Testing Hypotheses on Simulated Data: Why Traditional Hypotheses-Testing Statistics Are Not Always Adequate for Simulated Data, and How to Modify Them

Richard Aló
Center for Computational Sciences and
Advanced Distributed Simulation
University of Houston-Downtown
One Main Street, Houston, TX 77002
Email: ralo@uh.edu

Vladik Kreinovich
and Scott A. Starks
Pan-American Center for
Earth and Environmental Studies
University of Texas at El Paso
El Paso, TX 79968, USA
Emails: {vladik,sstarks}@utep.edu

***Abstract*— To check whether a new algorithm is better, researchers use traditional statistical techniques for hypotheses testing. In particular, when the results are inconclusive, they run more and more simulations ($n_2 > n_1$, $n_3 > n_2$, ..., $n_m > n_{m-1}$) until the results become conclusive. In this paper, we point out that these results may be misleading. Indeed, in the traditional approach, we select a statistic and then choose a threshold for which the probability of this statistic "accidentally" exceeding this threshold is smaller than, say, 1%. It is very easy to run additional simulations with ever-larger n. The probability of error is still 1% for each n_i, but the probability that we reach an erroneous conclusion for at least one of the values n_i increases as m increases. In this paper, we design new statistical techniques oriented towards experiments on simulated data, techniques that would guarantee that the error stays under, say, 1% no matter how many experiments we run.**

I. Hypotheses Testing: An Important Applied Problem

One of the main uses of statistics is to compare two (or more) hypotheses. For example, we would like to check whether a new medical treatment is better than the previously known one.

Let us describe this problem in more precise terms.

Usually, an efficiency of a method can be described by an appropriate numerical quantity x. For example, an efficiency of an anti-cholesterol medicine can be described by the average amount to which its use lowers the patient's originally high cholesterol level during a certain period of time.

So, we arrive at the following problem:

- we know the average amount μ corresponding to the original hypothesis (e.g., the original treatment);
- we have the results $x_1, \ldots, x_n$ of the experiments with the new method.

Based on these results, we would like to check whether the new method is indeed better, i.e., whether for the new method, the mean value μ_x is larger than μ.

II. Hypotheses Testing: How It Is Currently Done

There are many known statistical methods for hypotheses testing; see, e.g., [5], [6]. One of these methods is as follows: we compute the population average

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

and the population standard deviation

$$\widetilde{s} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

and check whether the ratio

$$t \stackrel{\text{def}}{=} \frac{\sqrt{n} \cdot (\bar{x} - \mu)}{\widetilde{s}}$$

(called "t statistic") exceeds a certain threshold t_0.

According to the Central Limit Theorem, for large n, the distribution of the difference $\bar{x} - \mu_x$ is almost Gaussian, with 0 average and standard deviation $\sigma/\sqrt{n}$, where σ is the (unknown) standard deviation of the actual distribution, and the sample standard deviation is almost equal to σ. Thus, if $\mu_x = \mu$ (i.e., if the new method is not better than the existing one), then, for large n, the t statistic is normally distributed with 0 average and standard deviation 1. For a normal distribution, the probability of exceeding 2σ is $\approx 2.5\%$. So, if the new method is not better, we get $t > 2$ with probability $\approx 2.5\%$.

Thus, if $t > 2$, we conclude that the new method is better with certainty $100\% - 2.5\% = 97.5\%$. Similarly, if we $t > 3$, then we conclude that the new method is better with the certainty $\approx 99.95\%$, etc.

III. Theoretically, We Can Always Distinguish Between Two Hypotheses

Theoretically, if a new method is better, i.e., if the corresponding mean μ_x is larger than μ, then we will eventually

find it out as long as we perform a sufficient number of experiments. Indeed, if $\mu_x > \mu$, then, taking into consideration that $\widetilde{s} \approx \sigma$, we can describe t as the sum of the following two terms:

$$t \approx \frac{\sqrt{n} \cdot (\bar{x} - \mu_x)}{\sigma} + \frac{\sqrt{n} \cdot (\mu_x - \mu)}{\sigma}.$$

As we have mentioned, the first term is, for large n, normally distributed with 0 average and standard deviation 1; thus, it is bounded by 2 with a certainty 95%. On the other hand, when $\mu_x > \mu$, the second terms grows with n: it tends to ∞ as $n \to \infty$.

So, e.g., when n is so large that the second term is > 4, the sum is greater than 2, and hence, the above method recognizes that the new method is better. For the second term in the above sum to be larger than 4, it is sufficient to take

$$n \geq \left(\frac{4\sigma}{\mu_x - \mu} \right)^2.$$

IV. In Practice, What If Results Are Inconclusive?

In practice, in most experimental areas, each data point requires either a lot of work (as in foundational experiments related to theoretical physics) and/or a lot of risk (as in medical testing). So, researchers try to perform only as many experiments as necessary to test the new hypothesis.

With the small number of tests, the result of testing is often inconclusive (especially if a new medicine is expected to be only marginally better than the previously used one). In this case, if researchers have a reason to believe that a new method is indeed better, a reasonable approach is to perform more experiments – hoping that with more comparisons, we will be able to detect the difference between the two techniques.

Since, as we have mentioned, there is a severe limitation on the amount of experimental data, we can usually make at most one or two more iterations of this process.

V. Hypotheses Testing for Simulated Data

In computer science, we have a similar task: checking whether a new algorithm for solving a practical problem is better than the previously known algorithms.

A natural way to solve this task is to compare the performance of both algorithms on simulated data. Because of the similarity with the above situation, researchers use traditional statistical techniques for comparing algorithms. This approach was pioneered in 1990s (see, e.g., [3]); for later development, see, e.g., [1], [2], [4].

VI. What If Results Are Inconclusive? Case of Simulated Data

When the results are inconclusive, the computer science researchers currently follow the recommendations from the traditional hypothesis testing. Namely, if after n_1 simulations, the results are inconclusive, but the researchers have reasons to believe that the new algorithm is better than the old one, the researchers follow the recommendations developed for hard-to-get experimental data: they run more and more simulations ($n_2 > n_1$, $n_3 > n_2$, ...) until the results become conclusive.

VII. The Difference Between Real and Simulated Data

The main difference between real and simulated data is as follows.

For real data, each data point requires a lot of effort to get. As a result, there is a strong limitation on the amount of data that we can acquire.

In contrast, for simulated data, new data points are easy to generate. In many cases, we can easily get thousands, millions, and even billions of simulated data points.

VIII. The Resulting Problem

If after many repeated experiments with $n_1 < n_2 < \ldots < n_m$ simulated data points, researchers finally arrive at a simulation for which $t > 2$, they follow the recommendations of the traditional hypothesis testing techniques and conclude that the new algorithm is better with certainty $\geq 97.5\%$.

As we will show, this is an erroneous conclusion. Specifically, we will show that even if the new method is of exactly the same quality as the previous one, eventually, after sufficiently many simulation, we will get $t > 2$ (or $t > t_0$ for whatever threshold t_0 we select).

Indeed, for every i, let t_i describe the value of the t statistic corresponding to n_i simulations. When $\mu_x = \mu$, then, as we have mentioned, for large n, the corresponding value of t is normally distributed with 0 average and standard deviation 1. When $n_i \ll n_{i+1}$, then, as one can easily check, the corresponding random variables t_i and t_{i+1} are almost independent. So, when the sample sizes n_i grow fast enough, the resulting values $t_1, t_2, \ldots, t_m, \ldots$ are close to independent normally distributed random variables with 0 average and standard deviation 1.

Hence, in this sequence, for every i, the probability that $t_i > 2$ is approximately equal to the probability that a normally distributed random variable with 0 average and unit standard deviation exceeds the value 2 – i.e., to $\approx 2.5\% = 1/40$. Thus, on average, one out of 40 iterations leads to $t > 2$.

So, when the new method is not better ($\mu_x = \mu$), but we repeat the experiment $\gg 40$ times, with larger and larger size of simulated samples, we are guaranteed to encounter at least one case when t_i will be greater than 2 – and when, therefore, the traditional hypothesis testing technique will lead us to an erroneous conclusion that the new method is better.

Comment. The general fact that traditional statistical methods are not always adequately applicable to the statistical processing of simulated data was emphasized in several papers (see, e.g., [1]) and highlighted at several conferences, including the major 2000 International Conference on Artificial Intelligence in Austin, Texas [2].

IX. Seemingly Natural Solution to this Problem

We have mentioned that if we perform a large number of experiments m and in one of the experiments, the corresponding value of the t statistic exceed the threshold ($t_i > t_0$), we should not conclude that the new method is better, because

even when the new method is not better, t_i will eventually exceed t_0 – as long as sufficiently many experiments m are performed.

At first glance, it may look like all we have to do to rectify this situation is to limit the number of experiments. For example, for $t_0 = 2$, if we only allow $m < 40$ experiments, we should be able to avoid the above problem.

X. Does the Above Seemingly Natural Solution Work? Additional Problem Revealed

Does the above seemingly natural solution work? If we restrict ourselves to experiments with only $m < 40$ iterations, will this make the conclusions of the traditional hypothesis testing technique justified?

Let us analyze this question. According to the traditional techniques, if we encounter the case when $t_i > t_0$, we conclude that the new method is better with certainty $\geq 97.5\%$.

This conclusion makes sense if we performed a single experiment with n_1 data points. In this case, if n_1 is sufficiently large, the distribution of t_1 is normal, so the probability that $t_1 > 2$ is indeed $\approx 2.5\%$ and the probability that $t_1 \leq 2$ is $\approx 97.5\%$. In other words, if $\mu_x = \mu$, then the probability that we accidentally get $t_1 > 2$ is 2.5%. So, if we do observe that $t_1 > 2$, we conclude that the new method is better with probability $100\% - 2.5\%$.

Let us now consider what happens when we perform several (m) experiments. In each of the experiments, the probability that $t_i > 2$ is still 2.5%. However, in contrast to the traditional case of a single experiment, we now make a conclusion that the new method is better when at least one of the m values of t exceeds 2. What is the probability p that this accidentally happens when actually $\mu_x = \mu$? This probability p can be estimated as $1 - q$, where q is the probability that none of the values t_i exceeds 2, i.e., that $t_i \leq 2$ for all i.

For each i, the probability that $t_i \leq 2$ is equal to 0.975. Since the values t_i are almost independent, this probability q is approximately equal to the product of the corresponding m probabilities, i.e., to 0.975^m. Hence, $p \approx 1 - 0.975^m$:

- for $m = 1$, we get $p = 2.5\%$;
- for $m = 2$, we get $p \approx 5\%$;
- for still larger m, we get $p \approx m \cdot 2.5\%$ – up to $\approx 25\%$ for $m = 10$.

So, if perform $m = 10$ experiments and get $t_i > 2$, the probability of this accidentally happening when $\mu_x = \mu$ is not 2.5% (as researchers may erroneously conclude), it is actually 25%, 10 times larger. Thus, if we get $t_i > 2$, our confidence that the new method is better is not 97.5% – it is $100\% - 25\% = 75\%$, much smaller and much less reliable than one may think based on the uncritical application of the traditional hypothesis testing techniques.

XI. What We Propose to Do: Derivation of the New Method

We have already mentioned that when the new method is not better than the old one (i.e., when $\mu_x = \mu$), then, for large n, the distribution of the statistic t is, in effect, normal, with 0 average and standard deviation 1. Therefore, the probability that in one experiment, t exceeds some value v, is equal to $1 - F_0(t)$, where $F_0(v) \stackrel{\text{def}}{=} \text{Prob}(t < v)$ is the cumulative distribution function (cdf) of a Gaussian distribution with 0 average and standard deviation 1.

Thus, for a sample size n_i, if we select a threshold v_i, then the probability that for $\mu_x = \mu$, accidentally, the corresponding t-value t_i exceeds this threshold, is equal to $1 - F_0(v_i)$, and the probability that $t_i \leq v_i$ is equal to $F_0(v_i)$.

If we repeat experiments with several sample sizes

$$n_1 \ll \ldots \ll n_m,$$

then the corresponding statistics t_i are practically independent. Therefore, if we select the corresponding thresholds $v_1, \ldots, v_m$, then the probability q that all the corresponding t-values $t_1, \ldots, t_m$ do not exceed the selected thresholds can be estimated as the product of the of the corresponding probabilities:

$$q \approx \prod_{i=1}^{m} F_0(v_i).$$

Thus, the probability p that for $\mu_x = \mu$, one of the t values t_i accidentally exceeds the corresponding threshold v_i can be estimated as

$$p = 1 - q \approx 1 - \prod_{i=1}^{m} F_0(v_i).$$

So, if thus computed probability p is smaller than or equal to the required probability of error p_0 ($p \leq p_0$), we conclude that the new method is better with certainty $\geq 1 - p_0$.

As a result, we arrive at the following method:

XII. New Method: Summary

To test whether a new method is better than the previously used one, we do the following:

- we select a sequence of increasing sizes $n_1 \ll \ldots \ll n_m$;
- we select the certainty level $p_0 \ll 1$, and
- we select the corresponding threshold levels $v_1, \ldots, v_m$ in such a way that

$$1 - \prod_{i=1}^{m} F_0(v_i) \leq p_0,$$

i.e., equivalently, that

$$\prod_{i=1}^{m} F_0(v_i) \geq 1 - p_0.$$

Then, we sequentially perform experiments with samples of sizes n_1, n_2, ...

After performing each experiment, we compute the corresponding value t_i of the t statistic, and then:

- If we get $t_i > v_i$, we stop the experiments and conclude that the new method is better, with certainty $\geq 1 - p_0$.
- If we get $t_i \leq v_i$ and $i < m$, we continue the experiment with the sample of size n_{i+1}.

If in all m experiments, we get $t_i \leq v_i$, then we conclude that the new method is not better than the previously used one.

XIII. Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0365, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, by the Army Research Laboratories grant DATM-05-02-C-0046, and by the NIH grant 3T34GM008048-20S1.

The authors are very thankful to Hung T. Nguyen and Karen Ward for valuable discussions, and to the anonymous referees for valuable suggestions.

References

[1] P. R. Cohen, *Empirical Methods for Artificial Intelligence*, MIT Press, Cambridge, Massachusetts, 1995.

[2] P. R. Cohen, I. Gent, and T. Walsh, *Empirical Methods for Artificial Intelligence and Computer Science*, Tutorial at the 17th National Conference on Artificial Intelligence AAAI'2000, Austin, TX, July 30–August 3, 2000.

[3] I. Gent and T. Walsh, "An Empirical Analysis of Search in GSAT", *Journal of Artificial Intelligence Research*, 1993, Vol. 1, pp. 47–59.

[4] C. McGeoch, P. Sanders, R. Fleischer, P. R. Cohen, and D. Precup, "Using Finite Experiments to Study Asymptotic Performance", In: R. Fleischer, B. Moret, and M. Schmidt (eds.), *Experimental Algorithmics*, Springer-Verlag, Berlin, Heidelberg, New York, 2002, pp. 93–124.

[5] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[6] H. M. Wadsworth Jr., *Handbook of statistical methods for engineers and scientists*, McGraw-Hill, N.Y., 1990.