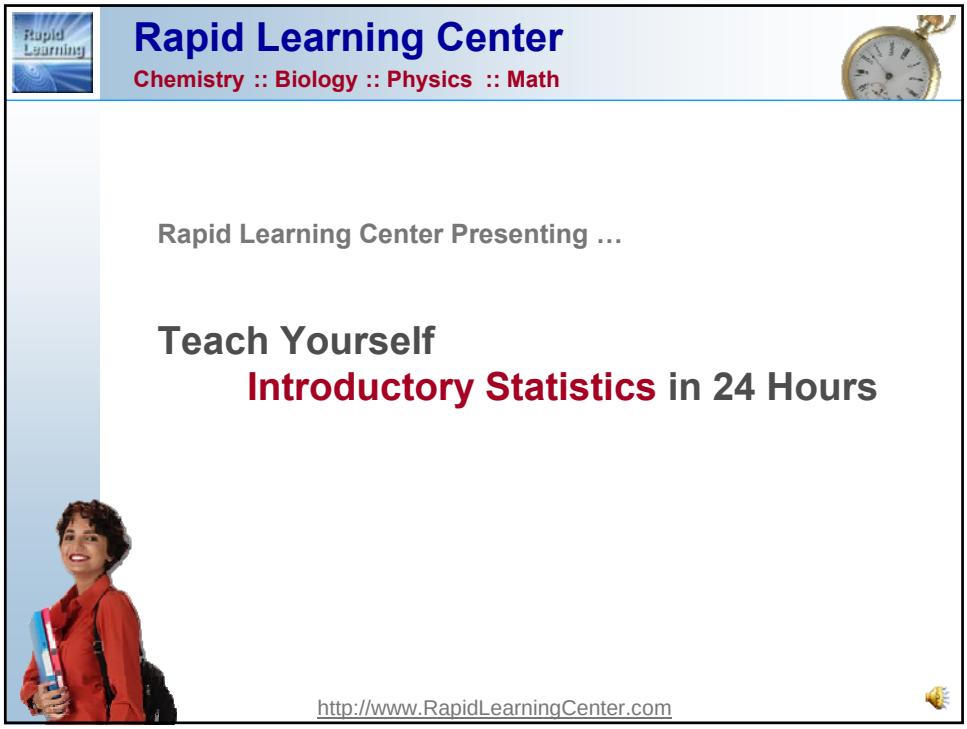
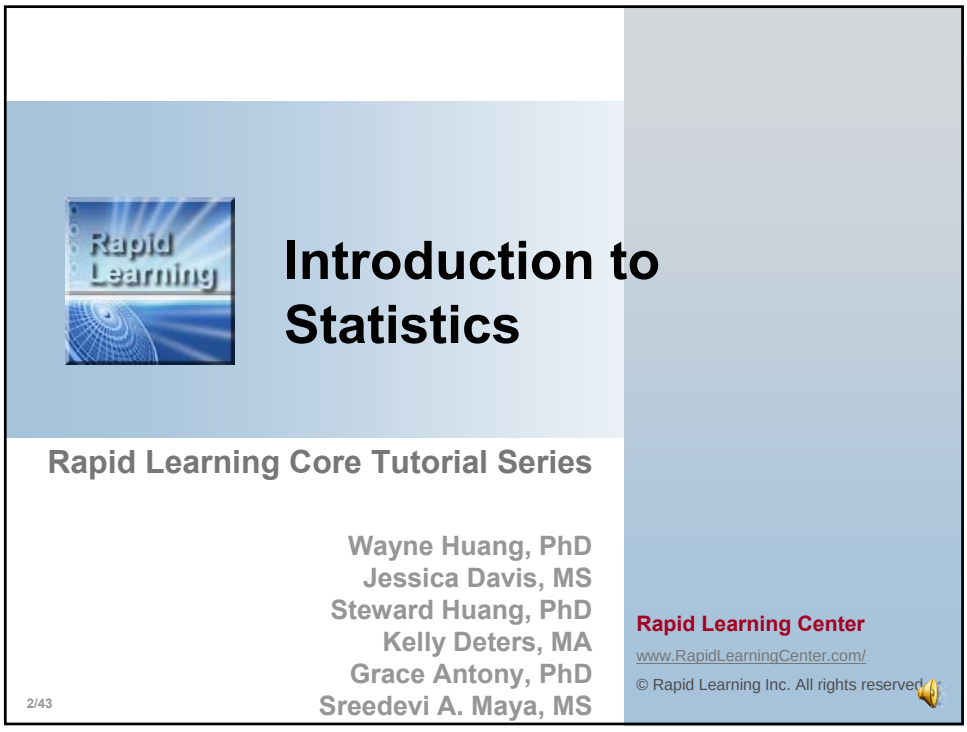




© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

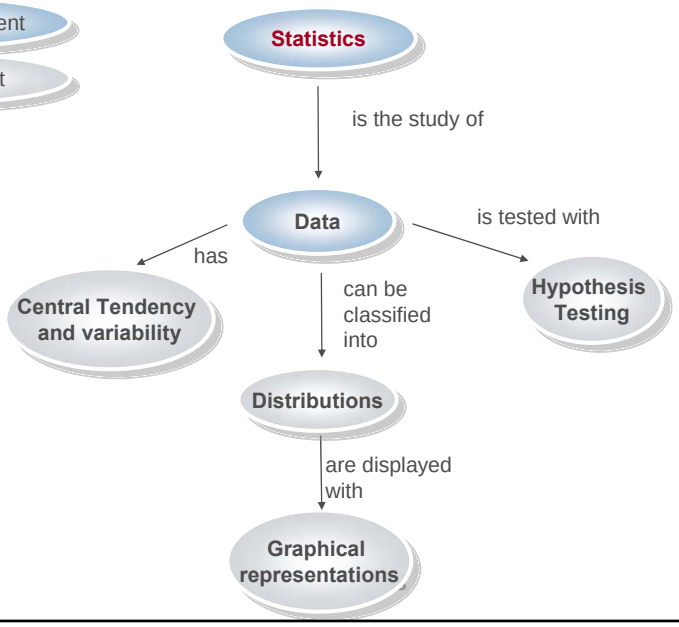
Learning Objectives

- **By completing this tutorial, you will learn how to:**
  - ✓ **Define Statistics**
  - ✓ **Study Statistics**
  - ✓ **Identify measures of central tendency and variability**
  - ✓ **Calculate probability**
  - ✓ **Understand hypothesis testing**
  - ✓ **Use the Central Limit Theorem**
  - ✓ **Identify probability distributions**

Introduction to Statistics

Previous content

New content



4/43

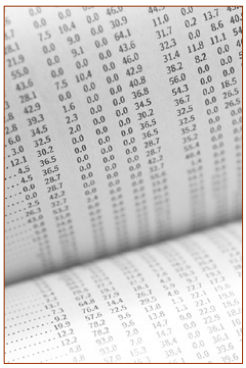
© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

What is Statistics?

Statistics is a branch of mathematics that deals with the effective management and analysis of data.

In statistics:

- **Data is made more manageable and presented in a logical form**
- **Patterns can be seen from organized data:**
  - **In Frequency tables**
  - **Using Graphical techniques**
  - **Measuring Central Tendency**
  - **Measuring Spread (variability)**



5/43

Study Tips – Learning Statistics

The following study tips will help you learn the material presented in this course:

✓ **Read the introduction and objectives for each lesson in this course guide.**

✓ **Have a pencil ready and work through the examples.**

✓ **Make sure you understand the steps to each solution**

✓ **Work problems regularly (several days per week) to help you master the concepts.**

6/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Prerequisite Review

Many basic math concepts are used in statistics.

Here are some of the key math concepts (from intermediate algebra) used in this course:

- ✓ Numbers, Equations and Inequalities
- ✓ Functions and Their Graphs
- ✓ Exponential and Logarithmic Functions
- ✓ The Trigonometric Functions
- ✓ Sequences, Counting and Probability

7/43

Why Statistics?

- ✓ To develop an appreciation for variability and how it effects products and processes.
- ✓ Study methods that can be used to help solve problems, build knowledge and continuously improve products and processes.
- ✓ Build an appreciation for the advantages and limitations of informed observation and experimentation.
- ✓ Determine how to analyze data from designed experiments in order to build knowledge and continuously improve.
- ✓Develop an understanding of some basic ideas of statistical reliability and the analysis data.

8/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Problem Solving Methods

What is a problem?

A problem is a question that motivates you to search for a solution.

What is problem solving?

Finding a solution to a problem by developing an understanding of the problem through the creation and/or manipulation of processes and concepts.

- ✓ Understand and explore the problem;
- ✓ Find a strategy;
- ✓ Use the strategy to solve the problem;
- ✓ Look back and reflect on the solution.

9/43

Problem Solving Strategies

Problem solving strategies:

- ❖ Split problems into parts.
- ❖ Analyze the given values
- ❖ Draw (this includes drawing pictures and diagrams)
- ❖ Make a List (this includes making a table)
- ❖ Think (this means using skills you already know)
- ❖ Think about the statistical methods that are used to solve the problem.
- ❖ Analyze the efficiency of the result.

10/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

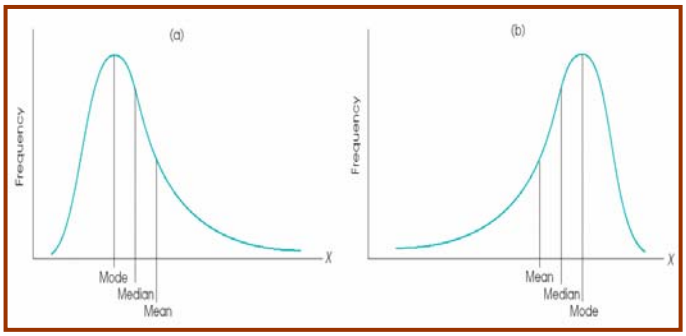
Measures of Central Tendency

Measures of central tendency are measures of the location of the middle or the center of a distribution.

The mean is the most commonly used measure of central tendency.

The following are measures of central tendency:

- Mean
- Median
- Mode
- Midrange



11/43

Mean

The mean is the sum of scores divided by the number of observations.

$$\mu = \frac{\sum x}{N}$$

N is the number of observations

Example: Set of numbers 2, 2, 3, 5, 5, 7, 8

2+2+3+5+5+7+8 = 32

There are 7 Values. So you divide the total by 7

32/7 = 4.57… So the mean is 4.57

12/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Median

The data must be ranked (sorted in ascending order) first. The median is the number in the middle.

To find the median, put the values in order, then find the middle value. If there are two values in the middle then find the average of these two values.

Example Set : 2, 2, 3, 5, 5, 7, 8

The numbers in order: 2 , 2 , 3 , (5) , 5 , 7 , 8

The middle value is marked in parentheses, and it is 5.

So the median is 5

13/43

Mode

The mode of a set of data is the value in the set that occurs most often.

Problem:

The number of points scored in a series of football games is listed below. Which score is the mode?

7, 13, 18, 24, 9, 3, 18

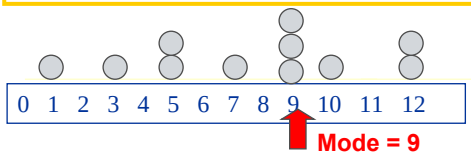
Solution:

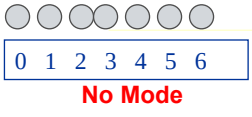
Ordering the scores from least to greatest, we get:

3, 7, 9, 13, 18, 18, 24

Answer:

The score which occurs most often is 18.





14/43

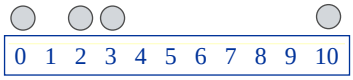
© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Midrange

The midrange is simply the midpoint between the highest and lowest values.

Midrange $= \dfrac{x_{largest} + x_{smallest}}{2}$

Example:



Midrange = 5

15/43

Measures of Variability

Variability describes in an exact quantitative measure how spread out/clustered together the scores are.

- Range
- Variance
- Standard Deviation

- These two distributions have the same symmetrical shape.
- They have the same mean value, not the same variability.
- Say these are graphs showing IQ from two different samples of people.
- In the left graph the spread of the scores is much smaller than the right graph.

16/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Range

The range of a set of data is the difference between the highest and lowest values in the set.

Range = χ*Highest* - χ*Lowest*

Problem:

Cheryl took 7 math tests in one marking period. What is the range of her test scores?
89, 73, 84, 91, 87, 77, 94

Solution:

Ordering the test scores from least to greatest, we get:

73, 77, 84, 87, 89, 91, 94

highest - lowest = 94 - 73 = 21

Answer:

The range of these test scores is 21 points.

17/43

Variance

The variance of a sample measures how the observations are spread around the mean. Large variance means the score is widely spread around the mean.

- **Population variance is designated by σ^2**

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

- **Sample Variance is designated by s^2**
  - Samples are less variable than populations: they therefore give biased estimates of population variability
  - Degrees of Freedom (*df*): the number of parameters that may be independently varied.
  - In a sample, the sample mean must be known before the variance can be calculated, therefore the final score is dependent on earlier scores. The formula is:

$$s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$$

18/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Standard Deviation

The most common measure of variability is the standard deviation or the square root of the variance.

Population (σ) and Sample (s) standard deviations:

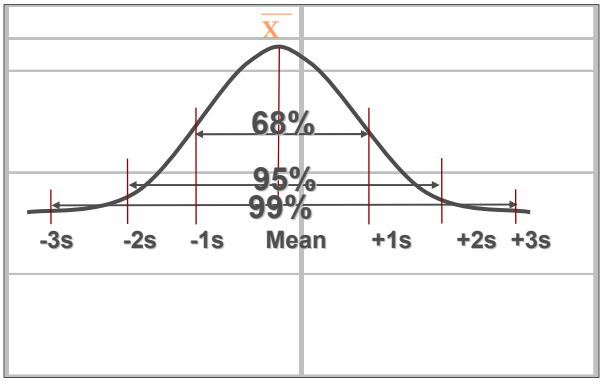
$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \qquad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

A good measure of variability must:

- **Be stable and reliable; not greatly affected by certain details in the data such as:**
  - **Extreme scores**
  - **Multiple sampling from the same population**
  - **Open-ended distributions**
- **Both the variance and SD are related to other statistical techniques**

19/43

The Empirical Rule



For a data set of normal distribution, a value will fall within a range of:

+/- 1 SD 68% of the time

+/- 2 SD 95% of the time

+/- 3 SD 99% of the time

20/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Probability

Probability of an event: A probability that provides a quantitative description of the likely occurrence of a particular event.

$$P(E) = \frac{\text{number of outcomes corresponding to event E}}{\text{total number of outcomes}}$$

Example

The probability of drawing a spade from a pack of 52 well-shuffled playing cards is:

$$\frac{13}{52} = \frac{1}{4} = 0.25$$

21/43

Conditional Probability

The probability that event B occurs, given that event A has already occurred is:

$$P(B|A) = P(A \text{ and } B) / P(A)$$

Example :
The question, "Do you smoke?" was asked of 100 people.

Results are shown in the table.

	Yes	No	Total
Male	19	41	60
Female	12	28	40
Total	31	69	100

What is the probability of a randomly selected individual being a male who smokes? This is a joint probability. The number of "Male and Smoke" divided by the total = 19/100 = 0.19

22/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Random Variable

A random variable is a function that associates a unique numerical value with every outcome of an experiment.

There are two types of random variable:
- **Discrete**
- **Continuous**

- **Discrete**: A coin is tossed ten times.
  - The random variable X is the number of tails that are noted.
  - X can only take the values 0, 1, ..., 10, so X is a discrete random variable.
- **Continuous**: A light bulb is burned until it burns out.
  - The random variable Y is its lifetime in hours.
  - Y can take any positive real value (even decimals), so Y is a continuous random variable.

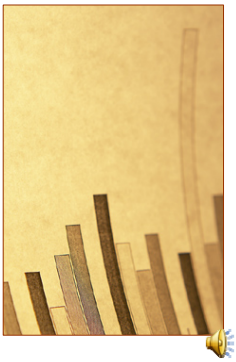
23/43

Frequency Distribution and Graph

A set of scores arranged in order of magnitude along the x-axis with the frequency of each score along the y-axis.

Use Graphs: To illustrate relative amounts
To specify the subject
To answer specific questions

Here are some commonly used graphs:
- Categorical Frequency Distribution
- Histogram
- Bar Chart
- Frequency Polygon
- Stem-and-Leaf plot



24/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Categorical Frequency Distributions

Categorical frequency distributions - can be used for data that can be placed in specific categories, such as nominal- or ordinal-level data.

Examples - political affiliation, religious affiliation, blood type etc.

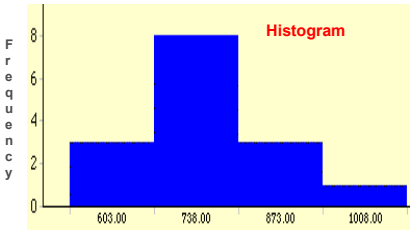
Class	Frequency	Percent
A	5	20
B	7	28
O	9	36
AB	4	16

Blood Type frequency distribution example

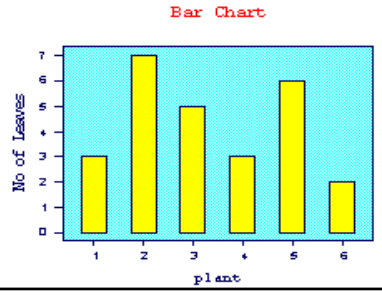
25/43

Histogram & Bar Chart

- Maintained to approximate the distribution of data according to numerical attributes.
- Constructed by partitioning the data into mutually disjoint subsets.
- Frequency is recorded on the y axis and the data intervals on the x axis.



Bar charts can be displayed horizontally or vertically.



26/43

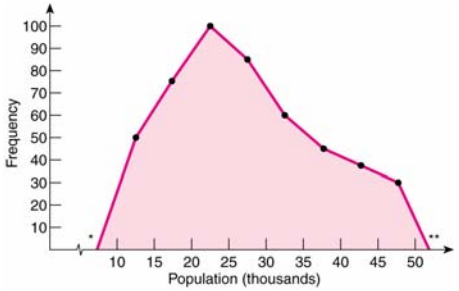
© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Frequency Polygon

A frequency polygon is a graph that represents the shape of the data.

It can be conceptualized as a connection of the midpoints of the classes at the height specified by the frequency.

A relative frequency polygon is similar to a frequency polygon, except that the height is dictated by the relative frequency.



27/43

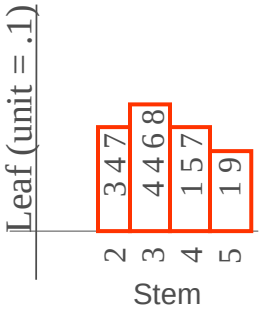
Stem-and-Leaf Plot

Stem-and-Leaf Plots were developed to summarize data without loss of information. The stem is every digit except the last, the leaf represents the last digit.

Reports of the after-tax profits of 12 companies are (recorded as cents per dollar of revenue) as follows:

3.4, 4.5, 2.3, 2.7, 3.8, 5.9, 3.4, 4.7, 2.4, 4.1, 3.6, 5.1

Stem	Leaf (unit = .1)
2	3 4 7
3	4 4 6 8
4	1 5 7
5	1 9



28/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Probability Distribution

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.

- The probability distribution is defined by a probability function, denoted by $f(x)$, which provides the probability for each value of the random variable.
- The required conditions for a discrete probability function are:

  $f(x) \geq 0$

  $f(x) = 1$
- We can describe a discrete probability distribution with a table, graph, or equation.

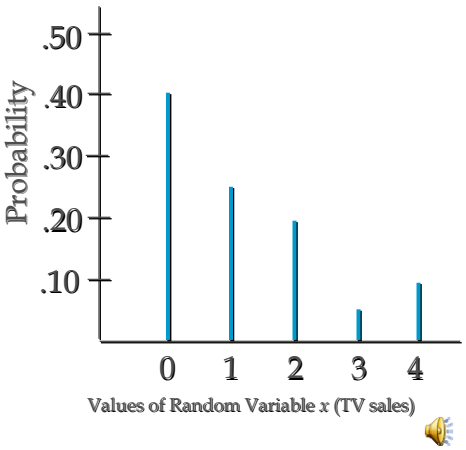
29/43

Probability Distribution Graph

Using data on TV sales (below left), a tabular representation of the probability distribution for TV sales (below right) was developed.

Units Sold	Number of Days
0	80
1	50
2	40
3	10
4	20
	200

x	$f(x)$
0	.40
1	.25
2	.20
3	.05
4	.10
	1.00



30/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Testing Hypotheses

Hypothesis testing "tests" whether the data supports the claim (hypothesis) or not.

The critical concepts of hypothesis testing:

- ✓ H_0 - the null hypothesis
  The statement of "no effect" or "no difference".
- ✓ *Ha* - the alternative hypothesis
  The statement we hope or suspect is true.

Example: Spin a coin 250 times
p: probability of getting a head during each spin

H_0: $p = .5$ against *Ha*: $p > .5$.
One-sided

H_0: $p = .5$ against *Ha*: $p \neq .5$.
Two-sided

31/43

Alternative and Null Hypothesis

A Mechanic is considering replacing his old equipment with new equipment.

- μ_0 is the average weekly maintenance cost of one of the old machines.
- μ is the average weekly maintenance cost he can expect for one of the new ones
- **We want to test the null hypothesis $\mu = \mu_0$.**

He will purchase the new equipment if it will reduce his average weekly maintenance cost. That is: $\mu < \mu_0$.

This is called a **one-sided alternative**, using inequalities <, >, ≤, and ≥.

He just wants to find out if the price of the new equipment is different (higher or lower than) the old equipment. That is: $\mu \neq \mu_0$.

This is called a **two-sided alternative**, using ≠.

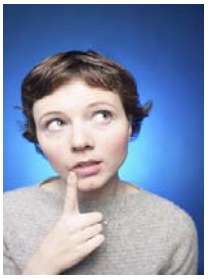
32/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Hypothesis Testing: Forms and Errors

Null and alternative hypotheses can take the following forms:

Null	Possible Alternatives
$\mu = \mu_0$	$\mu \neq \mu_0$, $\mu < \mu_0$, $\mu > \mu_0$
$\mu \geq \mu_0$	$\mu < \mu_0$
$\mu \leq \mu_0$	$\mu > \mu_0$



Now we are going to either reject the null hypothesis or not. It is important to realize that we can make two types of errors in rejecting the null hypothesis.

Type I error

Type II error

33/43

Type I and II Error

Type I error is rejecting the null hypothesis when it is true.

Type II error is not rejecting the null hypothesis when it is false.

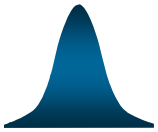
Truth / Conclusion	H_0 true (Do not reject H_0)	H_0 false (Reject H_0)
H_0 true (Do not reject H_0)	○ (correct)	✕ (Type II Error)
H_0 false (Reject H_0)	✕ (Type I error)	○ (correct)

34/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Normal Distribution

A **Normal Distribution** is:



- Single-peaked
- Bell-shaped
- Tails fall off quickly
- The mean, median, and mode are the same
- The points where there is a change in curvature are one standard deviation on either side of the mean.
- The mean and standard deviation completely specify the curve

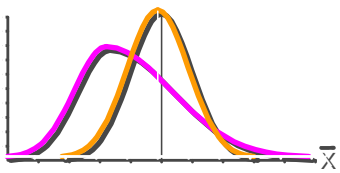
35/43

Central Limit Theorem

As the sample size increases the sampling distribution of the sample mean approaches the normal distribution with mean μ (0) and variance σ^2/n (1).

Note: As the sample size gets larger (n > 30), the sampling distribution becomes almost Normal regardless of the shape of the population.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$



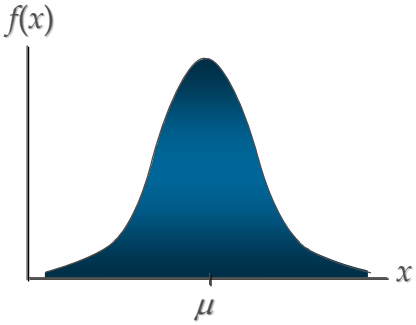
36/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Continuous Probability Distribution

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

- Uniform Probability Distribution
- Normal Probability Distribution
- Exponential Probability Distribution



37/43

Uniform Probability Distribution

- A random variable is uniformly distributed whenever the probability is proportional to the interval's length.
- Uniform Probability Density Function

$$f(x) = 1/(b - a) \quad \text{for } a \le x \le b$$

$$= 0 \quad \text{elsewhere}$$

where: a = smallest value the variable can assume

b = largest value the variable can assume

Expected Value of x:	$E(x) = (a + b)/2$
Variance of x:	$Var(x) = (n^2 - 1)/12$

38/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

Normal Probability Distribution

- The **normal probability distribution** is the most important distribution for describing a continuous random variable.
- It has been used in a wide variety of applications:
  - Heights and weights of people
  - Test scores
  - Scientific measurements
  - Amounts of rainfall
- It is widely used in statistical inference

Normal Probability Density Function

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

μ is Mean, σ is Standard Deviation
π is 3.14.., e is 2.718

39/43

Exponential Probability Distribution

- The **exponential probability distribution** is appropriate for modeling time between events at an average rate.
- The exponential random variables can be used to describe:
  - Time between vehicle arrivals at a toll booth
  - Time required to complete a questionnaire
  - Distance between major defects in a highway

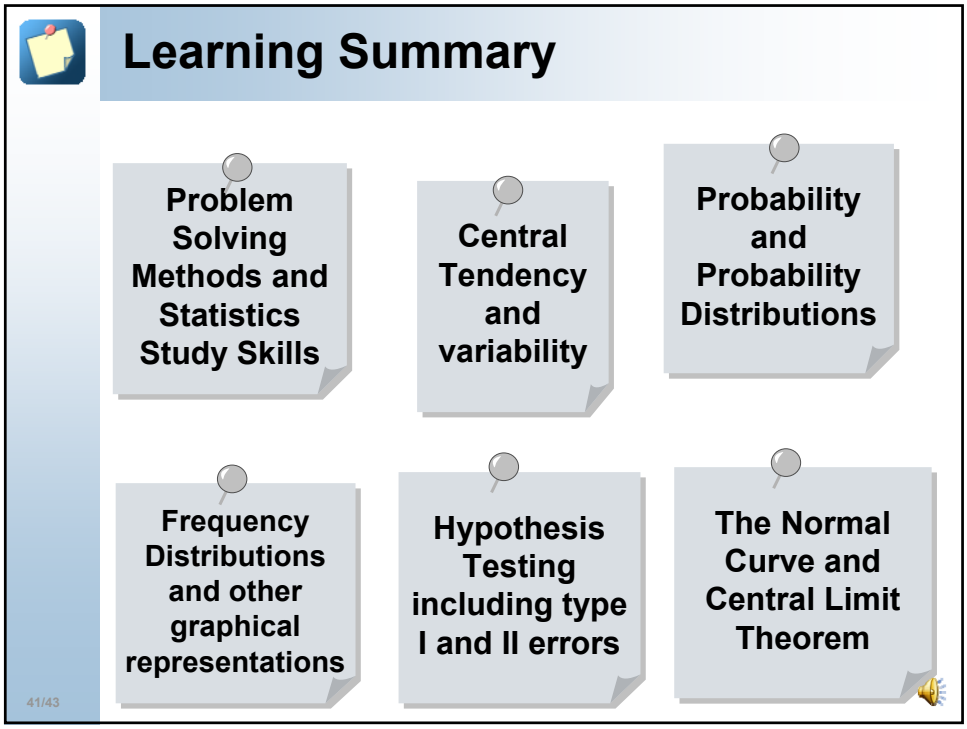
Exponential Probability Distribution Function:

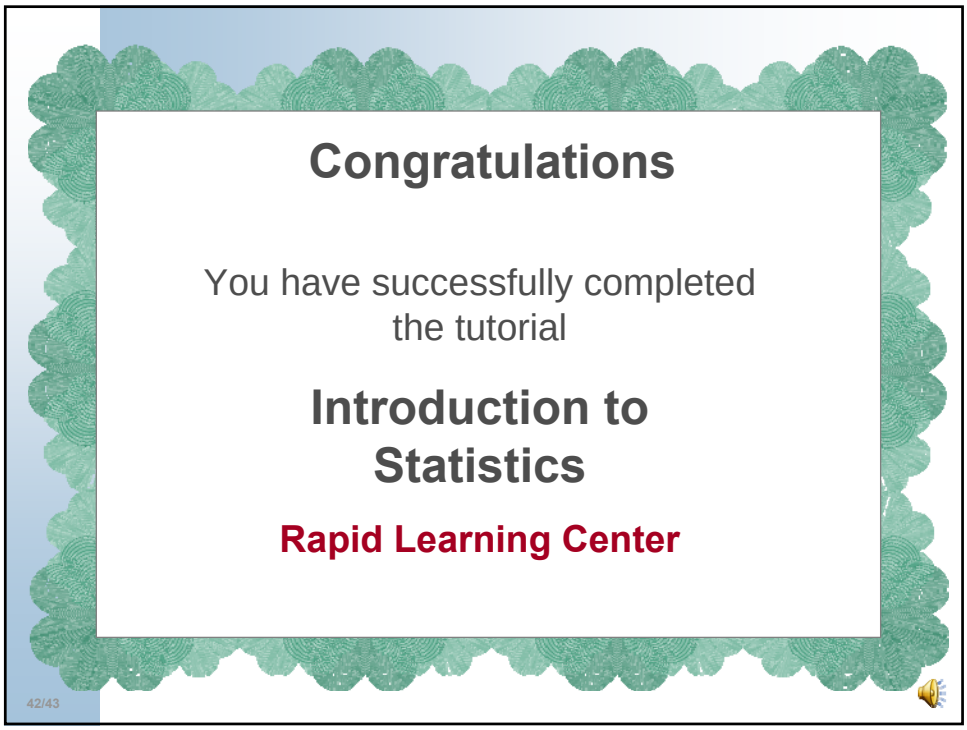
$$P(x) = \lambda e^{-\lambda x}$$

where λ is the rate of change

40/43

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com





© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com

© Rapid Learning Inc. All rights reserved. :: http://www.RapidLearningCenter.com