

Lec 1 - Basic of Excel.pptx

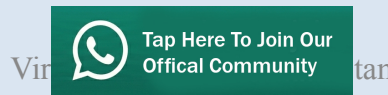
Software's

- MS EXCEL 2010
- SPSS 19

PDF version

By [Hanzla Ahmed](#)

WU TOOLKIT



Vir tan

Lecture 1

WU TOOLKIT

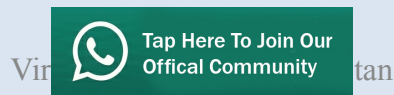


Tap Here To Join Our
Official Community

Pakistan

Contents

- Starting MS EXCEL
- Opening a file,
- Naming and defining variables,
- Data display,
- Entering data,
- Help system,
- tables ,
- Saving files in Excel ,
- Leaving Excel,
- Practical exercise in Data Preparation.



Introduction to Microsoft Excel

Objectives:

- To define spreadsheets and explain basic functionality
- To introduce the basic features of Excel
 - Vocabulary
 - Entering Data
 - Formatting Data
 - Precision vs. Display
 - Operators & Order of Precedence



Spreadsheet: Electronic sheet of paper organized by columns & rows

The advantage of an electronic spreadsheet is it allows you to easily change data and have all “related” calculations automatically update..

	A	B	C	D	E
1	Food Item	price per package	#pkgs	total	% total
2	cereal	2.50	2.0	\$ 5.00	24%
3	milk	1.50	1.0	\$ 1.50	7%
4	eggs	1.00	1.0	\$ 1.00	5%
5	cheese	3.50	2.0	\$ 7.00	33%
6	meat	3.75	1.0	\$ 3.75	18%
7	pasta	1.00	3.0	\$ 3.00	14%
8	totals	13.25	10.0	\$21.25	100%



Tap Here To Join Our
Official Community

Spreadsheets in Excel are referred to as *worksheets*.


A *workbook* file may contain many worksheets.

The screenshot shows the Microsoft Excel interface with the following components labeled:

- Quick Access Toolbar:** Located at the top left, containing Save, Undo, and Redo icons.
- Ribbon Tabs:** The tabs at the top of the ribbon, including File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Team.
- Home Ribbon:** The active ribbon, containing groups like Clipboard, Font, Alignment, Number, Styles, Cells, and Editing.
- Help Button:** A question mark icon in the top right corner.
- Sizing Buttons:** Buttons for window management (restore down, maximize, close) in the top right.
- Name Box:** Shows the active cell address, D2.
- Formula Bar:** Shows the formula being entered, $=B2*C2$.
- Column Letter Headings:** Letters A through I above the columns.
- Row Numbers:** Numbers 1 through 17 to the left of the rows.
- Contents of Active Cell displayed on Formula Bar:** The value \$5.00 from cell C2 is circled in red and pointed to the formula bar.
- Scroll Bars:** Horizontal and vertical bars at the bottom right for navigating through the spreadsheet.
- Sheet Tabs:** Tabs for Sheet1, Sheet2, and Sheet3 at the bottom.
- View Buttons:** Buttons for gridlines, gridlines and formulas, and gridlines and formulas hidden.
- Zoom:** A slider and percentage indicator (120%) for zooming in and out.
- Insert Worksheet Button:** A plus sign icon next to the sheet tabs.

	A	B	C	D	E	F	G	H	I
1	Food Item	price per package	#pkgs	total	% total				
2	cereal	2.50	2.0	\$ 5.00	24%				
3	milk	1.50	1.0	\$ 1.50	7%				
4	eggs	1.00	1.0	\$ 1.00	5%				
5	cheese	3.50	2.0	\$ 7.00	33%				
6	meat	3.75	1.0	\$ 3.75	18%				
7	pasta	1.00	3.0	\$ 3.00	14%				
8	totals	13.25	10.0	\$ 21.25	100%				
9									
10									
11	price per breakfast	\$ 1.63							
12									
13	price per lunch	\$ 3.75							
14									
15	Price per dinner	\$ 5.63							
16									
17	Available funds	\$ 14.00							



Vir  Tap Here To Join Our Official Community tan

Each box is referred to as a “*cell*”. Cells may contain *Labels, Values* or *Formulas* that result in a value or label. A cell is identified first by its column letter and then by its row number

Columns
Cell
D2
Contains
the
Formula
=
B2*C2

	A	B	C	D
1	Food Item	price per package	#pkgs	total
2	cereal	2.50	2.0	\$ 5.00
3	milk	1.50	1.0	\$ 1.50
4	eggs	1.00	1.0	\$ 1.00
5	cheese	3.50	2.0	\$ 7.00
6	meat	3.75	1.0	\$ 3.75
7	pasta	1.00	3.0	\$ 3.00
8	totals	13.25	10.0	\$21.25

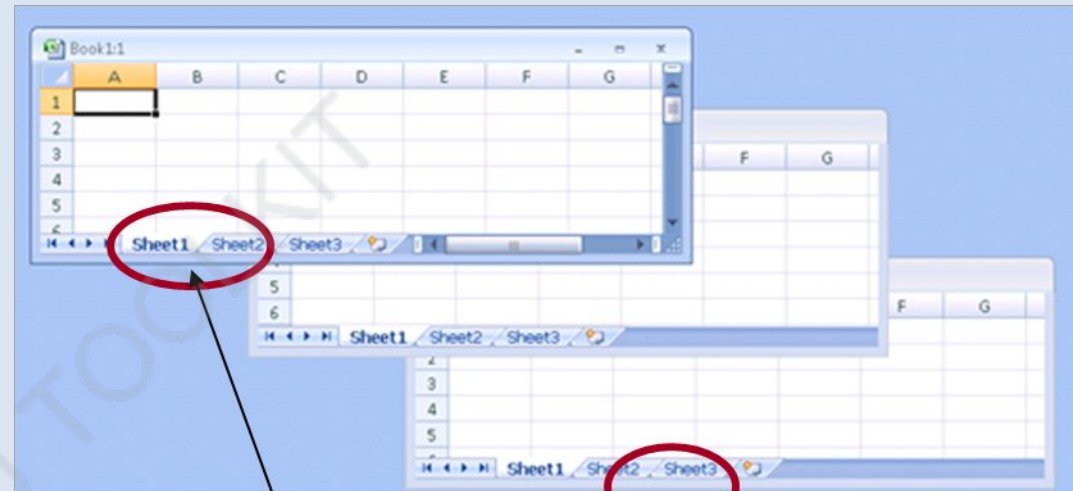
Rows

Labels



One can also write formulas that refer to cells on other worksheets – *Sheetname!*Cell-Reference

*input!B1*input!B3 + A1*
When referencing a cell on the same spreadsheet as the active cell the sheet name is not required.



Sheet tabs

*Sheets may be **named** and displayed with different **colors tabs**,
The **order** of the worksheets may be modified as well.*

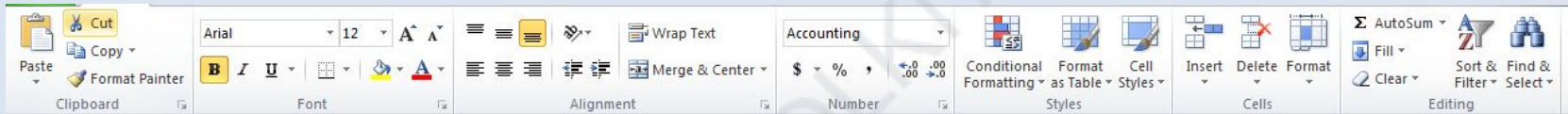




File tab – opens menus for opening and saving Files, and modifying Excel Options



Quick Access Toolbar can be customized to include icons to frequently Used features such as Print Preview

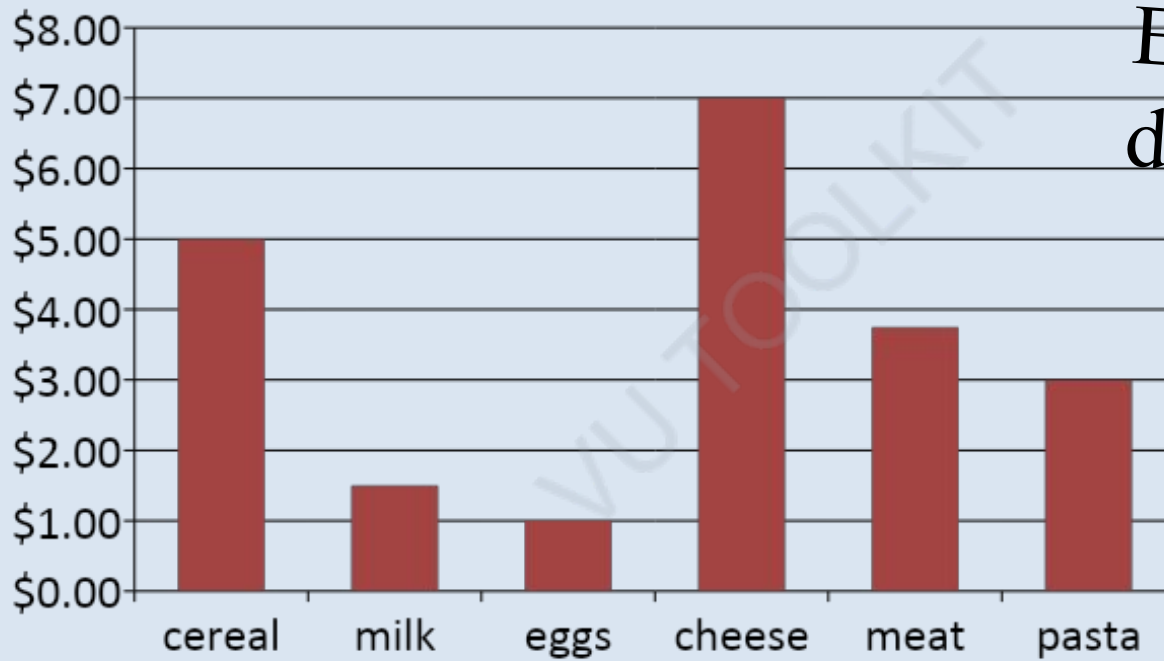


Home Ribbon use to change fonts, justify text, insert rows etc. Ribbons are organized into ***Groups*** of similar tasks such as the Font group or the Number group. In addition, there are other ribbons containing groups/buttons for laying out pages using the review features etc.





Highlight your data, select a Chart type and Edit & its done!

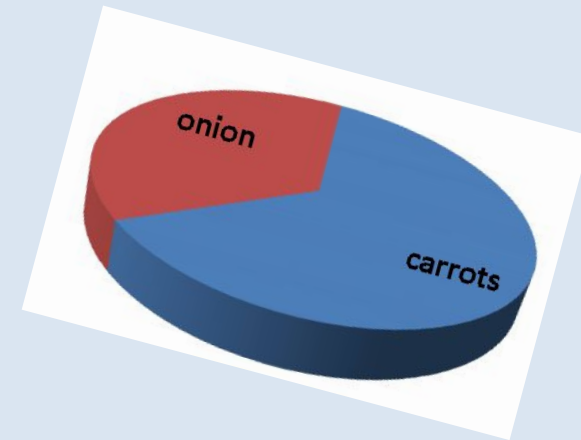


The “Power” of using Spreadsheet Applications

	A	B	C	D
1	Items	\$/each	Quantity	Total
2	carrots	\$ 1.00	4	\$ 4.00
3	onion	\$ 0.39	6	\$ 2.34
4	total			\$ 6.34
5				

=B2*C2

- Each entry can be related to other values by including cell referencing in *formulas*.
- Formula values are automatically updated when a referenced value changes
- Formulas can be copied
- Charts can be easily generated



Tap Here To Join Our
Official Community

Formulas

- A **formula** is a sequence of values, cell references and operators that produce a new value.

$$= E8 + 3*(E10 - E11)$$

- Formulas always start with an equal sign =
- In addition a formula can also contain built-in **functions** like SUM, AVERAGE, IF, COUNTIF, etc. **=Sum(A2:A8)*2**

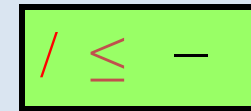


Things you need to know when writing formulas in Excel

- Data precision vs. cell display



- Types of operators that can be used



- Order of precedence of operators

An Excel formula bar with a pink background containing the formula $=B2+B3*B1/B8^2$. This illustrates the order of operations: exponentiation, multiplication and division, and addition.



In order to write Excel formulas we also need to use the correct Operator Symbols

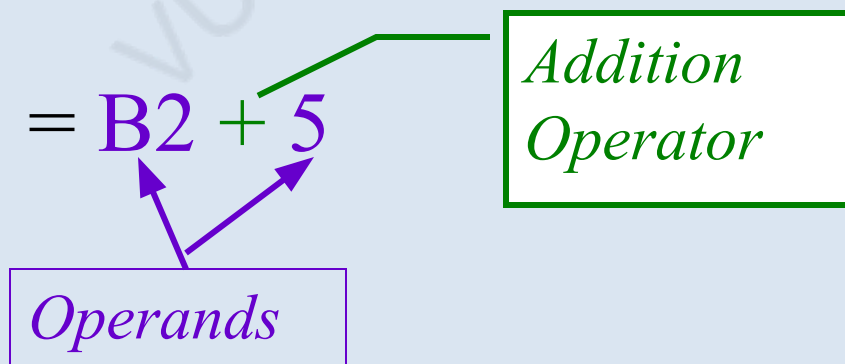
Formulas contain two types of components:

- **Operators**: Operations to be performed

*Arithmetic operators: * / + - ^*

Relational operators: >, <, <=, >=, < >, =

- **Operands**: Values to be operated on



Precedence of Operators

- *() Parenthesis* is a special operator that forces evaluation of the expression inside it first
- *Exponentiation* ($2^3 \square 8$)
- *Arithmetic operators: Multiplication & Division*
 - Multiplication & Division have equal precedence and are evaluated from left to right
- *Arithmetic operators: Addition & Subtraction*
 - Addition & Subtraction have equal precedence and are evaluated from left to right
- *Relational operators* have a lower precedence than arithmetic operators



Precision: number of decimal places stored in the computer.

Formatted Display: number of decimal places that appear in a cell

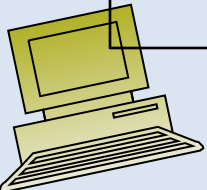
Type in a cell : =1/8
display in cell

	A	B
1	0 places	0
2	1 decimal	0.1
3	2 decimal	0.13
4	3 decimal	0.125
5		

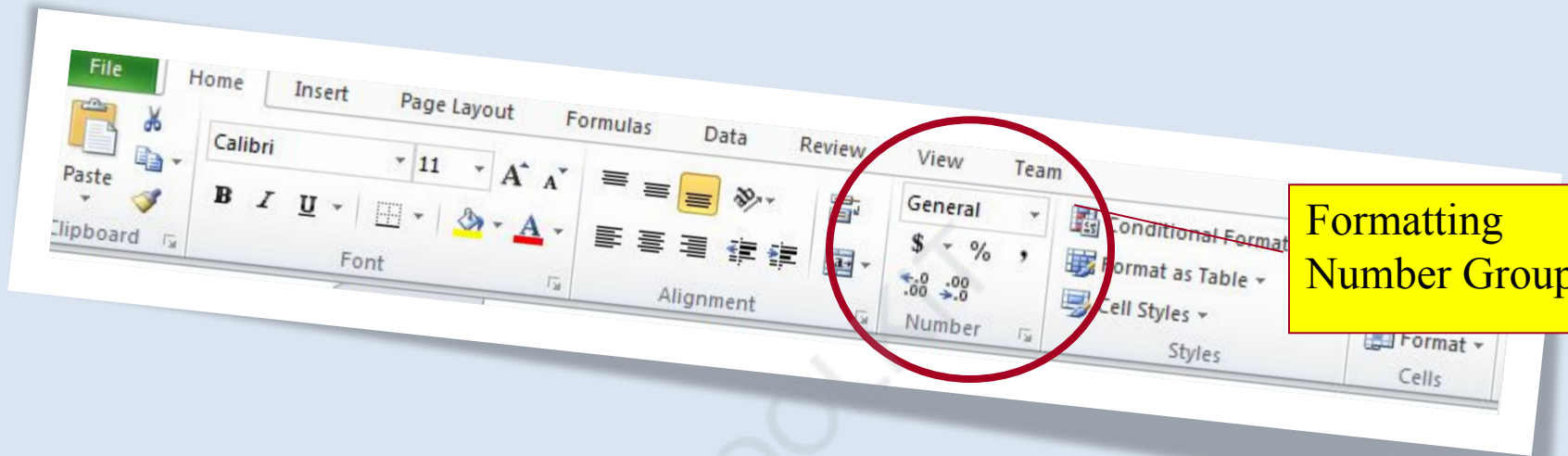
What value results for each - if multiplied by 1000?

Does the addition appear to be correct in col B?

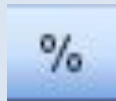
	A	B
1	21.4	21
2	51.3	51
3	98.1	98
4	170.8	171



Formatting affects display not the precise value:



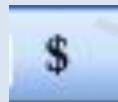
Percent



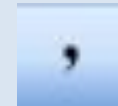
Decimal Display



Currency



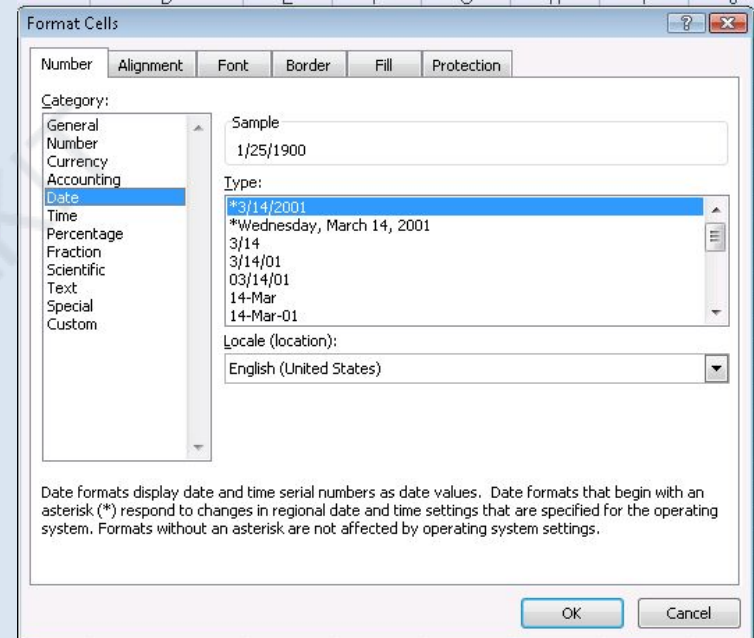
Commas



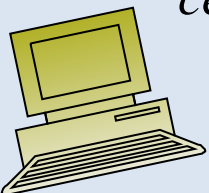
Values can also be used to display dates

	A	B
1	1/27/2013	Friday, February 01, 2013

- Dates are *values* that can be entered in several *formats*: January 27, 2013 or 1/27/2013
- Excel converts these dates to a numerical representation (1/22/2013 \square 41301)
- Thus dates may be used in formulas: =A1-B1 will result in the value 5



Note: To do arithmetic calculations with dates if you type =1/27/2013-1/22/2013 directly in a cell it does not interpret it a date – cell references must be used.



Tap Here To Join Our
Official Community

Walkthrough: Building a Simple Spreadsheet

- **Entering labels and values**
- **Formatting cells**
 - font, size, style, color, borders, alignment
 - Numeric Format, Currency, Decimal Places
 - text wrap, center titles
 - Column widths, row height
- **Inserting/Deleting rows and columns and sheets**
- **Writing a simple formula & modify decimal display**
- **Create a simple chart**
- **Sheet tabs**
 - Creating a new worksheets in a workbook (“new sections in a document”, Naming Sheets)



Microsoft Excel Vocabulary

- **Workbook** - an Excel file with one or more sheets or pages
- **Worksheet** - page in the workbook (*spreadsheet*)
- **Ribbon** – Tabbed section containing command icons
- **Row** - Horizontal (Row Number)
- **Column** - Vertical (Column Letter)
- **Cell** - Column/Row combination (ex: C3)
- **Values** - Numeric Entries used in calculations
- **Labels** - text that describes the data
- **Active Cell** - cell currently in use (highlighted)
- **Formula Bar** - top of spreadsheet where Excel displays the value or formula for that cell



Practical Exercise in Data Preparation

WU TOOLKIT



Practical Exercise in Data Preparation

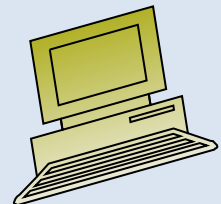
VU TOOLKIT



Lec 2 - Central Tendency.pptx

Lecture 2

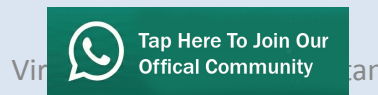
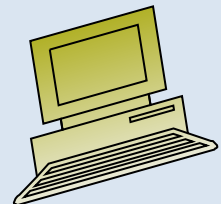
WU TOOLKIT



Tap Here To Join Our
Official Community Pakistan

Contents

- Measure of Central Tendency
- Mean
- Median
- Mode
- Grouping Data
- Frequency Distribution
- Group Mean
- Practical exercise in Data Preparation.



How to Use Excel to Find the Mean, Median & Mode

VU TOU KY

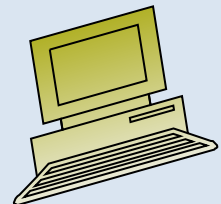


Tap Here To Join Our
Official Community

Objectives

After completing this lecture, you should be able to:

- Compute and interpret the mean, median, and mode for a set of data.
- Group data and frequency distribution.



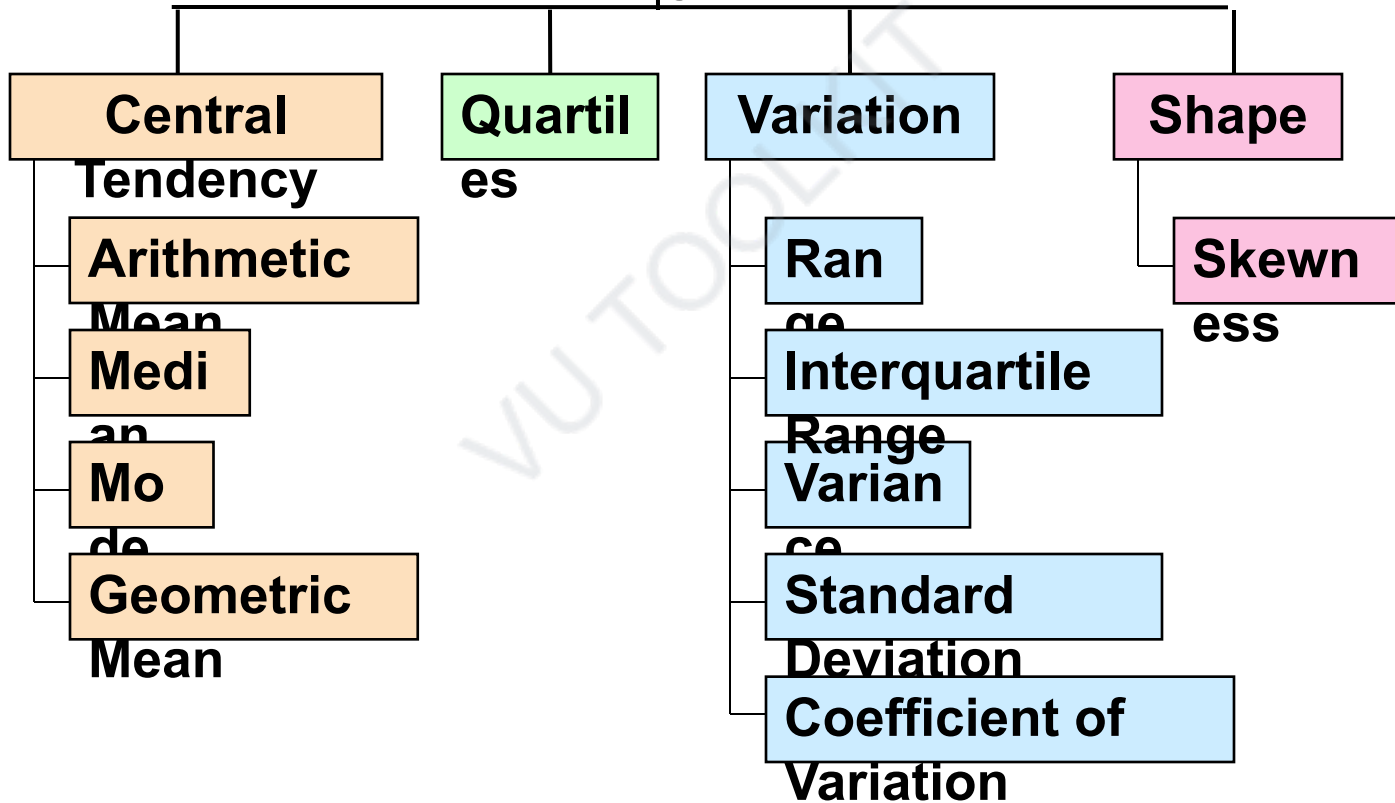
Central Tendency:

- In statistics a central tendency (or, more commonly, a measure of central tendency) is a central or typical value for a probability distribution. It may also be called a center or location of the distribution.
- The most common measures of central tendency are the arithmetic mean, the median and the mode.

A central tendency can be calculated for either a finite set of

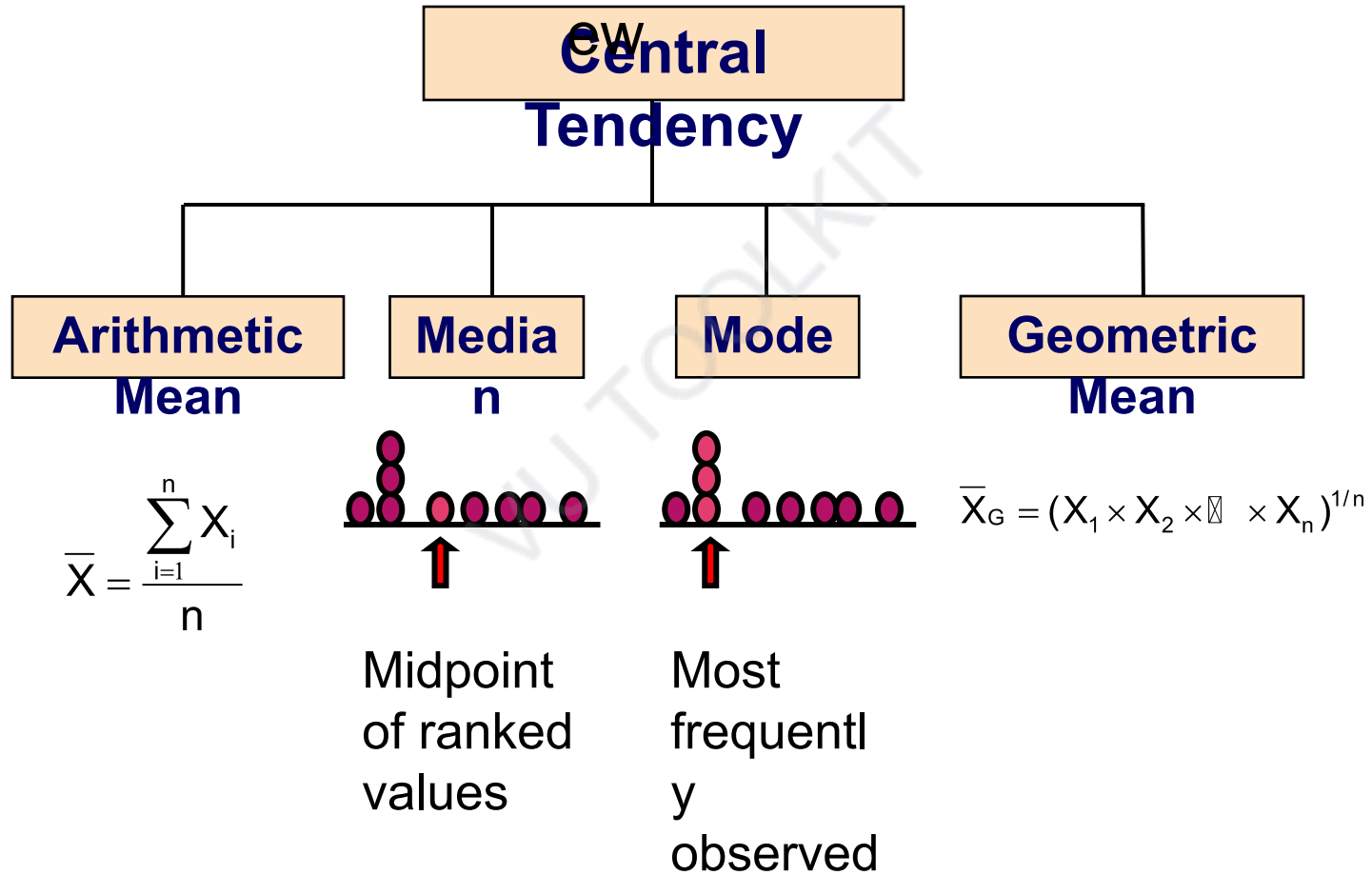


Describing Data Numerically



Measures of Central Tendency

Overview



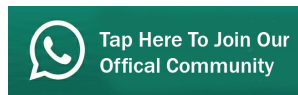
Arithmetic Mean

- ▶ The arithmetic mean (mean) is the most common measure of central tendency
- ▶ For a sample of size n:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Sample
size

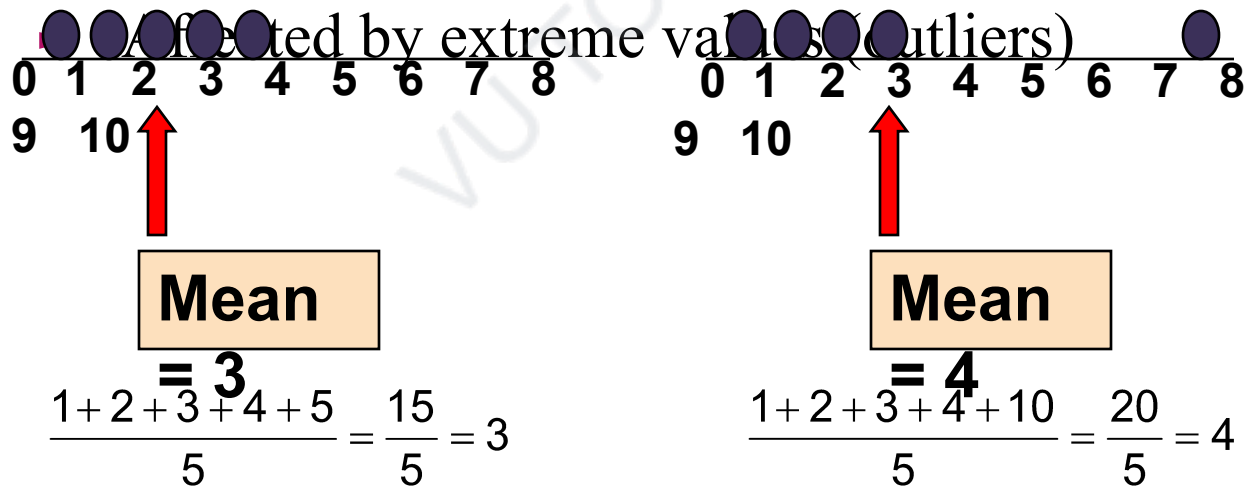
Observed
values



Arithmetic Mean

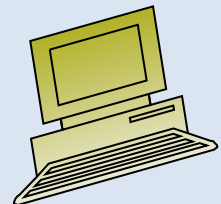
(continued)

- ▶ The most common measure of central tendency
- ▶ Mean = sum of values divided by the number of values



Introduction with excel:

- Microsoft Excel 2010 is designed to store numerical inputs and permit calculation on those numbers, making it an ideal program if you need to perform any numerical analysis such as computing the mean, median, mode and range for a set of numbers.
- Each of these four mathematical terms describes a slightly different way of looking at a set of numbers .
- Excel has a built-in function to determine each of them.



Tap Here To Join Our
Official Community

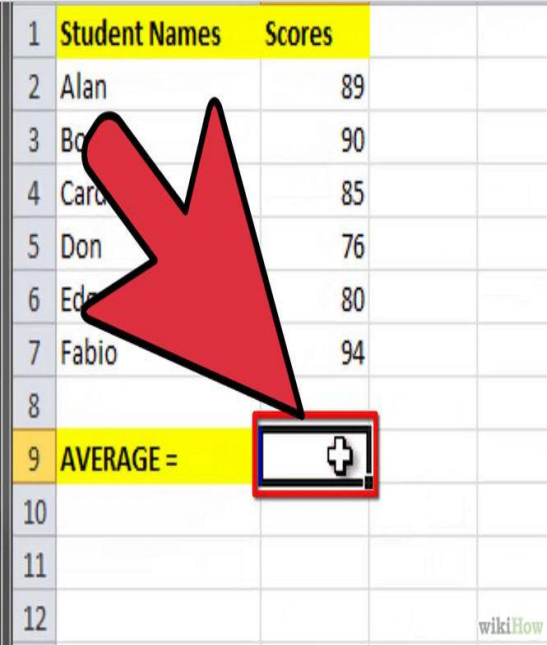
- Microsoft Excel includes a number of statistical functions, including the ability to figure the mean, median and mode of a data sample.
- While mean, the average of a group of numbers,
- Median, the midpoint number of a data group, are used
- Mode, the most frequently appearing number in a data group, can be useful as well, such as using the most frequent numeric grade score to provide feedback on the effectiveness of a teaching method.



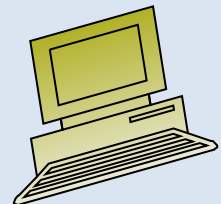
Mean in Excel:

Use the "AVERAGE" function in Excel to find the mean of a set of numbers.

- Enter the range of numbers in your Excel spreadsheet.
- Click where you want the mean (Average).

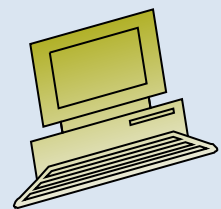
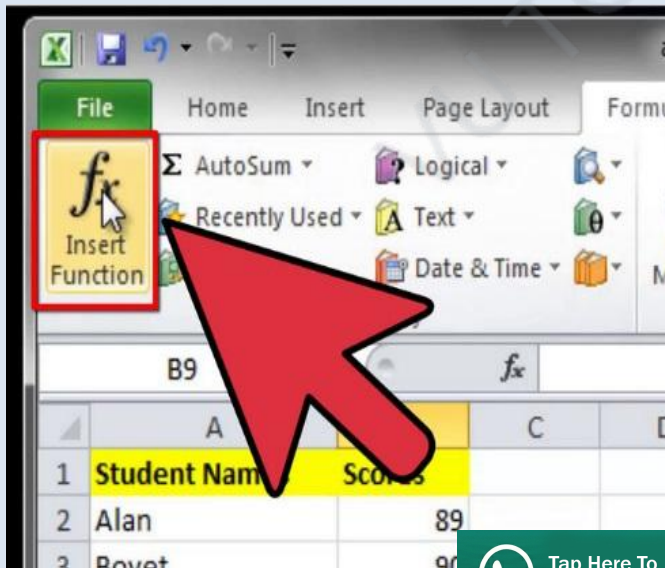


1	Student Names	Scores		
2	Alan	89		
3	Bob	90		
4	Carol	85		
5	Don	76		
6	Ed	80		
7	Fabio	94		
8				
9	AVERAGE =			
10				
11				
12				



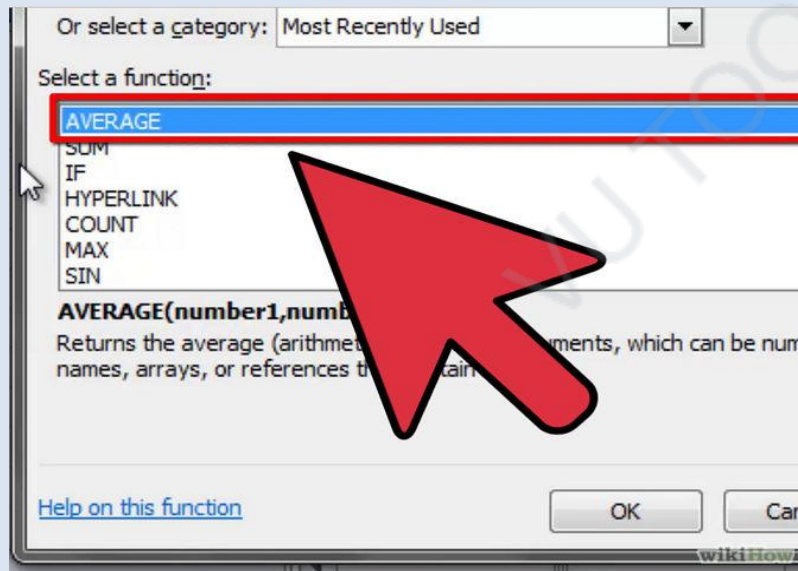
Tap Here To Join Our
Official Community

- Click "Formulas" and select the "Insert Function" tab.
- Enter the numbers in your Excel spreadsheet in either a row or a column.



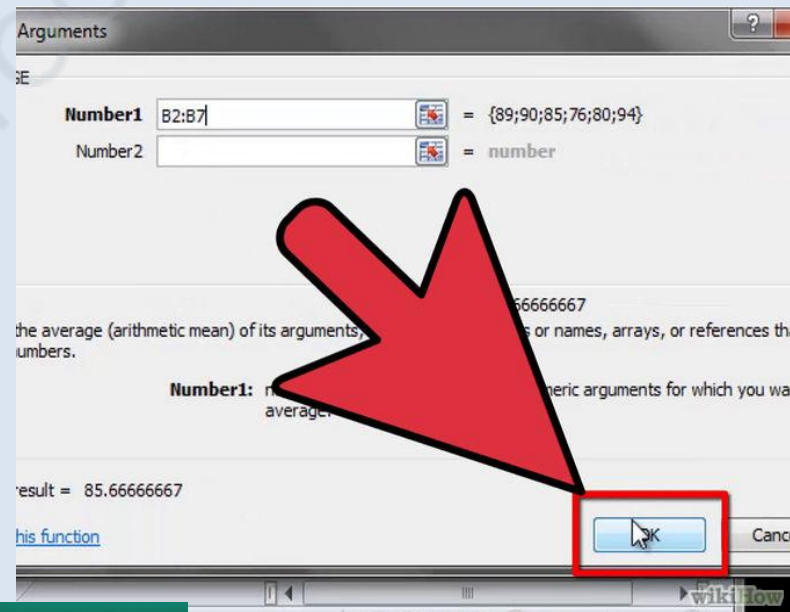
Select Formula:

- Scroll down and select the "Average" function.

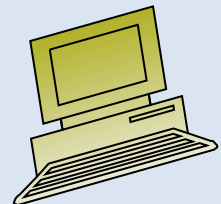


Select Data

- Select the range of data to find the value of mean.
- Enter the cell range for your list of numbers in the number 1 box, for instance B2:B7 and click "OK".

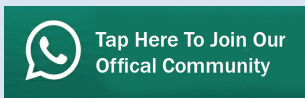
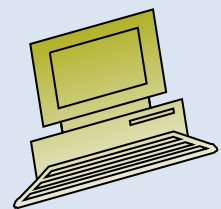
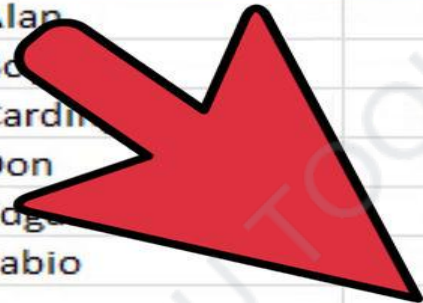


Tap Here To Join Our
Official Community



- The mean (average) for the list will appear in the cell you selected.

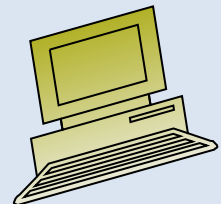
	A	B	C	D
1	Student Names	Scores		
2	Alan	89		
3	Bob	90		
4	Cardin	85		
5	Don	76		
6	Edgar	80		
7	Fabio	94		
8				
9	AVERAGE =	85.66667		
10				
11				wikiHow



Median

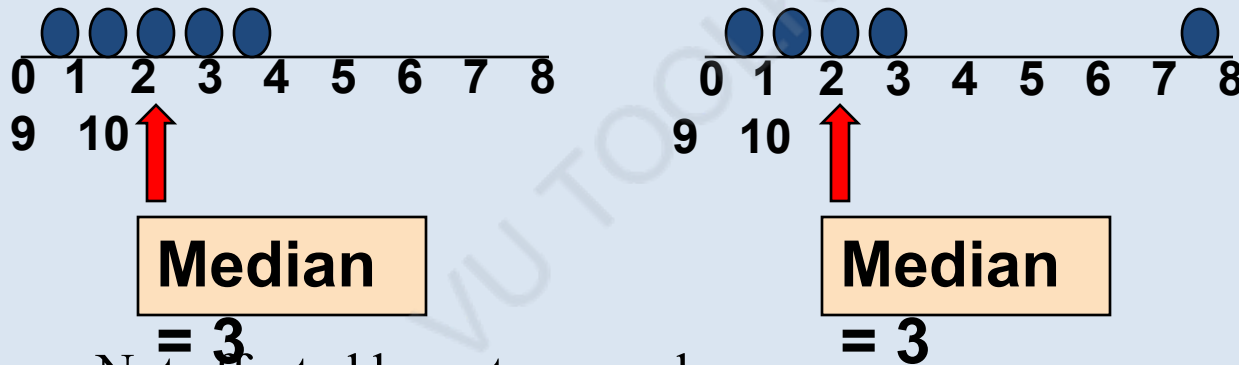
- To find the **median**, list the values of the data set in numerical order and identify which value appears in the middle of the list.

VU TOOLKIT

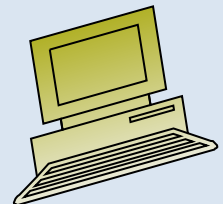


Median

- ▶ In an ordered array, the median is the “middle” number (50% above, 50% below)



- ▶ Not affected by extreme values



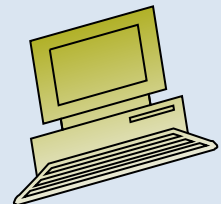
Finding the Median

- ▶ The location of the median:

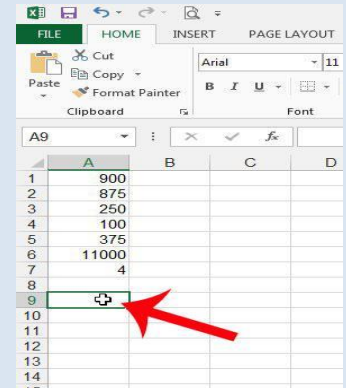
$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- ▶ If the number of values is odd, the median is the middle number
- ▶ If the number of values is even, the median is the average of the two middle numbers
$$\frac{\text{middle number 1} + \text{middle number 2}}{2}$$

- ▶ Note that $\frac{n+1}{2}$ is not the value of the median, only the position of the median in the ranked data



Median in Excel



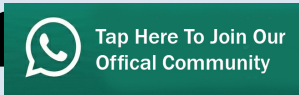
- **Step 1**

Open a new Microsoft Excel 2010 spreadsheet by double-clicking the Excel icon.

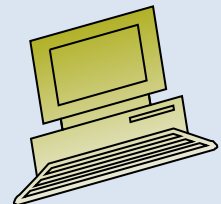
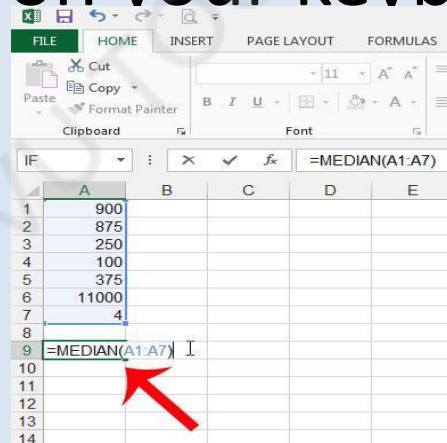
- **Step 2**

Click on cell A1 and enter the first number in the set of numbers that you are investigating. Press "Enter" and the program will automatically select cell A2 for you. Enter the second number into cell A2 and continue until you have entered the entire set of numbers into

column A



- Step 3: Type **=MEDIAN(AA:BB)** into the cell, where **AA** is the cell location of the first cell of your range, and **BB** is the cell location of the last cell. In my example image below, the formula would be **=MEDIAN(A1:A7)**. You can then press **Enter** on your keyboard to calculate the formula.

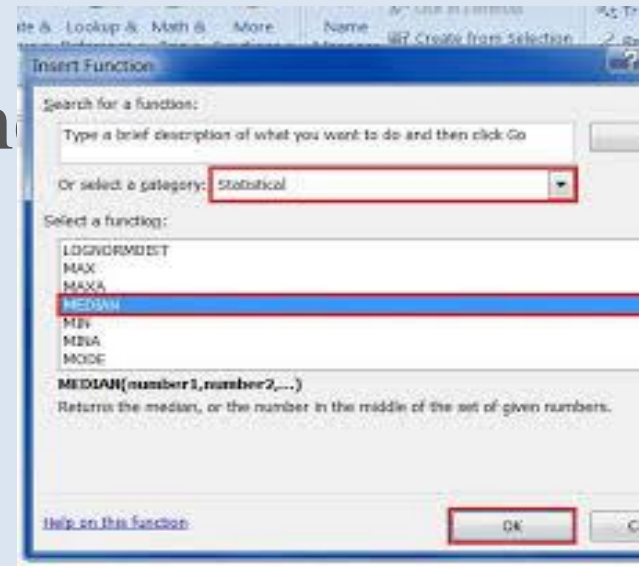
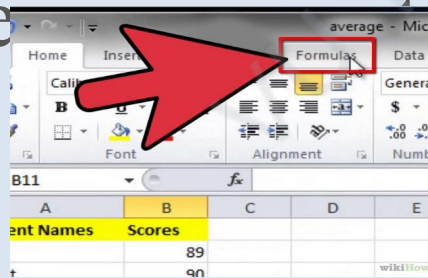


Tap Here To Join Our
Official Community

How to select function

Use the “Median” function in Excel to find the median of a set of numbers.

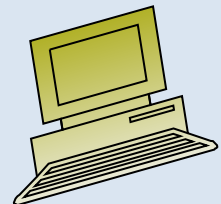
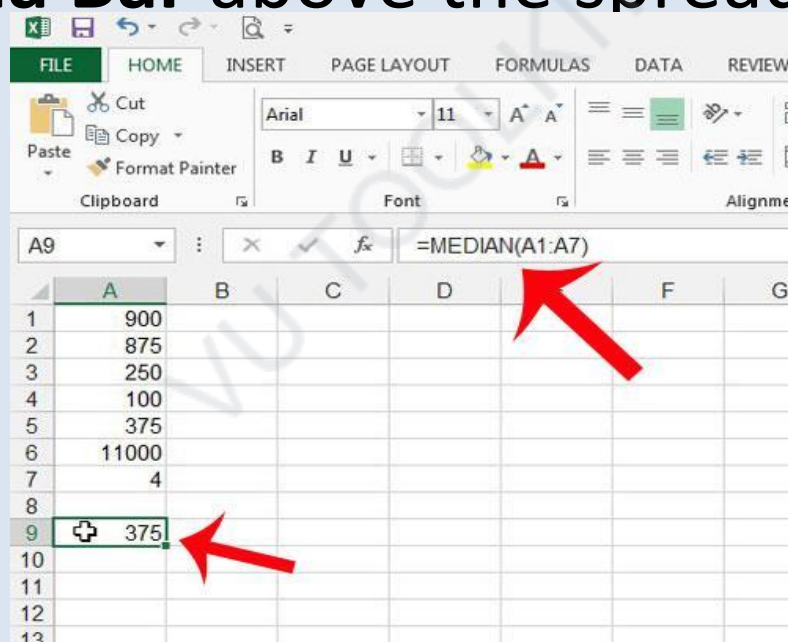
- Enter the range of numbers in your Excel spreadsheet.
- Click where the m



Tap Here To Join Our
Official Community



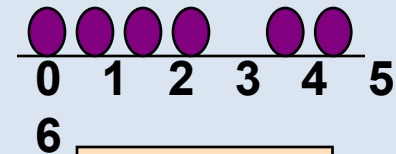
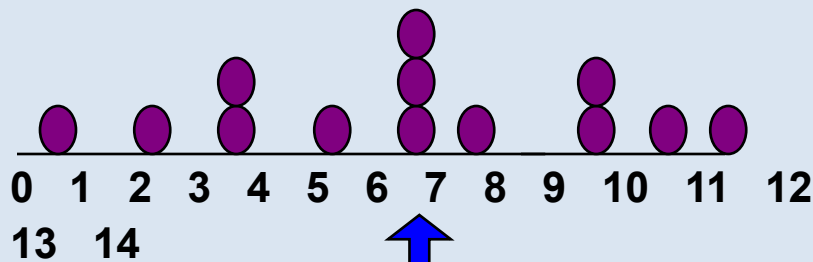
- Note that the median will be displayed in the cell, but that you can still view the formula by clicking on the cell, then looking at the **Formula Bar** above the spreadsheet.



Tap Here To Join Our
Official Community

Mode

- ▶ A measure of central tendency
- ▶ Value that occurs most often
- ▶ Not affected by extreme values
- ▶ Used for either numerical or categorical data
- ▶ There may may be no mode
- ▶ There may be several modes



No
Mode

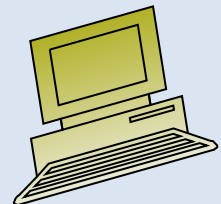
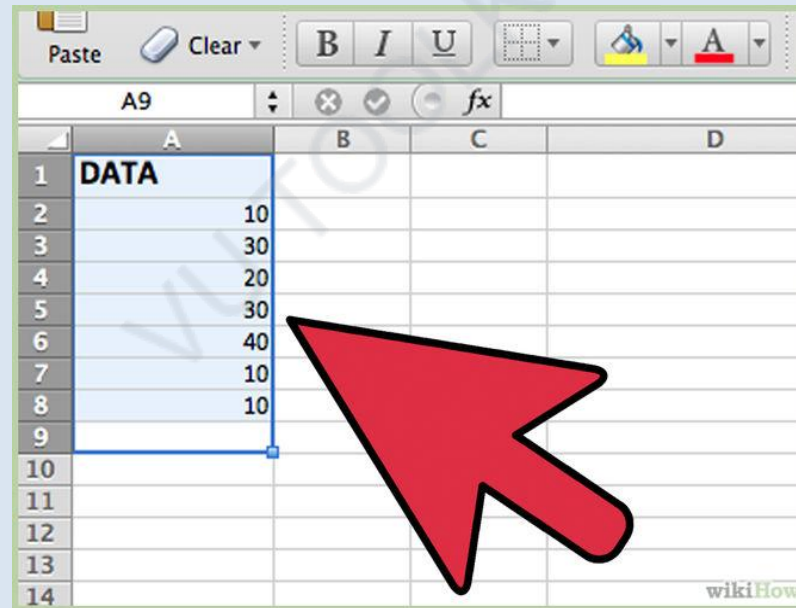


Tap Here To Join Our
Official Community

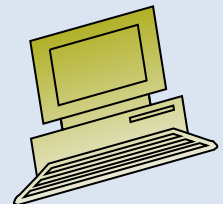
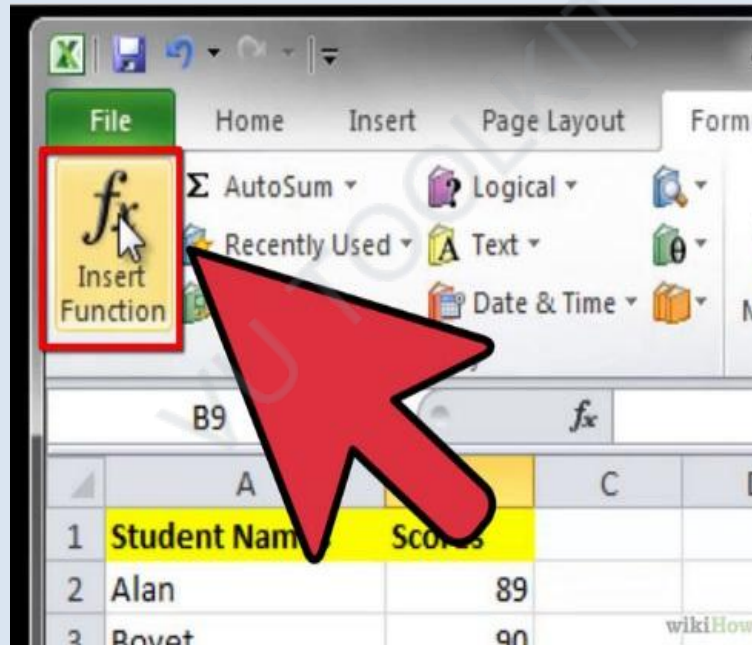


How to calculate mode using Excel.

- **Enter each number in the data set into its own cell.** For consistency, it helps to enter the number in consecutive cells in either a row or column, and for readability, a column is better.

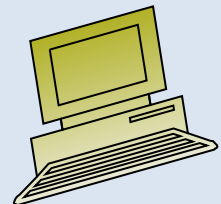


- Click "Formulas" and select the "Insert Function" tab.
- Enter the numbers in your Excel spreadsheet in either a row or a column.

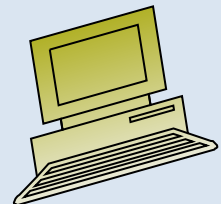
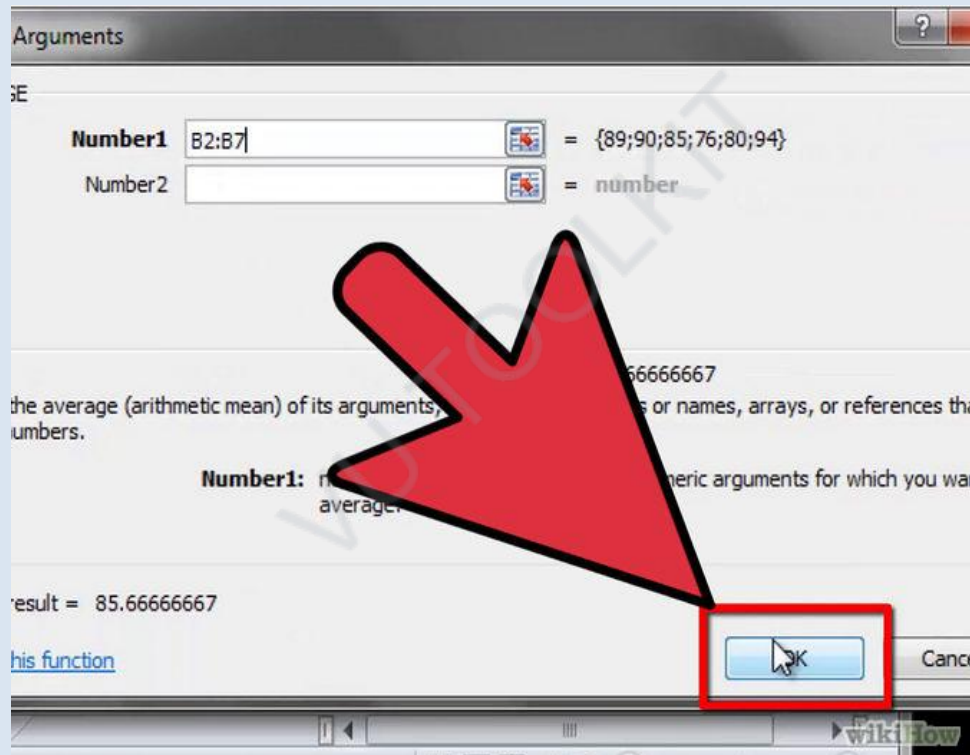


Mode

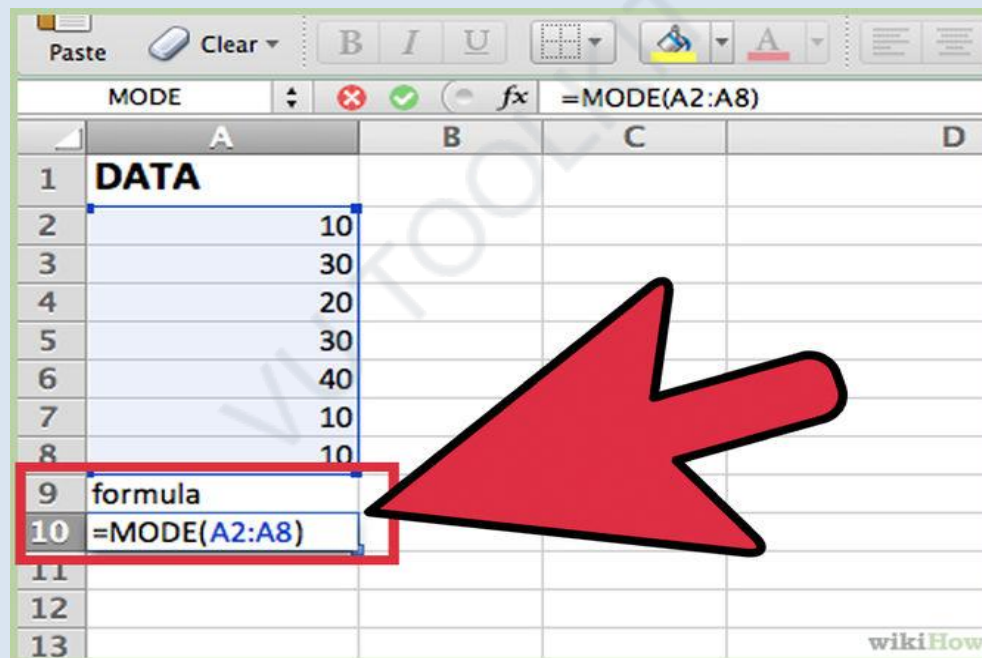
- Enter the **MODE** function into the cell in which you wish to display the result. The **MODE** function's format is "**=MODE(Cx:Dy)**," where C and D represent the letter of the column of the first and last cell in the range, and x and y represent the number of the first and last row in the range. (Although different letters are used in this example, you will use the same column letter for both the first and last cell if you entered the data in a column of cells or the same row number for both the first and last cell if you entered the data in a row of cells.)
- You can also specify each cell individually, up to 255 cells,
- as in "**=MODE(A1, A2, A3)**," but this is not advisable unless you have only a very small dataset and do not plan to add to it. You can also use the function with constants, for example, "**=MODE(4,4,6)**," which requires editing the function each time you wish to search for a different



- Enter the cell range for your list of numbers in the number 1 box, for instance A2:A8 and click "OK".



- Enter the **MODE** function into the cell in which you wish to display the result.

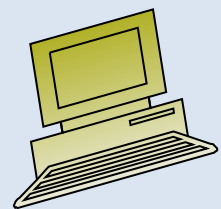



Paste Clear

MODE \downarrow \times \checkmark fx =mode(A1:A8)

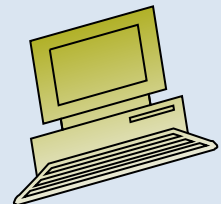
	A	B	C	D
1	DATA			
2		10		
3		7		
4		9		
5		8		
6		7		
7		0		
8		4		
9	formula			
10	=mode(A1:A8)			
11				
12				
13				
14				

wikiHow



 Tap Here To Join Our Official Community

- **Calculate and display the result.**
- For a dataset of 10, 7, 9, 8, 7, 0 and 4 entered in cells 1 through 8 of Column A, the function =MODE(A1:A8) will deliver a result of 7, because 7 appears more often in the data than any other number.



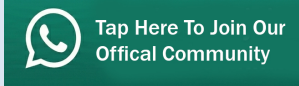
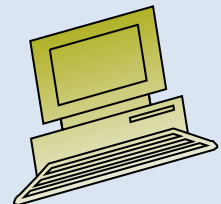
More than one mode

- If the data set contains more than one number that qualifies as the mode (such as 7 and 9 each appearing twice and every other number appearing only once), whichever mode number is listed first in the data set will be the result. If none of the numbers in the data set appear more often than any other, the MODE function will display the error result #N/A.
- The MODE function is available in all versions of Excel, including Excel 2010, which includes it for compatibility with spreadsheets created in earlier versions.
- Excel 2010 uses the MODE.SNGL function, which except for syntax (=MODE.SNGL(Cx:Dy)) works essentially the same as the MODE function in earlier versions of Excel.



- Mean is generally used, unless extreme values (outliers) exist
- Then median is often used, since the median is not sensitive to extreme values.

VU TOOLKIT



Practical Exercise in Data Preparation

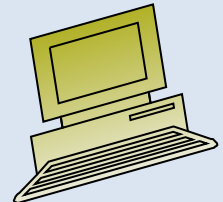
VU TOOKKIT



Lec 3 - Dispersion.pptx

Lecture 3

WU TOOLKIT

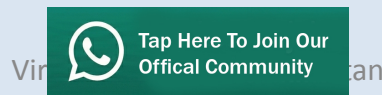
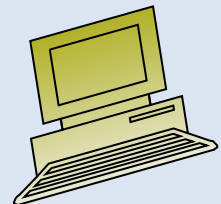


Tap Here To Join Our
Official Community Pakistan

Contents

Measure of Variation

- Range,
- Standard deviation,
- Variance,
- Inter Quartile Range,
- Coefficient of Variation,
- Box Plot ,
- 5-number summary,
- Real Statistics Add-in ,
- Practical exercise.

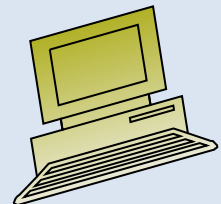


Objectives

After completing this lecture, you should be able to:

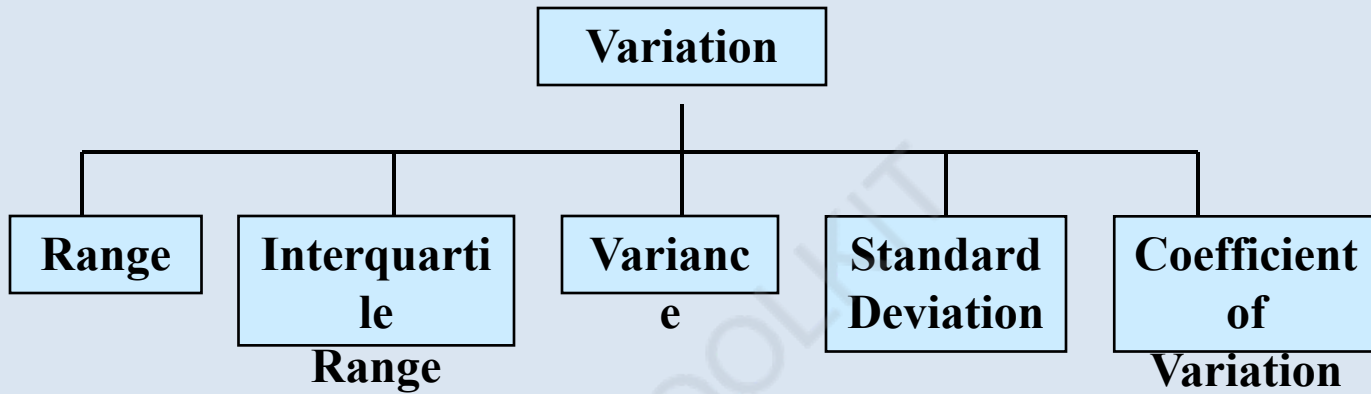
Compute and interpret :

- Range,
- Standard deviation,
- Variance,
- Inter Quartile Range,
- Coefficient of Variation,
- Box Plot.

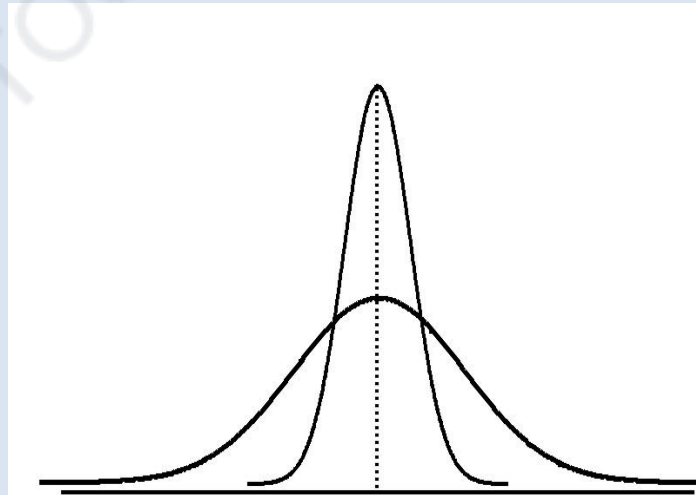


Tap Here To Join Our
Official Community

Measures of Variation:



- Measures of variation give information on the **spread** or **variability** of the data values.



Same center,
**different
variation**



Tap Here To Join Our
Official Community

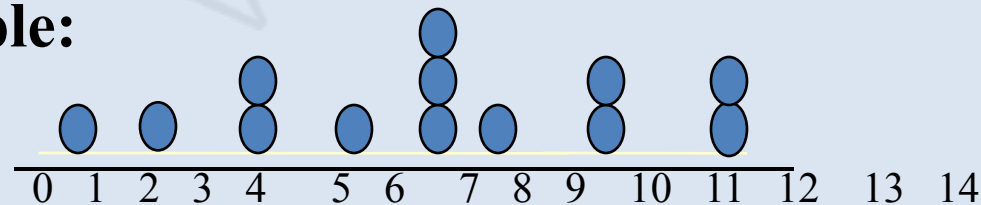


Range:

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

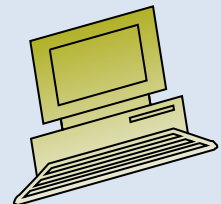
Example:



$$\text{Range} = 14 - 1 = 13$$

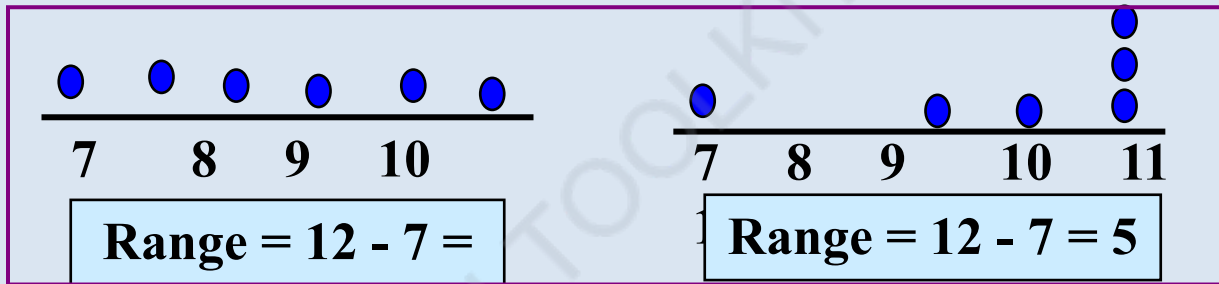


Tap Here To Join Our
Official Community



Disadvantages of the Range

- Ignores the way in which data are distributed



- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

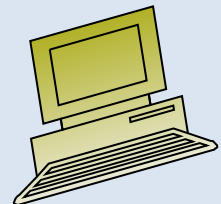
$$\text{Range} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 =$$



Tap Here To Join Our
Official Community



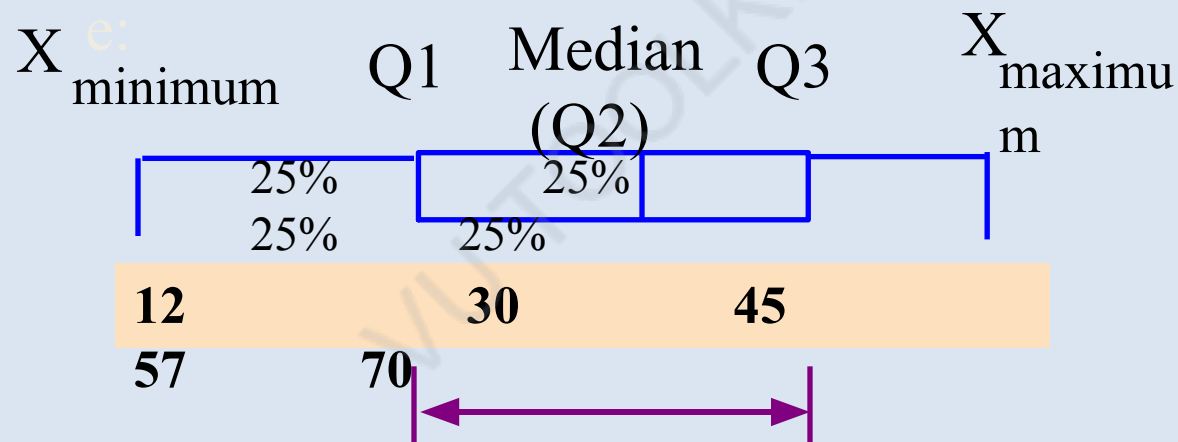
Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high- and low-valued observations and calculate the range from the remaining values
- Interquartile range = 3rd quartile – 1st quartile
$$= Q_3 - Q_1$$



Interquartile Range

Exempl

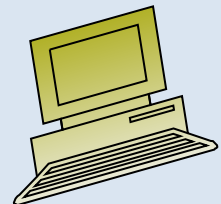


Interquartile
range

$$= 57 - 30 =$$



Tap Here To Join Our
Official Community



Variance

- Average (approximately) of squared deviations of values from the mean

– Sample variance:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Where \bar{X} = arithmetic mean

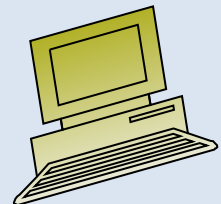
n = sample size

X_i = i^{th} value of the variable

X



Tap Here To Join Our
Official Community

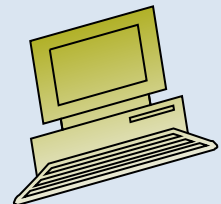


Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

– Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$



Calculation Example: Sample Standard Deviation

Sample

Data (X_i): 10 12 14 15 17 18 18

24

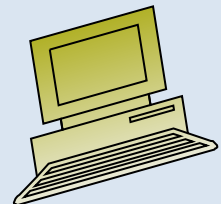
$n = 8$

Mean = \bar{X}

$$s = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{X})^2 + (14 - \bar{X})^2 + \dots + (24 - \bar{X})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \dots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{130}{7}} = 4.309$$



Tap Here To Join Our
Official Community

Population Variance

- Average of squared deviations of values from the mean

– Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Where μ = population mean

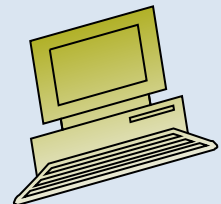
N = population size

X_i = i^{th} value of the variable

X



Tap Here To Join Our
Official Community

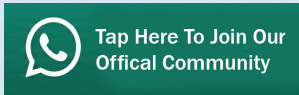
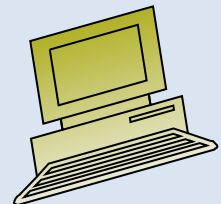


Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

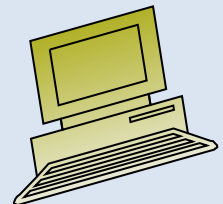
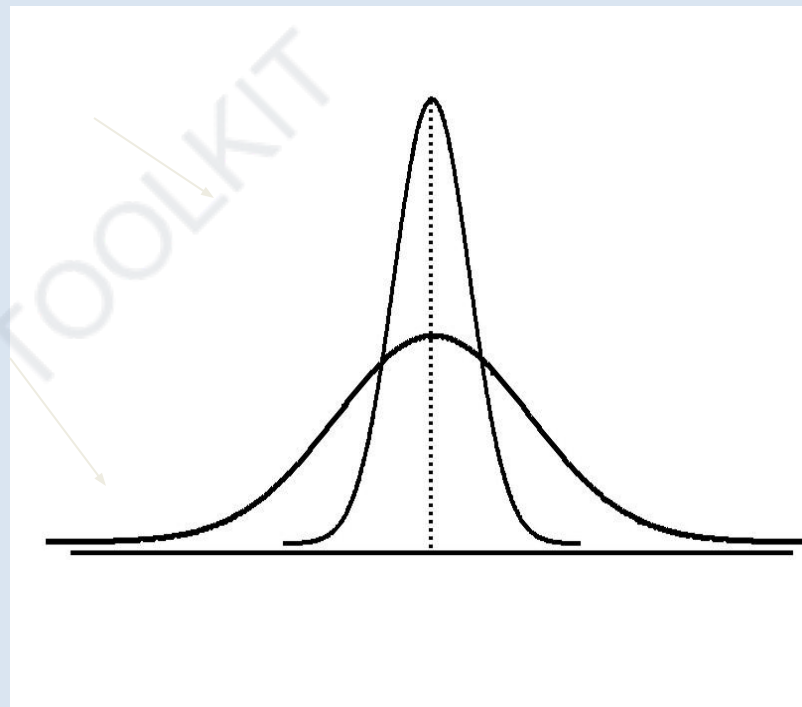
WU TOOLS



Measuring variation

Small standard deviation

Large standard deviation



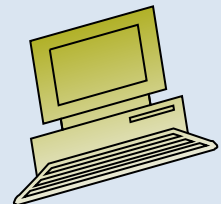
Enter Data:

- **Step 1**

Open a new Microsoft Excel 2010 spreadsheet by double-clicking the Excel icon.

- **Step 2**

Click on cell B1 and enter the first number in the set of numbers that you are investigating. Press "Enter" and the program will automatically select cell B2 for you. Enter the second number into cell B2 and continue until you have entered the entire set of numbers into column B.



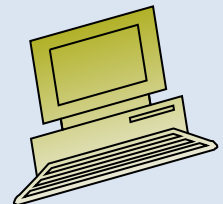
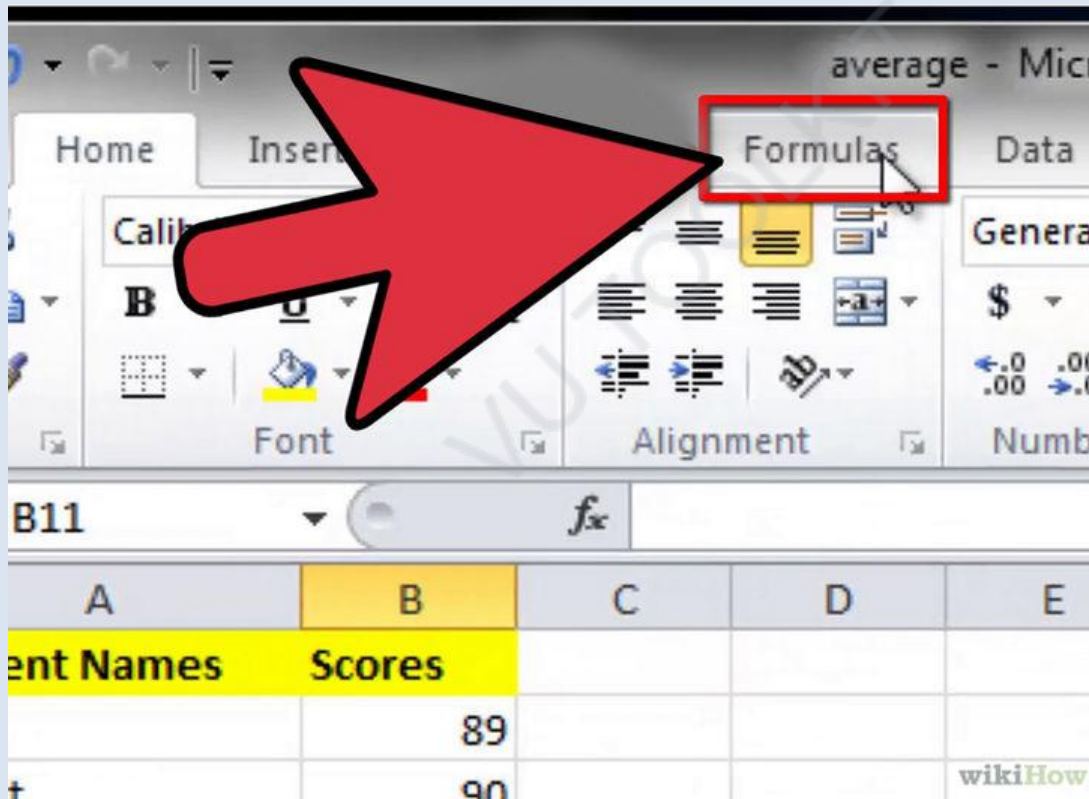
Tap Here To Join Our
Official Community

Advantages of Variance and Standard Deviation

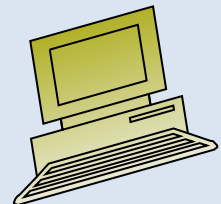
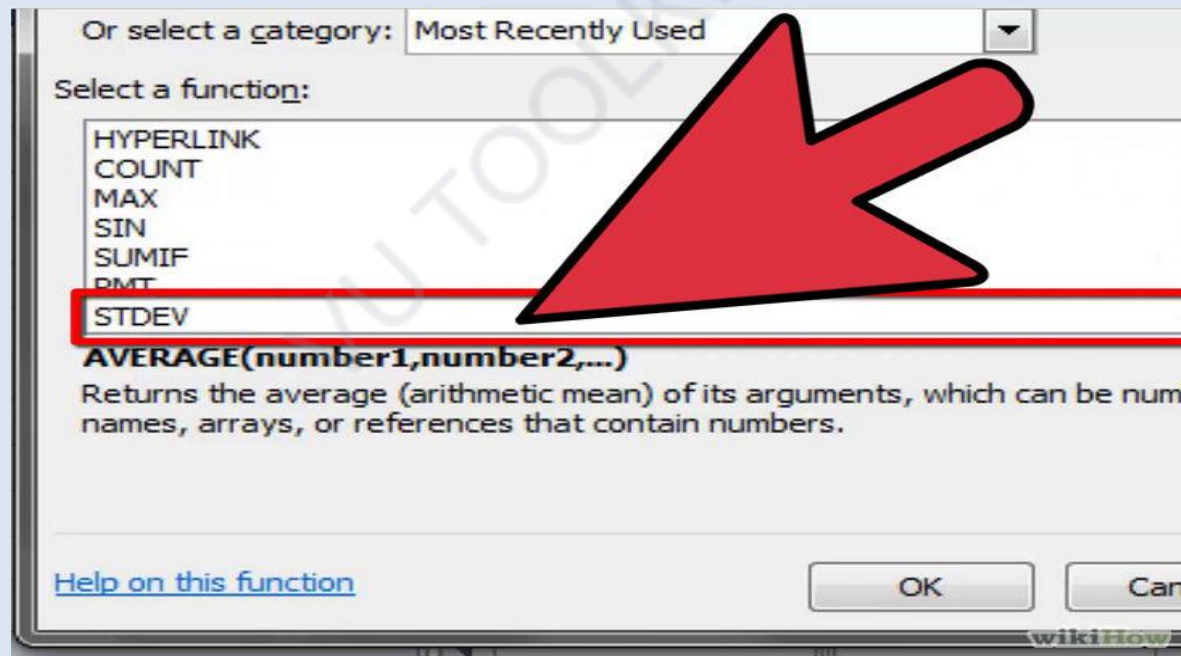
- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight
(because deviations from the mean are squared)




Click on "Formulas" and select the "Insert Function" (fx) tab again.

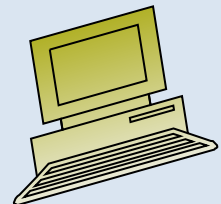
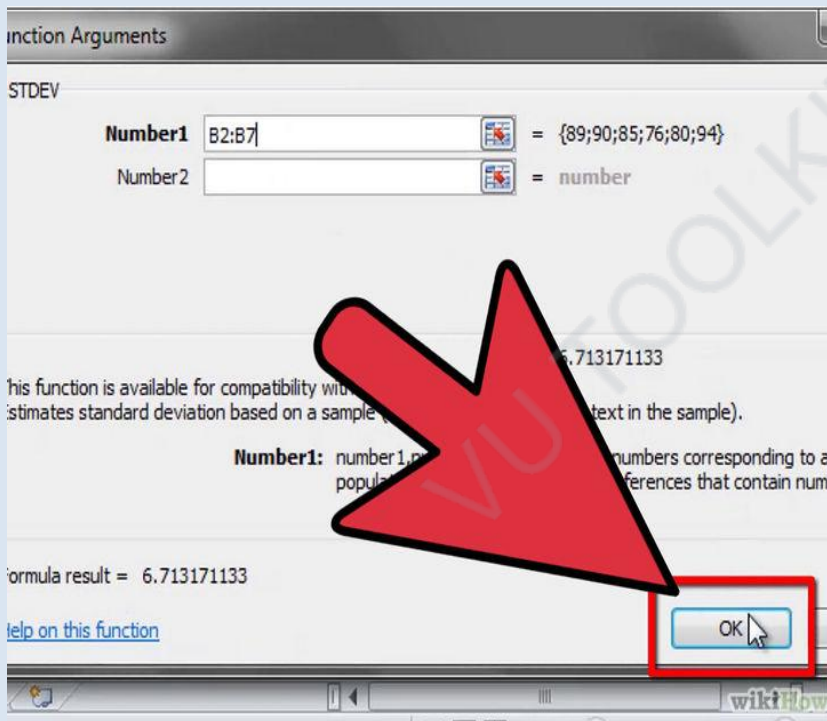


Scroll down the dialog box and select the STDEV function



 Tap Here To Join Our Official Community

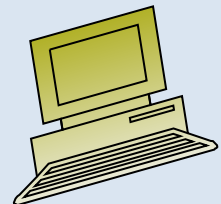
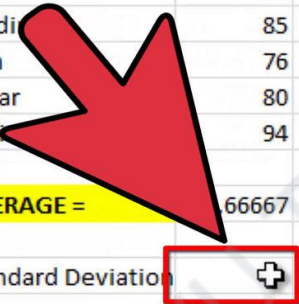
Enter the cell range for your list of numbers in the number 1 box and click OK



 Tap Here To Join Our Official Community


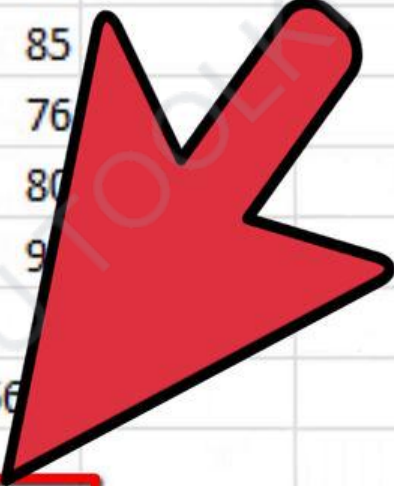
Use the STDEV function to compute the standard deviation. Place your cursor where you wish to have it appear.

1	Student Names	Scores		
2	Alan	89		
3	Boyet	90		
4	Cardi	85		
5	Don	76		
6	Edgar	80		
7	Fabr	94		
8				
9	AVERAGE =	66667		
10				
11	Standard Deviation			
12				wikiHow

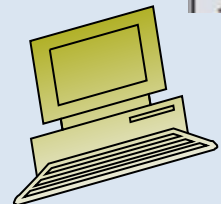


The standard deviation will appear in the cell you selected.

1	Student Names	Scores		
2	Alan	89		
3	Boyet	90		
4	Carding	85		
5	Don	76		
6	Edgar	80		
7	Fabio	9		
8				
9	AVERAGE =	85.666		
10				
11	Standard Deviation	6.713171		
12				

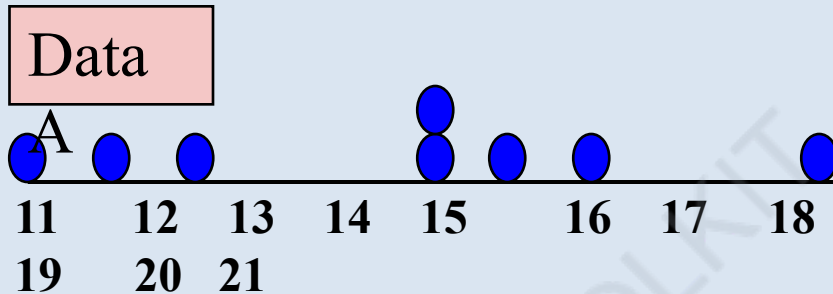


wikiHow

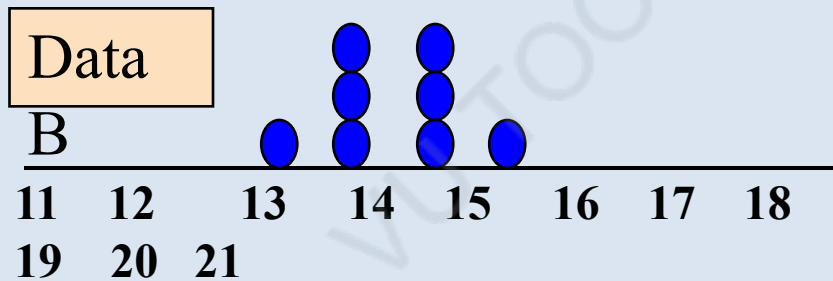


Tap Here To Join Our
Official Community

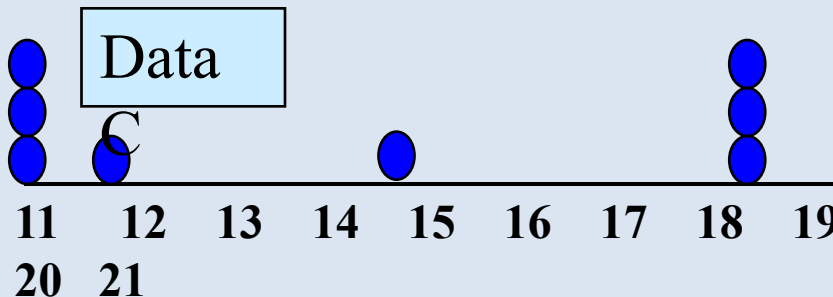
Comparing Standard Deviations:



Mean =
15.5
 $S = 3.34$



Mean =
15.5
 $S = 0.92$



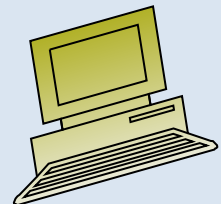
Mean =
15.5
 $S = 4.84$



Coefficient of Variation:

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$



Tap Here To Join Our
Official Community

Comparing Coefficient of Variation

- Stock A:

- Average price last year = \$50
- Standard deviation = \$5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:

- Average price last year = \$100
- Standard deviation = \$5

$$CV_B = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less

variable relative to its price

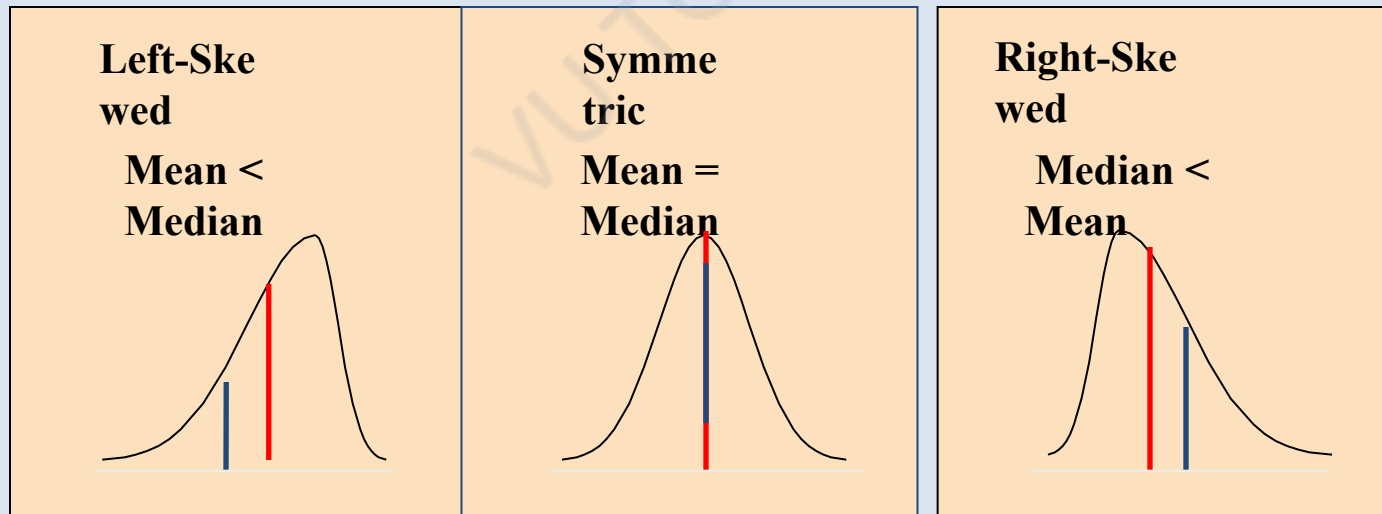


Tap Here To Join Our Official Community

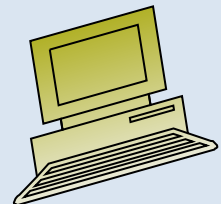


Shape of a Distribution:

- Describes how data is distributed
- Measures of shape
 - Symmetric or skewed

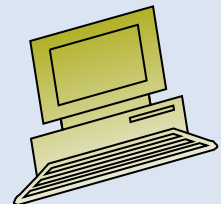


Tap Here To Join Our
Official Community

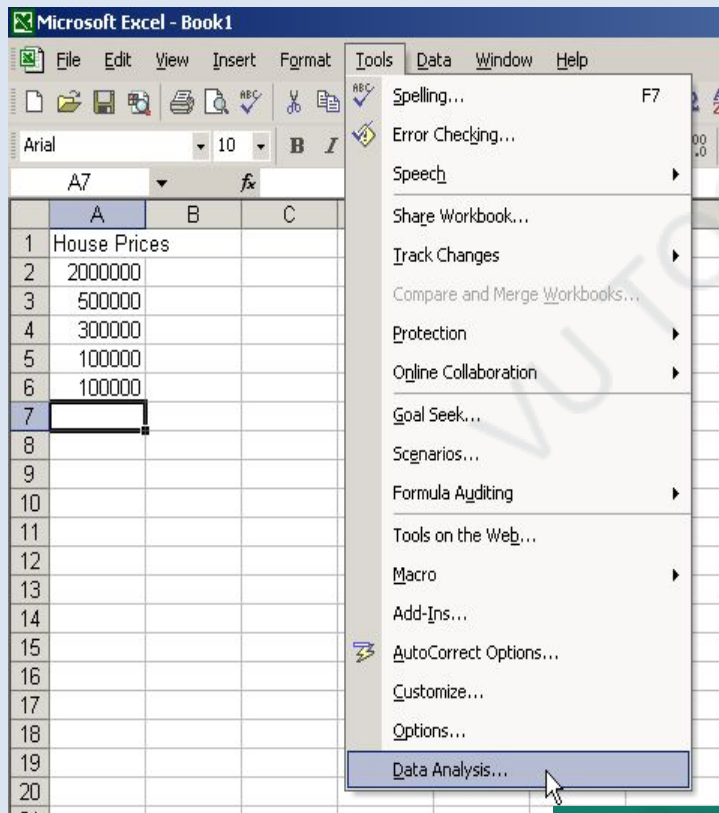


Using Microsoft Excel:

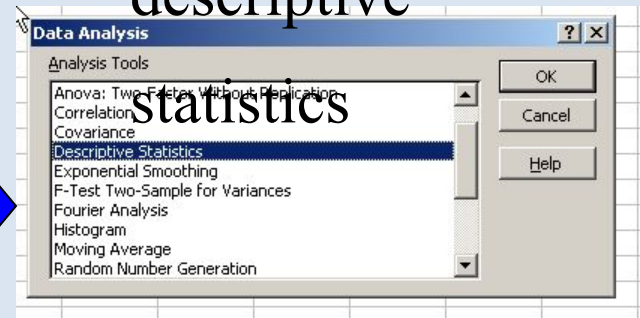
- Descriptive Statistics can be obtained from Microsoft® Excel
 - Use menu choice:
tools / data analysis / descriptive statistics
 - Enter details in dialog box



Using Excel:



- Use menu choice:
tools / data
analysis /
descriptive
statistics



Tap Here To Join Our
Official Community

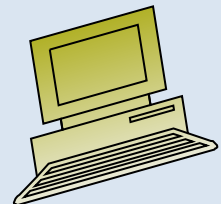
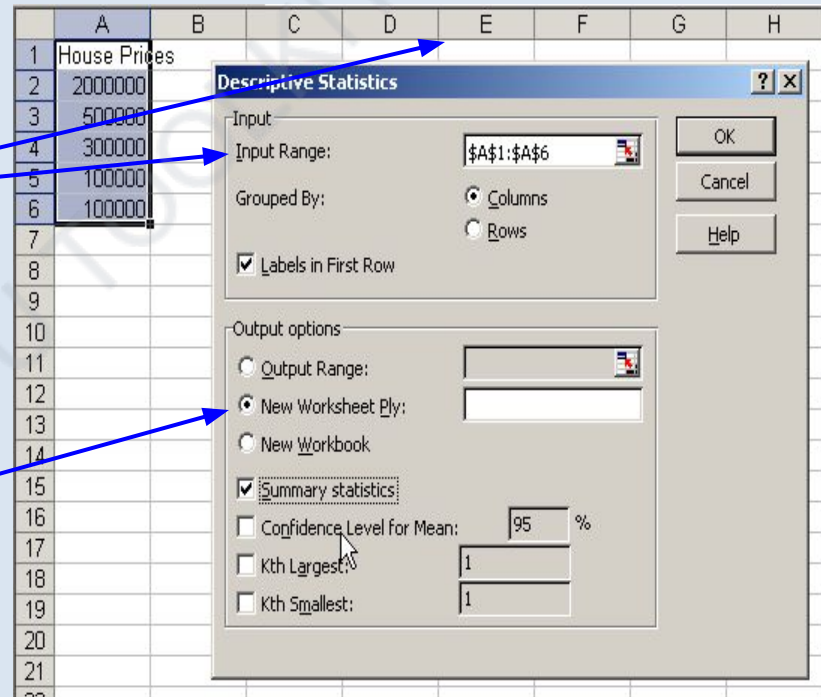


Using Excel:

- Enter dialog box details

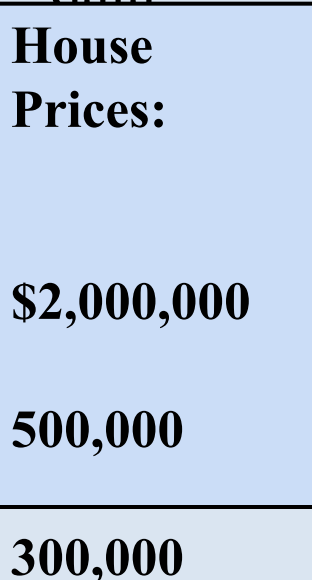
- Check box for summary statistics

- Click OK



Excel output

Microsoft Excel
descriptive statistics
output,
using the house price
data:



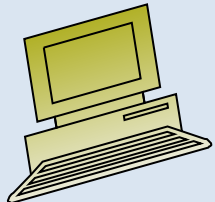
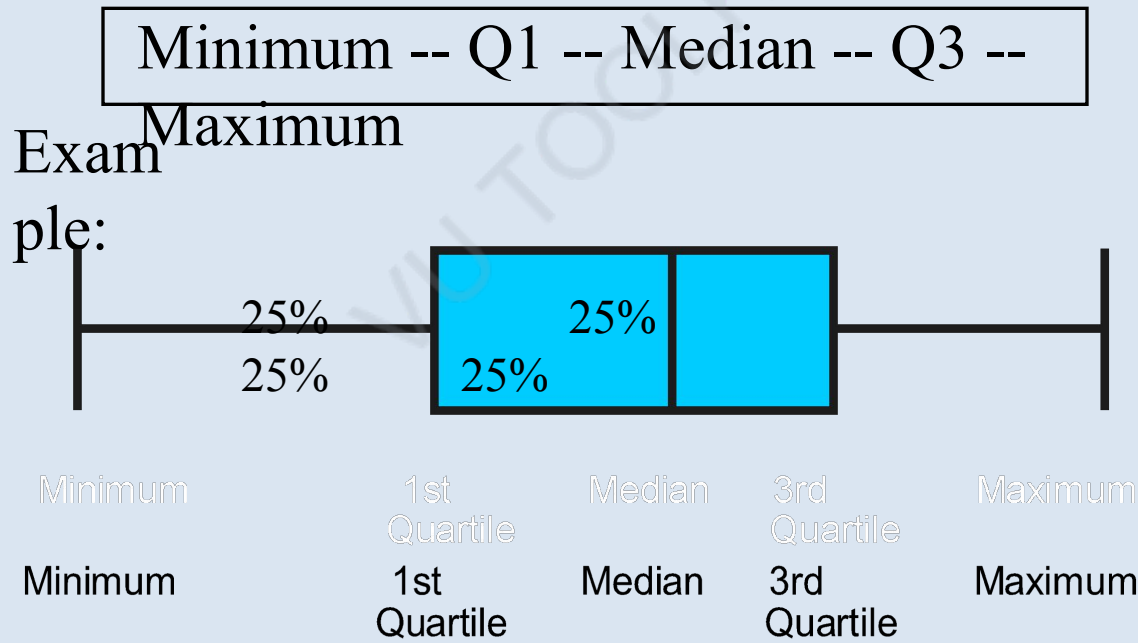
	A	B	
1	<i>House Prices</i>		
2			
3	Mean	600000	
4	Standard Error	357770.8764	
5	Median	300000	
6	Mode	100000	
7	Standard Deviation	800000	
8	Sample Variance	6.4E+11	
9	Kurtosis	4.130126953	
10	Skewness	2.006835938	
11	Range	1900000	
12	Minimum	100000	
13	Maximum	2000000	
14	Sum	3000000	
15	Count	5	
16			
17			



Tap Here To Join Our
Official Community

Exploratory Data Analysis:

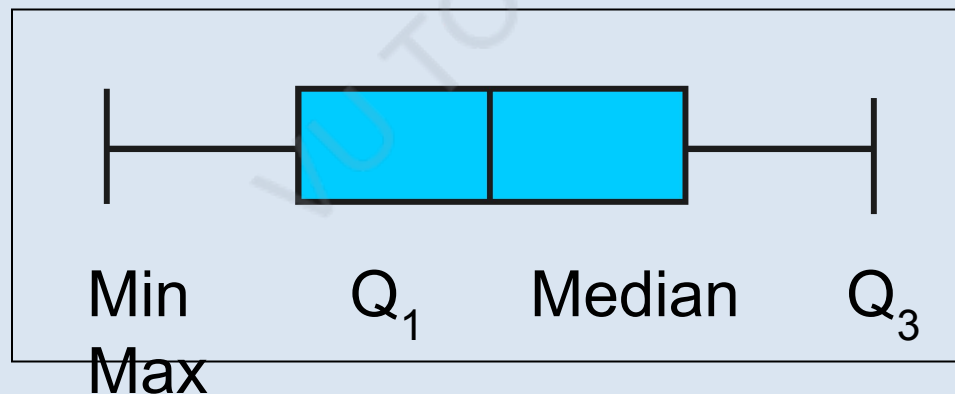
- Box-and-Whisker Plot: A Graphical display of data using 5-number summary:



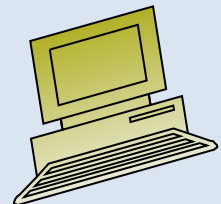
Tap Here To Join Our
Official Community

Shape of Box-and-Whisker Plots

- The Box and central line are centered between the endpoints if data are symmetric around the median

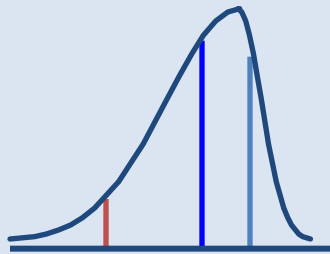


- A Box-and-Whisker plot can be shown in either vertical or horizontal format



Distribution Shape and Box-and-Whisker Plot

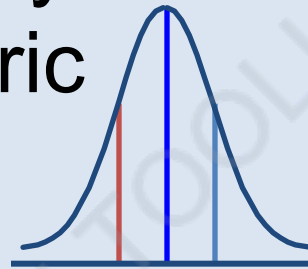
Left-Skewed



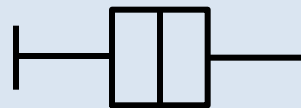
Q 1 Q 2 Q 3



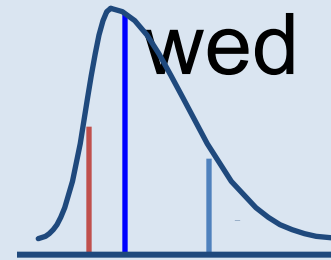
Symmetric



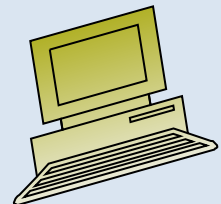
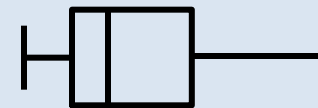
Q 1 Q 2 Q 3



Right-Skewed



Q 1 Q 2 Q 3

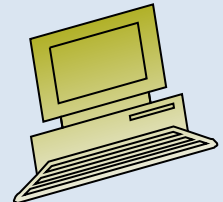


Tap Here To Join Our Official Community

Lec 4 - Graphs.pptx

Lecture 4

WU TOOLKIT



Tap Here To Join Our
Official Community Pakistan

Objectives

After completing this chapter, you should be able to:

- Construct and interpret a frequency distribution
- Construct a histogram
- Create and interpret bar charts, pie charts,
- Present and interpret categorical data in bar charts and pie charts

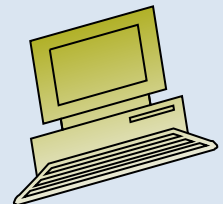


Tap Here To Join Our
Official Community

Topics to be covered

- Frequency distribution
- Histogram
- Bar charts
- Pie charts

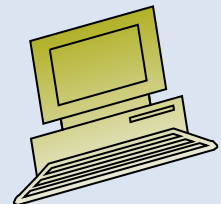
VU TOOLKIT



Tap Here To Join Our
Official Community

Organizing and Presenting Data Graphically

- Data in raw form are usually not easy to use for decision making
 - Some type of organization is needed
 - Table
 - Graph
- Techniques reviewed here:
 - Frequency Distributions and Histograms
 - Bar charts and pie charts



Class Intervals and Class Boundaries

- Each class grouping has the same width

- Determine the width of each interval

$$\text{Width of interval} \cong \frac{\text{range}}{\text{number of desired class groupings}}$$

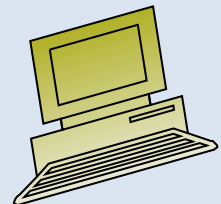
- Use at least 5 but no more than 15 groupings
- Class boundaries never overlap
- Round up the interval width to get desirable endpoint



Frequency Distribution Example

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

**24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27**



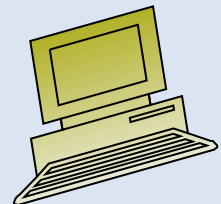
Frequency Distribution Example

(continued)

- Sort raw data in ascending order:
- 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 15)
- Compute class interval (width): 10 ($46/5$ then round up)
- Determine class boundaries (limits): 10, 20, 30, 40, 50, 60
- Compute class midpoints: 15, 25, 35, 45, 55
- Count observations and assign to classes



Tap Here To Join Our
Official Community



Frequency Distribution

Example

(continue
d)

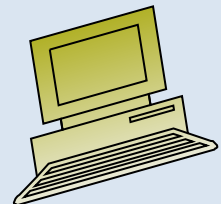
Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43,
44, 46, 53, 58

Class	Relative Frequency	Percentage
10 but less than 20	3	.15
20 but less than 30	6	.30
30 but less than 40	5	.25
40 but less than 50	4	.20
50 but less than 60	2	.10

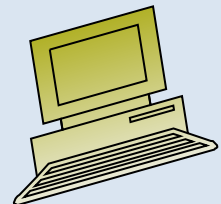


Tap Here To Join Our
Official Community



Graphing Numerical Data: The Histogram

- A graph of the data in a frequency distribution is called a **histogram**
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis
- The vertical axis is either **frequency, relative frequency, or percentage**
- Bars of the appropriate heights are used to represent the number of observations within each class

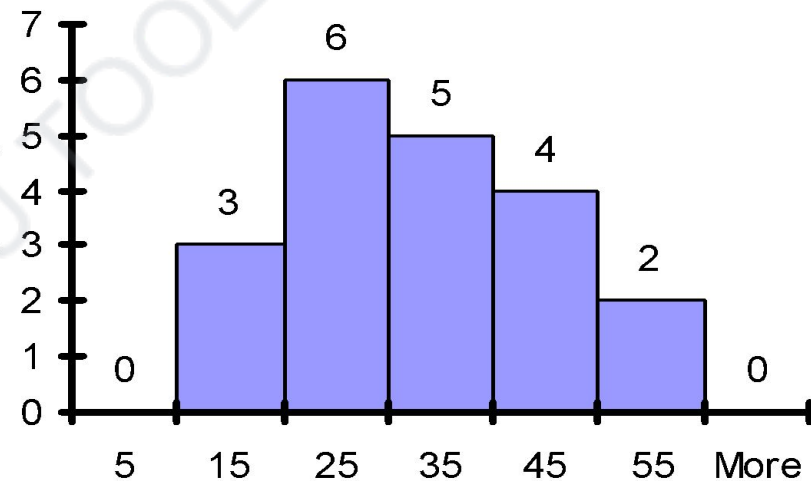


Tap Here To Join Our
Official Community

Histogram Example

Class	Class Midpoint	Frequency
10 but less than 20	20	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

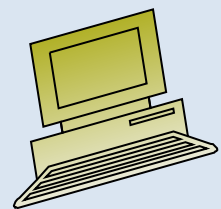
Histogram: Daily High Temperature



(No gaps between bars)

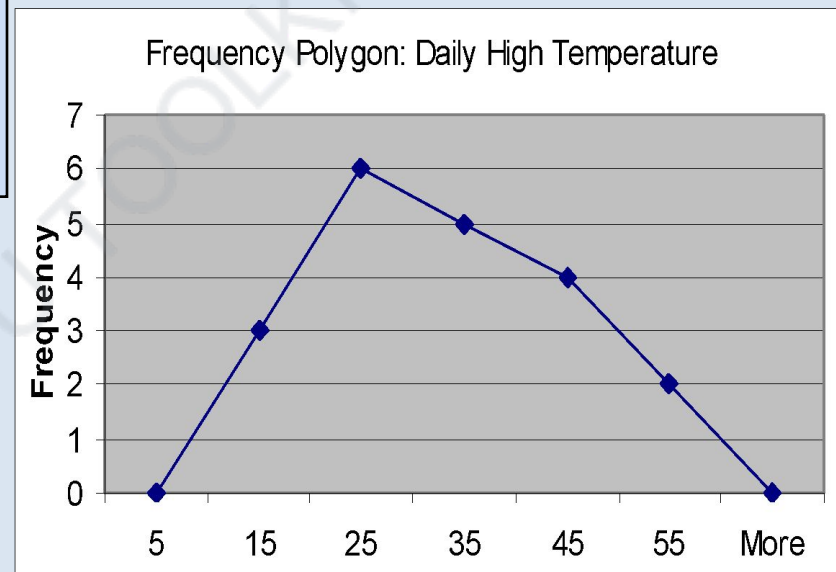
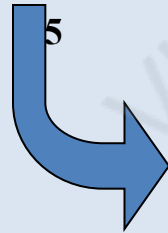


Class
Midpoints



Graphing Numerical Data: The Frequency Polygon

Class	Class Midpoint	Frequency
10 but less than 20 3	20	15
20 but less than 30 6	25	25
30 but less than 40 5	35	35
40 but less than 50 4	45	45
50 but less than 60 2	55	5

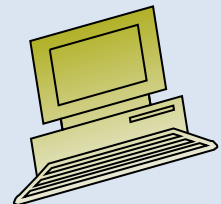


(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



Tap Here To Join Our Official Community

Class
Midpoints



Tabulating Numerical Data: Cumulative Frequency

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43,
44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 15	3	15	3	15
20 but less than 45	6	30	9	45
30 but less than 70	5	25	14	70
40 but less than 90	4	20	18	90

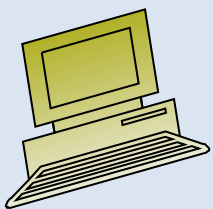
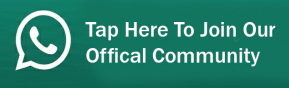
50 but less than 60

2

10

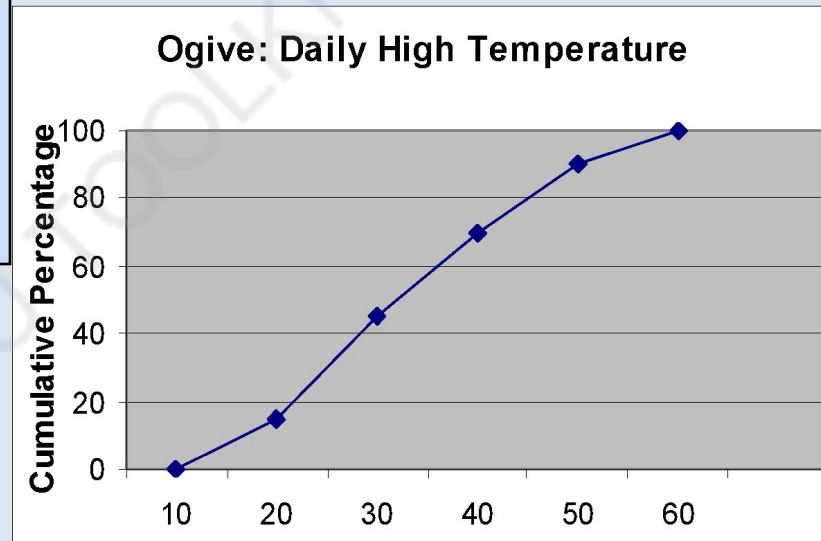
10

20



Graphing Cumulative Frequencies: The Ogive (Cumulative % Polygon)

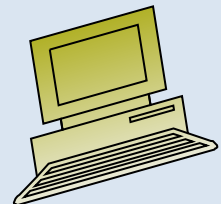
Class	Lower class bound	Cumulative Percentage
Less than 10	0	0
10 but less than 20	10	20
20 but less than 30	20	30
30 but less than 40	30	40
40 but less than 50	40	50
50 but less than 60	50	60



Class Boundaries (*Not*
limits)



Tap Here To Join Our
Official Community



Bar and Pie Charts

- Bar charts and Pie charts are often used for qualitative (category) data
- Height of bar or size of pie slice shows the frequency or percentage for each category



Tap Here To Join Our
Official Community

Examples

	Number of Candies
Blue	23
Green	35
Red	56
Orange	54

This spreadsheet has one data series in a column: Number of candies

	Bag #1	Bag #2	Bag #3
Blue	23	33	43
Green	35	22	24
Red	56	19	56
Orange	54	20	44

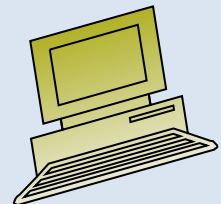
This spreadsheet has three data series in columns: Bag #1, Bag #2, and Bag #3

OR

This spreadsheet has four data series in rows: Blue,



Tap Here To Join Our Official Community



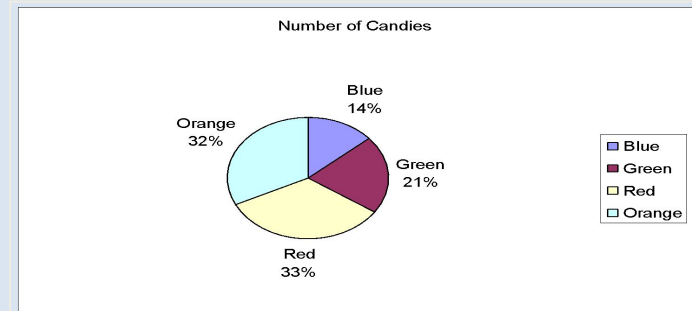
Pie Chart

- Shows the proportional size of items that make up a data series to the sum of the items
- Pie charts have only one data series

The title comes from this cell.

Each of the numbers become a slice of the pie.

The legend and labels come from the first column.

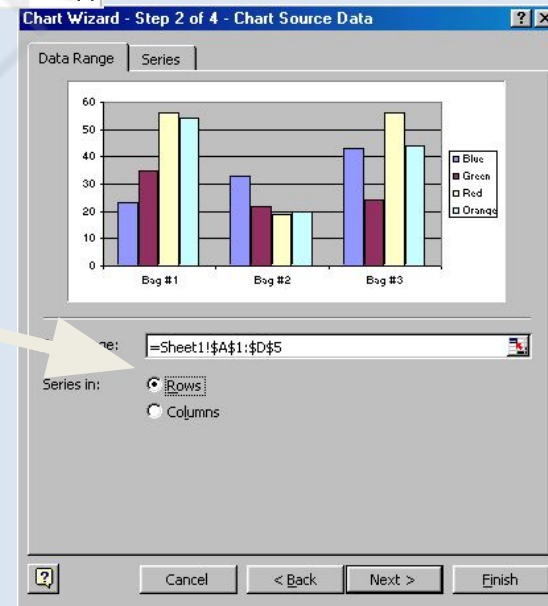
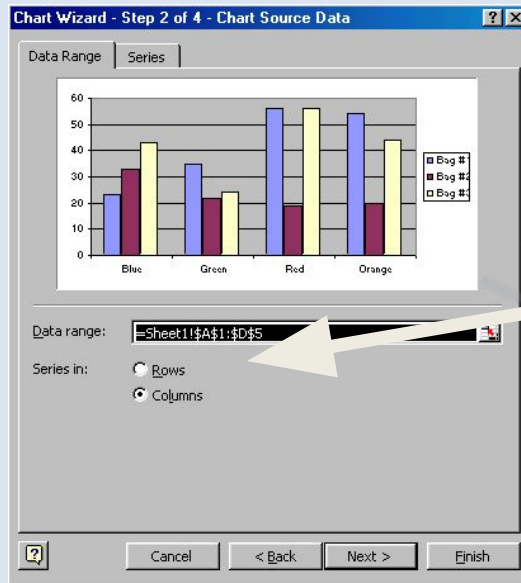


Tap Here To Join Our Official Community



Multiple Bar Chart #1

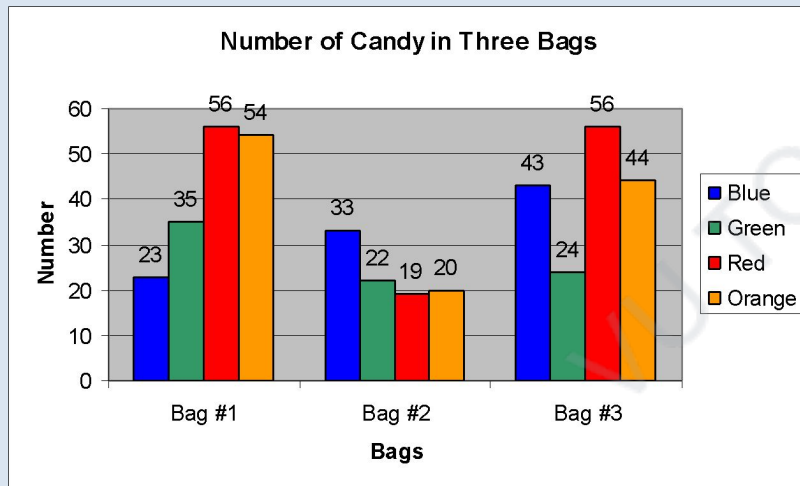
	Bag #1	Bag #2	Bag #3
Blue	23	33	43
Green	35	22	24
Red	56	19	56
Orange	54	20	44



The same set of data can yield two different bar charts if the data series are in rows or

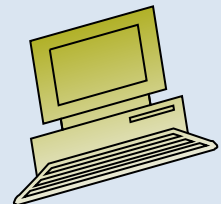


Multiple Bar Chart #2



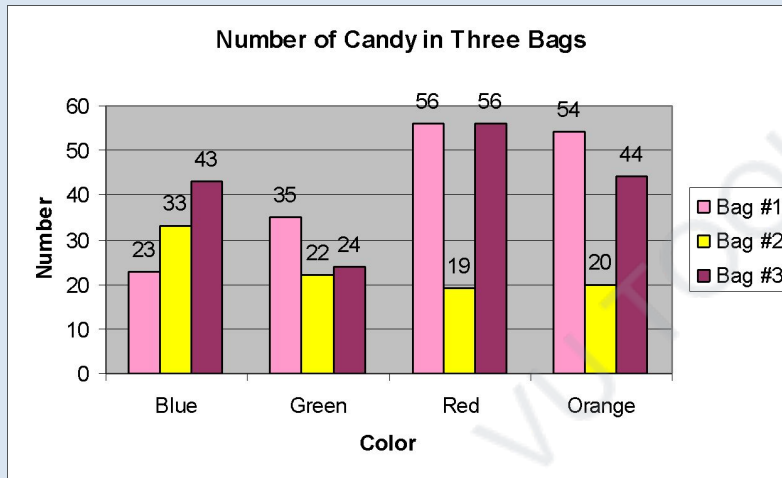
Data Set in Rows

- Each group of columns represents one bag.
- Each column in the group is a color.
- The colors are the DATA SET. They appear in the legend.



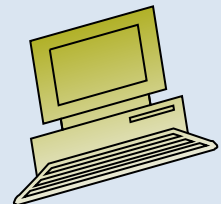
Tap Here To Join Our
Official Community

Multiple Bar Chart #3



Data Set in Columns

- Each group of columns represents one color.
- Each column in the group is a bag.
- The bags are the DATA SET. They appear in the legend.



Tap Here To Join Our
Official Community

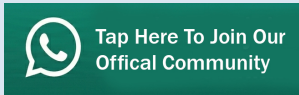
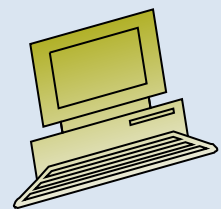
Lecture Summary

- Data in raw form are usually not easy to use for decision making -- Some type of organization is needed:
 - ◆ Table◆ Graph
- Techniques reviewed in this chapter:
 - Frequency distributions and histograms
 - Percentage polygons and ogives
 - Scatter diagrams for bivariate data
 - Bar charts, pie charts,



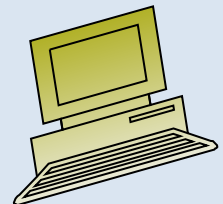
Practical Exercise in Data Preparation

WU TOOKKIT



Practical Exercise

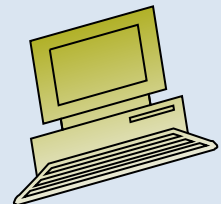
VU TOOLKIT



Lec 5 - SPC Charts.pptx

Lecture 5

WU TOOLKIT



Tap Here To Join Our
Official Community Pakistan

Control Charts

originally developed by

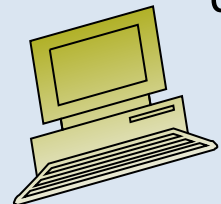
Walter A. Shewhart



Tap Here To Join Our
Official Community

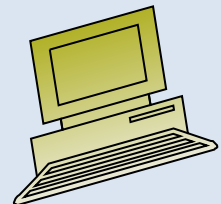
History - beginnings

- The control chart - invented by Walter Shewhart
- How to improve the reliability of telephony transmission systems?
- Shewhart framed the problem in terms of common and special causes of variation
- In 1924, May 16, in an internal memo Shewhart introduced the control chart as a tool for distinguishing between the two causes of variation.
- Shewhart's boss, George Edwards: "Dr. Shewhart prepared a little memorandum only about a page in length... all of the essential principles and considerations which are involved in what we know today as process quality control."



Control Charts

The *control chart* is a **statistical quality control** tool used in the monitoring variation in the characteristics of a product or service



Tap Here To Join Our
Official Community

Control Charts

The control chart focuses on the time dimension and the nature of the variability in the system.



Control Charts

The control chart may be used to study past performance and/or to evaluate present conditions.



Tap Here To Join Our
Official Community

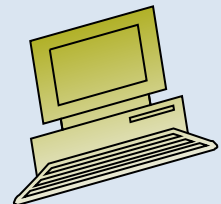
Control Charts

Data collected from a control chart may form the basis for process improvement.



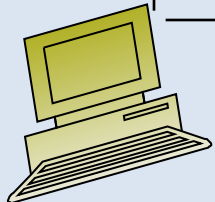
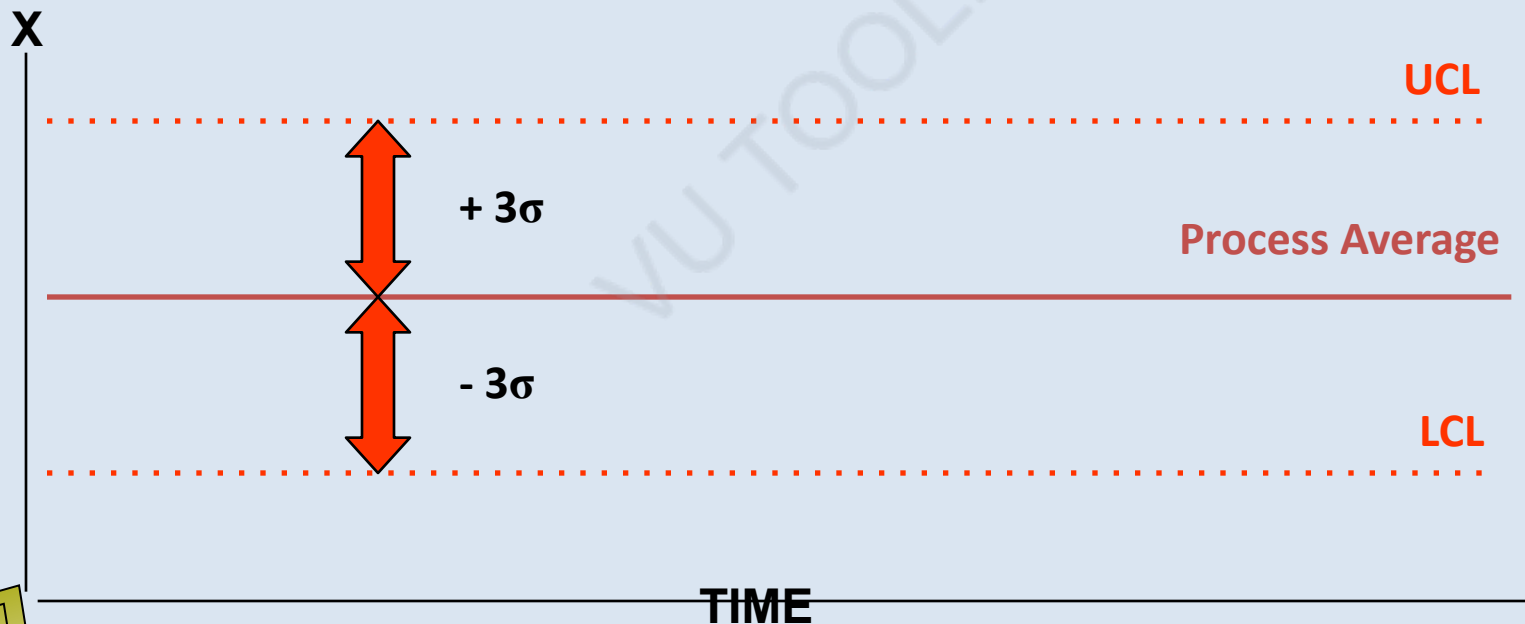
Tap Here To Join Our
Official Community

- Control chart - a graph used in SPC that shows whether a sample of data falls within the normal range of variation.
- Control chart: central line (CL), upper (UCL) and lower control limits (LCL).
 - Control limits separate common from assignable causes of variation
- Control charts for variables
 - monitor characteristics that can be measured
- Control charts for attributes
 - monitor characteristics that can be counted



Control Charts

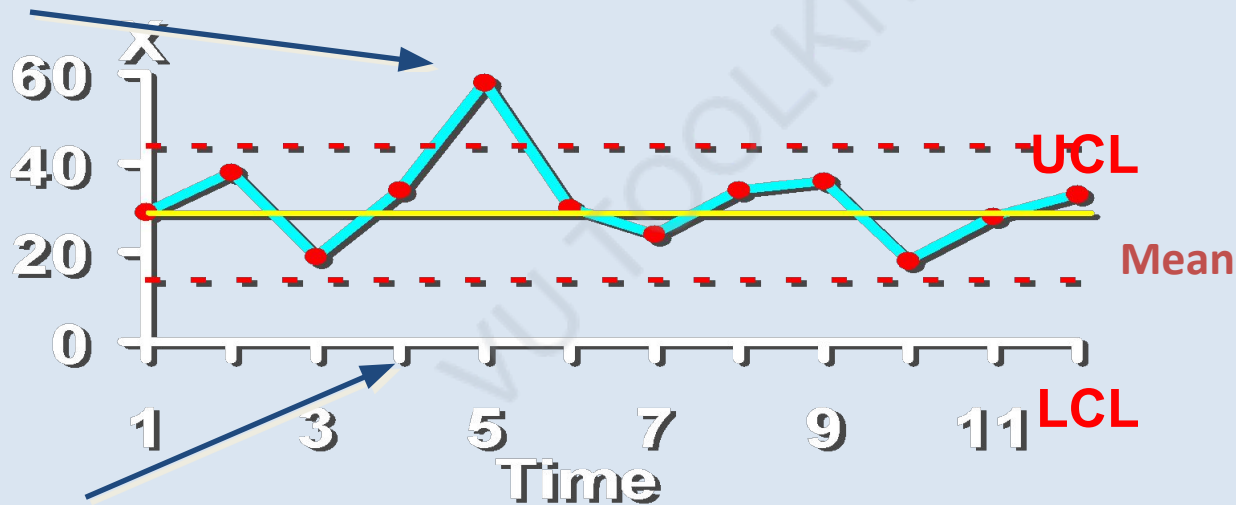
- **UCL** = Process Average + 3 Standard Deviations
- **LCL** = Process Average - 3 Standard Deviations



Control Charts

Graph of sample data plotted over time

• Assignable Cause Variation



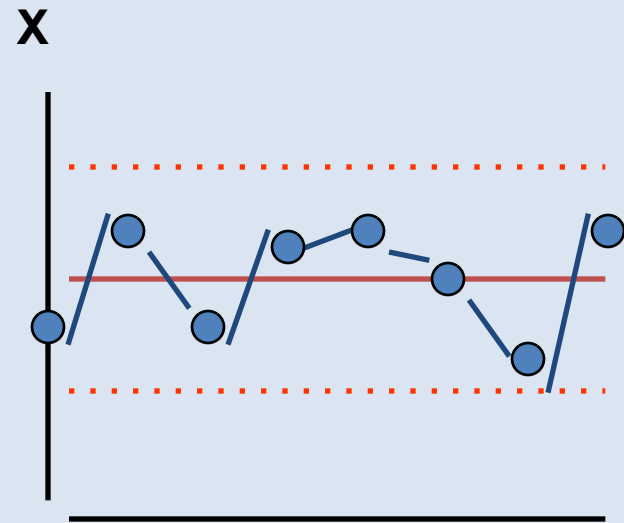
Process Average $\pm 3\sigma$

Random Variation

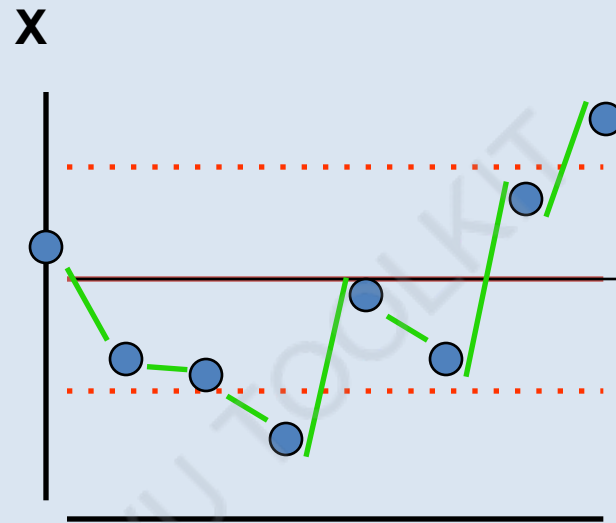


Tap Here To Join Our Official Community

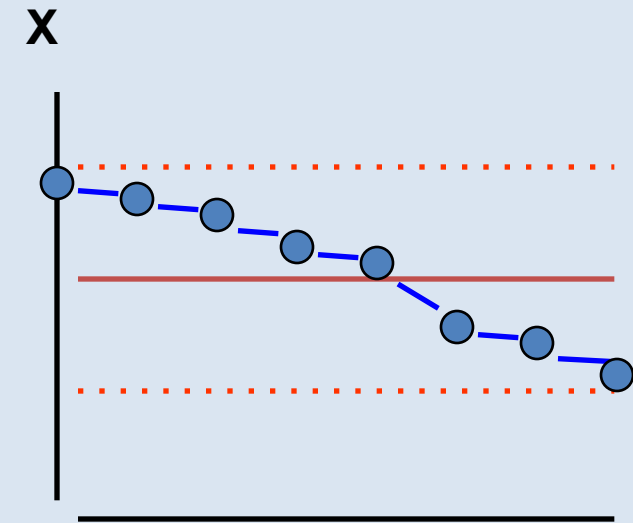
Control Charts



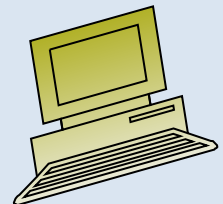
Common Cause Variation: no points outside control limit



Special Cause Variation: two points outside control limit



Downward Pattern: no points outside control limit; however, eight or more points in trend



Control Charts

- Attribute charts
- Variables charts



Tap Here To Join Our
Official Community

Control Charts

Charts may
be used for
categorical
variables.
[i.e.: attributes]



Tap Here To Join Our
Official Community

Control Charts

Whenever a character of interest is measured on a nominative or an interval, i.e.: categorical, scale...an *attributes* control chart is used to monitor a process.



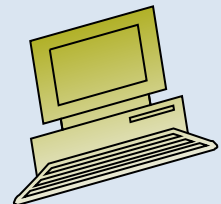
Tap Here To Join Our
Official Community

Control Charts

- **Attributes Control Charts**

counts [c-chart]

proportion [p-charts]



Tap Here To Join Our
Official Community

Control Charts

- **Attributes Control Charts**

when sample size are **not constant**

and/or are **unknown**

use counts charts

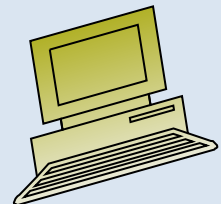
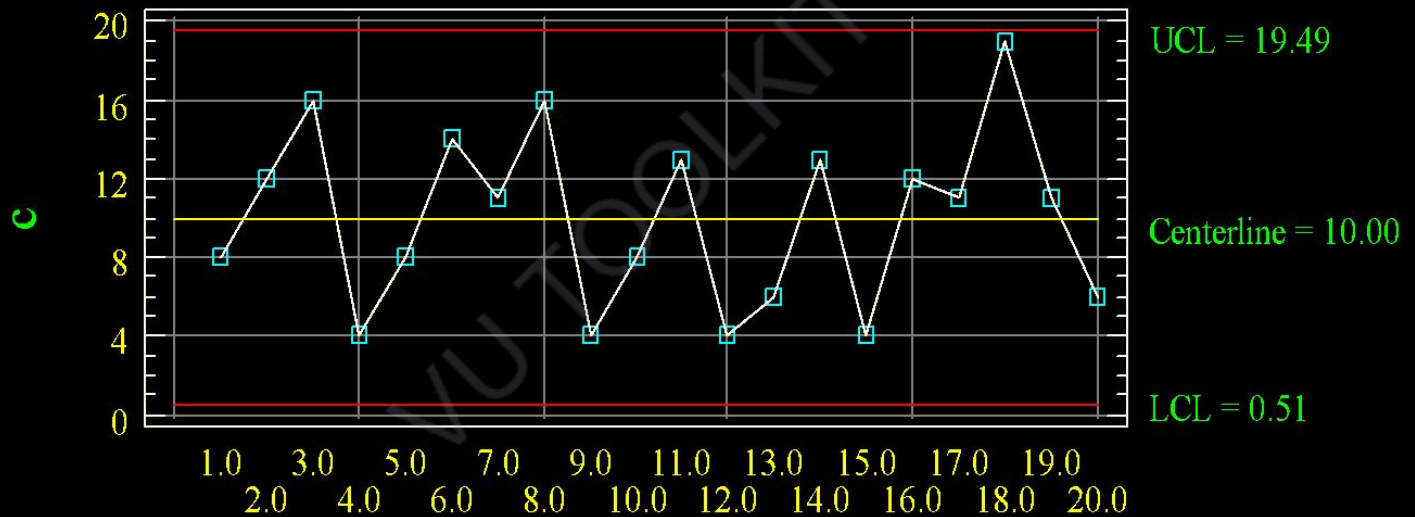
[c-charts]



Tap Here To Join Our
Official Community

Control Charts

c Chart for num_defect



Tap Here To Join Our
Official Community

Control Charts

- **Attributes Control Charts**

when sample size are **constant** and
are **known**

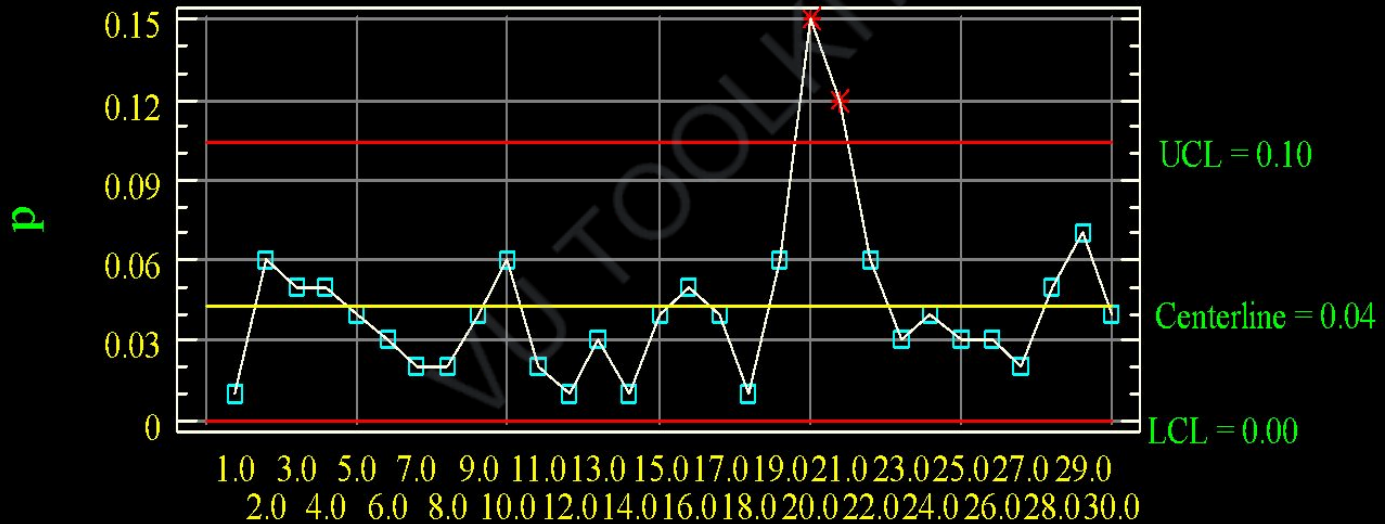
use proportion charts
[p-charts]



Tap Here To Join Our
Official Community

Control Charts

p Chart for defects/100



Tap Here To Join Our Official Community

Control Charts



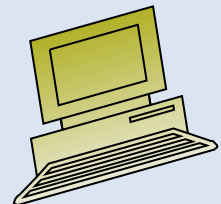
Charts may
be used for
interval or
ratio data [i.e.:
variables]



Tap Here To Join Our
Official Community

Control Charts

Whenever a character of interest is measured on an interval or a ratio scale, a *variables* control chart is used to monitor a process.



Control Charts

- **Variables Control Charts**

Mean and Range charts

[\bar{x} -bar & R charts]



Tap Here To Join Our
Official Community

Control Charts

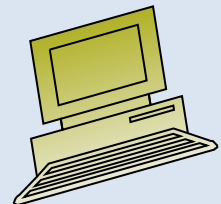
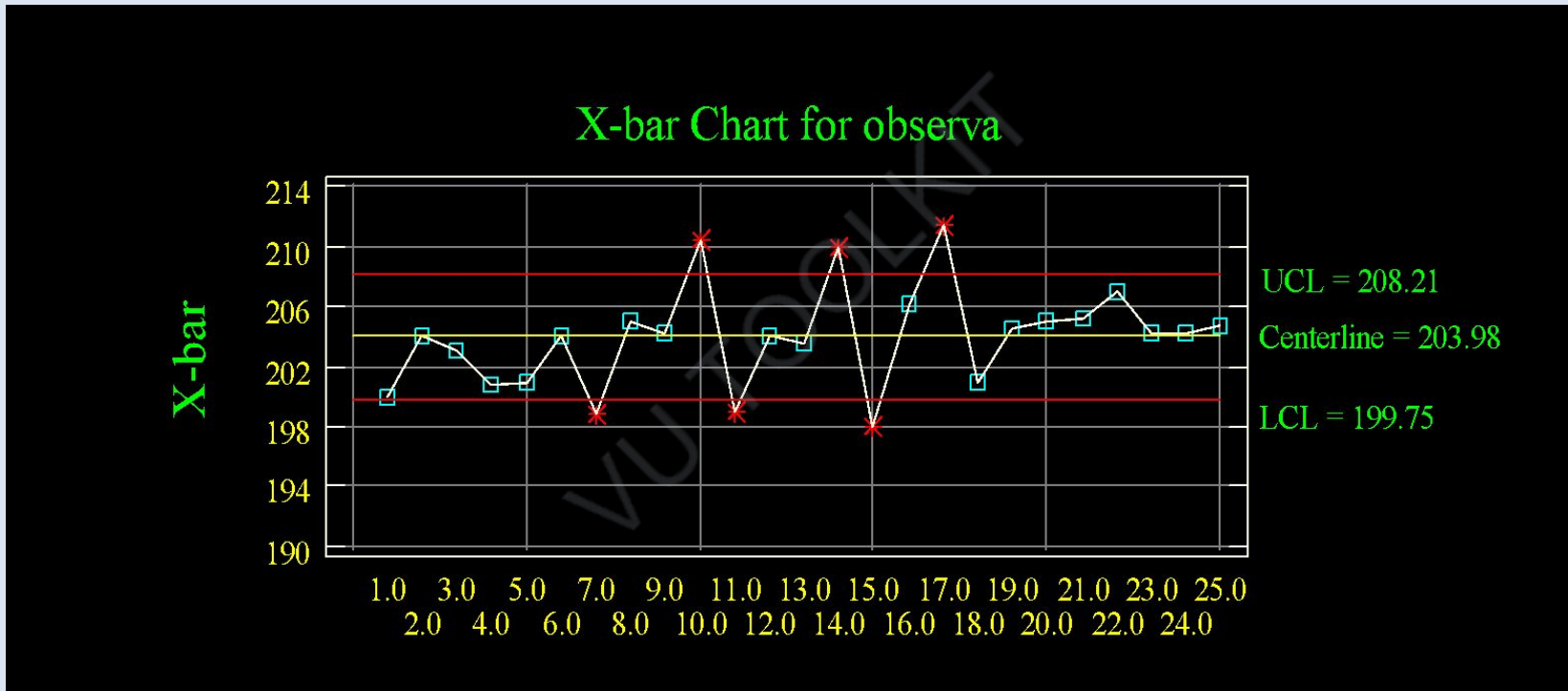
Variables control charts are typically used in pairs.


..... one chart monitors *process average* while the other monitors the *variation* in a process.



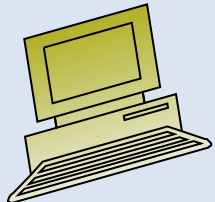
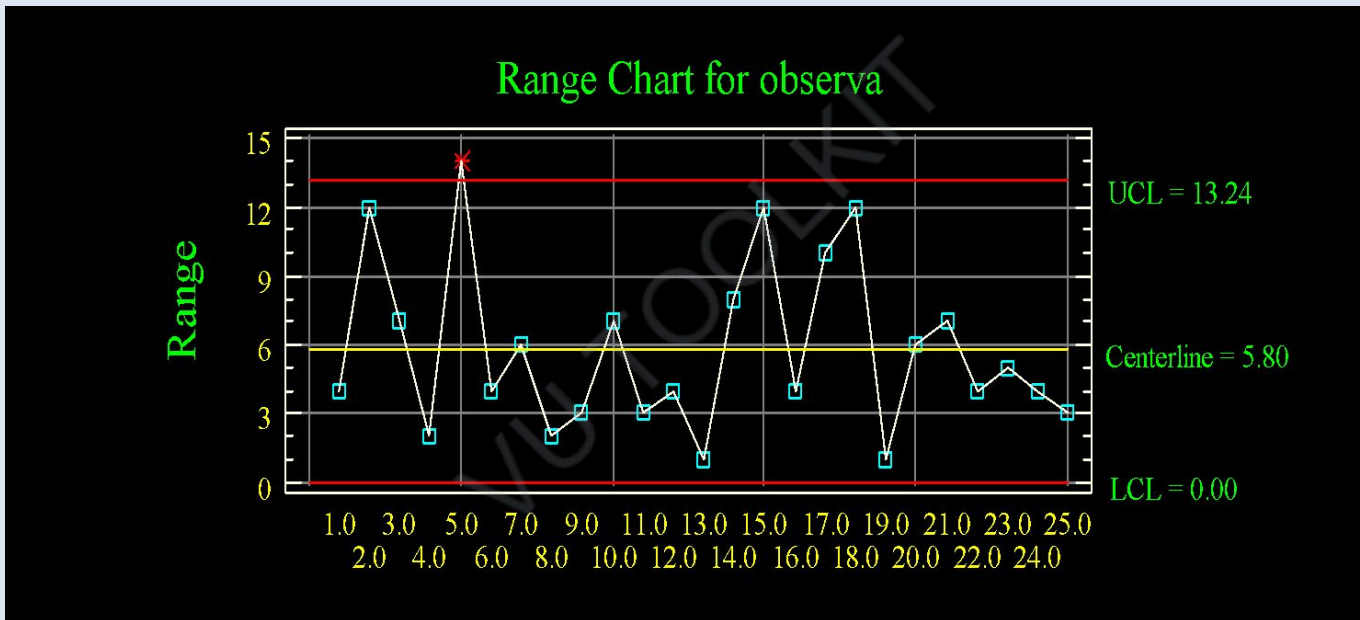
Tap Here To Join Our
Official Community

Control Charts



 Tap Here To Join Our Official Community

Control Charts

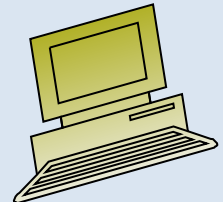


Tap Here To Join Our
Official Community

Average Run Length (ARL)

- Expected number of samples taken before shift is detected is called the Average Run Length (ARL)

$$ARL = \sum_{r=1}^{\infty} r \beta^{r-1} (1 - \beta) = \frac{1}{(1 - \beta)}$$



Performance of Any Shewhart Control Chart

- In-Control ARL:
 - Average number of points plotted on control chart before a false alarm occurs (ideally, should be large)

$$ARL_0 = \frac{1}{\alpha}$$

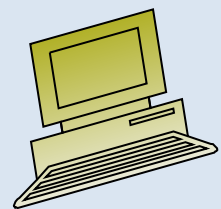
- Out-of-Control ARL:
 - Average number of points, after the process goes out-of-control, before the control chart detects it (ideally, should be small)



$$ARL_1 = \frac{1}{1 - \beta}$$



WU TOOLKIT

Practice



Vir  Tap Here To Join Our Official Community  an

Lec 6 - Regression and Correlation.ppt

Lecture 6

WU TOOLKIT



Introduction to Linear Regression and Correlation Analysis

VU TOOLKIT



Tap Here To Join Our
Official Community

Goals

After this, you should be able to:

- Calculate and interpret the simple correlation between two variables
- Determine whether the correlation is significant
- Calculate and interpret the simple linear regression equation for a set of data
- Understand the assumptions behind regression analysis
- Determine whether a regression model is significant



Tap Here To Join Our
Official Community

Goals

(continued)

After this, you should be able to:

- Calculate and interpret confidence intervals for the regression coefficients
- Recognize regression analysis applications for purposes of prediction and description
- Recognize some potential problems if regression analysis is used incorrectly
- Recognize nonlinear relationships between two variables



Tap Here To Join Our
Official Community

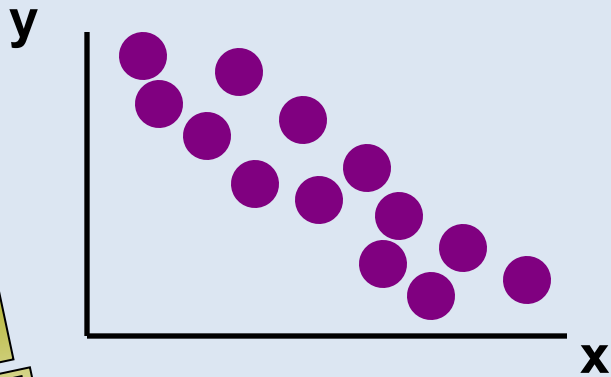
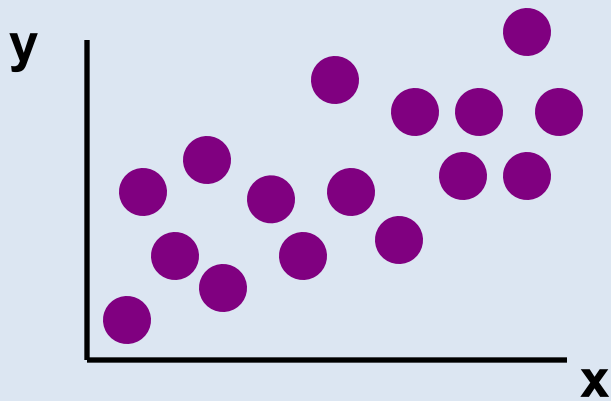
Scatter Plots and Correlation

- A **scatter plot** (or scatter diagram) is used to show the relationship between two variables
- **Correlation** analysis is used to measure strength of the association (linear relationship) between two variables
 - Only concerned with strength of the relationship
 - No causal effect is implied

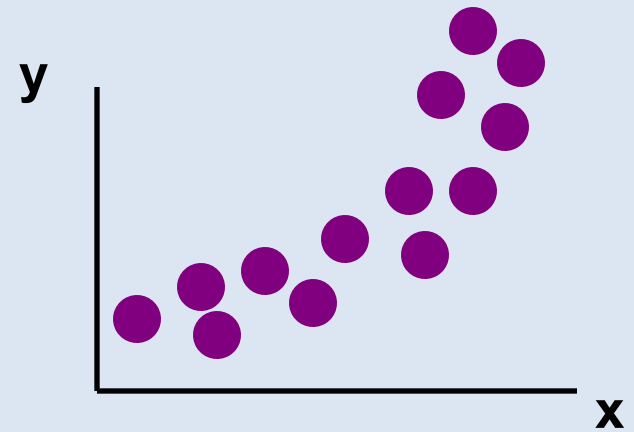
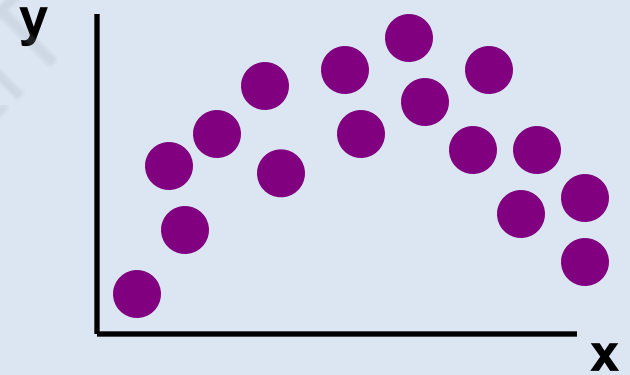


Scatter Plot Examples

Linear relationships



Curvilinear relationships



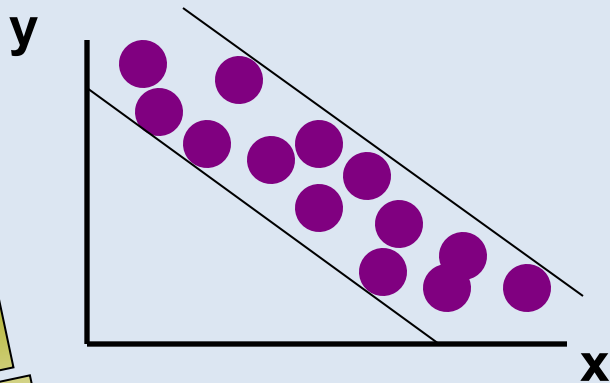
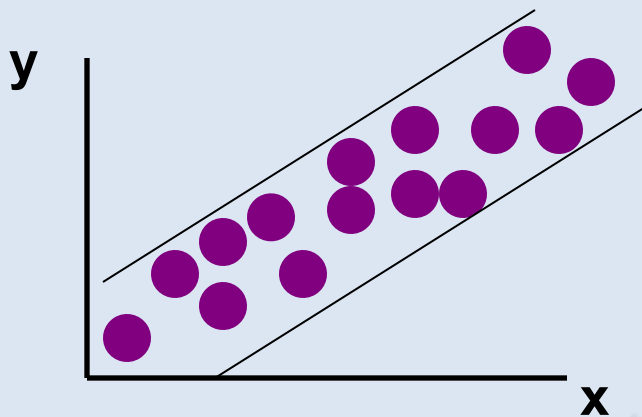
Tap Here To Join Our
Official Community



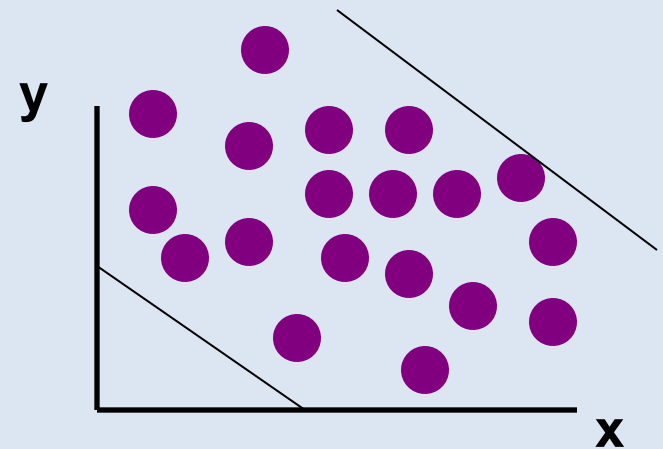
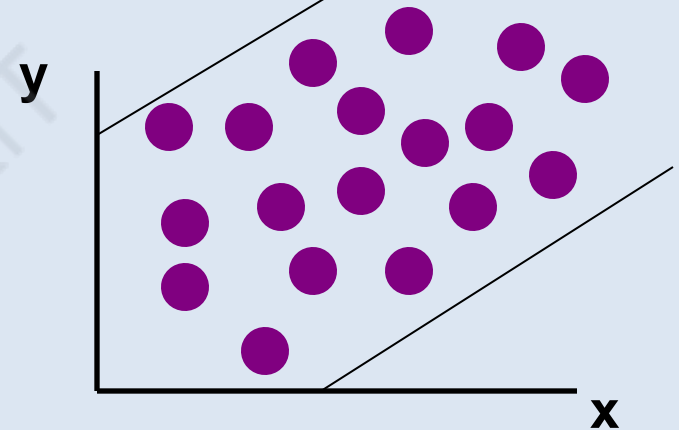
Scatter Plot Examples

(continued)

Strong relationships



Weak relationships



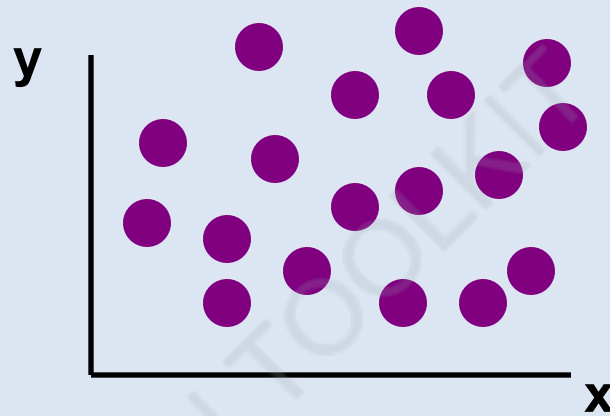
Tap Here To Join Our
Official Community



Scatter Plot Examples

(continued)

No relationship



Tap Here To Join Our
Official Community



Correlation Coefficient

(continued)

- The **population correlation coefficient ρ** (rho) measures the strength of the association between the variables
- The **sample correlation coefficient r** is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations



Tap Here To Join Our
Official Community

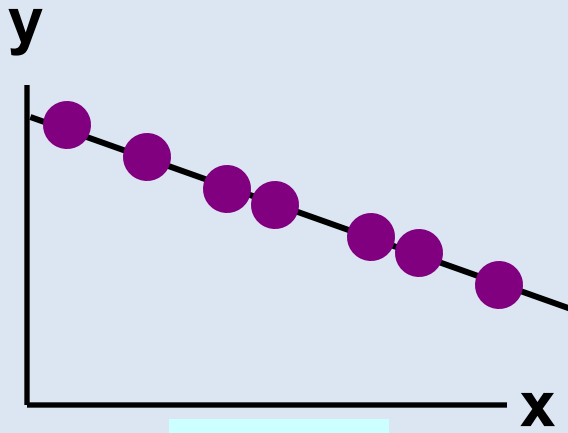
Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

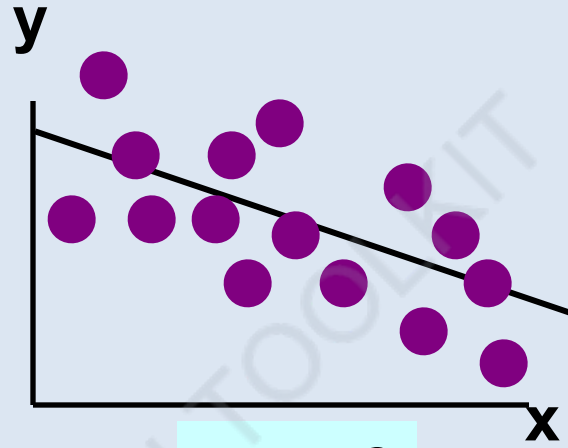


Tap Here To Join Our
Official Community

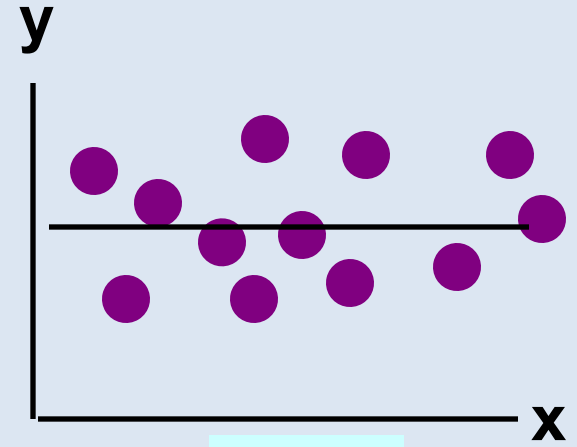
Examples of Approximate r Values



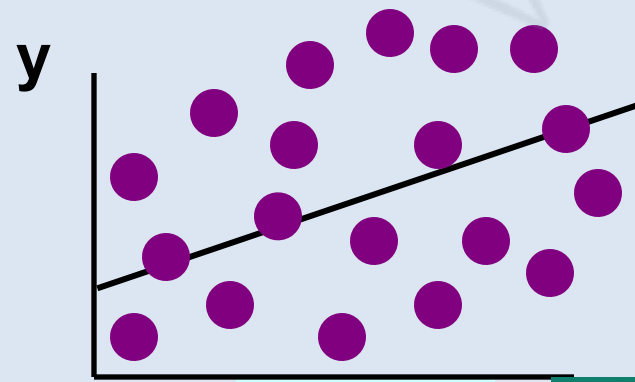
$r = -1$



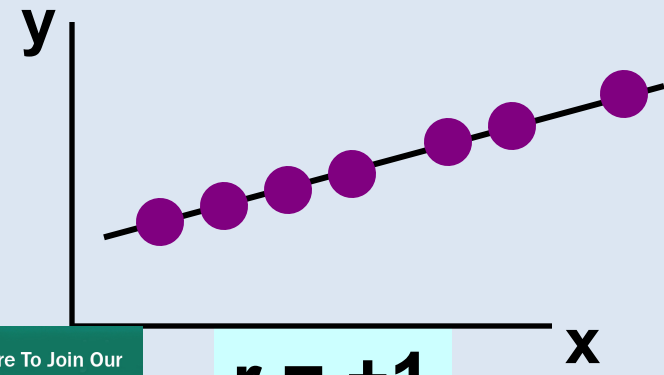
$r = -.6$




$r = 0$



$r = +.3$



$r = +1$

 Tap Here To Join Our Official Community



Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

r = Sample correlation coefficient

n = Sample size

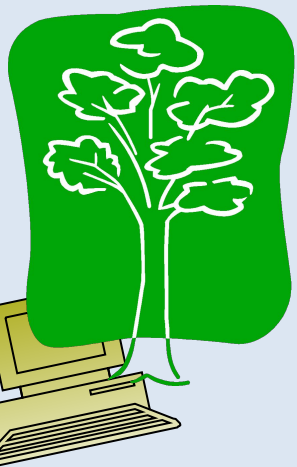
x = Value of the independent variable

y = Value of the dependent variable



Calculation Example

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\Sigma=321$	$\Sigma=142$	$\Sigma=14111$	$\Sigma=713$	

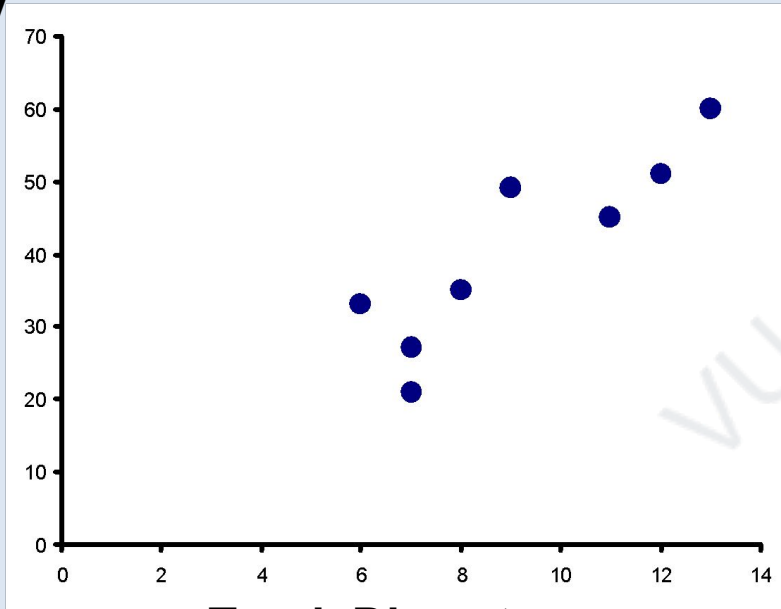



 Tap Here To Join Our Official Community

Calculation Example

(continued)

Tree Height, y



Trunk Diameter, x

$$\begin{aligned} r &= \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}} \\ &= 0.886 \end{aligned}$$

$r = 0.886 \rightarrow$ relatively strong positive linear association between x and y



Tap Here To Join Our Official Community

Excel Output

Excel Correlation Output

Tools / data analysis / correlation...

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between
Tree Height and Trunk Diameter



Tap Here To Join Our
Official Community



Significance Test for Correlation

- Hypotheses

$$H_0: \rho = 0 \text{ (no correlation)}$$

$$H_A: \rho \neq 0 \text{ (correlation exists)}$$

- Test statistic

$$- t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ (with } n-2 \text{ degrees of freedom)}$$



Tap Here To Join Our
Official Community



Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

$H_0: \rho = 0$ (No correlation)

$H_1: \rho \neq 0$ (correlation exists)

$$\alpha = .05, \quad df = 8 - 2 = 6$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$



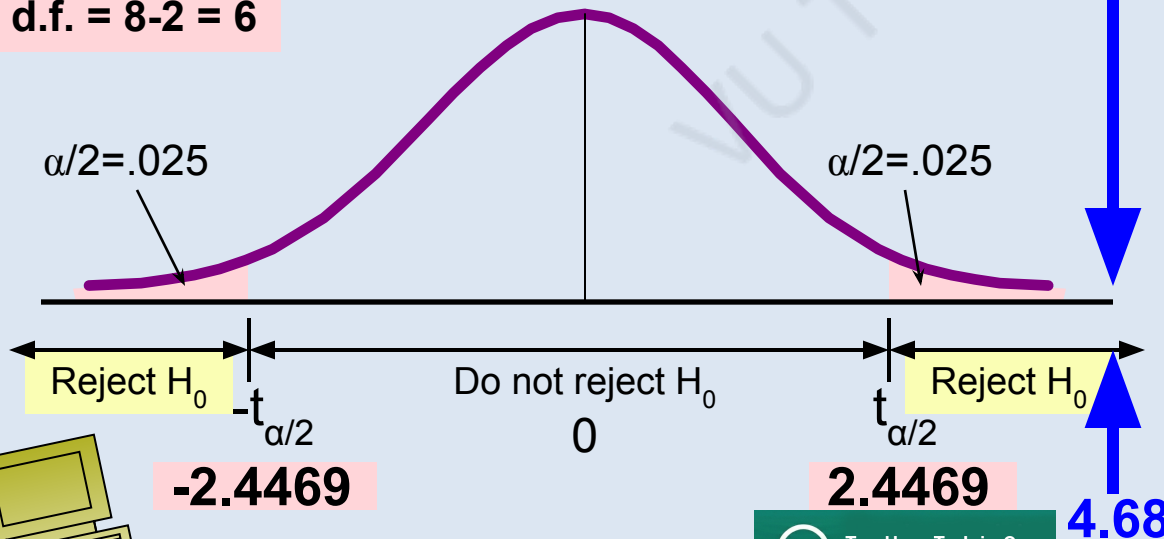
Tap Here To Join Our Official Community



Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

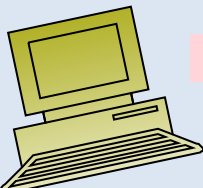
d.f. = 8-2 = 6



Decision:
Reject H_0

Conclusion:
There is **evidence** of a linear relationship at the 5% level of significance

Tap Here To Join Our Official Community



Introduction to Regression Analysis

- **Regression analysis** is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable



Tap Here To Join Our
Official Community

Simple Linear Regression Model

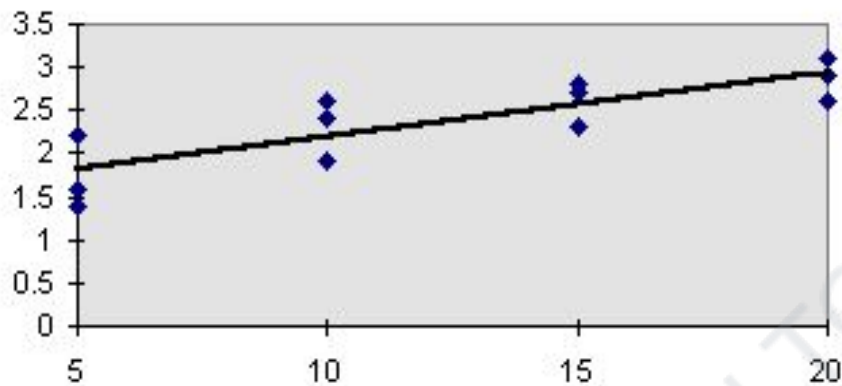
- Only **one independent variable**, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x



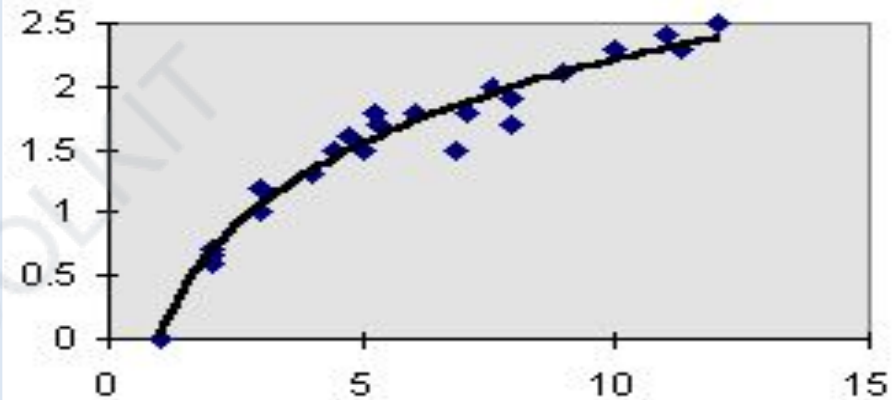
Tap Here To Join Our
Official Community

Types of Regression Models

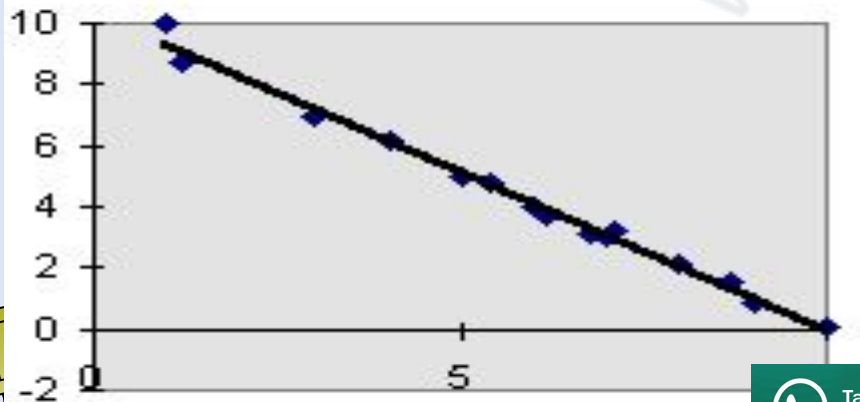
Positive Linear Relationship



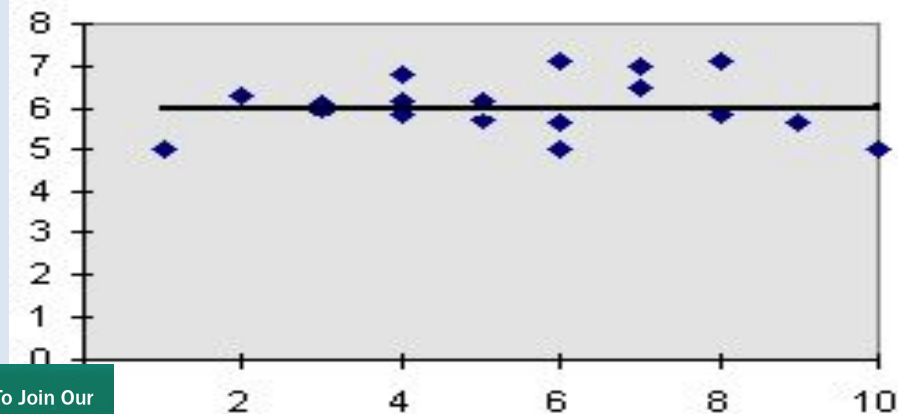
Relationship NOT Linear



Negative Linear Relationship

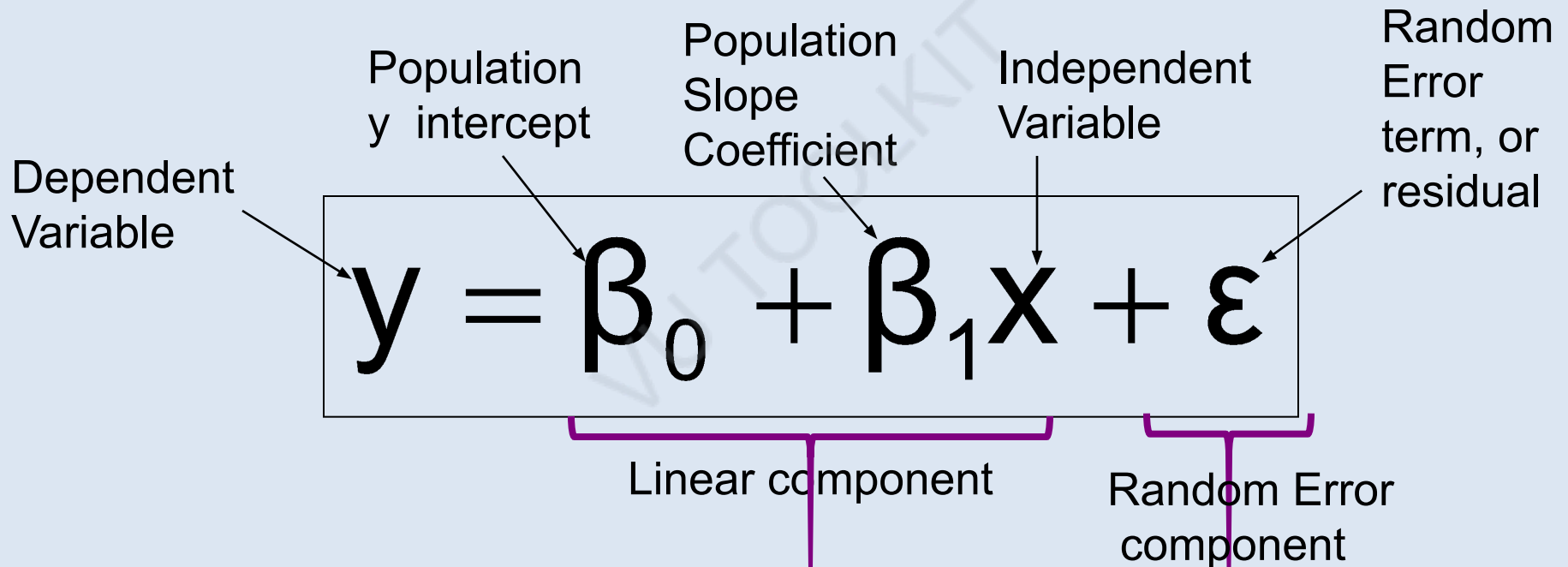


No Relationship



Population Linear Regression

The population regression model:



Tap Here To Join Our
Official Community

Linear Regression Assumptions

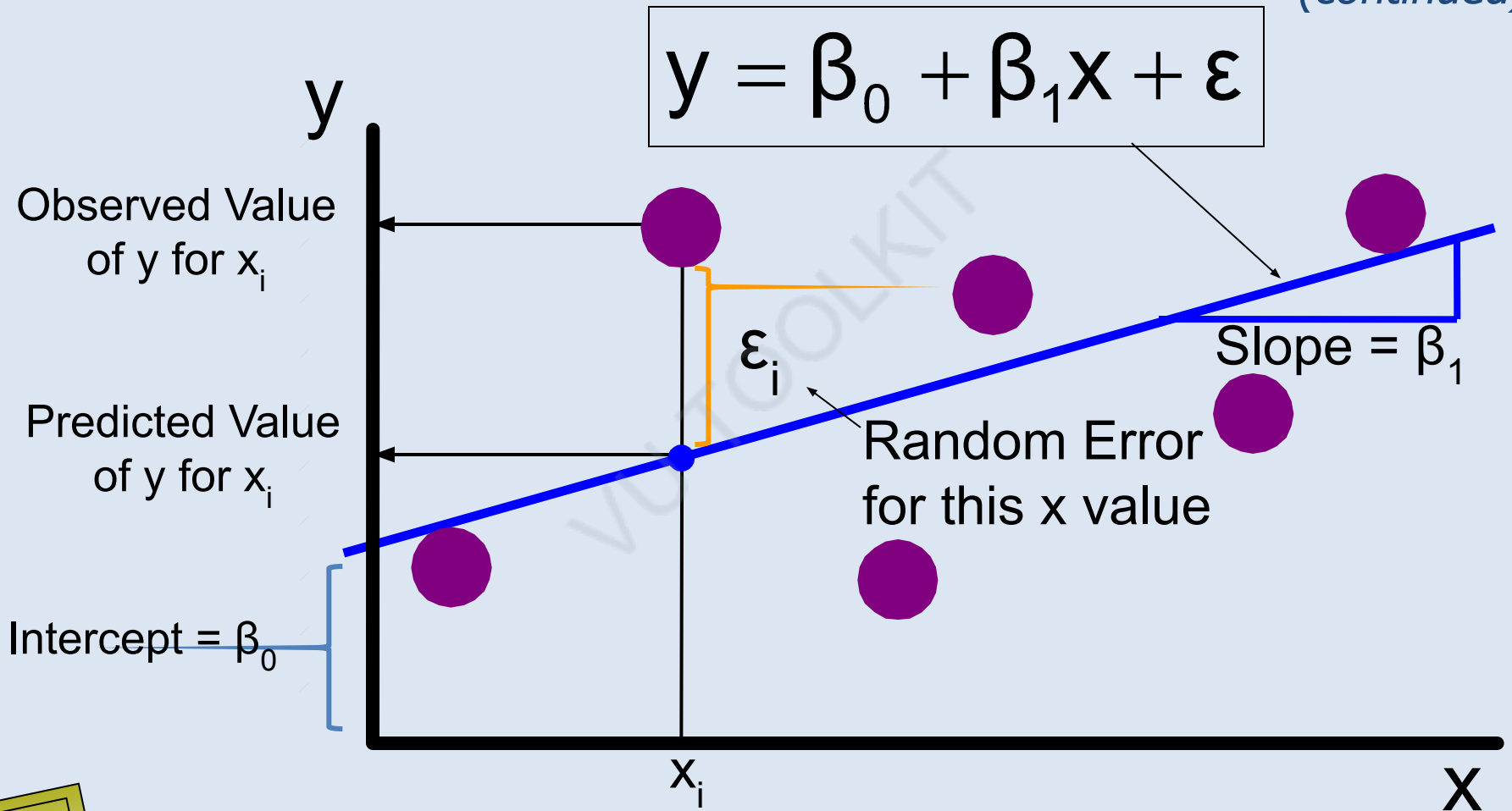
- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear



Tap Here To Join Our
Official Community

Population Linear Regression

(continued)



Tap Here To Join Our Official Community

Estimated Regression Model

The sample regression line provides an **estimate** of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero



Tap Here To Join Our
Official Community

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared residuals**

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1x))^2\end{aligned}$$



The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

algebraic

equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$



Tap Here To Join Our
Official Community



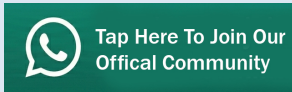
Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x



Finding the Least Squares Equation

- The coefficients b_0 and b_1 will usually be found using computer software, such as Excel
- Other regression measures will also be computed as part of computer-based regression analysis



Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
 - Dependent variable (y) = house price in \$1000s
 - Independent variable (x) = square feet




Tap Here To Join Our
Official Community

Sample Data for House Price Model

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



 Tap Here To Join Our Official Community



Regression Using Excel

- Tools / Data Analysis / Regression

Microsoft Excel - 13data.xls

File Edit View Insert Format Tools Data Window Help Acrobat

Chart 1

	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

Regression

Input

Input Y Range: \$A\$1:\$A\$11

Input X Range: \$B\$1:\$B\$11

Labels Constant is Zero

Confidence Level: 95 %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

OK Cancel Help



Tap Here To Join Our
Official Community

Excel Output

Regression Statistics

Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

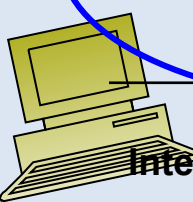
The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.934	11.084	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

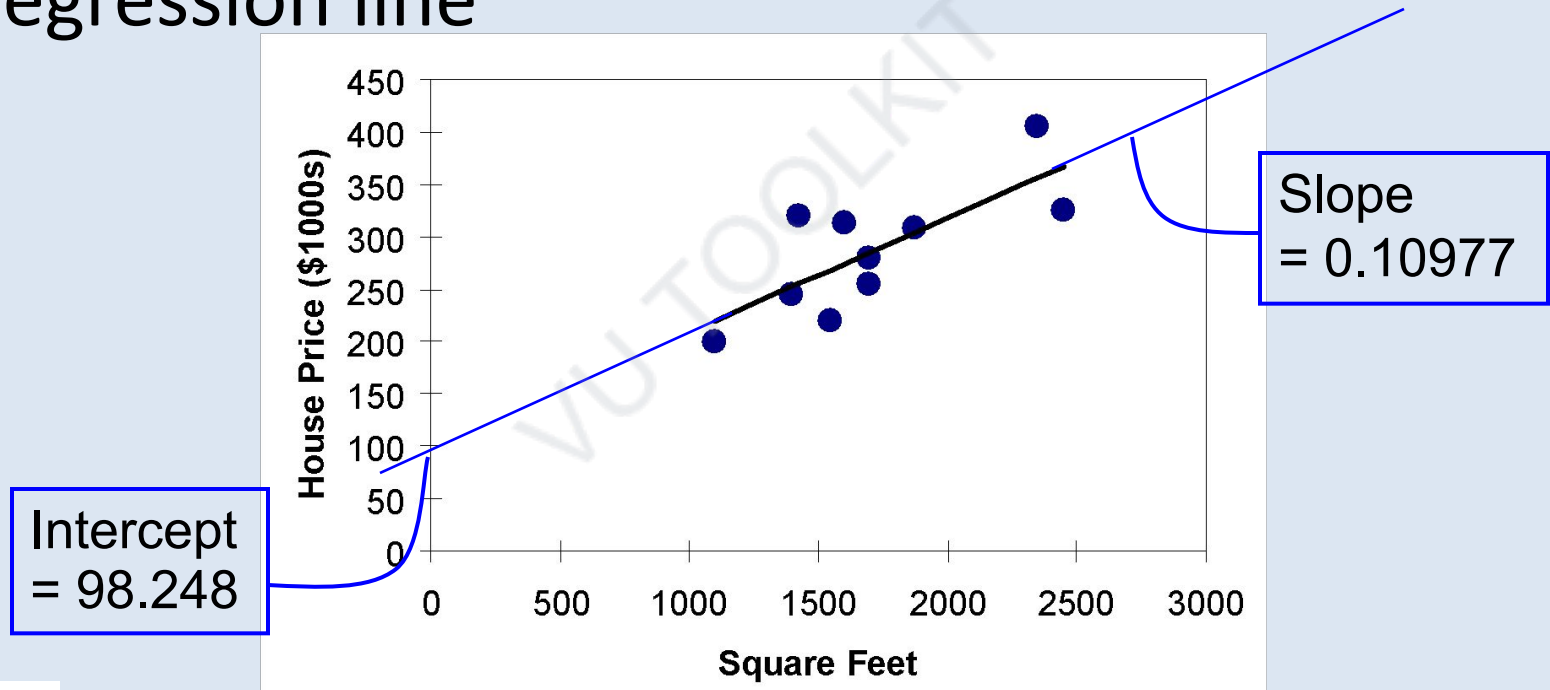
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	9296	0.01289	2	-35.57720	232.0738

Tap Here To Join Our Official Community



Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$



Tap Here To Join Our
Official Community

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if $x = 0$ is in the range of observed x values)

– Here, no houses had 0 square feet, so $b_0 =$
98.24833 just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet



Tap Here To Join Our
Official Community



Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X

– Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(\$1000) = \109.77 , on average, for each additional one square foot



Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 () $\sum (y - \hat{y}) = 0$
- The sum of the squared residuals is a minimum (minimized) $\sum (y - \hat{y})^2$
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased

estimates of β_0 and



Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum
of Squares

Sum of
Squares Error

Sum of
Squares

Regression

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

$$SSR = \sum (\hat{y} - \bar{y})^2$$

where:

\bar{y} = Average value of the dependent variable

y = Observed values of the dependent variable

\hat{y} = Estimated value of y for the given x value



Tap Here To Join Our
Official Community



Explained and Unexplained Variation

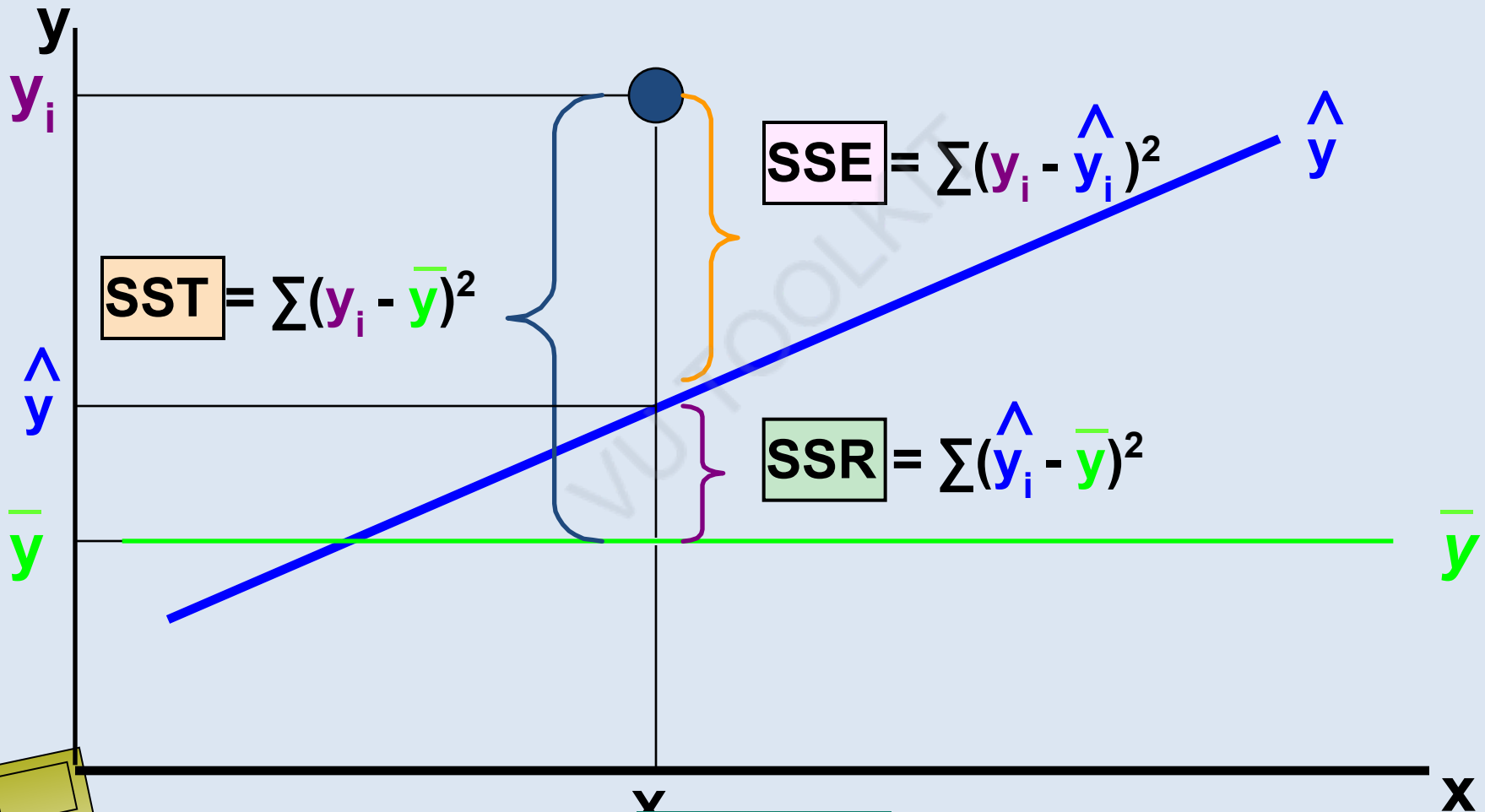
(continued)

- SST = total sum of squares
 - Measures the variation of the y_i values around their mean y
 - SSE = error sum of squares
 - Variation attributable to factors other than the relationship between x and y
- SSR = regression sum of squares
 - Explained variation attributable to the relationship between x and y



Explained and Unexplained Variation

(continued)



 Tap Here To Join Our Official Community



Coefficient of Determination, R^2

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **R-squared** and is denoted as R^2

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

where

$$0 \leq R^2 \leq 1$$



Coefficient of Determination, R^2

(continued)

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Note: In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

where:

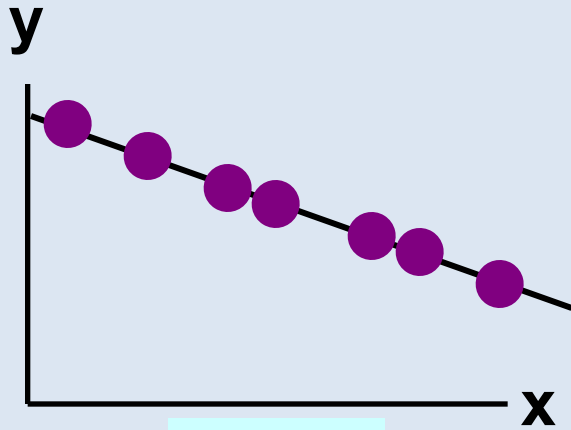
R^2 = Coefficient of determination

r = Simple correlation coefficient



Tap Here To Join Our
Official Community

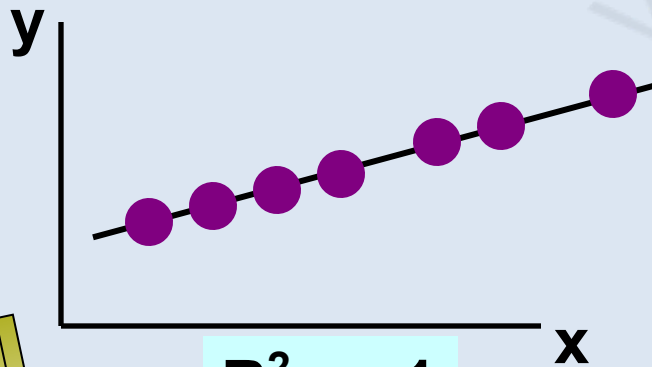
Examples of Approximate R^2 Values



$$R^2 = 1$$

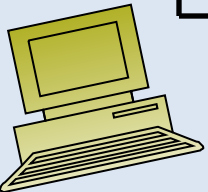
$$R^2 = 1$$

Perfect linear relationship between x and y:



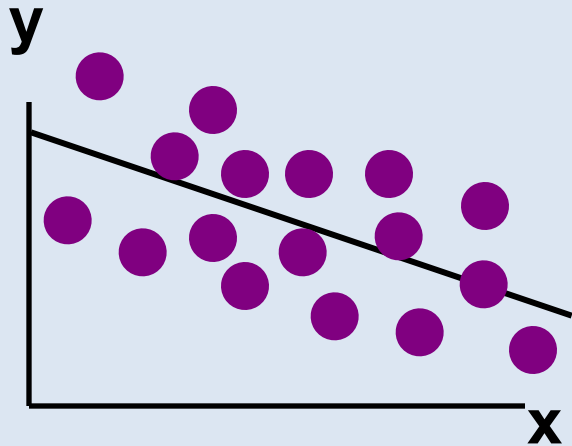
$$R^2 = +1$$

100% of the variation in y is explained by variation in x



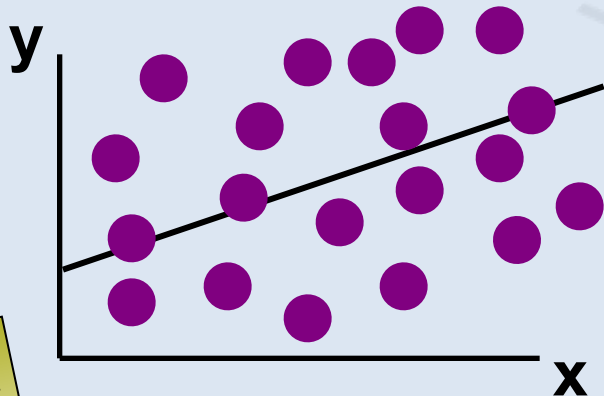
Tap Here To Join Our
Official Community

Examples of Approximate R^2 Values



$$0 < R^2 < 1$$

Weaker linear relationship between x and y:

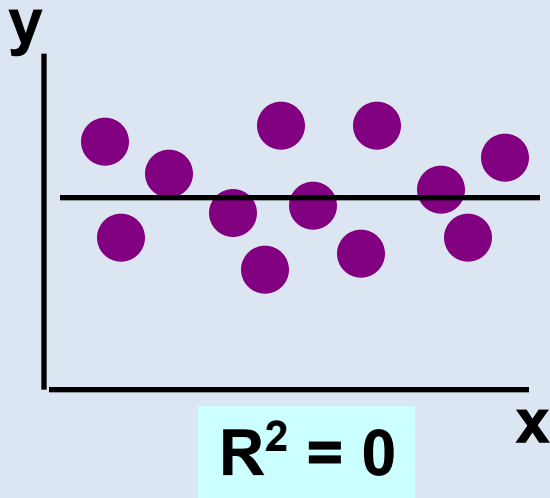


Some but not all of the variation in y is explained by variation in x



Tap Here To Join Our
Official Community

Examples of Approximate R^2 Values



$$R^2 = 0$$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)



Tap Here To Join Our
Official Community

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n - k - 1}}$$

Where

SSE = Sum of squares error

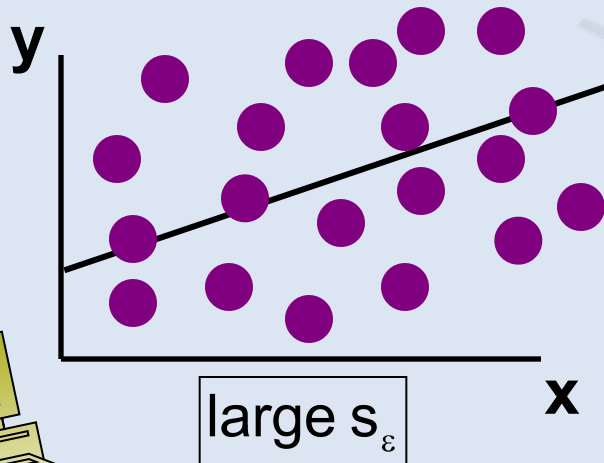
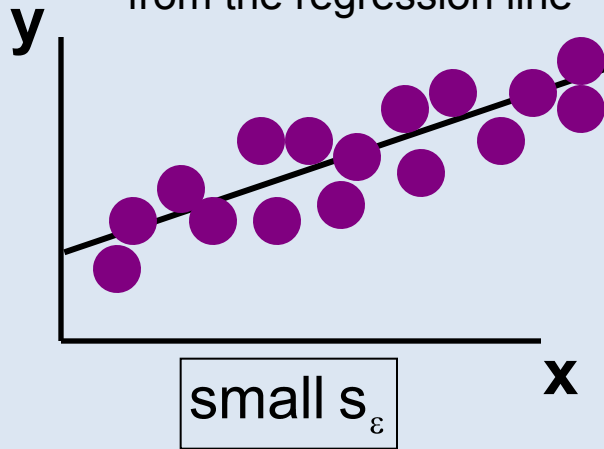
n = Sample size

k = number of independent variables in the model

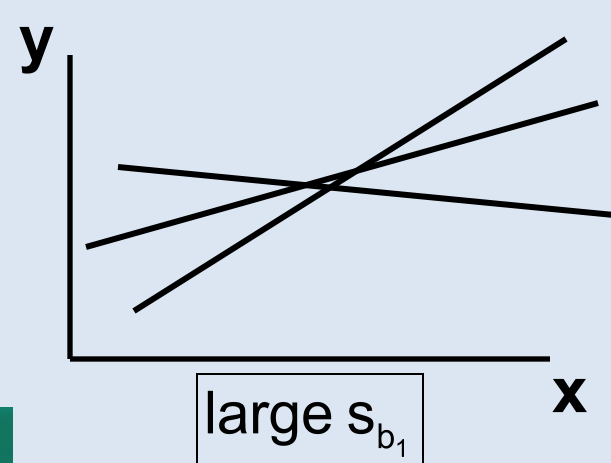
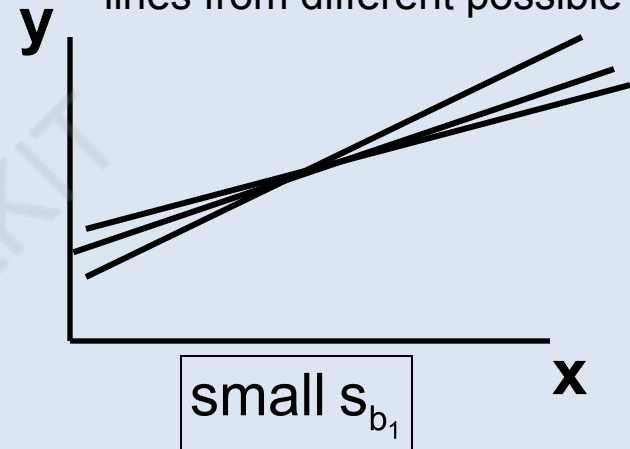


Comparing Standard Errors

Variation of observed y values from the regression line



Variation in the slope of regression lines from different possible samples



Tap Here To Join Our Official Community



Inference about the Slope: t Test

- t test for a population slope
 - Is there a linear relationship between x and y?
- Null and alternative hypotheses
 - $H_0: \beta_1 = 0$ (no linear relationship)
 - $H_1: \beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b_1} = Estimator of the standard error of the slope

d.f. = n - 2



Tap Here To Join Our
Official Community



Inference about the Slope: t Test

(continued)

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098 (\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house
affect its sales price?



Tap Here To Join Our
Official Community

Inferences about the Slope: t Test Example

Test Statistic: **t = 3.329**

$$H_0: \beta_1 = 0$$

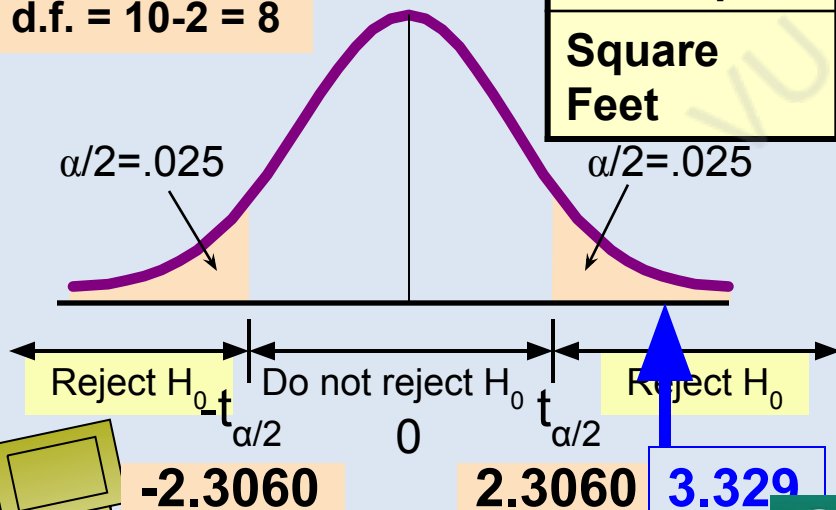
$$H_A: \beta_1 \neq 0$$

From Excel output:

	Coefficient s	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

b_1 s_{b_1} t

d.f. = 10 - 2 = 8



Reject H_0

There is sufficient evidence that square footage affects house price

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

$$d.f. = n - 2$$

Excel Printout for House Prices:

	<i>Coefficient s</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)



Tap Here To Join Our
Official Community

Regression Analysis for Description

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is \$1000, we are 95% confident that the average impact on sales price is between \$33.70 and \$185.80 per square foot of house size

This 95% confidence interval **does not include 0**.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance



Tap Here To Join Our
Official Community

Lec 7 - Hypothesis Testing.pptx

Lecture 7

WU TOOLKIT



- Random Numbers
- Checking Normality
- Hypothesis testing

VU TOOLKIT



Tap Here To Join Our
Official Community

HYPOTHESIS TESTING

VU TOOLKIT



Tap Here To Join Our
Official Community

HYPOTHESIS

What do you mean by a Hypothesis?

A supposition or explanation that is provisionally accepted in order to interpret certain events or phenomena, and to provide guidance for further investigation.

OR

A hypothesis is a tentative statement about the relationship between two or more variables.



Tap Here To Join Our
Official Community

HYPOTHESIS TESTING

Statistically, an assumption about certain characteristics of a population.

The statistical procedure for testing a hypothesis requires some understanding of the null hypothesis and alternative hypothesis.



Tap Here To Join Our
Official Community

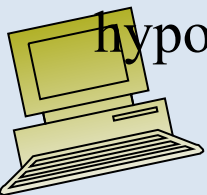
TYPES OF HYPOTHESIS

Null Hypothesis

A null hypothesis is "the hypothesis that there is no relationship between two or more variables. In a mathematical formulation of the null hypothesis there will typically be an equal sign. This hypothesis is denoted by H_0 .

Alternate Hypothesis

The alternative hypothesis proposes a relationship between two or more variables. In a mathematical formulation of the alternative hypothesis there will typically be an inequality, or not equal to symbol. This hypothesis is denoted by either H_a or by H_1 .



EXAMPLE OF HYPOTHESES

For example, you might have come up with a measurable hypothesis that children have a higher IQ if they eat oily fish for a period of time.

Null hypothesis: Children who eat oily fish for six months do not show a higher IQ.

Alternative hypothesis: Children who eat oily fish for six months will show a higher IQ



Tap Here To Join Our
Official Community

A common statistical method is to compare a population to the mean.

H_0 : The children will show no increase in mean intelligence.

i.e.

$$H_0 : \mu = 100$$

H_1 : The children will show an increase in mean intelligence.

i.e.

$$H_1 : \mu > 100$$



TYPES OF ERRORS

What are errors in Hypothesis Testing?

The purpose of Hypothesis Testing is to reject or not reject the Null Hypothesis based on statistical evidence

Hypothesis Testing is said to have resulted in an error when the decision regarding treatment of the Null Hypothesis is wrong

There are basically **two types of errors** we can make:

TYPE I Error

TYPE II Error



Tap Here To Join Our
Official Community

TYPES OF ERRORS

Type-I Error (Ho right but rejected)

When Null Hypothesis is rejected despite the test on data showing that the outcome was true

Type-II Error (Ho wrong but not rejected)

When Null Hypothesis is not rejected despite the test on data showing that the outcome was false



Tap Here To Join Our
Official Community

Type I Error

- A **type I error**, also known as an **error of the first kind**, occurs when the null hypothesis (H_0) is true, but is rejected.
- A type I error may be compared with a so called *false positive*.
- The rate of the type I error is called the *size* of the test and denoted by **the Greek letter α (alpha)**.
- It usually equals the **significance level of a test**.
- If type I error is fixed at 5 %, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true.



Type II Error

- **Type II error**, also known as an **error of the second kind**, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis **which should have been rejected**.
- A type II error may be compared with a so-called *False Negative*.
- The rate of the type II error is denoted by the **Greek letter β (beta)** and related to the power of a test (**which equals $1-\beta$**).



In the tabular form two error can be presented as follows:

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False decision	Correct True decision
Fail to reject null hypothesis	Correct True decision	Type II error False decision



Tap Here To Join Our
Official Community

Reducing Type I Errors

The chances of making a Type I error is reduced by increasing the level of confidence.

VU TOOLKIT



Tap Here To Join Our
Official Community

Reducing Type II Errors

Test condition and acceptance criteria are in turn reduces Type II errors. This increases the number of times we reject the Null hypothesis – **with a resulting increase *in the number of Type I errors.***

VU TOOKIT



Tap Here To Join Our
Official Community

Type II Error and Power

- “**Power**” of a test is the probability of rejecting null when alternative is true.
- “**Power**” = $1 - P(\text{Type II error})$
- To minimize the $P(\text{Type II error})$, we equivalently want to maximize power.
- But power depends on the value under the alternative hypothesis .



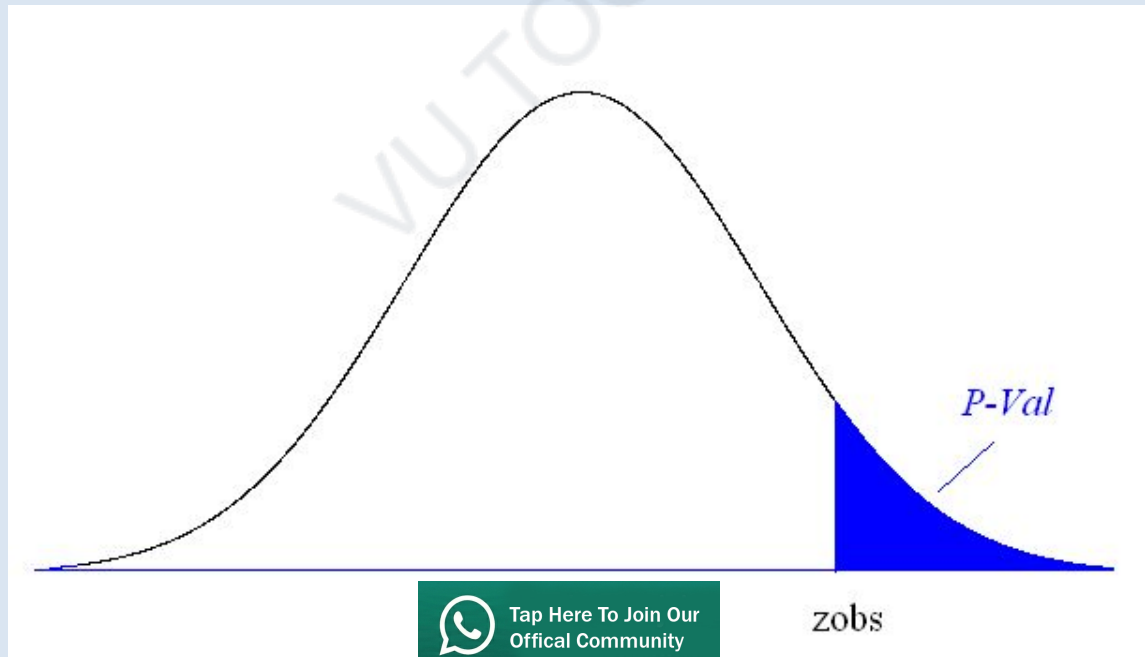
Tap Here To Join Our
Official Community

P-value

P-value - Measure of the strength of evidence the sample data provides against the null hypothesis:

$P(\text{Evidence This strong or stronger against } H_0 \mid H_0 \text{ is true})$

$$P\text{-val} : p = P(Z \geq z_{obs})$$



INTERPRETING RESULTS

Interpreting the weight of evidence against the Null Hypothesis for rejecting / not rejecting H_0

If the p -value for testing H_0 is less than –

- **< 0.10**, we have some evidence that H_0 is false
- **< 0.05**, we have strong evidence that H_0 is false
- **< 0.01**, we have very strong evidence that H_0 is false
- **< 0.001**, we have extremely strong evidence that H_0 is false



Simple and Composite Hypothesis

A **simple hypothesis** is one in which all parameters of the distribution are specified. For example, if the heights of college students are normally distributed with $\sigma^2 = 4$ the hypothesis that its mean μ is, say, 62", that is $H: \mu = 62$, we have stated a simple hypothesis, as the mean and variance together specify a normal distribution completely.



Tap Here To Join Our
Official Community

Simple and Composite Hypothesis

A hypothesis which is not simple (i.e. in which not all of the parameters are specified) is called a **composite hypothesis**. For instance, if we hypothesize that $H : \mu > 62$ (and $\sigma^2 = 4$) or $H : \mu = 62$ and $\sigma^2 < 4$, the hypothesis becomes a composite hypothesis because we cannot know the exact distribution of the population in either case. Obviously, the parameters $\mu > 62$ and $\sigma^2 < 4$ have more than one value and no specified values are being assigned.



Tap Here To Join Our
Official Community

Critical Region (or Rejection Region)

The critical region CR, or rejection region RR, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test. That is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H_0 , the other will not. So, if the observed value of the test statistic is a member of the critical region, we conclude "Reject H_0 "; if it is not a member of the critical region then we conclude "Do not reject H_0 ".




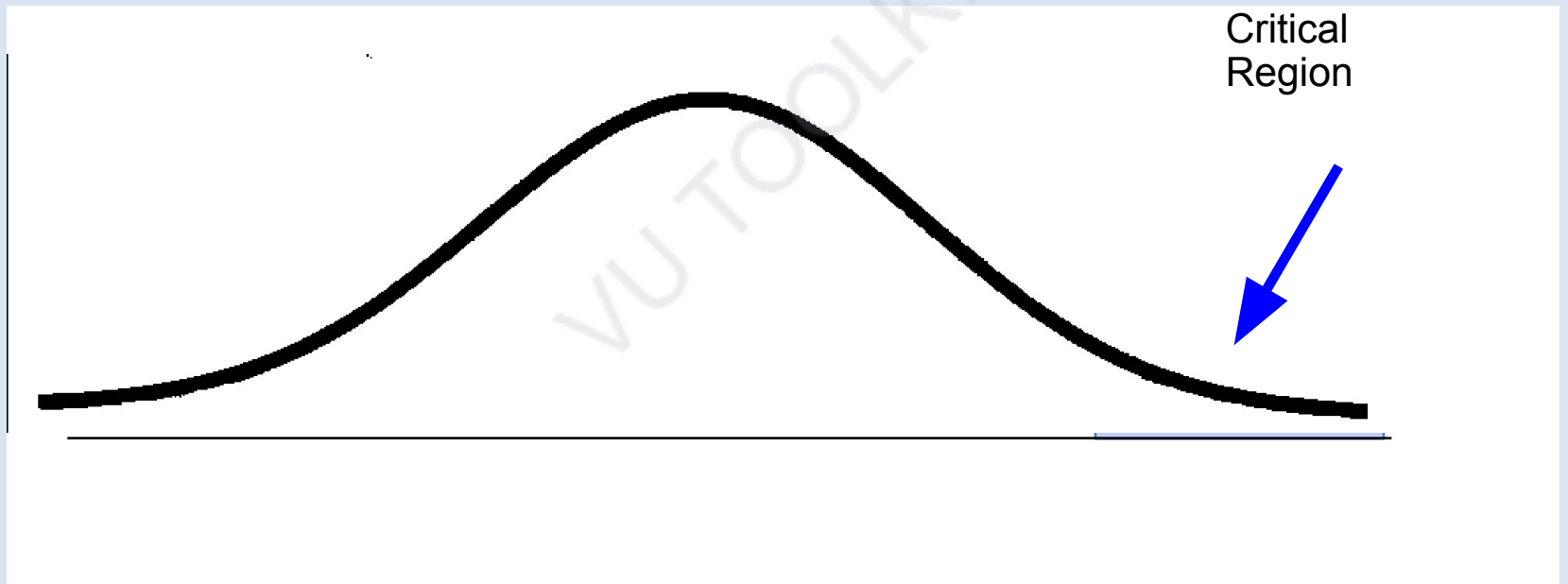
Critical Region



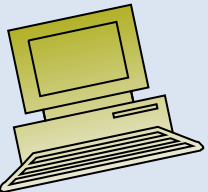
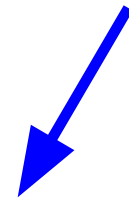
VU TOOLKIT



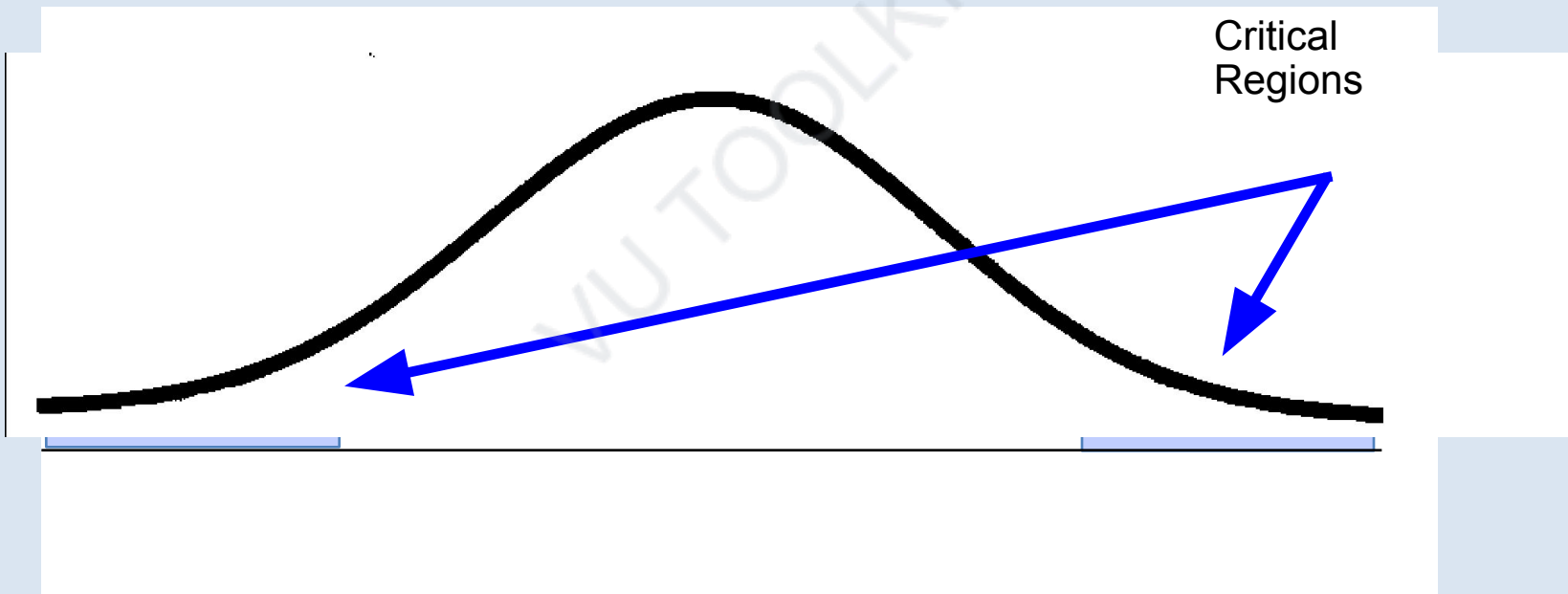
 Tap Here To Join Our Official Community



Critical
Region




Tap Here To Join Our
Official Community



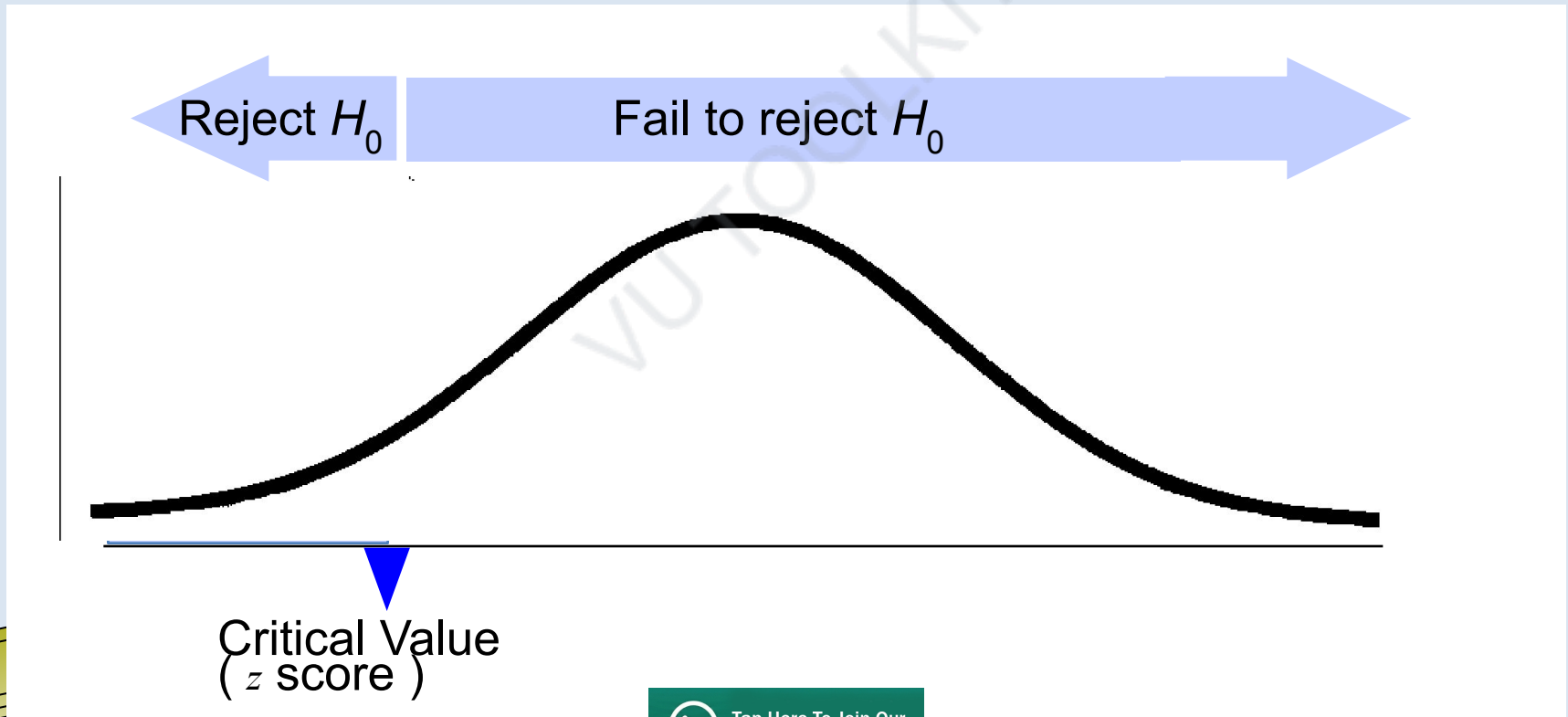
Critical
Regions



 Tap Here To Join Our
Official Community

Critical Value

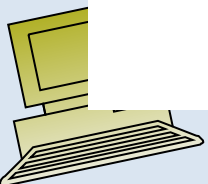
Any value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to a rejection of the null hypothesis



Critical Value
(z score)



Tap Here To Join Our
Official Community



Level of Significance, α and the Rejection Region

$$H_0: \mu \geq 3$$

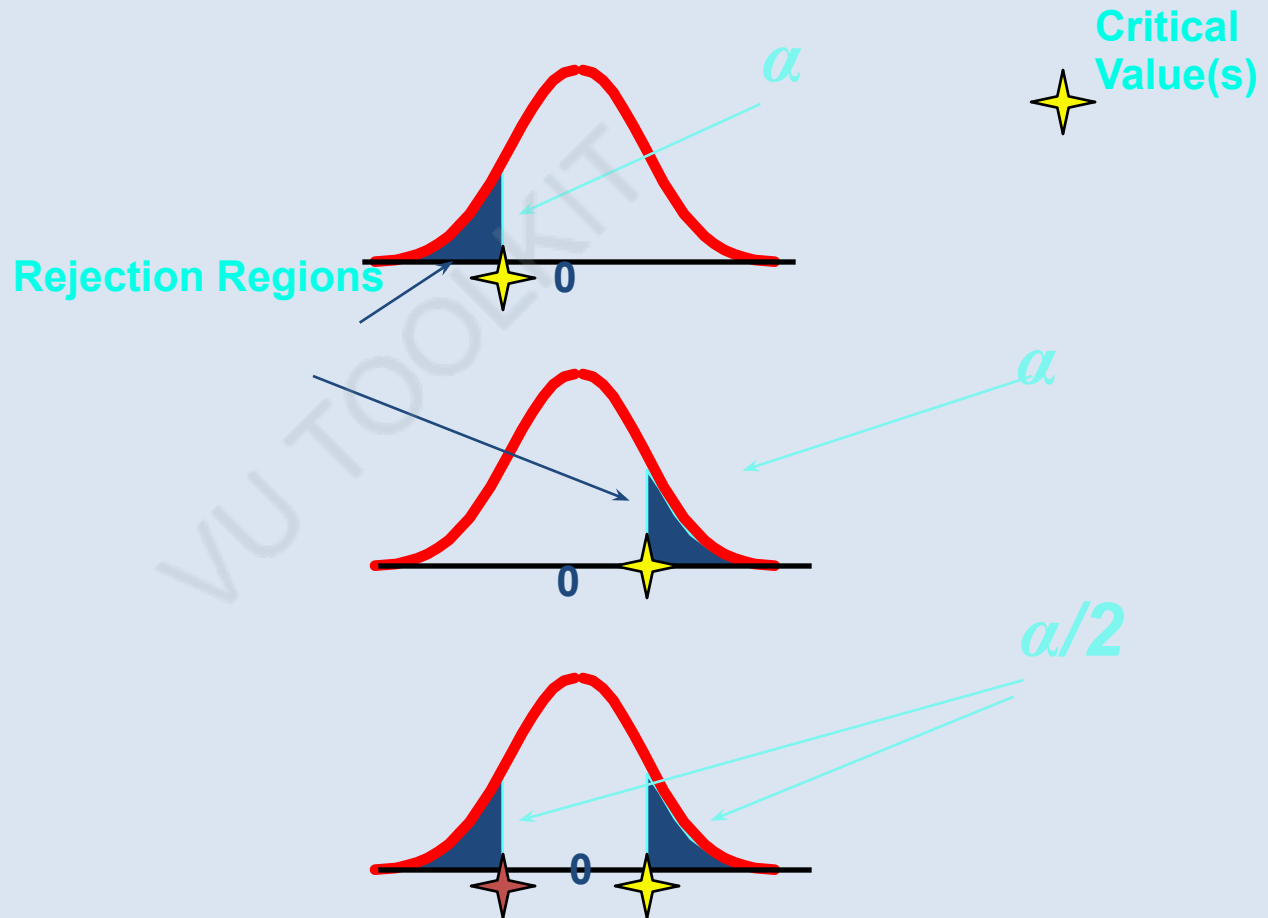
$$H_1: \mu < 3$$

$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$

$$H_0: \mu = 3$$

$$H_1: \mu \neq 3$$



Tap Here To Join Our
Official Community



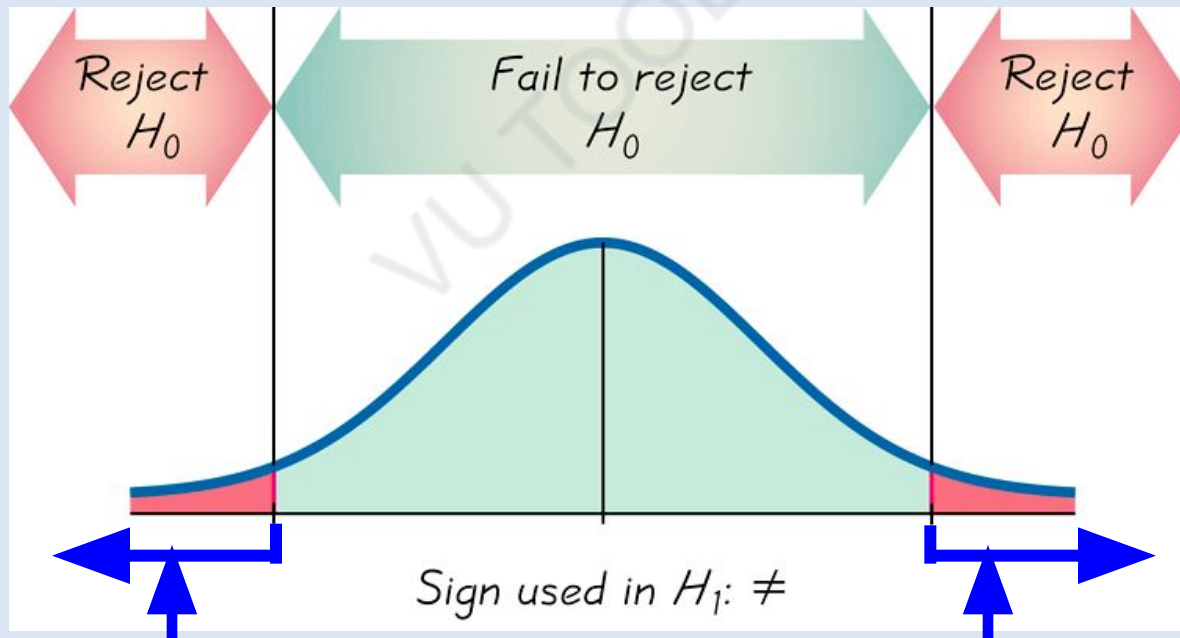
Two-tailed Test

$$H_0: =$$

$$H_1: \neq$$

α is divided equally between the two tails of the critical region

Means less than or greater than



Values that differ significantly from

H_0



Tap Here To Join Our Official Community

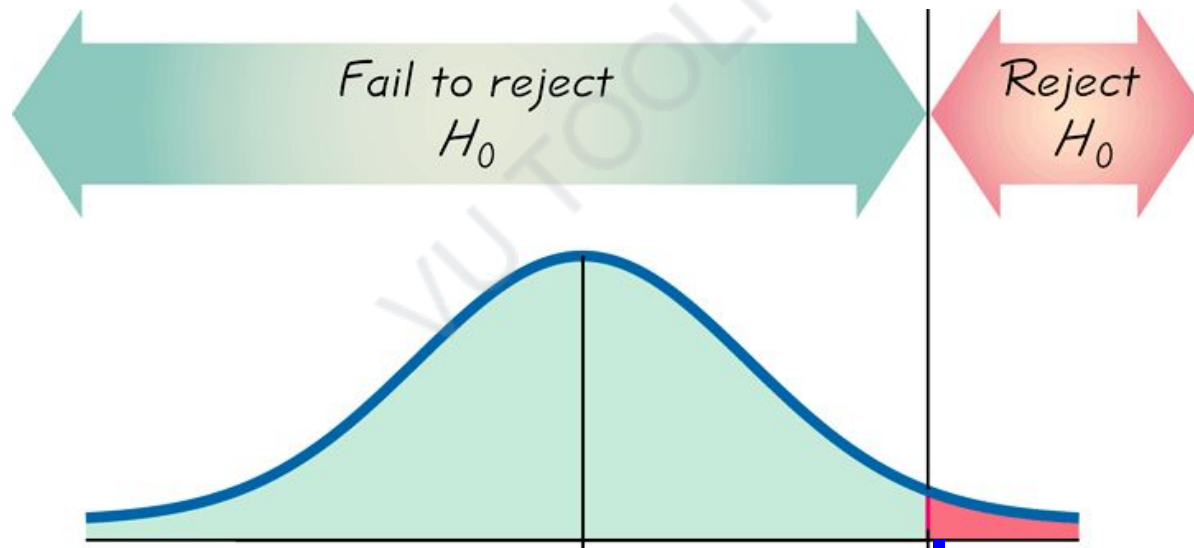


Right-tailed Test

$$H_0: =$$

$$H_1: >$$

Points Right

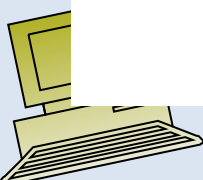


Values that differ significantly from H_0

Sign used in $H_1: >$



Tap Here To Join Our Official Community

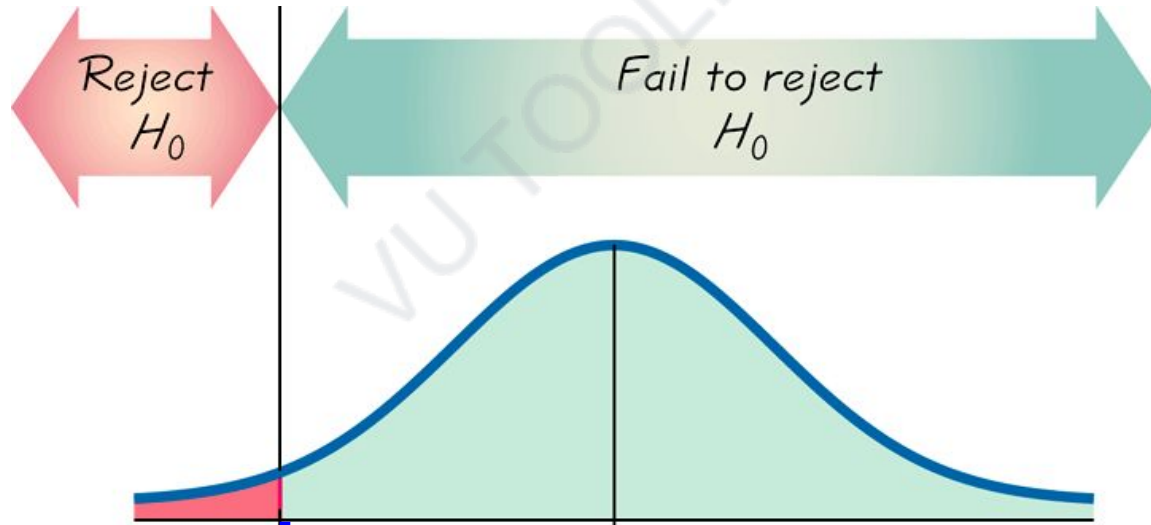
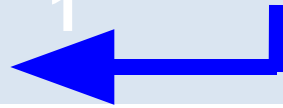


Left-tailed Test

$$H_0 :=$$

$$H_1 := <$$

Points Left



Values that differ significantly from H_0

Sign used in $H_1: <$



Tap Here To Join Our Official Community

Lec 7 - Hypothesis Testing.pptx

VU TOOLKIT



Tap Here To Join Our
Official Community

Lec 8 - Z T F Testing.pptx

Lecture 8

WU TOOLS



Tap Here To Join Our
Official Community

kistan

- Random Numbers
- Checking Normality
- Hypothesis testing (Continued)....




Tap Here To Join Our
Official Community

HYPOTHESIS TESTING

- Z-test
- T-test
- F-test

VU TOOLKIT



 Tap Here To Join Our Official Community

- **Notation**

- Test Statistic = $z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- Significance Level = α (in decimal form)
- Critical Values = z_{α} or $\pm z_{\alpha/2}$

- **Finding Critical Values**

Here we use the NORM.S.INV function.

NORM.S.INV stands for the inverse of the standard normal distribution (z -distribution).

Usage: NORM.S.INV(area to the left of the critical value)

This function returns the critical value from the z -distribution provided you put in the appropriate area.

Left-Tailed Tests: $z_{\alpha} = \text{NORM.S.INV}(\alpha)$

Right-Tailed Tests: $z_{\alpha} = \text{NORM.S.INV}(1 - \alpha)$

Two-Tailed Tests: $z_{\alpha/2} = \pm \text{NORM.S.INV}(\alpha/2)$

- **Finding P -Values**

Here we use the NORM.S.DIST function.

NORM.S.DIST stands for the standard normal distribution (z -distribution).



- Notation

- Test Statistic = $z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- Significance Level = α (in decimal form)
- Critical Values = z_{α} or $\pm z_{\alpha/2}$



- Finding Critical Values

Here we use the NORM.S.INV function.

NORM.S.INV stands for the inverse of the standard normal distribution (z -distribution).

Usage: NORM.S.INV(area to the left of the critical value)

This function returns the critical value from the z -distribution provided you put in the appropriate area.

Left-Tailed Tests: $z_{\alpha} = \text{NORM.S.INV}(\alpha)$

Right-Tailed Tests: $z_{\alpha} = \text{NORM.S.INV}(1 - \alpha)$

Two-Tailed Tests: $z_{\alpha/2} = \pm \text{NORM.S.INV}(\alpha/2)$



Tap Here To Join Our
Official Community

- Finding P -Values

Here we use the NORM.S.DIST function.

NORM.S.DIST stands for the standard normal distribution (z -distribution).

Usage: NORM.S.DIST(z , Cumulative?)

This function returns the area under the curve to the left of z when Cumulative = TRUE.

Left-Tailed Tests: P -value = NORM.S.DIST($z_{\bar{x}}$, TRUE) $z_{\bar{x}}$ should be < 0 .

Right-Tailed Tests: P -value = $1 - \text{NORM.S.DIST}(z_{\bar{x}}, \text{TRUE})$ $z_{\bar{x}}$ should be > 0 .

Two-Tailed Tests: P -value = $2(1 - \text{NORM.S.DIST}(|z_{\bar{x}}|, \text{TRUE}))$



Tap Here To Join Our
Official Community

23		1.354666
24		2.330544
25		4.036379
26		0.848363
27		1.268207
28		3.496795
29		4.40355
30		2.339433
31		0.845808
32		4.942728
33		0.05767
34		3.805368
35		3.839795
36		1.032443
37		1.255797
38		2.631889
39		1.995978
40		1.467479
41		
42	X-bar	2.662353
43		
44	S.D	1.310511
45		
46		

VU TOOLKIT

29	4.40355
30	2.339433
31	0.845808
32	4.942728
33	0.05767
34	3.805368
35	3.839795
36	1.032443
37	1.255797
38	2.631889
39	1.995978
40	1.467479

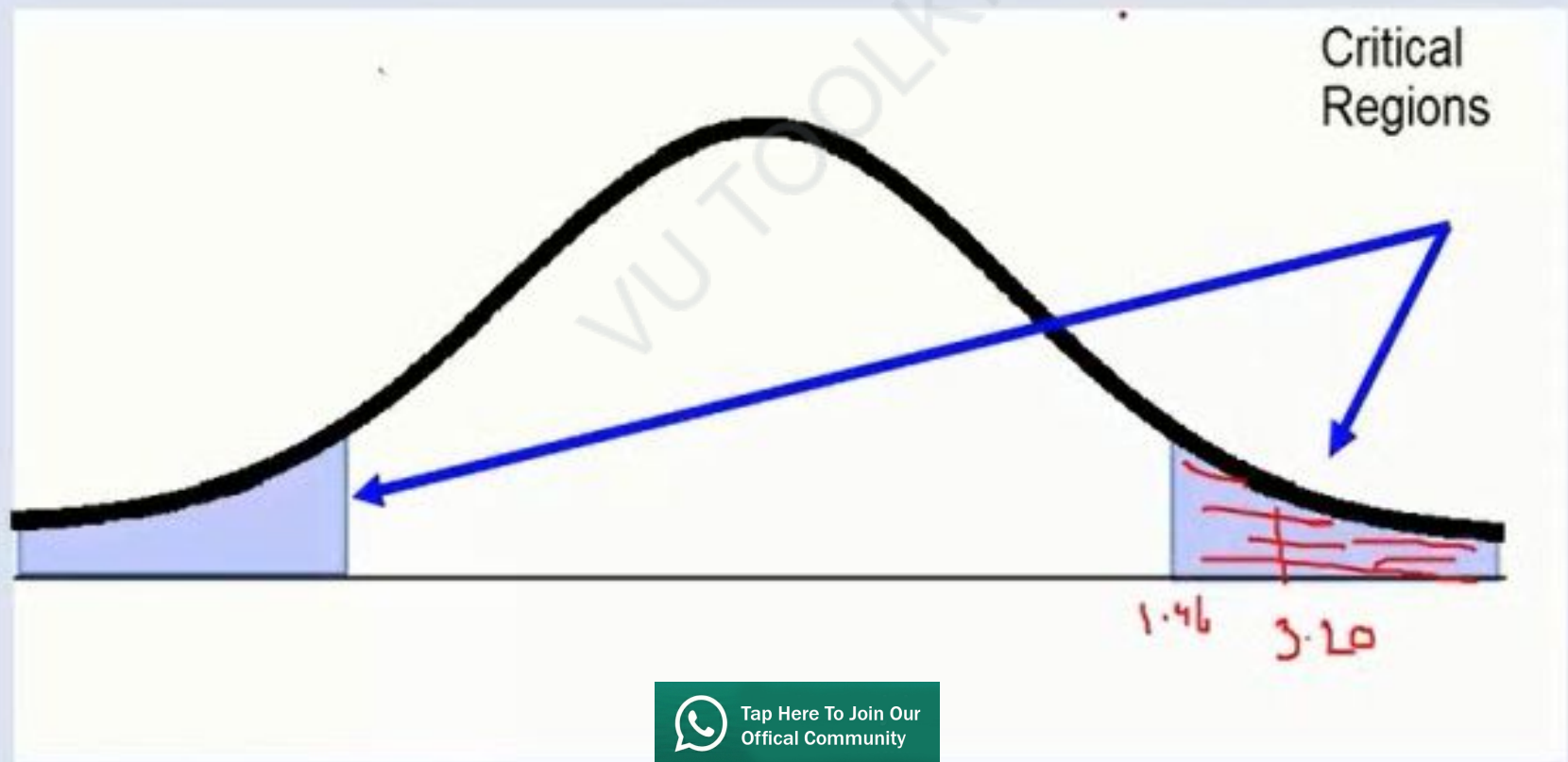
42	X-bar	2.662353
43		
44	S.D	1.310511

Z=
$$= \frac{D42-2}{(D44/SQRT(40))}$$

	A	B	C	D	E	F	G	H	I	J
30				2.339433						
31				0.845808						
32				4.942728						
33				0.05767						
34				3.805368						
35				3.839795						
36				1.032443						
37				1.255797						
38				2.631889						
39				1.995978						
40				1.467479						
41										
42			X-bar	2.662353			Z cal=	3.196529		
43										
44			S.D	1.310511						
45										
46										
47							Z cr=	=NORMSINV(0.025)		
48								NORMSINV(probability)		
49										
50										
51										

$$H_0 = \mu = 2$$

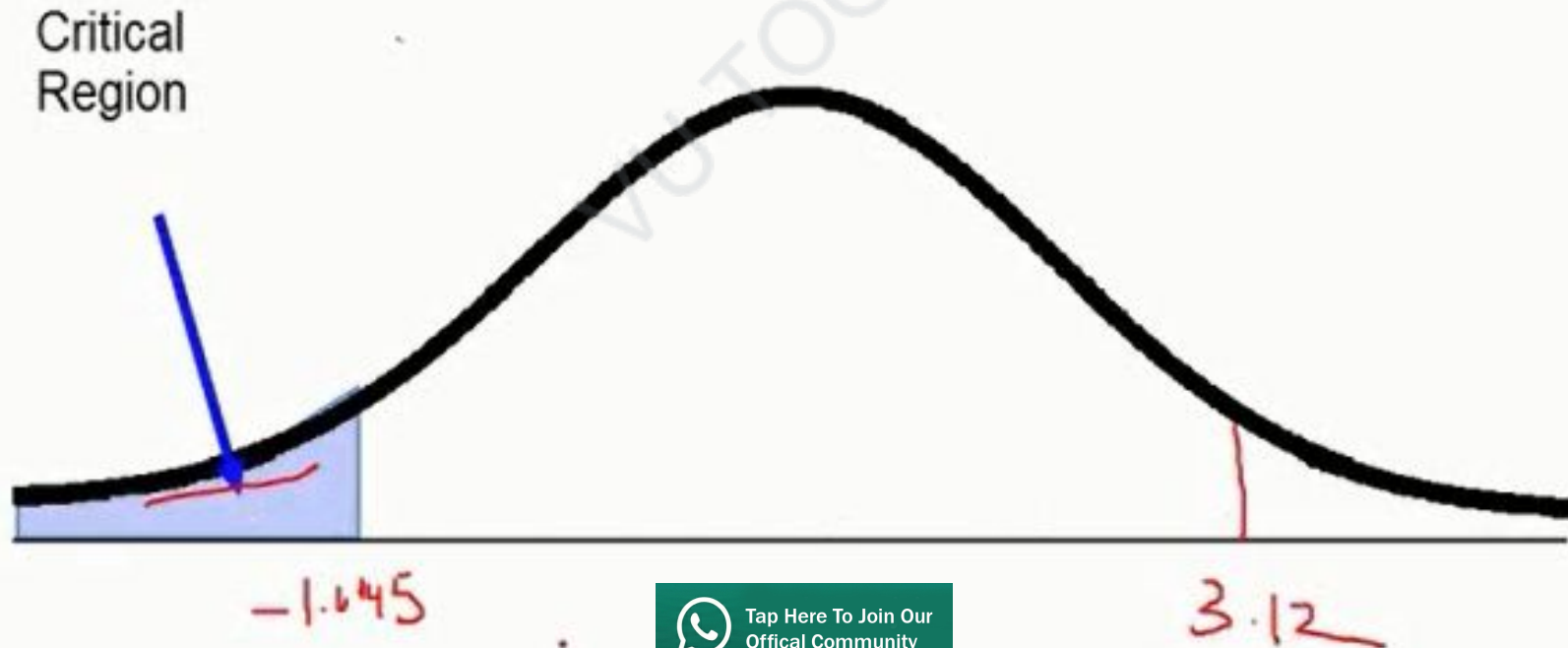
$$H_1 = \mu \neq 2$$



	A	B	C	D	E	F	G	H	I	J	K
30				2.339433							
31				0.845908							
32				4.942728							
33				0.05767							
34				3.805368							
35				3.839795							
36				1.032443							
37				1.255797							
38				2.631889							
39				1.995978							
40				1.467479							
41											
42			X-bar	2.662353			Z cal=	3.196529			
43											
44			S.D	1.310511							
45											
46											
47							Zcr=	-1.95996			
48											
49											
50											
51											

Critical Region

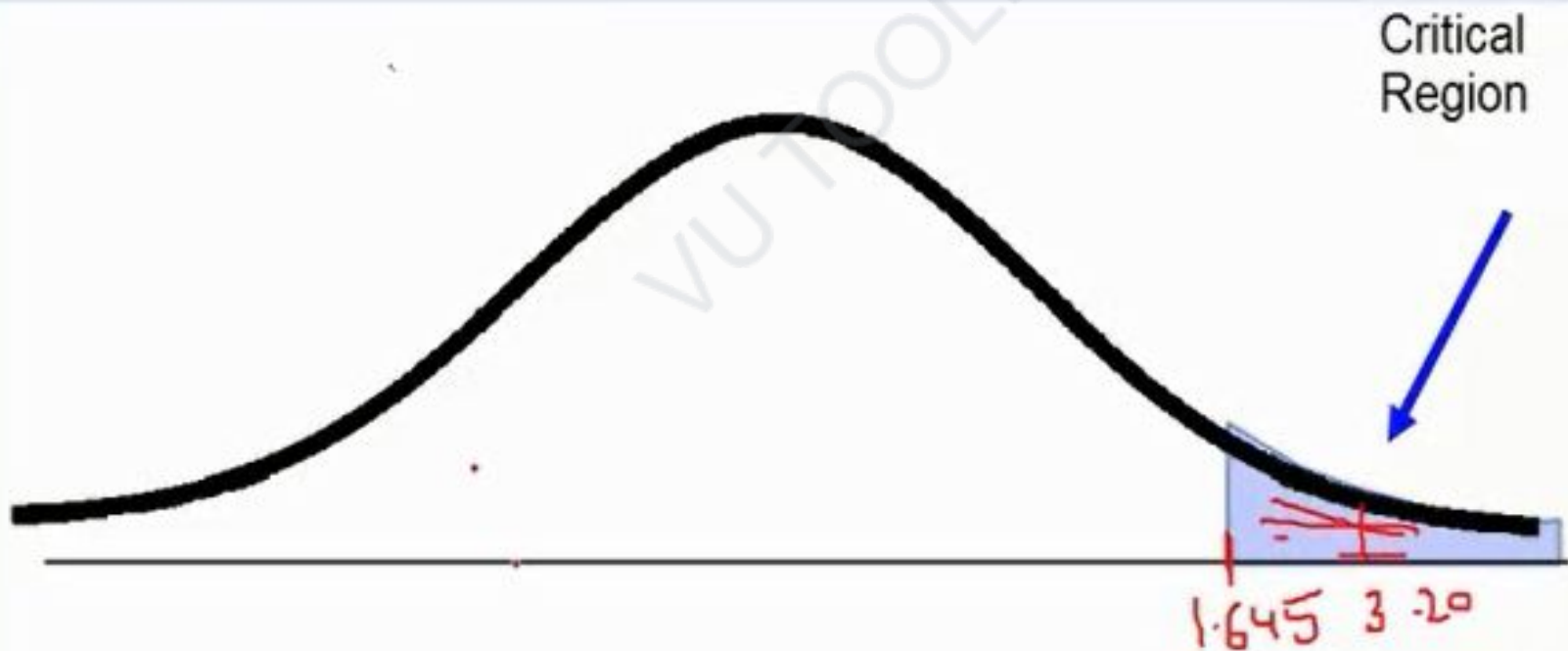
$$H_1: \mu < 2$$



Tap Here To Join Our
Official Community

$$H_0: \mu = 2$$

$$H_1: \mu > 2$$



Tap Here To Join Our
Official Community