

**King Fahd University of  
Petroleum & Minerals  
Computer Engineering Dept**

---

**COE 587 – Performance Evaluation and  
Analysis**

**Term 091**

**Dr. Ashraf S. Hasan Mahmoud**

**Rm 22-148-3**

**Ext. 1724**

**Email: [ashraf@kfupm.edu.sa](mailto:ashraf@kfupm.edu.sa)**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

1

---

**Slides are based on the  
textbook:**

**R. Jain, "Art of Computer  
Systems Performance  
Analysis," Wiley, 1991,  
ISBN:0471503363**

**Book website:**

**<http://www.cse.wustl.edu/~jain/books/perfbook.htm>**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

2

## The Cumulative Distribution Function

---

- The cumulative distribution function (cdf) of a random variable  $X$  is defined as the probability of the event  $\{X \leq x\}$ :

$$F_X(x) = \text{Prob}\{X \leq x\} \quad \text{for } -\infty < x < \infty$$

i.e. it is equal to the probability the variable  $X$  takes on a value in the set  $(-\infty, x]$

- A convenient way to specify the probability of all semi-infinite intervals

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

3

## Properties of the CDF

---

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $F_X(x)$  is a nondecreasing function  $\rightarrow$  if  $a < b \rightarrow F_X(a) \leq F_X(b)$
- $F_X(x)$  is continuous from the right  $\rightarrow$  for  $h > 0$ ,  
$$F_X(b) = \lim_{h \rightarrow 0} F_X(b+h) = F_X(b^+)$$
- $P[a < X \leq b] = F_X(b) - F_X(a)$
- $P[X = b] = F_X(b) - F_X(b^-)$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

4

## Example 1: Exponential Random Variable

---

- **Problem:** The transmission time  $X$  of a message in a communication system obey the exponential probability law with parameter  $\lambda$ , that is

$$\text{Prob } [X > x] = e^{-\lambda x} \quad x > 0$$

Find the CDF of  $X$ . Find  $\text{Prob } [T < X \leq 2T]$  where  $T = 1/\lambda$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

5

## Example 1: Exponential Random Variable – cont'd

---

- **Answer:**

The CDF of  $X$  is

$$\begin{aligned} F_X(x) &= \text{Prob } \{X \leq x\} = 1 - \text{Prob } \{X > x\} \\ &= 1 - e^{-\lambda x} \quad x \geq 0 \\ &= 0 \quad x < 0 \end{aligned}$$

$$\begin{aligned} \text{Prob } \{T < X \leq 2T\} &= F_X(2T) - F_X(T) \\ &= 1 - e^{-2} - (1 - e^{-1}) \\ &= 0.233 \end{aligned}$$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

6

## Example 2: Use of Bayes Rule

- **Problem:** The waiting time  $W$  of a customer in a queueing system is zero if he finds the system idle, and an exponentially distributed random length of time if he finds the system busy. The probabilities that he finds the system idle or busy are  $p$  and  $1-p$ , respectively. Find the CDF of  $W$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

7

## Example 2: cont'd

- **Answer:**

The CDF of  $W$  is found as follows:

$$\begin{aligned}F_X(x) &= \text{Prob}\{W \leq x\} \\ &= \text{Prob}\{W \leq x/\text{idle}\}p + \text{Prob}\{W \leq x/\text{busy}\}(1-p)\end{aligned}$$

**Note**  $\text{Prob}\{W \leq x/\text{idle}\} = 1$  for any  $x > 0$

→

$$\begin{aligned}F_X(x) &= 0 & x < 0 \\ &= p + (1-p)(1 - e^{-\lambda x}) & x \geq 0\end{aligned}$$

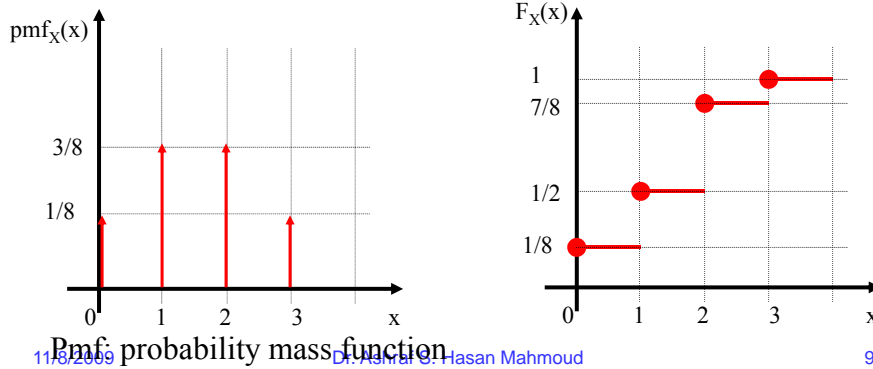
11/8/2009

Dr. Ashraf S. Hasan Mahmoud

8

## Types of Random Variables

- **(1) Discrete Random Variables**
  - CDF is right continuous, staircase function of  $x$ , with jumps at countable set  $x_0, x_1, x_2, \dots$



## Types of Random Variables

- **(2) Continuous Random Variables**
  - CDF is continuous for all values of  $x \rightarrow \text{Prob} \{ X = x \} = 0$  (recall the CDF properties)
  - Can be written as the integral of some non negative function

$$F_X(x) = \int_{-\infty}^{\infty} f(t) dt$$

Or

$$f(t) = \frac{dF_X(x)}{dx}$$

11/8 f(t) is referred to as the probability density function or PDF 10

## **Types of Random Variables**

---

- **(3) Random Variables of Mixed Types**

$$F_X(x) = p F_1(x) + (1-p) F_2(x)$$

## **Probability Density Function**

---

- **The PDF of X, if it exists, is define as the derivative of CDF  $F_X(x)$ :**

$$f_x(x) = \frac{dF_x(x)}{dx}$$

## Properties of the PDF

- $f_x(x) \geq 0$

- $P\{a \leq x \leq b\} = \int_a^b f_x(x) dx$

- $F_X(x) = \int_{-\infty}^x f_x(t) dt$

- $1 = \int_{-\infty}^{\infty} f_x(t) dt$

A valid pdf can be formed from any nonnegative, piecewise continuous function  $g(x)$  that has a finite integral:

$$\int_{-\infty}^{\infty} g(x) dx = c < \infty$$

By letting  $f_x(x) = g(x)/c$ , we obtain a function that satisfies the normalization condition.

This is the scheme we use to generate pdfs from simulation results!

## Conditional PDFs and CDFs

- If some event **A** concerning **X** is given, then conditional CDF of **X** given **A** is defined by

$$P([X \leq x] \cap A)$$

$$F_X(x/A) = \frac{\text{-----}}{P(A)} \quad \text{if } P(A) > 0$$

The conditional pdf of **X** given **A** is then defined by

$$f_X(x/A) = \frac{d}{dx} F_X(x/A)$$

## Mean or Expected Value

- Expectation of the random variable  $X$  can be computed by

$$\mu = E[X] = \sum_{\forall i} x_i P[X = x_i]$$

for discrete variables, or

$$\mu = E[X] = \int_{-\infty}^{\infty} x f_x(x) dx$$

for continuous variables.

## Expectation of a Function of the Random Variable

- Let  $g(x)$  be a function of the random variable  $x$ , the expectation of  $g(x)$  is given by

$$E[g(x)] = \sum_{\forall i} g(x_i) P[X = x_i]$$

for discrete variables, or

$$E[g(x)] = \int_{-\infty}^{\infty} g(t) f_x(t) dt$$

for continuous variables.



### Example 3:

- **Problem: For X nonnegative r.v. show that**

**for continuous X:**  $E[X] = \int_0^{\infty} (1 - F_x(t)) dt$  , and

**for discrete X:**  $E[X] = \sum_{k=0}^{\infty} P(X > k)$

**Prove the above formulas**

### Variance ( $\sigma^2$ ) – Standard Deviation ( $\sigma$ )

- **For continuous X:**

$$Var(x) = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

- **For discrete X:**

$$Var(x) = E[(x - \mu)^2] = \sum_{\forall i} (x_i - \mu)^2 \Pr[X = x_i]$$

- **Standard deviation ( $\sigma$ ) =  $\sqrt{Var(x)}$  =  $\sqrt{\sigma^2}$**
- **Variance or standard deviation is a measure of variability**

## Coefficient of Variation (COV)

- **COV = ratio of standard deviation to the mean**

$$COV = \frac{\sigma}{\mu}$$

- **COV is a measure of variability**
  - **What does it mean if COV = 0, < 1, or > 1?**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

19

## Covariance

- **Consider two random variables x and y, such that**

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

- **For independent x and y (i.e.  $E[xy] = E[x]E[y]$ ) →**

$$\sigma_{xy}^2 = 0$$

- **Two variables are independent →  $\sigma_{xy}^2 = 0$ , but the reverse it not always TRUE!!**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

20

## Correlation Coefficient

- **Correlation Coefficient: normalized value of the covariance**

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

- **The normalization is with respect to what?**
- **What is the range for  $\rho_{xy}$  ?**

## Mean/Variance of Sums

- **If  $x_1, x_2, \dots, x_k$  are  $k$  r.v. and  $a_1, a_2, \dots, a_k$  are  $k$  arbitrary constants, 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:**

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

## Quantile

---

- We know that  $F_X(x) \in [0, 1] \forall x$
- The value of  $x$  such that  $F_X(x) = \alpha$  is called the  $\alpha$ -quantile or  $100\alpha$ -percentile

$$\Pr[X \leq x_\alpha] = F_X(x_\alpha) = \alpha$$

- Quantile – percentile – fractile – quartile?

## Median

---

- The 50-percentile (or 0.5 quantile) for the r.v.
- i.e.  $x_{0.5}$  such that

$$\Pr[X \leq x_{0.5}] = F_X(x_{0.5}) = 0.5$$

## Mode

---

- **Mode: is the most likely value**
  - **x at which pmf or pdf is maximum**
- **i.e.  $x_m$  such that (for continuous r.v.)**

$$f_X(x_m) \geq f_X(x) \quad \forall x$$

- **Or (for discrete r.v.)**

$$p_{x_m} \geq p_{x_i} \quad \forall i$$

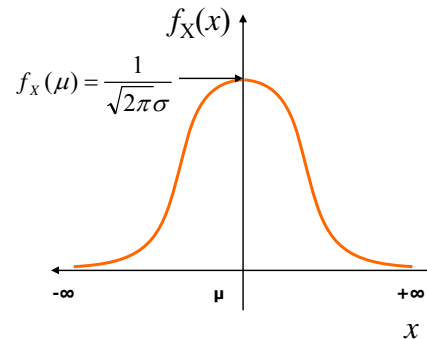
## Mean – Median - Mode

---

- **Figure 12.1**

## Normal Distribution - General

- More details on later slides
- One of the most commonly used distributions
- $X \sim N(\mu, \sigma)$  means
  - $X$  is a random variables taking values ranging from  $-\infty$  to  $+\infty$
  - $X$  has a mean of  $\mu$  and standard deviation of  $\sigma \rightarrow$  i.e.  $E[X] = \mu$ , and  $\text{Var}[X] = \sigma^2$ .
- The corresponding probability density function is given by



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

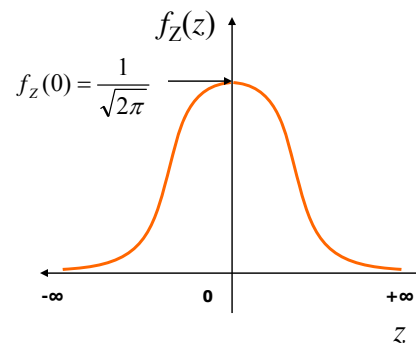
11/8/2009

Dr. Ashraf S. Hasan Mahmoud

27

## Normal Distribution – Zero Mean and unity Variance

- Referred to as Unit Normal or Standard Normal Distribution
- $\mu = 0$ , and  $\sigma = 1$ 
  - $\rightarrow Z \sim N(0, 1)$
- The corresponding probability density function is given by



$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty \leq z \leq +\infty$$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

28

## Normal Distribution – Zero Mean and unity Variance – cont'd

- If  $X \sim N(\mu, \sigma)$ , then  $(X - \mu)/\sigma$  is a standard normal distribution, i.e

$$\Pr[(x - \mu)/\sigma \leq z_\alpha] = \alpha$$

Or

$$\Pr[x \leq \mu + \sigma z_\alpha] = \alpha$$

- Prob  $[0 \leq Z \leq z]$  is listed in table A.1 (or evaluated to using Q-function or erfc function)

Show the PQRS tool

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

29

## Why is the Normal Distribution Important

- **Two reasons:**
  - The sum of  $n$  independent normal variates is a normal variate,
    - i.e, if  $x_1, x_2, \dots, x_n$  are  $n$  independent r.v. ( $x_i \sim N(\mu_i, \sigma_i)$ ), then
    - $Y = \sum a_i x_i$  is also a normal variable with  $Y \sim N(\mu, \sigma)$ , where  $\mu = \sum a_i \mu_i$  and  $\sigma^2 = \sum a_i^2 \sigma_i^2$
  - The sum of a large number of *independent* observations from *any distribution* tends to have a normal distribution – central limit theorem

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

30

## **Summarizing Data by a Single Number**

---

- **Referred to by an “average” of the data**
  - Should be representative of the major part of the data set
- **Choices (indices of central tendencies):**
  - Mean
  - Median
  - Mode
- **Which one to choose?**
  - Depends on the problem and the figure of interest

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

31

## **Common Misuses of Means**

---

- **Using mean of significantly different values**
- **Using mean without regard to skewness of Distribution (refer to table 12.1)**
- **Multiplying means to get the mean of a product (Example 12.1)**
- **Taking the mean of a ratio with different bases**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

32



## Example 12.1

- **Problem:** On a time sharing system, the total number of users and the number of subprocesses for each user are monitored. The average number of users is 23 while the average number of subprocesses per user is 2. What is the average number of subprocesses?
- **Solution:** The answer is NOT  $23 \times 2 = 46$ !  
The average number of subprocesses per user is dependent on the load or the number of users in the system  $\rightarrow$  i.e. the two r.v. are correlated and therefore  $E[xy] \neq E[x]E[y]$ !

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

33

## Geometric Mean

- The geometric mean for  $n$  values  $x_1, x_2, \dots, x_n$  is given by

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

- Another notation:

$$\dot{x} = gm(x_1, x_2, \dots, x_n)$$

- The mean dealt with previous is called the arithmetic mean
- When to use?
  - When the product of the observations is meaningful
- The multiplicative property: The geometric mean of a ratio is the ratio of the geometric means of the numerator and denominator

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

34

## Example 12.2

- **Problem:** The performance improvements in the latest version of seven layers of a new networking protocol was measured separately for each layer. The observations are as listed below. What is the average improvement per layer?

- **Solution:** The improvements work in a multiplicative manner

Average improvement per layer

$$= \frac{[(1.18)(1.13)(1.11)(1.08)(1.10)(1.28)(1.05)]^{1/7} - 1}{1}$$

$$= 0.13$$

i.e average improvement per layer = 13%

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

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

35

## Harmonic Mean

- We are interested in finding the average response time for a CPU
- We run n benchmarks of sizes: m1, m2, ..., mn – Let the elapsed time be t1, t2, ..., tn
- The average CPU response time is given by

$$\bar{x} = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n t_i}$$

where the numerator is the total size of all benchmarks and the denominator represents the total time

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

36

## Harmonic Mean – cont'd

- The previous expression can be written as

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

where:

- $w_i = m_i / (\sum m_j)$
- $x_i = m_i / t_i$
- Note that  $w_1 + w_2 + \dots + w_n = 1$
- The above called the weighted harmonic mean for the data set  $x_i$
- How would the above expression looklike if the weights for the n samples are equal?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

37

## Mean of a Ratio – 1<sup>st</sup> Case

- Given a set of n ratios – How would you summarize them in ONE number
- It depends on the physical meaning of the numbers involved

$$E\left[\frac{a_1}{b_1} + \frac{a_2}{b_2} + \dots + \frac{a_n}{b_n}\right] = \frac{(1/n)\sum_{i=1}^n a_i}{(1/n)\sum_{i=1}^n b_i} = \frac{E[a]}{E[b]}$$

- However, the above is suitable only if the numerator and the denominator do not follow the multiplicative property (i.e.  $a_i \approx c b_i$  where c is a constant).

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

38

## Example: Mean of a Ratio – 1<sup>st</sup> Case

- **Problem:** The CPU utilization of a system as measured over five different intervals is as shown in table. What is the average CPU utilization

- **Solution:**

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

Measurement Duration	CPU Busy (%)
1	45
1	45
1	45
1	45
100	20
sum	200%
Mean	$\neq 200/5$ or 40%

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

39

## Mean of a Ratio – 2<sup>nd</sup> Case

- If the numerator and the denominator do follow the multiplicative property (i.e.  $a_i \approx c b_i$  where  $c$  is a constant), then the geometric mean is used!

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

40

## Example: Mean of a Ratio – 2<sup>nd</sup> Case

- A number of benchmarks were run through a program optimizer. The static size of the program as measured before and after the optimization are shown in table. What is the mean optimization ratio?

- **Solution:**

- **Note:**
  - program sizes vary a lot (2 orders of magnitude between BubbleP and PuzzleP)
  - The after Size is expected to be a scaled version of the before size

- Therefore, geometric mean is used  
Geo Mean = 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	7072	0.99
QuickP	184	112	0.61
SieveP	2908	2879	0.99
Towers	433	307	0.71
Geometric Mean			0.82

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

41

## Summarizing Variability

- Summarizing a data set
  - Mean (discussed in the previous slides) – not enough
  - Variability of the data set
- **Indeces of Dispersion**
  - Range – min and max of observed Data
  - Variance
  - 10- and 90- percentiles
  - Semi-interquantile range
  - Mean absolute deviation

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

42

## Sample Variance

- For a set of n observations  $\{x_1, x_2, \dots, x_n\}$
- Sample variance,  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- Sample mean,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Sample standard deviation,  $s = \sqrt{s^2}$
- Coefficient of variation (COV) relates these two
- Mean absolute deviation:  $= \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

43

## Percentile

- A popular option for specifying dispersion
  - e.g. 5-percentile and 95-percentile
  - A quantile equal to  $\alpha$  is equal to  $\alpha \times 100$  percentile
- Quantile = fractile
- The percentiles at multiples of 10% are called deciles (e.g. first decile = 10% percentile)
- Quartiles: dividing the data into four parts at 25%, 50%, and 75%.

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

44

## **How to Estimate the $\alpha$ -Quantile?**

- Sort the observations
- Take the  $[(n-1)\alpha+1]$ th element in the ordered set
  - $[x]$  is the nearest integer to  $x$
  - For quantiles exactly halfway between two integers, use the lower integer

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

45

## **Semi-Interquartile Range (SIQR)**

- Interquartile range – range between Q3 and Q1
- SIQR – half the interquartile range

$$SIQR = \frac{Q3 - Q1}{2} = \frac{x_{0.75} - x_{0.25}}{2}$$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

46

## Example: 12.4

- In an experiment, which was repeated 32 times, the measured CPU time was found to {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}
- Calculate the 10-percentile?
- Calculate the 90-percentile?
- Calculate Q1, Q2 and SIQR?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

47

## Example: 12.4

- Solution:
- The sorted set = {1.9 2.7 2.8 2.8 2.8 2.9 3.1  
3.1 3.2 3.2 3.3 3.4 3.6 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}
- The 10-percentile is given by  $[(32-1)*0.1+1] = 4^{\text{th}}$  element = 2.8
- The 90-percentile is given by  $[(32-1)*0.9+1] = 29^{\text{th}}$  element = 5.1
- The 1<sup>st</sup> quartile (Q1) is given by  $[(32-1)*0.25+1] = 9^{\text{th}}$  element = 3.2
- The 2<sup>nd</sup> quartile (Q2 or median) is given by  $[(32-1)*0.5+1] = 16^{\text{th}}$  element = 3.9
- The 3<sup>rd</sup> quartile (Q3) is given by  $[(32-1)*0.75+1] = 24^{\text{th}}$  element = 4.5
- Thus  $SIQR = (Q3-Q1)/2 = (4.5 - 3.2)/2 = 0.65$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

48



## **Which Dispersion Index to Use?**

---

- **If variable is bounded – use range**
- **Else**
  - **If distribution is unimodal symmetric – use COV (mean and standard deviation)**
  - **Else use percentiles**
- **See figure 12.4**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

49

## **Determining the Distribution of Data**

---

- **Two methods:**
  - **Histograms**
  - **Quantile-Quantile plot**
- **The Histogram method:**
  - **Determine maximum and minimum**
  - **Divide range into subranges (cells or buckets)**
  - **Determine count of observations in each subrange**
  - **Normalize counts by dividing by the number of all observations**
  - **Plot cell frequencies as column charts**
- **Problems with histogram: How to determine cell size**
  - **Too small cell size – low count (in accurate)**
  - **To large cell size – details of histogram are lost**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

50

## Quantile-Quantile Plots

- Good for small sample size
- A plot of observed quantiles versus theoretical quantiles
- Procedure:
  - If  $y_{(i)}$  is the observed  $q_i^{\text{th}}$  quantile –
  - Using the theoretical distribution, the  $q_i^{\text{th}}$  quantile  $x_i$  is computed
  - Plot the points  $(x_i, y_{(i)})$
  - If the assumed distribution is correct – the plot will be linear
- How to use the theoretical distribution to get the  $q^{\text{th}}$  quantile?
  - Refer to slide 32
  - By definition  $q_i = F(x_i) \rightarrow x_i = F^{-1}(q_i)$  – i.e. we need to find the CDF inverse (refer to table 28.1 for CDF inverses for popular distributions)

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

51

## Example: Samples from $U(0,1)$

- Check whether the following samples follow the uniform distribution  $U(0,1)$ . The samples are {0.1820 0.4930 0.2909 0.7363 0.9375 0.9310 0.1080 0.5985}
- Solution:

The sorted samples are {0.1080 0.1820 0.2909 0.4930 0.5985 0.7363 0.9310 0.9375}.

For the  $U(0,1)$ , the PDF is given by  $f(x) = 1$ , while the CDF is given by  $F(x) = x$  for  $x$  in  $(0,1)$

This means the  $q_i^{\text{th}}$  quantile is given by

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

11/8/2009

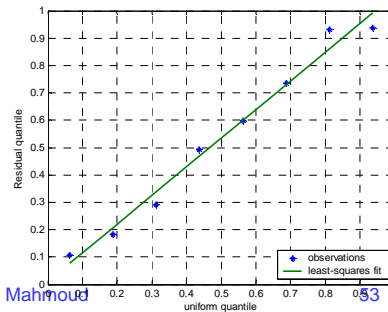
Dr. Ashraf S. Hasan Mahmoud

52

## Example: Samples from U(0,1) – cont'd

- Form the following table
- Plot  $(x_i, y_i)$  pairs
- Since the relation is close to linear – The samples appear to be uniformly distributed

$i$	$q_i = (i-0.5)/n$	$y_i$	$x_i$
1	0.0625	0.1080	0.0625
2	0.1875	0.1820	0.1875
3	0.3125	0.2909	0.3125
4	0.4375	0.4930	0.4375
5	0.5625	0.5985	0.5625
6	0.6875	0.7363	0.6875
7	0.8125	0.9310	0.8125
8	0.9375	0.9375	0.9375



11/8/2009

Dr. Ashraf S. Hasan Mahmoud

## Example: Samples from U(0,1) – cont'd

- The following Matlab code is used to generate this example:

```

0001 clear all
0002 %U(0,1)
0003 N = 8;
0004 y = rand(1,N);
0005 y_sorted = sort(y);
0006 qi = ([1:N]-0.5)/N;
0007 xi = qi;
0008 [P S] = polyfit(xi, y_sorted,1); % find the linear least squares fit
0009 Ye = polyval(P, xi); % evaluate the fitted polynomial
0010 figure(1);
0011 h = plot(xi,y_sorted,'*', xi, Ye,'-');
0012 set(h, 'LineWidth', 2);
0013 axis([0 1 0 1]);
0014 grid
0015 xlabel('uniform quantile');
0016 ylabel('Residual quantile');
0017 legend('observations', 'least-squares fit',4);
    
```

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

54

## Example: Samples from Exp(1)

- Check whether the following samples follow the exponential distribution Exp(1). The samples are {0.5956 0.3293 0.8846 1.0637 0.9959 0.1007 0.1867 0.4457}

- **Solution:**

The sorted samples are {0.1007 0.1867 0.3293 0.4457 0.5956 0.8846 0.9959 1.0637}.

For the Exp(1), the PDF is given by  $f(x) = \exp(-x)$ , while the CDF is given by  $F(x) = 1 - \exp(-x)$  for  $x$  in  $(0, \infty)$

This means the  $q_i^{\text{th}}$  quantile is given by

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

11/8/2009

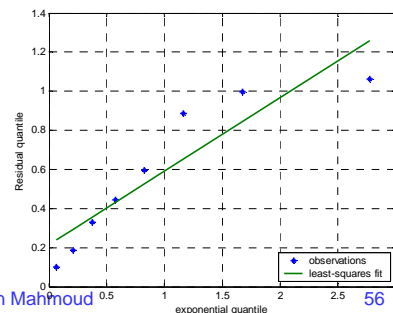
Dr. Ashraf S. Hasan Mahmoud

55

## Example: Samples from Exp(1) – cont'd

- Form the following table
- Plot  $(x_i, y_i)$  pairs
- Since the relation is close to linear – The samples appear to be exponentially distributed

$i$	$q_i = (i-0.5)/n$	$y_i$	$x_i$
1	0.0625	0.1007	0.0645
2	0.1875	0.1867	0.2076
3	0.3125	0.3293	0.3747
4	0.4375	0.4457	0.5754
5	0.5625	0.5956	0.8267
6	0.6875	0.8846	1.1632
7	0.8125	0.9959	1.6740
8	0.9375	1.0637	2.7726



11/8/2009

Dr. Ashraf S. Hasan Mahmoud

56

## Example: Samples from Exp(1) – cont'd

- The following Matlab code is used to generate this example:

```
0001 clear all
0002 %E(1)
0003 N = 8;
0004 y = -1*log(rand(1,N));
0005 y_sorted = sort(y);
0006 qi = ([1:N]-0.5)/N;
0007 xi = -log(1-qi);
0008 [P S] = polyfit(xi, y_sorted,1); % find the linear least squares fit
0009 Ye = polyval(P, xi); % evaluate the fitted polynomial
0010 figure(1);
0011 h = plot(xi,y_sorted,'*', xi, Ye,'-');
0012 set(h, 'LineWidth', 2);
0013 %axis([0 1 0 1]);
0014 grid
0015 xlabel('exponential quantile');
0016 ylabel('Residual quantile');
0017 legend('observations', 'least-squares fit',4);
```

This the same code for the previous example excepts for:

- Line 4 – the generation of the samples
- Line 7 – the calculation of the  $q_i^{\text{th}}$  quantile

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

57

## Example: 12.5 (from the textbook)

- The difference between values measured on a system and those predicted by a model is called the modeling error. The modeling error for eight predictions of a model were found to be -0.4, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, and 0.09.
- Does these sample appear to come from a normal ( $\sim N(0,1)$ ) distribution?

11/8/2009

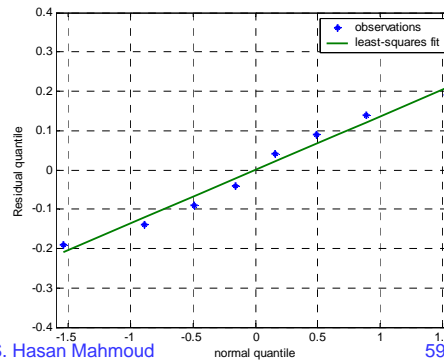
Dr. Ashraf S. Hasan Mahmoud

58

## Example: 12.5 – cont'd

- **Solution:**  
to find the  $q^{\text{th}}$  quantile for  $N(0,1)$  – we need to invert the CDF which is already not a closed form – the  $q^{\text{th}}$  quantile can be approximated by  $x_i = F^{-1}(q_i) \approx 4.19[q_i^{0.14} - (1-q_i)^{0.14}]$   
Therefore, one can build the following table and obtain the corresponding plot
- From the figure, the errors DO APPEAR to be normally distributed.

$i$	$q_i = (i-0.5)/n$	$y_i$	$x_i$
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



11/7/2009

Dr. Ashraf S. Hasan Mahmoud

59

## Sample Versus Population

- **Sample  $\{x_1, x_2, \dots, x_n\}$** 
  - Sample mean =  $\mu_s$
  - Population mean =  $\mu$
- **When  $n$  is extremely large, then sample mean approaches population mean**
- **Population characteristics  $\sim$  parameters**
- **Sample estimates  $\sim$  statistics**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

60

## Confidence Interval for the Mean

- It is NOT possible to get a perfect estimate of the population mean from a finite number of finite size samples
- The best we can do is get PROBABILISTIC bounds
- For example:  $\text{Prob}[c1 \leq \mu \leq c2] = 1 - \alpha$ 
  - With probability  $1 - \alpha$ , the population mean  $\mu$  is between  $c1$  and  $c2$
  - $(c1, c2)$  – confidence interval
  - $\alpha$  is the significance level
  - $100(1 - \alpha)$  is confidence level
  - $1 - \alpha$  is the confidence coefficient

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

61

## Confidence Interval for the Mean (2)

- Consider the sample  $\{x_1, x_2, \dots, x_n\}$ 
  - Assuming large sample (i.e.  $n \sim 30$ )
  - Independent samples
  - $x_i$  has mean  $\mu$  and standard deviation  $\sigma$
  - THEN sample mean  $\mu_s \sim N(\mu, \sigma/\sqrt{n})$  - using the central limit theorem
- Standard deviation of  $\mu_s$  is called standard error
- Note as  $n$  increases, the standard error approaches zero
- Using the central limit theorem, a  $100(1-\alpha)\%$  confidence interval for the population mean is given by
$$(\mu_s - z_{1-\alpha/2}s / \sqrt{n}, \mu_s + z_{1-\alpha/2}s / \sqrt{n})$$
  - $\mu_s$  is the sample mean,
  - $s$  is the sample standard deviation
  - $n$  is the sample size
  - $z_{1-\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of the unit normal variable – See table A.2 for listing of these quantiles

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

62

## Example: Confidence Interval for the Mean

- **Problem:** In an experiment, which was repeated 32 times, the measured CPU time was found to {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}
- Calculate the sample mean, the sample standard deviation and the
- Calculate the 90% confidence interval for the mean.
- Repeat the calculations for 95% and 99% confidence interval for the mean.

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

63

## Example: Confidence Interval for the Mean

### Solution:

Sample size,  $n = 32$

Sample mean,  $\mu_s = \Sigma xi / n = 3.90$

Sample standard deviation,  $s = \sqrt{(\Sigma(xi - \mu_s)^2)/(n-1)}$   
 $= 0.95$

The 90% confidence interval  $\rightarrow$  significance level,  $\alpha = 0.1$ ,  
therefore, the required quantile  $z_{1-\alpha/2} = z_{0.95}$

From table A.2,  $z_{0.95} = 1.645$

Therefore, confidence interval

$$3.90 \pm (1.645)(0.95)/\sqrt{32}$$
$$(3.62, 4.17)$$

- This means we take 100 samples and construct confidence interval for each sample, in 90% of cases the interval will include the population mean, and in 10% of the cases the interval would not include the population mean

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

64



## Example: Confidence Interval for the Mean – cont'd

**Solution: cont'd**

The 95% confidence interval → significance level,  $\alpha = 0.05$ , therefore, the required quantile  $z_{1-\alpha/2} = z_{0.975}$

From table A.2,  $z_{0.975} = 1.960$

Therefore, confidence interval

$$3.90 \pm (1.960)(0.95)/\sqrt{32}$$
$$(3.57, 4.23)$$

The 99% confidence interval → significance level,  $\alpha = 0.01$ , therefore, the required quantile  $z_{1-\alpha/2} = z_{0.995}$

From table A.2,  $z_{0.995} = 2.576$

Therefore, confidence interval

$$3.90 \pm (2.576)(0.95)/\sqrt{32}$$
$$(3.46, 4.33)$$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

65

## Example: Confidence Interval for the Mean – cont'd (Matlab Code)

```
clear all

ConfidenceLevel = 99;
Samples = [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];

n      = length(Samples);
Mue_s  = mean(Samples);
Sigma_s = sqrt(var(Samples));

p      = 1-(1 - ConfidenceLevel/100)/2;
z_p    = norminv(p, 0, 1);

Mue_L  = Mue_s - z_p*Sigma_s/sqrt(n);
Mue_H  = Mue_s + z_p*Sigma_s/sqrt(n);
fprintf('The %7.0f%% confidence interval for the mean = (%7.2f, %7.2f)\n', ...
        ConfidenceLevel, Mue_L, Mue_H);
```

Note: norminv() is the matlab function for computing the required quantiles

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

66

## **Example: Confidence Interval for the Mean – cont'd**

---

### **Figure 13.1 – Meaning of the confidence interval**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

67

## **Confidence Interval for the Mean (3)**

---

- For small sample size ( $n < 30$ ) and if the samples come from a normally distributed population, the  $100(1-\alpha)\%$  confidence interval is given by

$$(\mu_s - t_{[1-\alpha/2;n-1]}s / \sqrt{n}, \mu_s + t_{[1-\alpha/2;n-1]}s / \sqrt{n})$$

- $t_{[1-\alpha/2;n-1]}$  are tabulated in the textbook (A.4)

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

68

## Example: Confidence Interval for the Mean

- **Problem:** The difference between values measured on a system and those predicted by a model is called the modeling error. The modeling error for eight predictions of a model were found to be -0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, and 0.09.
- Calculate the 90% confidence interval for the measured error

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

69

## Example: Confidence Interval for the Mean – cont'd

- **Solution:**  
Sample mean,  $\mu_s = 0$   
Sample size,  $n = 8$   
Sample standard deviation,  $s = 0.138$   
The 90% confidence interval  $\rightarrow$  significance level,  $\alpha = 0.1$ , therefore, the required quantile  $t_{1-\alpha/2} = t_{0.95,7}$   
From table A.4,  $t_{0.95,7} = 1.895$   
Therefore, confidence interval  
 $0 \pm (1.895)(0.138)/\sqrt{8}$   
 $(-0.0926, 0.0926)$

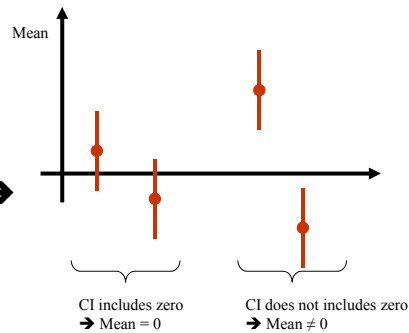
11/8/2009

Dr. Ashraf S. Hasan Mahmoud

70

## Testing For A SPECIFIC Mean Value

- Is the population mean equal to a specific value  $\theta$ ?
- It depends on the confidence interval
  - If the confidence interval contains  $\theta \rightarrow$  Yes
  - If the confidence interval does not contain  $\theta \rightarrow$  No



11/8/2009

Dr. Ashraf S. Hasan Mahmoud

71

## Example: Testing For A SPECIFIC Mean Value

- **Problem:** The difference in the processor times of two different implementations of the same algorithm was measured on seven similar workload. 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?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

72

## Example: Testing For A SPECIFIC Mean Value – cont'd

- **Solution:**

Sample size,  $n = 7$

Sample mean,  $\mu_s = 1.03$

Sample variance = 2.57  $\rightarrow s = 1.60$

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

$100(1 - \alpha) = 99\% \rightarrow \alpha = 0.01 \rightarrow 1 - \alpha/2 = 0.995$

Therefore,  $t_{0.995,6} = 3.707$

Hence, 99% confidence interval =  $(-1.21, 3.27)$  – includes the zero

Therefore, we can not say with 99% confidence that the mean difference is significantly different from zero

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

73

## Comparing Two Alternatives

- Often it is required to compare two or more systems
- If the requirement is to compare
  - TWO SYSTEMS under
  - Similar work loads
  - $\rightarrow$  Then we can use confidence intervals to perform the comparison
- ELSE use simulation techniques!!
- For two systems under similar work loads, the reading can be
  - Paired (i.e. follow the form  $(x,y)$ )
  - Unpaired

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

74

## Comparing Two Alternatives – cont'd

---

- For two systems under similar work loads, the reading can be
  - Paired (i.e. follow the form  $(x,y)$ )
  - Unpaired
- For the paired case:
  - Form the sample  $(x-y)$
  - If the confidence interval for the difference sample contain the zero, then the two systems are not significantly different!!
  - See the matlab function "signtest( )"
- For the unpaired case
  - Perform the t-test – to be explained in the coming slides
  - See the matlab function "ttest( )"

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

75

## Example: Paired Observations - Comparing Two Alternatives

---

- **Problem:** Six similar workloads were used on two systems. The observations are  $(\{15.3, 19.1\}, \{16.6, 3.5\}, \{0.6, 3.4\}, \{1.4, 2.5\}, \{0.6, 3.6\}, \{7.3, 1.7\})$
- Is one system better than the other?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

76

## Example: Paired Observations - Comparing Two Alternatives – cont'd

- **Solution:**

The performance difference constitute a sample of six observations {-13.7, 13.1, -2.8, -1.1, -3.0, 5.6}

Following the same procedure for testing for a zero mean, results in:

```
Sample size           = 6
Sample mean           = -0.317
Sample standard deviation = 9.034
Confidence level 100(1-a) = 90% ==> a = 0.100 and 1-a/2 = 0.9500
confidence interval for mean = -0.317 +- tp * 9.034 / sqrt( 6)
confidence interval for mean = -0.317 +- tp * ( 3.688)
the 0.9500-quantile of the t-variate with 5 degrees of freedom t = 2.0150
The 90% confidence interval is given by ( -7.749, 7.115)
Confidence interval ( -7.75, 7.12) contains the zero
```

Therefore, the two systems are not different

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

77

## Example: Paired Observations - Comparing Two Alternatives – cont'd (Matlab Code - 1)

- **Code that can be used for solving examples in this section:**

```
0001 clear all
0002 %
0003 % Code for generating confidence intervals - can be used also for
0004 % - testing for a specific mean value
0005 % - comparing paired observations
0006 ConfidenceLevel = 90; % required confidence level
0007 % put your samples here
0008 Samples = [-13.7, 13.1, -2.8, -1.1, -3.0, 5.6];
0009 n = length(Samples);
0010 Mue_s = mean(Samples);
0011 Sigma_s = sqrt(var(Samples));
0012
0013 p = 1-(1 - ConfidenceLevel/100)/2;
0014
0015 fprintf('Sample size           = %3d\n', n);
0016 fprintf('Sample mean           = %7.3f\n', Mue_s);
0017 fprintf('Sample standard deviation = %7.3f\n', Sigma_s);
0018 fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2
= %5.4f\n', ...
0019 ConfidenceLevel, 1- ConfidenceLevel/100, p);
```

Example\_13\_4.m

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

78

## Example: Paired Observations - Comparing Two Alternatives – cont'd (Matlab Code - 2)

- Code that can be used for solving examples in this section – cont'd:

```
0020 %
0021 % check whether to use the normal quantile or the t-distribution
0022 if (n>30)
0023     z_p = norminv(p, 0, 1);
0024     fprintf('confidence interval for mean = %7.3f +- z_p * %7.3f / sqrt(%3d)\n', ...
0025           Mue_s, Sigma_s, n);
0026     fprintf('confidence interval for mean = %7.3f +- z_p * (%7.3f)\n', ...
0027           Mue_s, Sigma_s/sqrt(n));
0028     fprintf('the %5.4f-quantile of the normal-variate z = %7.4f\n', p, z_p);
0029     Mue_L = Mue_s - z_p*Sigma_s/sqrt(n);
0030     Mue_H = Mue_s + z_p*Sigma_s/sqrt(n);
0031     fprintf('The %7.0f%% confidence interval is given by (%7.3f, %7.3f)\n', ...
0032           ConfidenceLevel, Mue_L, Mue_H);
0033 else
0034     t_p = tinv(p, n-1);
0035     fprintf('confidence interval for mean = %7.3f +- t_p * %7.3f / sqrt(%3d)\n', ...
0036           Mue_s, Sigma_s, n);
0037     fprintf('confidence interval for mean = %7.3f +- t_p * (%7.3f)\n', ...
0038           Mue_s, Sigma_s/sqrt(n));
0039     fprintf('the %5.4f-quantile of the t-variate with %2d degrees of freedom t = %7.4f\n', ...
0040           p, n-1, t_p);
0041     Mue_L = Mue_s - t_p*Sigma_s/sqrt(n);
0042     Mue_H = Mue_s + t_p*Sigma_s/sqrt(n);
0043     fprintf('The %7.0f%% confidence interval is given by (%7.3f, %7.3f)\n', ...
0044           ConfidenceLevel, Mue_L, Mue_H);
0045 end
0046 if (Mue_L*Mue_H < 0)
0047     fprintf('Confidence interval (%7.2f, %7.2f) contains the zero\n', ...
0048           Mue_L, Mue_H);
0049 else
0050     fprintf('Confidence interval (%7.2f, %7.2f) does NOT contain the zero\n', ...
0051           Mue_L, Mue_H);
0052 end
```

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

79

## Unpaired Observations - Comparing Two Alternatives

- **t**-test – refer to textbook

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

80



## Example: Unpaired Observations - Comparing Two Alternatives

- **Problem:** The processor time required to execute a task was measured on two systems. The times on system A were {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}. The times on system B were {19.2, 3.52, 3.38, 2.5, 3.60, 1.74}
- Are the two system significantly different?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

81

## Example: Unpaired Observations - Comparing Two Alternatives – cont'd

- **Solution:**  
Following the procedure for the  $t$ -test:  
System A:  
Sample size = 6  
Sample mean = 5.310  
Sample standard deviation = 6.158  
System B:  
Sample size = 6  
Sample mean = 5.657  
Sample standard deviation = 6.674  
Confidence level  $100(1-\alpha) = 90\% \Rightarrow \alpha = 0.100$  and  $1-\alpha/2 = 0.9500$   
Mean difference  $\mu_A - \mu_B = -0.347$   
Sigma for mean difference = 3.707  
Effective number of degrees of freedom,  $f = 11.910$  ( 12)  
confidence interval for mean =  $-0.347 \pm t_p * 3.707$   
the 0.9500-quantile of the  $t$ -variate with 12 degrees of freedom  $t = 1.7823$   
The 90% confidence interval is given by ( -6.954, 6.261)  
**The two systems ARE NOT significantly different**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

82

## Example: Unpaired Observations - Comparing Two Alternatives – cont'd (Matlab Code - 1)

- **Solution:**

- **Matlab code for performing the t-test:**

```

0001 clear all
0002 %
0003 % Code for t-test
0004 ConfidenceLevel = 90; % required confidence level
0005 % put your samples here
0006 Samples_A = [5.36, 16.57, 0.62, 1.41, 0.64, 7.26];
0007 Samples_B = [19.2, 3.52, 3.38, 2.5, 3.60, 1.74];
0008 n_A = length(Samples_A);
0009 Mue_s_A = mean(Samples_A);
0010 Sigma_s_A = sqrt(var(Samples_A));
0011
0012 n_B = length(Samples_B);
0013 Mue_s_B = mean(Samples_B);
0014 Sigma_s_B = sqrt(var(Samples_B));
0015
0016 Mean_Difference = Mue_s_A - Mue_s_B;
0017 Sigma_Mean_Difference = sqrt(Sigma_s_A*Sigma_s_A/n_A + Sigma_s_B*Sigma_s_B/n_B);
0018 Effective_number = (Sigma_Mean_Difference^4) / ...
0019 ((Sigma_s_A*Sigma_s_A/n_A)^2/(n_A+1) + (Sigma_s_B*Sigma_s_B/n_B)^2/(n_A+1)) ...
0020 - 2;
0021 Effective_number_rounded = round(Effective_number);
0022
0023 p = 1-(1 - ConfidenceLevel/100)/2;
0024 t_p = tinv(p, Effective_number_rounded);
0025
0026 fprintf('System A:\n');
0027 fprintf('Sample size = %3d\n', n_A);
0028 fprintf('Sample mean = %7.3f\n', Mue_s_A);
0029 fprintf('Sample standard deviation = %7.3f\n', Sigma_s_A);
0030 fprintf('System B:\n');
0031 fprintf('Sample size = %3d\n', n_B);
0032 fprintf('Sample mean = %7.3f\n', Mue_s_B);
0033 fprintf('Sample standard deviation = %7.3f\n', Sigma_s_B);

```

If  $n > 30$ , use the z-value (or the norminv() function) in line 0024

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

83

## Example: Unpaired Observations - Comparing Two Alternatives – cont'd (Matlab Code - 2)

- **Solution:**

- **Matlab code for performing the t-test: cont'd**

```

0034
0035 fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2 = %5.4f\n', ...
0036 ConfidenceLevel, 1- ConfidenceLevel/100, p);
0037
0038 fprintf('Mean difference Mue_A - Mue_B = %7.3f\n', Mean_Difference);
0039 fprintf('Sigma for mean difference = %7.3f\n', Sigma_Mean_Difference);
0040 fprintf('Effective number of degrees of freedom, f = %7.3f (%3d)\n', ...
0041 Effective_number, Effective_number_rounded);
0042
0043 fprintf('confidence interval for mean = %7.3f +- tp * %7.3f \n', ...
0044 Mean_Difference, Sigma_Mean_Difference);
0045 fprintf('the %5.4f-quantile of the t-variate with %2d degrees of freedom t = %7.4f\n', ...
0046 p, Effective_number_rounded, t_p);
0047 Mue_L = Mean_Difference - t_p*Sigma_Mean_Difference;
0048 Mue_H = Mean_Difference + t_p*Sigma_Mean_Difference;
0049 fprintf('The %7.0f%% confidence interval is given by (%7.3f, %7.3f)\n', ...
0050 ConfidenceLevel, Mue_L, Mue_H);
0051
0052 if (Mue_L*Mue_H < 0)
0053 fprintf('The two systems ARE NOT significantly different\n', ...
0054 Mue_L, Mue_H);
0055 else
0056 fprintf('The two systems ARE significantly different\n', ...
0057 Mue_L, Mue_H);
0058 end

```

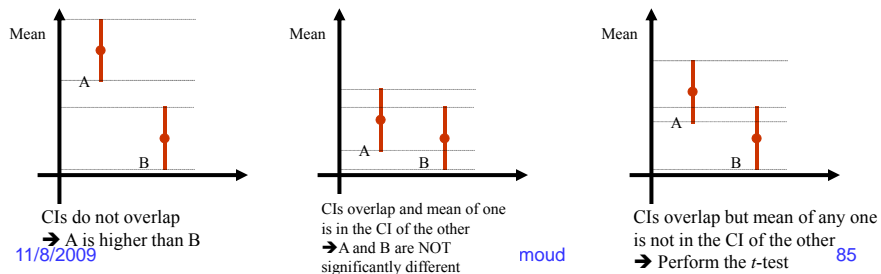
11/8/2009

Dr. Ashraf S. Hasan Mahmoud

84

## Approximate Visual Test - Comparing Two Alternatives

- **Simpler than t-test**
- **Procedure:**
  - **Compute confidence interval (CI) for each alternative**
  - **If CIs do not overlap → the two systems are significantly different**
  - **Else CIs overlap and mean of one is in the CI of the other → the two system are NOT significantly different**
  - **Else CIs overlap but mean of any one is not in the CI of the other → perform the  $t$ -test**



## One-Sided Confidence Interval

- **For a two-sided confidence level of  $100(1-\alpha)\%$** 
  - **There is a  $100\alpha/2\%$  chance the sample will be more than the upper confidence limit**
  - **There is a  $100\alpha/2\%$  chance the sample will be less than the upper confidence limit**
- **To test a hypothesis that the mean is greater than a certain value – use one-sided confidence interval**
  - **Given by  $(\mu_s - t_{[1-\alpha; n-1]}s / \sqrt{n}, \mu_s)$**
- **The one-sided upper confidence interval for the population mean**
  - **Given by  $(\mu_s, \mu_s + t_{[1-\alpha; n-1]}s / \sqrt{n})$**
- **For large ( $n > 30$ ) samples, z-values are used instead of t-values.**

## Example: One-Sided Confidence Interval

- **Problem: Refer to example 13.8 in textbook**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

87

## Confidence Interval for Proportions

- **For categorical variables, the statistical data often consist of probabilities associated with various categories**
  - Such probabilities are called **PROPORTIONS**
- **How to generate a confidence interval for an proportion estimate?**
- **Procedure:**
  - Sample proportion =  $p = n1/n$
  - CI for proportion =  $p \pm z_{1-\alpha/2} \sqrt{p(1-p)/n}$
- **Condition:  $np \geq 10$  (Binomial distribution  $\approx$  Normal distribution)**
  - If condition is not satisfied – can not use *t*-test
  - Procedure not defined at this stage

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

88

## Example: Confidence Interval for Proportions

---

- **Problem:** If 10 out 1000 pages printed on a laser printer are illegible.
- **Characterize the proportion of illegible pages using a 90% and 95% confidence intervals**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

89

## Example: Confidence Interval for Proportions – cont'd

---

- **Solution:**

**For 90% confidence:**

```
Sample proportion      = 0.010
n*p = 10.000 >= 10 is satisfied
Confidence level 100(1-a) = 90% ==> a = 0.100 and 1-a/2 = 0.9500
confidence interval for proportion = 0.0100 +- za * sqrt( 0.010 *
0.990 / 1000)
confidence interval for proportion = 0.0100 +- za * 0.003
the 0.9500-quantile of the normal-variate z = 1.6449
The 90% confidence interval is given by (0.0048, 0.0152)
```

**For 95% confidence:**

```
the 0.9750-quantile of the normal-variate z = 1.9600
The 95% confidence interval is given by (0.0038, 0.0162)
```

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

90

## Example: Confidence Interval for Proportions – cont'd (Matlab Code)

- **Matlab code for confidence interval for proportions:**

```
0001 clear all
0002 %
0003 % Code for generating confidence intervals for proportions
0004 ConfidenceLevel = 90; % required confidence level
0005 % put your samples here
0006 n1 = 10;
0007 n = 1000;
0008 p = n1/n;
0009
0010 a = 1-(1 - ConfidenceLevel/100)/2;
0011
0012 fprintf('Sample proportion = %7.3f\n', p);
0013 if (p*n<10)
0014     fprintf('n*p = %7.3f >= 10 is not satisfied\n',n*p);
0015 else
0016     z_a = norminv(a,0,1);
0017     fprintf('n*p = %7.3f >= 10 is satisfied\n',n*p);
0018     fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2 = %5.4f\n', ...
0019           ConfidenceLevel, 1- ConfidenceLevel/100, a);
0020     fprintf('confidence interval for proportion = %7.4f +- za * sqrt(%7.3f * %7.3f / %4d)\n', ...
0021           p, p, 1-p, n);
0022     fprintf('confidence interval for proportion = %7.4f +- za * %7.3f\n', ...
0023           p, sqrt(p*(1-p)/n));
0024     fprintf('the %5.4f-quantile of the normal-variate z = %7.4f\n', a, z_a);
0025     p_L = p - z_a*sqrt(p*(1-p)/n);
0026     p_H = p + z_a*sqrt(p*(1-p)/n);
0027     fprintf('The %7.0f%% confidence interval is given by (%5.4f, %5.4f)\n', ...
0028           ConfidenceLevel, p_L, p_H);
0029 end
```

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

91

## Example: Confidence Interval for Proportions – Testing for Zero

- **Problem:** A single experiment was repeated on two systems 40 times. System A was found superior to system B in 26 repetitions.
- **Can we state with 99% confidence that system A is superior?**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

92

## Example: Confidence Interval for Proportions – Testing for Zero – cont'd

- **Solution:**

- **For 99% confidence:**

```
Sample proportion = 0.650
n*p = 26.000 >= 10 is satisfied
Confidence level 100(1-a) = 99% ==> a = 0.010 and 1-a/2 = 0.9950
confidence interval for proportion = 0.6500 +- za * sqrt( 0.650 * 0.350 / 40)
confidence interval for proportion = 0.6500 +- za * 0.075
the 0.9950-quantile of the normal-variate z = 2.5758
The 99% confidence interval is given by (0.4557, 0.8443)
```

- **We note that 0.5 (the point of equality between two systems) is included in the interval → we can NOT say with 99% that A is superior**

- **For 90% confidence:**

```
Sample proportion = 0.650
n*p = 26.000 >= 10 is satisfied
Confidence level 100(1-a) = 90% ==> a = 0.100 and 1-a/2 = 0.9500
confidence interval for proportion = 0.6500 +- za * sqrt( 0.650 * 0.350 / 40)
confidence interval for proportion = 0.6500 +- za * 0.075
the 0.9500-quantile of the normal-variate z = 1.6449
The 90% confidence interval is given by (0.5260, 0.7740)
```

- **We note that 0.5 (the point of equality between two systems) is NOT included in the interval → we can say with 90% that A is superior**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

93

## Determining Sample Size

- **Previously, we were given a sample set and required to calculate the confidence interval for some confidence level**
- **The other side of the coin – Can you calculate the size of the samples set for a required confidence level?**
  - **E.g. how many iterations should you run your code for a 95% confidence in the collected mean throughput?**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

94

## **Determining Sample Size – cont'd**

- **Suppose we want to estimate the mean with an accuracy of  $\pm r\%$  and a confidence level of  $100(1-\alpha)\%$**
- **We know the confidence interval is given by  $\mu_s \pm z \times s / \sqrt{n} = \mu_s (1 \pm r/100)$**
- **Therefore,  $n = (100 z s / (r \mu_s))^2$**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

95

## **Example: Determining Sample Size**

- **Problem: Based on a preliminary test, the sample mean of the response time is 20 seconds, and the sample standard deviation is 5.**
- **How many repetitions are needed to get the response time accurate within 1 second at 90% confidence?**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

96



## Example: Determining Sample Size – cont'd

---

- **Solution:**

Required accuracy = 1 in 20 = 5%

$\mu_s = 20, s = 5, r = 5\%$

Confidence level  $100(1-a) = 95\% \implies a = 0.05$  and  $1-a/2 = 0.9750$

$z_{0.975} = 1.960$

Therefore, required repetitions

$$n = \frac{(100)(1.960)(5)}{(5)(20)}^2 = 9.8^2 = 96.04$$

A total of 97 observations are required.

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

97

## Sample Size for Determining Proportions

---

- The CI for proportions is given by

$$p \pm z\sqrt{p(1-p)/n}$$

- To get half-width (accuracy of)  $r$ ,

$$p \pm r = p \pm z\sqrt{p(1-p)/n}$$

$$p(1-p)$$

- Therefore,  $n = z^2 \frac{p(1-p)}{r^2}$

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

98

## Example: Determining Sample Size for Proportions

---

- **Problem:** A preliminary measurement of a laser printer showed an illegible print rate of 1 in 10,000.
- How many pages must be observed to get an accuracy of 1 per million at 95% confidence?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

99

## Example: Determining Sample Size for Proportions – cont'd

---

- **Solution:**

$$p = 1/10\,000 = 10^{-4}, r = 10^{-6}, z = 1.960$$

$$n = (1.960)^2 \frac{10^{-4}(1-10^{-4})}{(10^{-6})(10^{-6})}$$

$$= 384,160,000$$

**A total of 384.16 million pages must be observed.**

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

100

## Sample Size for Comparing Two Alternatives

---

- Utilizing the previous info, we need to make the CI for the two systems non overlapping (refer to "Visual Test" slide)
- Therefore, the upper edge of the lower confidence interval should be below the lower edge of the upper confidence interval

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

101

## Example: Sample Size for Comparing Two Alternatives

---

- **Problem:** Two packet-forwarding algorithms were measured. Preliminary measurements showed that algorithm A loses 0.5% of packets and algorithm B loses 0.6%.
- How many packets do we need to observed to state with 95% that algorithm A is better than algorithm?

11/8/2009

Dr. Ashraf S. Hasan Mahmoud

102

## Example: Sample Size for Comparing Two Alternatives – cont'd

---

- **Solution:**

$$\text{CI for algorithm A} = 0.005 \pm 1.960 (0.005(1-0.005)/n)^{1/2}$$
$$\text{CI for algorithm B} = 0.006 \pm 1.960 (0.006(1-0.006)/n)^{1/2}$$

For A to be better than B

upper edge of CI for A should be lower than lower edge of CI for B, ie.

$$0.005 + 1.960 (0.005(1-0.005)/n)^{1/2} < 0.006 - 1.960 (0.006(1-0.006)/n)^{1/2}$$

$$\rightarrow n > 84340$$

**We need to observe 85,000 packets**