Traceability Challenge 2013: Statistical Analysis for Traceability Experiments

Software Verification and Validation Research Laboratory (SVVRL) of the University of Kentucky

Mark Hays, Jane Huffman Hayes
Computer Science Department
University of Kentucky
Lexington, Kentucky, USA
mahays0@engr.uky.edu, hayes@cs.uky.edu

Arnold J. Stromberg, Arne C. Bathke
Statistics Department
University of Kentucky
Lexington, Kentucky, USA
astro@ms.uky.edu, arne@uky.edu

***Abstract*—An important aspect of traceability experiments is the ability to compare techniques. In order to assure proper comparison, it is necessary to perform statistical analysis of the dependent variables collected from technique application. Currently, there is a lack of components in TraceLab to support such analysis. The Software Verification and Validation Research Laboratory (SVVRL) and the Statistics Department of the University of Kentucky have developed a collection of such components as well as a workflow for determining what type of analysis to apply (parametric, non-parametric). The components use industry-accepted R algorithms. The components have been validated using independent standard statistical algorithms applied to publicly available datasets. This work addresses the Purposed grand challenge (research project 4) and Cost-Effective Grand Challenge (research project 4) as well as the Valued Grand Challenge - research project 6.**

***Index Terms*—Traceability experiment, statistical analysis, TraceLab component, parametric tests, non-parametric tests, Purposed grand challenge, Cost-Effective Grand challenge, Valued Grand Challenge**

I. INTRODUCTION

Early traceability papers rarely applied statistical analyses as the authors were only able to examine two or three datasets and knew that such a small sample could not lead to statistically significant results. With the advent of the use of Mean Average Precision (MAP) and other "per query" measures, traceability researchers now have many more data points (a dataset that has 50 queries searching into 150 elements now has at least 50 data points versus being considered one dataset). With larger sample sizes, it is now incumbent on traceability researchers to apply statistical analyses to the dependent variables when running experiments.

This leads to the next conundrum. What statistical techniques should be used? How can traceability researchers overcome the parade of criticism from reviewers such as: "your data did not conform to the assumptions of the statistical technique used," your test did not have sufficient power," and/or "you cannot use the mean with that type of data."

This Challenge paper seeks to address some of the aforementioned concerns by providing a collection of TraceLab components that take the dependent variables from experiments (such as MAP, F, recall, precision) and determine what tests are required, check the appropriate assumptions, and run the tests. This paper contains standard language that can be used in traceability papers to demonstrate to reviewers that proper statistical analysis, designed by statisticians, has been applied.

The paper is organized as follows. Section 2 discusses statistical tests for traceability. Section 3 presents some thoughts on statistical testing. Section 4 discusses the TraceLab components developed for statistical analysis. The standard language to be used in papers employing these Statistics components is provided in Section 5. Section 6 discusses evaluation of the TraceLab statistic components, and Section 7 concludes and discusses future work.

II. STATISTICAL TESTS FOR TRACEABILITY

Currently, it is becoming more commonplace to see non-parametric techniques applied to dependent variables (such as MAP) in various experiments. Examples include Kong, Hayes, Dekhtyar, and Dekhtyar (used Wilcoxon Signed Rank test) [1], Niu and Mahmoud (used Mann Whitney) [2], and Shin, Hayes, and Huang (examined correlation of commonly used measures and their analysis) [3].

It is rare for parametric tests such as student's t or ANOVA to be applied. It appears that this is due to author fear of reviewer criticism versus due to data not meeting required assumptions (such as normality). Yet when normality and equal variance assumptions are met, appropriately chosen parametric tests are more powerful than their non-parametric counterparts and thus should be considered first. Our TraceLab components support such consideration, making statistics accessible to all researchers, even those who may not feel comfortable working with statistics.

III. STATISTICAL TESTS

Selecting an appropriate inferential method for statistical analysis is a complex and highly interactive task. Typically, there is not one correct procedure, but there are

some that are more appropriate and others less and some simply inappropriate. An expert statistician will consult diagnostic plots, test statistics, p-values, and transformations, among other tools, in order to choose a method that is adequate and powerful. Automating the process of test selection may therefore draw criticism: no automated procedure will be able to substitute expertise and experience. On the other hand, with widespread availability of free statistical software packages, the application of statistical procedures is at the fingertips of many. Many researchers simply don't have advanced statistical expertise or experience, or even quick access to expert statistics knowledge to choose the most appropriate method, or to decide when a standard method is not appropriate. The MeansTest algorithm will be useful for this group of researchers. It is designed to imitate the major decisions a statistician would make when analyzing two-sample data.

Indeed, the first decision is whether the two samples are independent or paired. If paired, then for normal data, the paired t-test [7] is the method of choice, while the signed rank test [8] is its nonparametric alternate. For independent samples, even more important than checking normality is whether it is reasonable to assume that both samples come from distributions with equal variances. If not, there exist powerful approximate methods for the normal distribution case [9-11], and for the case in which normal distributions cannot be assumed (the Brunner-Munzel test [12]).

The latter is also an example that the statistics research community continues to derive and validate new and more powerful or more robust inferential procedures, so that updates on the decision trees may have to be made. For example, for the comparison of two independent samples of non-normal data, the rank-sum test [13], [8], [14] has been the method of choice for several decades. However, it assumes that under null hypotheses, the variances of both samples are equal. Just recently, the Brunner-Munzel [12] test has been devised and validated to provide a nonparametric test for location in the presence of unequal variances. How are the decisions regarding normality and unequal variances made? Normality can be assessed using the Shapiro-Wilk [15] test. However, since the t-test is rather robust against violations of the normality assumption, an alpha-level of 5% can be chosen as a threshold. In the case of two independent samples, neither should show strong evidence of non-normality. The assumption of equal variances in the case of two independent samples is rather important and is tested using the Levene-Brown-Forsythe [16, 17] test at the 5% level.

IV. TRACELAB STATISTICAL COMPONENTS

We implemented all of the above tests as individual TraceLab components. In this section, we describe the implementation details of our composite component, MeansTest, as well as our experiences with TraceLab.

A. R Implementation

It is straightforward to calculate many test statistics, such as the t statistic for the t-test. However, most researchers are interested in the p-value of the test statistic, which expresses the significance of the result. The computation generally does not have an easily computed closed form, so implementing this step by hand is undesirable. TraceLab already links with the commercial library ALGLIB [4] that provides a limited selection of hypothesis tests that return p values. R [5], a popular statistics language, supports several tests not in ALGLIB that are relevant to our research. In the interest of maximum code reuse, we wrote our statistics in R.

We wrote a TraceLab helper component, called Rscript, which returns an opaque reference representing the R runtime. TraceLab components can use this opaque reference to execute any R script. In a TraceLab experiment, the user simply specifies the path to his or her Rscript.exe in the helper's configuration. Then the user can write a component that takes the helper as input and can use the helper's API to easily invoke their R script, which they store in their component's DLL as an embedded resource.

Each of our R-based statistics components follows a shared workflow. First, the component loads the experiment's sample data from the TraceLab workspace. Second, the component extracts and executes the R script corresponding to the statistics test in question. Finally, the component stores the resulting test statistic and p-value in the TraceLab workspace.

We unknowingly developed the ability to run R in parallel to similar efforts at the College of William and Mary. Their work, RPlugin, uses a similar technique to run R scripts, but uses a singleton pattern instead of providing a workspace variable. We only discovered this overlap inadvertently [6] and very late in development, so it should be interesting to compare the two implementations in future work. For now, we turn our attention to the main novelty, the MeansTest component.

B. MeansTest Implementation

Although R provides implementations for all of the tests mentioned in section III, R assumes that the user is a statistics savant who is aware of all of the assumptions that the tests entail. Our goal is to reduce the user's burden by cataloging and automatically testing these assumptions. TraceLab helped us in this respect by providing a very useful feature called *composite components*. The composite component wizard in TraceLab enables researchers to take a subset of an existing experiment and encapsulate it as one component. Using this wizard, we developed a composite component called MeansTest. The MeansTest component takes just a few parameters: the two samples the user is comparing, a flag specifying whether the samples represent paired data, and the Rscript opaque reference. After executing an experiment containing a MeansTest, TraceLab stores the p-value of the test and the appropriate test statistic in the workspace. MeansTest then prints a human-readable summary of the steps and tests involved in the computation.

Figure 1 shows the TraceLab dependency graph of MeansTest. As can be seen, MeansTest automatically verifies all of the assumptions one would normally have to check before performing a comparison of location parameters. First,

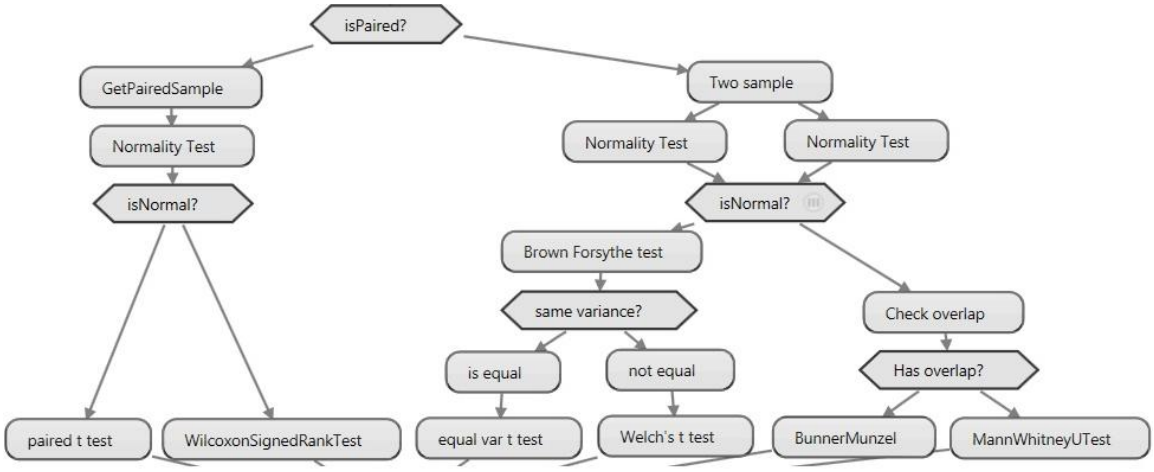

Fig. 1. Internals of the MeansTest composite component.

MeansTest checks whether the user provided paired data. In the paired case, MeansTest branches to the left to test the normality of the pairs with Shapiro-Wilks. If the difference between the pairs is normally distributed, MeansTest performs the paired t-test. Otherwise, MeansTest performs the nonparametric Wilcoxon signed-rank test.

In the case the samples were not paired, MeansTest branches to the right and tests the normality of the two sample groups separately. If both groups are normal, MeansTest compares the sample variances for equality. If there is no evidence against this assumption, it performs a classical t-test using the pooled sample standard deviation; otherwise, it makes Welch's adjustment to the standard deviation when performing the t-test.

If either of the groups is not normal, MeansTest branches into the nonparametric tests. MeansTest checks the assumption of overlap between the samples. If this assumption holds, MeansTest simply invokes the Brunner-Munzel non-parametric test, which is designed for testing location differences in the presence of possibly unequal variances. When there is no overlap, then clearly the difference is significant, so MeansTest invokes the two-sample rank sum test (Mann Whitney U test) only to provide the researcher with a non-zero p-value.

MeansTest is by no means a complete summary of all possible statistics tests, but is representative of the involved thought process we use in practice when comparing means. There are many other tests for normality, equal variance, and shift in location that fit specialized circumstances. There are also some assumptions, such as independence and identical distributions, which statisticians have yet to invent ways to numerically verify. We hope the TraceLab dependency graph of MeansTest will start a dialog in the statistics community to agree on a complete process for testing for shifts in means.

C. *Experiences*

In general, we found that the TraceLab tool was very stable and facilitated a wide variety of experiment procedures. As we mentioned earlier, composite components proved useful for bundling our massive statistics workflow into one comprehensive (and comprehensible!) statistics test. Besides applications in statistics, we found other uses for TraceLab. For instance, we were able to implement a classical mutation testing experiment comparing all-definitions testing to random testing. Mutation testing experiments are entirely outside the scope of TraceLab, yet the tool proved to be a plausible fit. Although the component framework adds extra work to experiment implementation, it is our experience that this extra work leads to portable experiments with reproducible results.

While the core tool provides a nice framework for developing experiments, we discovered several major issues indicating that the stock components are still experiencing growing pains. For instance, we identified a computation failure in the TraceLab vector space model where it reported that the cosine similarity between a vector and itself was far less than 1.0 [18]. This failure resulted from an error in the TF-IDF computation where the authors were normalizing the document vectors but not updating their pre-computed lengths. This mistake adversely affected the similarity measures; for example, Equation 1 gives the resulting erroneous cosine similarity:

$$\cos(q,d) = \frac{q \cdot d}{|q||d|^2}. \quad (1)$$

The other measures were similarly impacted. For obvious reasons, this error invalidates the results of every tracing experiment run in TraceLab 0.5 or earlier using the default tracing components.

Errors like these aside, we see the need for design and documentation improvements to the stock components as well. For instance, it is not possible to extract the per-artifact recall and precision scores from the experiment results data type; the data type hides these scores from the public API in the form of summary statistics. This API makes it impossible to perform meaningful analysis of experiment results. Also, the file importer and exporter descriptions provide no hint as to the expected file format. The worst offender is the "multiple dataset importer," which takes a "configuration file" as input. It would be useful to provide the expected formats in the components' descriptions.

Changes to the design of these components will definitely help improve productivity in TraceLab, but more work is needed. First, TraceLab needs to better advertise the

availability of third-party components to collaborators. The Rscript/RPlugin overlap mentioned earlier is a perfect example of this necessity. We hope that the new Component Directory on coest.org [19] will help improve code reuse to avoid collisions like these.

Another key obstacle to productivity is that all operations, no matter how trivial, need to be encapsulated in their own components. For instance, to test the normality assumption in the paired case, one usually computes the difference between the two paired samples and tests the normality of the resulting vector. In R, this is very easy; if you have two vectors x and y, the expression x-y will return the input vector. However, TraceLab did not have a component to compute x-y, so we had to write our own x-y as a separate TraceLab component (see GetPairedSample in Figure 1) consisting of 99% TraceLab boilerplate and 1% actual code. We postulate that the existing decision nodes, which support inline scripts for making branching decisions, can be repurposed to avoid this boilerplate. To this end, we would like to see better documentation of decision nodes describing the available variables, the process to save workspace variables, and the particular .NET dialect in which decision code is written. Perhaps the TraceLab developers could help us create a facility to script R code inline as well.

V. STANDARD LANGUAGE FOR PAPERS

"We used the statistical analysis components available in TraceLab. These were designed by computer scientists and statisticians at the University of Kentucky and use the well-respected statistical analysis toolkit R. The TraceLab components, collectively called MeansTest, first examine the paired or independent variables for the experiment and determine what statistical tests to apply by testing the appropriate assumptions. Next, the TraceLab components apply the appropriate statistical test. The components then report the appropriate p-value. This information has been included below. Details on the statistical analysis methodology applied by the TraceLab component can be found in an earlier publication by Hays et al." (with proper citation of this TEFSE paper).

In addition, the researcher shall use the output from MeansTest to describe the tests applied. For each outcome, we include standard text below that can be included in research papers (filling in the bracketed and dotted sections as appropriate), with proper citation of this TEFSE paper.

Outcome: Brunner-Munzel Test

"The two techniques were compared using two independent samples of {add more details}. We tested for normality in each sample, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that normality could not be assumed. Therefore, the nonparametric Brunner-Munzel test (Brunner and Munzel 2000) was chosen for further analysis. This test is specifically designed to compare the location of two samples in the possible presence of unequal variances. It does not assume normality. However, it assumes that the two underlying populations have overlapping support. This assumption is met since the observations in both samples overlap. The Brunner-Munzel test resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

Outcome: Two-Sample Rank-Sum Test

"The two techniques were compared using two independent samples of {add more details}. We tested for normality in each sample, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that normality could not be assumed. Also, the observations in both groups are totally separated. Therefore, the nonparametric two-sample rank-sum test (Deuchler 1914, Wilcoxon 1945, Mann and Whitney 1947) was chosen for further analysis. It compares the location of two samples without assuming normality. The two-sample rank-sum test resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

Outcome: two-sample t-test for equal variances

"The two techniques were compared using two independent samples of {add more details}. We tested for normality in each sample, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that there was no evidence against the normality assumption. Using the Levene-Brown-Forsythe test (Levene 1960, Brown and Forsythe 1974), we tested whether variances could be assumed equal for both groups. There was no evidence against this assumption (at the 5% level). Therefore, the t-test for two samples with equal variances (Gosset "Student" 1908) was chosen for analysis. It resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

Outcome: two-sample t-test for unequal variances (Satterthwaite-Smith-Welch approximation)

"The two techniques were compared using two independent samples of {add more details}. We tested for normality in each sample, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that there was no evidence against the normality assumption. Using the Levene-Brown-Forsythe test (Levene 1960, Brown and Forsythe 1974), we tested whether variances could be assumed equal for both groups. We concluded that this was not the case (at the 5% level). Therefore, the t-test for two samples with unequal variances (Satterthwaite-Smith-Welch approximation; Smith 1936, Welch 1938, Satterthwaite 1946) was chosen for analysis. It resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

Outcome: paired t-test

"The two techniques were compared using two paired samples of {add more details}. The differences between both samples were tested for normality, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that there was no evidence against the normality assumption. Therefore, the paired t-test (Gosset "Student" 1908) was

chosen for further analysis. It resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

Outcome: Wilcoxon signed rank test

"The two techniques were compared using two paired samples of {add more details}. The differences between both samples were tested for normality, using the Shapiro-Wilk test (Shapiro and Wilk 1965), and concluding (at the 5% level) that normality could not be assumed. Therefore, the Wilcoxon signed rank test for paired samples (Wilcoxon 1945) was chosen for further analysis. It resulted in a test statistic of … and a p-value of …, meaning that a significant difference between both techniques can be concluded {no evidence for a significant difference between both techniques was found}."

VI. EVALUATION

In order to vet the MeansTest components, we ran an independent evaluation. The dependent variable measures output by TraceLab from a typical traceability experiment on one dataset with comparison of techniques (collection of MAP values for a traceability dataset for TF-IDF with stopwords removed and MAP values for TF-IDF on that same dataset without stopword removal) was provided to the Statistics department co-authors of this paper. They independently analyzed the data using publicly available tools such as SAS (not R) and derived p-values (these are shown in Table 1). We generated t and p-values using MeansTest, also shown in Table 1. As expected, the values are within rounding error of each other. The t-distribution in this context is symmetric around zero, so the difference in sign simply reflects a minor implementation difference between their tools and R.

TABLE I. EVALUATION RESULTS.

	Statistics Department values	**MeansTest values**
t	0.346225	-0.3462248
p-value	0.7371	0.7371301

VII. CONCLUSIONS AND FUTURE WORK

As the traceability research community ushers in the era of TraceLab, it will be much easier to generate and try out new ideas. It is incumbent upon researchers to practice responsible experimentation and use proper techniques in ensuring that the obtained results are statistically significant. Toward that end, we present the MeansTest component as well as standard language that can be used in papers which employ this composite TraceLab component. We evaluated our component and found that the values generated match those of SAS.

In the future, we would like to expound on other statistical analyses such as power analysis and analysis of variance. As with the comparison of means, researchers performing other analyses of their results have many options readily available thanks to statistical software packages. Unfortunately, researchers often lack the required expertise to select the most appropriate option. While a fully automated solution to the selection process is not a panacea, we posit that such a solution can imitate the decisions of an expert in most applicable cases.

ACKNOWLEDGMENT

This work was funded in part by the National Science Foundation under grant CCF-0811140 (research) and ARRA-MRI-R2 500733SG067 (benchmark development for Tracelab). We thank Wenbin Li for assistance with the datasets.

REFERENCES

[1] W-K Kong, J H Hayes, A Dekhtyar, O Dekhtyar, "Process improvement for traceability: A study of human fallibility." Requirements Engineering Conference (RE), 2012 20th IEEE International, vol., no., pp.31-40.

[2] N Niu, A Mahmoud, "Enhancing candidate link generation for requirements tracing: The cluster hypothesis revisited," Requirements Engineering Conference (RE), 2012 20th IEEE International, vol., no., pp.81-90, 24-28 Sept. 2012.

[3] Y Shin, J H Hayes, J C Huang, "A Framework for Evaluating Traceability," DePaul University Technical Report, TR12-001, http://www.cdm.depaul.edu/SoC/research/Documents/TechnicalReports/2012/TR12-001.pdf.

[4] ALGLIB, Cross Platform Numerical Analysis Library, http://alglib.codeplex.com/, last accessed February 7, 2013.

[5] The R Project for Statistical Computing, http://www.r-project.org/, last accessed February 7, 2013.

[6] B Dit, E Moritz, D Poshyvanyk," A TraceLab-based solution for creating, conducting, and sharing feature location experiments," in Proceedings of the 20th IEEE International Conference on Program Comprehension (ICPC), 2012.

[7] W S Gosset "Student" (1908), "The probable error of a mean," Biometrika 6, 1, pp.1-25.

[8] F Wilcoxon (1945), "Individual comparisons by ranking methods," Biometrics Bulletin 1, 6, pp.80-83.

[9] F E Satterthwaite (1946), "An approximate distribution of estimates of variance components," Biometrics Bulletin 2, pp.110-114.

[10] H Smith (1936), "The problem of comparing the results of two experiments with unequal errors," Journal of the Council for Scientific and Industrial Research 9, 211-212.

[11] B L Welch (1938), "The significance of the difference between two means when the population variances are unequal," Biometrika 29, pp.350-362.

[12] E Brunner, U Munzel (2000), "The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation," Biometrical Journal 42, pp.17-25.

[13] G Deuchler (1914), "Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie," Zeitung für pädagogische Psychologie 15, pp.114-131, 145-159, 229-242.

[14] H B Mann, D R Whitney (1947), "On a test of whether one of two random variables is stochastically larger than the other," Annals of Mathematical Statistics 18, 1, pp. 50-60.

[15] S S Shapiro, M B Wilk (1965), "An analysis of variance test for normality (complete samples)," Biometrika 52, 3-4, pp.591-611.

[16] H Levene (1960), "Robust tests for equality of variances," In Ingram Olkin, Harold Hotelling, et alia. Stanford University Press, pp. 278-292.

[17] M B Brown,A B Forsythe (1974), "Robust tests for equality of variances," Journal of the American Statistical Association, pp.364-367.

[18] "Bug #200: VSM returns strange values - TraceLab - Projects." [Online]. Available: http://coest.org/coest-projects/issues/200. [Accessed: 12-Feb-2013].

[19] "Directory | Coest." [Online]. Available: http://coest.cstcis.cti.depaul.edu/index.php/tracelab/component-directory. [Accessed: 12-Feb-2013].