data at the top of the buzzword list not only calls for continuing development of analytical methodology but also ignites heated debates in statistical and scientific literature as well as in the media. Following March 14, 2014 article in *Science* magazine [1, 2] dissecting Google Flu Trends failure, commentaries and opinion pieces on the subject appeared in such publications as *New York Times* [3], *Financial Times* [4], and *Harvard Business Review* [5].

Discussions on misuse of statistical inference as well as reliability of conclusions have a very long history, their current and highest wave starting back in 2005 when the paper titled “Why Most Published Research Findings Are False” by J. Ioannidis [6] came out and quickly became a famous catalyst on the topic – “an instant cult classic” [7]. This wave culminated with American Statistical Association issuing a statement on March 7, 2016 that calls forth a new era of statistical literacy and reliability/reproducibility of scientific findings [8]. With an advent of Big Data, the enthusiasm that accompanied it and the limitless possibilities it opened were counter-balanced by the growing distrust in published findings. Concerns outlined in Ioannidis’ article did not disappear with technological advances and restructuring of social networking that brought about massive amounts of generated (produced and collected) data. Just the opposite – the number of dis-

*Corresponding author: Inna Gerlovina, Division of Biostatistics, University of California, Berkeley, 101 Haviland Hall, Berkeley, CA 94720, USA, E-mail: innager@berkeley.edu

Mark J. van der Laan, University of California, Berkeley 101 Haviland Hall, Berkeley, CA 94720, USA

Alan Hubbard, Division of Biostatistics, School of Public Health, UC Berkeley, Berkeley, CA 94720, USA

coveries in many fields were rising, and so was the skepticism about how trustworthy those results might be, reflected in opinion pieces and papers such as an eloquently titled “A vast graveyard of undead theories: Publication bias and psychological science’s aversion to the null” [9].

In describing data, adjectives *big* and *small* take on special meanings and are not necessarily mutually exclusive anymore – Big Data and small sample size can be the characteristics of the same dataset. Complexity of Big Data often involves multiple comparisons and thus necessitates testing a great number of hypotheses simultaneously, which only exacerbates problems posed by the small sample size. Much quoted words from David Spiegelhalter, professor at the Statistical Laboratory of the University of Cambridge, summarize the problem with blunt simplicity: “There are a lot of small data problems that occur in big data. They don’t disappear because you’ve got lots of the stuff. They get worse.” [4]

2 Theory and practice

Rejecting null hypotheses when true (in other words, finding signal where there is none) produces type I errors (false positives); traditional construction of null hypotheses puts more gravity on type I errors than on type II errors (failure to reject the null when it is false). To protect against proliferation of distracting and potentially misleading false positive results is one of the main points of rigorous statistical inference. In high-dimensional data analysis this is partly addressed by a variety of multiple testing procedures. Still, many other factors contribute to reliability of the conclusions. Given these factors, reliability can itself be explored, quantified, and possibly improved by the choice of an analysis method. This notion relates to the “inference on inference” issue, which has also received attention in the literature recently; for example exploring the fact that summaries such as p -values and confidence intervals are themselves statistics and carry a certain amount of noise [10]. When “Big Data” means that the number of tests gets progressively larger and information available for testing each of the hypotheses gets yet more limited, finite sample departures from asymptotic distributions become critical, possibly leading to biased inference and inflated false positive error rates that extend far beyond nominal level.

As an example of small sample high-dimensional data, consider a genomic study where the number of replicates, n , is very limited. Each replicate consists of many features, such as genes, RNA-seq, miRNA, etc.; analysis of such a study includes testing one or more hypotheses for each of the features bringing the number of tests, m , to hundreds of thousands or even millions. Accounting for multiple comparisons is commonly done by some multiple testing procedure aimed at controlling an error rate, e.g. familywise error rate (FWER) [11] or false discovery rate (FDR) [12]. All of these procedures move the critical value (rejection cut-off) further away from the center of the null distribution, which means that decisions in hypothesis testing are relying on accurate approximation of distal tails of that distribution; the greater the number of tests, the smaller the corresponding tail probabilities, and thus the more extreme tails of the distribution need to be approximated in order to derive statistical inference.

For a small sample size, true sampling distribution of even asymptotically normal estimators might be quite far from normal. For quantiles that are far away from the mean this can potentially present an even bigger problem and adversely affect reliability of inference; even apparently small departures from normality might have a significantly amplified proportional effect on the far tails resulting in poor approximation and, consequently, poor error rate control and ultimately faulty inferences.

One way to avoid making assumptions on a sampling distribution is to use permutation – a nonparametric exact method – when it is relevant. When the null hypothesis postulates independence of some kind, permutation can be used to estimate the null distribution of a test statistic. That, however, will present a problem of a different kind: the limited number of permutations available for a small sample will provide too coarse a grid to estimate very small probabilities. Then the smallest possible unadjusted p -value, which would be a result of a situation when observed data provides an extreme point of a permutation distribution, is an inverse of the number of possible permutations – e.g. if an independent variable is binary with the sample size $n = n_0 + n_1$, this smallest value is $1/\binom{n}{n_0}$. After multiple testing correction with large number of comparisons, the minimal *adjusted* p -value will be too large to pass

the significance threshold in most cases and thus will not allow the method to detect any significant associations.

How can we know the inference is to be trusted – what is needed to ensure a declared level of certainty? The most accurate inference is achieved using approaches with rigorous theoretical underpinnings, which in this case are provided by large deviation theory. The theory is used to determine the sample size that is large enough so that multivariate sampling distribution of a test statistic (such as sample average) is closely approximated by multivariate normal distribution at the appropriate quantiles (based on m) for proper error rate control. Some of the conditions that ensure the convergence of Cramér's large deviation expansion establish formal relationship between a sample size n and a quantile x of the distribution of an estimator: $x = o(n^{1/6})$ [13, 14]. As noted above, in high-dimensional data the critical value (cut-off quantile) x_c is directly related to the number of tests m through a multiple testing procedure, and that leads to the condition that is necessary and sufficient to guarantee error rate control: For a Normal or Student's t approximation of the distribution of the test statistic, error rate control (either FWER or FDR) can be achieved if and only if $\log(m) = o(n^{1/3})$ [15]. If this condition is satisfied, the actual error rates are close to the nominal ones and the null hypothesis is indeed rejected at the declared significance level α ; failure to achieve it puts reported results into the territory of hope rather than probability.

How often is this condition satisfied in high-dimensional data we encounter in practice? To illustrate, we gauge the sample size that would make error rate control possible for a given number of tests. Suppose $m = 10,000$, which is relatively modest; then, if we want $\log(m) = \frac{1}{10} n^{1/3}$, n should be 800,000. As a comparison, sample sizes of recently uploaded datasets on Gene Expression Omnibus (GEO) provide a striking reality check: sample sizes of the great majority of the studies are below 20 and very few reach above 80; the empirical density of the number of independent replicates peaks at the values between 1(!) and 3, then drops considerably after $n = 10$. Many of these studies are genome-wide and therefore the number of tests is usually high (starting at tens of thousands), which means error rate control is not guaranteed for such studies though it is still not clear how far off claimed results might be from the truth.

3 Edgeworth expansions

It would be illuminating to see what this difference between theory and practice translates to in terms of actual error rate control: to explore the effect that some departures from normality might have on type I error rates and show the extent of the resulting discrepancies between actual and nominal error rates. To do that, we employ Edgeworth expansions that provide a way of comparing cut-offs obtained with different orders of approximations for the same dataset and evaluating them under the true sampling distribution of a chosen test statistic. In the reality of a finite sample, Edgeworth series allow us to get closer to the critical value that would be calculated had the true distribution been known by obtaining higher-order approximations of that distribution. As such, they also suggest a promising direction for data analysis where empirical moments could be substituted for the true moments in the higher-order terms. We use these higher-order approximations to assess how sensitive the accuracy of the analyses involving multiple-testing procedures is to even slight departures from normality.

Edgeworth series is an asymptotic expansion that originally extended the idea of a Central Limit Theorem providing an expansion for the distribution of a standardized sample mean, as well as a general framework for obtaining expansions of the same type for other sample statistics such as a Studentized mean or a variance. It is a series of functions where the first function (which is usually denoted as a zero term) is a standard normal c.d.f. $\Phi(\cdot)$. Since it is a power series in $n^{-1/2}$, truncation after j 'th term provides an approximation to the distribution of interest as the remainder is of the order of $n^{-(j+1)/2}$. This truncated series is usually called a j -term expansion or a $(j + 1)$ 'th order approximation. Written in terms of cumulants κ_j and Gaussian p.d.f. $\phi(\cdot)$, the series has a form

$$F_{\hat{\theta}}(x) = P\left(\hat{\theta} \leq x\right) = \Phi(x) + n^{-1/2}p_1(x)\phi(x) + n^{-1}p_2(x)\phi(x) + \mathcal{O}_p\left(n^{-3/2}\right)$$

for a normalized test statistic $\hat{\theta}$, where $p_j(x)$'s are expressed in terms of cumulants of the data generating distribution.

Extensive body of work on Edgeworth expansions is part of the legacy of Peter Hall, one of the great and most influential figures in mathematical statistics. This work included a multitude of ideas, theoretical advancements, and analysis methods; among them methods based on empirical expansions, combination of bootstrap/resampling methods and Edgeworth expansions as well as possible applications to high-dimensional problems. He has formulated a framework for deriving these expansions for general statistics, especially those that could be used in practice, such as a t -statistic (presenting a third order expansion) and sample variance (second order expansion) [16, 17]. The methods that we use in our comparisons are based on this important work. In addition to Peter Hall's contributions, there is a vast literature on Edgeworth expansions; for an introduction to the theory behind them and practice of their use in asymptotic statistics see [18, Section 23.3].

4 Under the magnifying glass

As an illustration, we choose the following simple example: small sample ($n = 10$) of independent gamma distributed random variables with shape parameter 3 (this distribution is somewhat skewed but is still unimodal and fairly nice); the test statistic is a studentized mean and a one-sided test is used for simplicity. The distribution of this t -statistic is skewed to the left and we want to look at the left tail. In other words, we want to test " $E(X) - 3 \geq 0$ " against its alternative based on observing $X_1, \dots, X_n$ i.i.d. $\Gamma(3, 1)$ and using $\hat{\theta}_n = n^{1/2}(\bar{X}_n - 3)/s_n$ with $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $s_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

To obtain an actual error rate for such a test given the order j , sample size n , number of tests m , significance level α , distribution of X , and multiple testing procedure, we:

1. calculate the unadjusted probability cut-off corresponding to α : e.g. for Bonferroni multiple testing correction (MTC) $p = \frac{\alpha}{m}$ for the left tail and $p = \frac{1-\alpha}{m}$ for the right;
2. find the corresponding critical value (quantile) $q = F_{j,n}^{-1}(p)$, where $F_{j,n}$ is a j 'th order Edgeworth expansion;
3. find tail probability for this quantile that is based on the true sampling distribution F_0 of the estimator: $p_{true} = F_0(q)$ for the left tail and $p_{true} = 1 - F_0(q)$ for the right;
4. derive the corresponding actual error rate r : e.g. $r = \min(m p_{true}, 1)$ if Bonferroni MTC is used.

Here we use Bonferroni MTC for convenience since it is not a step-down procedure and thus does not require sorting of p -values. Note that r needs to be bounded by 1 since it is a probability; a value of r that is greater than 1 has no statistical interpretation; however, in some cases it might be helpful to look at that unsubstituted value to gauge the "magnitude of a disaster" and compare it with other unsubstituted values. We are specifically interested in how increasing the number of tests affects the accuracy of the analyses involving multiple testing procedures, with special focus on situations when $m \gg n$, so we conduct our error rate assessment across a wide range of m for the same sample size.

Figure 1 displays actual error rates for the sample size 10 and the number of tests ranging from 1 to about a million (the x-axis is on the log scale and marks the values of 1, 2, 4, $\dots$, $2^{20} = 1,048,576$). Dotted line at $y = 0.05$ indicates where the error rate would need to be to match the nominal (reported!) level. The blue line gives the rates for Student's t -distribution approximation that is customarily used in data analysis (at this sample size, normal approximation will not be anywhere near the truth at the far tails). It can be seen that for the values of m that are over a thousand (2^{10}), there is virtually no error rate control; while the rates are truncated at 1, the thickness of that truncation line at the highest values of m indicates that it is indeed "worse" than no control. Starting with second-order approximations, Edgeworth expansions show markedly improved results, and forth and fifth orders are well below the nominal line, which gives hope for a more reliable inference that could be achieved with incorporating higher empirical moments into data analysis. This approach will provide more power than dependable yet conservative methods that are based on finite

Figure 1: Actual error rates for the one-sided t -test of nominal level $\alpha = 5\%$ of “ $E(X) - 3 \geq 0$ ” against its alternative based on observing $X_1, \dots, X_n$ i.i.d. $\Gamma(3, 1)$ with $n = 10$.

sampling inequalities (Bernstein, Bennett, Hoeffding) [19]; however, these latter methods might be preferable in situations where false positives are highly undesirable and error rate control is crucial.

The disastrous numbers in this simple example do not reflect any of the misuses of the p -values addressed by the American Statistical Association statement; in fact, those would aggravate the situation that is already rather grim. It is meant to serve as an alert to the possible (un)reliability of inference where the error rate control is not guaranteed (which is the case in most of the studies involving multiple comparisons) and where assumptions of normality or near-normality are used to justify the certainty of the results while not being necessarily warranted. It also highlights one of the possible factors contributing to the “reproducibility crisis”; while not easy to overcome, this factor should be kept in mind as a cautionary tale.

Acknowledgment: This work was partially funded by NIEHS Award P42ES004705 (the Berkeley Superfund Research Program).

References

1. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science* 2014;343:1203–05.
2. Lazer D, Kennedy R, King G, Vespignani A. Google Flu trends still appears sick: an evaluation of the 2013–2014 flu season. Available at SSRN 2408560, 2014.
3. Marcus G, Davis E. Eight (no, nine!) problems with big data. *The New York Times* 2014;6.
4. Harford T. Big data: are we making a big mistake? *FT Magazine*, 2014.
5. Fung K. Google Flu trends' failure shows good data > big data. *Harvard Business Review* 2014;25.
6. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
7. Shaywitz D. *Science and shams*. Boston: Globe, 2006.
8. Wasserstein RL, Lazar NA. The ASA's statement on p -values: context, process, and purpose. *Am Stat* 2016;70(2):129–133.
9. Ferguson CJ, Heene M. A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspect Psychol Sci* 2012;7:555–61.
10. Gelman A. Commentary: P values and statistical practice. *Epidemiology* 2013;24:69–72.
11. Hochberg Y, Tamhane AC. *Multiple comparison procedures*. New York, NY: Wiley, 2009.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;57(1):289–300.
13. Cramér H. On the composition of elementary errors: first paper: mathematical deductions. *Scand Actuarial J* 1928;1928:13–74.

14. Petrov V. Sums of independent random variables, volume 82. New York, NY: Springer Science & Business Media, 2012.
15. Wang Q, Hall P. Relative errors in central limit theorems for student's t statistic, with applications. *Stat Sin* 2009;19:343–54.
16. Hall P. Edgeworth expansion for student's t statistic under minimal moment conditions. *Ann Probab* 1987;15:920–31.
17. Hall P. The bootstrap and Edgeworth expansion. New York, NY: Springer Science & Business Media, 2013.
18. van der Vaart AW. Asymptotic statistics, vol. 3. Cambridge, United Kingdom: Cambridge University Press, 2000.
19. van der Laan Mark J, Rosenblum M. Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. *Int J Biostat* 2009;5:1–46.