Overview

- *The paper in as few words as possible:* Stats 101 for computer scientists.
- *Key observation:* Managed runtime systems introduce a lot of variability.
- Performance evaluations need to consider two orthogonal factors:

  Experimental Design: be wary of (i.e. diversify) the benchmarks, inputs, VM, heap size, JIT settings, & hardware platform.

  Data Analysis: consider randomness due to JIT compilation, garbage collector, thread scheduling, VM timer-based sampling.

- The rest of the paper:
  - Review statistical methods
  - Recommend a suitable methodology
  - Use methodology to evaluate old measurements (examine 50 papers published in 2000 – 2007)

Recommended Methodology

- Measure startup *and* steady-state performance.
  - Startup performance is affected by class loading and JIT compilation.
- Try the JavaStats tool: it runs multiple VMs per invocation and monitors confidence levels.

  Disclaimer: JavaStats last updated in 2007. The benchmarkr repo from 2011 seems a little better.

Measuring Startup Performance

1. Measure invocation time of multiple VM instances
2. Compute the confidence interval of these execution times

Measuring Steady-State Performance

1. Run a many VMs for many iterations. Be willing to ignore some measurements per iteration.
2. Record when each VM reaches steady-state performance (i.e., once coefficient of variation — the measured variance over the measured mean — falls below your favorite epsilon).
3. Compute the (geometric?) mean of k benchmark iterations under steady-state.
4. Compute the confidence interval of the computed means.
5. Derive overall mean and confidence interval from the collected measurements.

Stats 101 Reference

Types of Error

Systematic Errors are due to the experimental setup.

Random Errors are out of our control, but we can identify their aggregate effect.

Confidence Intervals

A 95% confidence interval means that we are 95% certain that the *true proportion* falls within the given interval.

- This works because of the Central Limit Theorem. For $n \gtrsim 30$ measurements, the distribution of our *measurements* will model a normal (gaussian, bell) distribution, regardless of the sample space.
- For fewer than 30 measurements, use Student's t-test, which uses a heavy-tailed version of the normal distribution.

Comparing Two or More Alternatives

- If confidence intervals for two sets of measurements overlap, we *cannot conclude* a statistically significant difference between the alternatives.
  - i.e., the differences we observed may just be due to random fluctuations
- If confidence intervals *do not* overlap, we conclude there is *no evidence to suggest there is not* a statistically significant difference.
  - If the confidence interval of the difference between two observed means includes zero, then we may conclude there is no statistically significant difference between the alternatives.
- *Analysis of Variance* (ANOVA) separates total variation in a set of measurements into a component due to randomness and a component due to actual differences.
  - Randomness identified by observing variance within the measurements for one alternative.
  - Actual differences identified by comparing measured variances across alternatives.
  - If the actual differences exceed the random differences, we conclude that there is a statistically significant difference between the alternatives.
  - **Caveat:** ANOVA assumes that variance for measurement errors is uniform across alternatives, and that errors are independent and follow a Normal distribution.
- An F-Test is used to determine if two variances are statistically different. Computes the ratio between the variances, further from 1 is better.
- MANOVA permits variation of multiple inputs.
- The Coefficient of Variance is the measured variance divided by the measured mean. It is used to identify a steady state of measurements.
- Violin Plots have a *dot* at the median value, a *thick line* over the first and third quartiles of data, a *thin line* covering outliers, and width in proportion to the distribution's probability density at a point.