Benchmarking and Timing

James H. Davenport

6 March 2025

1 General comments on benchmarking

Laptops are dangerous for benchmarking, as they can change the frequency (and hence speed) depending on their temperature. A colleague saw a quite bizarre sinusoidal shape in his timing. Not there when he repeated the tests on a desktop, and the cause was apparently the laptop heated up, so slowed down (time increased) and hence cooled down, so sped up (time decreased) and thsi repeated.

Though it's tempting to do other work, you should leave the desktop alone which it's benchmarking, as other processes running (even if the OS were perfect at doing time allocation, which they aren't always) will take up cache in unpredictable ways. A Java Garbage collection can really trash another processes cache behaviour.

JHD is generally suspicious of times less than 10ms, as OS measurement granularity is not always much better than that.

2 Timing

Q How many measurements should I take?

A It depends on the variability of your configuration, and what you are measuring. Pure CPU times are liable to be much more accurate than network round-trips for example.

JHD's usual behaviour is thus.

1. Take five measurements and compute the mean m and standard deviation σ of this sample.
2. If $100\sigma < m$ (i.e. we are probably accurate to 1%), then JHD is generally satisfied with five repeats.
3. Let $N = \lceil \left(\frac{100\sigma}{m}\right)^2 \rceil$, since the standard deviation depends on the square root of the number of samples.

4. Take $5N$ measurements and compute the mean m_1 and standard deviation σ_1 of this sample. We would expect $100\sigma_1 \approx m_1$. If so, $5N$ is a good number of samples.

5. Otherwise adjust N and repeat.

3 Outliers

Occasionally, you will get measurements that "just seem wrong" — typically much larger than expected. But you shouldn't just decide "I don't believe this". My usual advice is "you should probably assume a normal distribution (since each "error" is normally a sum of a lot of small errors, so should be roughly normal) and throw away those that are further away than you'd expect" — "expect" in terms of standard deviation, abbreivated "s.d.".

- So for example 95% are expected to be within 2s.d. of the mean, so with < 20 samples, you could be pretty confident rejecting anything that was more than 3s.d. away

- And 99.7% are expected to be within 3s.d. of the mean, so with ¡300 samples, you could be pretty confident rejecting anything that was more than 4s.d. away.

Note that, after rejecting, you should recompute the s.d., as it will have decreased.

While checking this, I learn that this is formalised in Grubb's test: `https://en.wikipedia.org/wiki/Grubbs%27s_test`