# King Fahd University of Petroleum & Minerals Computer Engineering Dept

**CSE 642 – Computer Systems Performance**

**Term 043**

**Dr. Ashraf S. Hasan Mahmoud**

**Rm 22-148-3**

**Ext. 1724**

**Email: ashraf@ccse.kfupm.edu.sa**

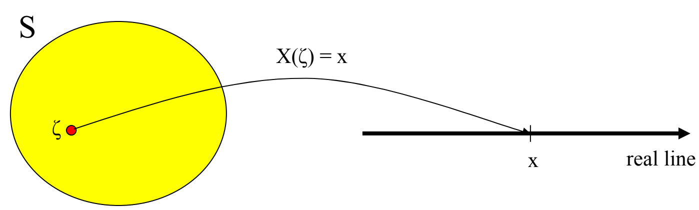


# Primer on Probability Theory

# What is a Random Variable?

- **Random Experiment**
- **Sample Space**

- **Def: A random variable X is a function that assigns a number of X(ζ) to each outcome ζ in the sample space of S of the random experiment**






# Set Functions

- Define **Ω** as the set of all possible outcomes
- Define **A** as set of events
- Define A as an event – subset of the set of all experiments outcomes
- Set operations:
  - Complement $A^c$: is the event that event A does not occur
  - Intersection $A \cap B$ : is the event that event A and B occur
  - Union $A \cup B$: is the event that event A or B occur
  - Inclusion $A \subseteq B$: An event A occurring implying events B occurs

## Set Functions

- Note:
  - Set of events **A** is closed under set operations
  - Φ – empty set
  - A ∩ B = Φ ➔ are mutually exclusive or disjoint



## Axioms of Probability

- Let P(A) denote probability of event A:
  1. For any event A belongs **A**, P(A) ≥ 0;
  2. For set of all possible outcomes **Ω**, P(**Ω**) = 1;
  3. If A and B are disjoint events, P(A un B) = P(A) + P(B)
  4. For countably infinite sets, $A_1$, $A_2$, … such that $A_i$ ins $A_j$ = Φ for i≠j

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

# Additional Properties

- For any event, $P(A) \leq 1$
- $P(A^C) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) \leq P(B)$ for $A \subseteq B$



# Conditional Probability

- Conditional probability is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

- P(A/B) probability of event A conditioned on the occurrence of event B
- Note:
  - A and B are *independent* if $P(A \cap B) = P(A)P(B) \rightarrow P(A/B) = P(A)$
  - Independent IS NOT EQUAL TO mutually exclusive

## The Law of Total Probability

- A set of events $A_i$, i = 1, 2, …, n partitions the set of experimental outcomes if

$$\bigcup_{i=1}^{n} A_i = \Omega$$

and

$$A_i \cap A_j = \Phi$$

Then we can write any event B in terms of $A_i$, i = 1, 2, …, n as

$$B = \bigcup_{i=1}^{n} A_i \cap B$$

Furthermore,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B)$$



## Bayes' Rule

- Using the total law of probability and applying it to the definition of the conditional probability, yields

$$P(A_i / B) = \frac{P(A_i \cap B)}{\sum_{i=1}^{n} P(A_i \cap B)}$$

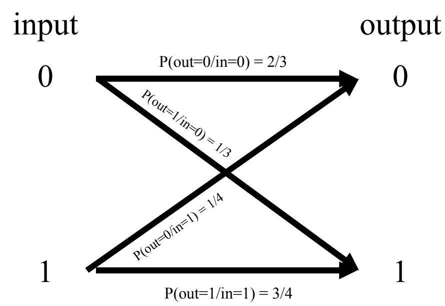
$$= \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}$$

# Example: Binary Symmetric Channel

- Given the binary symmetric channel depicted in figure, find P(input = j / output = i); i,j = 0,1. Given that P(input = 0) = 0.4, P(input = 1) = 0.6.

**Solution:**

**Do it yourself!**






# The Cumulative Distribution Function

- **The cumulative distribution function (cdf) of a random variable X is defined as the probability of the event {X ≤ 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**

## Properties of the CDF

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



## Example 1: Exponential Random Variable

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

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

**Find the CDF of X. Find Prob [T < X ≤ 2T] where T = 1/ λ**

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

- **Answer:**

**The CDF of X is**

$$F_X(x) = \text{Prob}\{X \le x\} = 1 - \text{Prob}\{X > x\}$$
$$= 1 - e^{-\lambda x} \quad x \ge 0$$
$$= 0 \quad x < 0$$

$$\text{Prob}\{T < X \le 2T\} = F_X(2T) - F_X(T)$$
$$= 1-e^{-2} - (1-e^{-1})$$
$$= 0.233$$



## 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**

## Example 2: cont'd

- **Answer:**

**The CDF of W is found as follows:**

$F_X(x) = \text{Prob}\{W \le x\}$
$\quad = \text{Prob}\{W \le x/\text{idle}\}p + \text{Prob}\{W \le x/\text{busy}\}(1-p)$

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

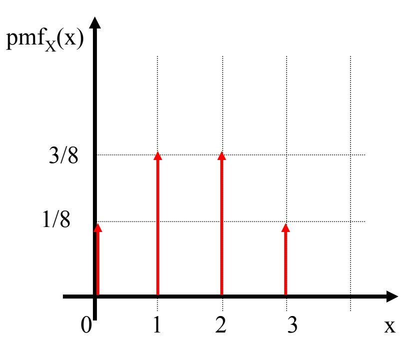
$\Rightarrow$

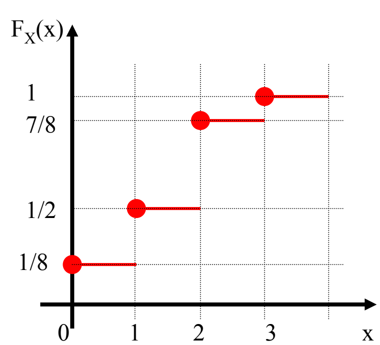
$F_X(x) = 0 \qquad x < 0$
$\quad = p+(1-p)(1- e^{-\lambda x}) \qquad x \ge 0$



## Types of Random Variables

- **(1) Discrete Random Variables**
  - **CDF is right continuous, staircase function of x, with jumps at countable set x0, x1, x2, …**




Pmf: probability mass function

## Types of Random Variables

- **(2) Continuous Random Variables**
  - **CDF is continuous for all values of x → 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**

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

f(t) is referred to as the probability density function or PDF



## 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

$$F_X(x/A) = \frac{P([X \le x] \cap A)}{P(A)} \quad \text{if } P(A) > 0$$

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: $E[X] = \int_{0}^{\infty} \left(1 - F_x(t)\right) dt$ , 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:**

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

- **For discrete X:**

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

- **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?**

## Covariance

- Consider two random variables x and y, such that

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

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

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

- Two variables are independent ➔ $\sigma^2_{xy} = 0$, but the reverse it not always TRUE!!



## Correlation Coefficient

- Correlation Coefficient: normalized value of the covariance

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

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

## Mean/Variance of Sums

- If x1, x2, …, xk are k r.v. and a1, a2, …, ak are k arbitrary constants, then

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

- For independent variables:

$$Var(a_1x_1 + a_2x_2 + \cdots + a_kx_k) = a_1Var(x_1) + a_2Var(x_2) + \cdots + a_kVar(x_k)$$



## 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

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

- Quantile – percentile – fractile – quartile?

## Median

- **The 50-percentile (or 0.5 quantile) for the r.v.**

- **i.e. $x_{0.5}$ such that**

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



## 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.)**

$$f_X(x_m) \ge f_X(x) \qquad \forall x$$

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

$$p_{x_m} \ge p_{x_i} \qquad \forall i$$

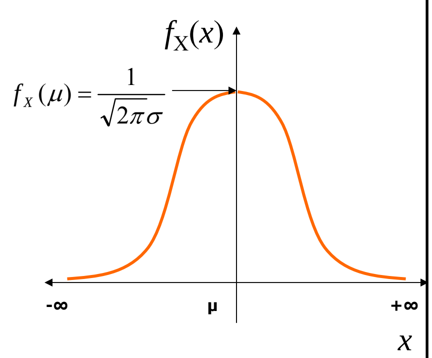
# Mean – Median - Mode

- **Figure 12.1**



# Normal Distribution - General

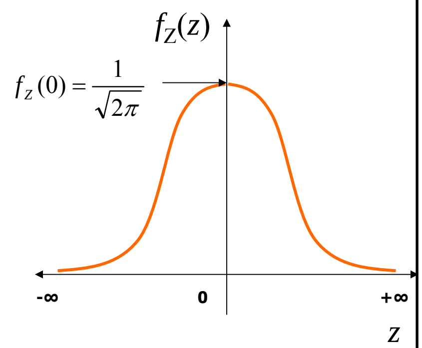
- **More details on later slides**
- **One of the most commonly used distributions**
- **X ~ N(μ, σ) means**
  - **X is a random variables taking values ranging from -∞ to +∞**
  - **X has a mean of μ and standard deviation of σ ➔ i.e. E[X] = μ, and Var[X] = σ².**
- **The corresponding probability density function is given by**




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

## Normal Distribution – Zero Mean and unity Variance

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




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

7/30/2005 Dr. Ashraf S. Hasan Mahmoud 37

## 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**

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

**Or**

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

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

Show the PQRS tool

7/30/2005 Dr. Ashraf S. Hasan Mahmoud 38

## Why is the Normal Distribution Important

- Two reasons:
  - The sum of n independent normal variates is a normal variate,
    - i.e, if x1, x2, …, xn are n independent r.v. ($x_i \sim N(\mu_i, \sigma_i)$), then
    - $Y = \Sigma a_i x_i$ is also a normal variable with $Y \sim N(\mu, \sigma)$, where $\mu = \Sigma a_i \mu_i$ and $\sigma^2 = \Sigma a_i^2 \sigma_i^2$
  - The sum of a large number of *independent* observations from *any distribution* tends to have a normal distribution – central limit theorem



## 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

## Common Misueses 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**



## 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 X 2 = 46!**

  **The average number of subprocesses per user is dependent on the load or the number of users in the system ➔ i.e. the two r.v. are correlated and therefore $E[xy] \neq E[x]E[y]$!**

# Geometric Mean

- **The geometric mean for n values x1, x2, …, xn is given by**

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

- **Another notation:**

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

- **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**



# Example 12.2

- **<u>Problem</u>: 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?**

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

- **<u>Solution</u>: The improvements work in a multiplicative manner**

**Average improvement per layer**

**= [(1.18)(1.13)(1.11)(1.08)(1.10)(1.28)(1.05)]^(1/7) – 1**

**= 0.13**

**i.e average improvement per layer = 13%**

## 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

$$\ddot{x} = \sum_{i=1}^{n} m_i \Big/ \sum_{i=1}^{n} t_i$$

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



## Harmonic Mean – cont'd

- The previous expression can be written as

$$\ddot{x} = \frac{1}{w_1/x_1 + w_2/x_2 + \cdots + w_n/x_1}$$

where:

- $w_i = m_i/(\Sigma m_j)$
- $x_i = m_i/t_i$

- Note that $w_1+w_2+ \ldots + w_n = 1$
- 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?

## Mean of a Ratio – 1st 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**

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

- **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).**



## Example: Mean of a Ratio – 1st Case

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

- **<u>Solution</u>:**

**Mean CPU utilization = sum of CPU busy times / sum of measurement duration**

**= (0.45 + 0.45 + 0.45 + 0.45 + 20) / (1 + 1 + 1 + 1 + 100)**

**= 21%**

| Measurement Duration | CPU Busy (%) |
|---|---|
| 1 | 45 |
| 1 | 45 |
| 1 | 45 |
| 1 | 45 |
| 100 | 20 |
| sum | 200% |
| Mean | ≠200/5 or 40% |

## Mean of a Ratio – 2nd 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!



## Example: Mean of a Ratio – 2nd 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?

| Program | Code Size Before | Code Size After | Ratio |
|---|---|---|---|
| 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 |

- **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**

## 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**



## Sample Variance

- **For a set of n oberservation {x1, x2, …, xn}**
- **Sample variance, $s^2$** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **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:** $= \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

## Percentile

- A popular option for specifying dispersion
  - e.g. 5-percentile and 95-percentile
  - A quantile equal to $\alpha$ is equal to $\alpha$X100 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%.



## How to Estimate the α-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

## Semi-Interquartile Range (SIQR)

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

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



## 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 10-percentile?
- Calculate Q1, Q2 and SIQR?

## 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] = 4th element = 2.8
- The 90-percentile is given by [(32-1)*0.9+1] = 29th element = 5.1
- The 1st quartile (Q1) is given by [(32-1)*0.25+1] = 9th element = 3.2
- The 2st quartile (Q2 or median) is given by [(32-1)*0.5+1] = 16th element = 3.9
- The 3rd quartile (Q3) is given by [(32-1)*0.75+1] = 24th element = 4.5
- Thus SIQR = (Q3-Q1)/2 = (4.5 – 3.2)/2 = 0.65



## 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

## 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



## Quantile-Quantile Plots

- Good for small sample size
- A plot of observed quantiles versus theoretical quantiles
- Procedure:
  - If $y_{(i)}$ is the observed $q_i^{th}$ quantile –
  - Using the theoretical distribution, the $q_i$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 $q^{th}$ 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)

## 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 $q_i^{th}$ quantile is given by

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

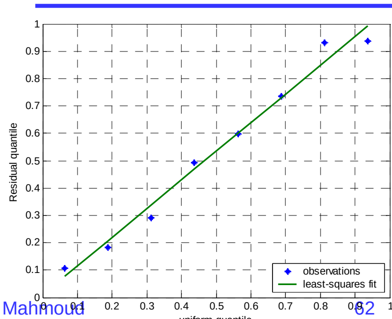


## Example: Samples from U(0,1) – cont'd

- Form the following table

| 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 |

- Plot $(x_i, y_i)$ pairs




- Since the relation is close to linear – The samples appear to be uniformly distributed

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

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

```
clear all
%U(0,1)
N = 8;
y = rand(1,N);
y_sorted = sort(y);
qi       = ([1:N]-0.5)/N;
xi       = qi;
[P S] = polyfit(xi, y_sorted,1); % find the linear least squares fit
Ye    = polyval(P, xi);      % evaluate the fitted polynomial
figure(1);
h = plot(xi,y_sorted,'*', xi, Ye,'-');
set(h, 'LineWidth', 2);
axis([0 1 0 1]);
grid
xlabel('uniform quantile');
ylabel('Residual quantile');
legend('observations', 'least-squares fit',4);
```



## 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 $q_i^{th}$ quantile is given by**

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

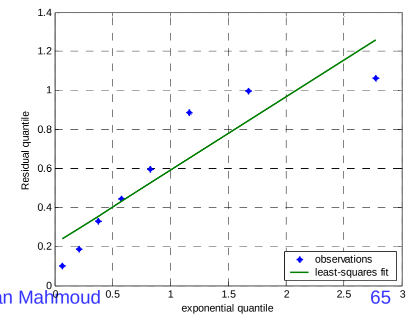
## Example: Samples from Exp(1) – cont'd

- **Form the following table**

| 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 |

- **Plot ($x_i$, $y_i$) pairs**

- **Since the relation is close to linear – The samples appear to be exponentially distributed**






## 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 $q_i^{th}$ quantile**

## 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.**

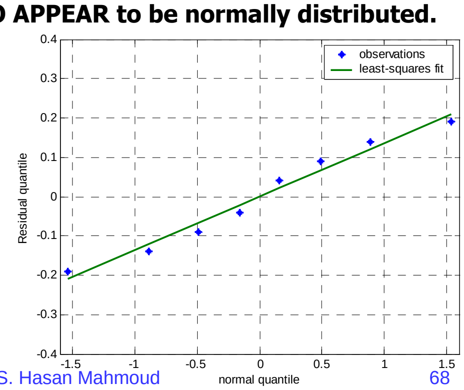
- **Does these sample appear to come from a normal (~N(0,1)) distribution?**



## Example: 12.5 – cont'd

- **Solution:**
  **to find the $q^{th}$ quantile for N(0,1) – we need to invert the CDF which is already not a closed form – the $q_i^{th}$ quantile can be approximated by $x_i = F\text{-}1(q_i) \approx 4.19[q_i\text{^}0.14 - (1-q_i)\text{^}0.14]$**
  **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 |



## Sample Versus Population

- Sample {x1, x2, …, xn}
  - Sample mean = $\mu_s$
  - Population mean = $\mu$
- When n is extremely large, then sample mean approaches population mean
- Population characteristics ~ parameters
- Sample estimates ~ statistics



## 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: 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

## Confidence Interval for the Mean (2)

- Consider the sample {x1, x2, ..., xn}
  - Assuming large sample (i.e. n ~ 30)
  - Independent samples
  - xi is 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-α)% 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- α/2)-quantile of the unit normal variable – See table A.2 for listing of these quantiles



## Example: Confidence Interval for the Mean

- <u>Problem</u>: 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.

## 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 ➔ 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 with include the population mean, and in 10% of the cases the interval would not include the population mean**



## 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)**

## 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



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

**<u>Figure 13.1 –Meaning of the confidence interval</u>**

## 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-α)% 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)**



## 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**

## 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 → 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**

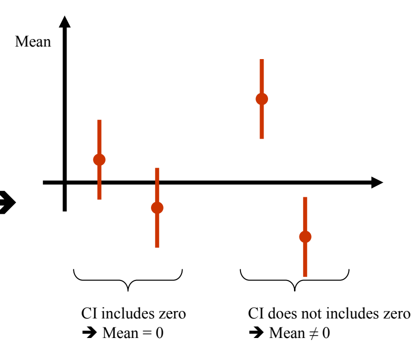
$$0 \pm (1.895)(0.138)/\sqrt{8}$$

$$(-0.0926, 0.0926)$$



## Testing For A SPECIFIC Mean Value

- **Is the population mean equal to a specific value θ?**
- **It depends on the confidence interval**
  - **If the confidence interval contains θ →** Yes
  - **If the confidence interval doe not contains θ →** No



## 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?



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

- **Solution**:

Sample size, $n = 7$
Sample mean, $\mu_s = 1.03$
Sample variance = 2.57 ➔ $s = 1.60$
Confidence interval = $1.03 \pm t \times 1.60 / \sqrt{7}$
= $1.03 \pm 0.605\, t$
$100(1-\alpha) = 99\%$ ➔ $\alpha = 0.01$ ➔ $1-\alpha/2 = 0.995$
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

## 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
  - ➔ 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))
  - Unpaird



## Comparing Two Alternatives – cont'd

- For two systems under similar work loads, the reading can be
  - Paired (i.e. follow the form (x,y))
  - Unpaird
- 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( )"

## 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?



## 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

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

- **<u>Code that can be used for solving examples in this section</u>:**

```
clear all
%
% Code for generating confidence intervals - can be used also for
% - testing for a specific mean value
% - comparing paired observations
ConfidenceLevel = 90;   % required confidence level
% put your samples here
Samples = [-13.7, 13.1, -2.8, -1.1, -3.0, 5.6];
n       = length(Samples);
Mue_s   = mean(Samples);
Sigma_s = sqrt(var(Samples));

p   = 1-(1 - ConfidenceLevel/100)/2;

fprintf('Sample size               = %3d\n', n);
fprintf('Sample mean               = %7.3f\n', Mue_s);
fprintf('Sample standard deviation = %7.3f\n', Sigma_s);
fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2 = %5.4f\n', ...
    ConfidenceLevel, 1- ConfidenceLevel/100, p);
```

Example_13_4.m



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

- **<u>Code that can be used for solving examples in this section – cont'd</u>:**

```
%
% check whether to use the normal quantile or the t-distribution
if (n>30)
    z_p = norminv(p, 0, 1);
    fprintf('confidence interval for mean = %7.3f +- zp * %7.3f / sqrt(%3d)\n', ...
        Mue_s, Sigma_s, n);
    fprintf('confidence interval for mean = %7.3f +- zp * (%7.3f)\n', ...
        Mue_s, Sigma_s/sqrt(n));
    fprintf('the %5.4f-quantile of the normal-variate z = %7.4f\n', p, z_p);
    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 is given by (%7.3f, %7.3f)\n', ...
         ConfidenceLevel, Mue_L, Mue_H);
else
    t_p = tinv(p, n-1);
    fprintf('confidence interval for mean = %7.3f +- tp * %7.3f / sqrt(%3d)\n', ...
        Mue_s, Sigma_s, n);
    fprintf('confidence interval for mean = %7.3f +- tp * (%7.3f)\n', ...
        Mue_s, Sigma_s/sqrt(n));
    fprintf('the %5.4f-quantile of the t-variate with %2d degrees of freedom t = %7.4f\n', ...
        p, n-1, t_p);
    Mue_L = Mue_s - t_p*Sigma_s/sqrt(n);
    Mue_H = Mue_s + t_p*Sigma_s/sqrt(n);
    fprintf('The %7.0f%% confidence interval is given by (%7.3f, %7.3f)\n', ...
         ConfidenceLevel, Mue_L, Mue_H);
end
if (Mue_L*Mue_H < 0)
    fprintf('Confidence interval (%7.2f, %7.2f) contains the zero\n', ...
         Mue_L, Mue_H);
 else
     fprintf('Confidence interval (%7.2f, %7.2f) does NOT contain the zero\n', ...
          Mue_L, Mue_H);
 end
```

## Unpaired Observations - Comparing Two Alternatives

- *t*-test – refer to textbook



## Example: Unpaired Observations - Comparing Two Alternatives

- <u>Problem</u>: 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?

# 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-a) =  90% ==> a = 0.100 and 1-a/2 = 0.9500
Mean difference Mue_A - Mue_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 +- tp *   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**



# 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**

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

- **Solution:**

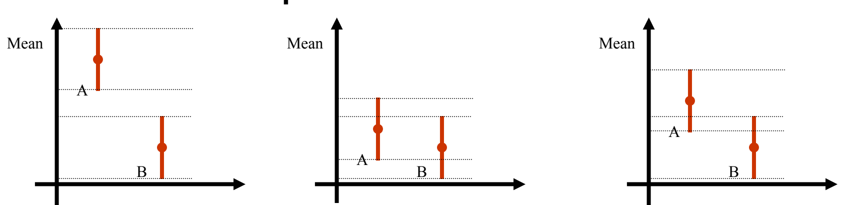
**Matlab code for performing the *t*-test: cont'd**

```
fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2 = %5.4f\n', ...
    ConfidenceLevel, 1- ConfidenceLevel/100, p);

fprintf('Mean difference Mue_A - Mue_B = %7.3f\n', Mean_Difference);
fprintf('Sigma for mean difference     = %7.3f\n', Sigma_Mean_Difference);
fprintf('Effective number of degrees of freedom, f = %7.3f (%3d)\n', ...
    Effective_number, Effective_number_rounded);

fprintf('confidence interval for mean  = %7.3f +- tp * %7.3f \n', ...
        Mean_Difference, Sigma_Mean_Difference);
fprintf('the %5.4f-quantile of the t-variate with %2d degrees of freedom t = %7.4f\n', ...
        p, Effective_number_rounded, t_p);
Mue_L = Mean_Difference - t_p*Sigma_Mean_Difference;
Mue_H = Mean_Difference + t_p*Sigma_Mean_Difference;
fprintf('The %7.0f%% confidence interval is given by (%7.3f, %7.3f)\n', ...
         ConfidenceLevel, Mue_L, Mue_H);

if (Mue_L*Mue_H < 0)
    fprintf('The two systems ARE NOT significantly different\n', ...
         Mue_L, Mue_H);
 else
     fprintf('The two systems ARE significantly different\n', ...
          Mue_L, Mue_H);
 end
```



# 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**




CIs do not overlap
→ A is higher than B

CIs overlap and mean of one is in the CI of the other
→ A and B are NOT significantly different

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-α)%
  - There is a 100α/2% chance the sample will be more than the upper confidence limit
  - There is a 100α/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

## 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 ≥ 10 (Binomial distribution ≈ Normal distribution)**
  - **If condition is not satisfied – can not use *t*-test**
  - **Procedure not defined at this stage**



## Example: Confidence Interval for Proportions

- **<u>Problem</u>: 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**

## 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)
```



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

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

```
clear all
%
% Code for generating confidence intervals for proportions
ConfidenceLevel = 90;   % required confidence level
% put your samples here
n1 = 10;
n  = 1000;
p  = n1/n;

a   = 1-(1 - ConfidenceLevel/100)/2;

fprintf('Sample proportion    = %7.3f\n', p);
if (p*n<10)
    fprintf('n*p = %7.3f >= 10 is not satisfied\n',n*p);
else
    z_a = norminv(a,0,1);
    fprintf('n*p = %7.3f >= 10 is  satisfied\n',n*p);
    fprintf('Confidence level 100(1-a) = %3.0f%% ==> a = %4.3f and 1-a/2 = %5.4f\n', ...
        ConfidenceLevel, 1- ConfidenceLevel/100, a);
    fprintf('confidence interval for proportion = %7.4f +- za * sqrt(%7.3f * %7.3f / %4d)\n', ...
        p, p, 1-p, n);
    fprintf('confidence interval for proportion = %7.4f +- za * %7.3f\n', ...
        p, sqrt(p*(1-p)/n));
    fprintf('the %5.4f-quantile of the normal-variate z = %7.4f\n', a, z_a);
    p_L = p - z_a*sqrt(p*(1-p)/n);
    p_H = p + z_a*sqrt(p*(1-p)/n);
    fprintf('The %7.0f%% confidence interval is given by (%5.4f, %5.4f)\n', ...
         ConfidenceLevel, p_L, p_H);
end
```

## 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?



## 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 99% 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

## 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?



## 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 $\mu_s \pm z \text{ X } s / \sqrt{n} = \mu_s (1 \pm r/100)$
- Therefore, $n = (100\, z\, s / (r\, \mu_s))$^2

## Example: Determining Sample Size

- **<u>Problem</u>: 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?**



## Example: Determining Sample Size – cont'd

- **<u>Solution</u>:**

**Required accuracy = 1 in 20 = 5%**
**$\mu_s$ = 20, s = 5, r = 5%**
**Confidence level 100(1-a) = 95% ==> a = 0.05 and 1-a/2 = 0.9750**
**$z_{0.95}$ = 1.960**

**Therefore, required repetitions**

```
       (100)(1.960)(5)
n = (------------------)^2 = 9.8^2 = 96.04
            (5) (20)
```

**A total of 97 observations are required.**

## 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,

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

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



## Example: Determining Sample Size for Proportions

- <u>Problem</u>: 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?

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

- <u>**Solution**</u>:

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

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

$$= 384{,}160{,}000$$

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



## 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**

## 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?



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

- **Solution**:

CI for algorithm A = 0.005 ± 1.960 (0.005(1-0.005)/n)^2
CI for algorithm B = 0.006 ± 1.960 (0.006(1-0.006)/n)^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)^2 < 0.006 - 1.960 (0.006(1-0.006)/n)^2

➔ n > 84340

We need to observe 85,000 packets

## Some Important Random Variables – Discrete Random Variables

- **Bernoulli**
- **Binomial**
- **Geometric**
- **Poisson**

**Identities to remember:**

$$\sum_{n=1}^{M} n = \frac{1}{2} M(M+1) \qquad \sum_{n=1}^{M} n^2 = M(M+1)(2M+1)/6 \qquad \sum_{n=0}^{\infty} r^n = \frac{1}{1-r}; |r| < 1$$

$$\sum_{n=0}^{\infty} n r^{n-1} = \frac{1}{(1-r)^2}; |r| < 1 \qquad \sum_{n=0}^{M} r^n = \frac{1-r^{M+1}}{1-r}; |r| < 1, M = 1,2,\ldots \qquad \sum_{n=0}^{M} \binom{M}{n} r^n = (1+r)^M; |r| < 1$$

$$\sum_{n=0}^{M} n r^{n-1} = \frac{1+(Mr-M-1)r^M}{(1-r)^2}; |r| < 1$$



## Bernoulli Random Variable

- **Let A be an event related to the outcomes of some random experiment. The indicator function for A is defined as**

$$I_A(\zeta) = 0 \quad \text{if } \zeta \text{ not in A}$$
$$= 1 \quad \text{if } \zeta \text{ is in A}$$

- **$I_A$ is random variable since it assigns a number to each outcome in S**
- **It is discrete r.v. that takes on values from the set {0,1}**
- **PMF is given by**

$$p_I(0) = 1\text{-}p, \; p_I(1) = p$$

**where P(A) = p**

- **Describes the outcome of a Bernoulli trial**

- **E[X] = p, VAR[X] = p(1-p)**

## Binomial Random Variable

- Suppose a random experiment is repeated n independent times; let X be the number of times a certain event A occurs in these n trials

  X = I1 + I2 + … + In

i.e. X is the sum of Bernoulli trials (X's range = {0, 1, 2, …, n})

- X has the following pmf

$$P[X=k]=\binom{n}{k}p^k(1-p)^{n-k}$$

for k=0, 1, 2, …, n

- E[X] = np, Var[X] = np(1-p)



## Geometric Random Variable

- Suppose a random experiment is repeated - We count the number of M of independent Bernoulli trials until the first occurrence of a success
- M is called geometric random variable
  - Range of M = 1, 2, 3, …
- X has the following pmf

$$\Pr[X=k]=(1-p)^{k-1}p$$

for k=1, 2, 3, …

- E[X] = 1/p, Var[X] = (1-p)/p$^2$

## Geometric Random Variable - 2

- Suppose a random experiment is repeated - We count the number of M of independent Bernoulli trials until the first occurrence of a success – not counting the successful trial
- M is called geometric random variable
  - Range of M = 0, 1, 2, 3, …
- X has the following pmf

$$\Pr[X = k] = (1-p)^k p$$

for k=0, 1, 2, 3, …

Note the different range for these two Geometric r.v.s

- E[X] = (1-p)/p, Var[X] = (1-p)/p²



## Poisson Random Variable

- In many applications we are interested in counting the number of occurrences of an event in a certain time period
- The pmf is given by $\Pr[X = k] = \frac{\alpha^k}{k!} e^{-\alpha}$

For k=0, 1, 2, … ;
α is the average number of event occurrences in the specified interval

- E[X] = α, Var[X] = α
- Poisson is the limiting case for Binomial as n→∞, p →0, such that np = α – remember $\lim_{n\to\infty, p\to 0} (1-\lambda/n)^n = e^{-\lambda}$

## Poisson Random Variable - 2

- **If the average rate of occurrence per time unit is λ, then the average number of occurrences in t seconds is equal to λt**
- **The probability of k occurrences in t seconds is given by**

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad k = 0,1,2,\ldots$$

Compared to previous slides – we have replaced α by λt



## Some Important Random Variables – Continuous Random Variables

- **Uniform**
- **Exponential**
- **Gaussian (Normal)**
- **Rayleigh**
- **Gamma**
- **Pareto**

## Uniform Random Variables

- Realizations of the r.v. can take values from the interval [a, b]
- PDF $f_X(x) = 1/(b-a)$  $a \le x \le b$
- $E[X] = (a+b)/2$, $Var[X] = (b-a)^2/12$



## Example 5: Analog-to-Digital Conversion

Problem: compute the SNR for a uniform quantizer using $2^N$ representation values?

## Exponential Random Variables

- **The exponential r.v. X with parameter λ has pdf**

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \ge 0 \end{cases}$$

- **And CDF given by**

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \ge 0 \end{cases}$$

- **Range of X: [0, ∞)**

- **$E[X] = 1/\lambda$, $Var[X] = 1/\lambda^2$**



## Exponential Random Variables – cont'd

- **The exponential r.v. is the only r.v. with the memoryless property!!**
- **Memoryless Property:**

$$P[X>t+h / X>t] = P[X>h]$$

**Proof:**

$$P[X>t+h / X>t] = \frac{P[(X > t+h) \cap (X > t)]}{P[(X > t)]}$$

$$= \frac{P[(X > t+h)}{P[X > t]} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}}$$

$$= e^{-\lambda h}$$

$$= P[X > h]$$

## Gaussian (Normal) Random Variable

- Rises in situations where a random variable X is the sum of a large number of "small" random variables – central limit theorem
- PDF $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

For $-\infty < x < \infty$; $\mu$ and $\sigma > 0$ are real numbers

- The characteristic function is given by

$$\Phi_X(\omega) = e^{j\mu\omega - \sigma^2\omega^2/2}$$

- E[X] = $\mu$, Var[X] = $\sigma^2$



## Gaussian (Normal) Random Variable - 2

- CDF given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/(2\sigma^2)} dt$$
$$= 0.5 + 0.5 erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$$

where

$$erf(x) = \int_{0}^{x} e^{-t^2/2} dt$$

Note – the CDF can also be written in terms of the Q-function, where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2/2} dt$$

## Rayleigh Random Variable

- **Rises in modeling of mobile channels**
- **Range: [0, $\infty$)**
- **PDF:** $$f_X(x) = \frac{x}{\alpha^2} e^{-x^2/(2\alpha^2)}$$
- **For $x \geq 0$, $\alpha > 0$**
- **$E[X] = \alpha\sqrt{(\pi/2)}$, $Var[X] = (2-\pi/2)\alpha^2$**



## Gamma Random Variable

- **Versatile distribution ~ appears in modeling of lifetime of devices and systems**
- **Has two parameters: $\alpha > 0$ and $\lambda > 0$**
- **PDF:** $$f_X(x) = \frac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$
- **For $0 < x < \infty$**
- **The quantity $\Gamma(z)$ is the gamma function and is specified by**
$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$
- **The gamma function has the following properties:**
  - **$\Gamma(1/2) = \sqrt{\pi}$**
  - **$\Gamma(z+1) = z\Gamma(z)$ for $z>0$**
  - **$\Gamma(m+1) = m!$ For m nonnegative integer**
- **$E[X] = \alpha/\lambda$, $Var[X] = \alpha/\lambda^2$**
- **$\Phi_X(\omega) = 1/(1-j\omega/\lambda)^\alpha$**

If $\alpha = 1$ ➔ gamma r.v. becomes exponential

## Pareto Random Variable

- Originally used by economists to model income and other soci-economic quantities.
- For α (shape parameter) > 0, β (scale parameter) > 0, the PDF is given by

$$f_X(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \qquad \beta \le x$$

- The CDF is given by

$$F_X(x) = 1 - \left(\frac{\beta}{x}\right)^\alpha \qquad \beta \le x$$



## Pareto Random Variable - 2

- $n^{th}$ moment (if it exists) is given by

$$E\left[x^n\right] = \frac{\alpha\beta^n}{\alpha - n} \qquad n < \alpha$$

- Expected value: $E[x] = \frac{\alpha\beta}{\alpha - 1} \qquad 1 < \alpha$
- Variance: $Var[x] = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \qquad 2 < \alpha$

## Example 6: Packet Size Modeling

- **Pareto distribution is used to model the packet size, P, in bytes for internet traffic as follows:** $P = \min(x, S_{\max})$

**where x is a Pareto random variable with the following PDF**

$$f_X(x) = \begin{cases} \dfrac{\alpha\beta^{\alpha}}{x^{\alpha+1}} & \beta \le x < S_{\max} \\ \theta & x = S_{\max} \end{cases}$$

**θ is given by** $\theta = 1 - F_X(S_{\max})$



## Example 7: Packet Size Modeling

- **Calculate the expected value for packet size using the model proposed in Example 3?**

- **Models proposed to test ETSI/UMTS networks use the following parameters: α = 1.1, β = 81.5 Bytes, Smax = 66,666 Byte (this results in a mean packet size of 480 Bytes)**