



EEC 686/785  
Modeling & Performance Evaluation of  
Computer Systems

---

Lecture 7

Wenbing Zhao

Department of Electrical and Computer Engineering  
Cleveland State University

wenbing@ieee.org

(based on Dr. Raj jain's lecture notes)



2

Outline

---

- Review of lecture 6
- Basic Probability and Statistics Concepts



## Review

---

- The art of data presentation
  - Gnuplot
  - Histogram
  - Gantt charts
  - Kiviat graphs
  - Schumacher charts
- Ratio games



## Histogram

---

- A **histogram** is defined as a bar graph that shows frequency data



## Gantt Charts

- **Gantt chart** can be used to show the relative duration of any number of Boolean conditions, i.e., conditions that are either true or false
  - A resource being used or being idle is an example of a Boolean condition
  - Each condition is shown to be a set of horizontal line segments
  - The total length of the line segments represents the relative duration of the condition
  - The position of various segments is arranged such that the overlap between different lines represents the overlap between the conditions



## Kiviat Graphs

- **Kiviat graph**: a circular graph in which several different performance metrics are plotted along radial lines
  - In the most popular version of the graph, an even number of metrics are used.
  - Half of these metrics are HB metrics so that a higher value of the metrics is considered better.
  - The other half of the metrics measure are LB metrics, and a lower value is considered better
  - **Kiviat graph for an ideal system is star**



## Ratio Games

---

- **Ratios** provide good opportunities for playing performance games with competitors
- Ratios have a *numerator* and *denominator (base)*
- **Two ratios with different bases are not comparable**
- However, many examples in published literature where computer scientists have knowingly or unknowingly compared ratios with different bases
- **Ratio game**: the technique of using ratios with incomparable bases and combining them to one's advantage



## Strategies for Winning a Ratio Game

---

- It is better to use your opponent's system as base for HB metric

## Objectives of Part III

- How should you report the performance as a single number? Is specifying the mean the correct way to summarize a sequence of measurements?
- How should you report the variability of measured quantities? What are the alternatives to variance and when are they appropriate?
- How should you interpret the variability? How much confidence can you put on data with a large variability?
- How many measurements are required to get a desired level of statistical confidence?
- How should you summarize the results of several different workloads on a single computer system?
- How should you compare two or more computer systems using several different workloads? Is comparing the mean performance sufficient?
- What model best describes the relationship between two variables? Also, how good is the model?

## Basic Probability and Statistics Concepts

- **Independent events:** two events are called independent if the occurrence of one event does not in any way affect the probability of the other event
- **Random variables:** a variable is called random variable if it takes one of a specified set of values with a specified probability
- **Cumulative Distributed Function (CDF):** the CDF of a random variable maps a given value  $a$  to the probability of the variable taking a value less than or equal to  $a$ :

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

## Basic Probability and Statistics Concepts

### ■ Probability Density Function:

- The derivative

$$f(x) = dF(x)/dx$$

of the CDF  $F(x)$  is called the probability density function (pdf) of  $x$

- Given a pdf  $f(x)$ , the probability of  $x$  being in the interval  $(x_1, x_2)$  can also be computed by integration:

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

## Basic Probability and Statistics Concepts

### ■ Probability Mass Function: for discrete random variable, the probability mass function (pmf) is used in place of pdf

- Consider a discrete random variable  $x$  that can take  $n$  distinct values  $\{x_1, x_2, \dots, x_n\}$  with probability  $\{p_1, p_2, \dots, p_n\}$  such that the probability of the  $i$ th value  $x_i$  is  $p_i$ . The pmf maps  $x_i$  to  $p_i$ :

$$f(x_i) = p_i$$

- The probability of  $x$  being in the interval  $(x_1, x_2)$  can also be computed by summation:

$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1) = \sum_{\substack{i \\ x_1 < x_i \leq x_2}} p_i$$

## Basic Probability and Statistics Concepts

- **Mean or Expected Value**

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

- **Variance:** the quantity  $(x-\mu)^2$  represents the square of distance between  $x$  and its mean. The expected value of this quantity is called the **variance**  $x$ :

$$Var(x) = E[(x-\mu)^2] = \sum_{i=1}^n p_i (x_i - \mu)^2 = \int_{-\infty}^{+\infty} (x_i - \mu)^2 f(x) dx$$

- **Variance** is denoted by  $\sigma^2$ . The square root of the variance is **standard deviation**, denoted as  $\sigma$

## Basic Probability and Statistics Concepts

- **Coefficient of Variation (C.O.V.):** the ratio of the standard deviation to the mean is called C.O.V. =  $\sigma/\mu$
- **Covariance:** given two random variables  $x$  and  $y$  with means  $\mu_x$  and  $\mu_y$ , their covariance is

$$Cov(x,y) = \sigma_{xy}^2 = E[(x-\mu_x)(y-\mu_y)] = E(xy) - E(x)E(y)$$

- For independent variables, covariance is zero since

$$E(xy) = E(x)E(y)$$

The reverse is not true

## Basic Probability and Statistics Concepts

- **Correlation Coefficient:** the normalized value of covariance is called the correlation coefficient or simply the correlation

$$\text{Correlation}(x,y) = \rho_{xy} = \sigma_{xy} / \sigma_x \sigma_y$$

- **Mean and Variance of Sums:** if  $x_1, x_2, \dots, x_k$  are  $k$  random variables and if  $a_1, a_2, \dots, a_k$  are  $k$  arbitrary constants (called weights), then

$$E(a_1x_1 + a_2x_2 + \dots + a_kx_k) = a_1E(x_1) + a_2E(x_2) + \dots + a_kE(x_k)$$

For independent variables,

$$\text{Var}(a_1x_1 + a_2x_2 + \dots + a_kx_k) = a_1^2\text{Var}(x_1) + a_2^2\text{Var}(x_2) + \dots + a_k^2\text{Var}(x_k)$$

## Basic Probability and Statistics Concepts

- **Quantile:** the  $x$  value at which the CDF takes a value  $\alpha$  is called the  $\alpha$ -quantile or  $100\alpha$ -percentile. It is denoted by  $x_\alpha$  and is such that the probability of  $x$  being less than or equal to  $x_\alpha$  is  $\alpha$ :

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

- **Medium:** the 50-percentile (or 0.5-quantile) of a random variable is called its medium
- **Mode:** the most likely value, that is,  $x_i$ , that has the highest probability  $p_i$ , or the  $x$  at which pdf is maximum, is called the mode of  $x$

## Normal Distribution

- **Normal distribution:** Also known as **Gaussian distribution**. It is the most commonly used distribution in data analysis.
  - The sum of a large number of independent observations from any distribution has a normal distribution.
  - Its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty \leq x \leq +\infty$$

## Normal Distribution

- Two parameters:  $\mu$  (mean) and  $\sigma$  (standard deviations) of  $x$
- A normal variate is denoted by  $N(\mu, \sigma)$
- A normal distribution with 0 mean and unit variance is called a **unit normal** or **standard normal distribution**, denoted as  $N(0, 1)$
- An  $\alpha$ -quantile of a unit normal variate  $z \sim N(0, 1)$  is denoted as  $z_\alpha$
- If a random variable  $x$  has a  $N(\mu, \sigma)$  distribution, then  $(x-\mu)/\sigma$  has a  $N(0, 1)$  distribution.  
Thus  $P((x-\mu)/\sigma \leq z_\alpha) = \alpha$ , or  $P(x \leq \mu + z_\alpha \sigma) = \alpha$



## Central Limit Theorem

---

- Main reasons for the popularity of normal distribution
  - The sum of  $n$  independent normal variates is a normal variate
  - The sum of a large number of independent observations from any distribution tends to have a normal distribution. This result is called **central limit theorem**
    - As a result of this property, experimental errors, which are contributed by many factors, are modeled with a normal distribution



## Summarizing Data by a Single Number

---

- In the most condensed form, a single number may be presented that gives the key characteristic of the data set
- This single number is usually called an **average** of the data. To be meaningful, this average should be representative of a major part of the data set
- Three popular alternatives to summarize a sample are to specify its **mean, median, or mode**
- These measures are also called **indices of central tendencies** because they specify the center of location of the distribution of the observations in the sample



## Indices of Central Tendencies

---

- **Sample mean:** obtained by taking the sum of all observations and dividing this sum by the number of observations in the sample
- **Sample median:** obtained by sorting the observations in an increasing order and taking the observation that is in the middle of the series. If the number of observations is even, the mean of the middle two values is used as a median
- **Sample mode:** obtained by plotting a histogram and specifying the midpoint of the bucket where the histogram peaks.
  - For categorical variables, mode is given by the category that occurs most frequently
- The word *sample* in the names of these indices signifies the fact that the values obtained are based on just one sample



## Indices of Central Tendencies

---

- **Mean and median** always exist and are unique. **Mode** may not exist
- The three indices are generally different



## Indices of Central Tendencies

---

- The main problem with the mean is that it is affected more by outliers than the median or mode.
  - A single outlier can make a considerable change in the mean. This is particularly true for small samples.
  - Median and mode are resistant to several outlying observations
- The mean gives equal weight to each observation and in this sense makes full use of the sample. Median and mode ignore a lot of the information
- The mean has an additivity or linearity property: the mean of a sum is a sum of means. This does not apply to mode or median



## Selecting among the Mean, Median and Mode<sup>24</sup>

---



## Selecting among the Mean, Median and Mode

25

- *Most used resource in a system*: resources are categorical and hence the mode must be used
- *Interarrival time*: total time is of interest and so the mean is the proper choice
- *Load on a computer*: the median is preferable due to a highly skewed distribution
- *Average configuration*: medians of number devices, memory sizes, and number of processors are generally used to specify the configuration due to the skewness of the distribution

22 September 2005

EEEC686/785

Wenbing Zhao



## Common Misuse of Means

26

- *Using mean of significantly different values*
  - When the mean is the correct index of central tendency for a variable, it does not automatically imply that a mean of any set of that variable will be useful
  - Usefulness depends upon the number of values and the variance, not only on the type of the variable
  - For example, it is not very useful to say that the mean CPU time per query is 505 milliseconds given two measurements of 10 and 1000 milliseconds

22 September 2005

EEEC686/785

Wenbing Zhao



## Common Misuse of Means

---

- *Using mean without regard to the skewness of distribution*
  - Mean is meaningful for system A, but not for system B



## Common Misuse of Means

---

- *Multiplying means to get the mean of a product*
  - This is correct only if the two random variables are independent
  - If  $x$  and  $y$  are correlated

$$E(xy) \neq E(x)E(y)$$

- *Taking a mean of a ratio with different bases*
  - This has been discussed in chapter 11 on ratio games

## Geometric Mean

- The **geometric mean** of  $n$  values  $x_1, x_2, \dots, x_n$  is obtained by multiplying the values together and taking the  $n$ th root of the product:

$$\dot{x} = \left( \prod_{i=1}^n x_i \right)^{1/n}$$

- The mean discussed in earlier sections is what should be termed the **arithmetic mean**
- The **arithmetic mean** is used if the *sum of the observations* is a quantity that is of interest
- The **geometric mean** is used if the *product of the observations* is a quantity of interest
  - Average error rate per hop on a multihop path in a network

## Geometric Mean

- The geometric mean can be considered as a function  $gm(\cdot)$ , which maps a set of responses  $\{x_1, x_2, \dots, x_n\}$  to a single number  $\dot{x}$
- It has the following multiplicativity property:

$$gm\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \dots, \frac{x_n}{y_n}\right) = \frac{gm(x_1, x_2, \dots, x_n)}{gm(y_1, y_2, \dots, y_n)} = \frac{1}{gm(y_1/x_1, y_2/x_2, \dots, y_n/x_n)}$$

## Geometric Mean

- The geometric mean of a ratio is the ratio of the geometric means of the numerator and denominator
  - The choice of the base does not change the conclusion => sometimes the geometric mean is recommended for ratios
  - However, if the geometric mean of the numerator or denominator does not have any physical meaning, the geometric mean of their ratio is meaningless

## Harmonic Mean

- The **harmonic mean** of a sample  $\{x_1, x_2, \dots, x_n\}$  is defined as follows:

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

- A harmonic mean should be used whenever an arithmetic mean can be justified for  $1/x_i$

## Harmonic Mean

- For example, repeated measurements are made for the elapsed time of a benchmark on a processor
  - In the  $i$ th repetition, the benchmark takes  $t_i$  second. Suppose the benchmark has  $m$  million instructions, the MIPS  $x_i$  computed from the  $i$ th repetition is  $x_i = m/t_i$
  - Sum of  $t_i$  has physical meaning  $\Rightarrow x_i$  should be summarized using harmonic mean
  - The average MIPS rate for the processor is

$$\bar{x} = \frac{n}{\frac{1}{m/t_1} + \frac{1}{m/t_2} + \dots + \frac{1}{m/t_n}} = \frac{m}{(1/n)(t_1 + t_2 + \dots + t_n)}$$

## Weighted Harmonic Mean

- If  $x_i$ 's represent the MIPS rate for  $n$  different benchmarks so that the  $i$ th benchmark has  $m_i$  million instructions, then the harmonic mean of  $n$  ratios  $m_i/t_i$  cannot be used
  - Since the sum of  $t_i/m_i$  does not have any physical meaning
  - The quantity  $\sum t_i/\sum m_i$  is a preferred average MIPS rate. This is a weighted harmonic mean of  $n$  observations

$$\bar{x} = \frac{1}{w_1/x_1 + w_2/x_2 + \dots + w_n/x_n}$$

where  $w_i$ 's are weights that add up to 1:  $w_1 + w_2 + \dots + w_n = 1$

## Weighted Harmonic Mean

- The harmonic mean is special case of the weighted harmonic mean with all weights being equal,  $w_i = 1/n$
- In the example of  $n$  observations of the MIPS rate, if the weights are chosen proportional to the size of the benchmark, that is,

$$w_i = \frac{m_i}{m_1 + m_2 + \cdots + m_n}$$

- Then the weighted harmonic mean would be

$$\ddot{x} = \frac{m_1 + m_2 + \cdots + m_n}{t_1 + t_2 + \cdots + t_n}$$

## Mean of a Ratio

- Given a set of  $n$  ratio, a common problem is to summarize them in a single number
- Considering the additivity of the numerator and denominator separately leads to the following rules for summarizing the ratios

## Rules for Summarizing the Ratios

- If we take the sum of numerators and the sum of denominators and both have a physical meaning, the average of the ratio is the ratio of the averages
- For example, if  $x_i = a_i/b_i$ , the average ratio is given by

$$\begin{aligned} \text{Average} \left( \frac{a_1}{b_1}, \frac{a_2}{b_2}, \dots, \frac{a_n}{b_n} \right) &= \frac{a_1 + a_2 + \dots + a_n}{b_1 + b_2 + \dots + b_n} \\ &= \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} = \frac{(1/n) \sum_{i=1}^n a_i}{(1/n) \sum_{i=1}^n b_i} = \frac{\bar{a}}{\bar{b}} \end{aligned}$$

## Example 12.3

- Illustrates a common application of this guideline in computing mean resource utilization
- CPU utilization of a system is tabulated =>
- Average utilization is not 40% as it may appear because the base of the ratios are not comparable

## Example 12.3

- The mean utilization is obtained by calculating the total CPU busy time and total time and taking the ratio of the two

$$\begin{aligned} \text{Mean CPU utilization} &= \frac{\text{sum of CPU busy times}}{\text{sum of measurement durations}} = \\ &= \frac{0.45 + 0.45 + 0.45 + 0.45 + 20}{1 + 1 + 1 + 1 + 100} = 21\% \end{aligned}$$

- This example also disproves the myth that ratios should always be summarized by a geometric mean

## Rules for Summarizing the Ratios

- When the arithmetic mean of the ratios can be used?
  - Denominator is a constant, so that the ratio has been taken with respect to a base that is constant across all observations,
  - The sum of the numerator has a physical meaning
- That is, if  $b_i = b$  for all  $i$ 's, then

$$\text{Average} \left( \frac{a_1}{b}, \frac{a_2}{b}, \dots, \frac{a_n}{b} \right) = \frac{1}{n} \left( \frac{a_1}{b} + \frac{a_2}{b} + \dots + \frac{a_n}{b} \right) = \frac{\sum_{i=1}^n a_i}{nb}$$

## Rules for Summarizing the Ratios

- When a harmonic mean of the ratio can be used
  - The sum of the denominators has a physical meaning
  - The numerators are constant
- That is, if  $a_i = a$  for all  $i$ 's, then

$$\text{Average} \left( \frac{a}{b_1}, \frac{a}{b_2}, \dots, \frac{a}{b_n} \right) = \frac{n}{b_1/a + b_2/a + \dots + b_n/a} = \frac{na}{\sum_{i=1}^n b_i}$$

- The problem of computing the mean MIPS rate for a processor using  $n$  observations of the same benchmark is an example of this case

## Rules for Summarizing the Ratios

- When a geometric mean can be used
  - The numerator and the denominator are expected to follow a multiplicative property such that  $a_i = cb_i$ , where  $c$  is approximately as constant that is being estimated,
  - $c$  can be estimated by the geometric mean of  $a_i/b_i$
- The case study illustrates the application of this guideline



## Case Study 12.1

---

- A number of benchmarks were run through a program optimizer.
  - The static size of the program is measured
- The total sum of code sizes does not make physical sense => can we use a ratio of sums to estimate the average ratio?



## Case Study 12.1

---

- We can't use the ratio of sums to estimate average ratio
  - Workload sizes selected for this study are widely different
    - The sizes should be weighted by their respective frequencies of usage.
    - Unfortunately we do not know the frequencies
  - The sizes before and after the optimization are expected to follow the following multiplicative model:  $a_i = cb_i$ 
    - where  $b_i$  and  $a_i$  are the sizes before and after the program optimization
    - $c$  is the effect of the optimization that is expected to be independent of the code size

## Case Study 12.1

- The best estimate of the effect  $c$  in this case is obtained by taking a log of the model

$$\log a_i = \log c + \log b_i \text{ or } \log c = \log a_i - \log b_i$$

- And estimating  $\log c$  as the arithmetic mean of  $\log a_i - \log b_i$ . This is equivalent to estimating  $c$  as the geometric mean of  $a_i/b_i = 0.82$ 
  - The assumption of the data following the multiplicative model is justified by the fact that the ratios  $a_i/b_i$  are all within the range of 0.61 and 0.99

## Summarizing Variability

- *Then there is the man who drowned crossing a stream with an average depth of six inches.*

– W. I. E. Gates

- Given a data set, summarizing it by a single number is rarely enough
  - It is important to include a statement about its variability in any summary
  - Most people would prefer the system with low variability



## Indices of Dispersion

---

- **Indices of dispersion:** variability is specified using one of the following measures
  - Range – minimum and maximum of the values observed
  - Variance or standard deviation
  - 10- and 90- percentiles
  - Semi-interquantile range
  - Mean absolute deviation



## Indices of Dispersion

---

- All the indices of dispersion apply only for quantitative data
- For qualitative data, the dispersion can be specified by giving the number of most frequent categories that comprise the given percentile, e.g., the top 90%

## Range

- **Range** – minimum and maximum of the values observed
- The range is useful if and only if the variable is bounded. The range gives the best estimate of these bounds
- Otherwise, the max goes on increasing with the number of observations, and the min goes on decreasing with the number of observations
  - There is no “stable” point that gives a good indication of the actual range

## Variance

- The variance of a sample of  $n$  observations  $\{x_1, x_2, \dots, x_n\}$  is calculated as follows:

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

- **Sample variance:**  $s^2$
- **Sample standard deviation:**  $s$
- The word “sample” can be dropped if there is no ambiguity
- **Degree of freedom:** the number of independent terms in a sum. In calculating the variance, the sum of squares is divided by  $n-1$

## Variance and Standard Deviation

- **Problem of variance:** it is expressed in units of square of the units of observations
- It is preferable to use the **standard deviation**
  - It is in the same unit as the mean, which allows us to compare it with the mean
- **Coefficient of variance (C.O.V.):** ratio of standard deviation to the mean
  - It takes the scale out of the variability consideration
  - E.g., a C.O.V of 5 is large, a C.O.V of 0.2 is small no matter what the unit is

## Percentile

- 5-percentile and the 95-percentile of a variable has the same impact of specifying its min and max
  - But it can be applied to variables without bounds
- When expressed as a fraction between 0 and 1, the percentiles are also called **quantiles**, or **fractiles**
  - 0.9 quantile is the same as 90% percentile
- The percentiles at multiples of 10% are called **deciles**
  - First decile is 10%, second decile is 20%, etc.

## Percentile

- **Quartiles** divide the data into four parts at 25, 50, 75%
  - 25% of the observations  $\leq$  the first quartile  $Q_1$
  - Second quartile  $Q_2$  is also the median
- $\alpha$ -quantiles can be estimated by sorting the observations and taking the  $[(n-1)\alpha+1]^{\text{th}}$  element in the ordered set

## SIQR

- The range between  $Q_3$  and  $Q_1$  is called **interquartile range of data**
- One half of this range is called **Semi-Interquartile Range (SIQR)**, that is,
 
$$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{x_{0.75} - x_{0.25}}{2}$$
- SIQR is very resistant to outliers, similar to the median
- SIQR is used as an index of dispersion whenever the median is used as an index of central tendency

## Mean Absolute Deviation

- **Mean absolute deviation:** an alternative measure of dispersion
  - It does not involve multiplication or square root

$$\text{Mean absolute deviation} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

## Example 12.4

- Measured CPU time was found to be {3.1, 4.2, 2.8, 5.1, 2.8, 4.4, 5.6, 3.9, 3.9, 2.7, 4.1, 3.6, 3.1, 4.5, 3.8, 2.9, 3.4, 3.3, 2.8, 4.5, 4.9, 5.3, 1.9, 3.7, 3.2, 4.1, 5.1, 3.2, 3.9, 4.8, 5.9, 4.2}
- The sorted set is {1.9, 2.7, 2.8, 2.8, 2.8, 2.9, 3.1, 3.1, 3.2, 3.2, 3.3, 3.4, 3.5, 3.7, 3.8, 3.9, 3.9, 3.9, 4.1, 4.1, 4.2, 4.2, 4.4, 4.5, 4.5, 4.8, 4.9, 5.1, 5.1, 5.3, 5.6, 5.9}



## Example 12.4

---

- Then:
  - 10-percentile is given by  $[1+(31)(0.10)] = 4^{\text{th}}$  element = 2.8
  - 90-percentile is given by  $[1+(31)(0.90)] = 29^{\text{th}}$  element = 5.1
  - First quartile  $Q_1$  is given by  $[1+(31)(0.25)] = 9^{\text{th}}$  element = 3.2
  - Median  $Q_2$  is given by  $[1+(31)(0.50)] = 16^{\text{th}}$  element = 3.9
  - The third quartile  $Q_3$  is given by  $[1+(31)(0.75)] = 24^{\text{th}}$  element = 4.5
- Thus:  $\text{SIQR} = (Q_3 - Q_1)/2 = (4.5 - 3.2)/2 = 0.64$



## Selecting the Index of Dispersion

---