Computer Systems Performance Analysis and Benchmarking (37-235)

Analytic Modeling
Simulation
Measurements / Benchmarking

Lecture/Assignments/Projects:
Dipl. Inf. Ing. Christian Kurmann

Textbook:
Raj Jain, “The Art of Computer Systems Performance Analysis”, 1991 Wiley & Sons, New York

Topic of Today:
- The ratio game
- Mean / Median / Mode
- Clustering

The Ratio Game

If you can’t convince them, confuse them.

Truman’s Law

Based on Mathematical Fact that...

	System A	System B
Program 0	A0	B0
Program 1	A1	B1
Average	$\frac{A0 + A1}{2}$	$\frac{B0 + B1}{2}$
Relative Average	1	$\frac{B0 + B1}{A0 + A1}$

... is not equal

	Sys A	Sys B
Program 0	1	B0/A0
Program 1	1	B1/A1
Relative Average	1	$\frac{1}{2}\left(\frac{B0}{A0} + \frac{B1}{A1}\right)$
Relative Average	1	$\frac{1}{2}\left(\frac{A1B0 + A0B1}{A0A1}\right)$

Graphical Version with%

TABLE 11.8 Two Tests on Two Systems

System A				System B			
Test	Total	Pass	% Pass	Test	Total	Pass	% Pass
1	300	60	20	1	32	8	25
2	50	2	4	2	500	40	8
	350	62	20.6		532	48	9

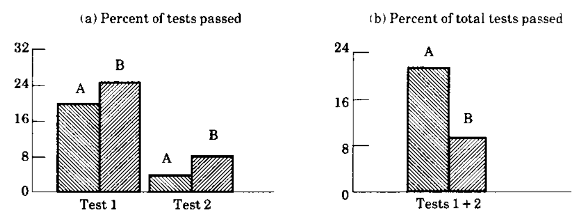


FIGURE 11.1 Ratio games with percentages.

Consequences/Strategies

- If one system is better on all benchmarks... no contradictions.
- Even if one system is better on all benchmarks... ratio game leads to better relative number
- If one system is better in some cases and worse in other cases contradictory conclusions can be drawn sometimes.
- If the metric is LB (lower better) then use your favorite system as a base.
- If the metric is HB (higher better) then use your opponents system as a base.
- Benchmarks that perform better should be elongated, those that perform worse should be shortened.

- Remember: Taking an average of a ratio is not a correct way to analyse data.

A mathematical Analysis.

TABLE 11.10 Derivation of the Rules

Raw Data			With A as a Base			With B as a Base		
	System			System			System	
Benchmark	A	B	Benchmark	A	B	Benchmark	A	B
I	a	ax	I	1	x	I	$\frac{1}{x}$	1
J	b	by	J	1	y	J	$\frac{1}{y}$	1
Average	$\frac{(a+b)}{2}$	$\frac{(ax+by)}{2}$		1	$\frac{(x+y)}{2}$		$\frac{1}{2}\left(\frac{1}{x}+\frac{1}{y}\right)$	1

Using raw data, System A is better if and only if:

$$\frac{a+b}{2} > \frac{ax+by}{2} \quad \text{or} \quad y < -\frac{a}{b}x + \frac{a+b}{b}$$

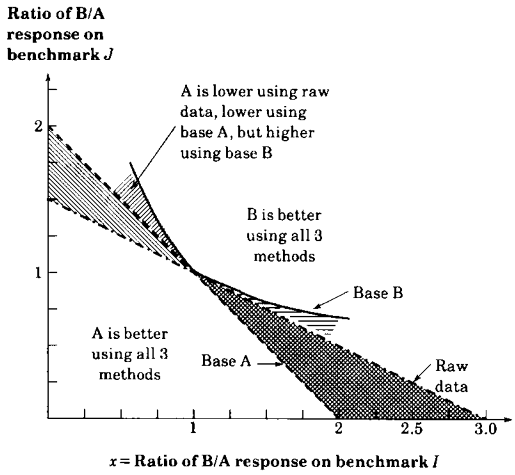
Using System A as a base, System A is better if and only if:

$$\frac{x+y}{2} < 1 \quad \text{or} \quad y < 2-x$$

Using raw data, System A is better if and only if:

$$\frac{1}{2}\left(\frac{1}{x}+\frac{1}{y}\right) > 1 \quad \text{or} \quad y < \frac{x}{2x-1}$$

A mathematical Analysis



Strategies for ratio games. Contradictory conclusions can be obtained using different methods only in shaded regions.

What is wrong in those games?

- Can’t take the mean value of ratios!

How can we fix a good analysis

Do your homework in statistics...

Rest of this lecture:

- Index - where to look it up
- Walk trough English terminology
- Recipes
- No proofs - no derivations

Summarizing Measured Data

- Independent Events
- Random Variable
- Cumulative Distribution Function (cdf)

$$F_x(a) = P(x \le a)$$

- Probability Density Function (pdf)

$$f(x) = \frac{dF(x)}{dx}$$

$$P(x_1 < x \le x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$$

- Probability Mass Function (pmf)

$$f(x_i) = p_i$$

$$P(x_1 < x \le x_2) = F(x_2) - F(x_1) = \sum_{x_1 < x_i < x_{2(1)}} p_i$$

- Mean or Expected Value

$$\text{Mean} = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{\infty} x f(x)\,dx$$

- Variance

$$\text{Var}(x) = E[(x-\mu)^2] = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

$$\int_{-\infty}^{\infty} (x_i - \mu)^2 f(x)\,dx$$

- Standard Deviation

$$\sigma = \sqrt{\text{Var}(x)}$$

- Coefficient of Variation

$$\text{C.O.V.} = \frac{\text{std.dev.}}{\text{mean}} = \frac{\sigma}{\mu}$$

- Covariance

$$\text{Cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] = E[xy] - E[x]E[y]$$

- Covariance Symbols

$$\text{Cov}(x,y) = \sigma_{xy}^2$$

- Correlation Coefficient

$$\text{Correlation}(x,y) = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

- Mean and Variance of Sums

  see Formula in Book

- Quantile/Percentile

$$P(x \le x_\alpha) = F(x_\alpha) = \alpha$$

- Median

  The 50 percentile or 0.5 quantile

- Mode

  Most likely value, that is x_i, that has the highest probability p_i or max of pdf(x_i).

- Normal Distribution N(μ,σ)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Standard Normal Distribution N(0,1)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

- α-quantile of Standard Normal Distrib.

$$z_\alpha = z \sim N(0, 1)$$

$$P\left(\frac{x-\mu}{\sigma} \le z_\alpha\right) = \alpha$$

  see tables at the end of the book

Central Limit Theorem

- The sum of a large number of independent observations from any distribution tens to have a normal distribution.
- The sum of a normal variate is a normal variate

Summarizing Data by a Single Number

Averages or (indices of central tendencies):

- Sample mean
- Sample median
- Sample mode

Selecting among them:

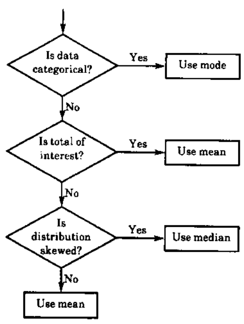


FIGURE 12.2 Selecting among the mean, meadian, and mode.

Examples:

Mode: Most used resource in a system.

Mean: Interarrival time of packets.

Median: Load on a computer.

Median: Average configuration (number of devices, memory size...).

Abuses of arithmetic mean

- Significantly different values
- Skewed distribution
- Multiplying means of dependent variables.

and once more....

- Taking means of ratio with different bases

Other means

Geometric mean

$$\dot{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- works also fine for ratios
- equal to exp[arithmetic mean of log[x_i]]).
- very commonly used in benchmarks.

Harmonic mean

$$\ddot{x} = \frac{n}{1/x_1 + 1/x_2 + \ldots + 1/x_n}$$

- works when 1/x are cumulative.
- work well for rates:
  - MIPS
  - MByte/s
  - MFlops

Summarizing Variability

Variability is specified the following measures or indices of dispersion

- Range - [min...max]
- Variance or standard deviation
- 10 and 90-percentile
- semi-interquantile range
- mean absolute deviation

Discussion of Variability Indices

- range

  extremely unstable, one outlier and you are gone

- sample variance and sample std.dev.

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

  only n-1 independent differences $(x_i - \bar{x})$, degree of freedom n-1.

- Variances and sample std.dev. are absolute measures.
- Variance has square as a unit.
- C.O.V is normalized by the mean and better
- C.O.V of 5 is bad,
  C.O.V.of 0.2 (or 20%) is good.

Specifying quantiles

Definition: How much of the distribution density is within a certain range.

- 5% - 95% is about equivalent to range.
- Decile, quartiles... fixed quantiles increment 0.1 or 0.25.
- Semi-Interquartile Range is difference between Q3 and Q1.
- mean absolute deviation: Sort of a Std. Dev. but with absolute value instead of square.

Selecting index of dispersion:

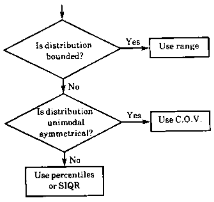


FIGURE 12.4 Selecting the correct index of dispersion.

Determining the Distribution

do a quantile/quantile plot

- determine the quantile of the suspected distribution

$$q_i = F(x_i)$$
$$x_i = F^{-1}(q_i)$$

- e.g. for N(0,1) approximation as follows:

$$x_i = 4.91[q_i^{0.14} - (1 - q_i)^{0.14}]_i$$

- Plot it against the quantiles of your distribution.
- see if you can match a linear function.

Example: Modelling error for 8 predictions of a model were found to be y_i. Are errors normally distributed?

TABLE 12.5 Data for Normal Quantile-Quantile Plot Example

i	$q_i = \frac{i - 0.5}{n}$	y_i	x_i
1	0.0625	−0.19	−1.535
2	0.1875	−0.14	−0.885
3	0.3125	−0.09	−0.487
4	0.4375	−0.04	−0.157
5	0.5625	0.04	0.157
6	0.6875	0.09	0.487
7	0.8125	0.14	0.885
8	0.9375	0.19	1.535

Result: Yes, see graphic.

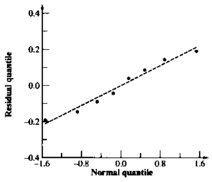


FIGURE 12.5 Normal quantile-quantile plot for the error data.

some matches on Figure 12.6 in the book

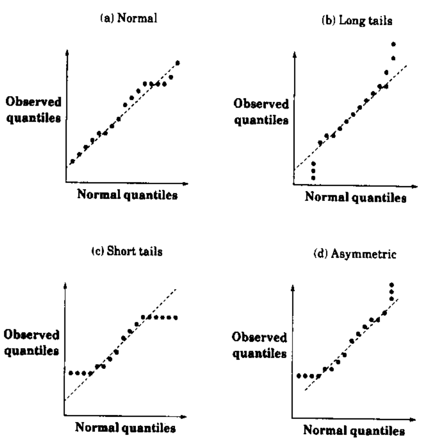


FIGURE 12.6 Interpretation of normal quantile-quantile plots.

Clustering

- Take a sample
  (subset of workload components)
- Select workload parameters
- Transform parameters (e.g. to log)
- Remove outliers
- Scale data (e.g. to $\mu=0, \sigma=1$)
- Select a distance metric (e.g. euclid)

$$d = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Do clustering algorithm
- Interpret results
- Change parameters (e.g. number of clusters) repeat from step transform.
- Select a representative of each cluster

Example: Simple Spanning Tree Method

Consider workload with five components and two parameters. The CPU time and number of I/O's were measured for five programs.

TABLE 6.6 Data for Clustering Example 6.3

Program	CPU Time	Disk I/O
A	2	4
B	3	5
C	1	6
D	4	3
E	5	2

- For more sophisticated approaches... read an AI or IT paper for more (keyword VC dimension).

Step 1: Consider five clusters with ith cluster consisting solely of the ith program.

Step 2: The centroids are {2,4}, {3,5}, {1,6}, {4,3}, and {5,2}. These are shown by the five points in Figure 6.6.

Step 3: Using the Euclidean distance measure, the distance matrix is

	Program				
Program	A	B	C	D	E
A	0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
B		0	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
C			0	$\sqrt{18}$	$\sqrt{32}$
D				0	$\sqrt{2}$
E					0

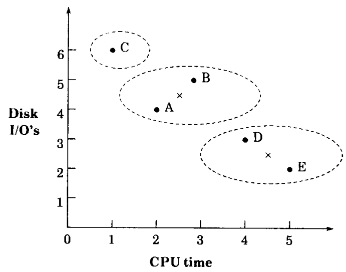


FIGURE 6.6 Clustering example.

Step 4: The minimum intercluster distance is $\sqrt{2}$ between A and B and between D and E. These two pairs are therefore merged.

Step 2: The centroid of cluster pair AB is $\{(2+3)\div 2, (4+5)\div 2\}$, that is, {2.5, 4.5}. Similarly, the centroid of pair DE is {4.5, 2.5}. The other centroids are the same as before.

Step 3: There are three clusters now, as shown in Figure 6.6, and the distance matrix is

	Program		
Program	AB	C	DE
AB	0	$\sqrt{4.5}$	$\sqrt{8}$
C		0	$\sqrt{24.5}$
DE			0

Step 4: The minimum intercluster distance is $\sqrt{4.5}$ between AB and C. These two clusters are therefore merged.

Step 2: The centroid of cluster ABC is $\{(2+3+1)\div 3, (4+5+6)\div 3\}$, that is, {2,5}.

Step 3: The distance matrix is

	Program	
Program	ABC	DE
ABC	0	$\sqrt{12.5}$
DE		0

Step 4: The minimum intercluster distance is $\sqrt{12.5}$. The merger of ABC and DE results in a single cluster ABCDE. □

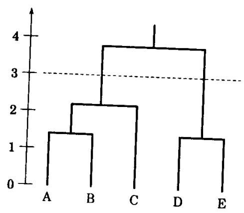


FIGURE 6.7 Dendrogram (spanning tree) for the clustering example.

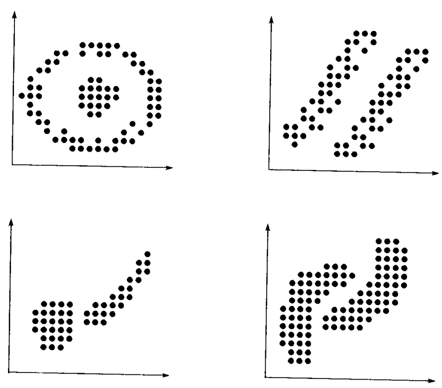


FIGURE 6.8 Problems with clustering.