மனோன்மணியம் சுந்தரனார் பல்கலைக்கழகம்

MANONMANIAM SUNDARANAR UNIVERSITY
TIRUNELVELI-627 012

தொலைநிலை தொடர் கல்வி இயக்ககம்

DIRECTORATE OF DISTANCE AND CONTINUING EDUCATION

B.Sc. MATHEMATICS

III YEAR

STATISTICS WITH EXCEL PROGRAMMING

Sub. Code: JNMA51

Prepared by

Dr. S. KALAISELVI

Assistant Professor

Department of Mathematics

Sarah Tucker College (Autonomous), Tirunelvei-7.

B.Sc. MATHEMATICS –III YEAR

JNMA51-STATISTICS WITH EXCEL PROGRAMMING

SYLLABUS

Unit I

Distribution of data: Characteristics of data-Frequency distribution- Procedure for Constructing a Frequency Distribution-Using Excel to Construct a Frequency Distribution-Relative Frequency Distribution -Cumulative Frequency.

Chapter 1: Sections 1.1 – 1.6.

Unit II

Histograms-Relative Frequency Histogram-Normal Distribution-Common Distribution Shapes-Skewness-Using XLSTAT for Histrograms - Graphs-Using Excel to Construct a Scatter plot-Correlation Coefficient.

Chapter 2: Sections 2.1 – 2.8.

Unit III

Time-Series Graph-Dot plots -Using XLSTAT for Stem plots -Bar Graphs-Using Excel to Create Bar Graphs-Pareto Charts-Pie Charts-Using Excel to Create Pie Charts-Frequency Polygon-Using Excel to Create Frequency Polygons.

Chapter 3: Sections 3.1 - 3.10

Unit IV

Descriptive statistics-Measures of Center-Mean -Using Excel to Calculate the Mean-Median-Using Excel to Find the Median.

Chapter 4: Sections 4.1 - 4.6

Unit V

Mode-Using Excel to Find the Mode-Midrange-Using Excel to Calculate the Midrange-Weighted Mean-Using Excel for Descriptive Statistics.

Chapter 5: Sections 5.1 - 5.6.

TEXT BOOK

Mario F. Triola, Elementary Statistics Using Excel, Fifth Edition, Pearson New International Edition, 2014.

JNMA51-STATISTICS WITH EXCEL PROGRAMMING

CONTENTS

Unit I

Distribution of data: Characteristics of data-Frequency distribution- Procedure for Constructing a Frequency Distribution-Using Excel to Construct a Frequency Distribution-Relative Frequency Distribution -Cumulative Frequency.

Chapter 1: Chapter 1: Sections 1.1 – 1.6.

1.Distribution of data:

In this chapter we are mainly concerned with the distribution of a data set, which is one of the following five characteristics that are typically most important. This chapter focuses mainly on the distribution of data.

1.1. Characteristics of Data:

1. Center: A representative value that indicates where the middle of the data set is located.
2. Variation: A measure of the amount that the data values vary.
3. Distribution: The nature or shape of the spread of the data over the range of values (such as bell-shaped).
4. Outliers: Sample values that lie very far away from the vast majority of the other sample values.
5. Time: Any change in the characteristics of the data over time.

Study Hint: Blind memorization is not effective in remembering information. To remember the above characteristics of data, it may be helpful to use a memory device (or mnemonic) for the first five letters CVDOT. Remembering the sentence “Computer Viruses Destroy or Terminate” is an easy way to help us remember the five key characteristics of data.

Critical Thinking and Interpretation: Going Beyond Formulas and Manual Calculations:

In the modern statistics course, it is not so important to memorize formulas or manually perform complex arithmetic calculations. Instead, we get results by using technology (a calculator or computer software), and then we focus on making practical sense of results through critical thinking. This chapter includes detailed steps for important procedures, but it is not necessary to master those steps in all cases. However, we recommend that in each case you perform a few manual calculations before using technology. This will enhance your understanding and help you acquire a better appreciation of the results obtained from the technology.

1.2. Frequency Distributions:

When one is working with large data sets, a frequency distribution (or frequency table) is often helpful in organizing and summarizing data. A frequency distribution helps us to understand the nature of the distribution of a data set.

Definition:

A frequency distribution (or frequency table) shows how data are partitioned among several categories (or classes) by listing the categories along with the number (frequency) of data values in each of them.

Consider the IQ scores of the low lead group listed in Table 1. Table 2 is a frequency distribution summarizing those IQ scores. The frequency for a particular class is the number of original values that fall into that class. For example, the first class in Table 2 has a frequency of 2, so 2 of the IQ scores are between 50 and 69 inclusive. The following standard terms are sometimes used in constructing frequency distributions and graphs.

Table 1 Full IQ Scores of Low Lead Group and High Lead Group															
Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		
High Lead Level (Group 3)															
82	93	85	75	85	80	101	89	80	94	88	104	88	88	83	104
96	76	80	79	75											

Table 2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Definitions:

Lower class limits are the smallest numbers that can belong to the different classes. (Table 2 has lower class limits of 50, 70, 90, 110, and 130.)

Upper class limits are the largest numbers that can belong to the different classes. (Table 2 has upper class limits of 69, 89, 109, 129, and 149.)

Class boundaries are the numbers used to separate the classes, but without the gaps created by class limits. Figure 1 shows the gaps created by the class limits from Table 2. In Figure 1 we see that the values of 69.5, 89.5, 109.5, and 129.5 are in the centers of those gaps, and following the pattern of those class

boundaries, we see that the lowest class boundary is 49.5 and the highest class boundary is 149.5. Thus the complete list of class boundaries is 49.5, 69.5, 89.5, 109.5, 129.5, and 149.5.

Class midpoints are the values in the middle of the classes. Table 2 has class midpoints of 59.5, 79.5, 99.5, 119.5, and 139.5. Each class midpoint is computed by adding the lower-class limit to the upper-class limit and dividing the sum by 2.

Class width is the difference between two consecutive lower-class limits (or two consecutive lower class boundaries) in a frequency distribution. Table 2 uses a class width of 20.

Caution

Finding the correct class width and class boundaries can be tricky. For class width, don't make the most common mistake of using the difference between a lower class limit and an upper class limit. See Table 2 and note that the class width is 20, not 19. For class boundaries, remember that they split the difference between the end of one class and the beginning of the next class, as shown in Figure 1.

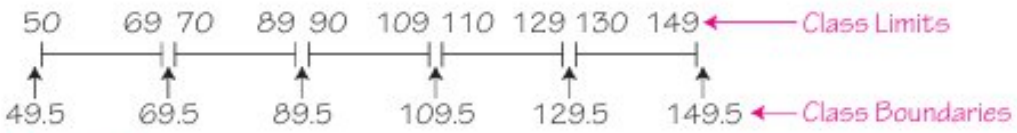


Figure 1 **Finding Class Boundaries from Class Limits in Table**

1.3. Procedure for Constructing a Frequency Distribution:

We construct frequency distributions (1) so that we can summarize large data sets, (2) so that we can analyze the data to see the distribution and identify outliers, and (3) so that we have a basis for constructing graphs (such as

histograms, introduced in the next section). Although technology can generate frequency distributions, the steps for manually constructing them are as follows:

1. Select the number of classes, usually between 5 and 20. The number of classes might be affected by the convenience of using round numbers.

2. Calculate the class width.

$$\text{Class width} \approx \frac{(\textit{maximum data value}) - (\textit{minimum data value})}{\textit{number of classes}}$$

Round this result to get a convenient number. (It's usually best to round up.) Using a specific number of classes is not too important, and it's usually wise to change the number of classes so that they use convenient values for the class limits.

3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.

4. Using the first lower class limit and the class width, list the other lower-class limits. (Add the class width to the first lower class limit to get the second lower class limit. Add the class width to the second lower class limit to get the third lower class limit, and so on.)

5. List the lower-class limits in a vertical column and then determine and enter the upper-class limits.

6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class. When constructing a frequency distribution, be sure the classes do not overlap. Each of the original values must belong to exactly one class. Include all classes, even those with a frequency of zero. Try to use the same width for all classes, although it is sometimes impossible to avoid open-ended intervals, such as "65 years or older."

Example 1: IQ Scores of Low Lead Group

Using the IQ scores of the low lead group in Table 1, follow the above procedure to construct the frequency distribution shown in Table 2. Use five classes.

Step 1: Select 5 as the number of desired classes.

Step 2: Calculate the class width. Note that we round 18.2 up to 20, which is a much more convenient number.

$$\text{Class width} \approx \frac{(\textit{maximum data value}) - (\textit{minimum data value})}{\textit{number of classes}}$$

$$= \frac{141-50}{5} = 18.2 \approx 20 \text{ (rounded up to a convenient number)}$$

Step 3: The minimum data value is 50 and it is a convenient starting point, so use 50 as the first lower class limit. (If the minimum value had been 52 or 53, we would have rounded down to the more convenient starting point of 50.)

Step 4: Add the class width of 20 to 50 to get the second lower class limit of 70. Continue to add the class width of 20 until we have five lower class limits. The lower class limits are therefore 50, 70, 90, 110, and 130.

Step 5: List the lower class limits vertically as shown in the margin. From this list, we identify the corresponding upper class limits as 69, 89, 109, 129, and 149.

Step 6: Enter a tally mark for each data value in the appropriate class. Then add the tally marks to find the frequencies shown in Table 2.

1.4. Using Excel to Construct a Frequency Distribution:

You can use Excel 2013, 2010, and 2007 to construct a frequency distribution from a list of sample data. (Excel for Mac 2011 does not include the Analysis Toolpak and cannot be used to construct a frequency distribution.) In Excel, the process of constructing a frequency distribution is called binning of the data, because each category acts like a separate bin into which we can pour some of the individual data values. Correct interpretation of an Excel frequency distribution requires that you know this important principle:

Excel's bins (classes) are based on upper class limits.

The following Excel procedure for constructing a frequency distribution uses the menu item of Histogram, which is a type of graph discussed in Section 3. Enter your sample data in a column of the Excel worksheet, or open an existing data set. To manually enter data, type the first value, and then press the Enter key. Type the second value, then press the Enter key, and so on.

1. Click on the Data tab in the Ribbon, and then click on Data Analysis in the Analysis section. If Data Analysis does not appear, you must install the Analysis add-in.

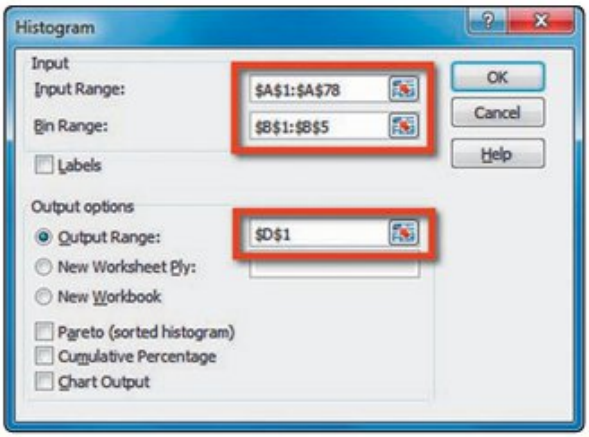
2. You should now see the Data Analysis dialog box. Select Histogram from the list of Analysis Tools and then click OK.

3. You should now see the Histogram dialog box.

a. First enter the Input Range. For example, if the data are listed in cells 1 through 78 of column A, select the cells that include the data or enter the input range of A1:A78.

b. For Bin Range, you have two options: (1) Leave the bin range blank and let Excel decide how to construct the frequency table. (2) If you want specific class limits, specify a range of cells that you have previously filled in with the values of the upper class limits. For the full IQ scores of the low lead group in Table 1 of the text, for example, the values range from 50 through 141, so you might enter these upper class limits in column B: 69, 89, 109, 129, 149 (as in Table 2 in the text). You would then select these cells with the bin values or enter a bin range of B1:B5. (Instead of typing B1:B5, you could click and drag the cells containing the data, and Excel will automatically insert the dollar sign symbols as shown in the accompanying display.) Excel would determine the frequency for the values up to 69, then the frequency for the values above 69 and up to 89, and so on.

c. Select Output Range under Output options, and then enter a cell location for the frequency table, such as D1. Alternatively, select New Worksheet Ply if you want the results to appear on a new worksheet. Finally, click OK.



Shown below is the Excel display that results from the IQ scores of the low lead group that are listed in Table 2 of the text. This Excel display results from specifying the class limits given in the second option in part b above (that is, the upper class limits of 69, 89, 109, 129, and 149 were entered in column B, and the bin range of B1:B5 was used). Compare Excel's display with a standard frequency table to see how Excel summarizes or "bins" the same data:

EXCEL

Bin	Frequency
69	2
89	33
109	35
129	7
149	1
More	0

← There are 33 values between 70 and 89

Table 2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

You can also use Excel to generate a frequency table with cumulative percentages. Follow the same three steps listed above, but in part c of Step 3, click on the box labeled Cumulative Percentage before clicking OK. So far we have

discussed frequency distributions using only quantitative data sets, but frequency distributions can also be used to summarize categorical (or qualitative or attribute) data, as illustrated in Example 2.

Example 2: East Haven Police Department Traffic Tickets

Table 3 summarizes the race/ethnic classifications recorded on traffic tickets issued by Connecticut's East Haven Police Department during a recent nine-month period. Here is an interesting and revealing fact about the data: Table 3 shows that 18 of those given tickets were classified by police as being Hispanic, but in fact, 209 of those given tickets had Hispanic names!

Table 3 East Haven Traffic Tickets

Race	Frequency
White	329
Black	15
Asian	0
Hispanic	18
White/Hispanic	4
Blank (no indication)	5

1.5. Relative Frequency Distribution:

A variation of the basic frequency distribution is a relative frequency distribution or percentage frequency distribution, in which each class frequency is replaced by a relative frequency (or proportion) or a percentage. In this text we use the term "relative frequency distribution" whether we use relative frequencies or percentages. Relative frequencies and percentages are calculated as follows. Relative frequency for a class = frequency for a class sum of all frequencies Percentage for a class = frequency for a class sum of all frequencies * 100,

Table 4 is an example of a relative frequency distribution. It is a variation of Table 2 in which each class frequency is replaced by the corresponding percentage value. Because there are 78 data values, divide each class frequency by 78, and then multiply by 100%. The first class of Table 2 has a frequency of 2, so divide 2 by 78 to get 0.0256, and then multiply by 100% to get 2.56%, which we rounded to 2.6%. The sum of the percentages should be 100%, with a small discrepancy allowed for rounding errors, so a sum such as 99% or 101% is acceptable. The sum of the per centages in Table 4 is 100.1%.

The sum of the percentages in a relative frequency distribution must be very close to 100%.

Table 4 Relative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2.6%
70–89	42.3%
90–109	44.9%
110–129	9.0%
130–149	1.3%

1.6. Cumulative Frequency Distribution:

Another variation of a frequency distribution is a cumulative frequency distribution in which the frequency for each class is the sum of the frequencies for that class and all previous classes. Table 5 is a cumulative frequency distribution based on Table 2. Using the original frequencies of 2, 33, 35, 7, and 1, we add 2 + 33 to get the second cumulative frequency of 35, then we add 2 + 33 + 35 to get the third, and so on. See Table 5, and note that in addition to the

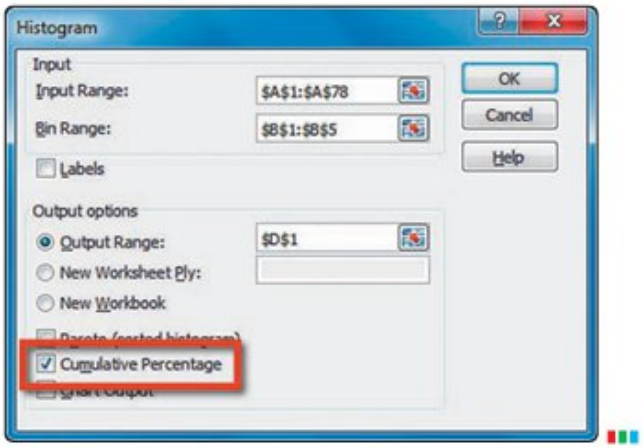
use of cumulative frequencies, the class limits are replaced by "less than" expressions that describe the new ranges of values.

Table 5 Cumulative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Cumulative Frequency
Less than 70	2
Less than 90	35
Less than 110	70
Less than 130	77
Less than 150	78

Using Excel for a Cumulative Frequency Distribution

Excel can generate a frequency table with cumulative percentages. Follow the same steps listed earlier in this section under Using Excel to Construct a Frequency Distribution, but in part c of Step 3, click on the box labeled Cumulative Percentage before clicking OK, as shown in the accompanying display.



Using Frequency Distributions to Understand Data Earlier, we noted that a frequency distribution can help us understand the distribution of a data set, which is the nature or shape of the spread of the data over the range of values (such as bell-shaped). In statistics we are often interested in determining whether the data have a normal distribution. Data that have an approximately normal distribution are characterized by a frequency distribution with the following features:

Normal Distribution

1. The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
2. The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

Table 6 satisfies these two conditions. The frequencies start low, increase to the maximum of 56, then decrease to a low frequency. Also, the frequencies of 1 and 10 that precede the maximum are a mirror image of the frequencies 10 and 1 that follow the maximum. Real data sets are usually not so perfect as Table 6, and judgment must be used to determine whether the distribution comes "close enough" to satisfying those two conditions.

Table 6 Frequency Distribution Showing a Normal Distribution

Score	Frequency	Normal Distribution
50–69	1	← Frequencies start low, . . .
70–89	10	
90–109	56	← Increase to a maximum, . . .
110–129	10	
130–149	1	← Decrease to become low again.

The following examples illustrate how frequency distributions are used to describe, explore, and compare data sets.

Example 3:

Describing Data: How Were the Weights Obtained in California?

When collecting weights of people, it's better to actually weigh people than to ask them what they weigh. People often tend to round way down, so that a weight of 196 lb might be reported as 170 lb. Table 7 summarizes the last digits of the weights of 100 people used in the California Health Interview Survey. If people are actually weighed on a scale, the last digits of weights tend to have frequencies that are approximately the same, but Table 6 shows that the vast majority of weights have last digits of 0 or 5, and this is strong evidence that people reported their weights and were not physically weighed. (Also, the word "interview" in the title of the California Health Interview Survey reveals that people were interviewed and were not physically measured.)

Table 7 Last Digits of Weights from the California Health Interview Survey

Last Digit of Weight	Frequency
0	46
1	1
2	2
3	3
4	3
5	30
6	4
7	0
8	8
9	3

Example 4:

Exploring Data: What Does a Gap Tell Us?

Table 8 is a frequency distribution of the weights (grams) of randomly selected pennies. Examination of the frequencies reveals a large gap between the lightest pennies and the heaviest pennies. This suggests that we have two different popula tions: Pennies made before 1983 are 95% copper and 5% zinc, but pennies made after 1983 are 2.5% copper and 97.5% zinc, which explains the large gap between the lightest pennies and the heaviest pennies in Table 8.

Table 8 Randomly Selected Pennies

Weight (grams) of Penny	Frequency
2.40–2.49	18
2.50–2.59	19
2.60–2.69	0
2.70–2.79	0
2.80–2.89	0
2.90–2.99	2
3.00–3.09	25
3.10–3.19	8

The presence of gaps can suggest that the data are from two or more different populations. The converse of this principle is not true, because data from different populations do not necessarily result in gaps.

Example 5:

Comparing IQ Scores of the Low Lead Group and the High Lead Group

Table 1, which is given with the Chapter Problem at the beginning of this chap ter, lists IQ scores from the low lead group and the high lead group. Because the sample sizes of 78 and 21 are so different, a comparison of frequency distribu tions is not easy, but Table 9 shows the relative frequency distributions for those two groups. By comparing those relative frequencies, we see that the majority of

children in the low lead group had IQ scores of 90 or higher, but the majority of children in the high lead group had IQ scores below 90. This suggests that perhaps high lead exposure has a detrimental effect on IQ scores.

Table 9 IQ Scores from the Low Lead Group and the High Lead Group

IQ Score	Low Lead Group	High Lead Group
50–69	2.6%	
70–89	42.3%	71.4%
90–109	44.9%	28.6%
110–129	9.0%	
130–149	1.3%	

Unit II

Histograms-Relative Frequency Histogram-Normal Distribution-Common Distribution Shapes-Skewness-Using XLSTAT for Histrograms - Graphs-Using Excel to Construct a Scatter plot-Correlation Coefficient.

Chapter 2: Sections 2.1 -2.8.

2.1. Histograms:

While a frequency distribution is a useful tool for summarizing data and investigating the distribution of data, an even better tool is a histogram, which consists of a graph that is easier to interpret than a table of numbers.

Definition:

A histogram is a graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.

A histogram is basically a graph of a frequency distribution. For example, Figure 2 shows the histogram corresponding to the frequency distribution given in Table 2. Class frequencies should be used for the vertical scale and that scale should be labeled as in Figure 2. The bar locations on the horizontal scale are usually labeled with one of the following: (1) class boundaries (as shown in Figure 2), (2) class midpoints, or (3) lower class limits. The first and second options are technically correct, while the third option introduces a small error. It is often easier for us mere mortals to use class midpoints for the horizontal scale. Histograms can usually be generated using technology.

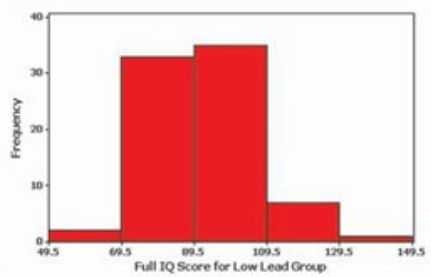


Figure 2 Histogram

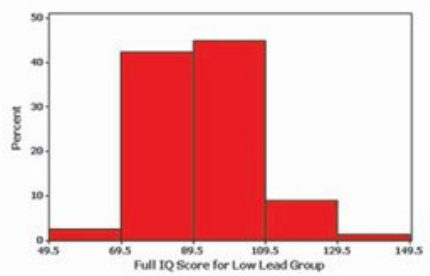


Figure 3 Relative Frequency Histogram

2.2. Relative Frequency Histogram:

A relative frequency histogram has the same shape and horizontal scale as a histo gram, but the vertical scale uses relative frequencies (as percentages or proportions) instead of actual frequencies. Figure 3 is the relative frequency histogram corresponding to Figure 2.

Critical Thinking: Interpreting Histograms

Even though creating histograms is more fun than human beings should be allowed to have, the ultimate objective is not creating a histogram, but rather understanding something about the data. Analyze the histogram to see what can be learned about "CVDOT": the center of the data, the variation, the distribution, and whether there are any outliers (values far away from the other values). Examining Figure 2, we see that the histogram is centered close to 90, the values vary from around 50 to 150, and the distribution is roughly bell-shaped.

2.3. Normal Distribution:

When graphed as a histogram, a normal distribution has a "bell" shape similar to the one superimposed in Figure 4. In a normal distribution, (1) the frequencies increase to a maximum and then decrease, and (2) the graph has symmetry, with the left half of the histogram being roughly a mirror image of the right half. Figure

4 shows that the histogram in Figure 2 roughly satisfies those two conditions, so we say that the

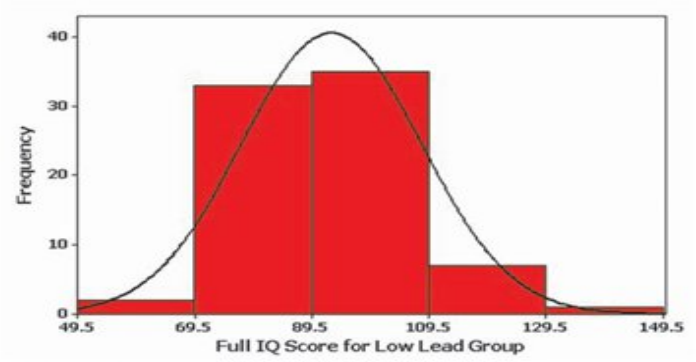


Figure 4 Bell-Shaped Distribution
Because this histogram is roughly bell-shaped, we say that the data have a *normal distribution*.

IQ scores are approximately normally distributed. (There are more advanced and less subjective methods for determining whether the distribution is a normal distribution.) Many statistical methods require that sample data come from a population having a distribution that is approximately a normal distribution, and we can often use a histogram to determine whether this requirement is satisfied.

2.4. Common Distribution Shapes

The histograms shown in Figure 5 depict four common distribution shapes. We have already discussed the characteristics of the normal distribution. With a uniform distribution, the different possible values occur with approximately the same frequency, so the heights of the bars in the histogram are approximately uniform, as in Figure 5(b). Figure 5(b) depicts outcomes of digits from state lotteries.

2.5. Skewness

A distribution of data is skewed if it is not symmetric and extends more to one side than to the other. Data skewed to the right (also called positively skewed) have a longer right tail, as in Figure 5(c), which depicts annual incomes (in

thousands of dollars) of adult Americans. Data skewed to the left (also called negatively skewed) have a longer left tail, as in Figure 5(d). Distributions skewed to the right are more common than those skewed to the left because it's often easier to get exceptionally large values than values that are exceptionally small. With annual incomes, for

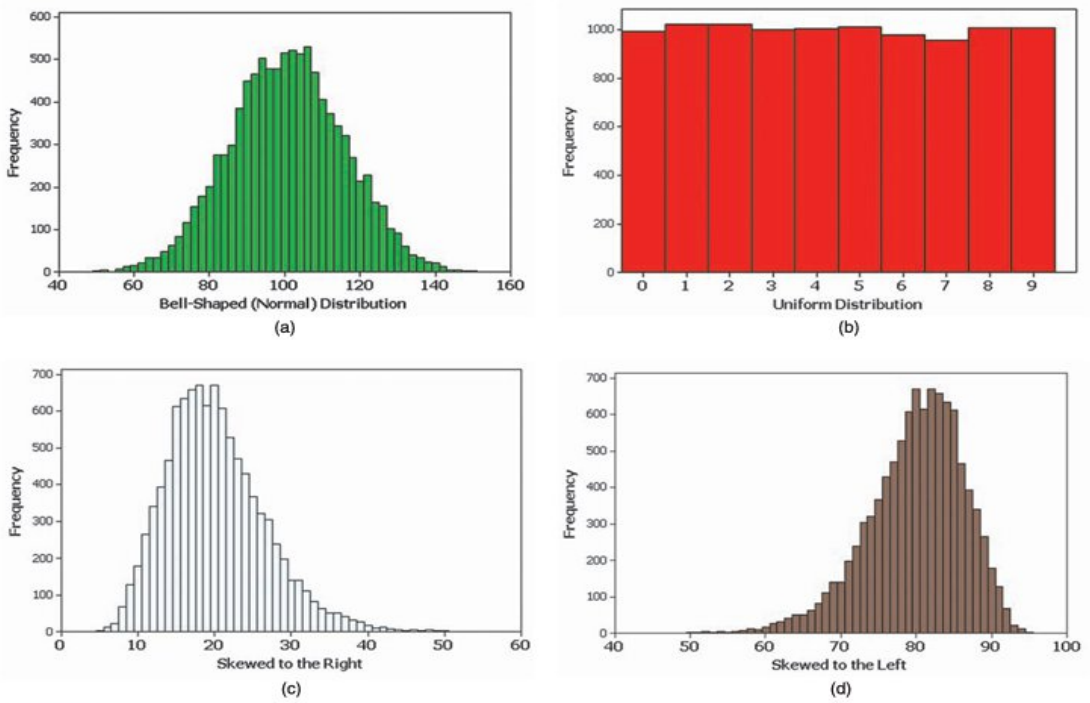
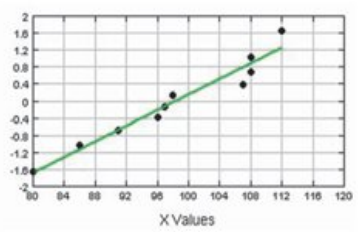


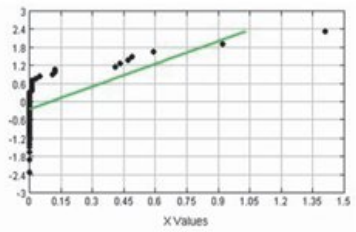
Figure 5 Common Distributions

example, it's impossible to get values below zero, but there are a few people who earn millions or billions of dollars in a year. Annual incomes therefore tend to be skewed to the right. Assessing Normality: Normal Quantile Plot Some really important methods have a requirement that sample data must be from a population having a normal distribution. We can see that a histogram is often helpful in determining whether the normality requirement is satisfied. However, histograms are not very helpful with small data sets. There are methods for assessing normality— that is, determining whether the sample data are from a

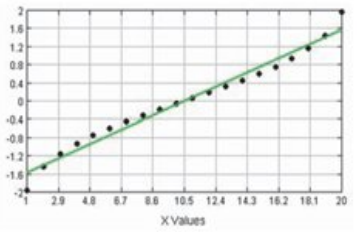
normally distributed popula tion. There is also a procedure for constructing normal quantile plots, which involve plotting transformed sample values. Normal quantile plots are easy to generate using technology such as XLSTAT. Interpretation of a normal quantile plot is based on the following criteria: Criteria for Assessing Normality with a Normal Quantile Plot Normal Distribution: The population distribution is normal if the pattern of the points in the normal quantile plot is reasonably close to a straight line, and the points do not show some systematic pattern that is not a straight-line pattern. Not a Normal Distribution: The population distribution is not normal if the normal quantile plot has either or both of these two conditions: The points show some systematic pattern that is not a straight-line pattern. The following are examples of normal quantile plots.



Normal Distribution: The points are reasonably close to a straight-line pattern, and there is no other systematic pattern that is not a straight-line pattern.



Not a Normal Distribution: The points do not lie reasonably close to a straight line.



Not a Normal Distribution: The points show a systematic pattern that is not a straight-line pattern.

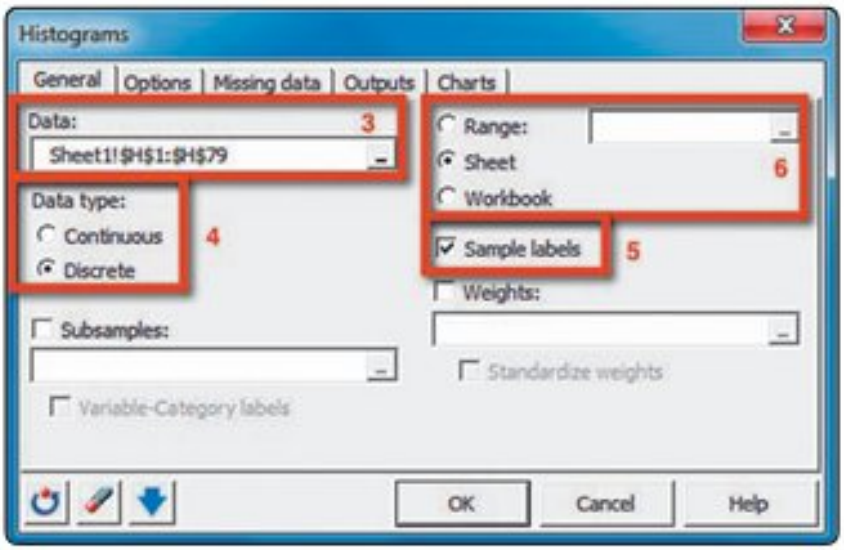
2.6. Using XLSTAT for Histograms

XLSTAT can be used to quickly and easily construct histograms. To generate a his togram using XLSTAT, first enter or open the data set. Use XLSTAT to generate the histogram as follows.

1. Click on the XLSTAT tab in the Ribbon. (If XLSTAT is not available, you must install it.)

2. Click on the somewhat creepy looking Visualizing Data button, and then select Histograms from the dropdown menu.

3. You should now see the Histograms dialog box. First enter the data range in the Data box. For example, in IQLEAD.xls the data (including sample label) are listed in cells 1 through 79 of column H. Select the cells that include the data or enter the input range of H1:H79.



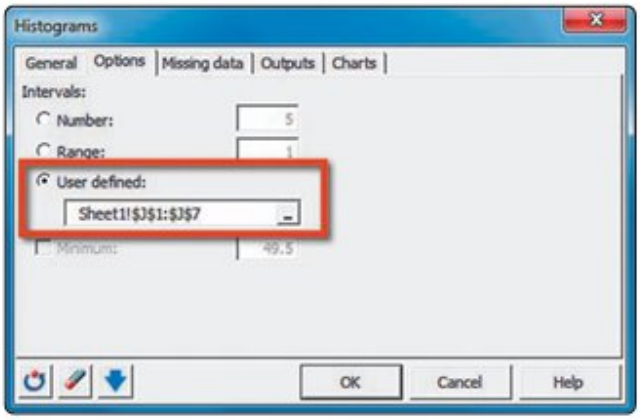
4. Next select Data type. In this example, select Discrete, since IQ scores are whole numbers only.

5. Check the Sample labels box if the first cell of your Data includes a name (or label) of the data instead of a data value. Uncheck the box if the first cell in cludes a data value instead of a label.

6. Select Sheet to display the histogram on a new worksheet. Select Range and specify a cell if you want the histogram to be displayed on the current worksheet.

7. XLSTAT will automatically define class boundaries. If you prefer to define your own class boundaries, first enter all of the desired class boundaries in a single column on your spreadsheet. These must be entered in ascending order, starting with the lower class boundary of the first class followed by the upper class boundaries of all classes (49.5, 69.5, 89.5, 109.5, 129.5, 149.5 in this example). If the Sample labels box is checked in Step 5, be sure that you also

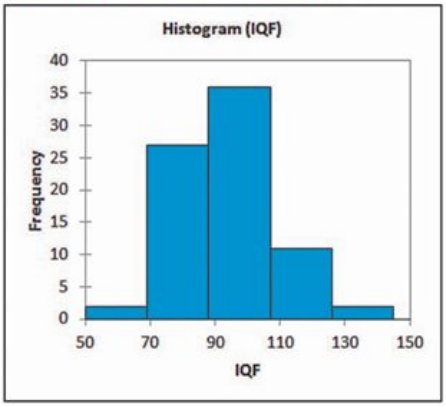
include a label for the class boundaries data. Click the Options tab, select User Defined, and select the cells that include the user-defined class boundaries (and data label if applicable).



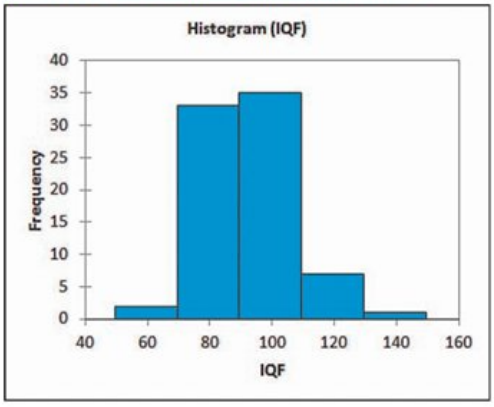
8. Click the Charts tab and confirm that the Histograms box is checked and the Bars option is selected. Also confirm that Frequency is selected in the Ordi nate of histograms box. If you want to construct a relative frequency histogram, simply select Relative Frequency instead.

9. Click OK. Click Continue on the XLSTAT–Selections dialog box, and the histogram will be displayed as shown on the next page. The histogram with default class boundaries is on the left, and the histogram with user-defined class boundaries (matching Figure 2) is on the right. The relative frequency histogram is also shown.

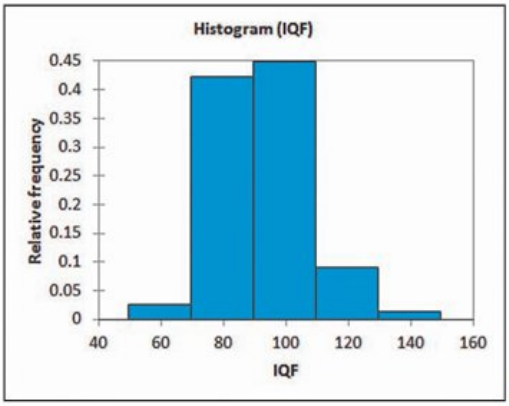
XLSTAT: Default Class Boundaries



XLSTAT: User-Defined Class Boundaries



XLSTAT: Relative Frequency Histogram



Graphs That Enlighten and Graphs That Deceive

The histogram is a graph that enlightens in the sense that it gives us better understanding of data. In this section we introduce other commonly used graphs that enlighten. We also discuss some graphs that deceive in the sense that they tend to create impressions about data that are somehow misleading or wrong.

The days of charming and primitive hand-drawn graphs are well behind us, and technology now provides us with powerful tools for generating a wide variety of different graphs.

Graphs That Enlighten

Scatterplots

A scatterplot (or scatter diagram) is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first

(x) variable, and the vertical axis is used for the second variable. The pattern of the plotted points is often helpful in determining whether there is a correlation (or relationship) between the two variables.

Example 1:

Correlation: Waist and Arm Circumference

Data Set 1 in Appendix: Data Sets includes the waist circumferences (cm) and arm circumferences (cm) of randomly selected males. Figure 6 is a scatterplot of the paired waist/arm measurements. The points show a pattern of increasing values from left to right. This pattern suggests that there is a correlation, or relationship, between waist circumference and arm circumference in males.

2.7. Using Excel to Construct a Scatterplot

Scatterplots can be generated from a list of paired sample data by using the XLSTAT add-in or by using Excel itself. Instructions for both methods are provided.

Using XLSTAT to Construct a Scatterplot

1. Enter or open two columns of paired data. As shown in Example 1, the data set MBODY from Data Set 1 in Appendix: Data Sets includes the waist sizes (WAIST) and arm circumferences (ARMC) of 40 males. We will illustrate the construction of a scatterplot using these paired data.

2. Click on the XLSTAT tab in the Ribbon.

3. Click on the Visualizing Data button, and then select Scatter plots from the dropdown menu.

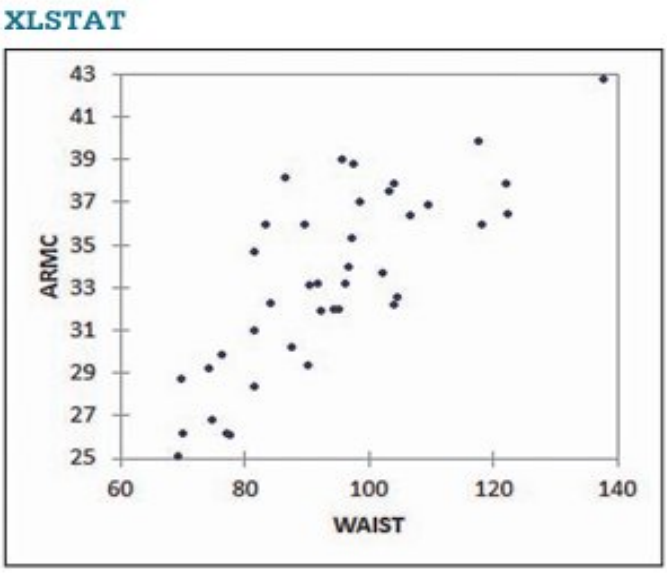
4. You should now see the Scatter plots dialog box. Click the X box, and select the cells that contain the data for the horizontal axis (WAIST) or enter the data

range L1: L41. Next, click the Y box, and select the cells that contain the data for the vertical axis (ARMC) or enter the data range M1:M41.

5. If the variable label is included in the data range, check the Variables labels box. In this example, that box should be checked.

6. Select Sheet to display the scatterplot on a new worksheet. Select Range, and specify a cell if you want the scatter plot to be displayed on the current worksheet.

7. Click OK. Click Continue on the XLSTAT – Selections dialog box, and the scatterplot will be displayed as shown below



Using Excel (Instead of XLSTAT) to Construct a Scatterplot:

The following steps provide basic instructions for creating a chart in Excel. For more comprehensive directions, click the HELP button in Excel and enter "Create a chart" in the search box.

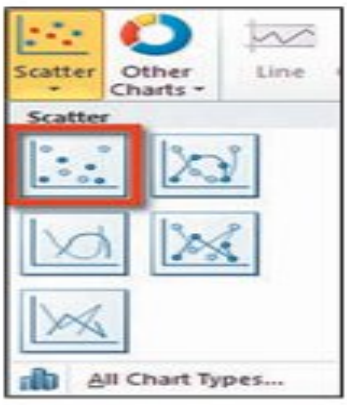
1. Enter or open two columns of paired data. For example, the data set MBODY from Data Set 1 in Appendix: Data Sets includes the waist sizes (WAIST) and arm circumferences (ARMC) of 40 males. We will illustrate the construction of a scatterplot using these paired data.

2. Select the cells that contain the data to be used in the chart by clicking on the top cell in the first data column and, while holding the mouse button down, dragging the cursor to the bottom cell in the final data column. In this example, cells L1:M41 are selected.

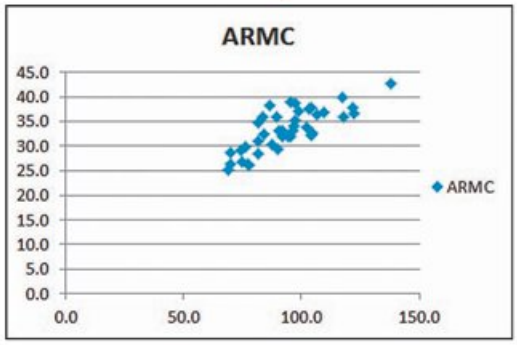
3. Click the Insert tab on the Ribbon, and then click Scatter in the Charts section of the Ribbon.

4. Select the Scatter with only markers graph type.

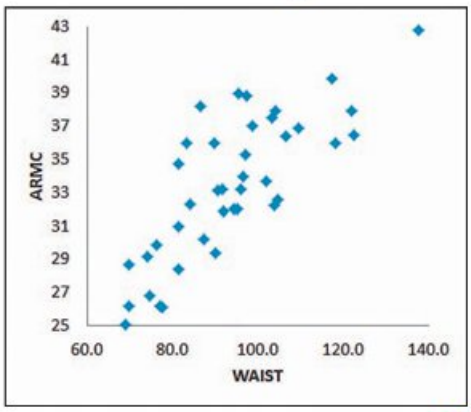
5. The scatterplot will appear with default formatting as shown. To further improve the scatterplot, delete the chart title and legend, add axis labels, and adjust the axis scale, using minimum and maximum values suitable for the scat terplot. Right click on the axis and select Format Axis…



EXCEL: Default Formatting



EXCEL: Modified Formatting



2.8. Correlation Coefficient

Examples 1 and 2 involve making decisions about a correlation based on subjective judgments of scatterplots, but there are more objective methods. Those methods involve calculating a value of a linear correlation coefficient r, which is a value between -1 and 1. If r is close to -1 or close to 1, there appears to be a correlation, but if r is close to 0, there does not appear to be a correlation. For the data depicted in the scatterplot of Figure 6, r = 0.788, and the data in the scatterplot of Figure 7 result in r =-0.213. Once you have studied the calculation and interpretation of a value of r, so that a decision about correlation is much more objective. Even though the methods based on calculations of r are much more objective than the subjective in perpetration of a scatterplot, it is always wise to construct a scatterplot first, so that we can see characteristics that cannot be seen by examining the list of paired data values.

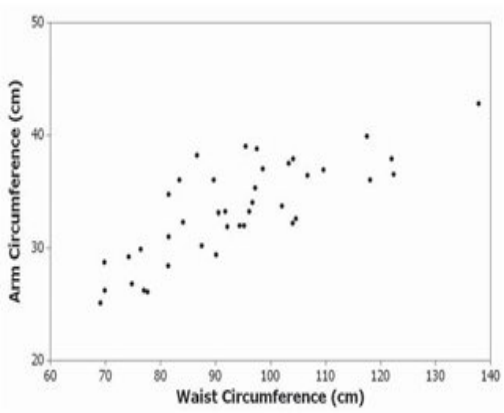


Figure 6 Waist Circumference and Arm Circumference in Males

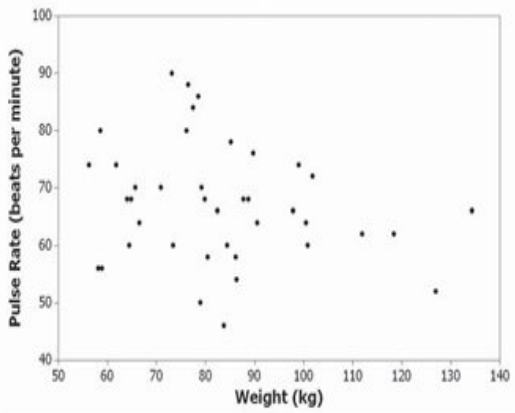


Figure 7 Weight and Pulse Rate in Males

Example 2:

No Correlation: Weight and Pulse Rate

Data Set 1 in Appendix: Data Sets includes weights (kg) and pulse rates (beats per minute) of randomly selected males. Figure 7 is a scatterplot of the paired weight/ pulse rate measurements. The points in Figure 7 do not show any obvious pattern, and this lack of a pattern suggests that there is no correlation, or relationship, between the weight and pulse rate of males.

Example 3:

Clusters and a Gap

Consider the scatterplot in Figure 8. It consists of paired data consisting of the weight (grams) and year of manufacture for each of 72 pennies. This scatterplot shows two very distinct clusters separated by a gap, which can be explained by the inclusion of two different populations: pre-1983 pennies are 97% copper and 3% zinc, whereas post-1983 pennies are 3% copper and 97% zinc. If we ignored the characteristic of the clusters, we might incorrectly think that there is a relationship between the weight of a penny and the year it was made. If we examine the two groups separately, we see that there does not appear to be a

relationship between the weights of pennies and the years in which they were produced.

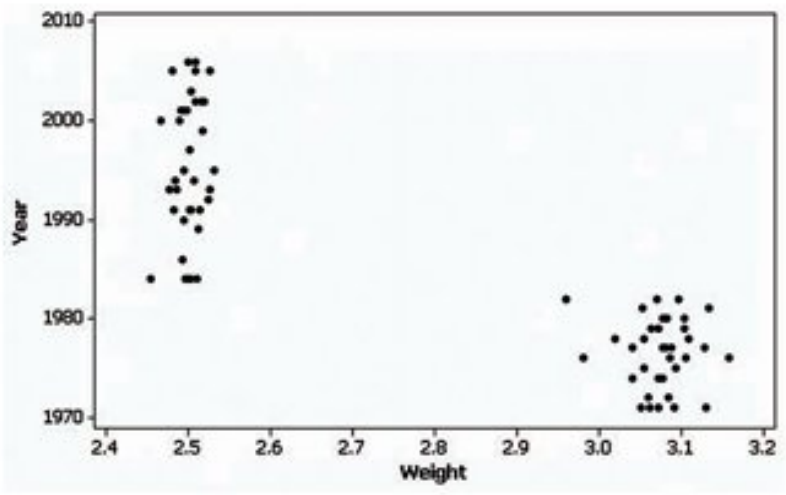


Figure 8 Weights (g) of Pennies and Years of Production

Unit III

Time-Series Graph-Dot plots -Using XLSTAT for Stem plots -Bar Graphs-Using Excel to Create Bar Graphs-Pareto Charts-Pie Charts-Using Excel to Create Pie Charts-Frequency Polygon-Using Excel to Create Frequency Polygons.

Chapter 3: Sections 3.1-3.10

3.1. Time-Series Graph:

A time-series graph is a graph of time-series data, which are quantitative data that have been collected at different points in time, such as monthly or yearly.

Example 4:

Time-Series Graph: Dow Jones Industrial Average

The time-series graph shown in Figure 9 depicts the yearly high values of the Dow Jones Industrial Average (DJIA) for the New York Stock Exchange. This graph shows a fairly consistent pattern of increases from 1980 to 1999, but the DJIA high values have been much more erratic in recent years.

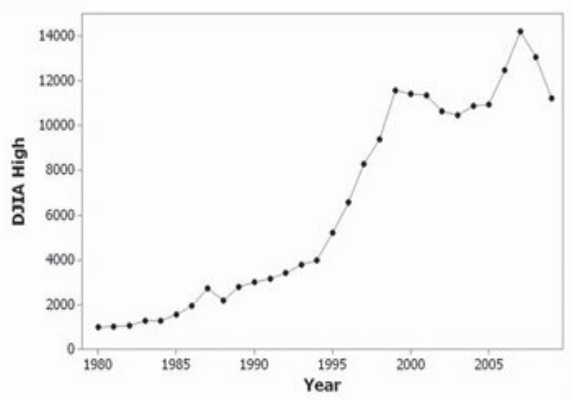


Figure 9 Dow Jones Industrial Average

Using Excel to Create Time-Series Graphs

Creating time-series graphs in Excel is simple and requires only a few steps. The procedure for creating time-series graphs is very similar to that for creating other types of charts in Excel, including pie charts and bar charts. The following steps

provide basic instructions for creating a chart in Excel. For more comprehensive directions, click the HELP button in Excel and enter “Create a chart” in the search box.

1.Enter the data values in a column of the spreadsheet.

2. Select the cells that contain the data to be used in the chart by clicking on the top cell in the data column and, while holding the mouse button down, dragging the cursor to the bottom cell in the data column.

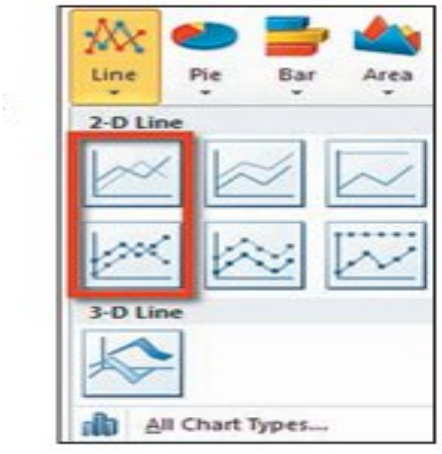
3. Click the Insert tab on the Ribbon, and then click Line in the Charts section of the Ribbon.

4. Select the Line or Line with markers graph type. The graph will appear with default formatting. The line graph can be edited for improved appearance.

a. Change the layout of the graph, using the Chart Layout section of the Ribbon.

b. Right click on the axis and select Format Axis… to adjust axis options and formatting.

c. Right click on any data point and select Format Data Series... to adjust line and marker formatting.



3.2. Dotplots:

A dotplot consists of a graph in which each data value is plotted as a point (or dot) along a horizontal scale of values. Dots representing equal values are stacked.

Example 5:

Dotplot: IQ Scores of Low Lead Group

Figure 10 shows a dotplot of the IQ scores of the low lead group from Table 1 included with the Chapter Problem at the beginning of this chapter. The five stacked dots above the position at 76 indicate that five of the IQ scores are 76. There are three dots stacked above 80, so three of the IQ scores are 80. This dotplot reveals the distribution of the IQ scores. It is possible to recreate the original list of data values, because each data value is represented by a single point.

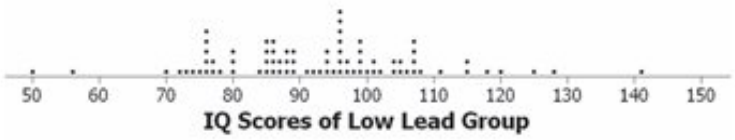


Figure 10 Dotplot: IQ Scores of Low Lead Group

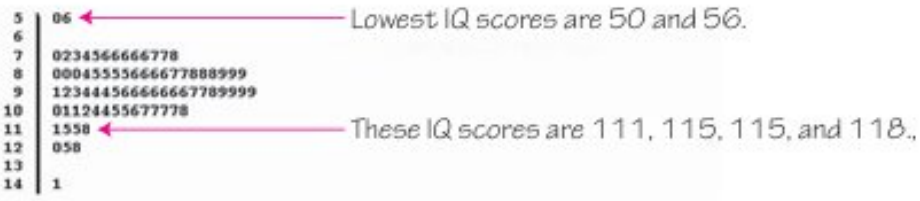
Stemplots:

A stemplot (or stem-and-leaf plot) represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit). Better stemplots are often obtained by first rounding the original data values. Also, stemplots can be expanded to include more rows and can be con densed to include fewer rows, as in Exercise 26. One advantage of the stemplot is that we can see the distribution of data while keeping the original data values. Another advantage is that constructing a stemplot is a quick way to sort data (arrange them in order), which is required for some statistical procedures (such as finding a median, or finding percentiles).

Example 6:

Stemplot: IQ Scores of Low Lead Group

The following stemplot displays the IQ scores of the low lead group in Table 1 given with the Chapter Problem. The lowest IQ score of 50 is separated into its stem of 5 and its leaf of 0, and each of the remaining values is separated in a similar way. The stems and leaves are arranged in increasing order, not the order in which they occur in the original list. Note that if you turn the stemplot on its side, you can see distribution of the IQ scores in the same way you would see it in a histogram.



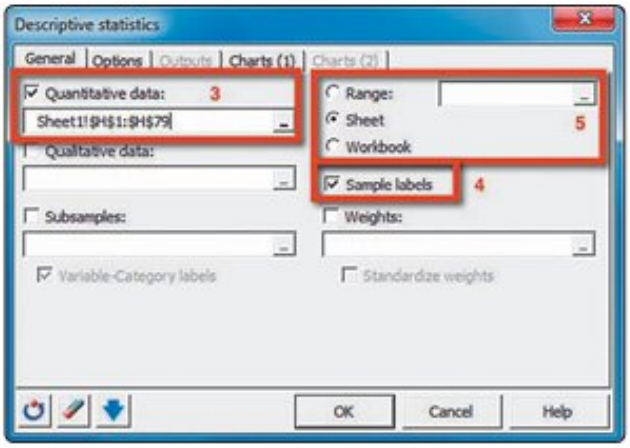
3.3. Using XLSTAT for Stemplots:

XLSTAT can be used to quickly and easily construct stemplots. To generate a stem plot using XLSTAT, first enter or open the data set. Use XLSTAT to generate the stemplot as follows.

1. Click on the XLSTAT tab in the Ribbon.

2. Click on the Describing Data button, then select Descriptive statistics from the dropdown menu.

3. You should now see the Descriptive statistics dialog box. First, check the Quantitative data box and enter the data range. For example, in



IQLEAD.xls the data (including sample label) are listed in cells 1 through 79 of column H. Select the cells that include the data or enter the input range of H1: H79. 4. Check the Sample Labels box if the first cell of your data column includes a description of the data rather than a data point. Uncheck that box if the first cell includes a data point. That box is checked in this example. 5. Select Sheet to display the histogram on a new worksheet. Select Range and specify a cell if you want the histogram to be displayed on the current worksheet. 6. Click the Charts (1) tab and confirm that the Stem-and-leaf-plots box is checked. 7. Click OK. Click Continue on the XLSTAT – Selections dialog box, and the stemplot will be displayed as shown below.

XLSTAT

```
 5 06
 6
 7 0234566666778
 8 000455556666778889999
 9 12344456666666778999 9
10 011244556777778
11 1558
12 058
13
14 1
```

3.4. Bar Graphs:

A bar graph uses bars of equal width to show frequencies of categories of categorical (or qualitative) data. The vertical scale represents frequencies or relative frequencies. The horizontal scale identifies the different categories of qualitative data. The bars may or may not be separated by small gaps. A multiple bar graph has two or more sets of bars and is used to compare two or more data sets.

3.5. Using Excel to Create Bar Graphs:

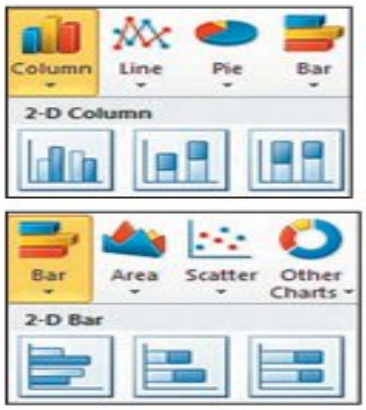
Creating bar graphs in Excel is simple and requires only a few steps. The procedure for creating bar graphs is very similar to that for creating time-series graphs and other types of charts in Excel.

1. Enter the names of categories in column A, and then enter the corresponding frequency or percentage values in column B.

2. Select the cells that contain the data to be used in the chart by clicking on the top cell in column A and, while holding the mouse button down, dragging the cursor to the bottom cell in column B.

3. Click the Insert tab on the Ribbon, and then click Column or Bar in the Charts section of the Ribbon.

4. Select column or bar chart type of your preference. The graph will appear with default formatting and can be edited for improved appearance.



For more comprehensive directions on creating and editing bar graphs, click the HELP button in Excel and enter "**Create a chart**" in the search box.

Example 7:

Multiple Bar Graph of Income by Gender

See Figure 11 for a multiple bar graph of the median incomes of males and females in different years. The data are from the U.S. Census Bureau, and the values for 2010 are projected. From this graph we see that males consistently have much higher median incomes than females, and that both males and females have steadily increasing incomes over time. Comparing the heights of the bars from left to right reveals that the ratios of incomes of males to incomes of females appear to be decreasing, which indicates that the gap between male and female median incomes is gradually becoming smaller.

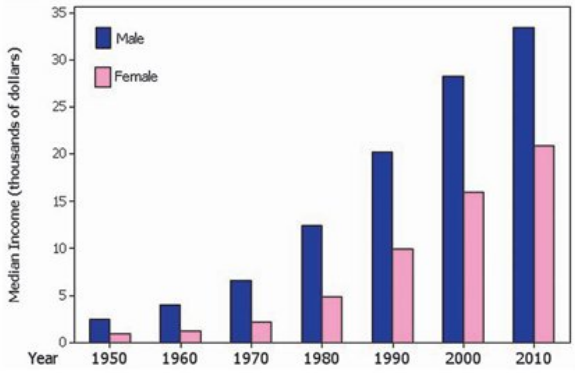


Figure 11 Multiple Bar Graph: Median Income by Gender

3.6. Pareto Charts:

When we want a bar graph to draw attention to the more important categories, we can use a Pareto chart, which is a bar graph for categorical data, with the added stipulation that the bars are arranged in descending order according to frequencies. The vertical scale in a Pareto chart represents frequencies or relative frequencies. The horizontal scale identifies the different categories of qualitative data. The bars de crease in height from left to right.

Example 8:

Pareto Chart: What Contributes Most to Happiness?

In a Coca-Cola survey of 12,500 people, respondents were asked what contributes most to their happiness. Figure 12 is a Pareto chart summarizing the results. We see that family or partner is by far the most frequently selected choice.

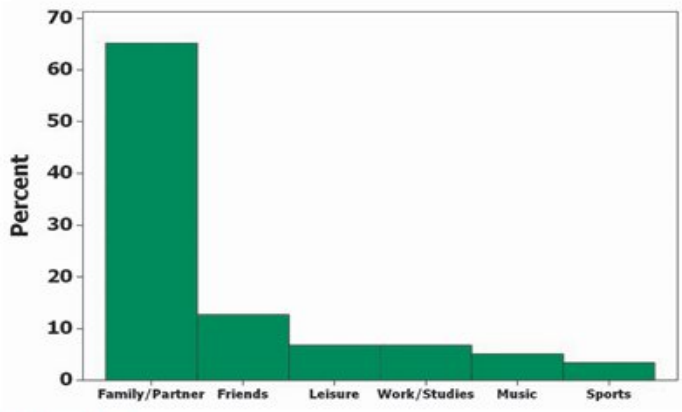


Figure 12 Pareto Chart: What Contributes Most to Happiness?

Using Excel to Create Pareto Charts

To create Pareto Charts in Excel, use the same procedure for creating bar graphs, but first arrange the frequencies in descending order.

3.7. Pie Charts:

A pie chart is a graph that depicts categorical data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category.

Example 9:

Pie Chart: What Contributes Most to Happiness?

Figure 13 is a pie chart corresponding to the same data from Example 8.

Construction of a pie chart involves slicing up the circle into the proper proportions that represent relative frequencies. For example, the category of friends accounts for 13% of the total, so the slice representing friends should be 13% of the total (with a central angle of $0.13 * 360^{o} = 47^{o}$).

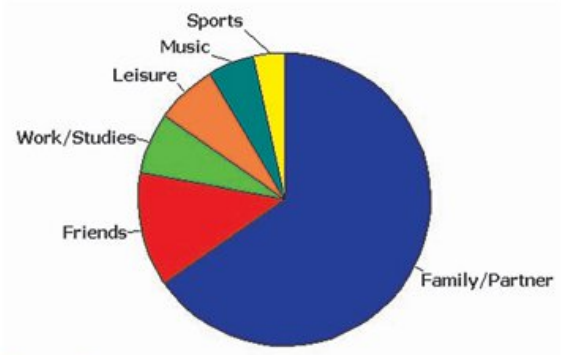


Figure 13 Pie Chart: What Contributes Most to Happiness?

The Pareto chart in Figure 12 and the pie chart in Figure 13 depict the same data in different ways, but the Pareto chart does a better job of showing the relative sizes of the different components. Graphics expert Edwin Tufte makes the following suggestion:

Never use pie charts because they waste ink on components that are not data, and they lack an appropriate scale.

3.8. Using Excel to Create Pie Charts:

The procedure for creating pie charts is very similar to that for creating time-series graphs and other types of charts in Excel.

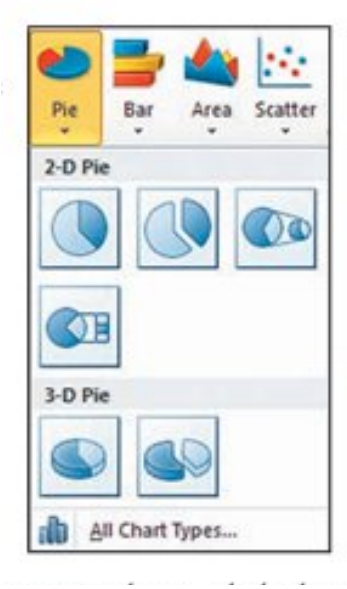
1. Enter the names of categories in column A, and then enter the corresponding frequency or percentage values in column B.

2. Select the cells that contain the data to be used in the chart by clicking on the top cell in column A and, while holding the mouse button down, dragging the cursor to the bottom cell in column B.

3. Click the Insert tab on the Ribbon, and then click Pie in the Charts section of the Ribbon.

4. Select the pie chart type of your preference. The graph will appear with default formatting and can be edited for improved appearance. For more comprehensive

directions on creating and editing pie charts, click the HELP button in Excel and enter “Create a chart” in the search box.



3.9. Frequency Polygon:

A frequency polygon uses line segments connected to points located directly above class midpoint values. A frequency polygon is very similar to a histogram, but a frequency polygon uses line segments instead of bars. We construct a frequency polygon from a frequency distribution as shown in Example 10.

Example 10:

Frequency Polygon: IQ Scores of Low Lead Group

See Figure 14 for the frequency polygon corresponding to the IQ scores of the low lead group summarized in the frequency distribution of Table 2 . The heights of the points correspond to the class frequencies, and the line segments are extended to the right and left so that the graph begins and ends on the horizontal axis. Just as it is easy to construct a histogram from a frequency distribution table, it is also easy to construct a frequency polygon from a frequency distribution table.

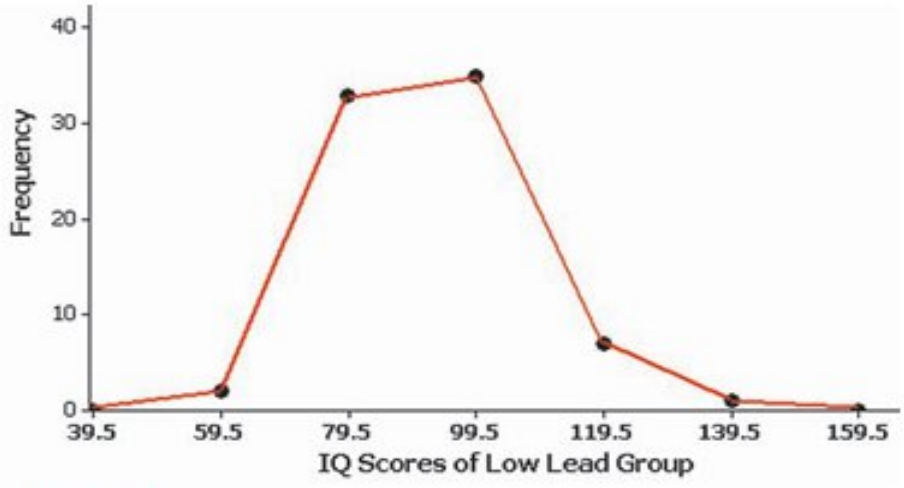


Figure 14 Frequency Polygon: IQ Scores of Low Lead Group

3.10. Using Excel to Create Frequency Polygons:

We'll start with the frequency distribution table below:

Lower Limit	Upper Limit	Frequency
61	70	5
71	80	4
81	90	7
91	100	9

Find the Midpoint

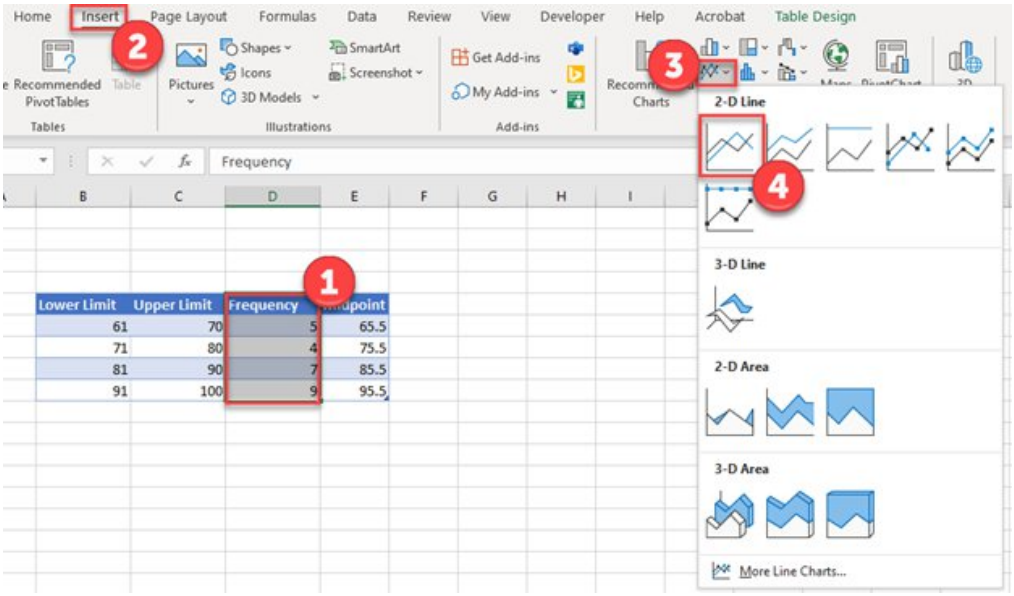
First, find the midpoint of the lower and upper limits with the formula:

=AVERAGE(B6,C6)

	A	B	C	D	E	F
1						
2						
3						
4						
5		Lower Limit	Upper Limit	Frequency	Midpoint	
6		61	70	5	=AVERAGE(B6,C6)	
7		71	80	4	75.5	
8		81	90	7	85.5	
9		91	100	9	95.5	

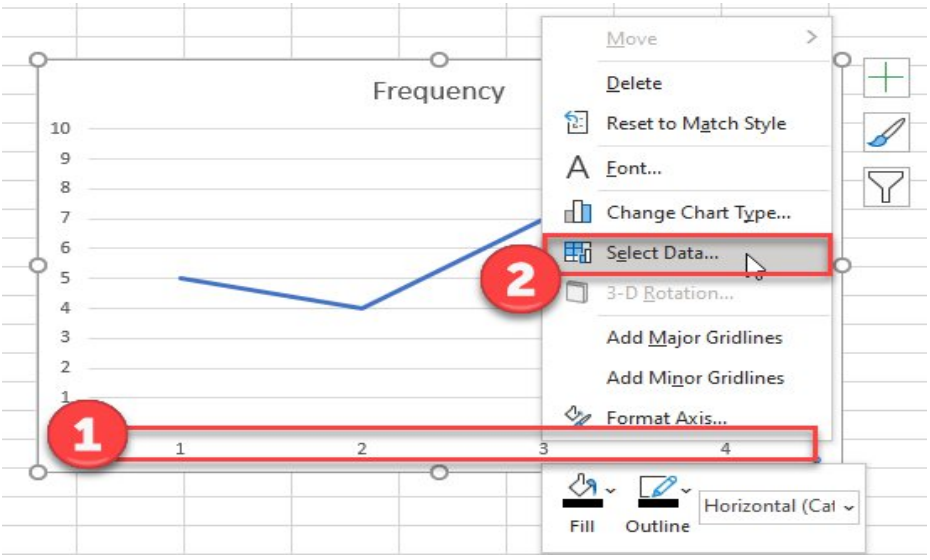
Create the Graph

1. **Select the Frequency Column**
2. **Select Insert**
3. **Click on the Line Graph Icon**
4. **Select the first Line Graph**

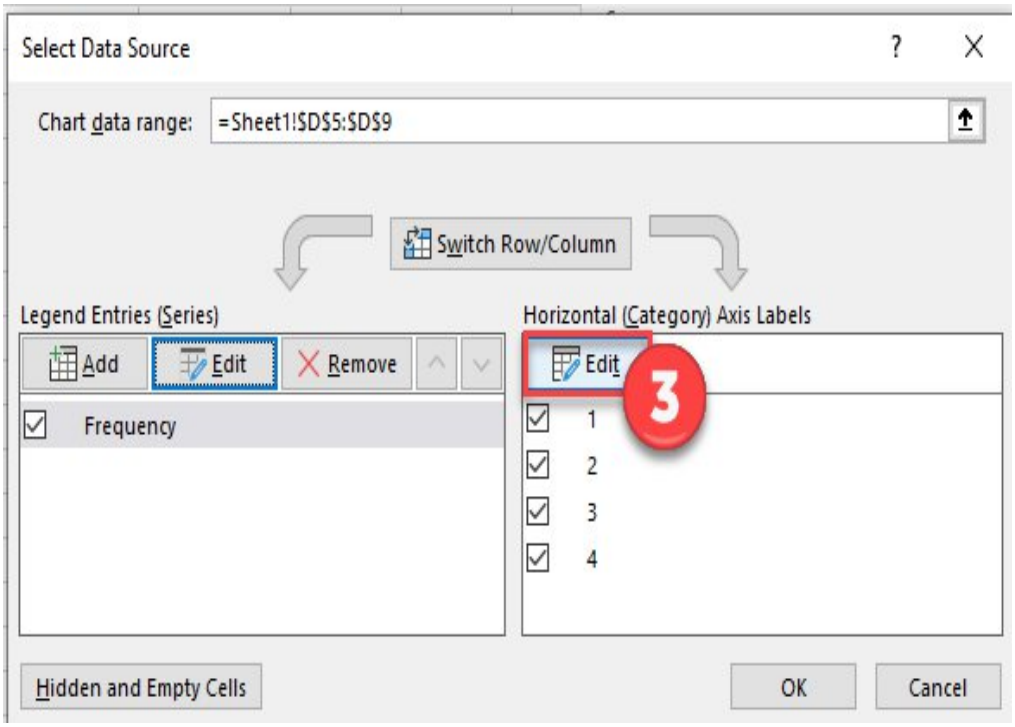


Update X Axis

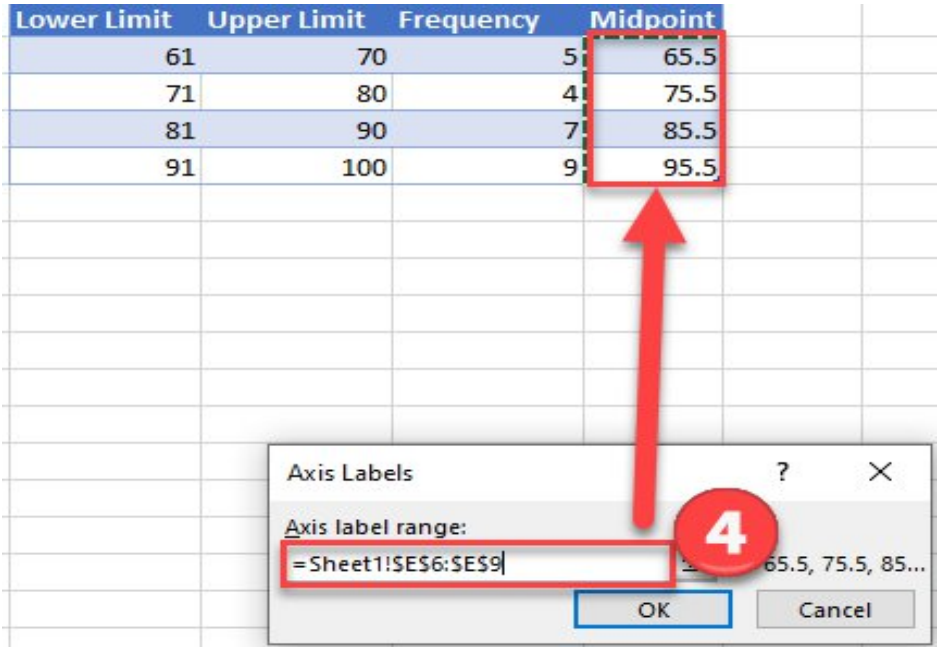
1. **Click on the X Axis**
2. **Click Select Data**



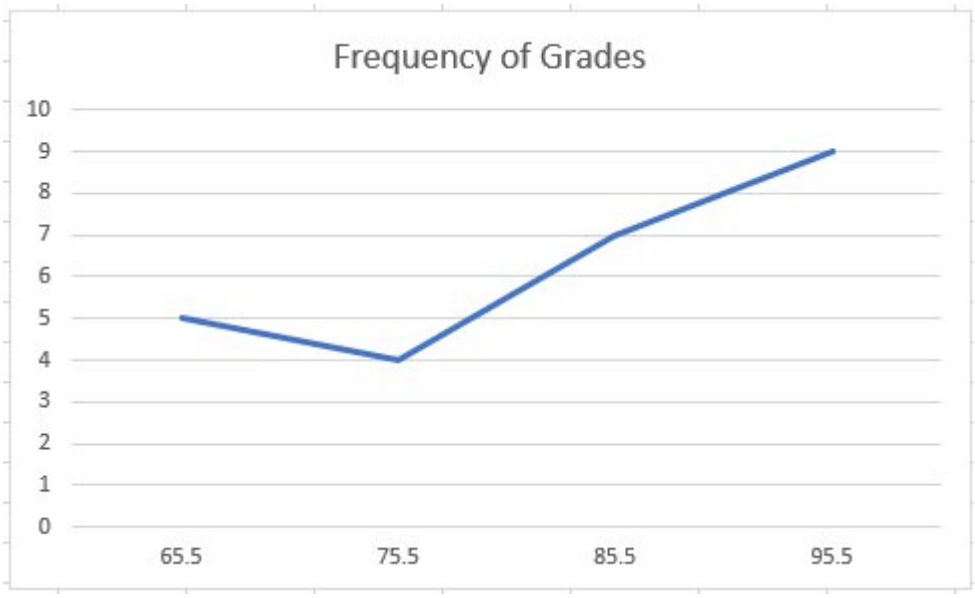
3. Select Edit



4. Highlight Midpoint Data points



5. Click OK and your graph will look like this:



Unit IV

Descriptive statistics-Measures of Center-Mean -Using Excel to Calculate the Mean-Median-Using Excel to Find the Median.

Chapter 4: Sections 4.1 - 4.6

4. Descriptive statistics:

In this chapter we present several formulas used to compute basic statistics. Because technology enables us to compute many of these statistics automatically, it is not as important for us to memorize formulas and manually perform complex calculations. Instead, we should focus on understanding and interpreting the values we obtain from them. The methods and tools presented in Chapter 2 and in this chapter are often called descriptive statistics, because they summarize or describe relevant characteristics of data. Later in this book, we will use inferential statistics to make inferences, or generalizations, about a population.

4.1. Measures of Center:

In this section we discuss the characteristic of center. In particular, we present measures of center, including mean and median, as tools for analyzing data. Our focus here is not only to determine the value of each measure of center, but also to interpret those values.

Definition:

A measure of center is a value at the center or middle of a data set.

There are several different ways to determine the center, so we have different definitions of measures of center, including the mean, median, mode, and midrange. We begin with the mean.

4.2. Mean

The (arithmetic) mean is generally the most important of all numerical measurements used to describe data, and it is what most people call an average.

Definition:

The arithmetic mean, or the mean, of a set of data is the measure of center found by adding the data values and dividing the total by the number of data values. This definition can be expressed as Formula 3-1, in which the Greek letter (uppercase sigma) indicates that the data values should be added. That is, $\sum x$ represents the sum of all data values. The symbol n denotes the sample size, which is the number of data values.

$$\text{mean} = \frac{\sum x}{n} \quad \begin{matrix} \leftarrow sum\ of\ all\ data\ values \\ \leftarrow number\ of\ data\ values \end{matrix}$$

If the data are a sample from a population, the mean is denoted by $\bar{x}$ (pronounced "x-bar"); if the data are the entire population, the mean is denoted by μ (lowercase Greek mu). (Sample statistics are usually represented by English letters, such as m population parameters are usually represented by Greek letters, such as μ.)

Notation

Σ	denotes the *sum* of a set of data values.
x	is the *variable* usually used to represent the individual data values.
n	represents the *number of data values* in a *sample*.
N	represents the *number of data values* in a *population*.
$\bar{x} = \frac{\sum x}{n}$	is the mean of a set of *sample* values.
$\mu = \frac{\sum x}{N}$	is the mean of all values in a *population*.

Example 1:

Mean The Chapter Problem refers to word counts from 186 men and 210 women. Find the mean of these first five word counts from men: 27,531; 15,684; 5,638; 27,997; and 25,433.

Solution:

The mean is computed by using Formula 3-1. First add the data values, then divide by the number of data values

$$\text{mean} = \frac{\sum x}{n} = \frac{27531+156384+5638+27997+25433}{5} = \frac{102283}{5} = 20456.63$$

Since $\bar{x} = 20456.63$ words, the mean of the first five word counts is 204563.6 words.

One advantage of the mean is that it is relatively reliable, so that when samples are selected from the same population, sample means tend to be more consistent than other measures of center. That is, the means of samples drawn from the same population don't vary as much as the other measures of center. Another advantage of the mean is that it takes every data value into account. However, because the mean is sensitive to every value, just one extreme value can affect it dramatically. Since the mean cannot resist substantial changes caused by extreme values, we say that the mean is not a resistant measure of center.

4.4. Using Excel to Calculate the Mean:

What is the Mean in Excel?

The Mean, also known as the Average, is a fundamental statistics measure that represents the central tendency of datasets. The term "mean" refers to the average value of a set of numbers when using Excel. It is a common method of obtaining the central value or concept of a set of statistical data. To calculate mean in Excel, one adds up all the numbers present in a dataset and divides them by the number

of values in that dataset. There are several ways which can be can be used in Excel to find mean and the easiest way is using the AVERAGE function which quickly computes the average value of a range selected.

How to Calculate Mean in Excel

The Arithmetic mean, commonly known as the average is likely a familiar concept to you. This measure is determined by summing a set of numbers and then dividing the total by the Count of those numbers to calculate mean in Excel.

For example, the numbers {1,2,2,3,4,6}. To know how to calculate mean in Excel, you add these numbers together and then divide the sum by 6, resulting in 3: (1+2+2+3+4+6)/6=3.

Step 1: Open MS Excel

Step 2: Select Data

Step 3: Enter AVERAGE Formula

Step 4: Press Enter

In Microsoft Excel, you can calculate the mean using one of the following functions:

AVERAGE: This Function returns the average of a range of numbers.

AVERAGE: This Function provides the average of cells containing various types of data, including numbers, Boolean values, and text.

AVERAGEIF: When you need to find the average based on a single criterion, this function can be used.

AVERAGEIFS: It can be used for calculating the average based on multiple criteria, you can employ this function.

AVERAGE function in Excel

In this method, we are going to use the AVERAGE function which returns the mean of the arguments. For example, the =AVERAGE(A1:A10) returns the average of the numbers in the range of A1 to A10.

Syntax:

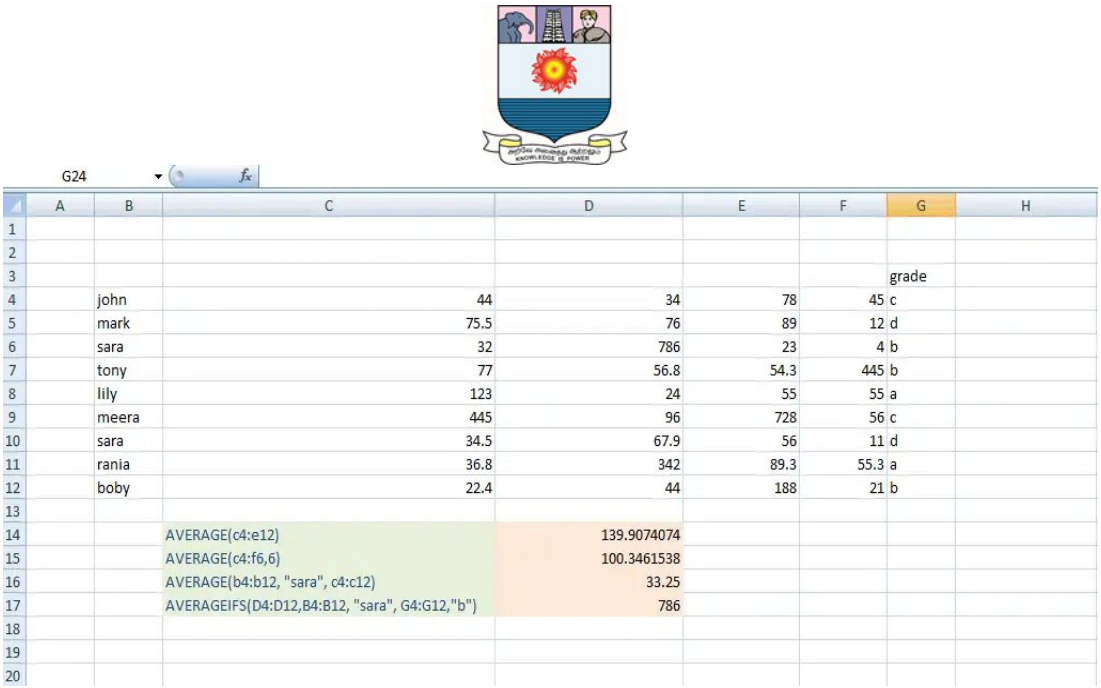
AVERAGE(number1,[number2],...)

Here,

- number1 (Required): The first cell reference or number for which you want the average
- number2 (optional): Additional cell references or numbers for which you want the average. The maximum limit is 255.

Notes:

- Arguments can either be numbers or names, ranges, or cell references that contain numbers.
- If the argument contains text or logical values or empty cells then those values are ignored. However, if the cell contains the value zero then it is included.
- If the argument contains error values or text that cannot be translated to numbers then it will cause errors.
- AVERAGE: This function returns the average of cells with any data (logical values or text representation of numbers)
- AVERAGEIF: This function is used to calculate the average of only the values that meet a single criterion.
- AVERAGEIFS: This function is used to calculate the average based on multiple criteria.



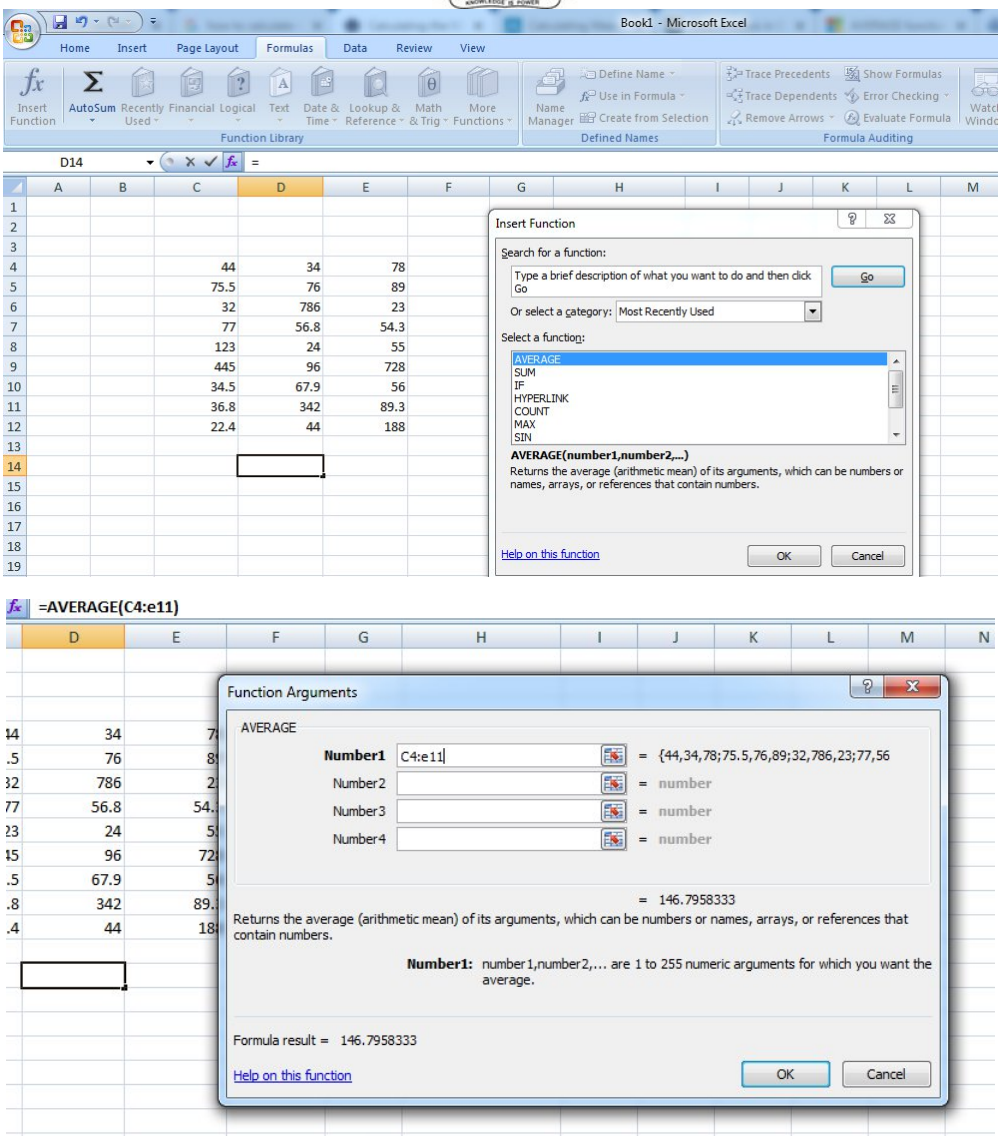
	A	B	C	D	E	F	G	H
1								
2								
3							grade	
4		john	44	34	78	45	c	
5		mark	75.5	76	89	12	d	
6		sara	32	786	23	4	b	
7		tony	77	56.8	54.3	445	b	
8		lily	123	24	55	55	a	
9		meera	445	96	728	56	c	
10		sara	34.5	67.9	56	11	d	
11		rania	36.8	342	89.3	55.3	a	
12		boby	22.4	44	188	21	b	
13								
14			AVERAGE(c4:e12)	139.9074074				
15			AVERAGE(c4:f6,6)	100.3461538				
16			AVERAGE(b4:b12, "sara", c4:c12)	33.25				
17			AVERAGEIFS(D4:D12,B4:B12, "sara", G4:G12,"b")	786				
18								
19								
20								

n the above example, we have used 4 different functions.

- **=AVERAGE(c4:e12)**: Returns the average of the numbers in cells c4 to e12
- **=AVERAGE(c4:f6,6)**: Computes the average of the values in cells C4 to F6 along with the number 6.
- **=AVERAGE(b4:b12, "sara", c4:c12)**: Calculates the average of the numbers in cells C4 to C12 by considering the instances where the name "Sara" appears within the range B4 to B12.
- **=AVERAGE(d4:d12,b4:b12,"sara",g4:g12,"b")**: Determines the average of the numbers in cells D4 to D12, taking into account two conditions: occurrences of the name "Sara" within the range B4 to B12, and occurrences of the grade "B" within the range G4 to G12. .

How to Calculate Mean Using Keyboard shortcuts

- In this method first, you need to select the cells for which you have to calculate the average. Then select the INSERT function from the formulas tab, a dialog box will appear.



4.5. Median:

Unlike the mean, the median is a resistant measure of center, because it does not change by large amounts due to the presence of just a few extreme values.

The median can be thought of loosely as a "middle value" in the sense that about half of the values in a data set are below the median and half are above it. The following definition is more precise.

Definition:

The median of a data set is the measure of center that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by $\tilde{x}$ (pronounced "x-tilde"). x '

To find the median, first sort the values (arrange them in order), then follow one of these two procedures:

1. If the number of data values is odd, the median is the number located in the exact middle of the list.
2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.

Example 2:

Median Find the median for this sample of data values used in Example 1: 27,531, 15,684, 5,638, 27,997, and 25,433.

Solution:

First sort the data values, as shown below: 5,638 15,684 25,433 27,531 27,997 Because the number of data values is an odd number (5), the median is the number located in the exact middle of the sorted list, which is 25,433. The median is there fore 25,433 words. Note that the median of 25,433 is different from the mean of 20,456.6 words found in Example 1.

Example 3:

Median Repeat Example 2 after including the additional data value of 8,077 words. That is, find the median of these word counts: 27,531, 15,684, 5,638, 27,997, 25,433, and 8,077.

Solution:

First arrange the values in order: 5,638 8,077 15,684 25,433 27,531 27,997 Because the number of data values is an even number (6), the median is found by computing the mean of the two middle numbers, which are 15,684 and 25,433.

$$\text{Median} = \frac{15{,}684 + 25{,}433}{2} = \frac{41{,}117}{2} = 20{,}558.5$$

The median is 20,558 words.

4.6. Using Excel to Find the Median:

What is Median in Excel?

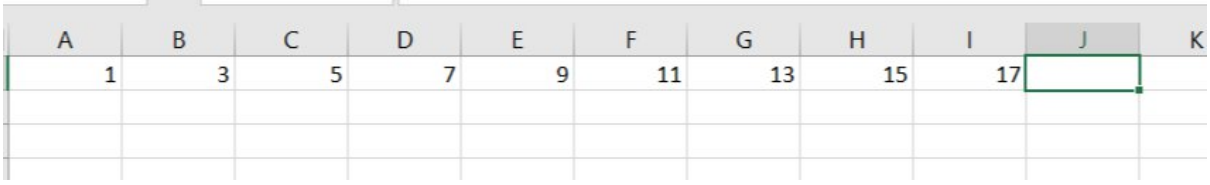
Median helps us to find the middle number in a bunch of numbers. It's the number that sits right in the middle when you put all the numbers in order from smallest to biggest. This middle number divides the group into two halves, with half the numbers being smaller and half being bigger.

In Microsoft Excel, you can find the median using the MEDIAN function. To illustrate, if you want to determine the median of all the sales amounts in our report, apply this formula:

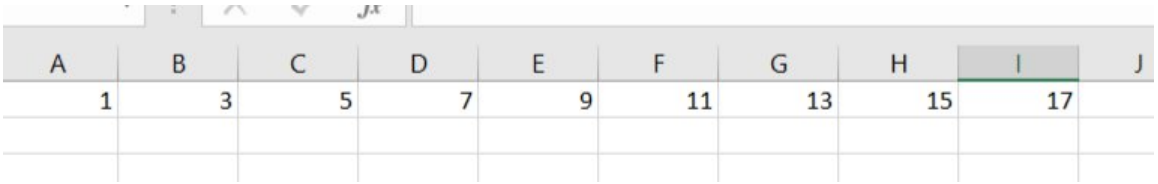
How to Calculate Median in Excel

Follow the below steps to calculate Median in Excel:

Step 1: Enter your data

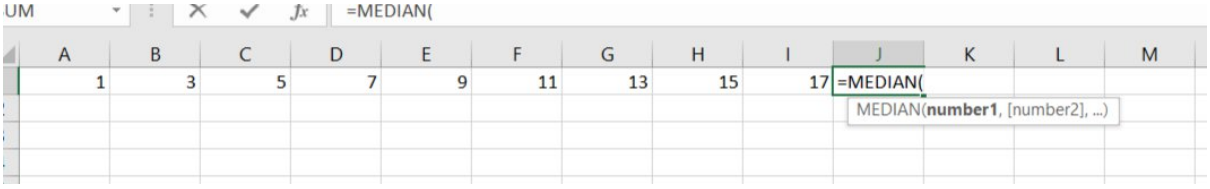


Step 2: Select a Cell



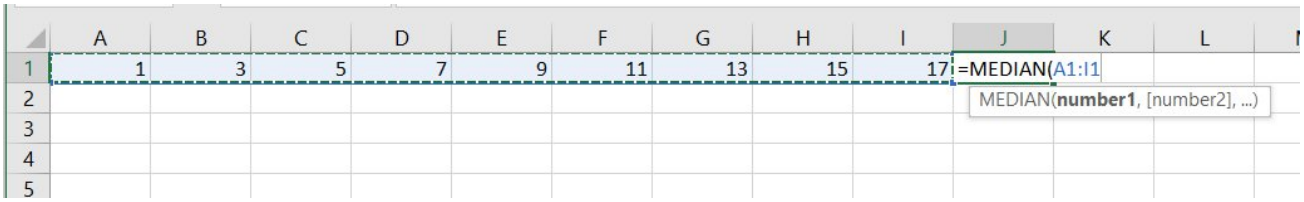
Step 3: Use the MEDIAN Function

In the selected cell, type "=MEDIAN("



Step 4: Select Your Data Range

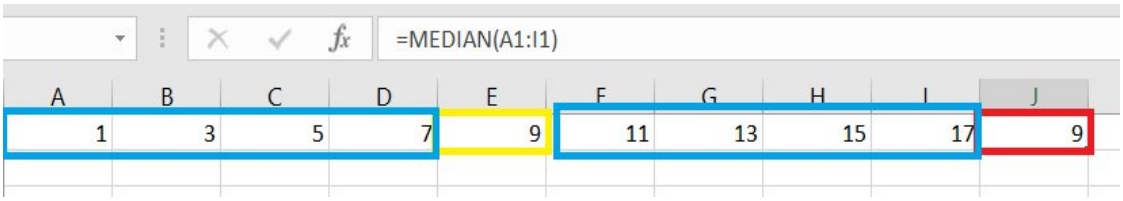
Now, select the range of cells that contain your data. In this example, you would select cells A1 to I1. You can either manually type this range or click and drag to select it.



Step 5: Close the Parenthesis and Press Enter

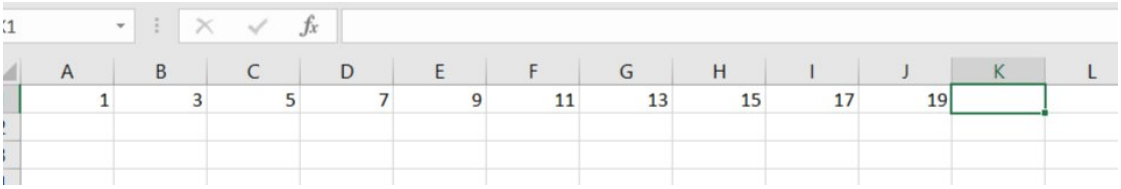
Once you've chosen your data range and closed the parenthesis, press Enter. Excel will then compute the median of the data you selected and show the result in your chosen cell.

Step 6: Preview the Result

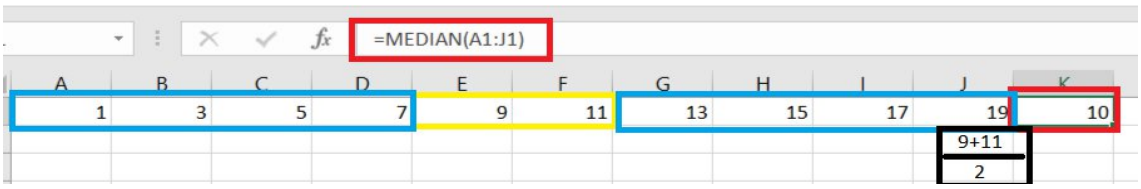


This method works well when dealing with an odd number of values in a dataset. However, what if you have an even number of values? In such a scenario, the median is the average (arithmetic mean) of the two middle values.

For example in the Below example, there are 10 values in the data set.



Repeat all the steps as of above and preview the result in the Below image:



Unit V

Mode-Using Excel to Find the Mode-Midrange-Using Excel to Calculate the Midrange-Weighted Mean-Using Excel for Descriptive Statistics.

Chapter 5: Sections 5.1 - 5.6

5.1. Mode:

The mode is another measure of center.

Definition:

The mode of a data set is the value that occurs with the greatest frequency.

A data set can have one mode, more than one mode, or no mode.

•When two data values occur with the same greatest frequency, each one is a mode and the data set is bimodal.

•When more than two data values occur with the same greatest frequency, each is a mode and the data set is said to be multimodal.

•When no data value is repeated, we say that there is no mode.

Example 1:

Mode Find the mode of these word counts: 18,360 18,360 27,531 15,684 5,638 27,997 25,433.

The mode is 18,360 words, because it is the data value with the greatest frequency.

In Example 4 the mode is a single value. Here are two other possible circumstances:

Two modes: The values of 0, 0, 0, 1, 1, 2, 3, 5, 5, 5 have two modes: 0 and 5.

No mode: The values of 0, 1, 2, 3, 5 have no mode because no value occurs more than once.

In reality, the mode isn't used much with numerical data. However, the mode is the only measure of center that can be used with data at the nominal level of measurement. (Remember, the nominal level of measurement applies to data that

consist of names, labels, or categories only.) Midrange Another measure of center is the midrange. Because the midrange uses only the maximum and minimum values, it is too sensitive to those extremes, so the midrange is rarely used. However, the midrange does have three redeeming features: (1) it is very easy to compute; (2) it helps to reinforce the important point that there are several different ways to define the center of a data set; (3) it is sometimes incorrectly used for the median, so confusion can be reduced by clearly defining the midrange along with the median.

5.2. Using Excel to Find the Mode:

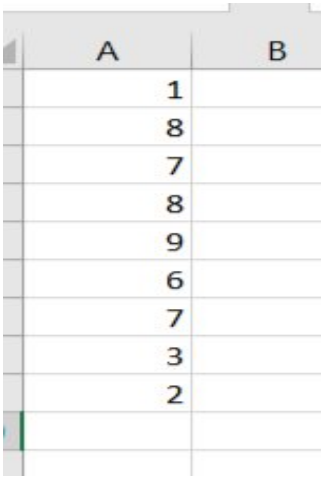
What is Mode in Excel?

The mode represents the number that appears most frequently within a given set of numbers. In simpler terms, it's the number that occurs the most times in the set is known as mode in Excel. Here are the steps to calculate mode in Excel.

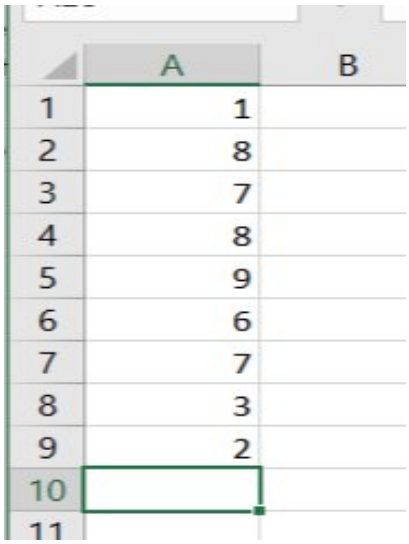
How to Find Mode in Excel

Among all the statistical measures, finding the mode is the simplest and requires the least mathematical computation. Essentially, you identify the mode by locating the score that appears most frequently in a dataset. Here are some steps on how to calculate mode in Excel.

Step 1: Enter the Data Set



Step 2: Select the Cell where you want the result to be displayed

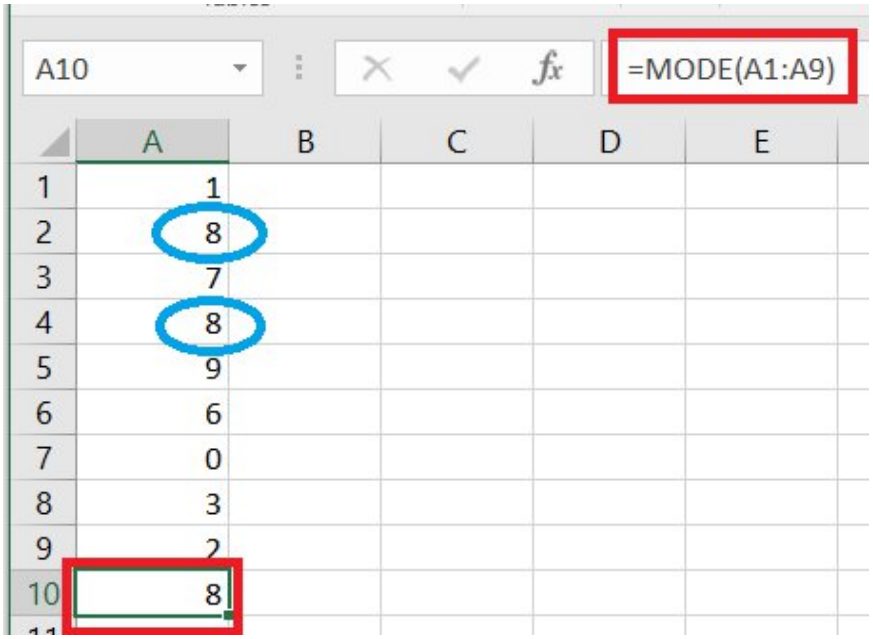


	A	B
1	1	
2	8	
3	7	
4	8	
5	9	
6	6	
7	7	
8	3	
9	2	
10		

Step 3: Enter the Formula and Press Enter

Use the Formula "=MODE (A1:A9)".

Step 4: Preview Result



5.3. Midrange:

Definition:

The midrange of a data set is the measure of center that is the value mid way between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2, as in the following formula:

$$\text{midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$

Example

Midrange Find the midrange of these values from Example 1: 27,531, 15,684, 5,638, 27,997, and 25,433.

Solution:

The midrange is found as follows:

$$\text{midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$

$$= \frac{27{,}997 + 5{,}638}{2} = 16{,}817.5$$

The midrange is 16,817.5 words.

The term average is often used for the mean, but it is sometimes used for other measures of center. To avoid any confusion or ambiguity we use the correct and specific term, such as mean or median. The term average is not used by statisticians and it will not be used throughout the remainder of this book when referring to a specific measure of center. When calculating measures of center, we often need to round the result. We use the following rule.

Because the midrange uses only the maximum and minimum values, it is too sensitive to those extremes, so the midrange is rarely used. However, the midrange does have three redeeming features: (1) it is very easy to compute; (2) it helps to reinforce the important point that there are several different ways to define the center of a data set; (3) it is sometimes incorrectly used for the median,

so confusion can be reduced by clearly defining the midrange along with the median.

5.4. Using Excel to Calculate the Midrange:

MS Excel is a spreadsheet developed by the company Microsoft. Excel provides various kinds of functions and we can insert the data in form of rows and columns and perform operations on the data and yield the results we desired.

The **Midrange** of the dataset in other terms can be specified as average or mean. the Midrange is also known as the **measure of center in statistics.**

Midrange of any dataset can be calculated as follows: (largest value + smallest value) / 2

In order to calculate using formulas of excel we follow 3 steps:

- Calculate Min value
- Calculate Max value
- Apply midrange formula

Consider the following data i.e. marks obtained by students of a class

When applying this rule, round only the final answer, not intermediate values that occur during calculations. For example, the mean of 2, 3, 5, is 3.333333..., which is rounded to 3.3, which has one more decimal place than the original values of 2, 3, 5. As another example, the mean of 80.4 and 80.6 is 80.50 (one more decimal place than was used for the original values). Because the mode is one or more of the original data values, we do not round values of the mode; we simply use the same original values

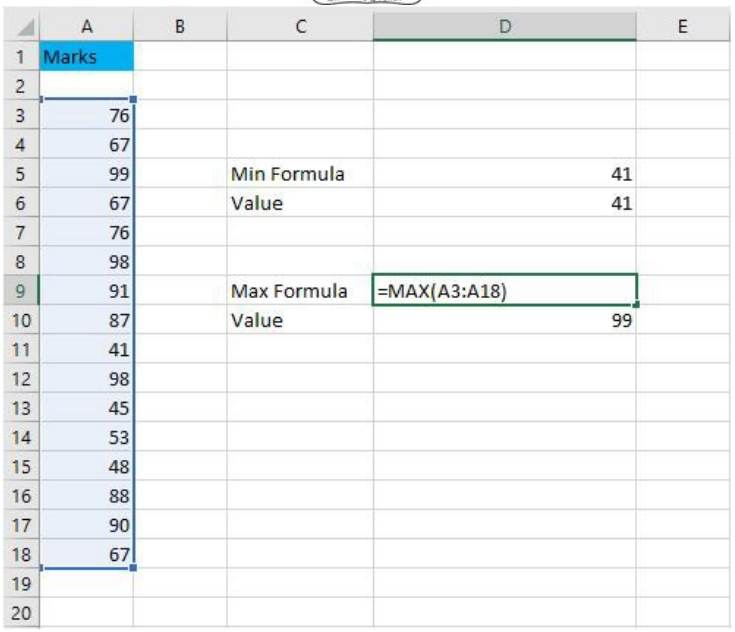
	A	B
1	Marks	
2		
3	76	
4	67	
5	99	
6	67	
7	76	
8	98	
9	91	
10	87	
11	41	
12	98	
13	45	
14	53	
15	48	
16	88	
17	90	
18	67	

To find the midrange of the given data we will follow the 3 steps

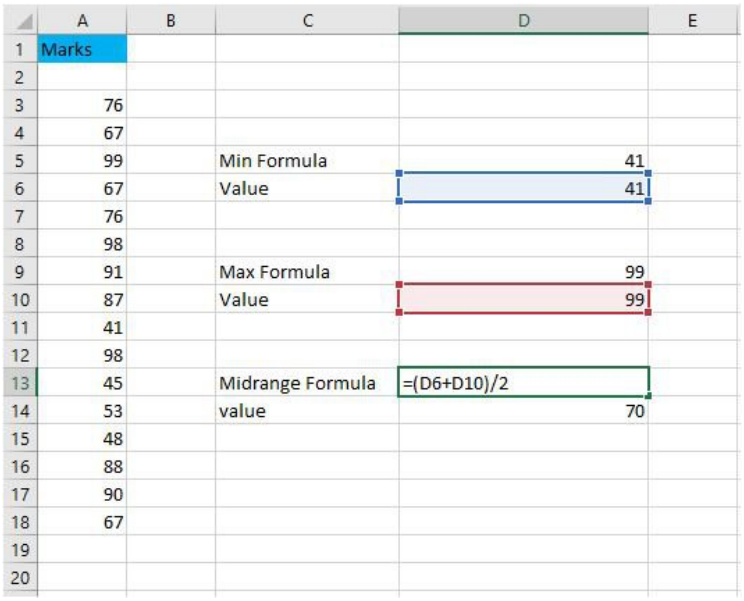
Step 1: Calculate the min value of the data by using the min function

	A	B	C	D	E
1	Marks				
2					
3	76				
4	67				
5	99		Min Formula	=MIN(A3:A18)	
6	67		Value	41	
7	76				
8	98				
9	91				
10	87				
11	41				
12	98				
13	45				
14	53				
15	48				
16	88				
17	90				
18	67				
19					
20					

Step 2: Calculate the max value of the data by using the max function



	A	B	C	D	E
1	Marks				
2					
3	76				
4	67				
5	99		Min Formula	41	
6	67		Value	41	
7	76				
8	98				
9	91		Max Formula	=MAX(A3:A18)	
10	87		Value	99	
11	41				
12	98				
13	45				
14	53				
15	48				
16	88				
17	90				
18	67				
19					
20					



	A	B	C	D	E
1	Marks				
2					
3	76				
4	67				
5	99		Min Formula	41	
6	67		Value	41	
7	76				
8	98				
9	91		Max Formula	99	
10	87		Value	99	
11	41				
12	98				
13	45		Midrange Formula	=(D6+D10)/2	
14	53		value	70	
15	48				
16	88				
17	90				
18	67				
19					
20					

Step 3: Now apply the midrange formula

Finally, we get the midrange of the data as 70.

Drawbacks:

- The midrange deviates hugely when there are outliers in our data
- We need to find first the maximum and minimum value to the data in order to find the midrange

- The midrange sometimes differs from mean and median hugely.

For example in our data, consider another field with the value of 1000 then the midrange will be 520.5 which is far away from the original midrange of 70. So this shows how large the midrange deviates from the outliers.

	A	B	C	D	E
1	Marks				
2					
3	76				
4	67				
5	99		Min Formula	"=MIN(A3:A19)"	
6	67		Value	41	
7	76				
8	98				
9	91		Max Formula	"=MAX(A3:A19)"	
10	87		Value	1000	
11	41				
12	98				
13	45		Midrange Formula	=(D6+D10)/2	
14	53		value	520.5	
15	48				
16	88				
17	90				
18	67				
19	1000				

Alternatives:

Instead of finding the midrange, we can directly calculate the mean, median of the data which points to the average or the midpoint of the data

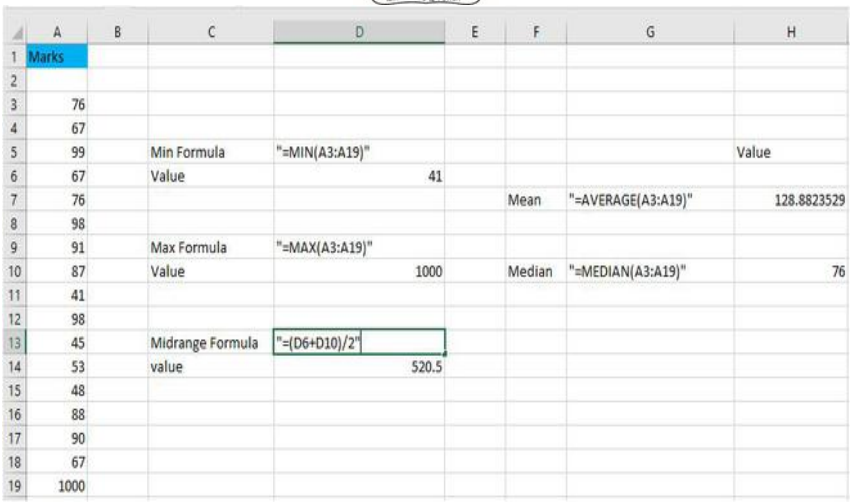
Mean: It is the average value of the data set. It is calculated by the formula

Mean = Sum of all the observations/Total number of Observations

In excel we can directly compute the mean by using the average function.

Median: It is the middle value of the dataset. We need to first arrange the dataset in either non-increasing or non-decreasing order and select the middlemost element. In case there are two middlemost elements we compute the average of two.

In excel we can compute it by using the median function



	A	B	C	D	E	F	G	H
1	Marks							
2								
3	76							
4	67							
5	99		Min Formula	"=MIN(A3:A19)"				Value
6	67		Value	41				
7	76					Mean	"=AVERAGE(A3:A19)"	128.8823529
8	98							
9	91		Max Formula	"=MAX(A3:A19)"				
10	87		Value	1000		Median	"=MEDIAN(A3:A19)"	76
11	41							
12	98							
13	45		Midrange Formula	"=(D6+D10)/2"				
14	53		value	520.5				
15	48							
16	88							
17	90							
18	67							
19	1000							

This image gives the mean, median and midrange of the dataset.

5.5. Weighted Mean:

Weighted Mean When data values are assigned different weights, we can compute a weighted mean. This formula can be used to compute the weighted mean, w.

$$\text{weighted mean: } \bar{x} = \frac{\sum (w.x)}{\sum w}$$

Above formula tells us to first multiply each weight w by the corresponding value x, then to add the products, and then finally to divide that total by the sum of the weights w.

Computing Grade Point Average In her first semester of college, a student of the author took five courses. Her final grades along with the number of credits for each course were: A (3 credits); A (4 credits); B (3 credits), C (3 cred its), and F (1 credit). The grading system assigns quality points to letter grades as follows: A = 4 ; B = 3 ; C = 2 ; D = 1 ; F = 0 . Compute her grade point average

Use the numbers of credits as weights: w = 3 , 4, 3, 3, 1. Replace the letter grades of A, A, B, C, and F with the corresponding quality points: x = 4 4, 3, 2, 0. We now use Formula 3-3 as shown below. The result is a first-semester grade point

average of 3.07. (Using the preceding round-off rule, the result should be rounded to 3.1, but it is common to round grade point averages with two decimal places

$$\bar{x} = \frac{\sum (w.x)}{\sum w}$$

$$= \frac{(3\times4)+(4\times4)+(3\times2)+(1\times0)}{3+4+3+3+1}$$

$$= \frac{43}{14} = 3.07$$

5.6. Using Excel for Descriptive Statistics:

What is Descriptive Statistics in Excel?

Descriptive statistics is all about describing the given data. To describe the data, we use measures of central tendency and measures of dispersion. In this article, we explain how to use "Data Analysis" in Excel for descriptive statistics in detail.

Measures of Central Tendency *[Mean, Median, and Mode*]

A single number about the centre of the data points. Where the majority of the values in the dataset are found is indicated by a measure of central tendency. It sits in the middle of the range of values. Three central tendency measures are provided by Excel.

- **Mean:** It is calculated by dividing the total number of observations by their number. It works best with symmetrically distributed data.
- **Median:** Your data is divided in half by this value. Half of the results are higher than the median, and the other half are lower.
- **Mode:** The value that appears the most frequently in your data is represented by this metric. For categorical and ordinal data, it works best.

What are Measures of Dispersion in Excel [*Range, Variance and Standard Deviation*]

Dispersion measurements show how tightly or loosely distributed the data points are around the centre. Three dispersion measures are provided by Excel. Generally, data points move away from the centre as their values rise.

- **Range:** It is the discrepancy between a dataset's largest and smallest values. The range is simple to comprehend, but it is highly prone to outliers because it is based simply on the two most extreme values in the dataset.
- **Variance:** It depicts the values' average squared deviation from the mean. The variance is expressed in squared units rather than the original data units because the calculations use squared differences.
- **Standard Deviation:** The normal or average deviation of each data point from the mean. This measure simplifies understanding because it utilizes the original units of the data. The variance's square root gives the standard deviation.

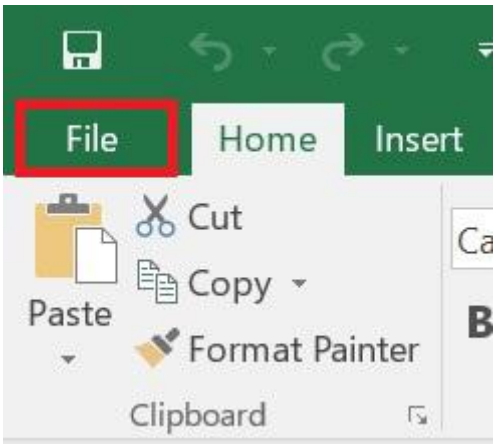
You must have the Data Analysis Toolpak enabled in order to access the descriptive statistics in Excel.

How to Enable Data Analysis Toolpak

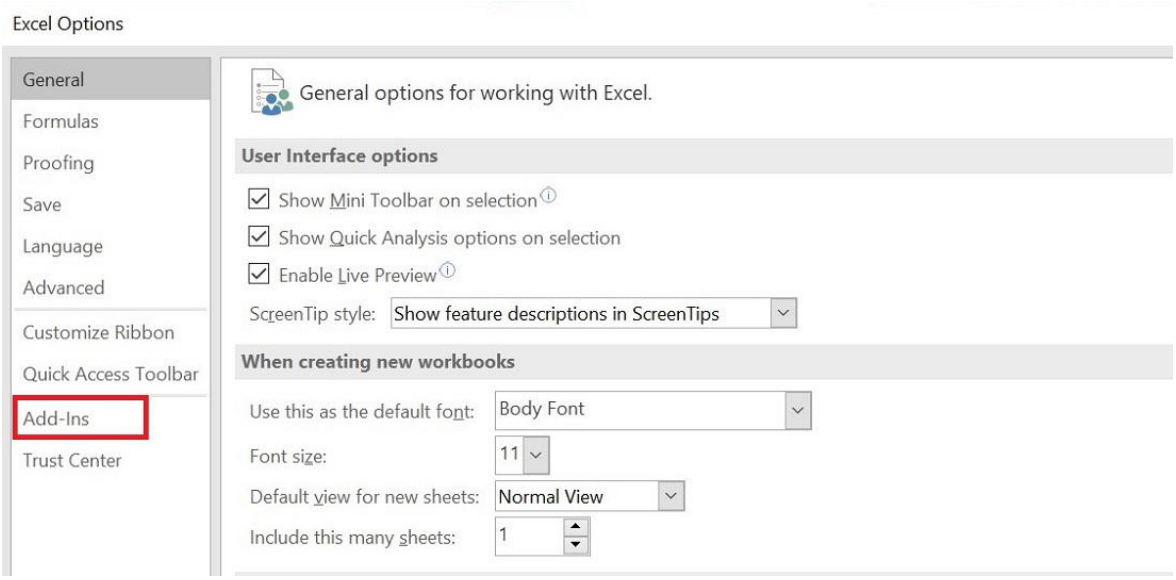
The procedures to enable the Data Analysis Toolpak in Excel are listed below:

Step 1. Launch any Excel file, Go to the File tab

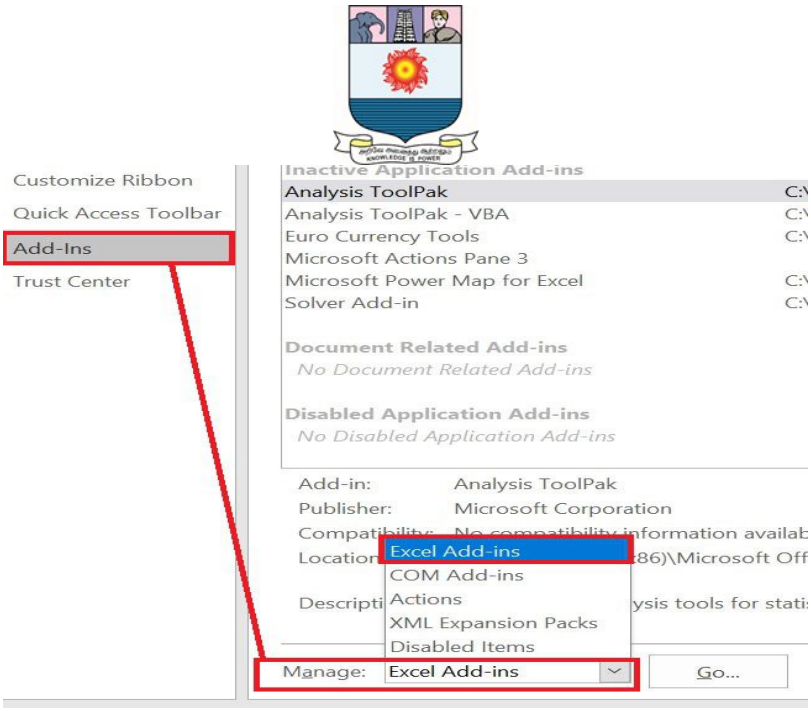
Step 2. Select Options, the Excel Options dialogue box will appear as a result



Step 3. In the left pane of the Excel Options dialogue box, select Add-ins ft pane of the Excel Options dialogue box, select Add-ins

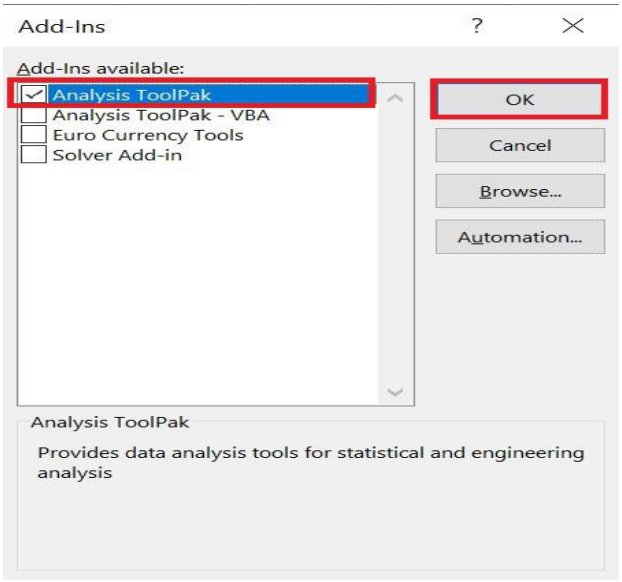


Step 4. Choose "Excel Add-ins" from the Manage drop-down menu



Step 5. Press the "Go" button

Step 6. Check the Analysis Toolpak box when the Add-ins dialogue box appears and Click OK



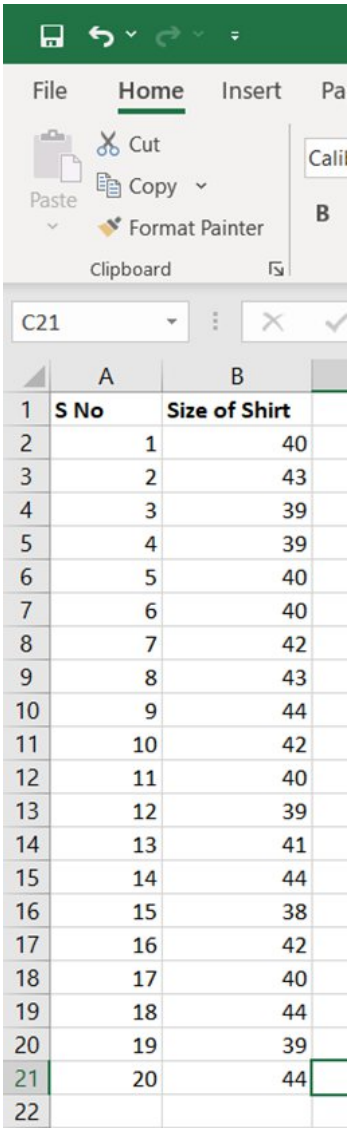
Well Done! You've now enabled Data Analysis Toolpak and is ready to use.

How to Get Descriptive Statistics in Excel?

Let's look at how to obtain the descriptive statistics using the Data Analysis Toolpak now that it is enabled.

Example: We have given shirt size of 20 men in the below table.

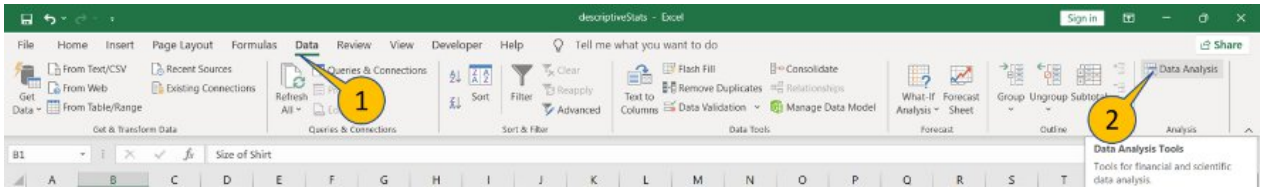
Sample data



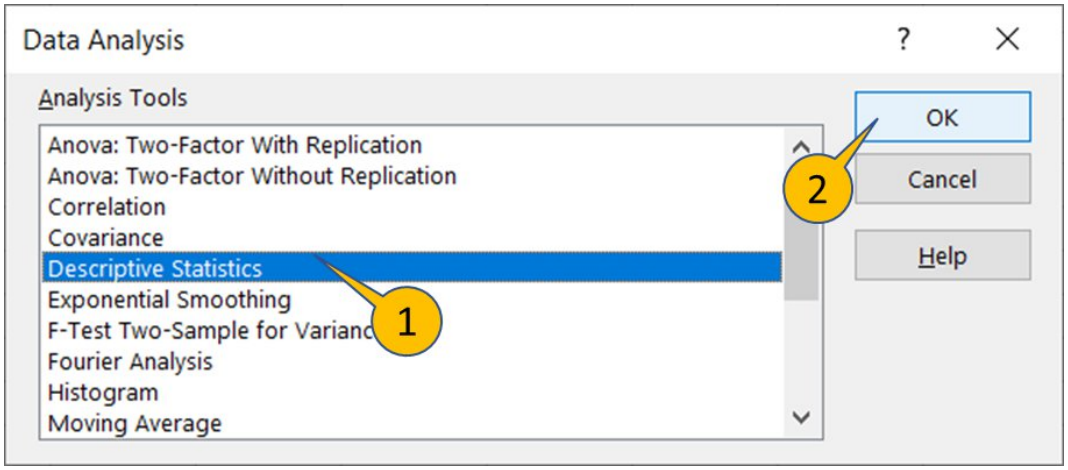
	A	B
1	S No	Size of Shirt
2	1	40
3	2	43
4	3	39
5	4	39
6	5	40
7	6	40
8	7	42
9	8	43
10	9	44
11	10	42
12	11	40
13	12	39
14	13	41
15	14	44
16	15	38
17	16	42
18	17	40
19	18	44
20	19	39
21	20	44
22		

Follow the below steps to implement descriptive statistics on sample data:

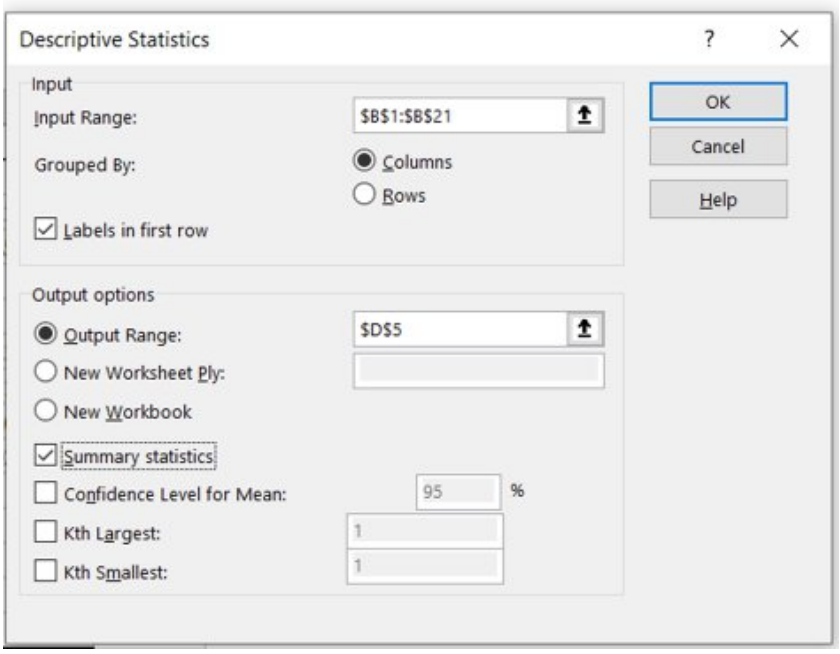
Step 1. Go to “Data” >> Click “Data Analysis” (Image 1) – to pop up the “Data Analysis” Dialogue box. If you cannot find “Data analysis” in Excel ribbon. The end of this article finds the steps [To provide “Data Analysis”]



Step 2. In Data Analysis, Select "Descriptive Statistics" and Press "OK" – To pop-up "Descriptive statistics" Dialogue box for further input

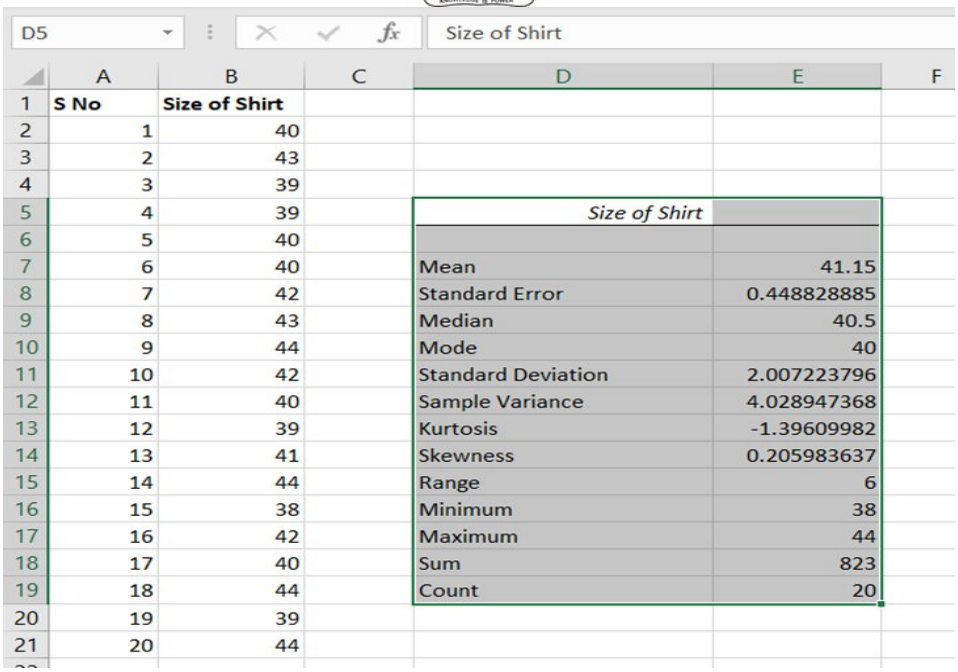


Step 3. Make sure the below options are selected in "Descriptive statistics" and Press "OK"



Output

Descriptive statistics Output in Table "D5:E19"



S No	Size of Shirt
1	40
2	43
3	39
4	39
5	40
6	40
7	42
8	43
9	44
10	42
11	40
12	39
13	41
14	44
15	38
16	42
17	40
18	44
19	39
20	44

Size of Shirt	
Mean	41.15
Standard Error	0.448828885
Median	40.5
Mode	40
Standard Deviation	2.007223796
Sample Variance	4.028947368
Kurtosis	-1.39609982
Skewness	0.205983637
Range	6
Minimum	38
Maximum	44
Sum	823
Count	20

Study Learning Material Prepared by

Dr. S. KALAISELVI M.SC., M.Phil., B.Ed., Ph.D.,

ASSISTANT PROFESSOR,

DEPARTMENT OF MATHEMATICS,

SARAH TUCKER COLLEGE (AUTONOMOUS),

TIRUNELVELI-627007.

TAMIL NADU, INDIA.