

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

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2004-09/0438.html>

From: Rithban (rithban_at_yahoo.com)

Date: 09/23/04

Date: Thu, 23 Sep 2004 11:06:10 -0600

> *What is that, 44 different test problems to benchmark?*

Yes. Each unique. Testing a large framework. Two object categories, seven configurations which include three important and independent functions, and one constant parameter (data set). So five items in 44 combinations.

Fortunately as I look at things deeper, about 1/4 of the combinations have no meaningful data because operating system effects overshadow the algorithms. For the most part about another 1/3 show no significant variance. Well, this is what I was testing for, but it's good to know.

> *18,000? Your program is writing out results for
> every pass through some loop? ...cases where...I/O took more CPU time*

Your objections are very, very reasoned. I need that many data points to squish a degree of intercontinental office politicking, where irrationality overrides reason.

No, I'm handling 18k separate examples of real-world data we swiped from the lab — with permission. :-) Technically I only need a subset of the 18k. The first part of the data has a linear drift in four of the seven configurations.

However, there are distinct behaviours that happen over time (varying by volume of data). We *do* have to have dump in around 10k experiments (each generating 44 data points) before the timings converge on an upper limit.

I readily split these into two groups and deal with the linear regression portion. I'm quite familiar with that since I used to be involved in data analysis that required this.

The tail portion of the remaining 8–10k data points are needed because we need to push several GB of data through the system because 4GB is a threshold for irrational paranoia. (No, there is no 4GB limit. See? This database holds 5GB. No, no problems. See? Error log reports nothing. All tests pass. The performance tests show no burps.)

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

- > *For equal Ns, the t-test is rather robust*
- > *against different variances.*

Good to know. I wrote a trivial Octave program to loop through and display a binary pass/fail based on t-test. It seems to work great.

- > *I think you are talking about ultimate presentation here,*

Yes, I have to boil it all down. Good ideas.

- >> *distributions ... that are heavily skewed*

...

- > *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.*

I've stared at things for a while longer, and realized that most of the very low data points contain no useful information. They're directly analogous to signals below the sensitivity of the sensor. There worst-case configurations do present strong means above the noise. They have long tails, but for our purposes they may not significant. Well, except in the baseline which shows a nasty worst-case behaviour that gets progressively worse over time. I'm glad that we can definitively show that it's gone.

- >> *I hope that these questions are not too vague.*

>

- > *I hope I've helped with that –*

Yes, you've given a lot to think about, and made some excellent points that cleared a lot of fog.