08 Experimental performance evaluation

EN0746 Computer Network Implementation

David Kendall

Northumbria University

Introduction

- Often need to determine the performance of some aspect of a computer network, e.g.: hardware component, algorithm, protocol
  - Will component X provide adequate performance for our requirements?
  - Is protocol Y better than protocol Z in these circumstances?
  - . . .
- Three main approaches to answering these sort of questions
  - Analysis
    - Build an abstract mathematical model of the component of interest and its interaction with the rest of the system and use the model to derive statements about its behaviour, e.g. queuing theory
  - Simulation
    - Build a computer model of the component of interest and its interaction with the rest of the system; collect and analyse data derived from simulated ’runs’ of the model to make predictions about the likely behaviour of the ’real’ component, e.g.: OPNET, ns2
  - Experiment (The focus of this lecture)
    - Observe (measure) the behaviour of the component of interest in a ’live’ environment. Analyse the data to make predictions about its likely behaviour ’in the long run’, e.g.: iperf

Questions for the experimenter

- What data to collect?
- How to collect the data?
- How to summarize the data?
- How confident that the summary of the sample data is a ’true’ representation of the population?
- How to test a hypothesis?
  - e.g. regarding alternative protocols, implementations etc.

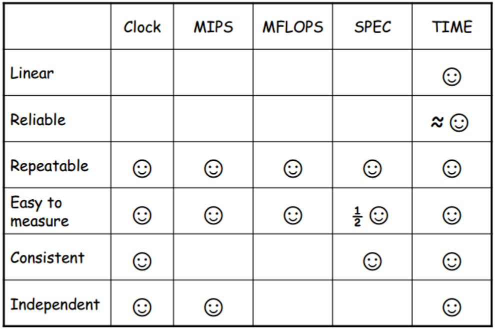
What data to collect (metrics)

- Count
  - of how many times an event occurs
- Duration
  - of a time interval
- Size
  - of some parameter
- Value derived from these fundamental measurements

Characteristics of a good metric

- Linear
  - i.e. if metric increases x2, performance increases x2
- Reliable
  - if metric of A > metric of B, then performance of A > performance of B
- Repeatable
  - Running the same experiment multiple times gives the same result
- Easy to use
  - Metric hard to obtain, then more likely to be wrong
- Consistent
  - e.g. units and definitions are consistent across systems
- Independent
  - resistant to pressure from manufacturers to optimize results for their systems

Summary of metrics



	Clock	MIPS	MFLOPS	SPEC	TIME
Linear					☺
Reliable					≈ ☺
Repeatable	☺	☺	☺	☺	☺
Easy to measure	☺	☺	☺	$\frac{1}{2}$ ☺	☺
Consistent	☺			☺	☺
Independent	☺	☺			☺

Figure from Setia, S., lecture 1 in CS 700 - Quantitative methods and experimental design in CS, George Mason University, 2009

Can also use *time-normalized* metrics: “rate”, “throughput”

- Transactions per second
- Bytes per second
- Queries per second

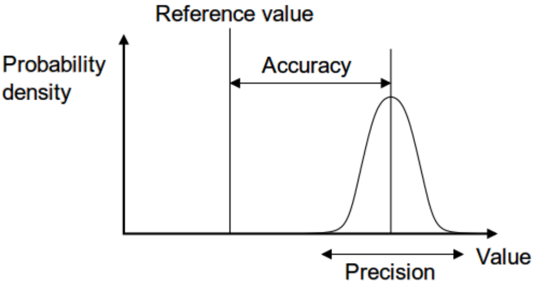
Summary of metrics

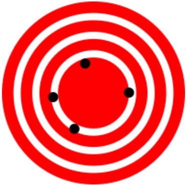
- Primarily interested in how much of some limited resource must be used in order to obtain some benefit
- Resource
  - Time – between start event and finish event (define events carefully, ensure time of occurrence can be detected accurately and precisely)
  - Space – processor registers, CPU cache, RAM, Flash, hard disk, tape, . . .
- Benefit
  - Computed result (output), e.g. result of program execution, response to request for service, completion of transaction . . .
  - Beware of metrics that measure "means" (e.g. clock cycles, MIPs, MFLOPS . . .) rather than "ends" (e.g. transactions, queries, results, . . .).

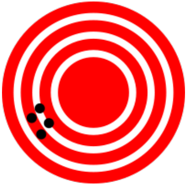
Accuracy, precision, and resolution

- Real-world effects introduce *uncertainty* into measurements.
- Three important features that characterise the quality of measurements are:
  - Accuracy: an indication of the closeness of measurements of a quantity to that quantity's actual value according to some well-defined standard.
  - Precision: the degree to which repeated measurements under unchanged conditions give the same results.
  - Resolution: the smallest change in the underlying physical quantity that produces a noticeable response in the measurement.

Distinguishing accuracy and precision




Accurate but not precise

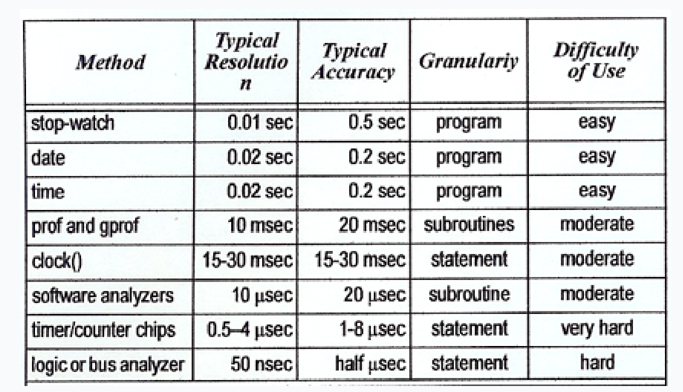

Precise but not accurate

Figures from http://en.wikipedia.org/wiki/Accuracy_and_precision

Measurement errors

- Difficult to separate individual contributions made to error by accuracy, precision, and resolution
  - Quantifying accuracy is hard, e.g. quantifying accuracy of interval timer requires calibration of clock source with standard measurement of time
  - Use variance of measurements to quantify precision
  - Resolution usually easy to quantify, e.g. resolution of interval timer can introduce an error of ± 1 clock period.
- Other sources of error include:
  - Perturbing the quantity to be measured by the act of measuring it, e.g. program statements added to a program to access a timer change the behaviour of the program being measured
  - Other non-deterministic events may also perturb the quantity to be measured
- Useful to classify errors as either systematic or random

Methods for measuring execution time



Method	*Typical Resolution*	*Typical Accuracy*	*Granulariy*	*Difficulty of Use*
stop-watch	0.01 sec	0.5 sec	program	easy
date	0.02 sec	0.2 sec	program	easy
time	0.02 sec	0.2 sec	program	easy
prof and gprof	10 msec	20 msec	subroutines	moderate
clock()	15-30 msec	15-30 msec	statement	moderate
software analyzers	10 µsec	20 µsec	subroutine	moderate
timer/counter chips	0.5–4 µsec	1-8 µsec	statement	very hard
logic or bus analyzer	50 nsec	half µsec	statement	hard

Figure from David B. Stewart, Measuring execution time and real-time performance (part 1), Dr. Dobbs Journal, November 2006 (available here)

What to do with the data?

- A set of experimental measurements is a *sample* of the underlying system that is being measured
- We want to make correct inferences from our sample about the underlying system
- We use *statistical techniques* to do this
- We also use statistical techniques to attempt to quantify the imprecision in our measured data due to random experimental error

Summarizing measured data

How

- Central tendency: arithmetic mean, harmonic mean, geometric mean, median, mode, . . .
- Variability: variance, standard deviation, . . .
- Distribution

Why

- Convenience of reducing performance to a single number
- Simplicity of comparing different systems
- BUT performance is often multi-dimensional and systems are often specialized
- When a single number must be given as a summary, you should also show a graph of the distribution of measured data, e.g. histogram

Central tendency: Arithmetic mean

- Assume x_i are values observed for discrete random variable X generated by some random process
- Expected value of $X = E[X]$

$$E[X] = \sum_{i=1}^{n} x_i p_i$$

  where n is the number of measurements and $p_i = Pr(X = x_i)$
- In absence of further information, assume probabilities are all the same, i.e. $p_i = 1/n, 1 \leq i \leq n$
- Arithmetic mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

More on means

- Arithmetic mean is appropriate for summarizing times
  - Directly proportional to total time, e.g. time doubles $\Longrightarrow$ mean value doubles
- Arithmetic mean is *not* appropriate for summarizing rates
  - Gives value that is proportional to sum of inverse of times
  - We want value that is inversely proportional to sum of times
- Harmonic mean

$$\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- Harmonic mean *is* appropriate for summarizing rates
  - times doubles $\Longrightarrow$ mean value halves (assume constant work)
- See (Lilja, 2000) for further explanation

Variability: Standard deviation

- Means hide information about variability
    - How spread out are the values?
    - How much spread relative to the mean?
- Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

- Gives a summary of the difference between measured data values and the arithmetic mean

Distribution of data

- A *distribution* is a summary of the values that appear in a sample and the frequency or probability of each
- A distribution on a continuous variable X can be defined by a *probability density function*
- $p(x)$ is a probability density function if the probability that tne value x of the variable X lies between the values a and b is given by the area under the graph of $p(x)$ between a and b, i.e.

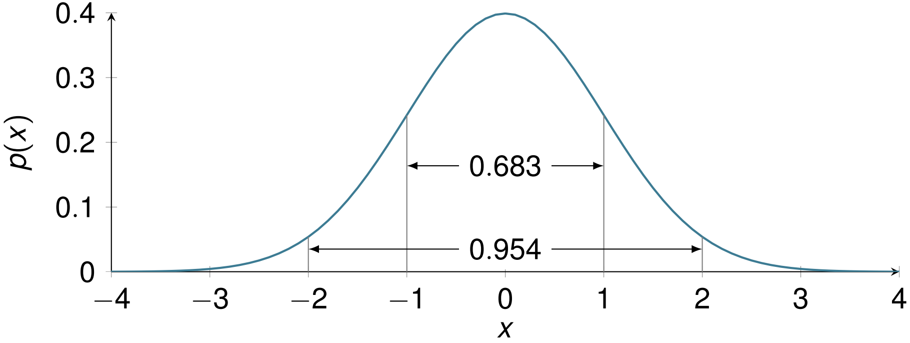
$$Pr[a <= x <= b] = \int_a^b p(x)dx$$

and that the area under the graph of $p(x)$ between $-\infty$ and ∞ is 1, i.e.

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Standard Normal Distribution

- Mean: 0
- Standard deviation: 1
- Probability density function



- Probability that value is within 1 SD of mean is 0.683
- Probability that value is within 2 SDs of mean is 0.954

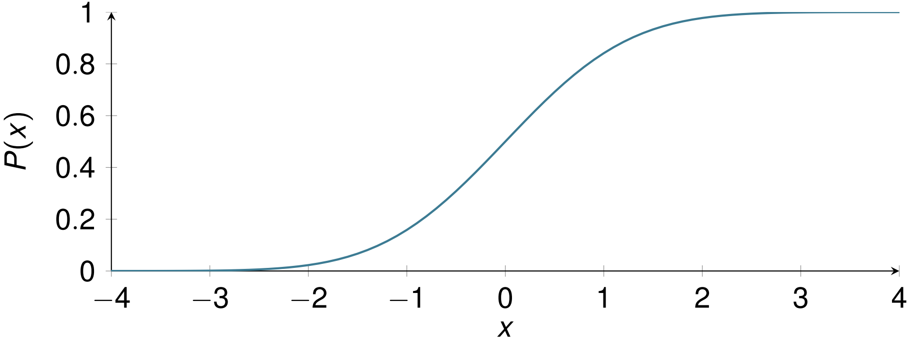
Cumulative distribution function

- The probability that the continuous random variable X takes on a value less than or equal to x, $Pr[X \leq x]$, is given by the *cumulative distribution function*, P, defined for probability density function, p, by:

$$P(p, x) = \int_{-\infty}^{x} p(t)dt$$

- i.e. the cumulative distribution function gives the area under the curve of the probability density function, p, from $-\infty$ to x

Cumulative distribution function for $N(0,1)$



- $NORM.DIST(x, 0, 1, TRUE)$ in Excel

Confidence interval for the mean

- If the distribution of random errors in our measurements can be approximated by a normal distribution, we can use the unique properties of this distribution to determine how well our estimate of the true value approximates the actual true value
- Specifically, we use *confidence intervals* to find a range of values that has a given probability of including the actual value
- Assume the number of measurements in our sample is large ($n \geq 30$), the measurements are independent and come from the same population whose mean is μ and standard deviation is σ:
  - the central limit theorem tells us that our sample mean, $\overline{x}$, is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
  - we assume that μ is the true value that we are trying to measure, that $\overline{x}$ is our best estimate of the true value and that our measurements, x_i, are normally distributed about the mean, $\overline{x}$

Confidence interval for the mean

- To quantify the precision of our measurements, we want to find two values, c_1 and c_2, such that the probability of the mean value being between c_1 and c_2 is $1 - \alpha$.
- This probability is simply the area under the curve, $p(x)$, between c_1 and c_2, where $p(x)$ is the probability density function for the normal distribution with mean, $\overline{x}$, and standard deviation, $s/\sqrt{n}$, s being the standard deviation of our sample of size n.

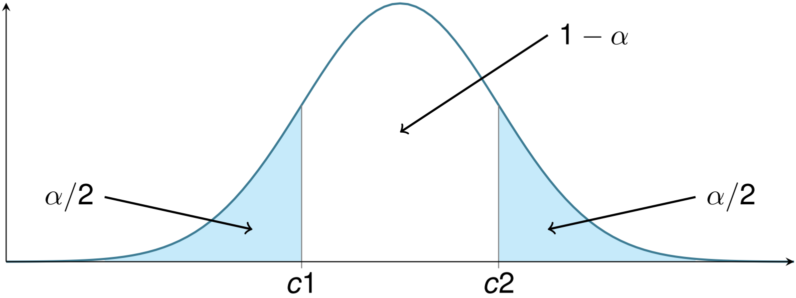
$$Pr[c_1 \leq x \leq c_2] = \int_{c_1}^{c_2} p(x)dx = 1 - \alpha$$

- Typically, we choose c_1 and c_2 to form a symmetric interval such that

$$Pr[x < c_1] = Pr[x > c_2] = \frac{\alpha}{2}$$

Confidence interval for the mean

- The interval $[c_1, c_2]$ is called the *confidence interval* for the mean value, $\overline{x}$
- α is called the *significance level*
- $\kappa = (1 - \alpha) \times 100$ is called the *confidence level*



- Note that our conclusion is always expressed in the form that we're κ% confident that the population mean lies somewhere in the range $[c_1, c_2]$

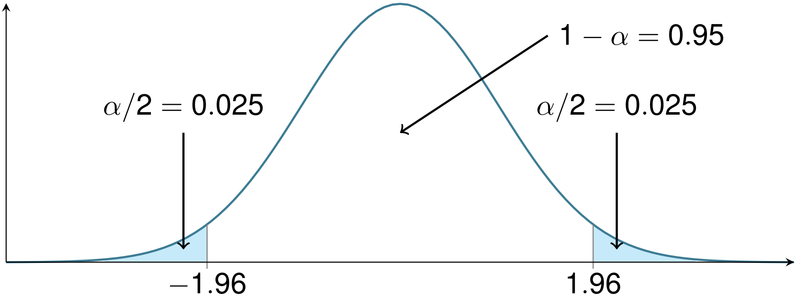
Confidence interval for the mean – example

- Assume
  - sample size: $n = 100$
  - sample mean: $\overline{x} = 0$
  - sample standard deviation: $s = 10$
  - required confidence level: $\kappa = 95$
- Then
  - Significance level, $\alpha = 0.05$ and $\alpha/2 = 0.025$
  - $z_{1-\alpha/2}$ is the value of the standard unit normal distribution that has an area of $1 - \alpha/2$ to the left of it
    - *NORM.INV*$(1 - \alpha/2, 0, 1)$ in Excel
  - and c_1 and c_2 are calculated by

$$c_1 = \overline{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} = -1.96$$

$$c_2 = \overline{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}} = 1.96$$

Confidence interval for the mean – example



- We can conclude that we're 95% confident that the population mean lies in the range $[-1.96, 1.96]$
- Note that, given the same sample,
  - if we want a *higher* level of confidence, the range will grow *wider*
  - if we want a *narrower* range, our level of confidence will be *lower*

Comparing alternatives

- Measurements that we make of a computer system are subject to error, i.e. are ‘noisy’
- If the errors are Gaussian (normally distributed) we can use confidence intervals to quantify the precision of the measurements
- We’d usually like to use our measurements to help us reach a conclusion about some aspect of the performance of one or more computer systems, e.g. to answer a question about whether changing some aspect of the implementation of a system leads to a performance improvement
- Since our measurements are noisy, we need some technique to determine whether any changes that we see are due to random fluctuations in our measurements or whether they are statistically significant
- One approach to this is to use confidence intervals

Using confidence intervals to compare alternatives

- Simplest approach: do confidence intervals for two sets of measurements overlap?
  - if yes then it is impossible to say that differences in the mean value are not due to chance
  - if no, then there is no evidence to suggest that the differences are not statistically significant
- Note the careful phrasing of the conclusion above: there remains the probability, α, that the differences that we observed are due to chance.

Before-and-after comparisons

- Used to determine whether some to change to a system has a statistically significant effect on its performance
- Construct a set of test data $t_1, t_2, \ldots, t_n$
- Obtain measurements before the change $b_1, b_2, \ldots, b_n$ using the test data
- Make the change and obtain measurements for the system after the change, using the same test data, giving $a_1, a_2, \ldots, a_n$
- Find the confidence interval for the mean of the differences of the paired observations, $d_1 = a_1 - b_1, d_2 = a_2 - b_2, \ldots, d_n = a_n - b_n$
- If the confidence interval includes 0, then we can conclude with some level of confidence that there is no significant difference in performance
- If the confidence interval does not include 0, then we conclude that there is no evidence to suggest that the difference in performance is not significant

How many measurements are enough?

The confidence interval formula can help us to determine how many measurements/observations we need to make to achieve a given level of confidence for a sufficiently narrow range about the mean

$$c_1 = \overline{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

$$c_2 = \overline{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

We can see that the size of the interval is inversely dependent on the square root of the number of measurements.

Typically we'd like to minimise the number of measurements required

How many measurements are enough?

Assume that we have

- An estimate for the mean, $\overline{x}$,
- an estimate for the standard deviation, s,
- a required confidence level, κ, and
- a tolerance of error, e.

Now imagine that we want to know how many measurements we need in order that there is a probability of $1 - \alpha$, where $\alpha = 1 - \frac{\kappa}{100}$, that the actual value x is within the interval $(c_1, c_2) = ((1 - e)\overline{x}, (1 + e)\overline{x})$.

How many measurements are enough?

Just consider

$$c_2 = (1 + e)\overline{x} = \overline{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

So we have

$$e\overline{x} = z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

and solving for *n* gives

$$n = \left(\frac{z_{1-\alpha/2}\ s}{e\overline{x}}\right)^2$$

We can use this as a general formula for calculating how many measurements we need to achieve a given confidence that the actual value of *x* is within some given tolerance of our sample mean $\overline{x}$

How many measurement are enough: Example

Suppose we did some experiments to determine the latency of a file transfer across a computer network. Assume that we observed that the mean file transfer time is 7.94 s with a standard deviation of 2.14 s. Approximately how many measurements would be required if we wanted to be 90% confident that the mean value is within 7% of the actual value?

- For a 90% confidence level, $\alpha = 0.10$, so $1 - \alpha/2 = 0.95$.
- To find a value within 7% of the actual value means that we are allowed an error of $\pm 3.5\%$, so $e = 0.035$.

So, we have

$$n = \left(\frac{z_{1-\alpha/2}\, s}{e\bar{x}} \right)^2 = \left(\frac{1.895(2.14)}{0.035(7.94)} \right)^2 = 212.95$$

Therefore, we need to make approximately 213 measurements to achieve the desired confidence that our sample mean is within acceptable limits of the actual mean.

Examples

- See this spreadsheet for worked examples involving confidence intervals.

Sources of information

- Lilja, D., *Measuring Computer Performance: A Practitioner’s Guide*, Cambridge University Press, 2000
- Klein, G. and Dabney, A., *The Cartoon Introduction to Statistics*, Hill and Wang, 2013
  Don’t be put off by the title. This is a gentle and lucid introduction to the main ideas of statistical analysis.
- Spiegel, M., Schiller, J. and Srinivasan, R., *Probability and Statistics*, 4th ed., Schaum’s Outline Series, McGraw Hill, 2013