# 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:**

- **Confidence Intervals**
- **Linear Regression**



# Confidence Intervals

## Sample versus population

- A sample is a random subset taken as an estimate. It can be varied depending on size.
- A sample has statistics like sample mean or sample standard deviation.

- Population is the distribution as it really is. It is fixed but might be unknown.
- A population has parameters like population mean or population standard deviation.

- Population characteristics are called parameters, sample characteristics are called statistics
- Statistics are estimates of parameters.



# Confidence Interval for a Mean

## Definitions:

$$P\{c_1 \le \mu \le c_2\} = 1 - \alpha$$

- $\alpha$: significance level (0.05 or 0.10)
- 100(1-$\alpha$): confidence level (95 or 90)[%]
- Interval (c1,c2): confidence interval

## Central Limit Theorem:

If observations in a sample are independent and come from a population with mean $\mu$ and std.dev. $\sigma$, then the sample mean for large samples is normally distributed with:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Std.dev. of sample mean is called standard error



## Calculation of the confidence Interval:

100(1-$\alpha$)% confidence interval for population mean is therefore:

$$\left(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

x = sample mean

s = sample std.dev.

n = number of large samples (n>30!)

$z_{1-\alpha/2}$ = (1-$\alpha$/2)-quantile of unit normal variate (read from Table A.2 in Book)

## Confidence Interval for small samples:

only when normally distributed population:

$$\left(\bar{x} - t_{[1-\alpha/2, n-1]}\frac{s}{\sqrt{n}}, \bar{x} + t_{[1-\alpha/2, n-1]}\frac{s}{\sqrt{n}}\right)$$

x = sample mean

s = sample std.dev.

n = number of samples (n<30)

t's = (1-$\alpha$/2)-quantile of t variate with n-1 degrees of freedom (read from Table A.4)

## Testing for zero mean

- Calculate confidence interval
- Is zero part of it?
- Example: Measured differences in the processor times of two different implementations {1.5,2.6,-1.8,1.3,-0.5,1.7,2.4}. Is one implementation better with 99% confidence?

**Example 13.3** The difference in the processor times of two different implementations of the same algorithm was measured on seven similar workloads. The differences are $\{1.5, 2.6, -1.8, 1.3, -0.5, 1.7, 2.4\}$. Can we say with 99% confidence that one implementation is superior to the other?

$$\text{Sample size} = n = 7$$

$$\text{Mean} = 7.20/7 = 1.03$$

$$\text{Sample variance} = (22.84 - 7.20 * 7.20/7)/6 = 2.57$$

$$\text{Sample standard deviation} = \sqrt{2.57} = 1.60$$

$$\text{Confidence interval} = 1.03 \mp t \times 1.60/\sqrt{7} = 1.03 \mp 0.605t$$

$$100(1-\alpha) = 99, \qquad \alpha = 0.01, \qquad 1-\alpha/2 = 0.995$$

From Table A.4 in the Appendix, the $t$-value at six degrees of freedom is $t_{[0.995;6]} = 3.707$, and

$$99\% \text{ confidence interval} = (-1.21, 3.27)$$

The confidence interval includes zero. Therefore, we cannot say with 99% confidence that the mean difference is significantly different from zero. □



## Comparing two Alternatives

Paired observations: n experiments on each of two systems A,B, *i*th test on A with same workload as *i*th test on B.

- just compute difference
- subject to zero mean test

Unpaired observations:

- more complicated
- procedure called the t-test, see next page
- approximate visual test
  Visualize the Confidence Intervals!

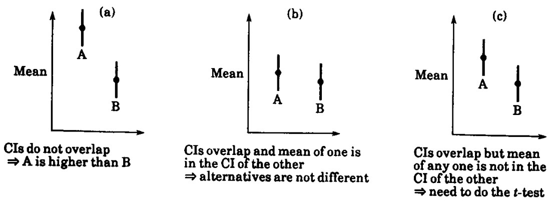


**FIGURE 13.4** Comparing two alternatives.



## Unpaired Observation: t-test

1. Compute the sample means:

$$\bar{x}_a = \frac{1}{n_a}\sum_{i=1}^{n_a} x_{ia}$$

$$\bar{x}_b = \frac{1}{n_b}\sum_{i=1}^{n_b} x_{ib}$$

2. Compute the sample standard deviations:

$$s_a = \left\{\frac{\left(\sum_{i=1}^{n_a} x_{ia}^2\right) - n_a\bar{x}_a^2}{n_a - 1}\right\}^{1/2}$$

$$s_b = \left\{\frac{\left(\sum_{i=1}^{n_b} x_{ib}^2\right) - n_b\bar{x}_b^2}{n_b - 1}\right\}^{1/2}$$

3. Compute the mean difference: $\bar{x}_a - \bar{x}_b$.
4. Compute the standard deviation of the mean difference:

$$s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

5. Compute the effective number of degrees of freedom:

$$\nu = \frac{\left(s_a^2/n_a + s_b^2/n_b\right)^2}{\frac{1}{n_a+1}\left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b+1}\left(\frac{s_b^2}{n_b}\right)^2} - 2$$

6. Compute the confidence interval for the mean difference:

$$(\bar{x}_a - \bar{x}_b) \mp t_{[1-\alpha/2;\nu]}s$$

Here, $t_{[1-\alpha/2;\nu]}$ is the $(1-\alpha/2)$-quantile of a $t$-variate with $\nu$ degrees of freedom.

7. If the confidence interval includes zero, the difference is not significant at $100(1-\alpha)\%$ confidence level. If the confidence interval does not include zero, then the sign of the mean difference indicates which system is better.



## Determining Sample Size:

The larger the sample, the higer is the associated confidence.

However, larger samples means more effort and resources.

Goal: Find minimal sample size that will provide confidence.

### Sample size for determining mean:

- Formulas can be inverted to tell us the minimum sample size.
- Example: Estimate the mean performance of a system with accuracy of r% and confidence level of 100(1-α)%. How many experiments are needed?:

r% implies that the confidence interval should be: $(\bar{x}(1-r/100), \bar{x}(1-r/100))$. We know that $\bar{x} \mp z(s/\sqrt{n})$. Equate these:

$$n = \left(\frac{100 z_{1-\alpha/2} s}{r\bar{x}}\right)^2$$

# Simple Linear Regression Model

## Good - Bad Model

**Hint: Visualize and check against errors on the plotted graph.**

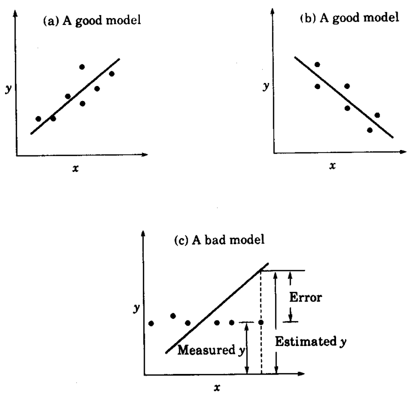


FIGURE 14.1 Good and bad regression models.

- measured in terms of residual or modeling error, or simple error



## Criterion of the least squares

**Model and Error:**

$$\hat{y} = b_0 + b_1 x$$
$$\hat{y}_i = b_0 + b_1 x_i$$
$$e_i = \hat{y}_i - y_i$$

**Sum of Squared Errors:**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0$$

**It can be shown that this constrained minimization problem is equivalent to minimizing the variance of errors.**



# Computing the Parameters

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$$
$$b_0 = \bar{y} - b_1\bar{x}$$

**Derivation:**

Simple substitute and differentiate-> book.

**Allocation of variation:**

SSY: Sum of Squares of y

SS0: Sum of Squares of $\bar{y}$, equal to $n\bar{y}^2$

SST: Sum of Squares Total

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSY - SS0$$

SST is called **variation** of y

**Quality of linear regression**

SSE: Sum of Squared Errors (see. above)

SSR: The Sum of Squares explained by regression = SST-SSE



SST is therefore = SSR+SSE and can be devided into two parts:

SSE: indicates the variation that was not explained by regression.

SSR: indicates the variation explained by regression.

**Coefficient of Determination:**

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

- determines the goodness of the regression
- 1 for perfect regression - 0 for a bad one