Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Chapter 2

Frequency Distributions and Graphs

Objectives

2-1

- Organize data using frequency distributions.
- Represent data using pie chart and bar chart.
- Represent data using histogram, frequency polygon and ogive.
- Other types of graphs such as time series graph and stem and leaf plot are presented.

Introduction

2-2

- In order to conduct any statistical study about any phenomenon, you must gather data for variables that describe that phenomenon.
- After gathering the data, you must organize them in a meaningful way to describe the situation, draw conclusions or make inferences.
- The common way for organizing the data is by constructing a frequency distribution.
- After organizing the data, you can present them using different types of charts or graphs.

Organizing Data

2-3

- When data are collected in their original form, they are called *raw data*.
- Example: Blood Type Data

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

- Then, the *raw data* can be organized into a table form that is called *frequency distribution* using classes and frequencies.
- Types of frequency distributions are *categorical frequency distribution*, *ungrouped frequency distribution* and *grouped frequency distribution*

Categorical Frequency Distributions

2-4

- When the sample size (n) is large, the data must be grouped into categories.
- Categorical Frequency Distributions are used for data that can be placed in specific categories, such as ***nominal*** or ***ordinal*** level data.

Categorical Frequency Distributions

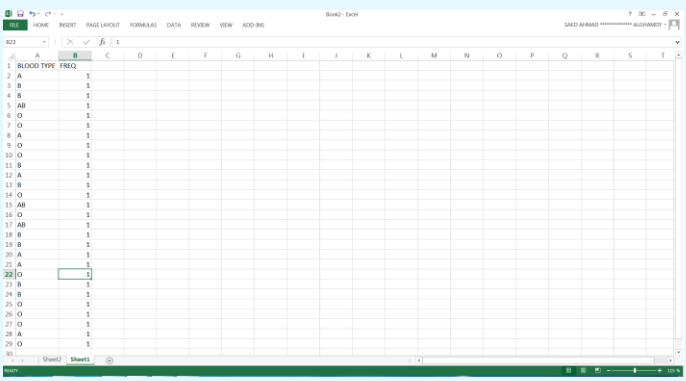
2-5

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Open Excel
- Enter the data in column A
- Enter the values of 1's in column B



Categorical Frequency Distributions

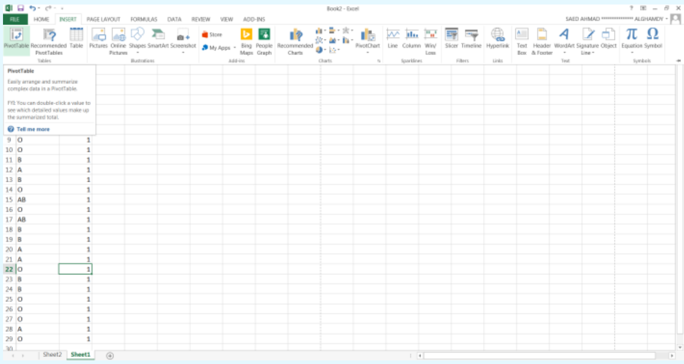
2-6

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Select INSERT
- Select PivotTable



Categorical Frequency Distributions

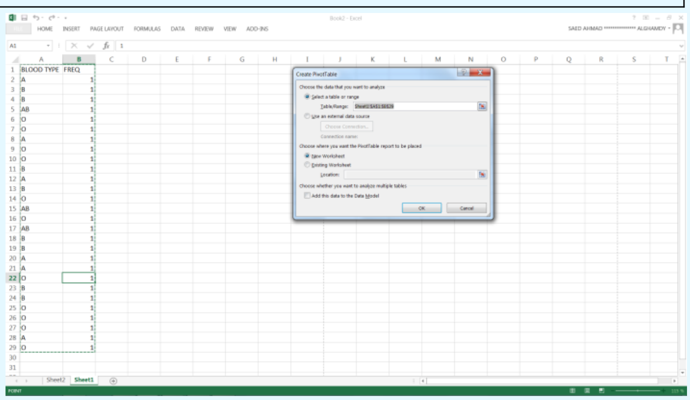
2-7

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Select OK



© All rights are preserved for Dr. Saeed A. Dobbah Alghamdi, Department of Statistics, Faculty of Sciences, KAU

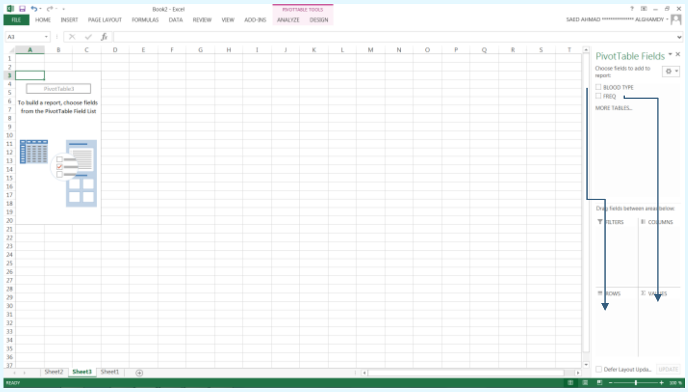
Categorical Frequency Distributions

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Drag the variable BLOOD TYPE into the ROWS area
- Drag the constant FREQ into the VALUES area



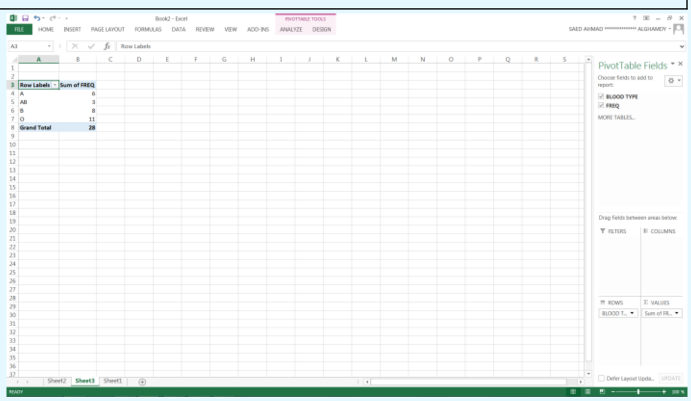
Categorical Frequency Distributions

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Rename the table columns as
  - Blood Type
  - Frequency



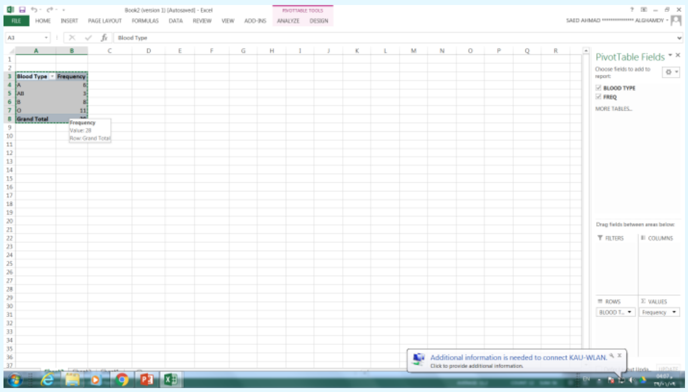
Categorical Frequency Distributions

2-10

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

<u>Instructions:</u>
- Select the table
- Press <u>Ctrl+C</u> to copy the table
- Press <u>Ctrl+V</u> to paste the table into your document
- Edit the table



Categorical Frequency Distributions

2-11

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Blood Type	Frequency
A	6
AB	3
B	8
O	11
Grand Total	**28**

Sample Size (n)

Categorical Frequency Distributions

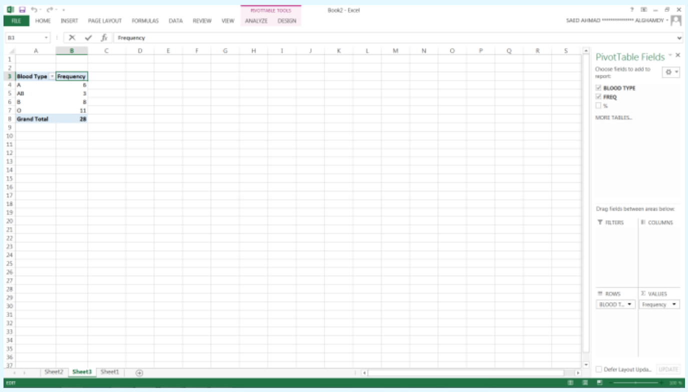
2-12

□ Example: Blood Type Frequency Distribution (Adding Percentages)

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Select ANALYZE



Categorical Frequency Distributions

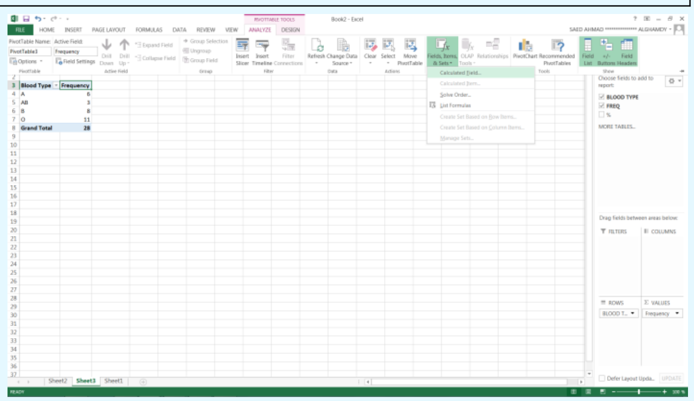
2-13

□ Example: Blood Type Frequency Distribution (Adding Percentages)

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Select Fields, Items & Sets
- Select Calculated Field



© All rights are preserved for Dr. Saeed A. Dobbah Alghamdi, Department of Statistics, Faculty of Sciences, KAU

Categorical Frequency Distributions

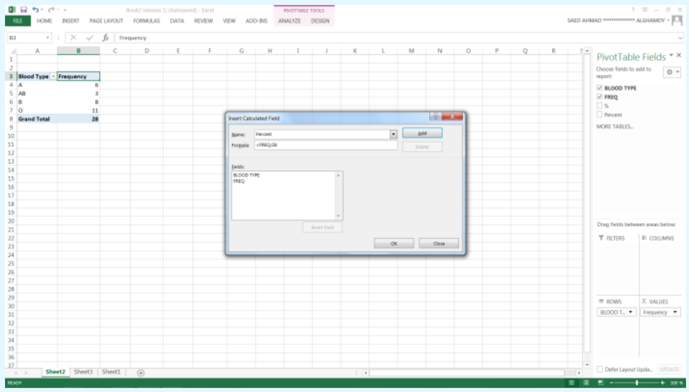
2-14

□ Example: Blood Type Frequency Distribution (Adding Percentages)

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Enter the name of the variable as Percent
- Enter the Formula for calculating the percent as FREQ/28
- Select OK



Categorical Frequency Distributions

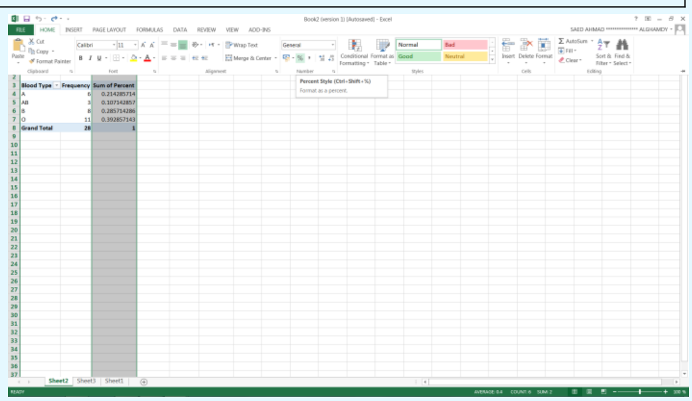
2-15

□ Example: Blood Type Frequency Distribution (Adding Percentages)

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Instructions:

- Select column C
- Select HOME
- Select % to change the format of the column to %.



Categorical Frequency Distributions

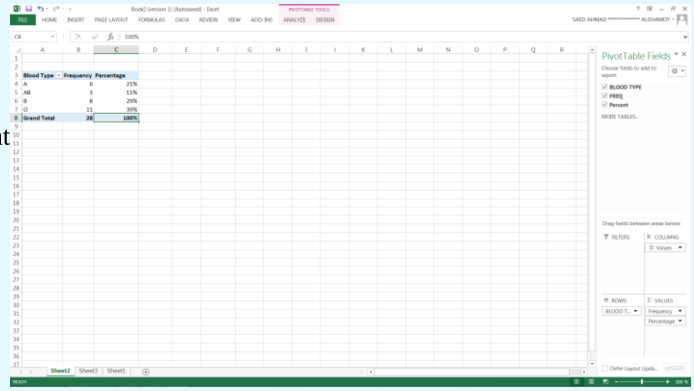
2-16

- Example: Blood Type Frequency Distribution (Adding Percentages)

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

<u>Instructions:</u>

- Select the table
- Copy and paste it into your document
- Edit the table



Categorical Frequency Distributions

2-17

- Example: Blood Type Frequency Distribution

A	B	B	AB	O	O	A	O	O	B	A	B	O	AB
O	AB	B	B	A	A	O	B	B	O	O	O	A	O

Blood Type	Frequency	Percent
A	6	21%
AB	3	11%
B	8	29%
O	11	39%
Grand Total	**28**	**100%**

$$\frac{frequency}{Total}\times 100=\frac{f}{\sum f}\times 100$$

$$=\frac{f}{n}\times 100\ \%$$

Sample Size

The Most Common Graphs for Categorical Data

2-18

- The *pie chart* is a circle that is divided into sections according to the frequencies in each category of the distribution, $Degree = \frac{f}{n} \times 360$. It is the best graph for displaying the nominal level of qualitative data.

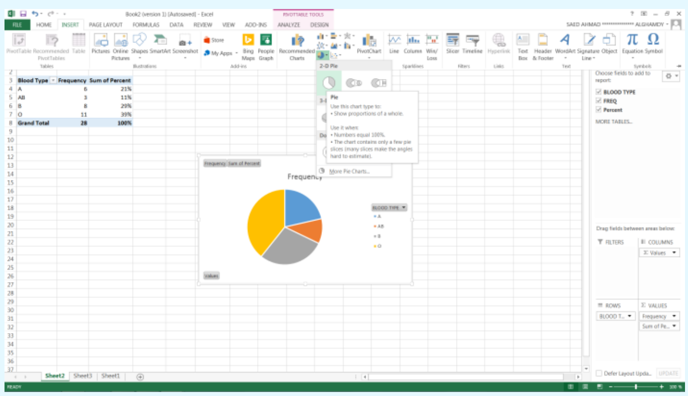
Class	Frequency	Percentage	Degree
A	6	21.43%	77.14
B	8	28.57%	102.86
O	11	39.29%	141.43
AB	3	10.71%	38.57
Total	**28**	**100%**	**360**

The Most Common Graphs for Categorical Data

2-19

- *Pie chart* using Excel.

Instructions:
- Select the table
- Select INSERT
- Select Pie chart



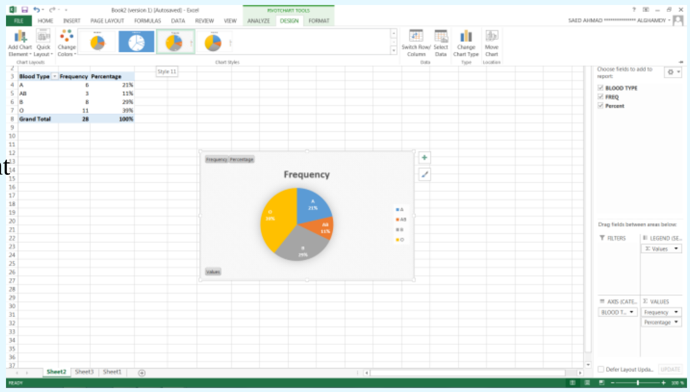
The Most Common Graphs for Categorical Data

2-20

- *Pie chart* using Excel.

Instructions:

- Click on the Chart
- Select DESGIN
- Format the chart
- Copy and paste it into your document

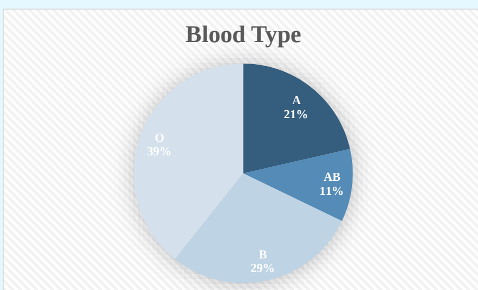


The Most Common Graphs for Categorical Data

2-21

- *Pie chart* using Excel.

Edit the graph



The Most Common Graphs for Categorical Data

2-22

- The *bar charts* display the data using vertical bars of various heights to reflect the frequencies of the categories. It is the best graph for displaying the ordinal level or discrete type of data.
- Example: Education levels

Education Level	Frequency
Level 1	47
Level 2	15
Level 3	12
Level 4	7
Level 5	19
Total	**100**

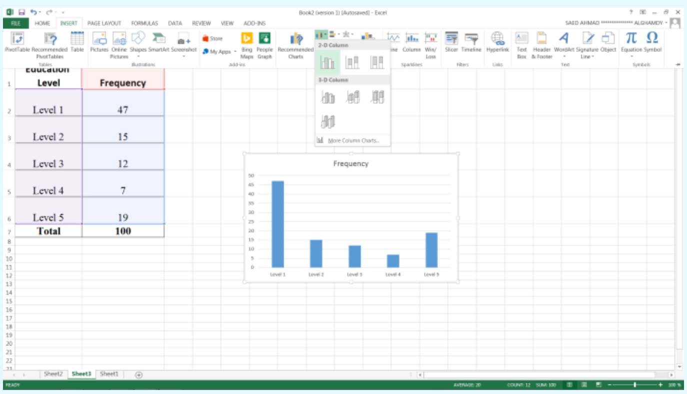
The Most Common Graphs for Categorical Data

2-23

- *Bar chart* using Excel.

Instructions:

- Copy the table to Excel
- Select the table without the total
- Select INSERT
- Select Bar chart



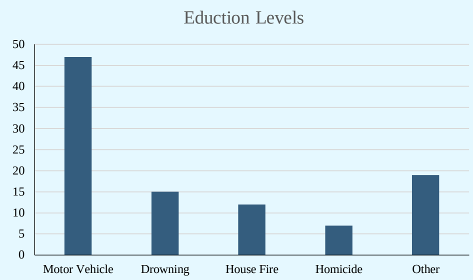
The Most Common Graphs for Categorical Data

2-24

- *Bar chart* using Excel.

Instructions:

- Edit the chart
- Copy and paste it into your document



Ungrouped Frequency Distributions

2-25

- Ungrouped frequency distributions are used for data that can be enumerated and when the range of values in the data set is small (discrete data) and the sample size (n) is large.
- Examples:
  - Number of children per family
  - Number of cars in a parking lot
  - Number of houses
  - Number of students

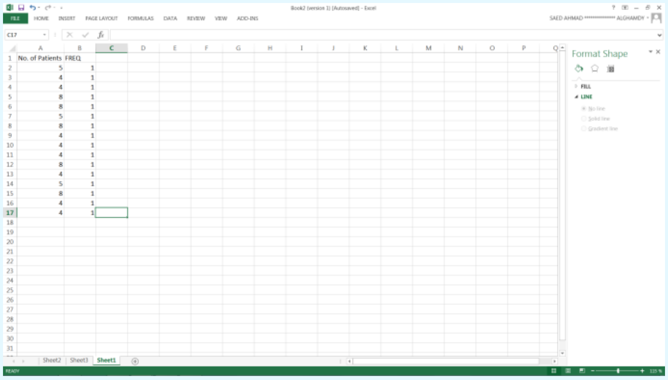
Ungrouped Frequency Distributions

2-26

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

Instructions:
- Open Excel
- Enter the data in column A
- Enter the values of 1's in column B



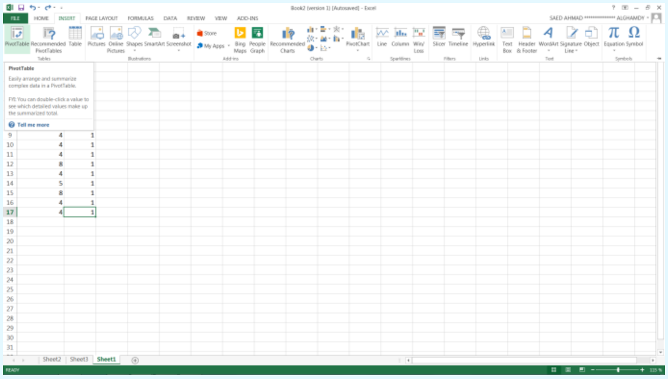
Ungrouped Frequency Distributions

2-27

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

Instructions:
- Select INSERT
- Select PivotTable



Ungrouped Frequency Distributions

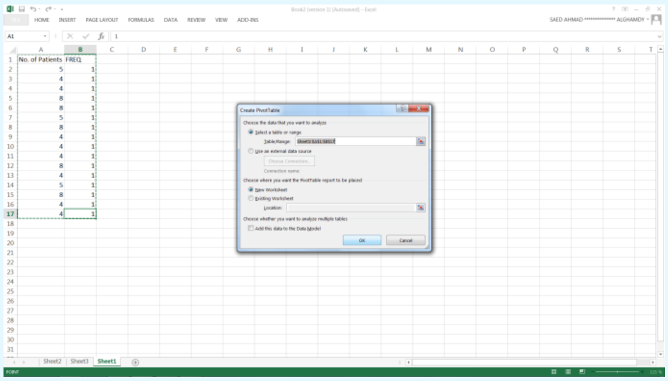
2-28

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

Instructions:

- Select OK



Ungrouped Frequency Distributions

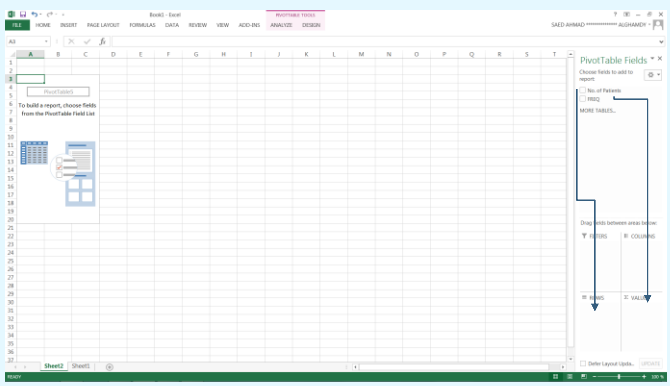
2-29

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

Instructions:

- Drag the variable No. of Patients into the ROWS area
- Drag the constant FREQ into the VALUES area



Ungrouped Frequency Distributions

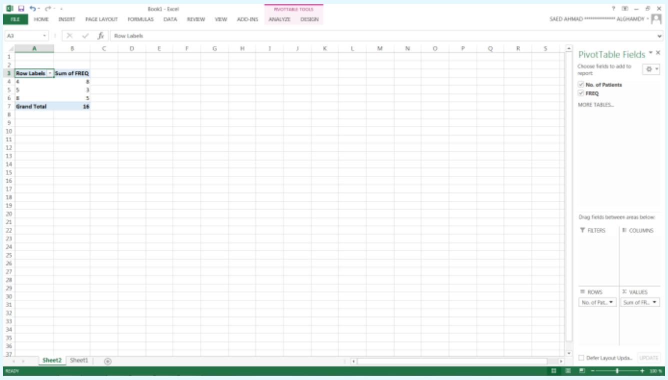
2-30

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

<u>Instructions:</u>

- Rename the table columns as
  No. of Patients
  Frequency



Ungrouped Frequency Distributions

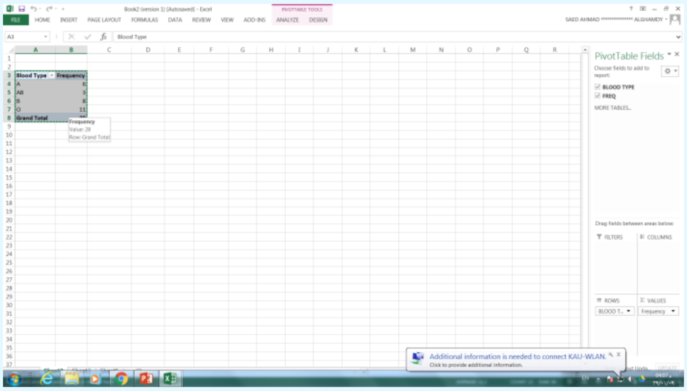
2-31

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

<u>Instructions:</u>

- Select the table
- Press <u>Ctrl+C</u> to copy the table
- Press <u>Ctrl+V</u> to paste the table into your document
- Edit the table



Ungrouped Frequency Distributions

- Example: Number of patients in the waiting rooms of 16 clinics within a hospital at a specific time.

5 4 4 8 8 5 8 4 4 4 8 4 5 8 4 4

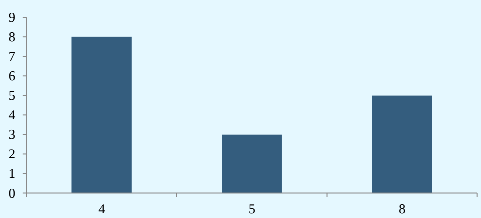
No. of Patients	Frequency
4	8
5	3
8	5
Grand Total	16

Sample Size (n)

The Most Common Graph for Discrete Data

- The *bar charts* display the data by using vertical bars of various heights to reflect the frequencies of the categories. It is the best graph for displaying the discrete type of data or ordinal level of qualitative data.
- Example: Number of patients

No. of Patients	Frequency
4	8
5	3
8	5
Grand Total	16



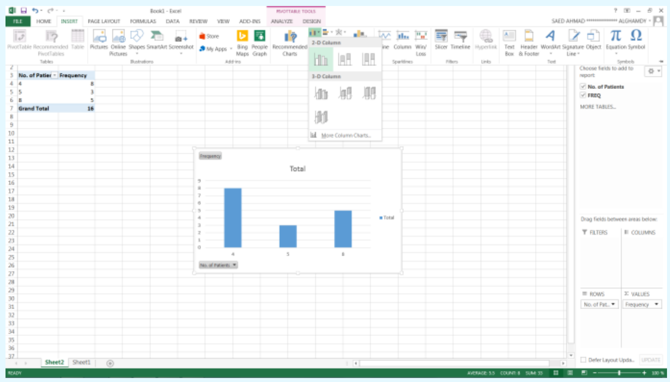
The Most Common Graphs for Categorical Data

2-34

- *Bar chart* using Excel.

Instructions:

- Select the table without the total
- Select INSERT
- Select Bar chart



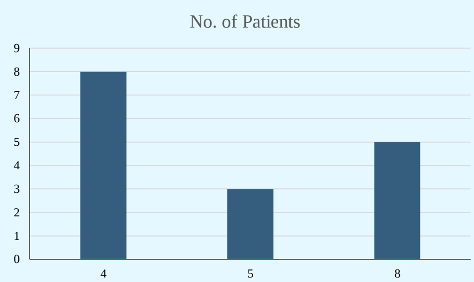
The Most Common Graphs for Categorical Data

2-35

- *Bar chart* using Excel.

Instructions:

- Edit the chart
- Copy and paste it into your document



Grouped Frequency Distributions

2-36

- When the range of values in a data set is large (continuous data), the data must be grouped into classes that are more than one unit in width, e.g., 24 – 30.
- The *lower class limit* represents the smallest data value that can be included in a class, e.g., 24 for the class limit 24 – 30.
- The *upper class limit* represents the largest value that can't be included in the class, e.g., 30 for the class limit 24 – 30.

Grouped Frequency Distributions

2-37

- The *class width* for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class.
- The *class midpoint* is found by adding the lower and upper boundaries (or limits) and dividing by 2.

Grouped Frequency Distributions

2-38

Class Rules

- There should be between 5 and 20 classes.

  As a guide line, the number of classes can be found using

  $$\boxed{\text{Number of Classes} \approx 1 + 3.3 \times \log(n)}$$

- The first class lower limit usually is the lowest value in the data set
- The class width should be an odd number.
- The classes must be mutually exclusive.
- The classes must be continuous.
- The classes must be exhaustive.
- The classes must be equal in width.

Grouped Frequency Distributions

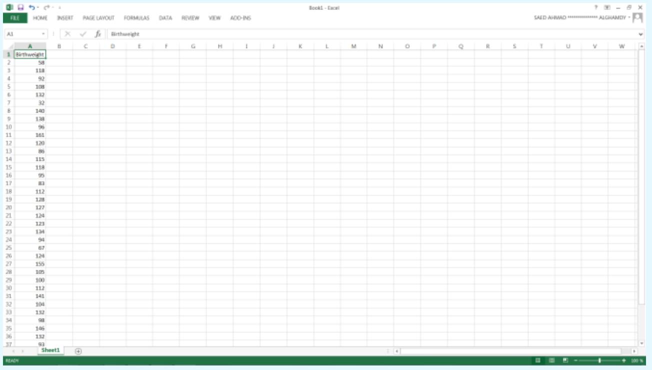
2-39

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Open Excel
- Enter the data in column A



Grouped Frequency Distributions

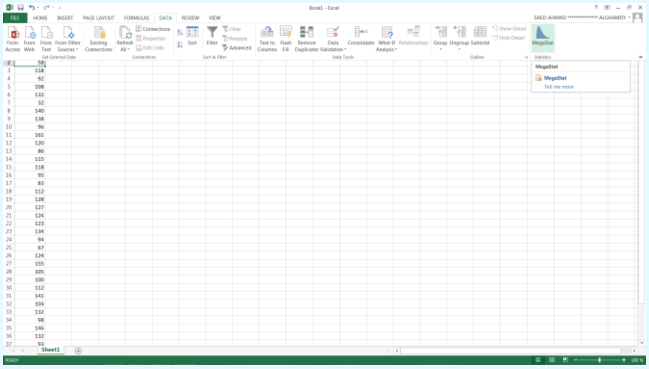
2-40

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Select DATA
- Select MegaStat



Grouped Frequency Distributions

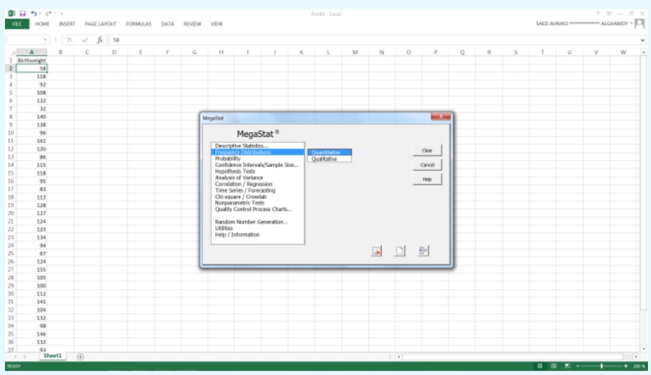
2-41

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Select Frequency Distribution
- Select Quantitative



Grouped Frequency Distributions

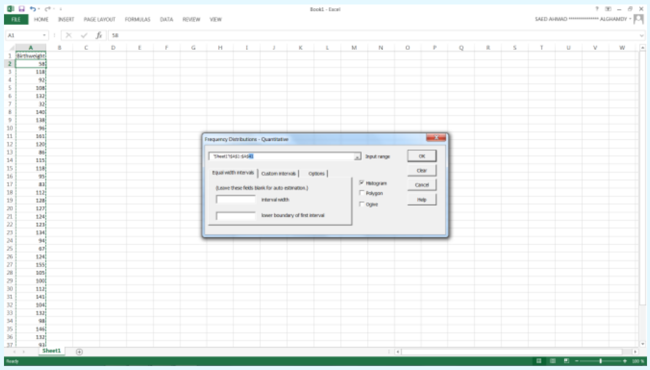
2-42

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Click on Input range
- Select your data "column A"



Grouped Frequency Distributions

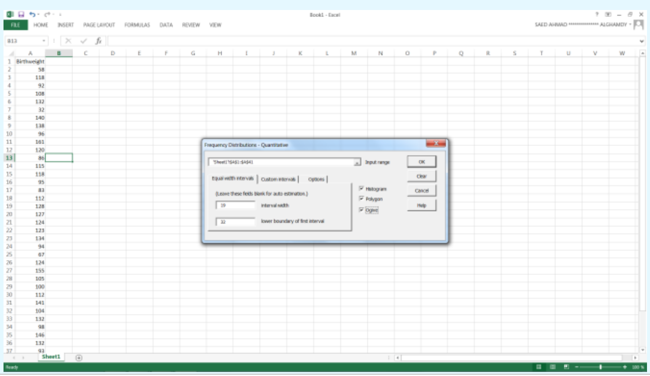
2-43

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

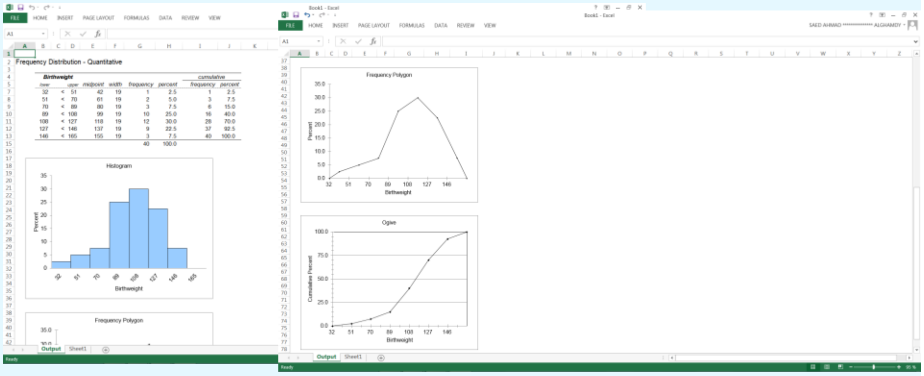
- Enter Interval width
- Enter Lower boundary of the first interval
- Select the desired graphs
  - Histogram
  - Polygon
  - Ogive
- Click OK



Grouped Frequency Distributions

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					



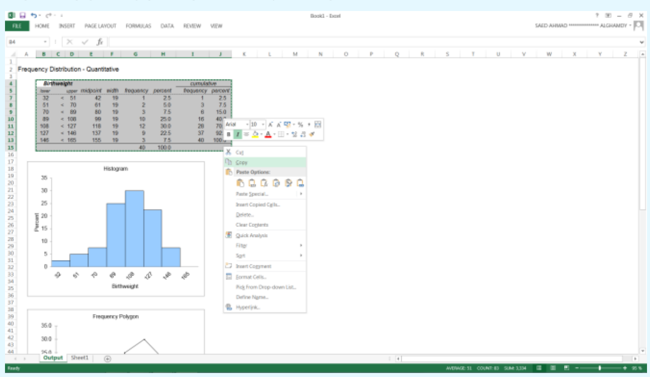
Grouped Frequency Distributions

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Select the frequency distribution table and copy it into your document.



Grouped Frequency Distributions

2-46

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

<u>Instructions:</u>

- Edit your table

Birthweight							cumulative	
lower		upper	Midpoint	width	frequency	percent	frequency	percent
32	<	51	42	19	1	2.5	1	2.5
51	<	70	61	19	2	5.0	3	7.5
70	<	89	80	19	3	7.5	6	15.0
89	<	108	99	19	10	25.0	16	40.0
108	<	127	118	19	12	30.0	28	70.0
127	<	146	137	19	9	22.5	37	92.5
146	<	165	155	19	3	7.5	40	100.0
					40	100.0		

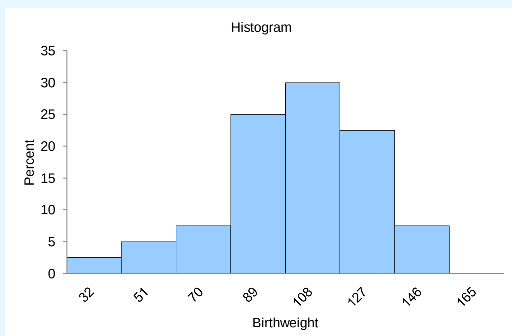
The Most Common Graphs for Continuous Data

2-47

- The ***histogram*** displays the <u>continuous</u> data that are organized in a grouped frequency distribution by using vertical bars of various heights to represent the frequencies.
- Example: Sample of birthweight

<u>Instructions:</u>

- Copy the <u>histogram</u> graph from MegaStat output and paste it into your document



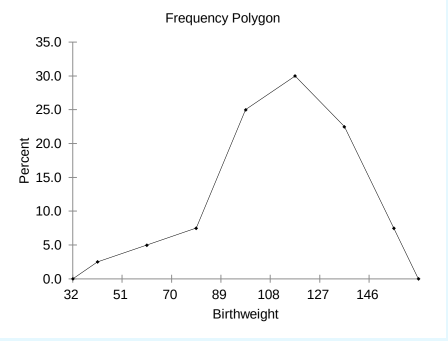
The Most Common Graphs for Continuous Data

2-48

- The *frequency polygon* displays the continuous data that are organized in a grouped frequency distribution by using lines that connect points plotted for the frequencies at the midpoints of the classes.
- Example: Sample of birthweight

Instructions:

- Copy the Frequency polygon graph from MegaStat output and paste it into your document



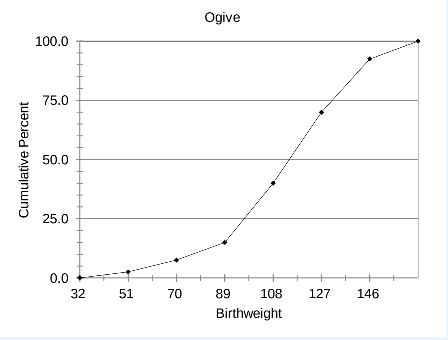
The Most Common Graphs for Continuous Data

2-49

- The *cumulative frequency graph* or *ogive* represents the cumulative frequencies for the classes in a **grouped frequency distribution**.
- Example: Sample of birthweight

Instructions:

- Copy the Ogive graph from MegaStat output and paste it into your document



Other Types of Graphs

2-50

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:

- Copy the data table into Excel sheet

Year	Cargo and Mail Traffic
1998	199368
1999	203402
2000	209102
2001	192702
2002	214587
2003	203480
2004	222418
2005	221344
2006	198063
2007	209119

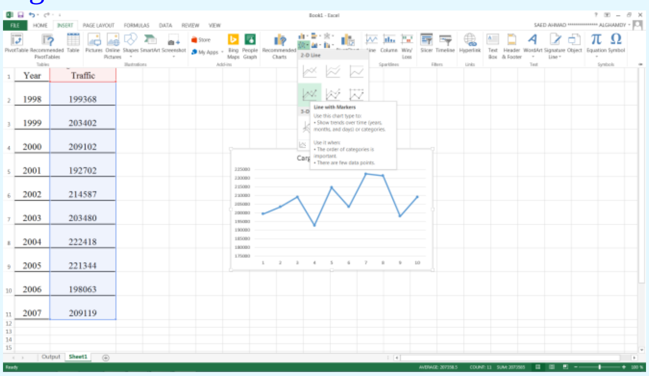
Other Types of Graphs

2-51

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:

- Edit the table
- Select Cargo and Mail Traffic, column B
- Select INSERT
- Select Line Chart

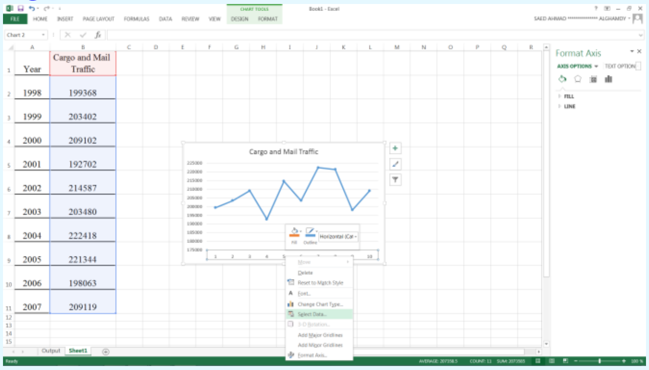


Other Types of Graphs

2-52

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:
- Edit the graph by selecting the horizontal axis
- Click right on the mouse
- Select Select Data

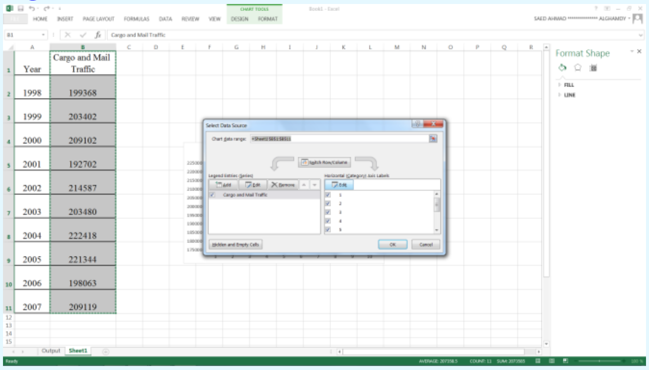


Other Types of Graphs

2-53

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:
- Select Edit on the right box

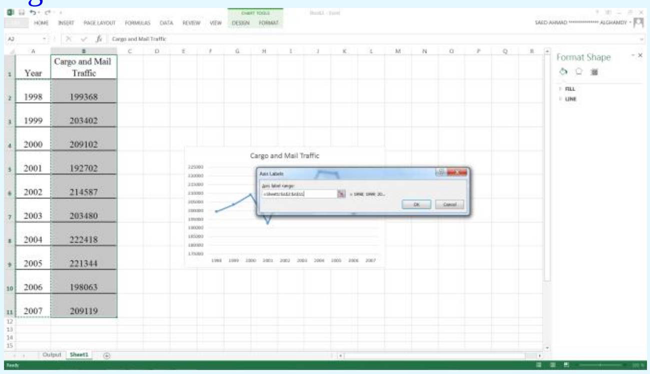


Other Types of Graphs

2-54

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:
- Select the data in Year, column A
- Click OK
- Click OK again
- Copy the graph into your document

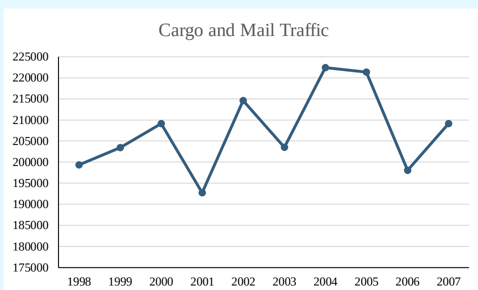


Other Types of Graphs

2-55

- The *time series graph* represents data that occur over a specific period of time.
- Example: Yearly cargo and mail traffic of an airline

Instructions:
- Edit the graph



Other Types of Graphs

2-56

- A *stem-and-leaf plot* is a data plot that uses part of a data value as the leaf, the less significant digits (the 'units'), and the other part of the data value as the stem, the most significant digit (i.e. the 'tens'), to form groups or classes.
- It has the advantage over grouped frequency distribution of retaining the actual data while showing them in a graphic form.

Grouped Frequency Distributions

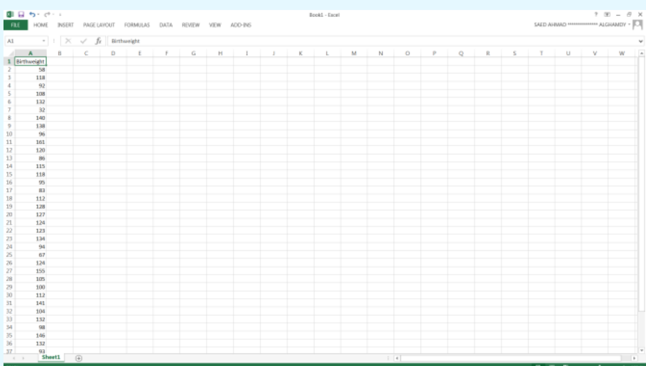
2-57

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Open Excel
- Enter the data in column A



Grouped Frequency Distributions

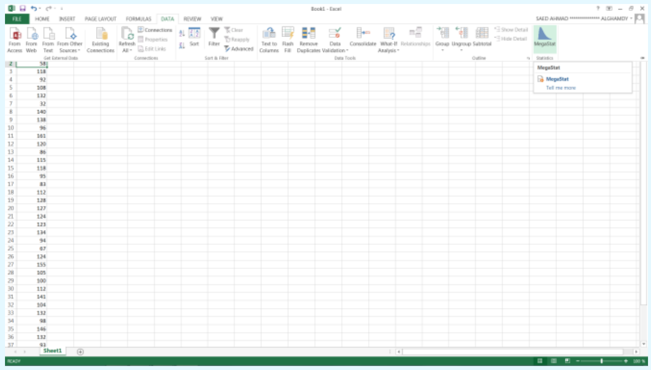
2-58

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Select DATA
- Select MegaStat



Grouped Frequency Distributions

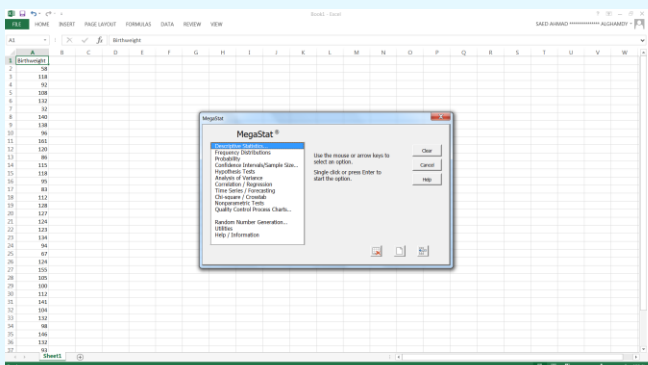
2-59

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Select Descriptive Statistics …



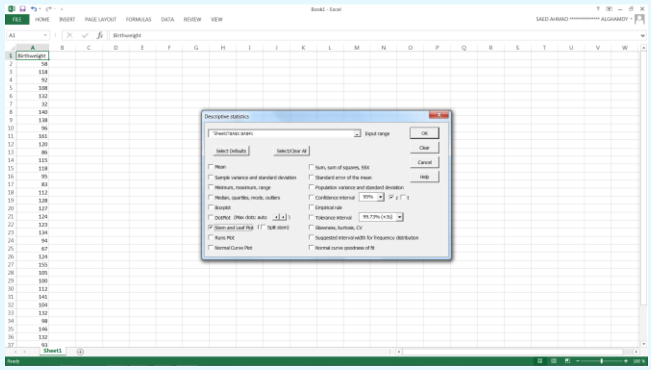
Grouped Frequency Distributions

2-60

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:
- Click on Input range
- Select your data "column A"
- Unselect all choices
- Select Stem and Leaf Plot
- Click OK



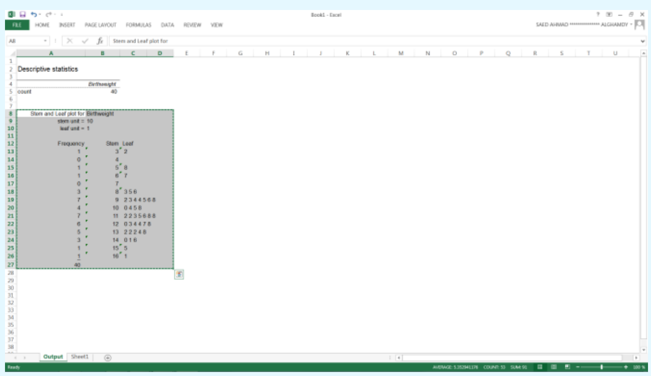
Grouped Frequency Distributions

2-61

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:
- Click on Input range
- Select your data "column A"
- Unselect all choices
- Select Stem and Leaf Plot
- Click OK
- Copy the graph into your document



Grouped Frequency Distributions

2-62

- Example: Sample of birthweight (oz) from 40 consecutive deliveries

58	118	92	108	132	32	140	138	96	161	120	86	115	118	95
83	112	128	127	124	123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113					

Instructions:

- Edit the graph

Stem and Leaf plot for Birthweight

Frequency	Stem	Leaf
1	3	2
0	4	
1	5	8
1	6	7
0	7	
3	8	3 5 6
7	9	2 3 4 4 5 6 8
4	10	0 4 5 8
7	11	2 2 3 5 6 8 8
6	12	0 3 4 4 7 8
5	13	2 2 2 4 8
3	14	0 1 6
1	15	5
1	16	1
40		

Summary

2-63

- When data are collected in the original form, they called raw data.
- Since little information can be obtained from raw data, they must be organized in a frequency distribution.
- Categorical frequency distribution is used for qualitative data (nominal or ordinal)
- Ungrouped frequency distribution is used for discrete data.
- Grouped frequency distribution is used for continuous data.
- Pie chart is used mostly to represent nominal data.
- Bar chart are used mostly to represent discrete and ordinal data.
- Histogram, frequency polygon and ogive graphs are used to represent continuous data.

Summary

2-64

- Time series graph is used to represent data that occur over specific period of time.
- Stem and leaf plot is a combination of sorting and graphing. It retains the actual data while showing them graphically.

Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Chapter 3

Data Description

Objectives

3-1

- Summarize data using measures of central tendency, such as the mean, weighted mean, median and mode.
- Describe data using the measures of variation, such as the range, variance, and standard deviation.
- Identify the position of a data value in a data set using quartiles.
- Use the techniques of exploratory data analysis such as the five-number summary and boxplot to discover various aspects of data.

Introduction

3-2

- Data can be summarized numerically using measures of central tendency, measures of variation and measures of position.
- *Measures of central tendency* such as the *mean*, *weighted mean, median* and *mode* are used to describe a data set with a single value that represents the middle or center of the data's distribution.
- *Measures of variation* such as the *range*, *variance*, and *standard deviation* are used to describe a data set with a single value that represents the spread of the data's distribution.
- *Measures of position* such as *standard scores, percentiles* and *quartiles* are used to tell where a specific value falls within a data set .
- Another type of statistics is called *exploratory data analysis* which is used to get information about the center, spread, symmetry and outliers of a data set.

Introduction

3-3

Before going any further, we need to distinguish between the measures that are calculated using the values of a sample and using all values of a population.

- A *statistic* is a characteristic or measure calculated using the data values of a sample, e.g., the sample mean.
- A *parameter* is a characteristic or measure calculated using all the data values of a specific population, e.g., the population mean.

Measures of Central Tendency

3-4

- The *mean* is the sum of the values divided by the number of values.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

- The *weighted mean* is used when the values in a data set are not all equally represented. It is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

Formula	In Excel
$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$	$= (x_1 * w_1 + x_2 * w_2 + \cdots + x_n * w_n)/(w_1 + w_2 + \cdots + w_n)$

- Where $w_1, w_2, \ldots, w_n$ are the weights for $X_1, X_2, \ldots, X_n$

Measures of Central Tendency

3-5

- The *median* (MD) is the halfway point in a data set. It is found by arranging the data in order and selecting the middle point.
- The *mode* value is the value that occurs most often in a data set.
  - A data set with one value that occurs with greatest frequency is said to be *unimodal* ,*e.g., (3,2,1,2,4,5,6)* .
  - A data set with two values that occur with greatest frequency is said to be *bimodal* , *e.g., (3,2,1,2,4,5,1).*
  - A data set with more than two values that occur with greatest frequency is said to be *multimodal* , *e.g., (4,5,3,3,2,1,2,6,1)*
  - When all the values in a data set occur with the same frequency is said to have *no mode* , *e.g., (3,2,3,2,1,5,5,1)*

Properties of Central Tendency Measures

3-6

- The *mean*
  - is computed by using all the values of a data set.
  - varies less than the *median* or *mode*.
  - is unique, and not necessarily one of the data values.
  - is affected by extremely high or low values and may not be the appropriate average.
- The *median*
  - is used when one must find the center or middle value of a data set.
  - is used when one must determine whether the data values fall into the upper half or lower half of the distribution.
  - is affected less than the *mean* by extremely high or extremely low values.

Properties of Central Tendency Measures

3-7

- The *mode*
  - is used when the most typical case is desired.
  - is the easiest average to compute.
  - can be used when the data are nominal, such as religious preference or gender.
  - is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

Distribution Shapes

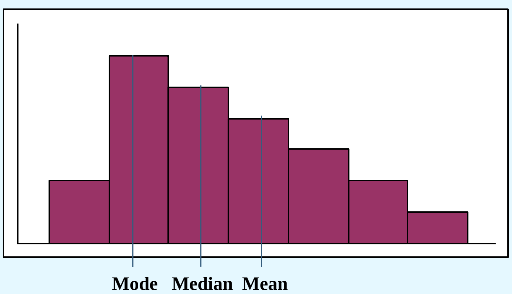
3-8

- Frequency distributions may assume different shapes. The most important shapes are *positively skewed* , *symmetrical* , *negatively skewed* .
- A coefficient of *skewness* may calculated to measure the degree of the skewness in a distribution. It ranges from -3 to 3. Negative value means negative skewness, zero means symmetrical and positive value means positive skewness. As the coefficient of skewness gets further away from zero as the shape became more skewed.
- A comparison of the mean, median and mode values indicates the type of skewness in the shape of a distribution.

Distribution Shapes

3-9

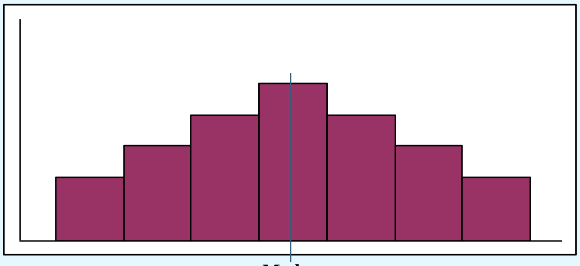
- In a *positively skewed* or *right skewed distribution*, the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution. Mode < Median < Mean



Distribution Shapes

3-10

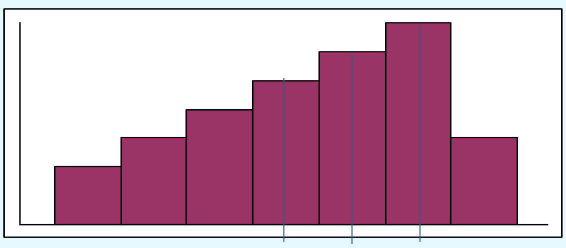
- In a *symmetrical distribution*, the data values are evenly distributed on both sides of the mean. $\boxed{\text{Mean} = \text{Median} = \text{Mode}}$



Distribution Shapes

3-11

- In a *negatively skewed* or *left skewed distribution*, the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution. $\boxed{\text{Mean} < \text{Median} < \text{Mode}}$



Measures of Variation

3-12

- The *range* is the highest value minus the lowest value in a data set.
- The *variance* is the average of the squares of the distance each value is from the mean.

Population Variance	*Sample Variance*
$\sigma = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$

- The *standard deviation* is the square root of the variance.

Measures of Variation

3-13

- The *coefficient of variation* is a measure of the dispersion of data values around the mean value of the data. It is calculated by dividing the standard deviation by the mean expressed as a percentage.
- The *coefficient of variation* is mostly used to compare standard deviations of two variables or more when the units or the values of the means are different.
- Large coefficient of variation means large variability.

Measure of Position

3-14

- A *standard score* or *z score* is used when direct comparison of raw scores is impossible.
- The z score represents the number of standard deviations a data value falls above or below the mean.

Population z-score	*Sample z-score*	In Excel
$z = \frac{x-\mu}{\sigma}$	$z = \frac{x-\bar{x}}{s}$	$=(x-\bar{x})/s$

- Positive z value means that the value is above the mean and negative z value mean that the value is below the mean

Measure of Position

3-15

- *Percentiles* divide the data set into 100 equal groups.
- The percentile corresponding to a given value X is computed using the following formula:

Formula	*In Excel*
$P = \frac{(number\ of\ values\ below\ X)+0.5}{total\ number\ of\ values} * 100$	$= (\boldsymbol{countif}(\boldsymbol{data};\text{"<X"}) + \mathbf{0.5})/\boldsymbol{n} * \mathbf{100}$

Measure of Position

3-16

Finding a data value corresponding to a given percentile

1- Arrange the data in order from lowest to highest.

In Excel
Select Data ⟶ HOME ⟶ Sort & Filter ⟶ Sort Smallest to Largest

2- Substitute into the formula

Formula	In Excel
$c = \frac{n * p}{100}$	$= n * p/100$

3- If c is not a whole number, round up to the next number. Starting at the lowest value, count over to the value that corresponds to the rounded-up number.

4- if c is a whole number, use the mid-value between the cth and (c+1)st values when counting up from lowest value.

Measures of Position

3-17

- *Quartiles* divide the distribution into four groups, denoted by Q_1, Q_2, Q_3. Note that Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile or the median; and Q_3 corresponds to the 75th percentile.
- An *outlier* is an extremely high or an extremely low data value when compared with the rest of the data values.
- *Outliers* can be identified using the *interquartile range (IQR)* which is a *measure of variation* that can be used when the data contains outlier values. IQR = $Q_3 - Q_1$
- Outliers can be the result of measurement or observational error.

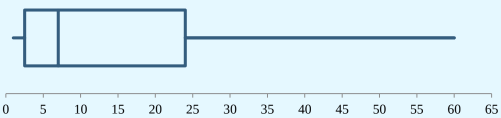
Exploratory Data Analysis

3-18

Exploratory data analysis includes the *box plot* and the *five-number summary*.

- *Boxplots* are graphical representations of a *five-number summary* of a data set.
- The five specific values that make up a *five-number summary* are

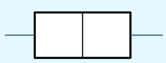
minimum, Q_1, Q_2, Q_3 and maximum.



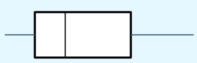
Skewness and boxplot

3-19

- Using the box:
  - If the median is near the center of the box, the distribution is approximately symmetric.



  - If the median is to the left of the box, the distribution is positively skewed.



  - If the median is to the right of the box, the distribution is negatively skewed.

Applying the Concepts

3-20

- The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

- If you work for the school board and do not wish to increase salaries. Compute the measures of central tendency and decide which one would best support your position.
- If you work for the teachers' union and want a raise for the teachers. Use the best measure of central tendency to support your position.
- Explain how outliers can be used to support one or the other position.
- If the salaries represented every teacher in the school district, would the averages be parameters or statistics?
- Which measure of central tendency can be misleading when a data set contains outliers?
- When you are comparing the measures of central tendency, does the distribution display any skewness? Explain.

Applying the Concepts

3-21

- The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

- If you work for the school board and do not wish to show a large variation in salaries. Compute the measures of variation and decide which one would best support your position.
- If you work for the teachers' union and want to show a large variation in salaries. Use the best measure of variation to support your position.
- Which measure of variation can be misleading when a data set contains outliers?
- From the coefficient of skewness, does the distribution display any skewness? Explain.
- If you take the first eight salaries as one group and the rest as another group, which group is more variable?
- Find the z score for a teacher's salary of 14000 and for a teacher's salary 18000. Explain.
- Find the percentile rank of a teacher's salary of 15400. Explain.

Applying the Concepts

3-22

- The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

- What value corresponds to the 30th percentile?
- What value corresponds to the 50th percentile?
- Calculate the values of Q_1, Q_2 and Q_3 and decide in which quartile a teacher's salary of 17000 falls.
- How many observations falls between the minimum and the median?
- Is the data contains any outlier values?
- From the boxplot, comment on the skewness of the distribution.

Applying the Concepts

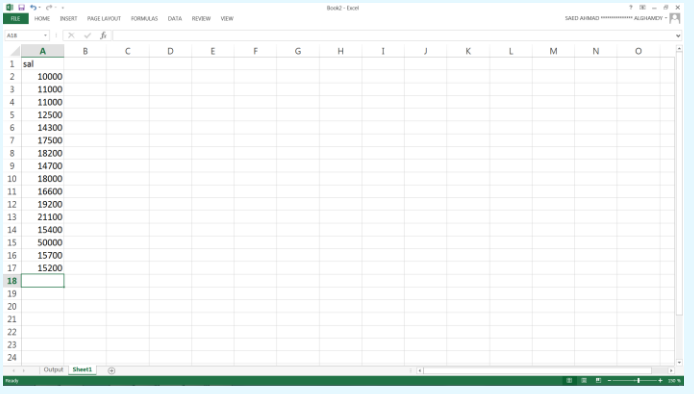
2-23

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Open Excel
- Enter the data in column A



Applying the Concepts

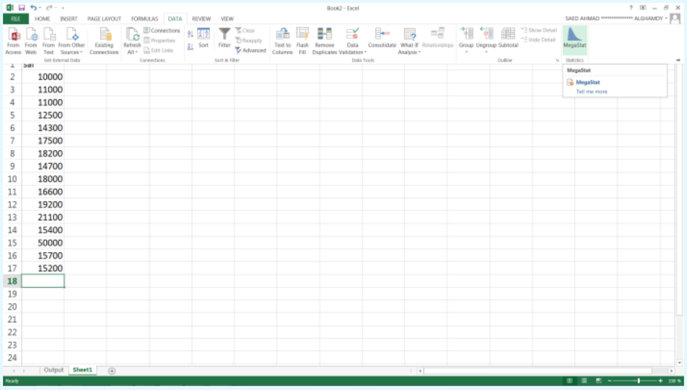
2-24

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Select DATA
- Select MegaStat



Applying the Concepts

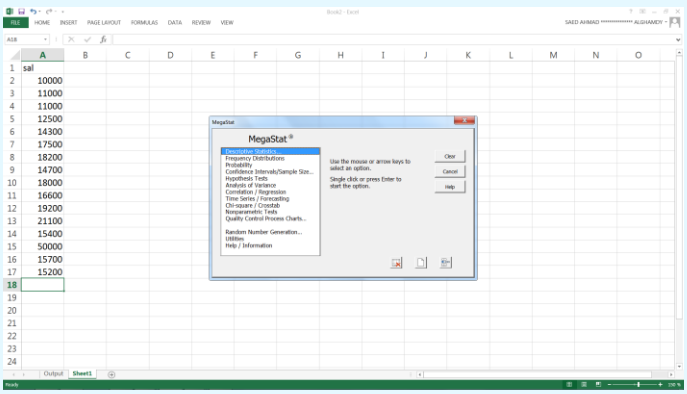
2-25

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Select Descriptive Statistics



Applying the Concepts

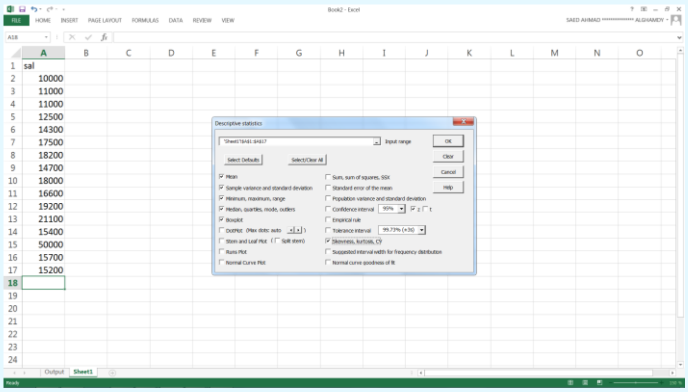
2-26

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Click on Input range
- Select your data "column A"
- Check the boxes in front of
  - Mean
  - Sample var …
  - Minimum …
  - Median …
  - Boxplot
  - Skewness …
- Click OK

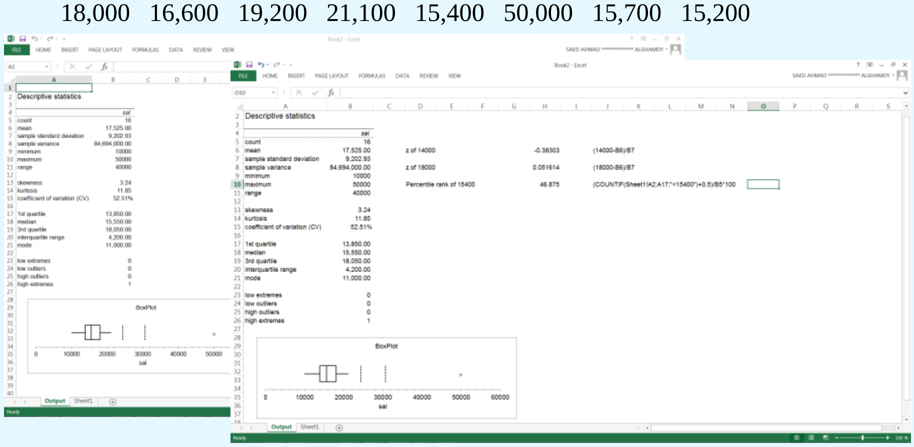


Applying the Concepts

2-27

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200



Applying the Concepts

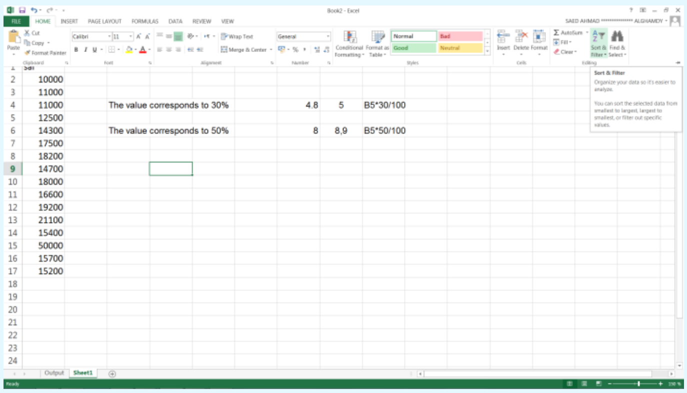
2-28

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Select HOME
- Select Sort & Filter



Applying the Concepts

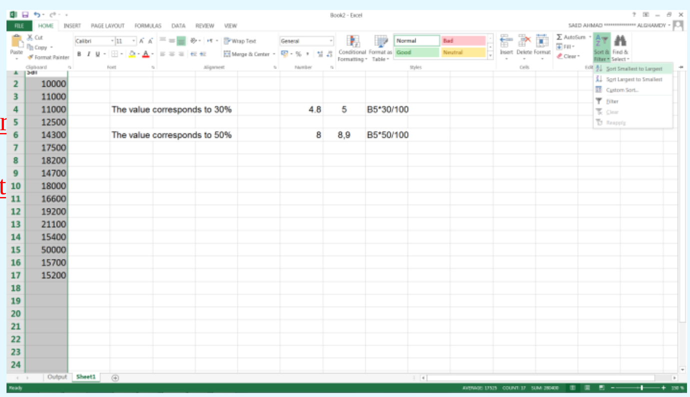
2-29

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

Instructions:

- Select your Data
- Select HOME
- Select Sort & Filter
- Select Sort Smallest to Largest

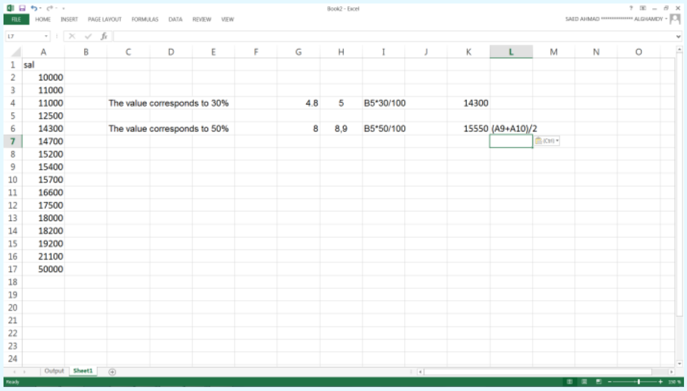


Applying the Concepts

2-30

- Example: The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200



Summary

3-31

- Measures of central tendency such as the mean, median and mode are used to summarize data.
  - The mean is the sum of values divided by the total number of values.
  - The median is the middle value of an ordered data set.
  - The mode is the most frequent data value.
- Measures of variation such as the range, variance and standard deviation are used to describe the spread of data.
  - The range is the distance between highest value and lowest value.
  - The variance is the average of the squares of the distance between the mean and each value in a data set.
  - The standard deviation is the square root of the variance.

Summary

3-32

- Measure of position such as the standard scores, percentiles and quartiles are used to identify the position of a data value.
  - Standard score is the number of standard deviation that a data value is above or below the mean.
  - Percentile is the position in hundredths that a data value holds in the distribution.
  - Quartile is the position in fourths that a data value holds in the distribution.

Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Chapter 4

Probability and Counting Rules

Objectives

4-1

- Determining sample space.
- Finding the probability of an event and compound events.
- Finding the total number of outcomes in a sequence of events.
- Finding the total number of selecting r objects from n objects.

Introduction

4-2

- *Probability* as a general concept can be defined as the chance of an event occurring.
- Probability are used in games of chance, insurance, investments, weather forecasting and in various other areas.
- Rules such as the fundamental counting rule, permutation rule and combination rule allow us to count the number of ways in which events can occur.

Basic Concepts

4-3

- A *probability experiment* is a chance process that leads to well-defined results called outcomes.
- An *outcome* is the result of a single trial of a probability experiment.
- A *sample space* is the set of all possible outcomes of a probability experiment.
- An *event* consists of a set of outcomes of a probability experiment.
- An event with one outcome is called a simple event and with more than one outcome is called compound event.

Basic Concepts

4-4 **Example**

Find the sample space for the gender of the children if a family has three children and give an example for a simple event and another one for a compound event. Use B for boy and G for girl.

There are two genders and three children, so there are $\mathbf{2^3 = 8}$ possibilities as shown here;

BBB BBG BGB GBB BGG GBG GGG GGB

So, the sample space is

S={BBB, BBG, BGB, GBB, BGG, GBG, GGB, GGG}

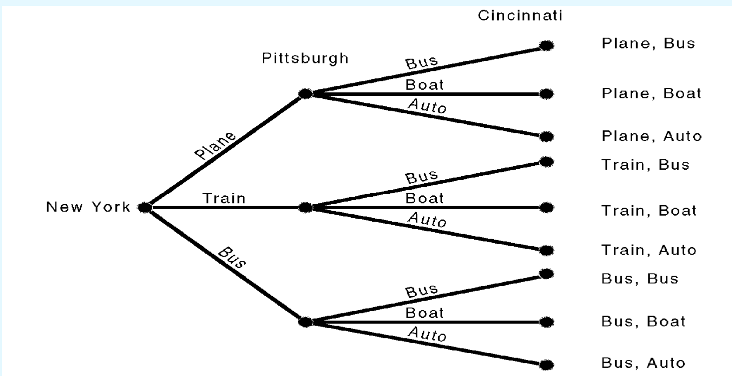
- Simple event as E={BBB}

- Compound event as E={BBG, BGB, GBB}

Basic Concepts

4-5

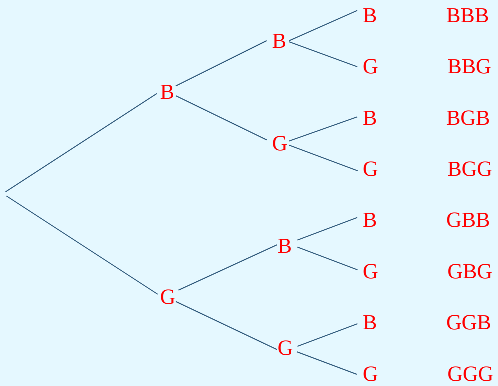
- A *tree diagram* is a device used to list all possibilities of a sequence of events in a systematic way.



Basic Concepts

4-6 **Example**

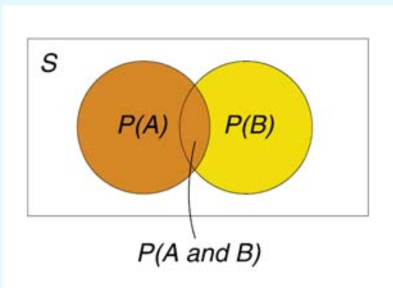
Find the sample space for the gender of the children if a family has three children. Use B for boy and G for girl.



Basic Concepts

4-7

- *Equally likely events* are events that have the same probability of occurring.
- *Venn diagrams* are used to represent probabilities pictorially.



Classical Probability

4-8

- *Classical probability* uses sample spaces to determine the numerical probability that an event will happen. It assumes that all outcomes in the sample space are equally likely to occur.
- The probability of an event E can be defined as

$$P(E) = \frac{n(E)}{n(S)} = \frac{Number\ of\ outcomes\ in\ E}{Number\ of\ outocmes\ in\ the\ sample\ space}$$

Classical Probability

4-9 **Example**

If a family has three children, find the probability that two of the children are girls .

The sample space is

S ={BBB, BBG, BGB, GBB, GGG, GGB, GBG, BGG}

The event of two girls is

E ={GGB, GBG, BGG}

Hence, the probability that two of the children are girls is

$$\boldsymbol{P(E) = \frac{n(E)}{n(S)} = \frac{3}{8} = 0.375}$$

Probability Rules

4-10

1. The probability of an event E is a number (either a fraction or decimal) between and including 0 and 1. Thus, $0 \le P(E) \le 1$.
2. If an event E cannot occur (i.e., the event contains no members in the sample space), the probability is zero.
3. If an event E is certain, then the probability of E is one.
4. The sum of the probabilities of the outcomes in the sample space is one.

Probability Rules

4-11 **Example**

When a single die is rolled, find the probability of getting a nine.

Since the sample space is S={1, 2, 3, 4, 5, 6}, it is impossible to get a 9. Hence,

$$P(9) = \frac{0}{6} = 0$$

When a single die is rolled, what is the probability of getting a number less than 7?

Since all outcomes in the sample space are less than 7, then

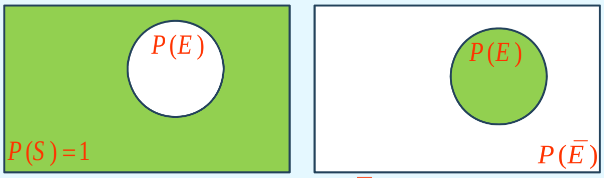
$$P(< 7) = \frac{6}{6} = 1$$

Complementary Events

4-12

- The *complement of an event* E is the set of outcomes in the sample space that are not included in the outcomes of event E. The complement of E is denoted by $\bar{E}$.
- Rule for Complementary Events, $P(E) + P(\bar{E}) = 1$

  Thus, $P(\bar{E}) = 1 - P(E)$ or $P(E) = 1 - P(\bar{E})$.
- Complementary events are mutually exclusive.



Simple Probability

$P(\bar{E}) = 1 - P(E)$

Complementary Events

4-13 **Example**

Find the complement of each event.

a. Rolling a die and getting a 4.

Getting 1,2,3,5 or 6

b. Selecting a letter of the alphabet and getting a vowel.

Getting a consonant

c. Selecting a month and getting a month that begins with a J.

Getting February, March, April, May, August, September, October, November or December

d. Selecting a day of the week and getting a weekday.

Getting Friday or Saturday

Complementary Events

4-14 **Example**

If the probability that a person lives in an industrialized country of the world is $\frac{1}{5}$, find the probability that a person does not live in an industrialized country.

$$P(not\ living\ in\ an\ industrialized\ country) = 1 - \frac{1}{5} = \frac{4}{5} = 0.8$$

Empirical Probability

4-15

- *Empirical probability* relies on actual experience to determine the likelihood of outcomes. It doesn't assumes that all outcomes in the sample space are equally likely to occur.
- Given a frequency distribution, the probability of an event being in a given class is:

$$P(E) = \frac{frequency\ for\ the\ class}{sample\ size} = \frac{f}{n}$$

Empirical Probability

4-16 **Example**

In a sample of 50 people, 21 had type O blood. 22 had type A, 5 had type B blood and 2 had type AB blood. Set up a frequency distribution and find the following probabilities:

Blood Type	Frequency
A	22
B	5
O	21
AB	2
Total	50

a. A person has type O blood.

$$P(O) = \frac{21}{50} = 0.42$$

b. A person has type A or type B blood.

$$P(A\ or\ B) = \frac{22+5}{50} = 0.54$$

c. A person has neither type A nor type O blood.

$$P(neither\ A\ nor\ O) = \frac{5+2}{50} = 1 - \frac{22+21}{50} = 0.14$$

d. A person does not have type AB blood.

$$P(not\ AB) = 1 - 2/50 = 0.04$$

Mutually Exclusive Events

4-17

- Two events are *mutually exclusive* if they cannot occur at the same time (i.e., they have no outcomes in common).
- Example: rolling a die and getting an even or an odd number.
- In the case of mutually exclusive events, the probability of two or more events can be determined by the *addition rules*.

Mutually Exclusive Events

4-18 **Example**

Determine which events are mutually exclusive and which are not, when a single die is rolled.

a. Getting a 3 and getting an odd number.

The events are not mutually exclusive, since the first event is a 3 and then second event is 1, 3 or 5. Hence, 3 is contained in both events.

b. Getting an odd number and getting a number less than 4.

The events are not mutually exclusive, since the first event can be 1, 3 or 5 and the second event is 1, 2 or 3. Hence, 1 and 3 are contained in both events.

c. Getting a number greater than 4 and getting a number less than 4.

The events are mutually exclusive, since the first event is 5 or 6 and the second event is 1, 2 or 3.

Addition Rules

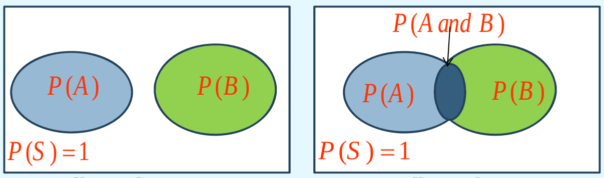
4-19

- When two events A and B are *mutually exclusive*, the probability that A or B will occur is

$$P(A\ or\ B) = P(AP) + P(B)$$

- When two events A and B are *not mutually exclusive*, the probability that A or B will occur is

$$P(A\ or\ B) = P(AP) + P(B) - P(A \,\&\, B)$$



Mutually Exclusive Events **Nonmutually Exclusive Events**

Addition Rules

4-20 **Example**

A box contains 3 glazed doughnuts, 4 jelly doughnuts and 5 chocolate doughnuts. If a person selects a doughnut at random, find the probability that it is either a glazed doughnut or a chocolate doughnut.

The total number of doughnuts in the box is 12 and the events are mutually exclusive, so

$$\boldsymbol{P(G\ or\ C) = P(G) + P(C) = \frac{3}{12} + \frac{5}{12} = \frac{8}{12} = 0.667}$$

Addition Rules

4-21 **Example**

A day of the week is selected at random. Find the probability that it is a weekend day (Friday or Saturday)

The total number of days in a week is 7 (5 weekdays and 2 weekend) and the events are mutually exclusive, so

$$\boldsymbol{P(F\ or\ S) = P(F) + P(S) = \frac{1}{7} + \frac{1}{7} = \frac{2}{7} = 0.286}$$

Addition Rules

4-22 **Example**

In a hospital unit there are 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. If a staff is selected, find the probability that the subject is a nurse or a male.

The events are not mutually exclusive and the sample space is

Staff	Female	Males	Total
Nurses	7	1	8
Physicians	3	2	5
Total	10	3	13

$$\boldsymbol{P(N\ or\ M) = P(N) + P(M) - P(N\ \&\ M) = \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13} = 0.769}$$

Independent and Dependent Events

4-23

- Two events A and B are *independent* if the fact that A occurs does not affect the probability of B occurring.
- When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be *dependent*.
- The *multiplication rules* can be used to find the probability of two or more events that occur in sequence.
- When two events are independent, the probability of both occurring is:

$$\boldsymbol{P(A\ and\ B) = P(A) * P(B)}$$

Independent and Dependent Events

4-24 **Example**

An urn contains 3 red balls, 2 blue balls and 5 white balls. A ball is selected and its color noted. Then it is replaced. A second ball is selected and its color noted. Find the probability of each of these.

a. Selecting 2 blue balls

$$P(B \text{ and } B) = P(B) * P(B) = \frac{2}{10} * \frac{2}{10} = \frac{4}{100} = 0.04$$

b. Selecting 1 blue ball and then 1 white ball

$$P(B \text{ and } W) = P(B) * P(W) = \frac{2}{10} * \frac{5}{10} = \frac{10}{100} = 0.1$$

c. Selecting 1 red ball and then 1 blue ball

$$P(R \text{ and } B) = P(R) * P(B) = \frac{3}{10} * \frac{2}{10} = \frac{6}{100} = 0.06$$

Independent and Dependent Events

4-25 **Example**

Approximately 9% of men have a type of color blindness that prevents them from distinguishing between red and green. If 3 men are selected at random, find the probability that all of them will have this type of red-green color blindness.

Let C denote red-green color blindness. Then

$$P(C \text{ and } C \text{ and } C) = P(C) * P(C) * P(C) = \frac{9}{100} * \frac{9}{100} * \frac{9}{100} = 0.000729$$

Hence, the rounded probability is 0.0007

© All rights are preserved for Dr. Saeed A. Dobbah Alghamdi, Department of Statistics, Faculty of Sciences, KAU

Counting Rule

4-26

- The multiplication rule can be used to determine the total number of outcomes in a sequence of events.
- Fundamental counting rule

  In a sequence of n events in which the first one has k_1 possibilities and the second event has k_2 and the third has k_3 and so forth, the total number of possibilities of the sequence will be:

$$k_1 * k_2 * k_3 * \cdots * k_n$$

- Note: "And" in this case means to multiply.

Counting Rule

4-27 **Example**

A paint manufacturer whishes to manufacture several different paints. The categories include

Color	Red, blue, white, black, green, brown, yellow
Type	Latex, oil
Texture	Flat, semi gloss, high gloss
Use	Outdoor, indoor

How many different kinds of paint can be made if a person can select one color, one type, one texture and one use?

Since there are 7 color choices, 2 type choices, 3 texture choices and 2 use choices, then the total number of possible different paints is

$$\mathbf{7 * 2 * 3 * 2 = 84}$$

Counting Rule

4-28 **Example**

The digits 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9 are to be used in a four-digit ID card. How many different cards are possible if

- repetitions are permitted?

Since there are 4 spaces to fill and 10 choices for each space, then the number of possible different cards is

$$\mathbf{10 * 10 * 10 * 10 = 10^4 = 10000}$$

- repetitions are not permitted?

Since there are 4 spaces to fill and 10 choices for first space, 9 choices for the second space, 8 choices for the third space and 7 choices for fourth space, then the number of possible different cards is

$$\mathbf{10 * 9 * 8 * 7 = 5040}$$

Permutations

4-29

- The arrangement of n objects in a specific order using r objects at a time is called a *permutation of n objects taking r objects at a time*. It is written as ${}_nP_r$, and the formula is

$${}_nP_r = \frac{n!}{(n-r)!}$$

where

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1$$
$$0! = 1$$

Permutations

4-30 **Example**

Suppose a business owner has a choice of five locations in which to establish his business. He decide to rank each location according to certain criteria, such as price of the store and parking facilities. How many different ways can he rank the five locations?

Since there are 5 choices for the first location, 4 choices for the second location, 3 choices for the third location, 2 choices for the fourth location and 1 choice for the last location, then the number of ways is

$$ {}_5P_5 = \frac{5!}{(5-5)!} = 5! = 5 * 4 * 3 * 2 * 1 = 120 $$

Application Using Excel

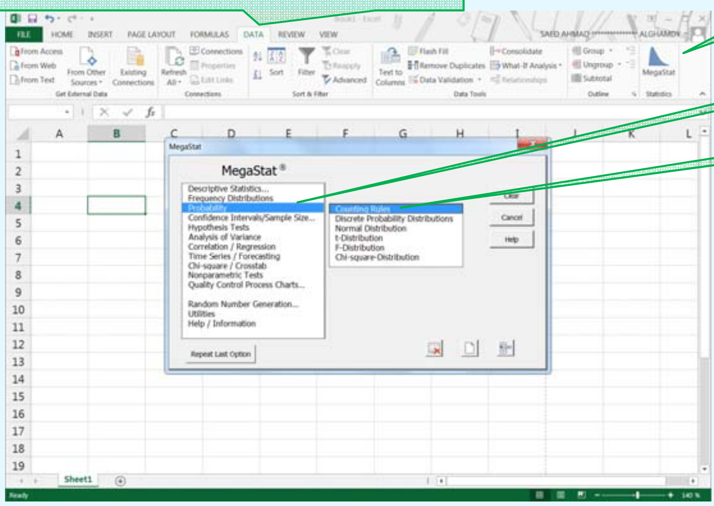
4-31 **Example "Permutation"**

Step 1: Select "DATA" tab or "ADD-INS" tab in older versions

Step 2

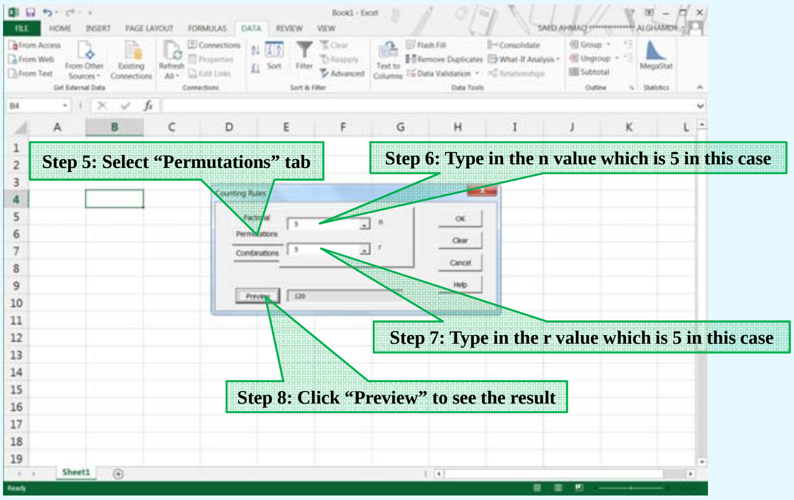
Step 3

Step 4



Application Using Excel

4-32 Example "Permutation"



Permutations

4-33 Example

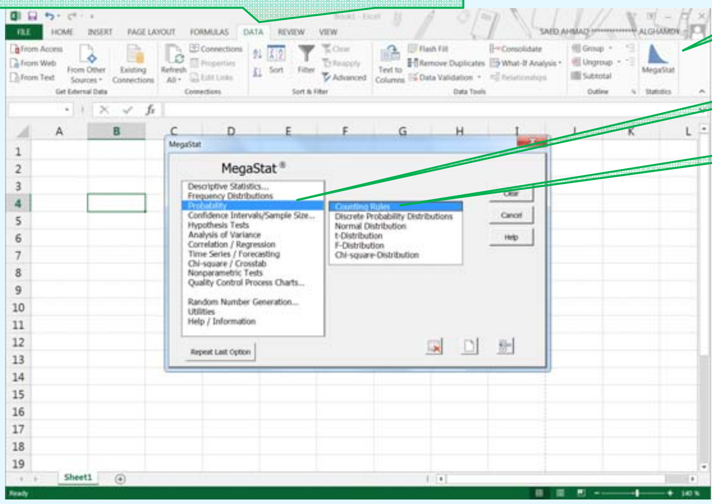
A television news director wishes to use three news stories on an evening show. One story will be the lead story, one will be the second story and the last will be a closing story. If the director has a total of eight stories to choose from, how many possible ways can the program be set up?

Since the order is important, then the number of ways to set up the program is

$$_8P_3 = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{8*7*6*5!}{5!} = 8*7*6 = 336$$

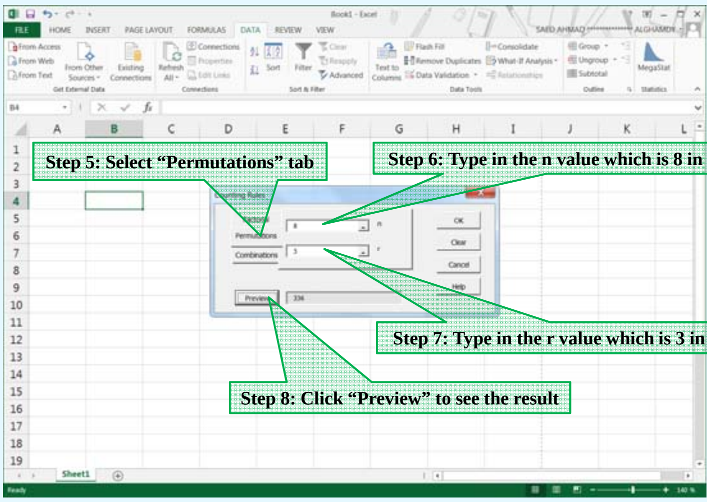
Application Using Excel

4-34 **Example “Permutation”**



Application Using Excel

4-35 **Example “Permutation”**



Combinations

4-36

- A selection of distinct objects without regard to order is called a *combination*.

- The number of combinations of r objects selected from n objects is denoted ${}_nC_r$ and is given by the formula

$${}_nC_r = \frac{n!}{(n-r)! * r!}$$

Combinations

4-37 **Example**

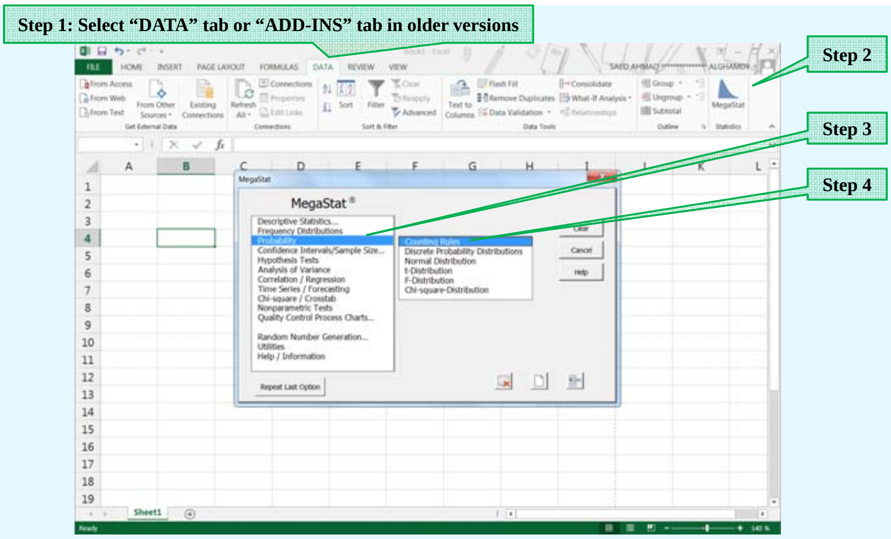
How many combination of 4 objects are there, taken 2 at a time?

Since this is a combination problem, then

$${}_4C_2 = \frac{4!}{(4-2)! * 2!} = \frac{4 * 3 * 2!}{2! * 2!} = \frac{4 * 3}{2 * 1} = 6$$

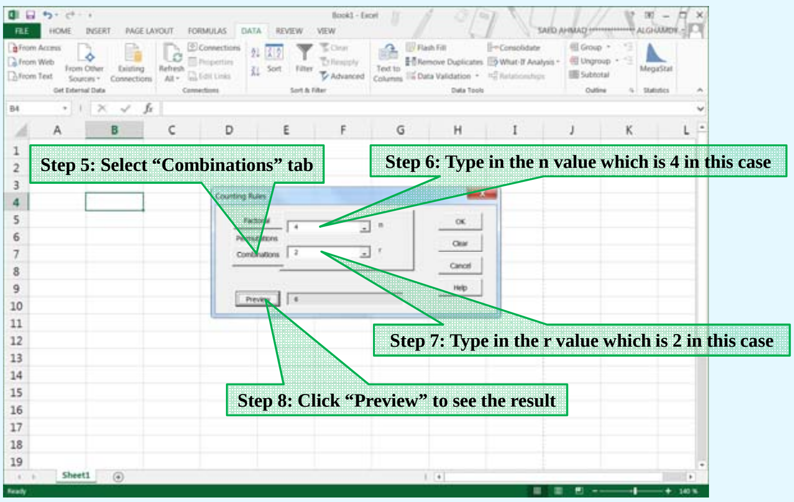
Application Using Excel

4-38 Example “Combination”



Application Using Excel

4-39 Example “Combination”



Combinations

4-40 **Example**

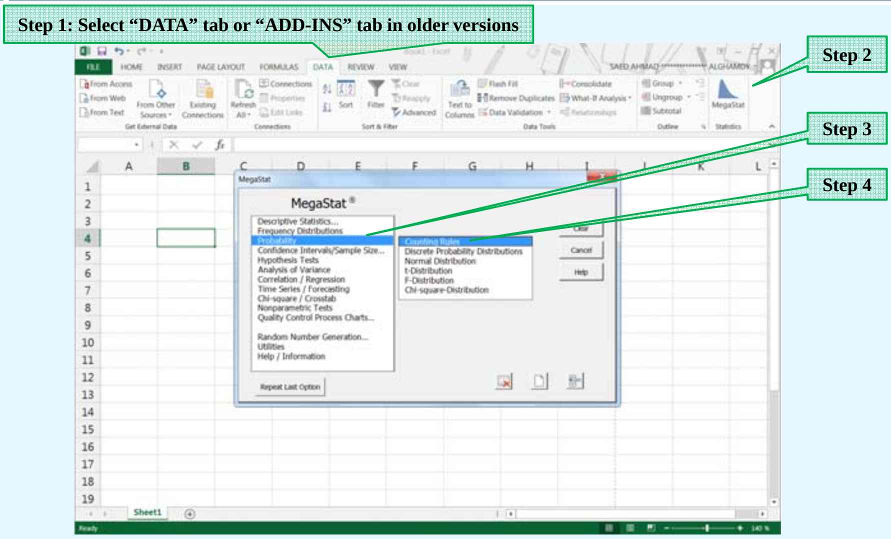
In a club there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

Here, one must selects 3 women from 7 women and selects 2 men from 5 men. Then, using the fundamental counting rule we can find the total number of different possibilities.

$$ {}_7C_3 * {}_5C_2 = \frac{7!}{(7-3)! * 2!} * \frac{5!}{(5-2)! * 2!} = 350 $$

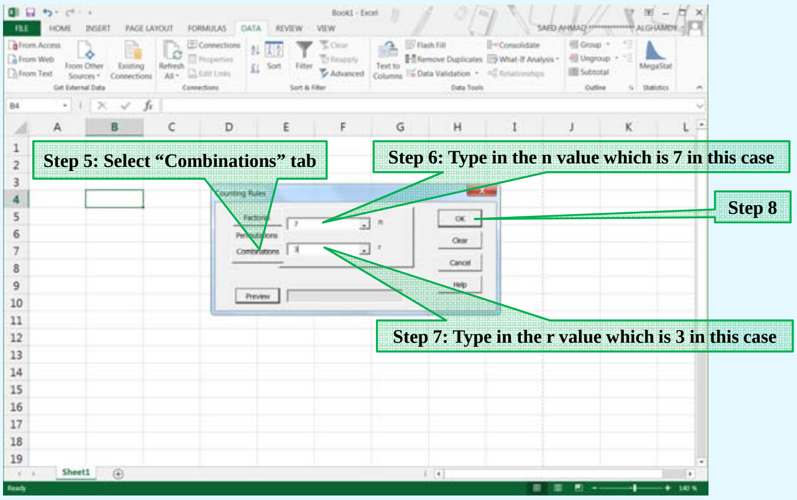
Application Using Excel

4-41 **Example "Combination"**



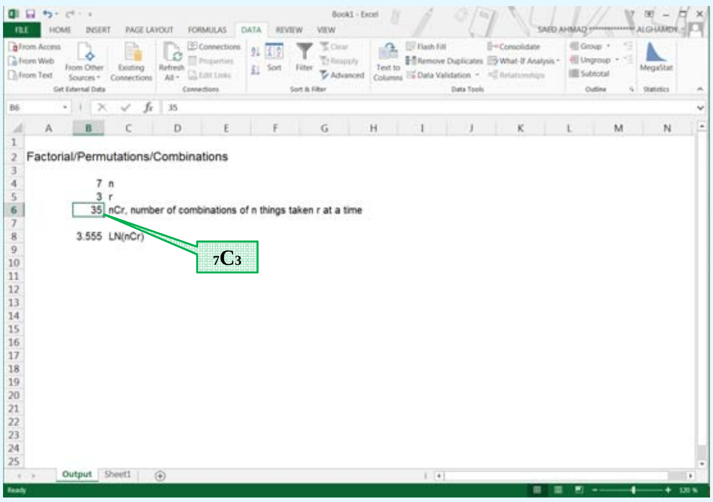
Application Using Excel

4-42 Example “Combination”



Application Using Excel

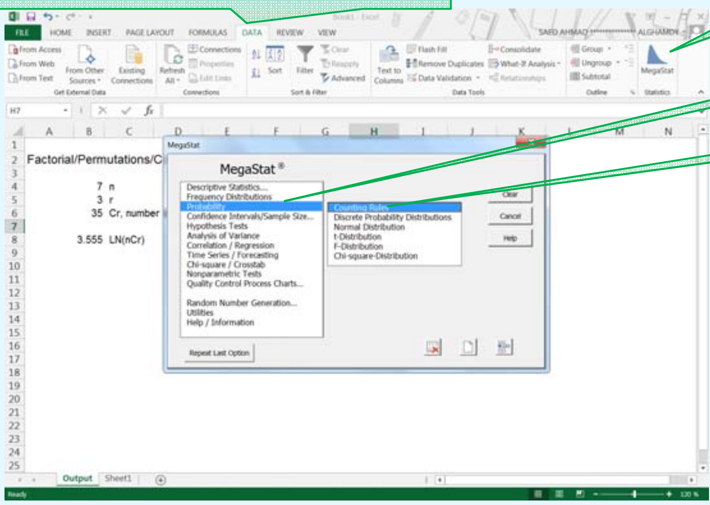
4-43 Example “Combination”



Application Using Excel

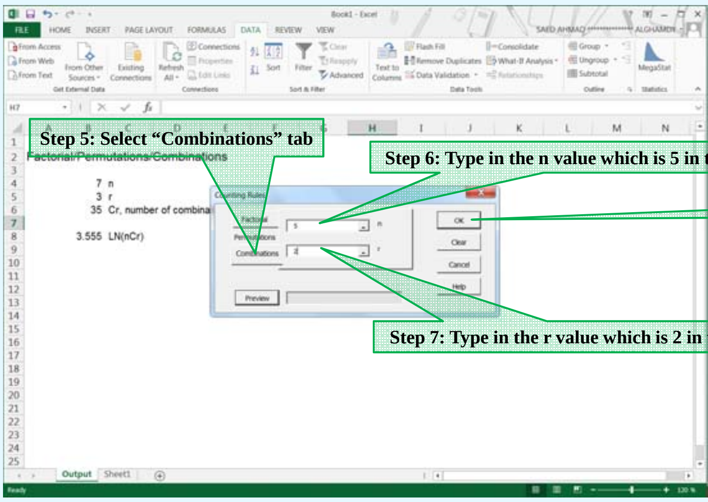
4-44 **Example "Combination"**

Step 1: Select "DATA" tab or "ADD-INS" tab in older versions



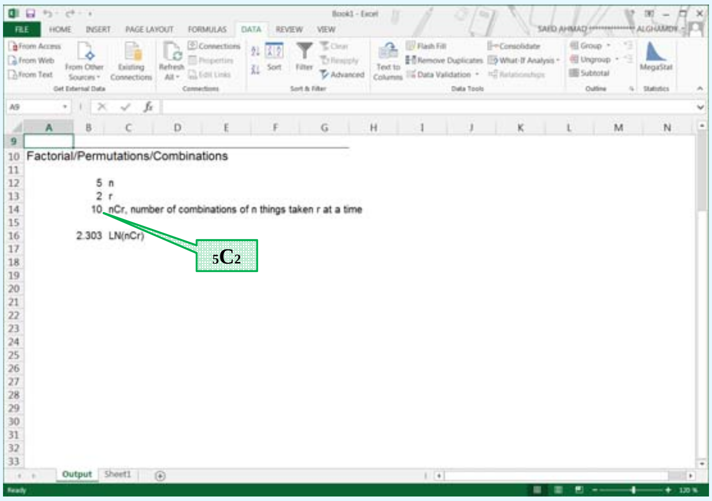
Application Using Excel

4-45 **Example "Combination"**



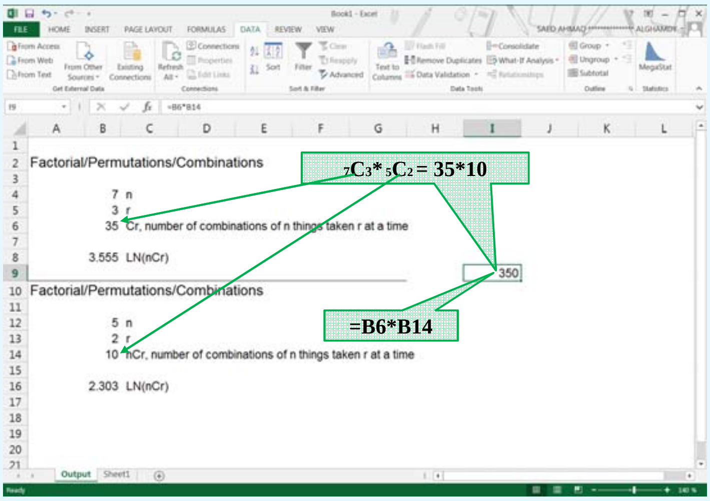
Application Using Excel

4-46 **Example “Combination”**



Application Using Excel

4-47 **Example “Combination”**



Summary

4-48

- The two types of probability are *classical* and *empirical*.
- Classical probability uses sample spaces and assumes that all outcomes in the sample space are equally likely to occur.
- Empirical probability uses frequency distributions and is based on observations.
- Two events are said to be mutually exclusive if they cannot occur together at the same time.
- Events can be independent or dependent if they occur in sequence.
- If events are independent, whether or not the first event occurs does not affect the probability of the next event occurring.
- If the probability of the second event occurring is changed by the occurrence of the first event, then the events are dependent.

Summary

4-49

Rule	Definition
Multiplication rule $k_1 \cdot k_2 \cdot k_3 \cdot \cdots \cdot k_n$	The number of ways a sequence of n events can occur; if the first event can occur in k_1 ways, the second event can occur in k_2 ways, etc.
Permutation rule $_nP_r = \frac{n!}{(n-r)!}$	The arrangement of n objects in a specific order using r objects at a time (order is important)
Combination rule $_nC_r = \frac{n!}{(n-r)!r!}$	The number of combinations of r objects selected from n objects (order is not important)

Review Examples

A combination lock consists of the 26 letters of the alphabet. If a 3-letter combination is needed, find the probability that the combination will consist of the letters ABC in that order. The same letter can be used more than once.

Review Examples

There are 8 married couples in a tennis club. If 1 man and 1 woman are selected at random to plan the summer tournament, find the probability that they are married to each other.

Review Examples

Given the letters A, B, C, and D, list the permutations and combinations for selecting two letters.

1-52

Review Examples

A school musical director can select 2 musical plays to present next year. One will be presented in the fall, and one will be presented in the spring. If she has 9 to pick from, how many different possibilities are there?

1-53

Review Examples

Suppose a business owner has a choice of 5 locations in which to establish his business. He decides to rank only the top 3 of the 5 locations. How many different ways can he rank them?

Review Examples

There are four blood types, A, B, AB, and O. Blood can also be +ve or -ve. Finally, a blood donor can be classified as either male or female. How many different ways can a donor have his or her blood labeled?

Review Examples

A coin is tossed and a die is rolled. Find the number of outcomes for the sequence of events.

1-56

Review Examples

A Harris poll found that 46% of Americans say they suffer great stress at least once a week. If three people are selected at random, find the probability that all three will say that they suffer great stress at least once a week.

1-57

Review Examples

A coin is flipped and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

1-58

Review Examples

The probability of a person driving with a friend is 0.32, the probability of a person having a driving accident is 0.09, and the probability of a person having a driving accident while driving with a friend is 0.15. What is the probability of a person driving with a friend or having a driving accident?

1-59

Review Examples

The corporate research and development centers for three local companies have the following number of employees:

U.S. Steel	110
Alcoa	750
Bayer Material Science	250

If a research employee is selected at random, find the probability that the employee is employed by U.S. Steel or Alcoa.

1-60

Review Examples

A city has 9 coffee shops: 3 Starbuck’s, 2 Caribou Coffees, and 4 Crazy Mocho Coffees. If a person selects one shop at random to buy a cup of coffee, find the probability that it is either a Starbuck’s or Crazy Mocho Coffees.

1-61

Review Examples

Hospital records indicated that knee replacement patients stayed in the hospital for the number of days shown in the distribution.

Number of days stayed	3	4	5	6	7	Σ
Frequency	15	32	56	19	5	127

Find these probabilities.

a. A patient stayed exactly 5 days.

b. A patient stayed at most 4 days.

c. A patient stayed less than 6 days.

d. A patient stayed at least 5 days.

Review Examples

Find the probability of getting identical number of spots when rolling two dice.

Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Chapter 5

Discrete Probability Distributions

Objectives

5-1

- Constructing a probability distribution for a random variable.
- Finding the mean, variance, and expected value for a discrete random variable.
- Finding the exact probability for x successes in n trials of a binomial experiment.
- Finding the mean, variance, and standard deviation for the variable of a binomial distribution.

Introduction

5-2

- Many decisions in business, insurance, and other real-life situations are made by assigning probabilities to all possible outcomes pertaining to the situation and then evaluating the results.
- This part explains the concepts and applications of probability distributions. In addition, a special probability distribution called *binomial distribution* is explained.

Discrete Probability Distribution

5-3

- A random variable is a variable whose values are determined by chance.
- A *discrete probability distribution* consists of the values a random variable can assume and the corresponding probabilities of the values. The probabilities are determined theoretically or by observation.

Discrete Probability Distribution

5-4 Example

Construct a probability distribution for rolling a single die.

Since the sample space is S={1,2,3,4,5,6} and each outcome has a probability $\frac{1}{6}$, the distribution will be

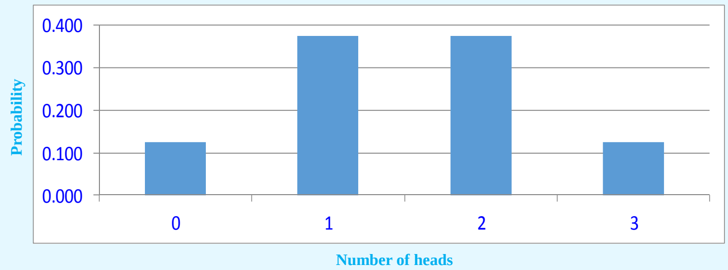
Outcome X	1	2	3	4	5	6
Probability P(X)	1/6	1/6	1/6	1/6	1/6	1/6

Discrete Probability Distribution

5-5 Example

Represent graphically the probability distribution for the sample space for tossing three coins.

Number of heads X	0	1	2	3
Probability P(X)	1/8	3/8	3/8	1/8



Discrete Probability Distribution

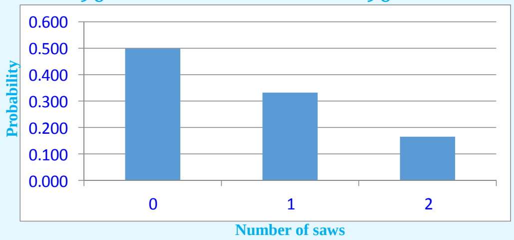
5-6 **Example**

During the summer months, a rental agency keeps track of the number of chain saws it rents each day during a period of 90 days. The number of saws rented per day is represented by the variable *X*. The results are shown here. Compute the probability *P(X)* for each X and construct a probability distribution and graph for the data.

X	0	1	2	Total
# of days	45	30	15	90

$$P(X=0)=\frac{45}{90}=0.5\,,P(X=2)=\frac{30}{90}=0.333\,,P(X=2)=\frac{15}{90}=0.167$$

X	0	1	2
P(X)	0.5	0.333	0.167



Requirements for a Probability Distribution

5-7

- The sum of the probabilities of all the events in the sample space must equal 1;

$$\sum X=1$$

- The probability of each event in the sample space must be between or equal to 0 and 1;

$$0\le P(X)\le 1$$

Requirements for a Probability Distribution

5-8 **Example**

Determine whether each distribution is a probability distribution.

a.

X	0	5	10	15	20
P(X)	1/5	1/5	1/5	1/5	1/5

b.

X	1	2	3	4
P(X)	1/4	1/8	1/16	9/16

c.

X	0	2	4	6
P(X)	-1.0	1.5	0.3	0.2

d.

X	2	3	7
P(X)	0.5	0.3	0.4

a. Yes, it is a probability distribution.

b. No, it is not a probability distribution, since P(X) cannot be 1.5 or -1.0.

c. Yes, it is a probability distribution.

d. No, it is not, since

Mean of a Probability Distribution

5-9

In order to find the mean for a probability distribution, one must multiply each possible outcome by its corresponding probability and find the sum of the products.

$$\mu = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum [X \times P(x)]$$

Mean of a Probability Distribution

5-10 **Example**

In a family with two children, find the mean of the number of children who will be girls.

The probability distribution is

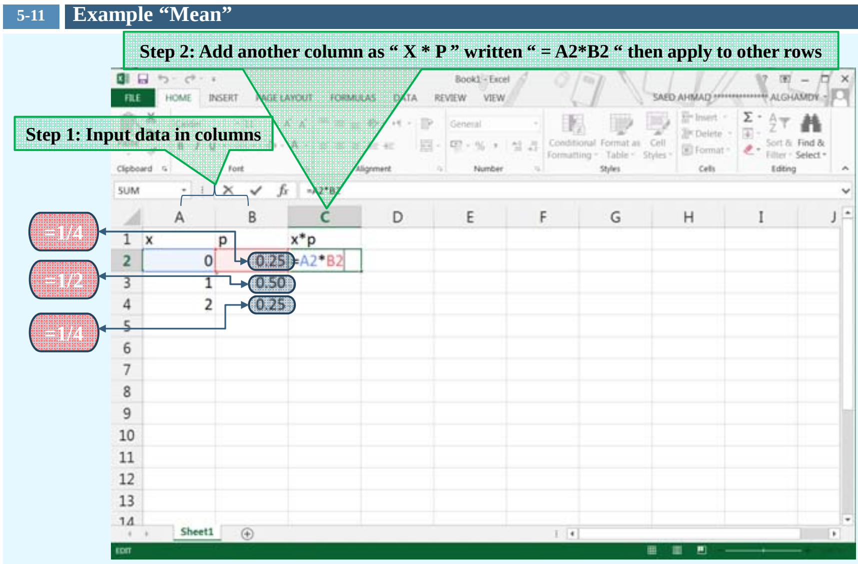
# of girls X	0	1	2
P(X)	1/4	1/2	1/4

Hence,

$$\mu = \sum [X \times P(x)] = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

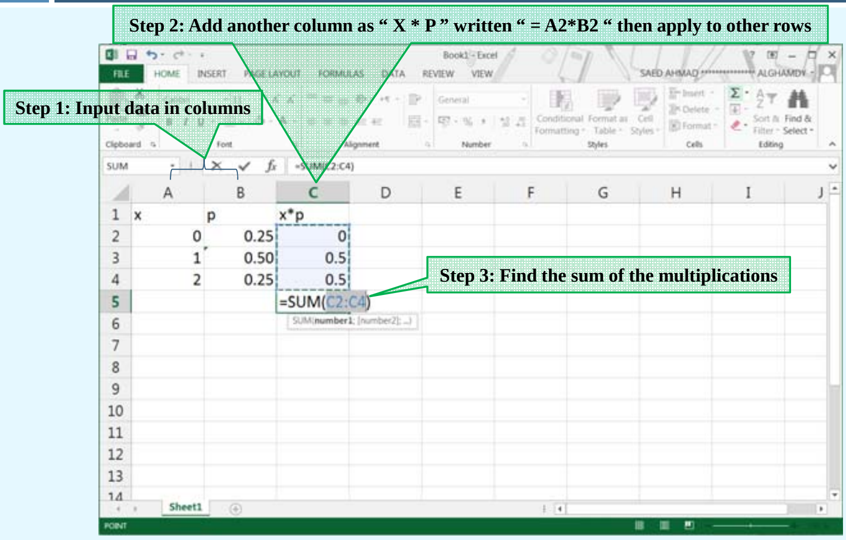
Application Using Excel

5-11 **Example "Mean"**



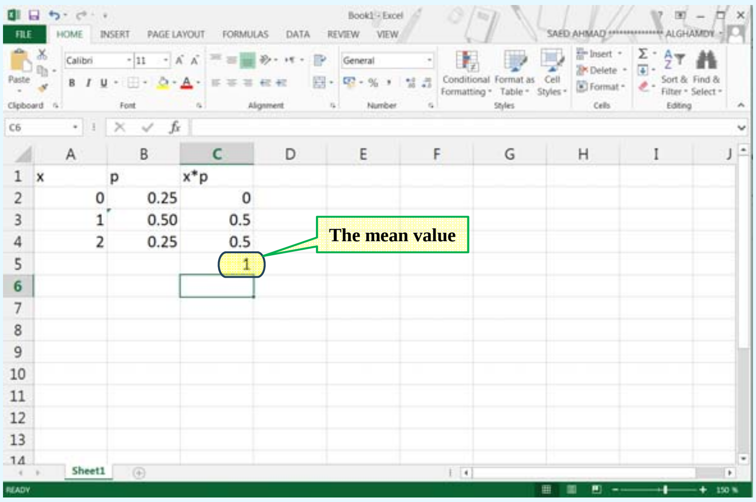
Application Using Excel

5-12 **Example "Mean"**



Application Using Excel

5-13 **Example "Mean"**



Mean of a Probability Distribution

5-14 **Example**

If three coins are tossed, find the mean of the number of heads that occur.

The probability distribution is

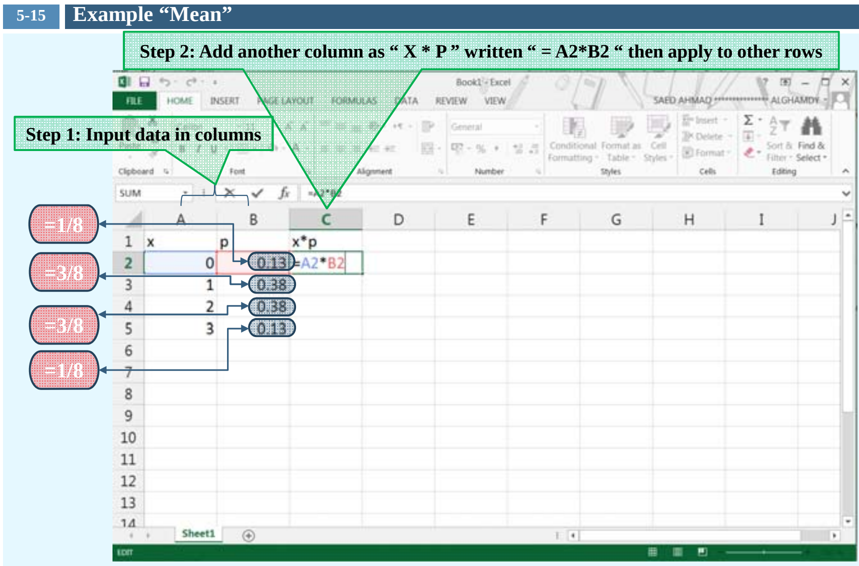
# of heads X	0	1	2	3
P(X)	1/8	3/8	3/8	1/8
X P(X)	0	3/8	6/8	3/8

Hence,

$$\mu = \sum [X \times P(x)] = 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = 1.5$$

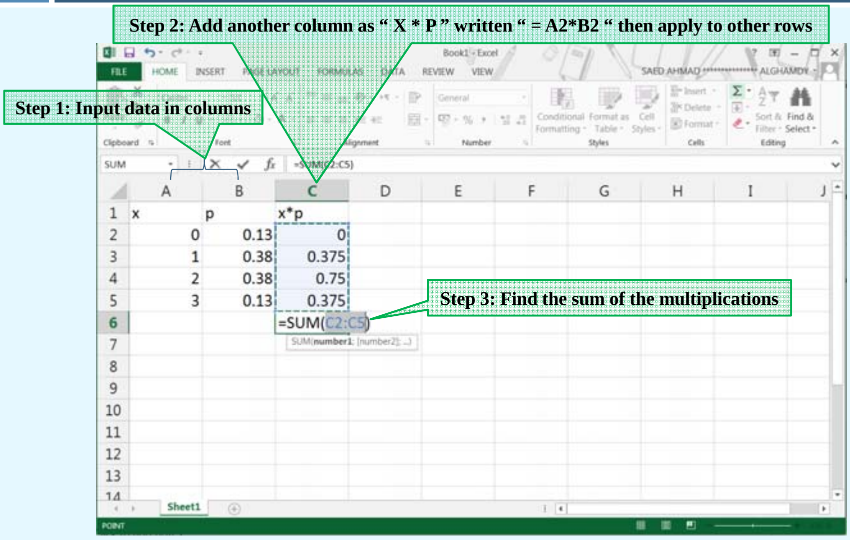
Application Using Excel

5-15 **Example "Mean"**



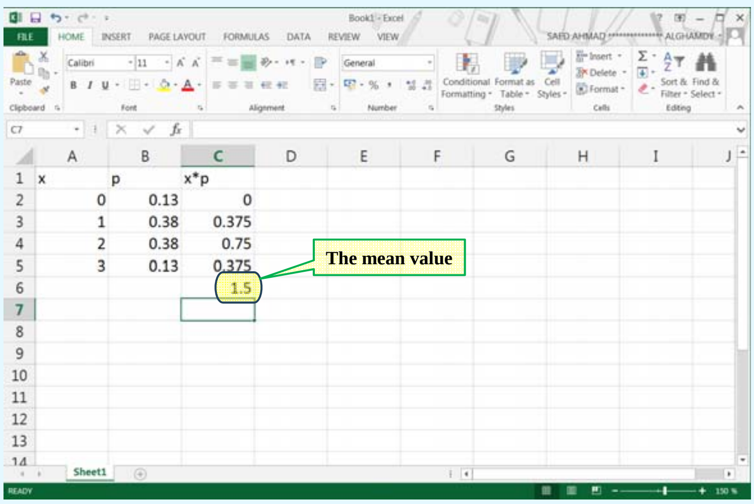
Application Using Excel

5-16 **Example “Mean”**



Application Using Excel

5-17 **Example “Mean”**



Mean of a Probability Distribution

5-18 Example

The probability distribution shown represents the number of trips of five nights or more that American adults take per year. Find the mean.

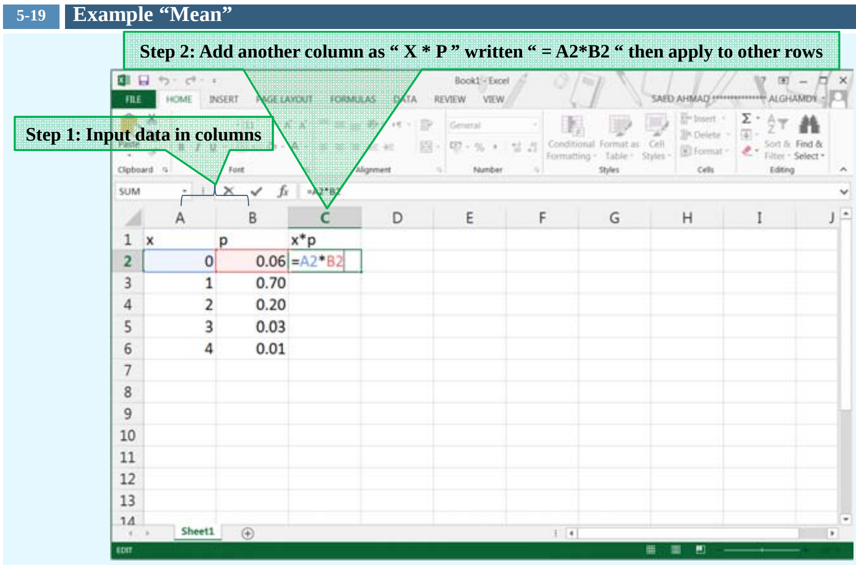
# of trips X	0	1	2	3	4
P(X)	0.06	0.70	0.20	0.03	0.01
X P(X)	0	0.70	0.40	0.09	0.04

Hence,

$$\mu = \sum[X \times P(x)] = 0 + 0.7 + 0.4 + 0.09 + 0.04 = 1.23$$

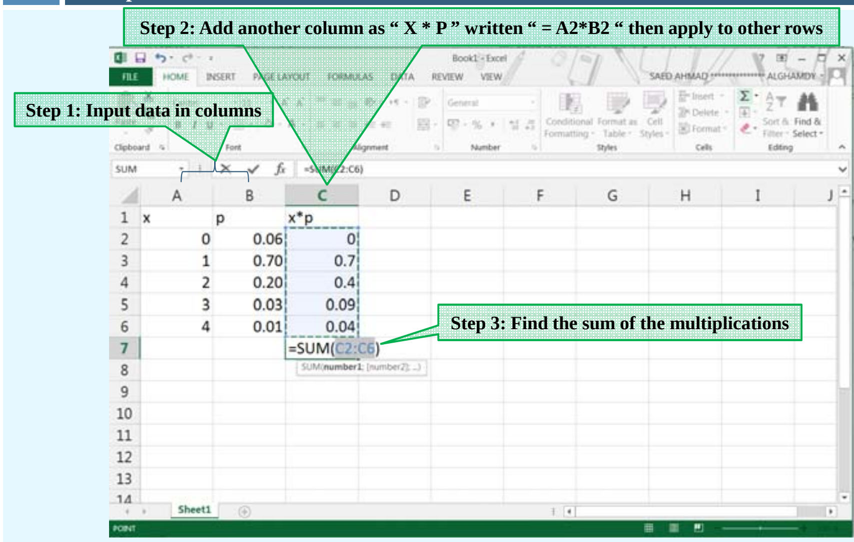
Application Using Excel

5-19 Example “Mean”



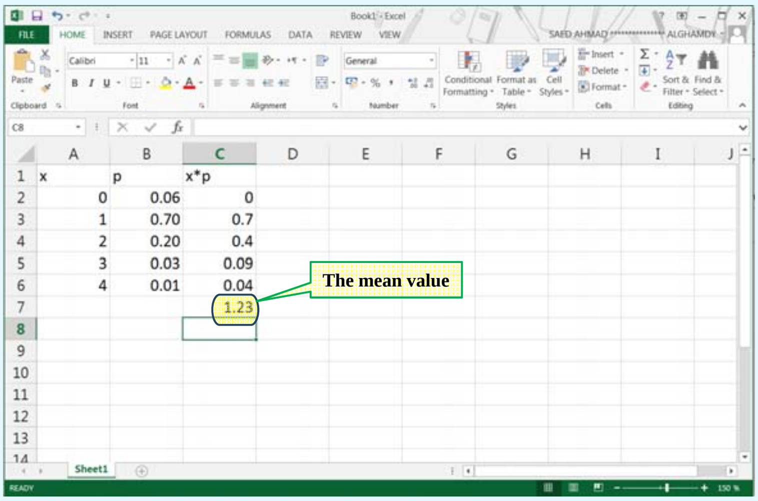
Application Using Excel

5-20 **Example “Mean”**



Application Using Excel

5-21 **Example “Mean”**



Variance of a Probability Distribution

5-22

- The variance of a probability distribution is found by multiplying the square of each outcome by its corresponding probability, summing those products, and subtracting the square of the mean.
- The formula for calculating the variance is:

$$\sigma^2 = \sum[X^2 \times P(X)] - \mu^2$$

- The formula for the standard deviation is:

$$\sigma = \sqrt{\sigma^2}$$

Variance of a Probability Distribution

5-23 **Example**

Find the mean, variance and standard deviation for the probability distribution for the number of spots that appear when a die is tossed is

Outcome X	1	2	3	4	5	6	Σ
Probability P(X)	1/6	1/6	1/6	1/6	1/6	1/6	1
$X \times P(X)$	1/6	2/6	3/6	4/6	5/6	6/6	21/6
$X^2 \times P(X)$	1/6	4/6	9/6	16/6	25/6	36/6	91/6

Hence,

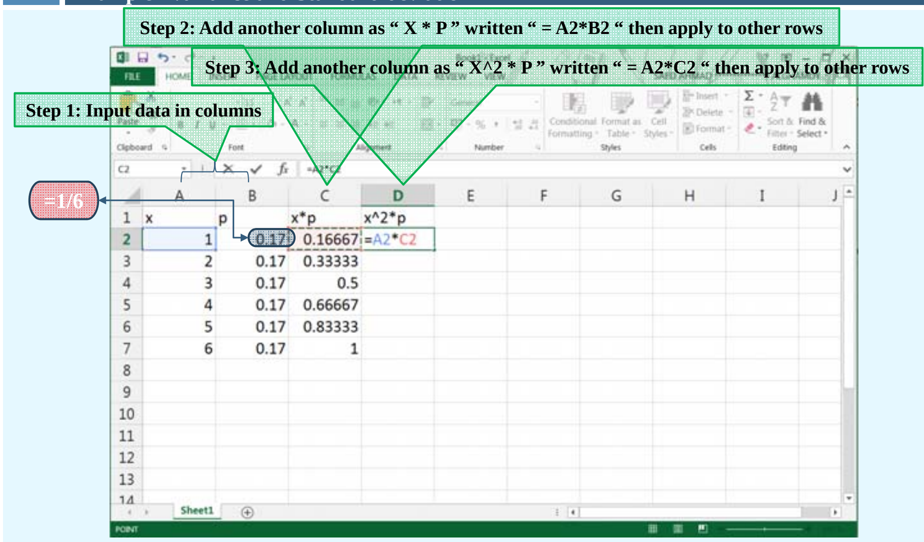
$$\mu = \sum[X \times P(x)] = 21/6$$

$$\sigma^2 = \sum[X^2 \times P(X)] - \mu^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = 2.917$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.917} = 1.708$$

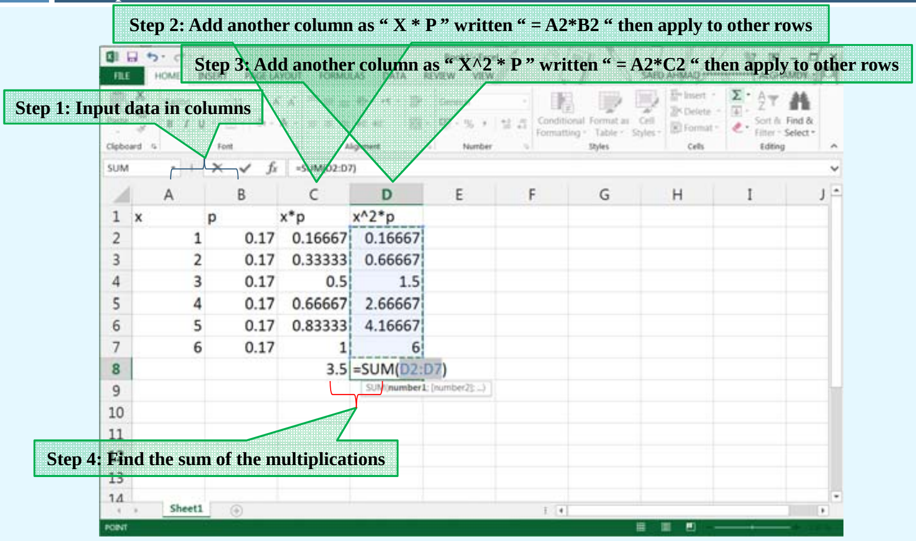
Application Using Excel

5-24 **Example “Variance and Standard deviation”**



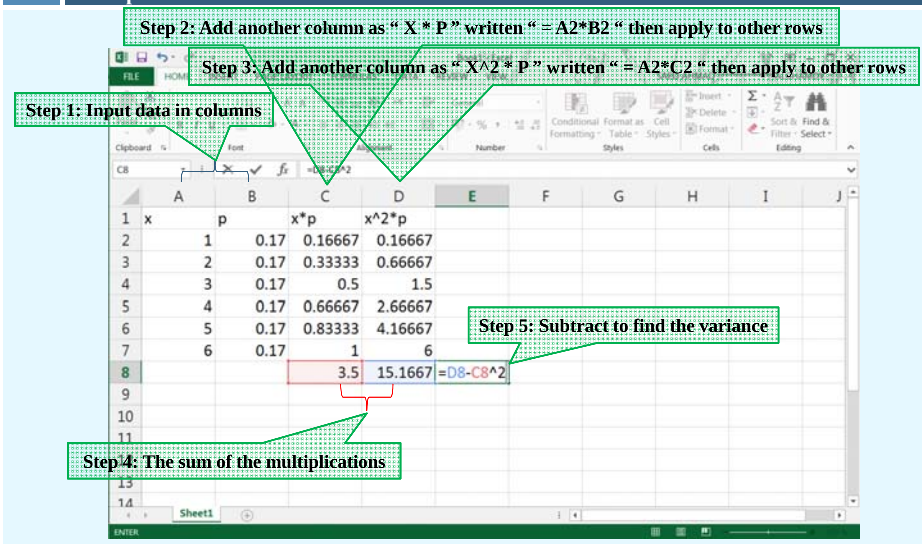
Application Using Excel

5-25 **Example “Variance and Standard deviation”**



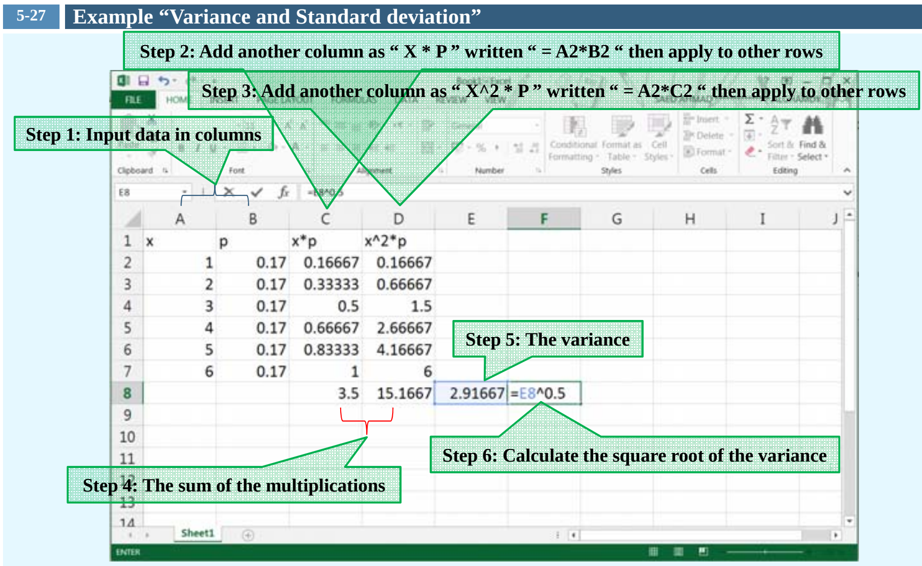
Application Using Excel

5-26 Example “Variance and Standard deviation”



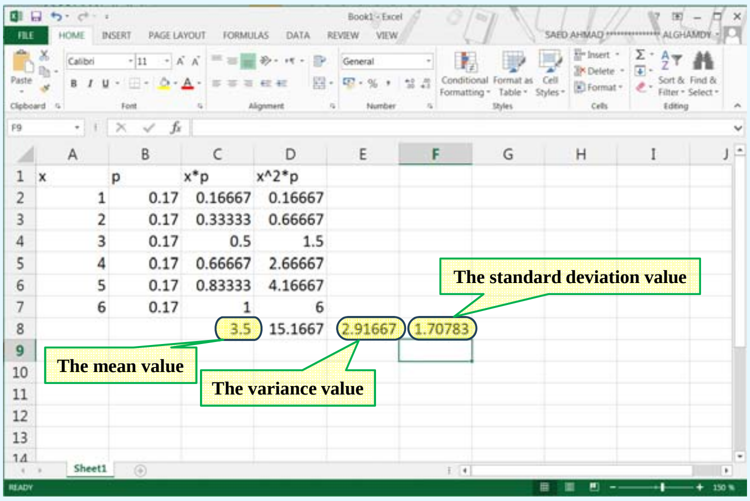
Application Using Excel

5-27 Example “Variance and Standard deviation”



Application Using Excel

5-28 Example "Variance and Standard deviation"



	A	B	C	D	E	F
1	x	p	x*p	x^2*p		
2	1	0.17	0.16667	0.16667		
3	2	0.17	0.33333	0.66667		
4	3	0.17	0.5	1.5		
5	4	0.17	0.66667	2.66667		
6	5	0.17	0.83333	4.16667		
7	6	0.17	1	6		
8			3.5	15.1667	2.91667	1.70783

Variance of a Probability Distribution

5-29 Example

Five balls numbered 0, 2, 4, 6 and 8 are placed in a bag. After the balls are mixed, one is selected, its number is noted and then it is replaced. If this experiment is repeated many times, find the variance and standard deviation of the numbers on the balls.

# on ball X	0	2	4	6	8	Σ
$P(X)$	1/5	1/5	1/5	1/5	1/5	1
$X \times P(X)$	0	2/5	4/5	6/5	8/5	20/5
$X^2 \times P(X)$	0	4/5	16/5	36/5	64/5	120/5

Hence,

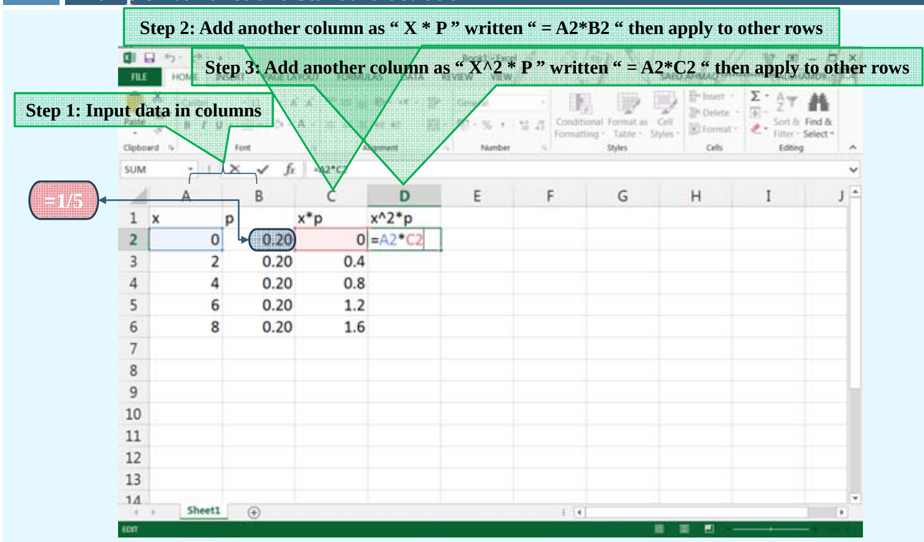
$$\mu = \sum [X \times P(x)] = \frac{20}{5} = 4$$

$$\sigma^2 = \sum [X^2 \times P(X)] - \mu^2 = \frac{120}{5} - 4^2 = 8$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{8} = 2.828$$

Application Using Excel

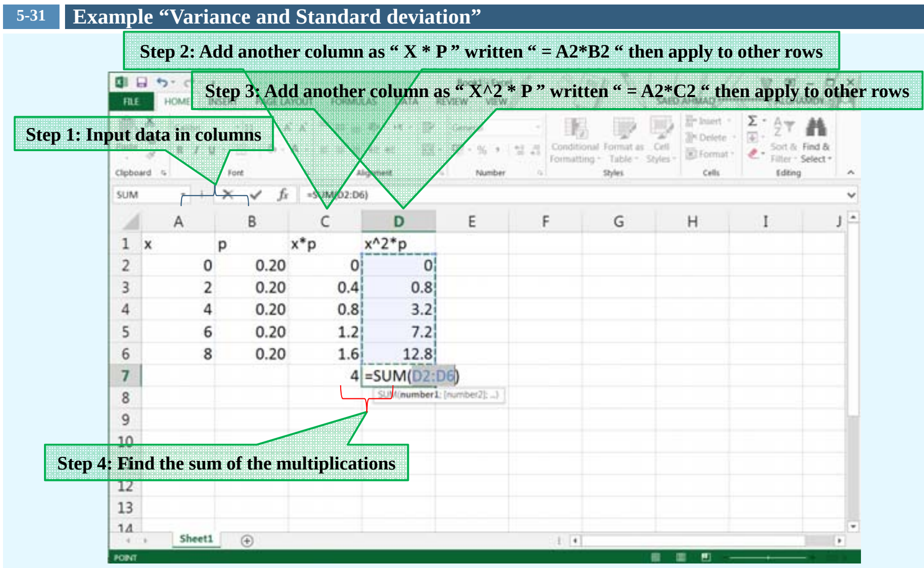
5-30 **Example “Variance and Standard deviation”**



	A	B	C	D
1	x	p	x*p	x^2*p
2	0	0.20	0	=A2*C2
3	2	0.20	0.4	
4	4	0.20	0.8	
5	6	0.20	1.2	
6	8	0.20	1.6	

Application Using Excel

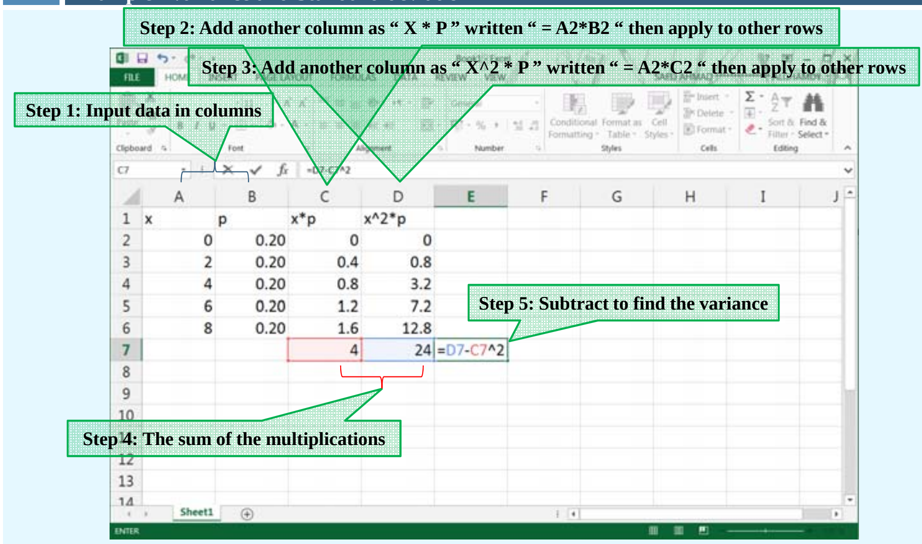
5-31 **Example “Variance and Standard deviation”**



	A	B	C	D
1	x	p	x*p	x^2*p
2	0	0.20	0	0
3	2	0.20	0.4	0.8
4	4	0.20	0.8	3.2
5	6	0.20	1.2	7.2
6	8	0.20	1.6	12.8
7			4	=SUM(D2:D6)

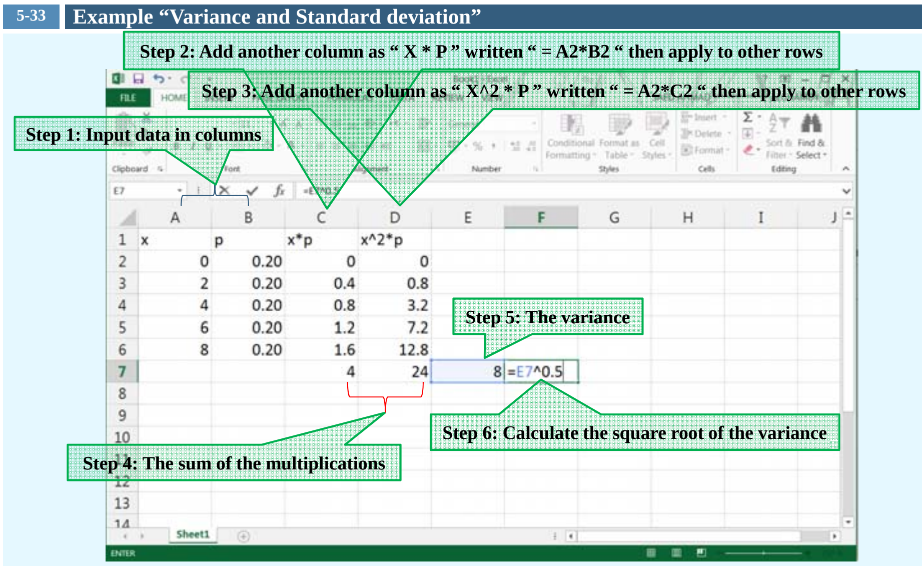
Application Using Excel

5-32 **Example “Variance and Standard deviation”**



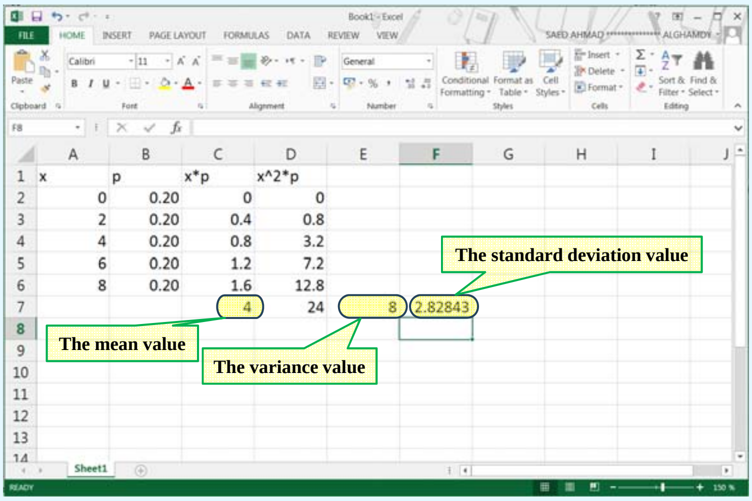
Application Using Excel

5-33 **Example “Variance and Standard deviation”**



Application Using Excel

5-34 **Example "Variance and Standard deviation"**



	A	B	C	D	E	F
1	x	p	x*p	x^2*p		
2	0	0.20	0	0		
3	2	0.20	0.4	0.8		
4	4	0.20	0.8	3.2		
5	6	0.20	1.2	7.2		
6	8	0.20	1.6	12.8		
7			4	24	8	2.82843

Expected Value

5-35

- *Expected value* or *expectation* is used in various types of games of chance, in insurance, and in other areas, such as decision theory.
- The *expected value* of a discrete random variable of a probability distribution is the theoretical average of the variable. The formula is:

$$E(X) = \mu = \sum [X \times P(x)]$$

- The symbol *E(X)* is used for the expected value.

Expected Value

5-36 **Example**

One thousand tickets are sold at $1 each for a color television valued at $350. What is the expected value of the gain if a person purchases one ticket?

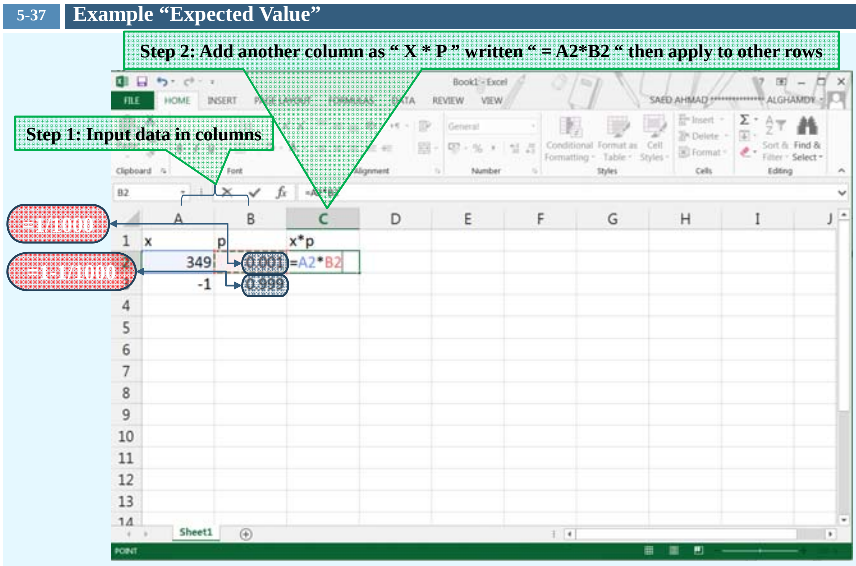
	Win	Lose	
X	\$349	-\$1	Σ
$P(X)$	1/1000	999/1000	1
$X \times P(X)$	349/1000	-999/1000	-650/1000

Hence, $E(X) = \sum[X \times P(x)] = -\frac{650}{1000} = -0.65$

The meaning of this value is that if a person purchased one ticket each week over a long time, the average loss would be $0.65 per ticket, since theoretically, on average, that person would win the set once for each 1000 tickets purchased.

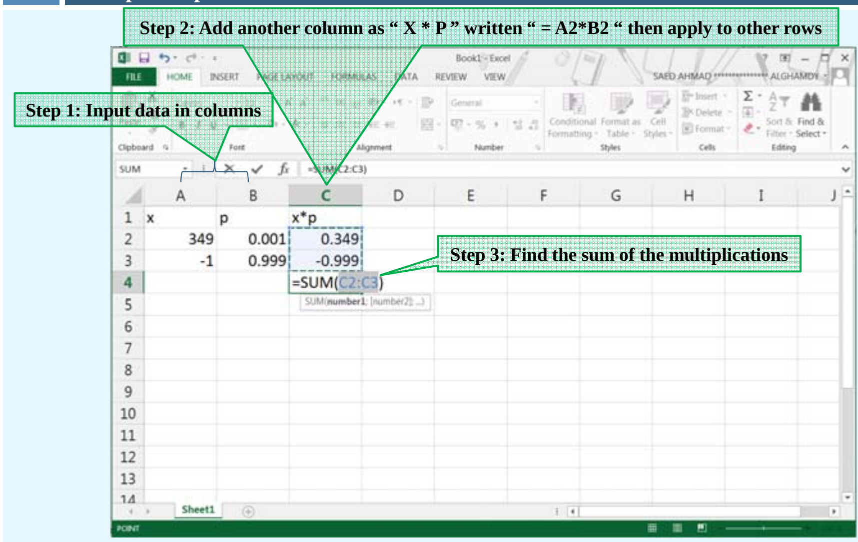
Application Using Excel

5-37 **Example "Expected Value"**



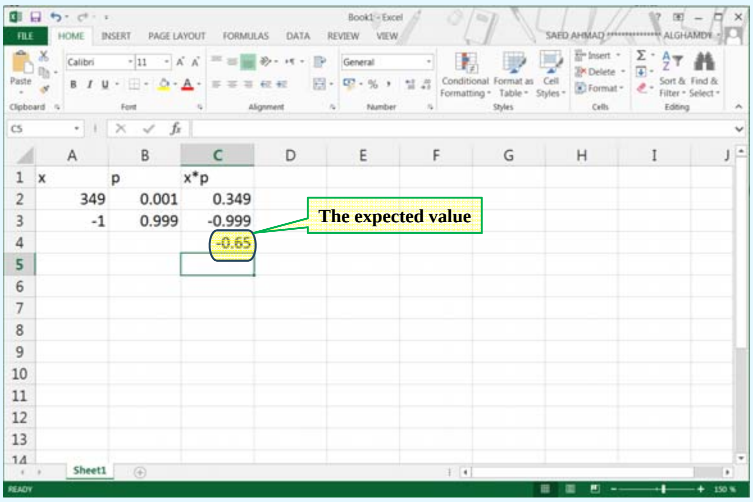
Application Using Excel

5-38 **Example “Expected Value”**



Application Using Excel

5-39 **Example “Expected Value”**



The Binomial Distribution

5-40

- Many types of probability problems have only two possible outcomes or they can be reduced to two outcomes.
- Examples:

  - When a coin is tossed it can land on heads or tails,

  - When a baby is born it is either a boy or girl,

  - A multiple-choice question can be classified as correct or incorrect.

The Binomial Experiment

5-41

- The binomial experiment is a probability experiment that satisfies these requirements:

1. Each trial can have only two possible outcomes; success or failure.
2. There must be a fixed number of trials.
3. The outcomes of each trial must be independent of each other.
4. The probability of a success must remain the same for each trial.

The Binomial Distribution

5-42

- The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a *binomial distribution* which is the probability of exactly x successes in n trials

$$P(x) = {}_nC_x \times p^x \times q^{n-x}$$

where

p the symbol for the numerical probability of success

q the symbol for the numerical probability of failure

$$p + q = 1$$

n the number of trials

x the number of successes; $x = 0,1,2,\dots,n$

The Binomial Distribution

5-43 **Example**

A coin is tossed 3 times. Find the probability of getting exactly two heads.

- This can solved using the sample space

  HHH,<u>HHT,HTH,THH</u>,HTT,THT,TTH,TTT

  There are three ways of getting 2 heads.

  So, $P(getting\ 2\ heads) = \frac{3}{8} = 0.375$

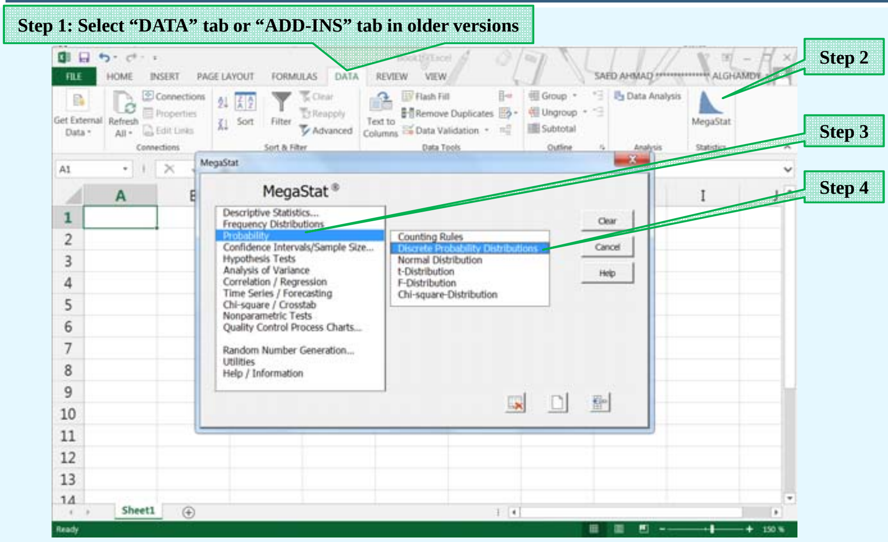
- Using the binomial distribution as following
  - we have fixed number of trials (three), so n=3
  - there are two outcomes for each trial, H or T
  - the outcomes are independent of one another
  - the probability of success (heads) is 1/2, so p=1/2

here X=2 since we need to find the probability of getting 2 heads.

Thus, $P(getting\ 2\ heads) = {}_3C_2 \times 0.5^2 \times 0.5^{3-2} = \frac{3}{8} = 0.375$

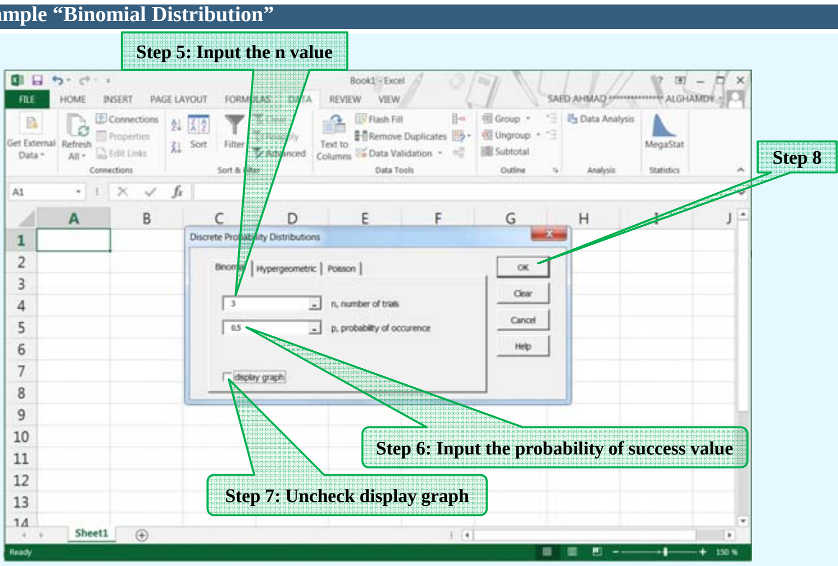
Application Using Excel

5-44 **Example "Binomial Distribution"**



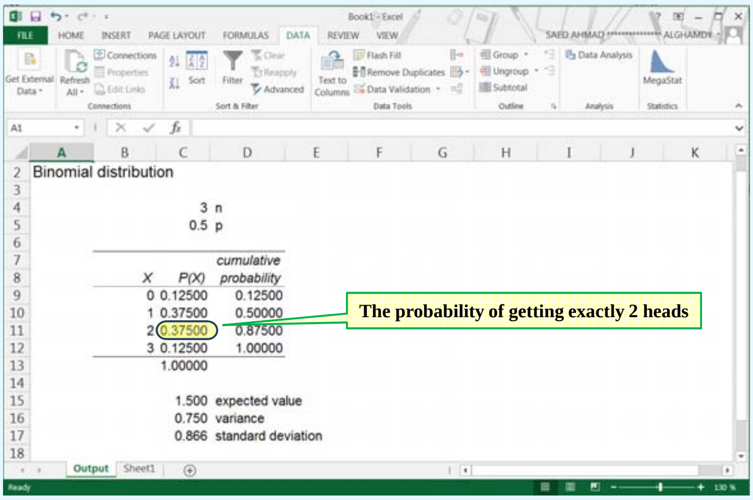
Application Using Excel

5-45 **Example "Binomial Distribution"**



Application Using Excel

5-46 **Example "Binomial Distribution"**



The Binomial Distribution

5-47 **Example**

A survey found that one out of five Americans says he or she has visited a doctor in any given month. If 10 people are selected at random, find the probability that exactly 3 would have visited a doctor last month.

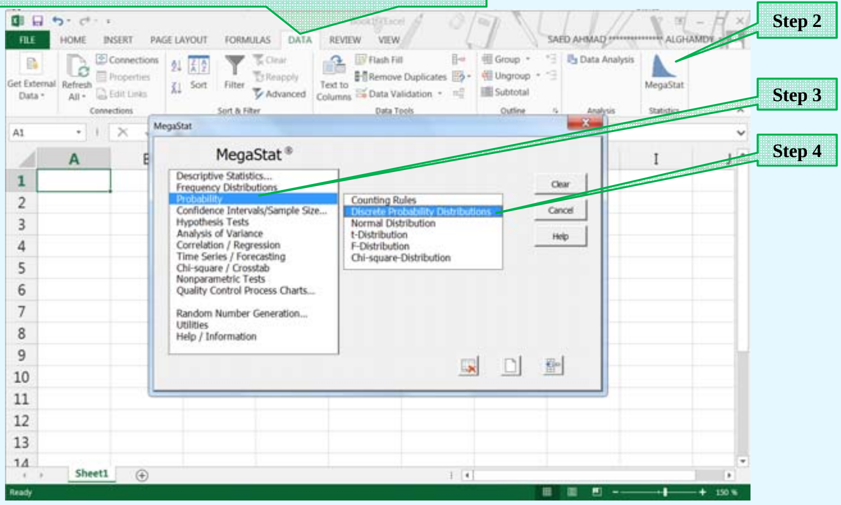
In this case *n=10, X=3, p=1/5*. So q=4/5 and

$$P(3) = \ _{10}C_3 \times \left(\frac{1}{5}\right)^3 \times \left(\frac{4}{5}\right)^{10-3} = 0.201$$

Application Using Excel

5-48 **Example “Binomial Distribution”**

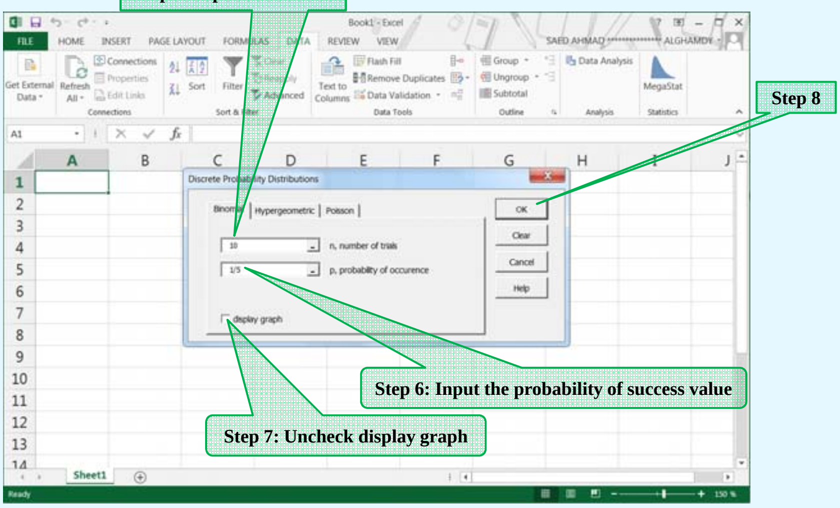
Step 1: Select “DATA” tab or “ADD-INS” tab in older versions



Application Using Excel

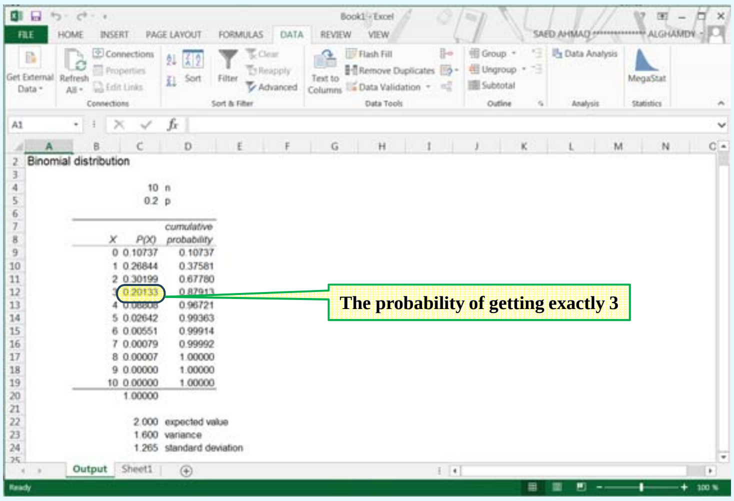
5-49 **Example “Binomial Distribution”**

Step 5: Input the n value



Application Using Excel

5-50 **Example "Binomial Distribution"**



The Binomial Distribution

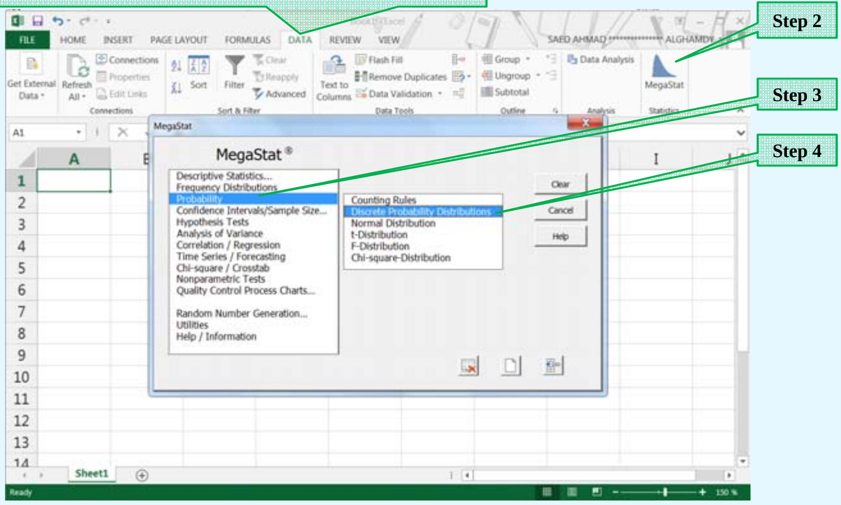
5-51 **Example**

A survey from Teenage Research Unlimited found that 30% of teenage consumers receive their spending money from part-time jobs. If 7 teenagers are selected at random, find

- the probability that exactly 3 of them will have part-time jobs.
- the probability that less than 3 of them will have part-time jobs.
- the probability that at most 3 of them will have part-time jobs.
- the probability that 5 or 6 of them will have part-time jobs.
- the probability that greater than 3 of them will have part-time jobs.
- the probability that at least 3 of them will have part-time jobs.

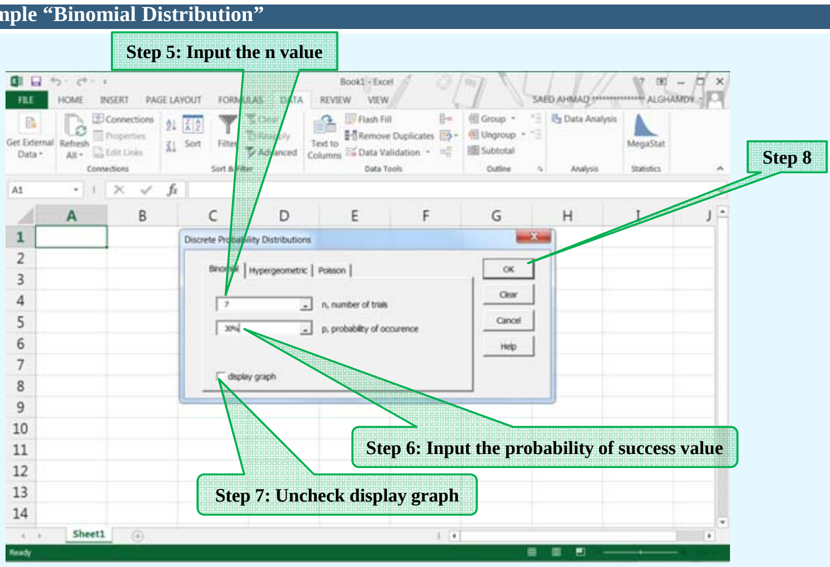
Application Using Excel

5-52 **Example “Binomial Distribution”**



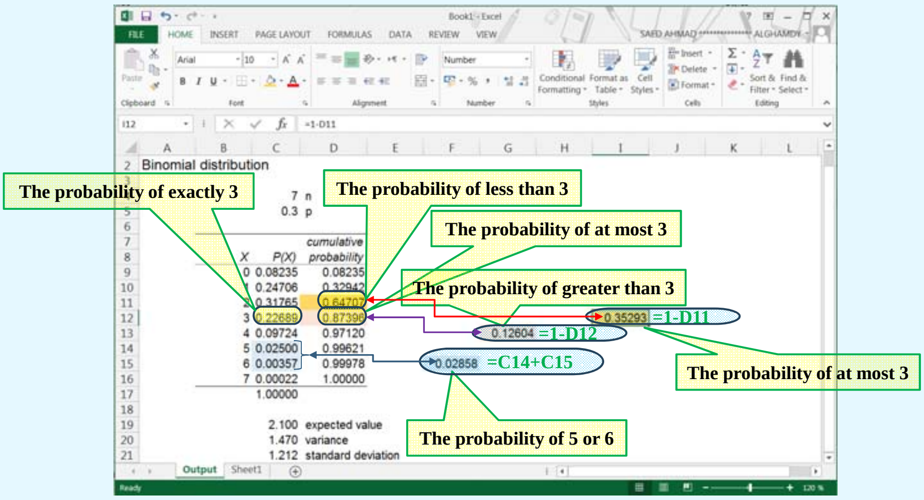
Application Using Excel

5-53 **Example “Binomial Distribution”**



Application Using Excel

5-54 **Example “Binomial Distribution”**



The Binomial Distribution

5-55

- The mean, variance, and standard deviation of a variable that has the binomial distribution can be found by using the following formulas.

mean $\mu = n \times p$

variance $\sigma^2 = n \times p \times q$

standard deviation $\sigma = \sqrt{\sigma^2}$

The Binomial Distribution

5-56 Example

A survey from Teenage Research Unlimited found that 30% of teenage consumers receive their spending money from part-time jobs. If 7 teenagers are selected at random, find the mean , variance and standard deviation of the number of teenagers who will have part time jobs.

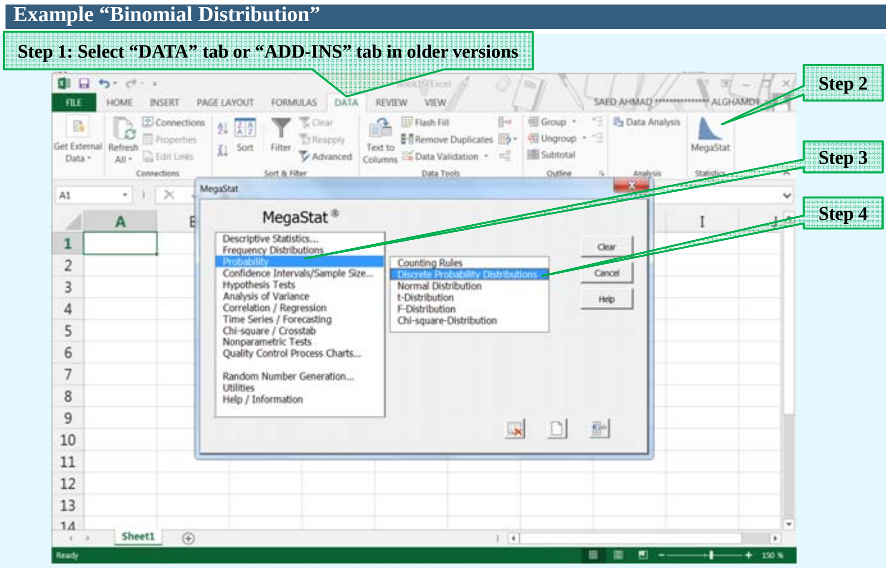
mean $\mu = n \times p = 7 \times 0.3 = 2.1$

variance $\sigma^2 = n \times p \times q = 7 \times 0.3 \times 0.7 = 1.47$

standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{1.47} = 1.212$

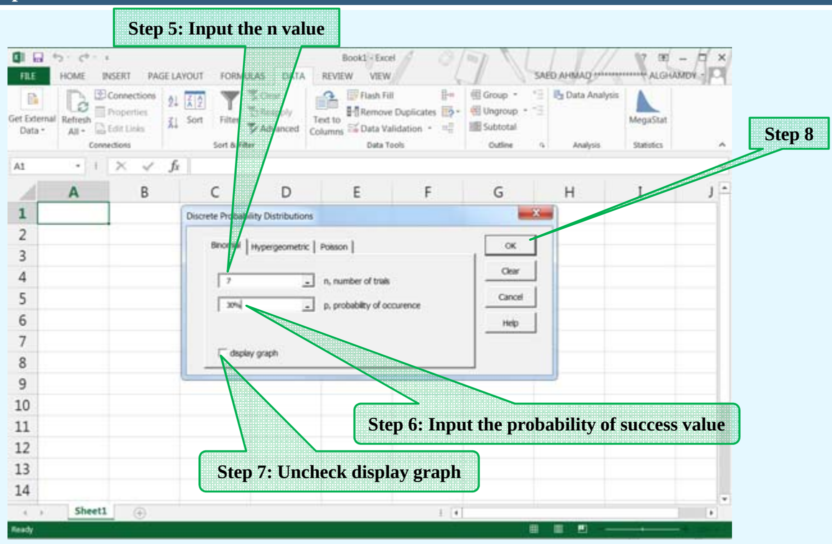
Application Using Excel

5-57 Example "Binomial Distribution"



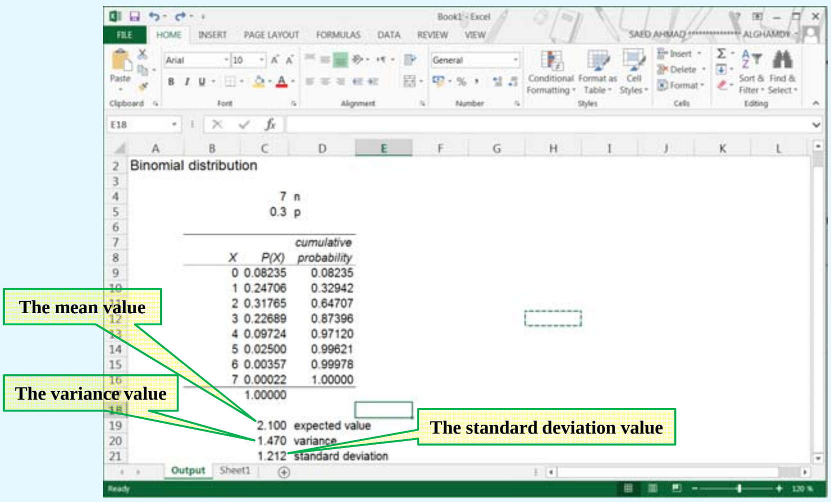
Application Using Excel

5-58 **Example “Binomial Distribution”**



Application Using Excel

5-59 **Example “The mean , variance and standard deviation of the Binomial Distribution”**



Summary

5-60

- A discrete probability distribution can be graphed using bar chart.
- The mean, variance, and standard deviation can be found for a probability distribution.
- The mathematical expectation can also be calculated for a probability distribution.
- The binomial distribution is used when there are only two outcomes for an experiment, a fixed number of trials, the probability is the same for each trial, and the outcomes are independent of each other.

Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Chapter 6

The Normal Distribution

Objectives

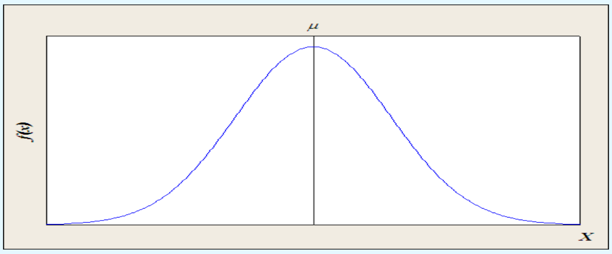
6-1

- Identifying the properties of the normal distribution.
- Finding the area under the standard normal distribution, given various z values.
- Finding the probability of a normally distributed variable by transforming it into a standard normal variable.
- Using the central limit theorem to solve problems involving sample means for large samples.

Introduction

6-2

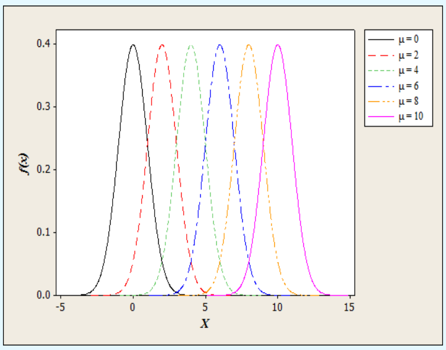
- Many continuous variables have distributions that are bell-shaped and are called *approximately normally distributed variables*, such as the heights, cholesterol level, etc...
- A *normal distribution* is a continuous, symmetric, bell-shaped distribution of a variable, which is also known as the *bell curve* or the *Gaussian distribution*.

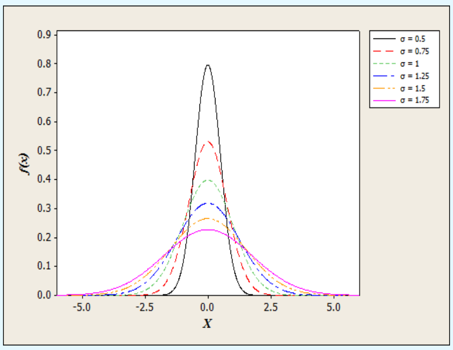


Normal Distribution Properties

6-3

- The normal distribution curve depends on two parameters, the *mean* (the position parameter) and the *standard deviation* (the shape parameter).





Normal Distribution Properties

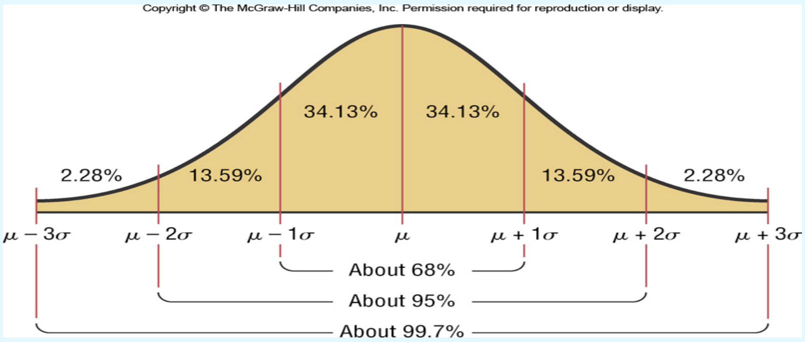
6-4

- The mean, median, and mode of the normal distribution are equal and located at the center of the distribution.
- The normal distribution curve is *unimodal* (i.e., it has only one mode).
- The curve of the normal distribution is continuous, i.e., there are no gaps. Thus, for each value of X, there is a corresponding value of Y.
- The total area under a normal distribution is equal to 1 or 100%. This fact may seem unusual, since the curve never touches the x-axis, but one can prove it mathematically by using calculus.

Normal Distribution Properties

6-5

- The area under the part of the normal curve that lies within 1 standard deviation of the mean is approximately 0.68 or 68%, within 2 standard deviations, about 0.95 or 95%, and within 3 standard deviations, about 0.997 or 99.7%.



Standard Normal Distribution

6-6

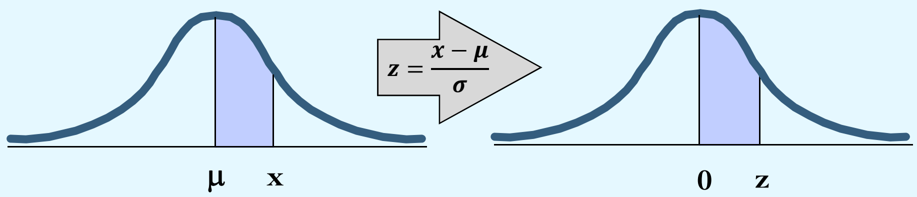
- Because the total area under the normal distribution is 1, there is a correspondence between area and probability
- Since each normal distribution is determined by its own mean and standard deviation, we would have to have a table of areas for each possibility !!!
- To simplify this situation, we use a common standard that requires only one table.
- The *standard normal distribution* is a normal distribution with a mean of 0 and a standard deviation of 1.

Standard Normal Distribution

6-7

- All normally distributed variables can be transformed into the standard normally distributed variable by using the z value which is the number of standard deviations that a particular x value is away from the mean

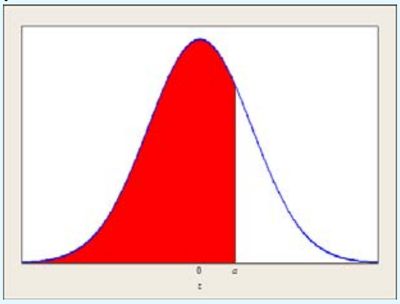
$$z = \frac{Value - Mean}{Standard\ deviation} = \frac{x-\mu}{\sigma}$$



Area Under the Standard Normal distribution Curve

6-8

- The table of the standard normal distribution gives the probability to the left of the values, thus $P(z < a)$.



- Note the following:
  - $P(z < a) = P(z \le a)$
  - $P(a < z < b) = P(a \le z \le b) = P(z < b) - P(z < a)$
  - $P(z > a) = P(z \ge a) = 1 - P(z < a)$

Area Under the Standard Normal distribution Curve

6-9 **Example**

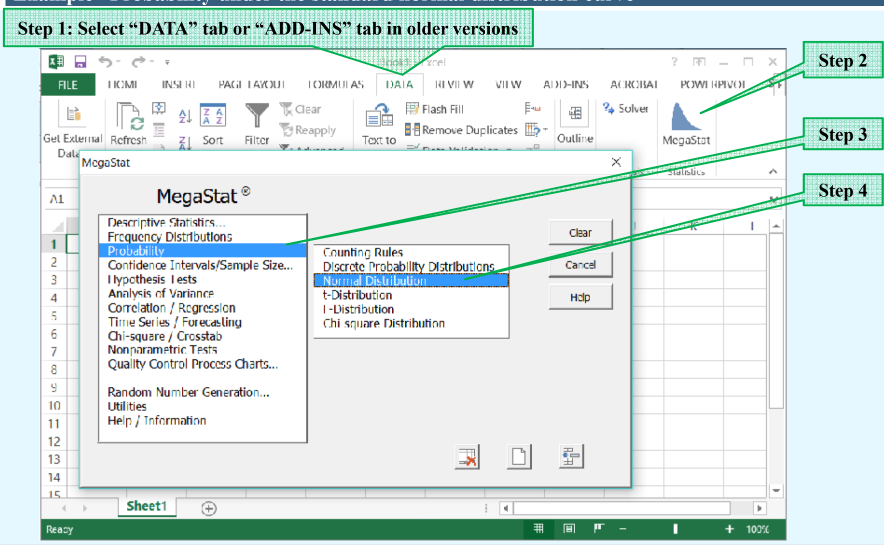
$P(z < 2.32) = 0.9898$

$P(z > 2.32) = 1 - P(z < 2.32) = 1 - 0.9898 = 0.0102$

z	.00	.01	.02	.03	…	.09
⋮						
2.0						
2.1						
2.3			0.9898			
2.4						
2.5						
⋮						

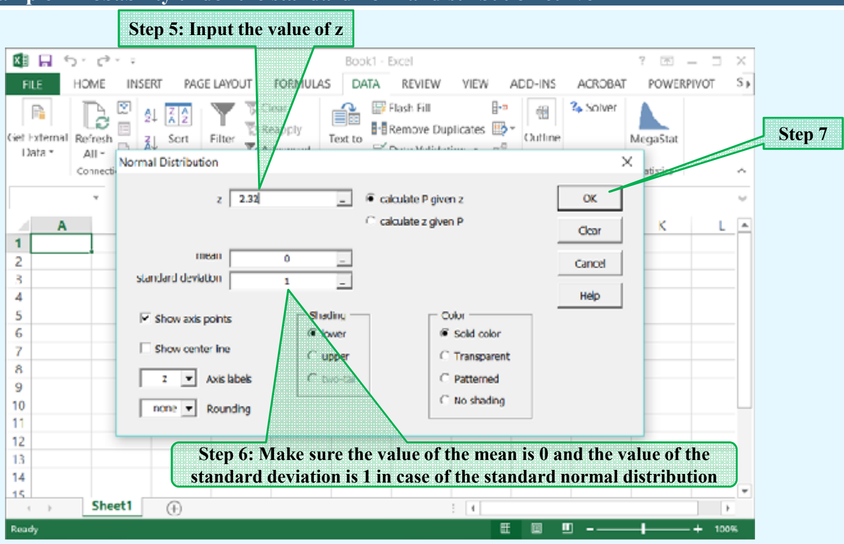
Application Using Excel

6-10 **Example “Probability under the standard normal distribution curve”**



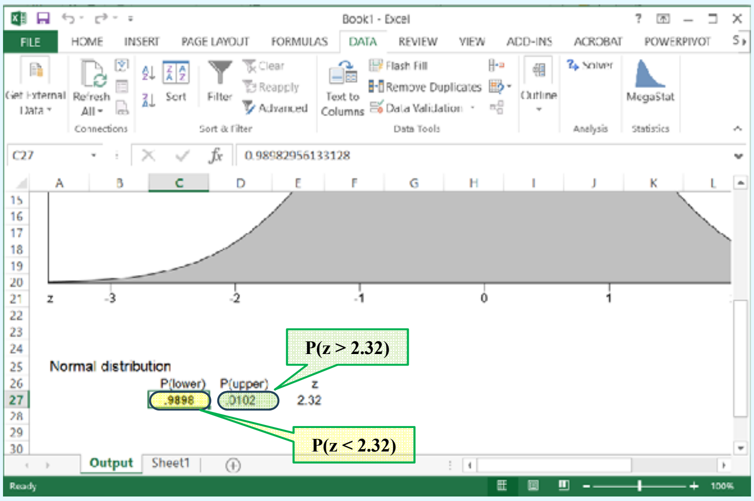
Application Using Excel

6-11 **Example “Probability under the standard normal distribution curve”**



Application Using Excel

6-12 Example "Probability under the standard normal distribution curve"



Area Under the Normal distribution Curve

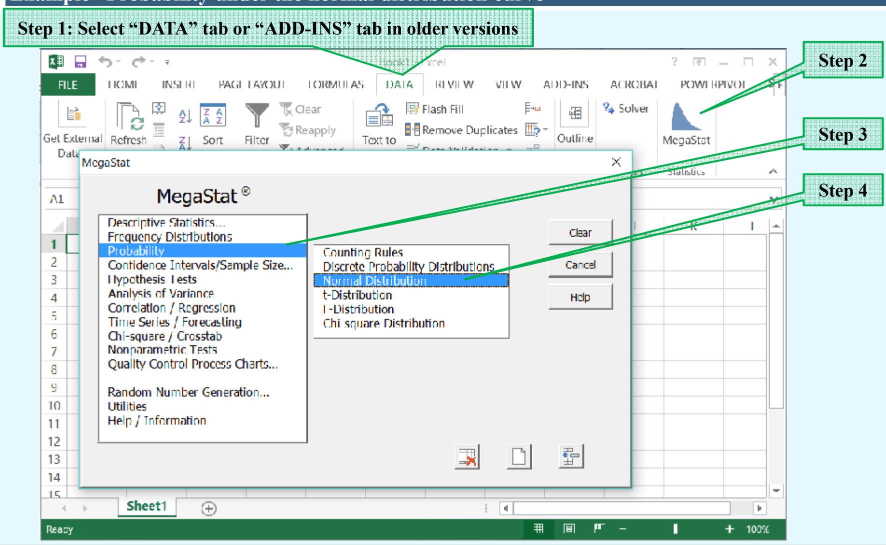
6-13 Example

If the income of 10000 family follows a normal distribution with mean 1800 SAR and standard deviation 300 SAR, find

- The probability of a family income is less than 2550.
$$P(x < 2550) = 0.9938$$
- The probability of a family income is greater than 1500.
$$P(x > 1500) = 0.8413$$
- The probability of a family income is between 1650 and 2100,
$$P(1650 < x < 2100) = P(x < 2100) - P(x < 1650) = 0.5328$$
- The number of families that have income greater than 1500,
$$P(x > 1500) * 10000 = 0.8413 * 10000 = 8413$$

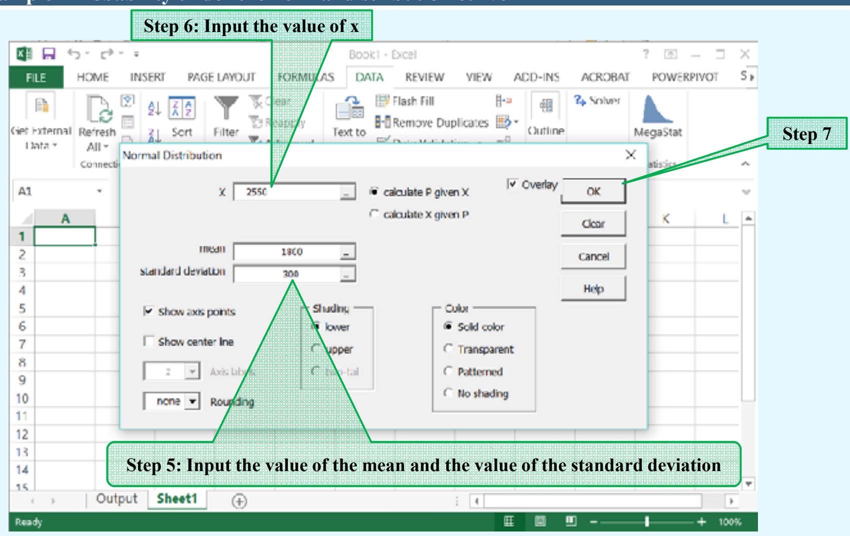
Application Using Excel

6-14 **Example “Probability under the normal distribution curve”**



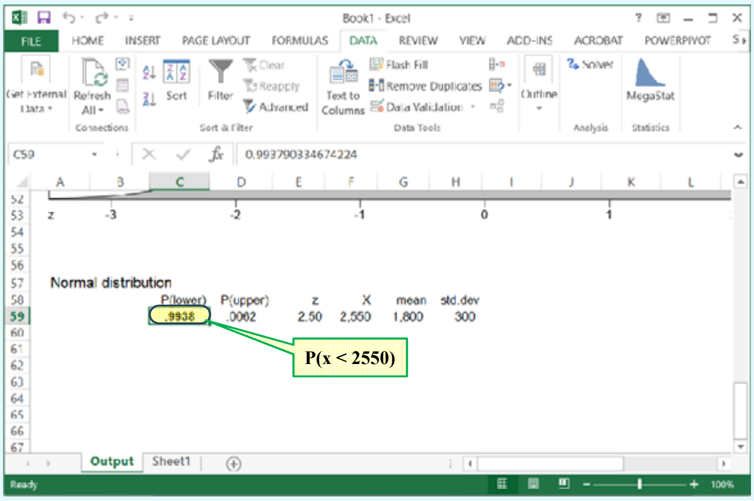
Application Using Excel

6-15 **Example “Probability under the normal distribution curve”**



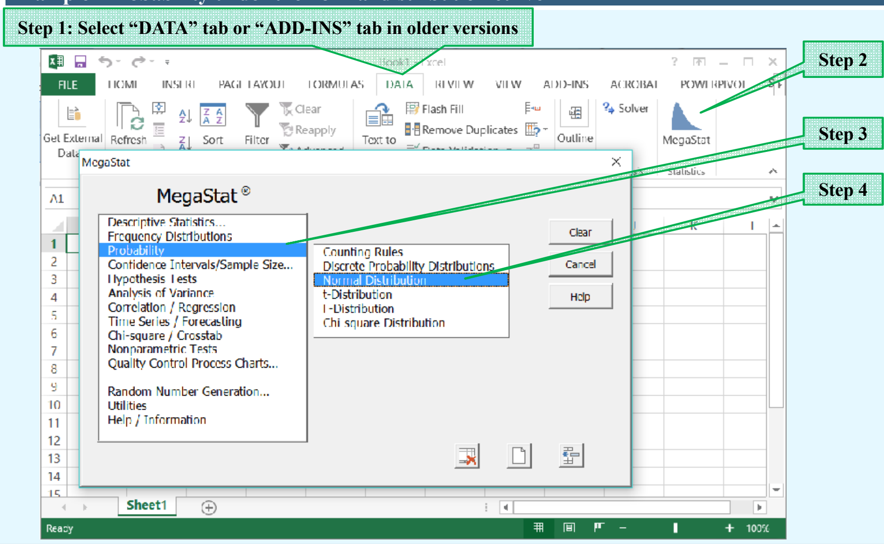
Application Using Excel

6-16 **Example “Probability under the normal distribution curve”**



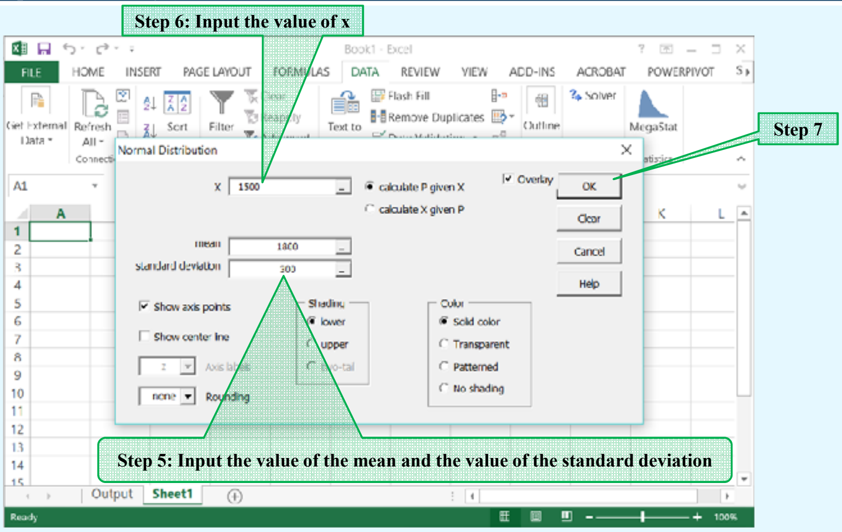
Application Using Excel

6-17 **Example “Probability under the normal distribution curve”**



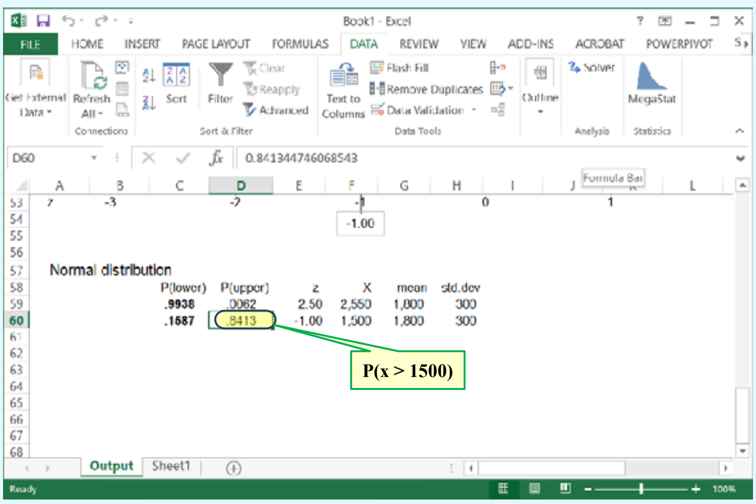
Application Using Excel

6-18 **Example "Probability under the normal distribution curve"**



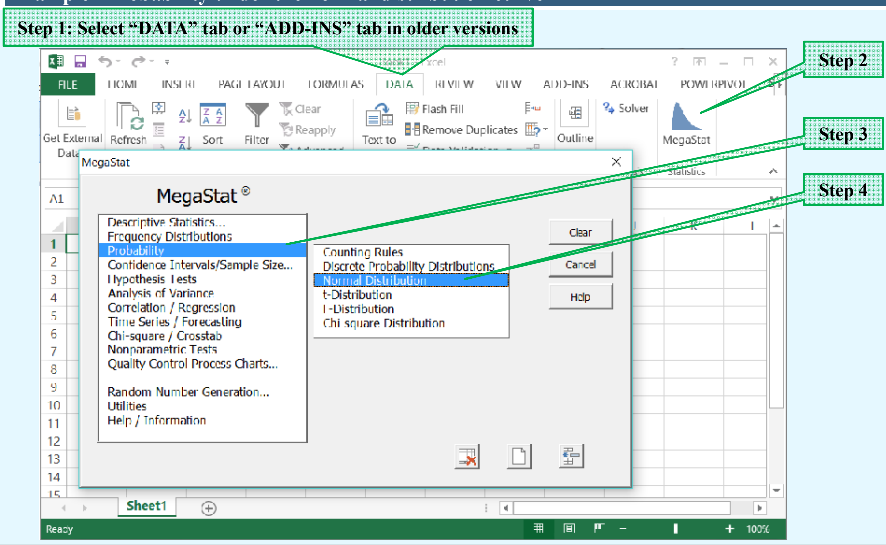
Application Using Excel

6-19 **Example "Probability under the normal distribution curve"**



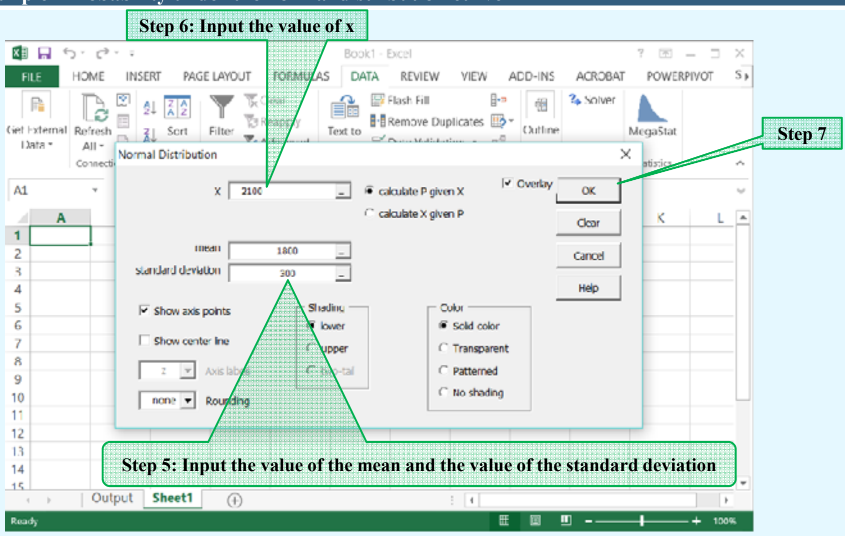
Application Using Excel

6-20 **Example "Probability under the normal distribution curve"**



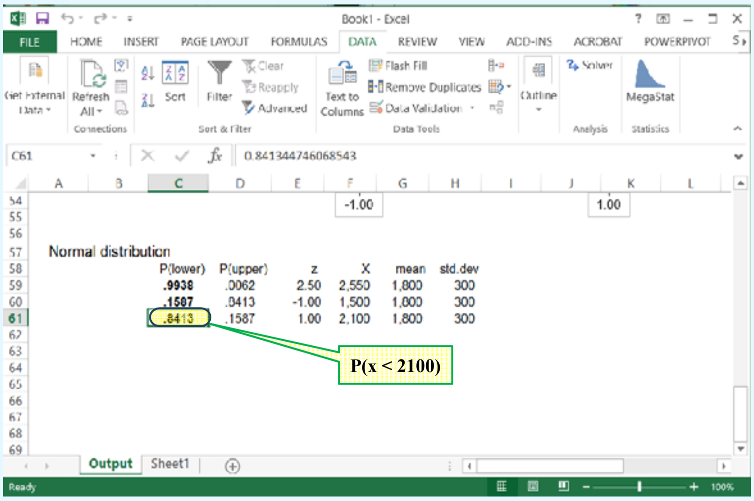
Application Using Excel

6-21 **Example "Probability under the normal distribution curve"**



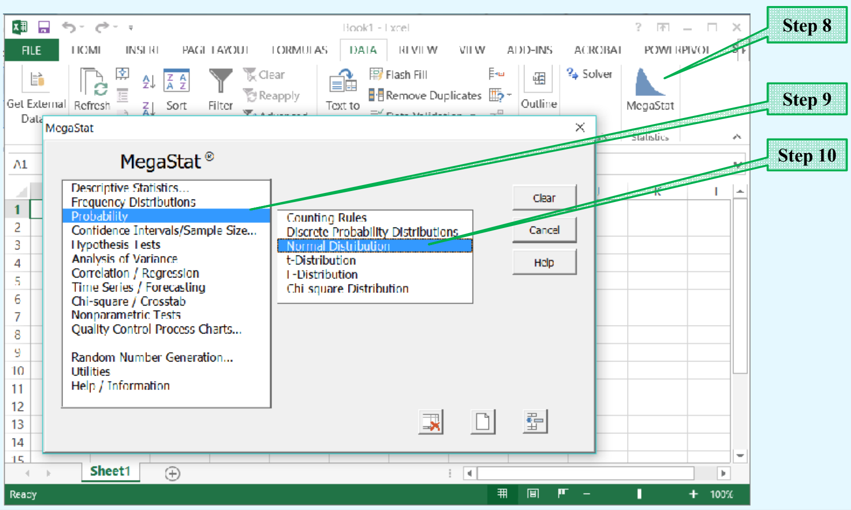
Application Using Excel

6-22 **Example "Probability under the normal distribution curve"**



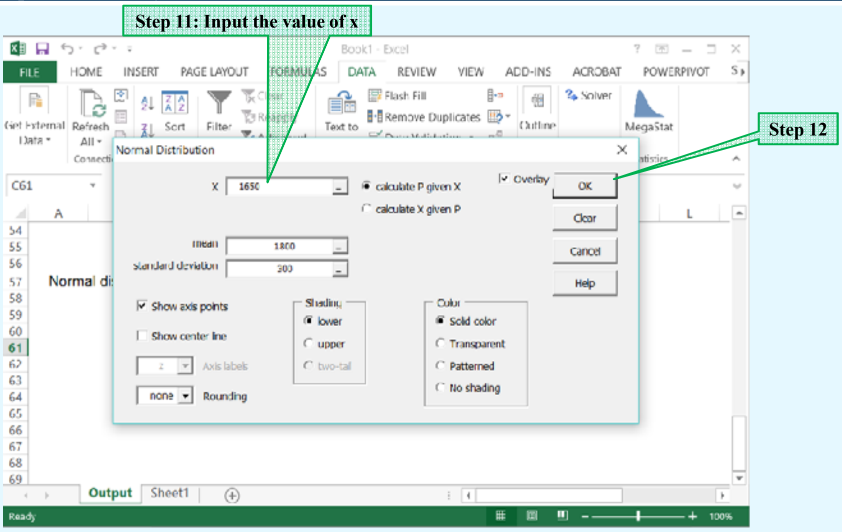
Application Using Excel

6-23 **Example "Probability under the normal distribution curve"**



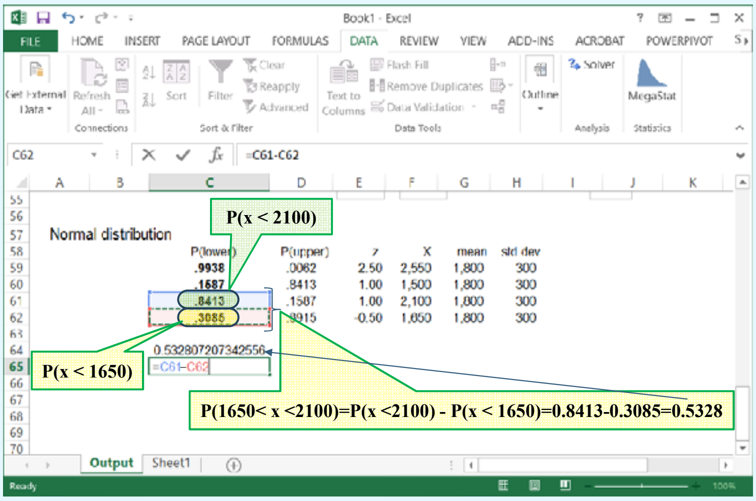
Application Using Excel

6-24 **Example "Probability under the normal distribution curve"**



Application Using Excel

6-25 **Example "Probability under the normal distribution curve"**



Area Under the Normal distribution Curve

6-26 **Example**

The lifetime of a type of microwaves follows a normal distribution with mean 3 years and standard deviation 1 year. If one microwave was chosen randomly,

- What is the probability that its lifetime will be greater than 2 years?

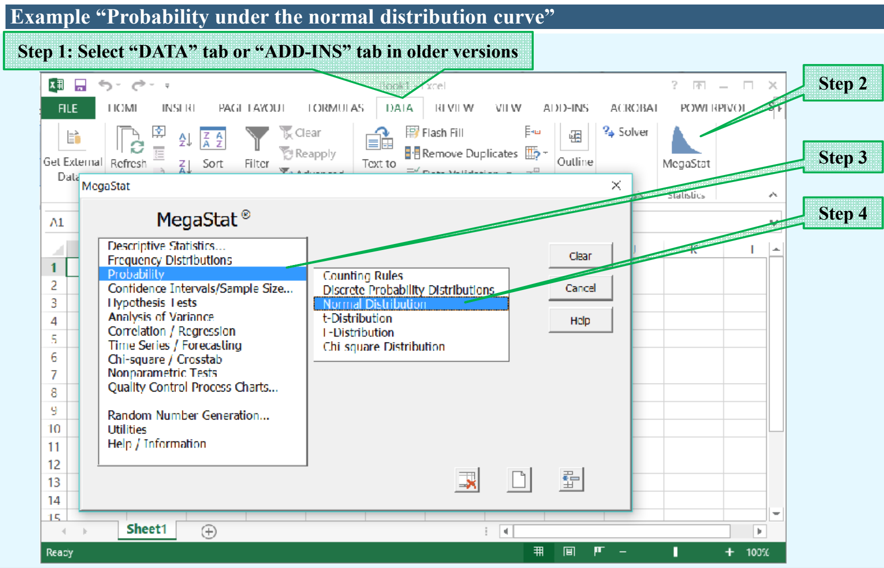
$$P(x > 2) = 0.8413$$

- If the microwaves have warranty for one year, what is the percentage of microwaves that the factory has to exchange with new ones.

$$P(x < 1) * 100 = 0.0228 * 100 = 2.28\%$$

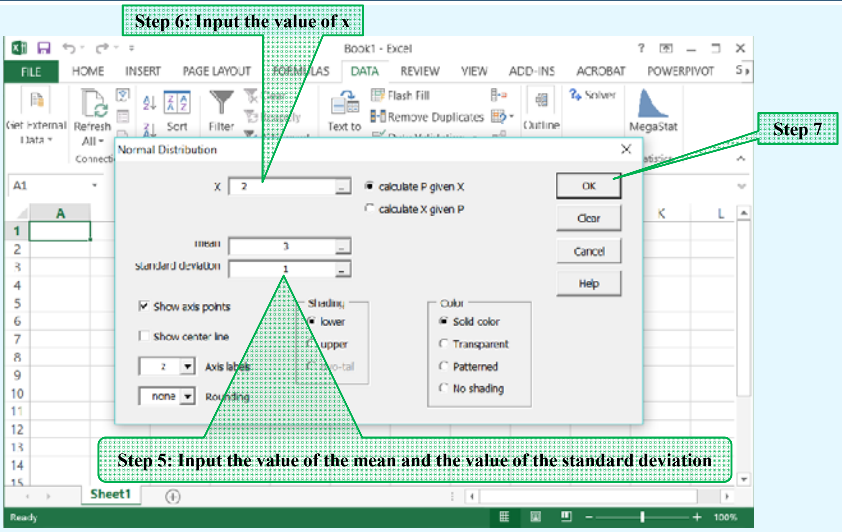
Application Using Excel

6-27 **Example "Probability under the normal distribution curve"**



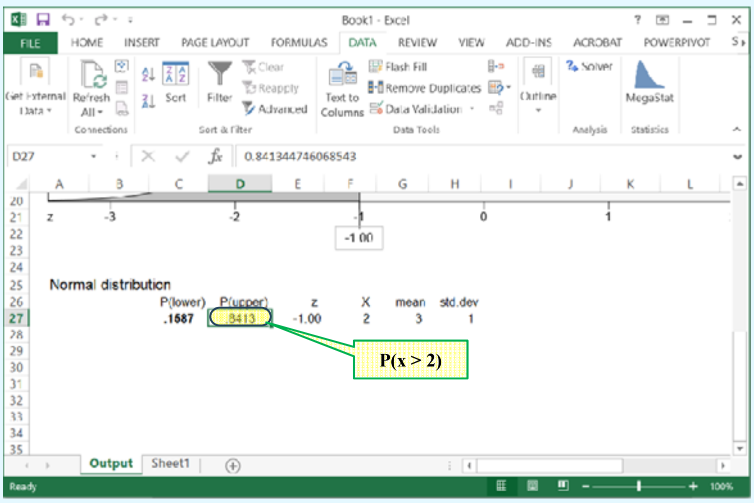
Application Using Excel

6-28 **Example "Probability under the normal distribution curve"**



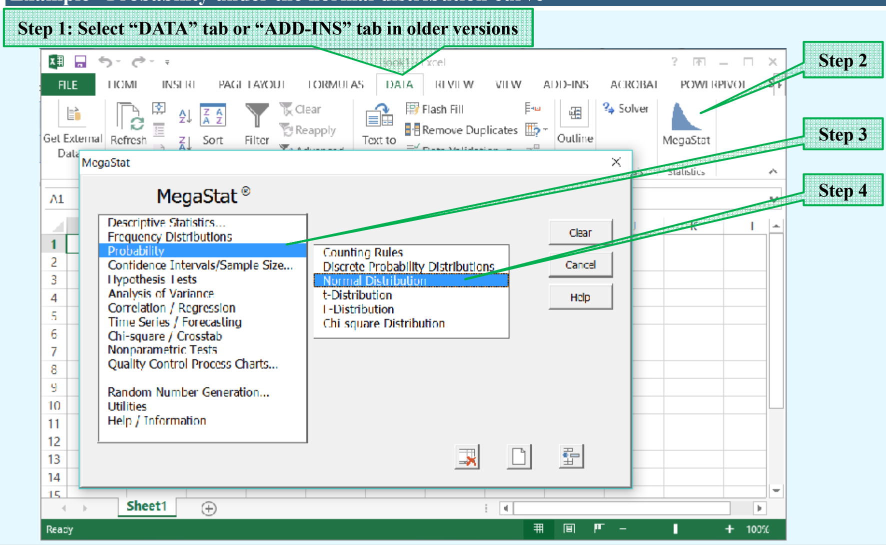
Application Using Excel

6-29 **Example "Probability under the normal distribution curve"**



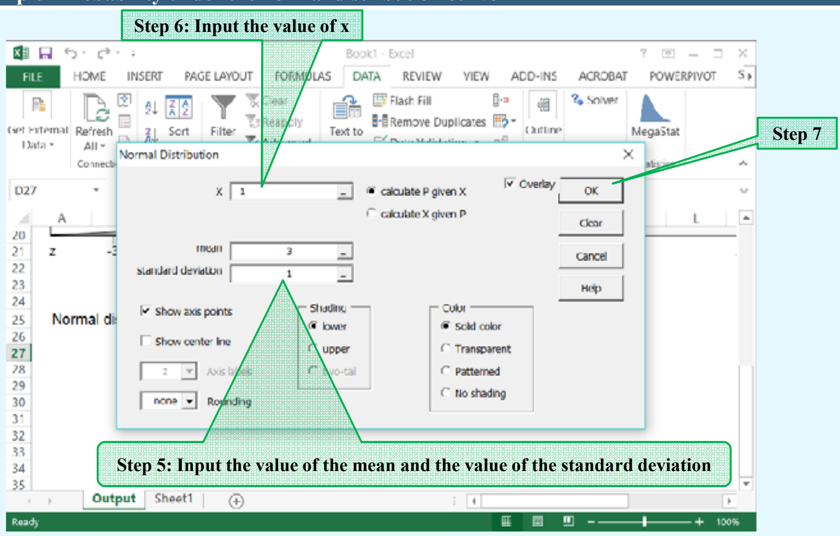
Application Using Excel

6-30 **Example “Probability under the normal distribution curve”**



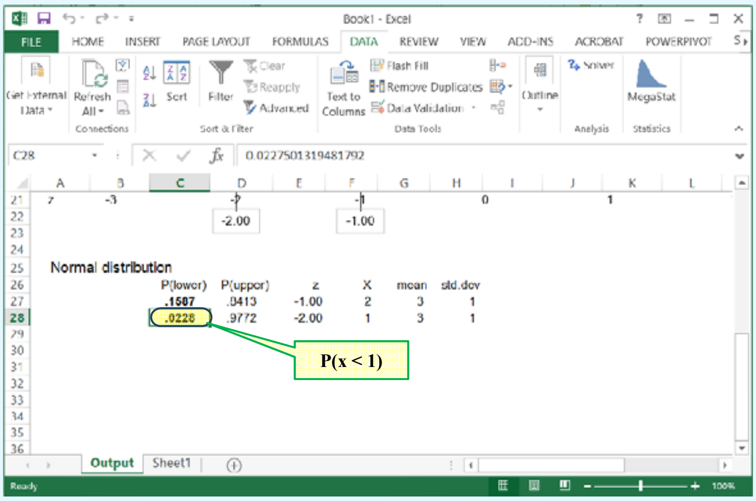
Application Using Excel

6-31 **Example “Probability under the normal distribution curve”**



Application Using Excel

6-32 **Example “Probability under the normal distribution curve”**



Distribution of Sample Means

6-33

- A *sampling distribution of sample means* is a distribution obtained by using the means computed from random samples of a specific size taken from a population.
- *Sampling error* is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

The Central Limit Theorem

6-34

- As the sample size n increases, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution.
- Thus, the mean of the sample means equals the population mean, $\mu_{\bar{x}} = \mu$, and the standard deviation of the sample means which is called the *standard error of the mean* is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- The central limit theorem can be used to answer questions about sample means in the same manner that the normal distribution can be used to answer questions about individual values.
- A new formula must be used for the z values:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Standard Error of The Mean

6-35 **Example**

Find the standard error of the mean for a sample of 49 that has been drawn from a population with standard deviation equals to 14.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{49}} = 2$$

The Central Limit Theorem

6-36 **Example**

A.C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of TV per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 32 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch TV is greater than 26.3 hours.

$$P(\bar{x} > 26.3) = 0.0071$$

Application Using Excel

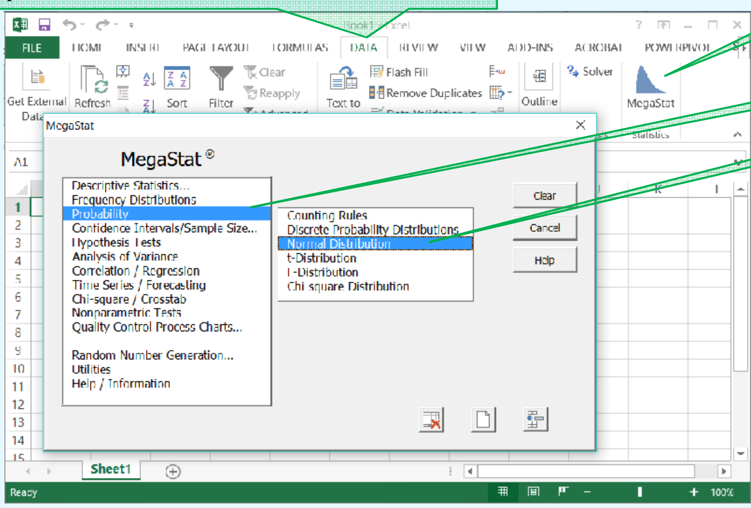
6-37 **Example "Central limit theorem"**

Step 1: Select "DATA" tab or "ADD-INS" tab in older versions

Step 2

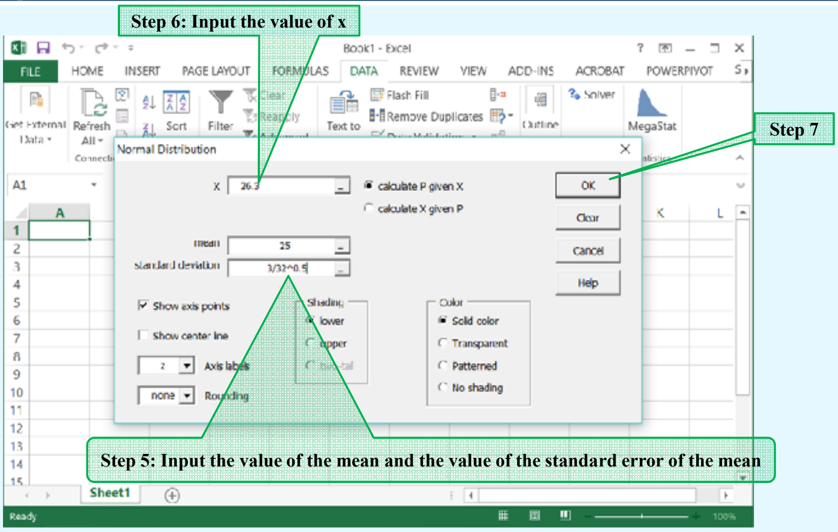
Step 3

Step 4



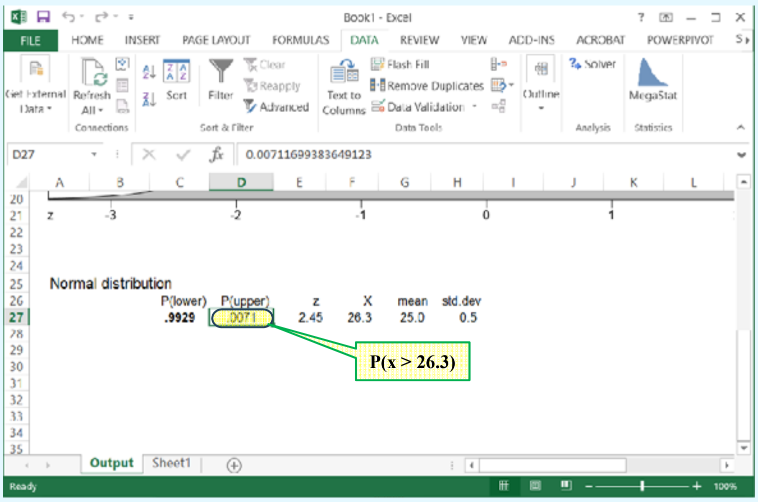
Application Using Excel

6-38 **Example “Central limit theorem”**



Application Using Excel

6-39 **Example “Central limit theorem”**



The Central Limit Theorem

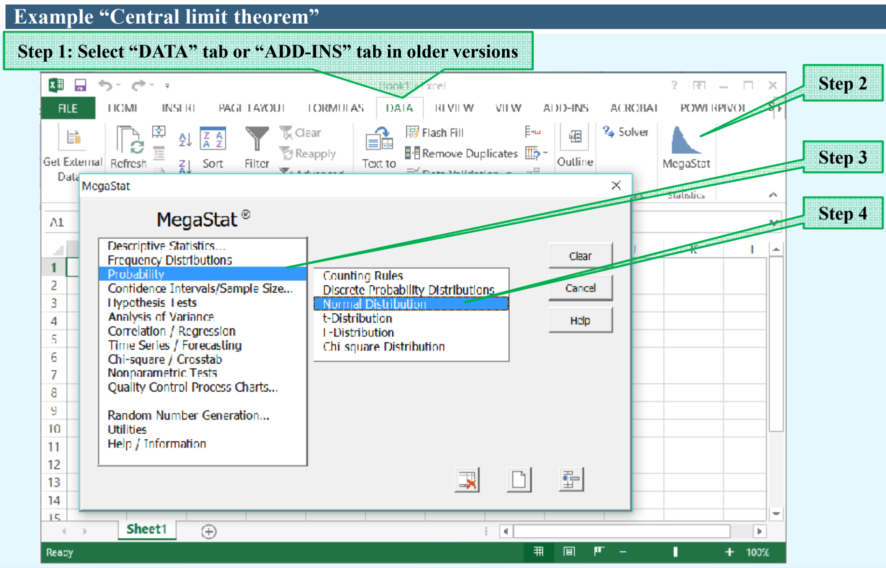
6-40 **Example**

The average age of a vehicle registered in the United States is 8 years. Assume the standard deviation is 16 months. If a random sample of 36 cars is selected, find the probability that the <u>mean</u> of their age is between 90 and 100 months.

$$P(90 < \bar{x} < 100) = P(\bar{x} < 100) - P(\bar{x} < 90) = 0.921$$

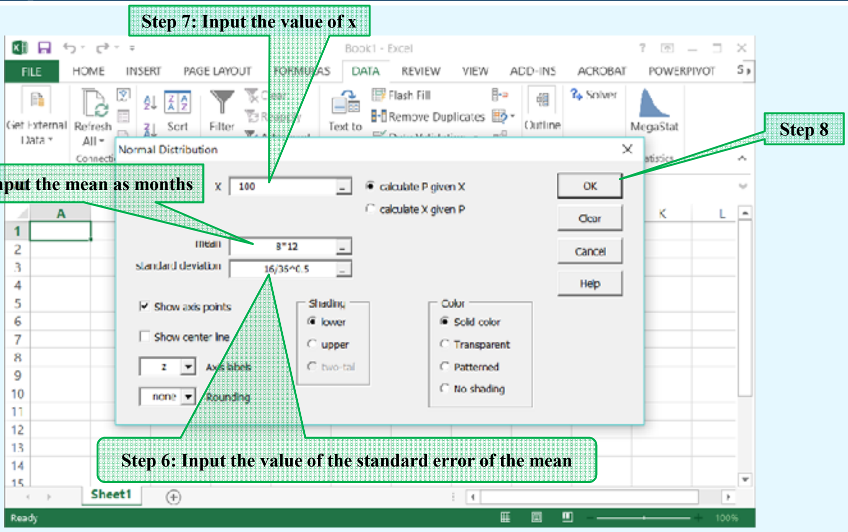
Application Using Excel

6-41 **Example "Central limit theorem"**



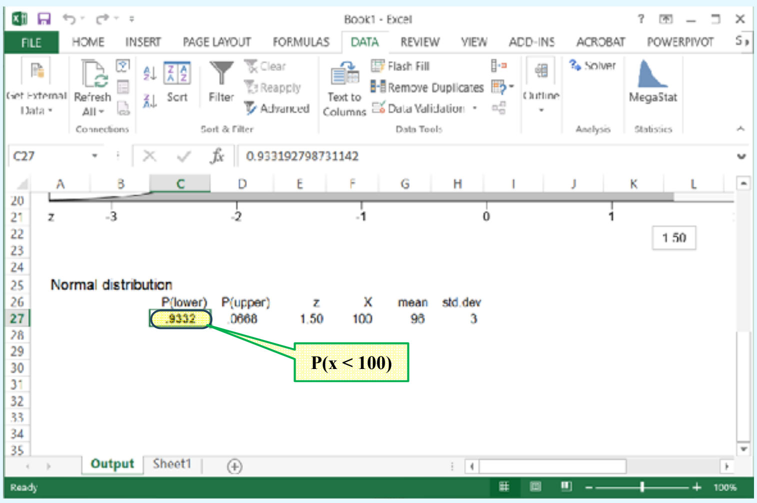
Application Using Excel

6-42 **Example "Central limit theorem"**



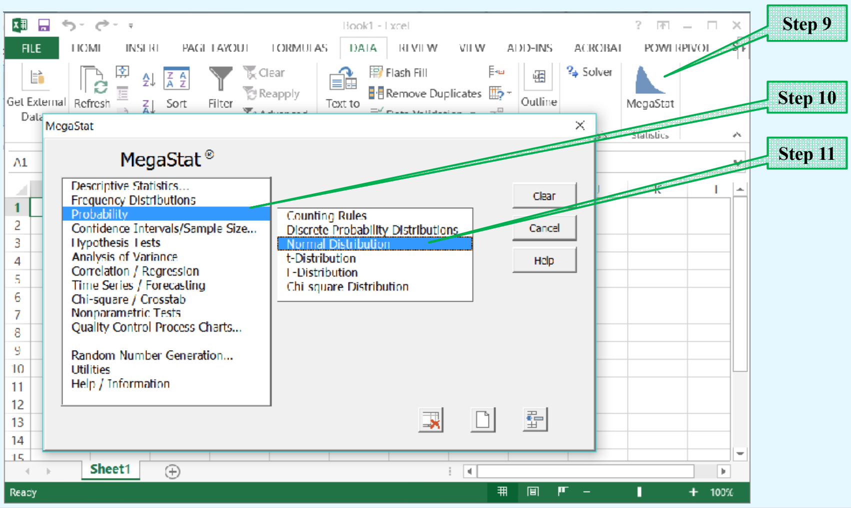
Application Using Excel

6-43 **Example "Central limit theorem"**



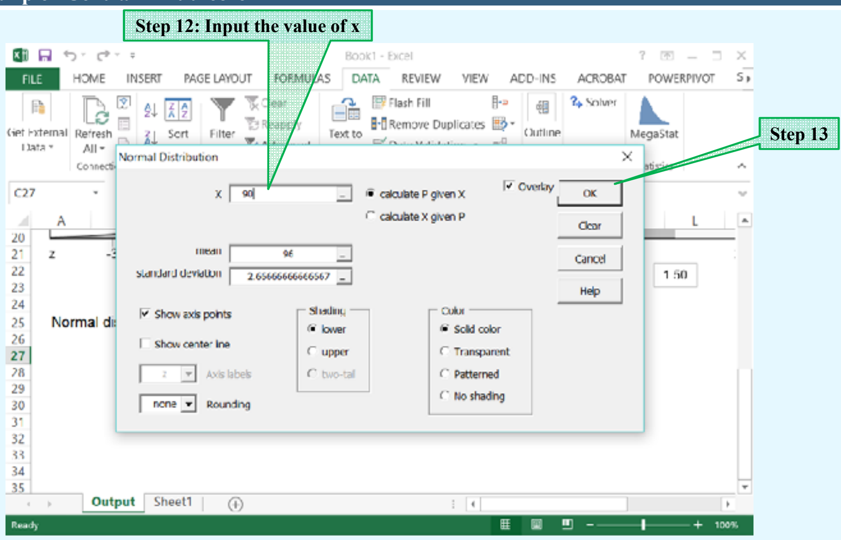
Application Using Excel

6-44 **Example “Central limit theorem”**



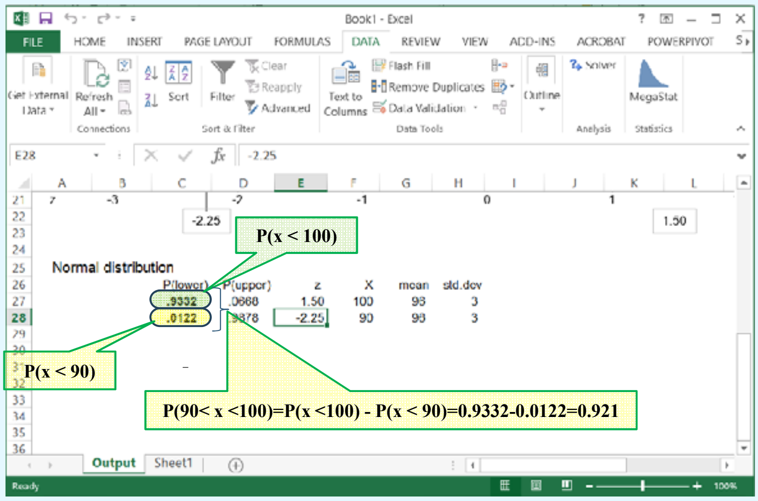
Application Using Excel

6-45 **Example “Central limit theorem”**



Application Using Excel

6-46 **Example "Central limit theorem"**



Summary

6-47

- The normal distribution can be used to describe a variety of variables, such as heights, weights, and temperatures.
- The normal distribution is bell-shaped, unimodal, symmetric, and continuous; its mean, median, and mode are equal.
- The normal distribution can be used to approximate other distributions.
- Mathematicians use the standard normal distribution which has a mean of 0 and a standard deviation of 1.
- The normal distribution can be used to describe a sampling distribution of sample means. These samples must be of the same size and randomly selected with replacement from the population.

Summary

6-48

- The central limit theorem states that as the size of the samples increases, the distribution of sample means will be approximately normal.
- The distribution of sample means is much less variable than the distribution of individual data value.

Dr. Saeed Alghamdi
Room 26F41
Building 90
Statistics Department
Faculty of Sciences
King Abdulaziz University
web sit
http://saalghamdy.kau.edu.sa

Elementary Statistics

A Step by Step Approach
Eighth Edition

by
Allan G. Bluman

SLIDES PREPARED
BY

Dr. Saeed A Dobbah Alghamdi
Statistics Department
King Abdulaziz University

Chapter 1

1-1

The Nature of Probability and Statistics

Objectives

1-2

- Demonstrate knowledge of statistical terms.
- Differentiate between the two branches of statistics.
- Identify types of data.
- Identify the measurement level for each variable.
- Identify the four basic sampling techniques.
- Explain the difference between an observational and an experimental study.

Introduction

- We are exposed to numerical information that describes some aspects of the world around us daily.
- For instance the daily news is filled with numerical information about phenomenon such as sports, health or social activities. For example, the winning number of a football team, the most reliable cars, the daily allowance of students, the census of Saudi Arabia, and so on.

Introduction

- Statistics is used in almost all fields of human endeavor, such as sports, public health and education.
- Furthermore, it is used for analysis and as a tool in scientific researches.
- Other use of it includes operation research, quality control, estimation and prediction.

For more information see page 2 of Bluman's Book

Basic Statistical Terminology

Before going any further in this subject, we need first to define the word STATISTICS.

- *Statistics* is the science of collecting, organizing, summarizing, analyzing and draw conclusions from data.
- EXAMPLE

  The 2010 Census of Saudi Arabia reveals the following:

  - Population is 27,136,977 persons.
  - 69% of the population are Saudi citizens.
  - Number of houses is 4,643,151.

For more information see page 3 of Bluman's Book

Basic Statistical Terminology

Now, in the stage of planning any statistical study, we need to determine the POPULATION of our study.

- *Population* consists of all elements for the phenomenon under a study.
- EXAMPLE

  The 2010 Census of Saudi Arabia reveals the following:

  - Population is 27,136,977 persons.
  - Number of houses is 4,643,151.

Basic Statistical Terminology

Sometimes it is not possible to collect the information from each element in the population because it is impractical or too expensive. Hence, we may select a representative portion from the entire population which is called SAMPLE.

- *Sample* is a group of elements selected randomly from a population.
- EXAMPLE

  A group of students is selected randomly from all students studying in King Abdulaziz University (KAU) for a research study.

Basic Statistical Terminology

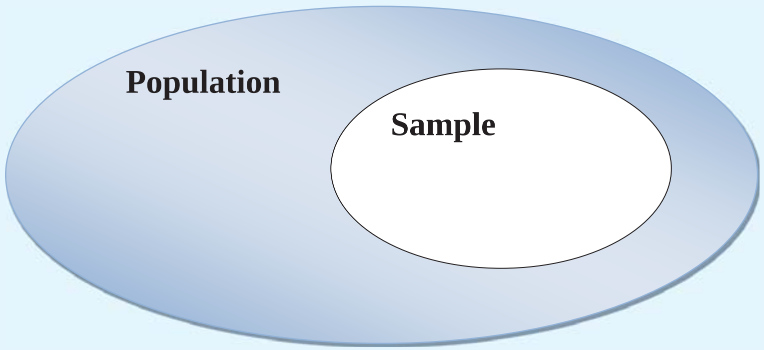


Figure 1.1: Population versus Sample

For more information see page 4 of Bluman's Book

Basic Statistical Terminology

In order to gain knowledge about any phenomenon, DATA are collected for VARIABLES that describe that phenomenon.

- ***Data*** are the values that variables can assume to describe a phenomenon.
- EXAMPLE

  The values that the variable weight can take when measuring the weights of students who are studying Statistics in KAU.

For more information see page 3 of Bluman's Book

Basic Statistical Terminology

In order to gain knowledge about any phenomenon, DATA are collected for VARIABLES that describe that phenomenon.

- ***Variable*** is a characteristic or an attribute of an item or individual that can assume different values.
- EXAMPLE

  Weight, height, gender, nationality, color, income, country, ...

For more information see page 3 of Bluman's Book

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Descriptive and Inferential Statistics

Depending on how data are used, the subject of **statistics** is divided into two main branches **descriptive statistics** and **inferential statistics**. However, in any statistical study, you should start with the descriptive statistics and then the inferential statistics which is usually the goal of any study.

- *Descriptive statistics* consists of the collection, organization, summarization, and presentation of data.
- *Inferential statistics* consists of generalizing from samples to populations by performing estimations and hypothesis testing, determining relationships among variables, and making predictions.

Descriptive and Inferential Statistics

- EXAMPLE (Descriptive)

  The 2010 Census of Saudi Arabia reveals the following:

  - Population is 27,136,977 persons.
  - 69% of the population are Saudi citizens.
  - Number of houses is 4,643,151.

Descriptive and Inferential Statistics

- EXAMPLE (Inferential)
  - The estimated average height of KAU students is 1.75m
  - There is a relationship between the performance of students in mathematics classes and statistics classes.
  - The population of Saudi Arabia is estimated to be around 37 million persons by the year 2020.

For more information see page 4 of Bluman's Book

Variables and Types of Data

- *Qualitative variables* can be placed into distinct categories according to some characteristic or attribute, e.g., flight classes, departments, gender, color, ….
- *Quantitative or scale variables* are numerical in nature and can be ordered or ranked, e.g., number of passengers, weight, age, ….

For more information see page 6 of Bluman's Book

Variables and Types of Data (cont'd.)

Quantitative or scale variables can also be classified into two groups:

- *Discrete variables* assume values that can be counted, e.g., number of rooms in buildings, number of students in Stat 110, number of children in families, ….
- *Continuous variables* can assume all values between any two specific values, e.g., heights, weights, temperature,….

For more information see page 6 of Bluman's Book

Variables and Types of Data (cont'd.)

Qualitative variables can be classified into two groups:

- *Nominal*—classifies data into categories that can not be ordered or ranked, e.g., gender, color, departments, ….
- *Ordinal*—classifies data into categories that can be ranked, e.g., flight classes, ranking, grade letters, ….

For more information see pages 7 and 8 of Bluman's Book

Data Collection

Data can be collected in a variety of ways. One of the most common ways is the use of *surveys* that can done by using a variety of methods. Three of the most common methods are:

- Telephone surveys
- Mailed questionnaire surveys
- Personal interview surveys

For more information see pages 9 and 10 of Bluman's Book

Sampling Techniques

- *Random samples* are selected using chance methods or random methods.
- EXAMPLE

A group of 10 students is selected using random numbers from 30 students to check the performance of a class.

1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

→

3 13 21 8
6 25 30 7
11 17 20

For more information see pages 10 and 11 of Bluman's Book

Sampling Techniques

1-19

- *Systematic samples* are selected by using every k^{th} number after the first subject is randomly selected from 1 through k .
- EXAMPLE

Suppose we have 30 mobiles and a sample of 6 is needed. Since 30/6=5, then k=5, thus every 5th mobile would be selected. However, the first mobile numbered between 1 and 5 would be selected at random. Suppose mobile 3 were the first mobile selected; then the sample would be

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

→

3	8
13	18
23	28

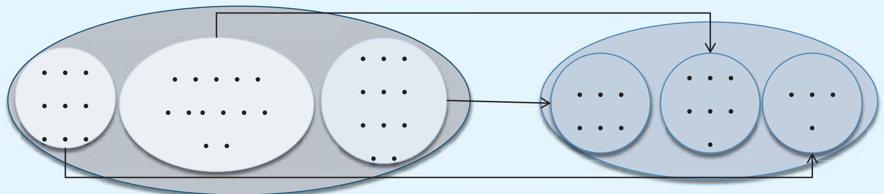
For more information see pages 11 and 12 of Bluman's Book

Sampling Techniques

1-20

- *Stratified samples* are selected by dividing the population into groups (*strata*), then subjects are randomly selected from each group or strata.
- EXAMPLE

Suppose we want to get a sample form a high school but we need to be sure that the sample contains subjects from each level in the school.

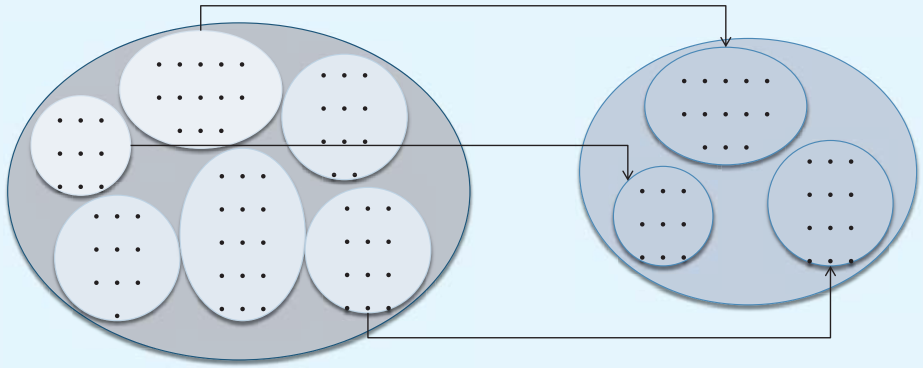


For more information see pages 12 of Bluman's Book

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Sampling Techniques

- *Cluster samples* are selected by using intact groups called clusters. Thus, dividing the population into groups and then **taking samples of the groups**.



For more information see pages 12 of Bluman's Book
Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Observational and Experimental Studies

- In an *observational study*, the researcher observes what is happening or what has happened and tries to draw conclusions based on these observations.

- In an *experimental study*, the researcher manipulates one of the variables and tries to determine how that influences other variables.

For more information see pages 13, 14 of Bluman's Book
Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Observational and Experimental Studies

- EXAMPLE (Observational Study)
  "Motorcycle owners are getting older and richer."
  Data were collected on the ages and incomes of motorcycle owners for the years 1980 and 1998 and then compared. The findings showed considerable differences in the ages and incomes of motorcycle owners for the two years.
  In this study, the researcher observed what had happened to the motorcycle owners over a period of time.

Observational and Experimental Studies

- EXAMPLE (Experimental Study)
  Students are divided into two groups and had the students perform as many sit-ups as possible in 90 sec. The first group was told only to "Do your best," while the second group was told to try to increase the actual number of sit-ups done each day by 10%. After four days, the subjects in the group who were given the vague instructions to "Do your best" averaged 43 sit-ups, while the group that was given the more specific instructions to increase the number of sit-ups by 10% averaged 56 sit-ups by the last day's session. The conclusion then was that athletes who were given specific goals performed better than those who were not given specific goals.

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Observational and Experimental Studies

Statistical studies usually include one or more independent variables and one dependent variable.

- The ***independent variable*** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the ***explanatory variable***
- The ***dependent variable or outcome variable*** is the resultant variable. Thus, The outcome variable is the variable that is studied to see if it has changed significantly due to the manipulation of the independent variable.

For more information see pages 14,15 of Bluman's Book

Observational and Experimental Studies

- EXAMPLE

  In the sit-up study, the researchers gave the groups two different types of instructions, general and specific. Hence, the independent variable is the type of instruction. The dependent variable, then, is the resultant variable, that is, the number of sit-ups each group was able to perform after four days of exercise.

Statistical Packages

- Excel, SPSS, MINITAB, SAS and the TI-83 graphing calculator can be used to perform statistical computations.
- Students should realize that the computer and calculator merely give numerical answers and save time and effort of doing calculations by hand.

Summary

- The two major areas of statistics are *descriptive* and *inferential*. Inferential statistics is based on probability theory.
- Data can be classified as *qualitative* or *quantitative*
- Quantitative data can be *discrete* or *continuous* depending on the values they can assume.
- Qualitative data can be *nominal* or *ordinal* depending on the category they can assume.
- When the *populations* to be studied are large, statisticians use subgroups called *samples*.

Summary

- The four basic methods for obtaining samples are *random*, *systematic*, *stratified*, and *cluster*.
- The two basic types of statistical studies are *observational* and *experimental*.
- When conducting an experimental study, researchers manipulate one or more of the independent or explanatory variables and see how this manipulation influences the dependent or outcome variable.

Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Main Reference

Elementary Statistics
A Step by Step Approach
By
Allan Bluman

Parts of Chapter 10 and 13

Correlation and Regression

Objectives

10-1

- Draw a scatter plot for a set of ordered pairs.
- Compute the correlation coefficient.
- Compute the equation of the regression line.

Introduction

10-2

- Inferential statistics involves determining whether a relationship between two or more numerical variables exists.
- *Correlation* is a statistical method used to determine whether a relationship between variables exists.
- *Regression* is a statistical method used to describe the nature of the relationship between variables.
- When you have two or more variables in your data, you may need to know
  1. Are two or more variables related?
  2. If so, what is the strength of the relationship?
  3. What type or relationship exists?
  4. What kind of predictions can be made from the relationship?

Introduction

10-3

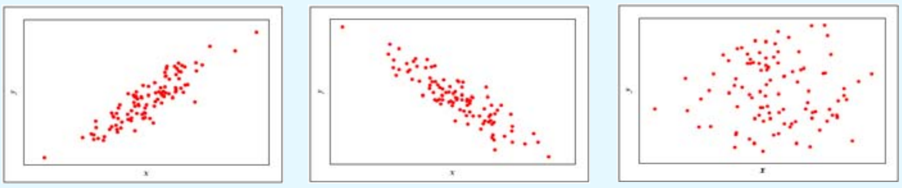
- In a *simple relationship*, there are only two types of variables under study;
  1. An *independent variable* (explanatory variable or predictor variable) is the variable that is being manipulated by the researcher and used to predict the dependent variable.
  2. A *dependent variable* (outcome variable or response variable) is the resultant variable.
- EXAMPLE

  A manager may wish to see whether the number of years the salespeople have been working for the company has anything to do with the amount of sales they make.

Scatter Plots

10-4

- The simple relationship can be positive or negative:
  - A *positive relationship* exists when both variables increase or decrease at the same time.
  - A *negative relationship* exists when one variable increases and the other variable decreases.
- A *scatter plot* is a visual way to describe the nature of the relationship between the independent and dependent variables.



Positive Linear Relationship **Negative Linear Relationship** **No Linear Relationship**

Correlation Coefficient

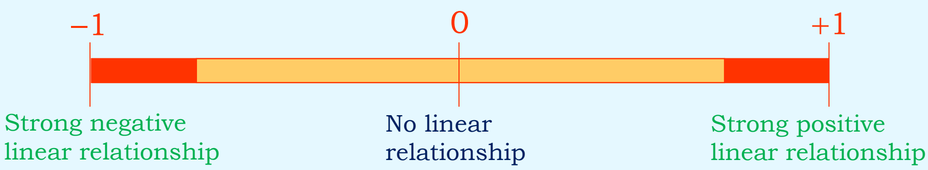
10-5

- The *correlation coefficient* is a measure of how variables are related, it measures the strength and direction of a linear relationship between two variables.
- The symbol for the population correlation coefficient is ρ (rho).
- The symbol for the sample correlation coefficient is r.
- The range of the correlation coefficient is from -1 to $+1$.

Correlation Coefficient

10-6

- If there is a *strong positive linear relationship* between the variables, the value of r will be close to +1.
- If there is a *strong negative linear relationship* between the variables, the value of r will be close to −1.
- When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.



Correlation Coefficient

10-7

Correlation Coefficient Value	Meaning
+1	Complete Positive Linear Relationship
0.70 — 0.99	Strong Positive Linear Relationship
0.50 — 0.69	Moderate Positive Linear Relationship
0.01 — 0.49	Weak Positive Linear Relationship
0	No Linear Relationship
-0.01 — -0.49	Weak Negative Linear Relationship
-0.50 — -0.69	Moderate Negative Linear Relationship
-0.70 — -0.99	Strong Negative Linear Relationship
-1	Complete Negative Linear Relationship

Correlation Coefficient

10-8

- *Pearson linear correlation coefficient*

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

- where n is the number of data pairs(sample size).

Application Using Excel

10-9 **Example**

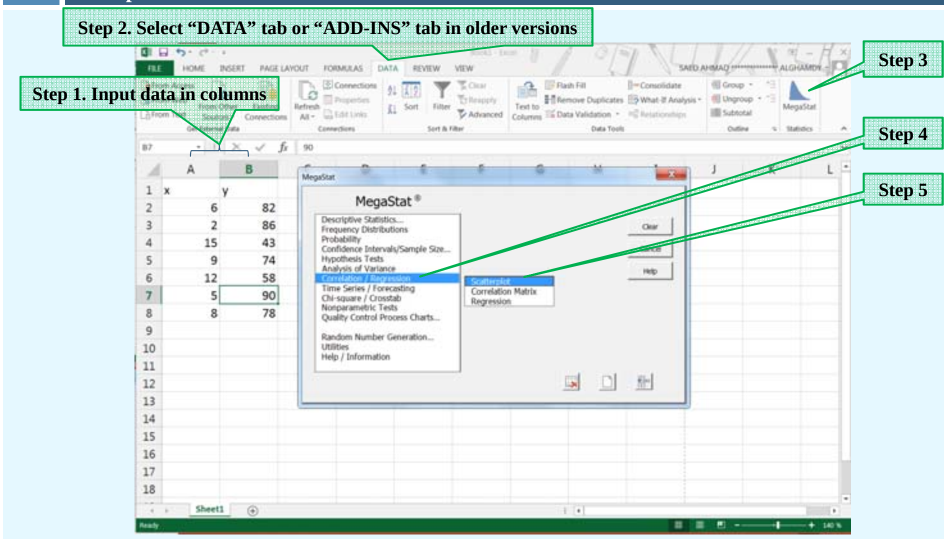
The following data are the number of absences and the final grades of seven randomly selected students from a statistics class.

- Draw the scatter plot for the variables.
- Compute the value of the Pearson correlation coefficient.

Number of absences ***x***	6	2	15	9	12	5	8
Final grades ***y***	82	86	43	74	58	90	78

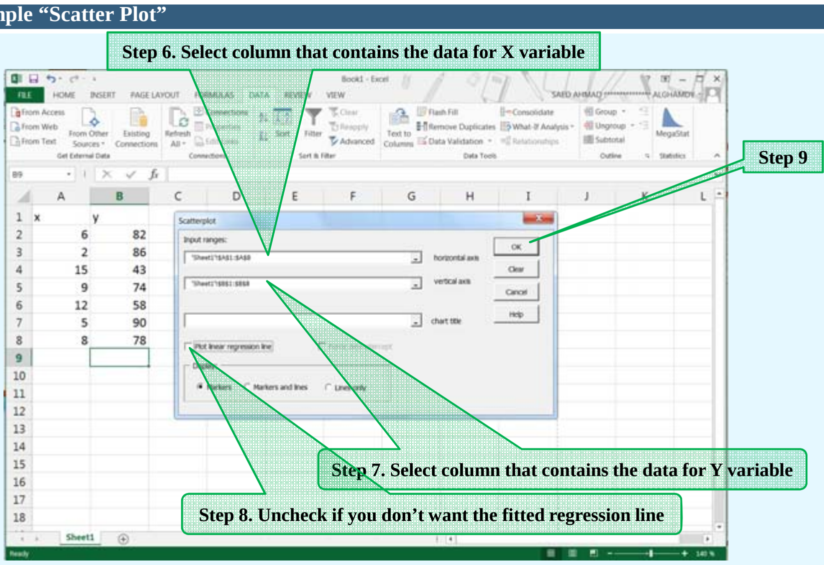
Application Using Excel

10-10 Example “Scatter Plot”



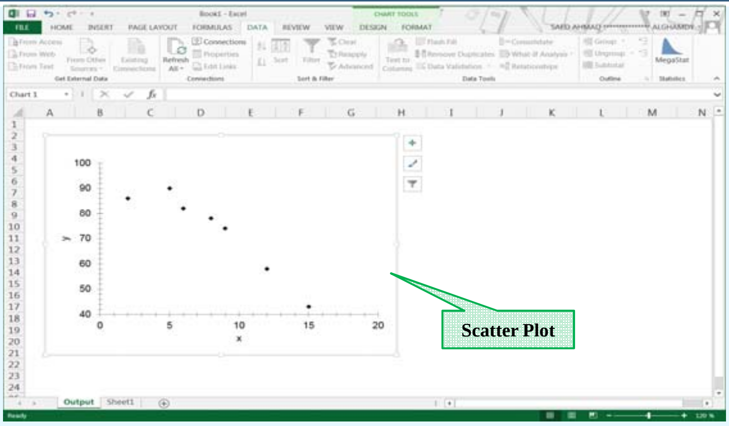
Application Using Excel

10-11 Example “Scatter Plot”



Application Using Excel

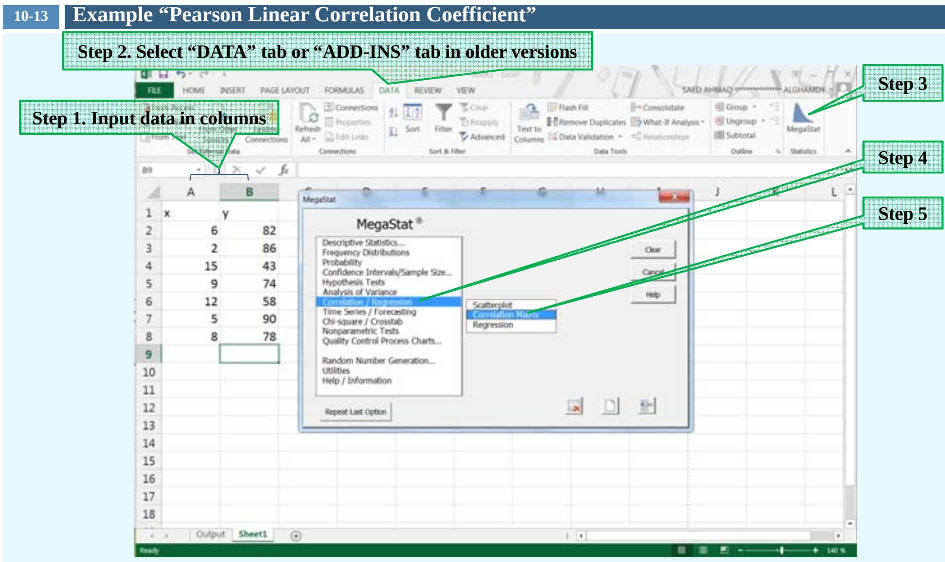
10-12 Example “Scatter Plot”



The plot indicates that there is a negative linear relationship. Thus, as the absences increased the final grades decreased on average.

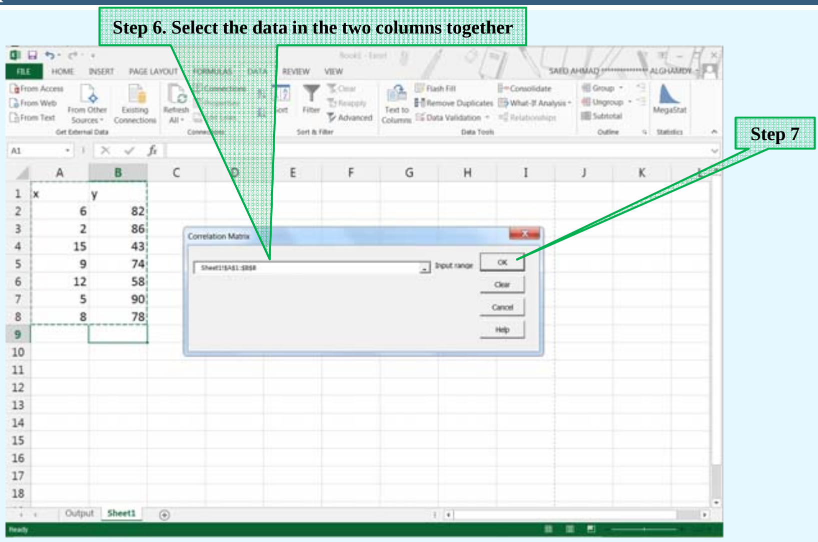
Application Using Excel

10-13 Example “Pearson Linear Correlation Coefficient”



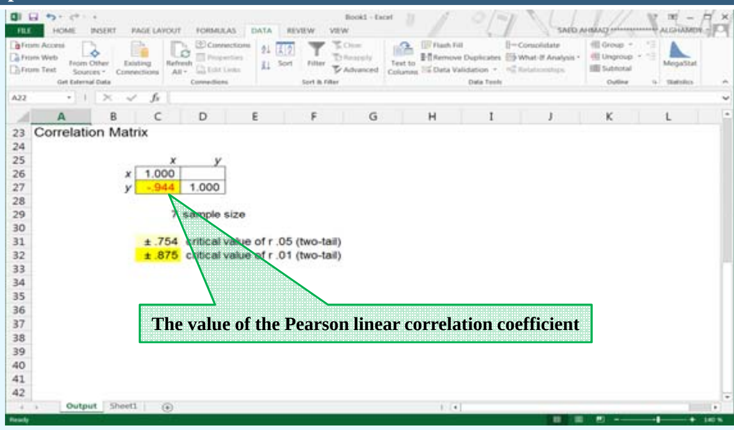
Application Using Excel

10-14 **Example "Pearson Linear Correlation Coefficient"**



Application Using Excel

10-15 **Example "Pearson Linear Correlation Coefficient"**



The value indicates that there is a strong negative linear relationship. Thus, the more absence a student has the lower is his final grade on average.

Correlation Coefficient

13-16

- *Spearman rank correlation coefficient*.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where d = difference in the ranks and

n = number of data pairs

Application Using Excel

13-17 **Example**

The table shows the total number of tornadoes that occurred in states from 1962 to 1991 and the record high temperatures for the same states.

Use the Spearman rank correlation coefficient to determine the relationship between the number of tornadoes and the record high temperatures.

State	Tornadoes	Record High Temp
AL	668	112
CO	781	118
FL	1590	109
IL	798	117
KS	1198	121
NY	169	108
PA	310	111
TN	360	113
VT	21	105
WI	625	114

Application Using Excel

13-18 **Example “Spearman rank Correlation Coefficient”**

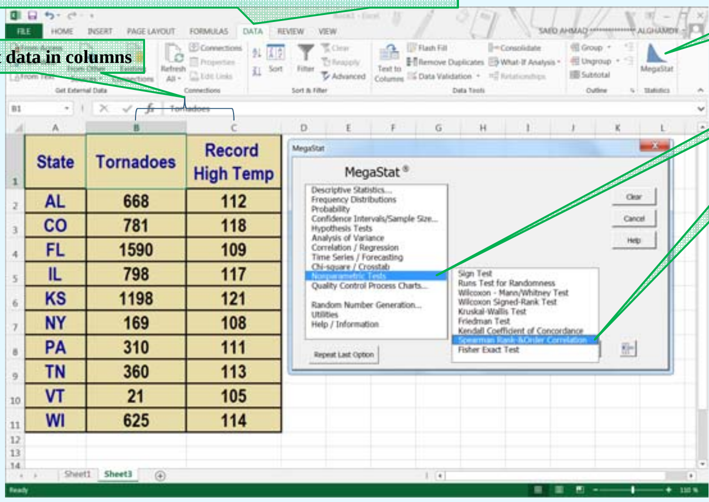
Step 2. Select “DATA” tab or “ADD-INS” tab in older versions

Step 1. Input data in columns

Step 3

Step 4

Step 5

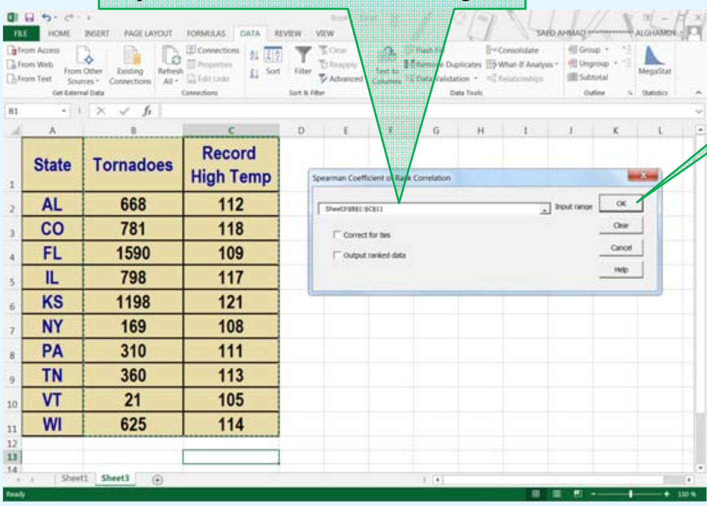


Application Using Excel

13-19 **Example “Spearman rank Correlation Coefficient”**

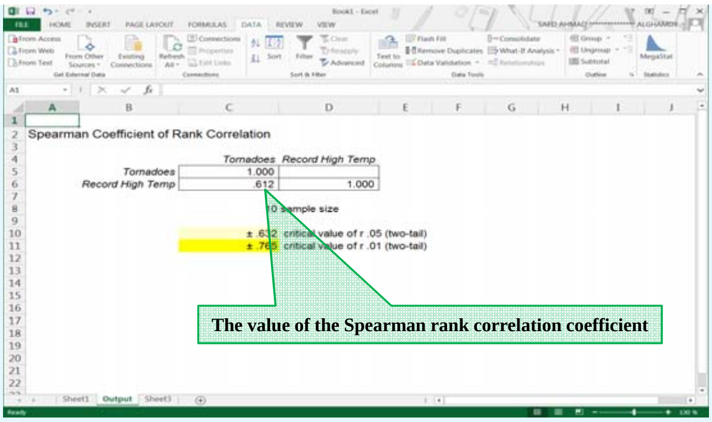
Step 6. Select the data in the two columns together

Step 7



Application Using Excel

13-20 **Example “Spearman rank Correlation Coefficient”**

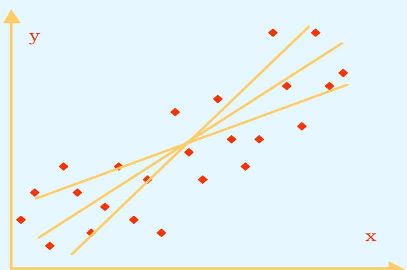


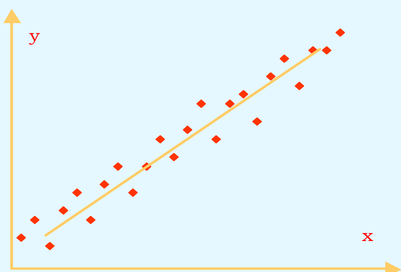
The value indicates that there is a moderate positive linear relationship between the number of tornados and the record high temperatures. Thus, high temperatures means more tornados on average.

Regression Line

10-21

- If the value of the correlation coefficient is significant (will not be discussed here), the next step is to determine the equation of the *regression line* which is the data’s line of best fit.
- *Best fit* means that the sum of the squares of the vertical distance from each point to the line is at a minimum.





For more information see pages 551-552 of Bluman's Book

Equation of the Regression Line

10-22

- The equation of the regression line is written as $y' = a + bx$
  where b is the slope of the line and a is the y' intercept.
- The regression line can be used to predict a value for the dependent variable (y) for a given value of the independent variable (x).
- Caution: Use x values within the experimental region when predicting y values.

Equation of the Regression Line

10-23

- Formulas for the regression line $y' = a + bx$

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

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

where a is the y' intercept and b is the slope of the line.

Application Using Excel

10-24 **Example**

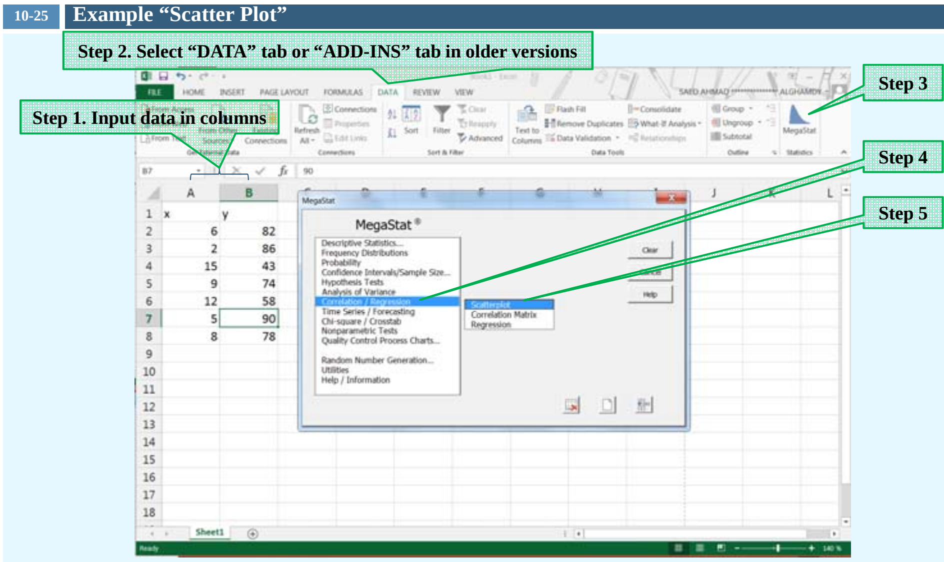
The following data are the number of absences and the final grades of seven randomly selected students from a statistics class.

- ➢ Determine the equation of the regression line. Remember that no regression should be done when r is not significant.
- ➢ Find the expected grade for a student who has been absent for 10 lectures.

Number of absences x	6	2	15	9	12	5	8
Final grades y	82	86	43	74	58	90	78

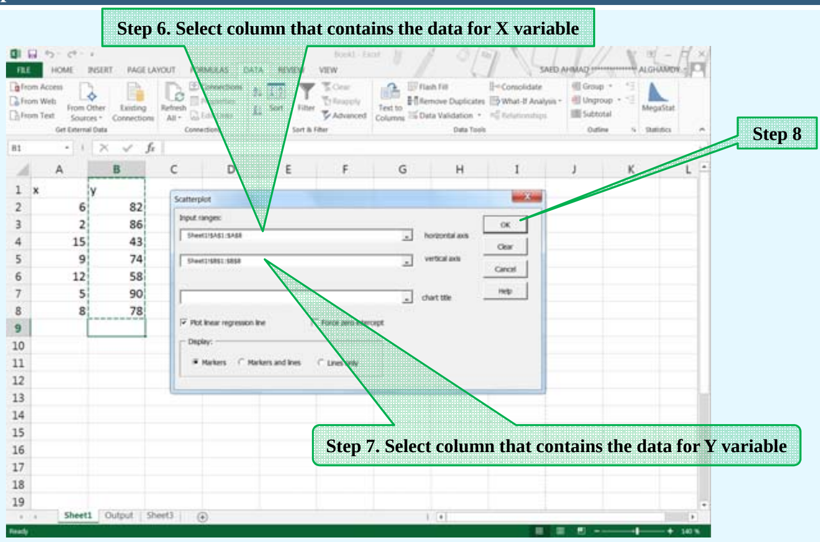
Application Using Excel

10-25 **Example "Scatter Plot"**



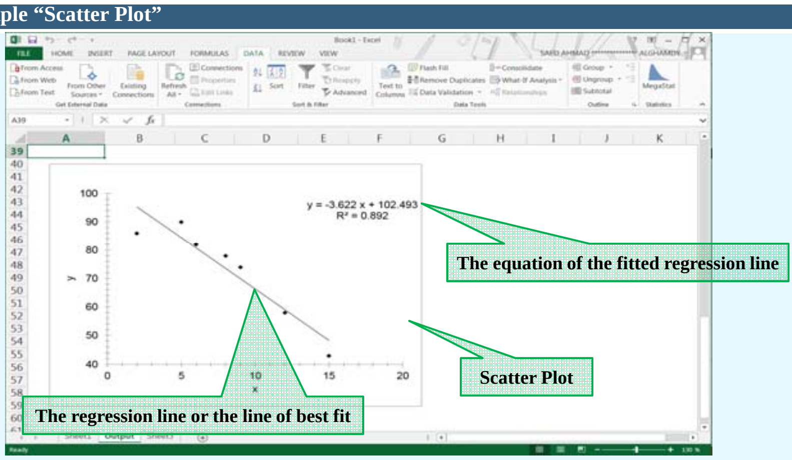
Application Using Excel

10-26 **Example "Scatter Plot"**



Application Using Excel

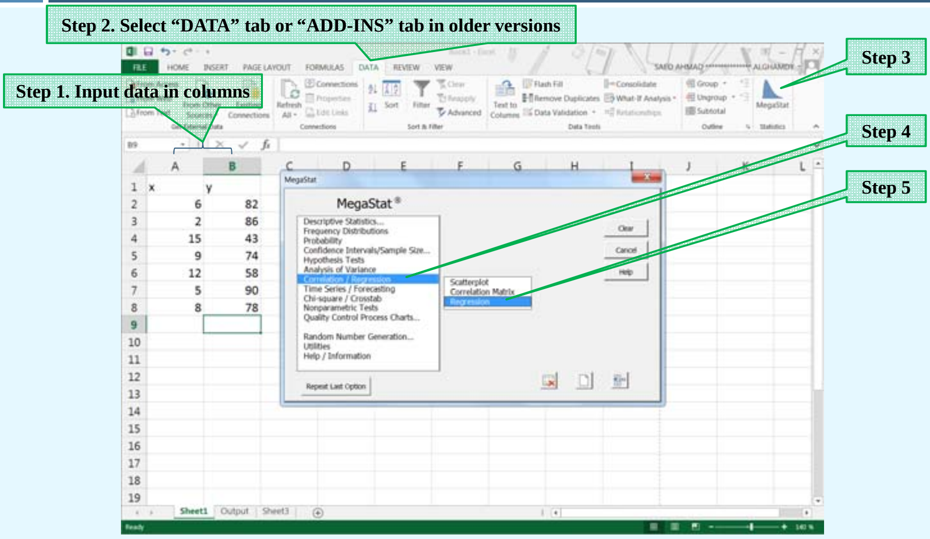
10-27 **Example "Scatter Plot"**



The plot indicates that there is a negative linear relationship. Thus, as the absences increased the final grades decreased on average.

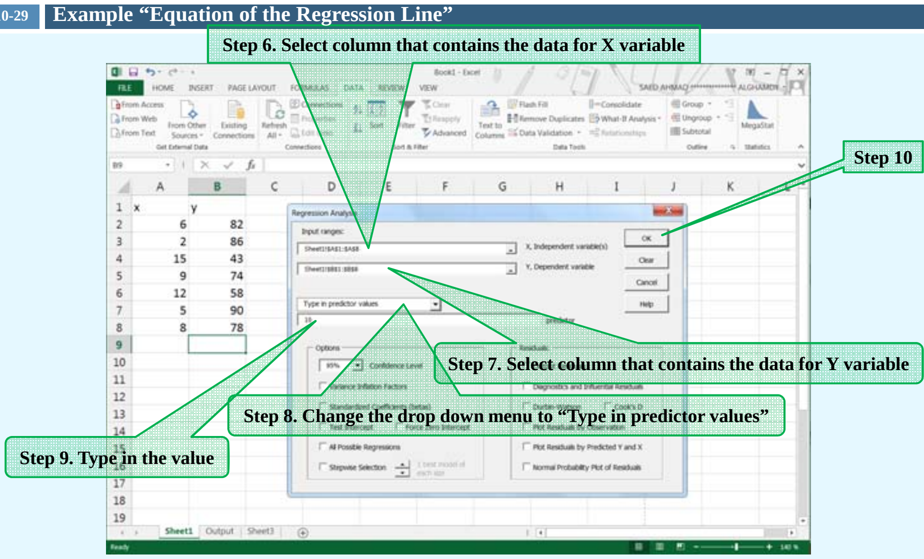
Application Using Excel

10-28 **Example “Equation of the Regression Line”**



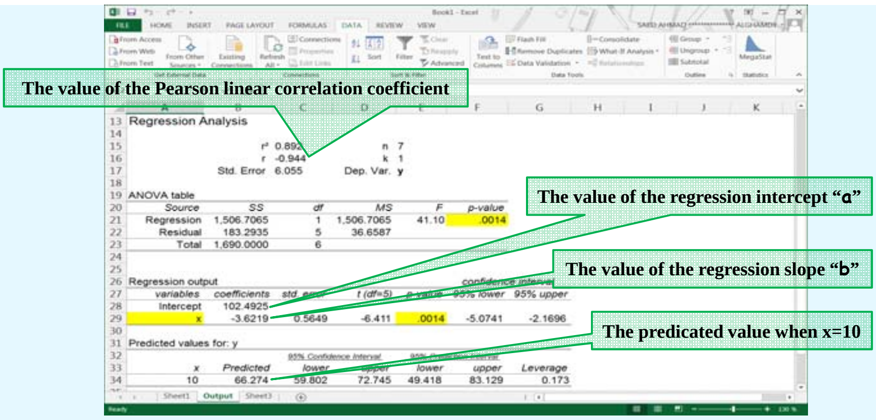
Application Using Excel

10-29 **Example “Equation of the Regression Line”**



Application Using Excel

10-30 **Example "Equation of the Regression Line"**



The equation of the fitted regression line is $y' = 102.493 - 3.622x$

Hence, a student who has been absent for 10 lectures, we expect his final grade to be 66 on average in his statistics class.

Summary

10-31

- One way to determine whether a relationship between variables exists is to use the statistical techniques known as correlation and regression.
- The strength and direction of the relationship are measured by the value of the correlation coefficient that assumes values between and including -1 and $+1$.
- The closer the value of the correlation coefficient is to -1 or $+1$, the stronger the linear relationship is between the variables.
- A value of -1 or $+1$ indicates a perfect linear relationship.

Summary

10-32

- To determine the shape of a relationship, one draws a scatter plot of the variables. If the relationship is linear, the data can be approximated by a straight line, called the *regression line,* or the *line of best fit.*
- The closer the value of the correlation coefficient is to -1 or $+1$, the closer the points will fit the line.
- The sign of the slope of the regression line indicates the direction of the relationship. Positive slope value means positive relationship and negative slope value means negative relationship.