

Re: Not sure how to approach this. (did stats 15 years ago)

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2004-09/0421.html>

From: Richard Ulrich (*Rich.Ulrich_at_comcast.net*)

Date: 09/23/04

Date: Thu, 23 Sep 2004 01:34:49 -0400

On Wed, 22 Sep 2004 21:15:05 -0600, Rithban <rithban@yahoo.com> wrote:

- > *I'm doing performance testing on different software configurations, and need*
- > *to defend conclusions and decisions. I did stats in college about 15 years*
- > *ago, but have spent a fair amount of time reviewing basics from on-line*
- > *material.*
- >
- > *I have three of questions regarding how I should approach testing*
- > *null-hypothesis re: whether means have changed significantly between*
- > *populations. In my terms I want to know "I tweaked this aspect of the*
- > *software. Did it do any good or was it a waste of time?" Since these require*
- > *programming changes and not just changing constants, it's important to our*
- > *project.*
- >
- > *I ran a baseline off of the original software. Collected 44 sets of 18,000*
- > *data points (timing results).*

What is that, 44 different test problems to benchmark?

18,000? Your program is writing out results for every pass through some loop? -- I've seen cases where the loop-reporting I/O took more CPU time than the algorithm.

Wherever it comes from, and whatever its immediate bad effects -- even if no bad effects are there -- you want to figure how to reduce the 18,000 to something much smaller. Does it make sense to take a few totals or averages?

- >
- > *I'm systematically tweaking different parameters to compare against the*
- > *baseline and each other. Again, about 44x18k timing results per run. The*
- > *sequences are identical and produce an identical number of data points.*
- >
- > *QUESTION #1*

sci.stat.math: Re: Not sure how to approach this. (did stats 15 years ago)

- > *Given, say, two populations of data points with a roughly normal*
- > *distribution, is it appropriate to do a simple t-test? From what I*
- > *understand, t-test assumes equal variances, which is not necessarily true*
- > *between the populations. The variances are not horribly different, but they*
- > *seem to shift according to mean value. I guess I don't know how concerned I*
- > *should be about differences in variance between populations.*

If variances are proportional to means, and if you are looking at computer timings, it does make sense that performance should be measured as proportional. That suggest taking the logarithms, especially if some numbers are 10 or 20 or 100 times as large as some other numbers in the data.

On the other hand: For equal Ns, the t-test is rather robust against different variances.

- >
- > *QUESTION #2*
- > *Given multiple populations of data points (3 now, more coming) with roughly*
- > *normal distributions, should I do one test between each set (geometric*
- > *explosion of combinations) to determine whether there is a significant shift*
- > *in mean time? I can do this programmatically.*

Mean of 18,000? Mean of 18,000 times 44?

Means are handiest when they are means of numbers that represent the same central value. Sums may be more indicative, if you are looking for "total performance", and you want to make it clear that the smallest numbers have little contribution.

- >
- > *I'm still not clear about where ANOVA fits in; I've read through about six*
- > *different sets of introductory material on-line. It seems relevant but still*
- > *don't get it. Given populations A, B and C, I want to test the null*
- > *hypothesis of same population means between A-B, A-C, and B-C. When the next*
- > *set of runs finish, I'll have A, B, C, and D, which mean A-B, A-C, A-D, B-C,*
- > *B-D, and C-D. We have two more after that planned (!!!).*

An overall ANOVA tests whether the variation among the four is more than chance. If you know that the overall differences exist, you might be able to skip that. Or, if you are doing 44 different benchmarks, you do the 44 ANOVAs to see which ones do matter. If you have an ANOVA stat-package, it should allow special contrasts, which you can select.

I think you are talking about ultimate presentation here, almost as much as the testing. How does this sound – Test everything separately against "Baseline." Ask for Tukey's test, which orders the means, and then draws lines underneath sets (or something functionally

Re: Not sure how to approach this. (did stats 15 years ago)

sci.stat.math: Re: Not sure how to approach this. (did stats 15 years ago)

similar) to show which ones are **not** different. That gives all the tests.

>

> *QUESTION #3*

> *I have some tests that produce distributions with means near zero that are heavily skewed (tails to the right along the X axis), kind of like some Poisson distributions. I'm kind of clueless on how to test a null-hypothesis that the population means are not significantly different.*

This is an problem for your interpretation. If it is the same for all 3 groups, then maybe it is not serious. If it's not skewed after taking a log, then maybe it is not serious. But if just **one** algorithm is subject to long tails, then it could be that you have de-tuned the overall performance, and you really do want to have the raw outliers in there, stretching out the mean for the poor group.
– What do you think of an algorithm that is 'usually' faster, but 'occasionally' gets hung up? What can you figure about the cause of being slow?

The t-test will still give a test, but the question could be whether the **mean** is still meaningful. When I suggested taking a sum of some of the 18000, I had in mind that it could smooth out some of the idiosyncracies of performance.

>

> *I hope that these questions are not too vague. I don't even know enough to know how to phrase the questions with any degree of precision.*

I hope I've helped with that –

--

Rich Ulrich, wpilib@pitt.edu
<http://www.pitt.edu/~wpilib/index.html>