COMP 190-088: Systems Performance Analysis

# Analyzing Measured Data-I

**Jasleen Kaur**

Department of Computer Science
The University of North Carolina at Chapel Hill

Spring 2005



## Review

- Independent Events, Random Variables
- CDF, PDF, PMF, Quantile
- Mean, Variance, Coefficient of Variation
- Covariance, Correlation Coefficient
- Median, Mode
- Mean and Variance of Sums

# Normal Distribution

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

- Also known as a Gaussian distribution
- Unit normal distribution: N(0,1)
  - $z_\alpha$: $\alpha$-quantile of N(0,1)
    - $P(x \le \mu + z_\alpha\sigma) = \alpha$
    - Well-known and tabulated values
- Most commonly used distribution in data analysis
  - Sum of n independent normal variates is a normal variate
    - Normal processes remain normal after passing through linear systems
  - Central limit theorem:
    - Sum of a large number of independent observations from any distribution tends to have a normal distribution
    - Used to model experimental errors, which are contributed by many factors



# Adding Two Normal Distributions

$$\boxed{N(\mu,\sigma) \sim f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}}$$

- Let: $X \sim N(\mu_x, \sigma_x)$; $Y \sim N(\mu_y, \sigma_y)$;
- Let: $Z = X + Y$

$$P(X \le x_0) = \int_{-\infty}^{x_0} f_X(x)dx;$$

$$P(Y \le y_0) = \int_{-\infty}^{y_0} f_Y(y)dy;$$

$$\begin{aligned}
P(Z \le z_0) &= P(X + Y \le z_0) \\
&= \int_{-\infty}^{\infty} P(X \le z_0 - y) f_Y(y)dy \\
&= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{z_0 - y} f_X(x)dx \right) f_Y(y)dy \\
&= \frac{1}{(\sigma_X\sqrt{2\pi})(\sigma_Y\sqrt{2\pi})} \int_{-\infty}^{\infty}\int_{-\infty}^{z_0 - y} e^{-(x-\mu_X)^2/2\sigma_X^2} . e^{-(y-\mu_Y)^2/2\sigma_Y^2} dx.dy
\end{aligned}$$

## Analysis of Measured Data

- **Summarizing the Average**
- Summarizing the Variability
- Comparing Systems Using Sample Data
- Simple Linear Regression Models



## Indices of Central Tendencies

- Average used to specify the center of location of the distribution of sample observations
  - Sample Mean
  - Sample Median
  - Sample Mode
- Mode may not exist or be unique
  - Mean and median always exist and are unique

## Selecting Between Mean, Median, and Mode

- Comparison
  - Mean is most affected by outliers
    - Especially for small samples
  - Mean gives equal weight to each observation
    - Median and mode ignore a lot of information
  - Mean has additivity property
    - Mean of a sum is the sum of the means
- Selection:
  - If data is categorical, use mode only
    - eg, most used resource in a system
  - If distribution is skewed, use median
    - eg, load on a computer, average configuration
  - If total is of interest, use mean
    - eg, inter-arrival time



## Mean of a Product

| Day | # of Users | # of Processes per User | # of Processes |
|---|---|---|---|
| 1 | 1 | 10 | 10 |
| 2 | 3 | 5 | 15 |

- Random Variables:
  - X: number of users
  - Y: number of processes per user
  - Z: number of processes
    - Z=X*Y
- Means:
  - Mean(X) = (1+3)/2 = 1.5
  - Mean(Y) = (10+5)/2 = 7.5
  - Mean(Z) = (10+15)/2 = 12.5
    - Mean(X)*Mean(Y) = 11.25

## Geometric Mean

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

| Protocol Layer | Performance Improvement |
|---|---|
| 7 | 18 % |
| 6 | 13 % |
| 5 | 11 % |
| 4 | 8 % |
| 3 | 10 % |
| 2 | 28 % |
| 1 | 5 % |

- Used if product of the observations is of interest
- Other examples:
  - Average cache hit/miss ratios over several layers of cache
  - Average per-hop error rate on a multi-hop path in a network
- $gm(x_1/y_1, \ldots, x_n/y_n) = gm(x_1, \ldots, x_n) / gm(y_1, \ldots, y_n)$
  - Recommended for ratios



## Harmonic Mean

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

- Used when an arithmetic mean for $1/x_i$ can be justified
- Example
  - $t_i$ : measurement of elapsed time of a benchmark on a processor
  - MIPS = $x_i = m/t_i$, if benchmark has m instructions
  - Average MIPS = $hm(x_i)$

## Mean of a Ratio

- If sum of numerators and sum of denominators has a physical meaning, the average of the ratio is the ratio of the averages

$$Average\left(\frac{a_1}{b_1},\frac{a_2}{b_2},\cdots,\frac{a_n}{b_n}\right)=\frac{\overline{a}}{\overline{b}}$$

  - If ratios have been taken with respect to a common base, arithmetic mean of ratio can be used

$$Average\left(\frac{a_1}{b},\frac{a_2}{b},\cdots,\frac{a_n}{b}\right)=\frac{1}{n}\left(\frac{a_1}{b}+\frac{a_2}{b}+\cdots+\frac{a_n}{b}\right)$$

  - If numerators are constant, harmonic mean of ratio should be used

$$Average\left(\frac{a}{b_1},\frac{a}{b_2},\cdots,\frac{a}{b_n}\right)=\frac{n}{b_1/a+b_2/a+...+b_n/a}$$



## Mean of a Ratio

- If numerator and denominator are expected to follow a multiplicative property, then geometric mean of ratio can be used
- Example: performance of benchmarks run with a program optimizer
  - Workload sizes are widely different
    - Should be weighted by frequency of usage
  - Sizes before and after optimization expected to follow a multiplicative model
    - Multiplicative constant can be estimated by the gm of ratios
    - 0.82

| Program | Code Size | | Ratio |
|---|---|---|---|
| | Before | After | |
| BubbleP | 119 | 89 | 0.75 |
| IntmmP | 158 | 134 | 0.85 |
| PermP | 142 | 121 | 0.85 |
| PuzzleP | 8612 | 7579 | 0.88 |
| QueenP | 7133 | 7062 | 0.99 |
| QuickP | 184 | 112 | 0.61 |
| SieveP | 2908 | 2879 | 0.99 |
| TowersP | 433 | 307 | 0.71 |