



CHAPTER 2: SAMPLING AND DATA

This presentation is based on material and graphs from Open Stax and is copyrighted by Open Stax and Georgia Highlands College.

OUTLINE

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

2.2 Histograms, Frequency Polygons, and Time Series Graphs

2.3 Measures of Locations

2.4 Box Plots

2.5 Measures of Center of the Data

IMPORTANT CHARACTERISTICS OF DATA

- Center: a representative or average value that indicates where the middle of the data set is located
- Variation: a measure of the amount that the values vary among themselves
- Distribution: the nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)
- Outliers: Sample values that lie very far away from the majority of other sample values
- Time: Changing characteristics of data over time

Computer Viruses Destroy Or Terminate

SECTION 2.1

STEM-AND-LEAF GRAPHS (STEMPLOTS), LINE GRAPHS, AND BAR GRAPHS

STEM-AND-LEAF GRAPH OR STEM PLOT

One simple graph, the stem-and-leaf graph or stem plot, comes from the field of exploratory data analysis.

It is a good choice when the data sets are small.

To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit.

PRACTICE

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47;
47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53;
54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

SIDE BY SIDE STEM-AND-LEAF GRAPH

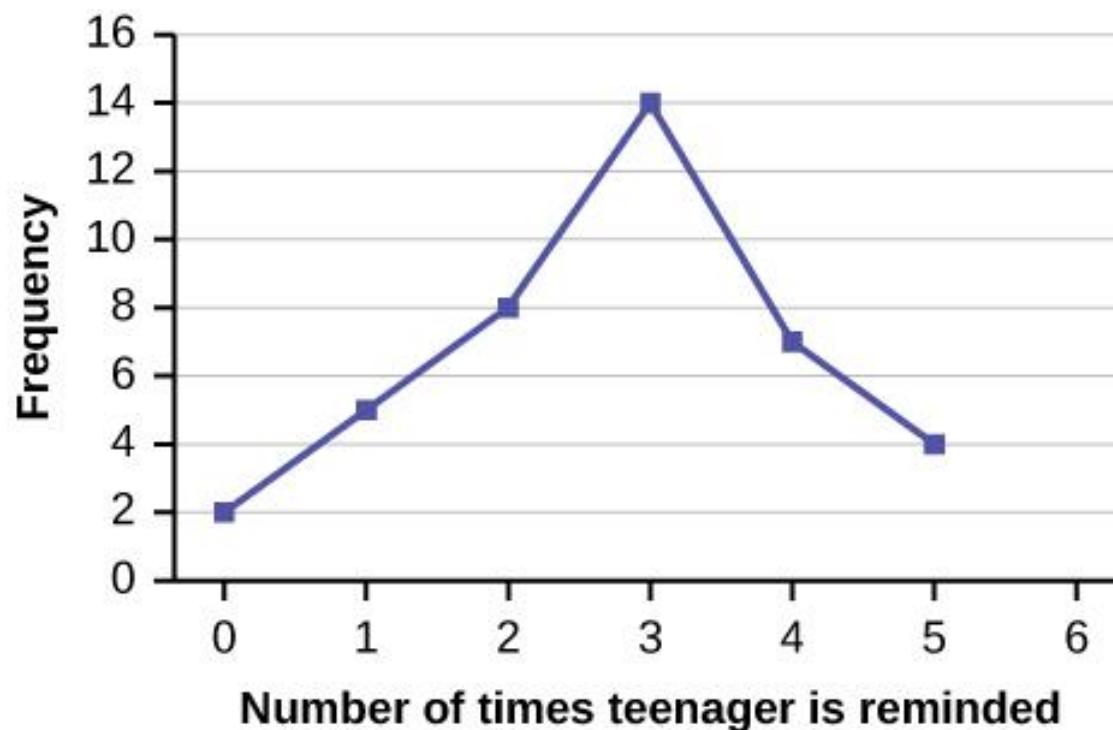
A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns.

In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem.

The leaves are to the left and the right of the stems.

Construct a side by-side stem-and-leaf plot using this data.

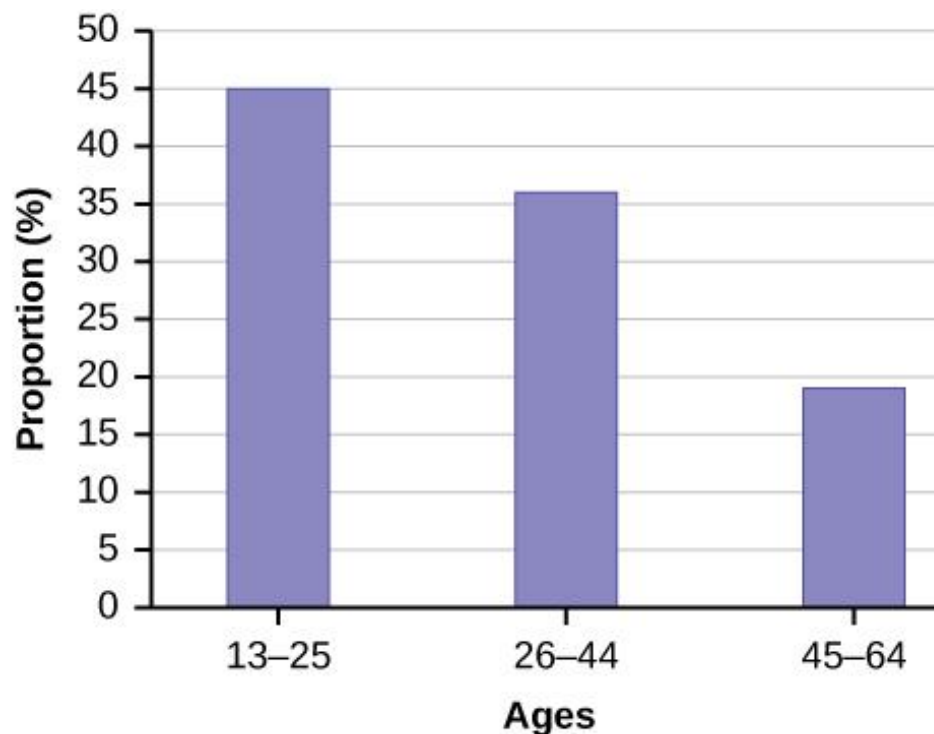
LINE GRAPHS



Another type of graph that is useful for specific data values is a **line graph**.

In the particular line graph shown above, the **x-axis** (horizontal axis) consists of **data values** and the **y-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

BAR GRAPHS



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The bar graph shown above has age groups represented on the x-axis and proportions on the y-axis.

SECTION 2.2

HISTOGRAMS, FREQUENCY POLYGONS, AND TIME SERIES GRAPHS

HISTOGRAM

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis.

The horizontal axis is labeled with what the data represents (for instance, distance from your home to school).

The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability).

The graph will have the same shape with either label.

The histogram (like the stemplot) can give you

- the shape of the data
- the center
- the spread of the data

RELATIVE FREQUENCY

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.

(Remember, frequency is defined as the number of times an answer occurs.)

- f = frequency
- n = total number of data values (or the sum of the individual frequencies)
- RF = relative frequency, then: $RF = \frac{f}{n}$

For example, if three students in Mr. Ahab's English class of 40 students

received from 90% to 100%, then, $f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$.

7.5% of the students received 90–100%. 90–100% are quantitative measures.

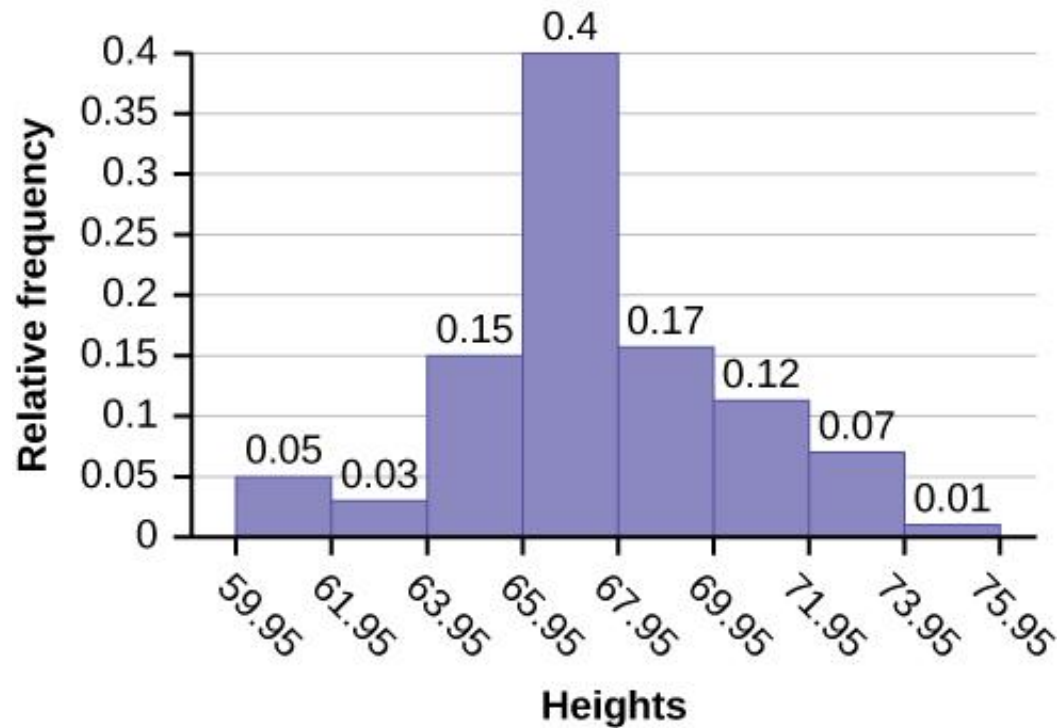
CONSTRUCT A HISTOGRAM

- To construct a histogram, first decide how many bars or intervals, also called classes, represent the data.
- Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen.
- Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places.

CONSTRUCT A HISTOGRAM

- For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 ($6.1 - 0.05 = 6.05$). We say that 6.05 has more precision.
- If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 ($1.5 - 0.005 = 1.495$).
- If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 ($1.0 - 0.0005 = 0.9995$).
- If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 ($2 - 0.5 = 1.5$).
- Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

EXAMPLE OF HISTOGRAM



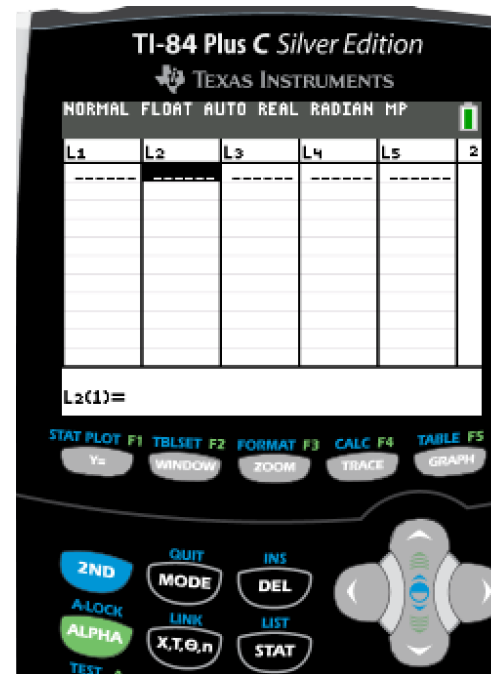
ENTERING DATA AND MAKING HISTOGRAMS USING THE TI-83/84 CALCULATOR

ONE VARIABLE DATA ENTRY

Go to



Choose Option 1: Edit

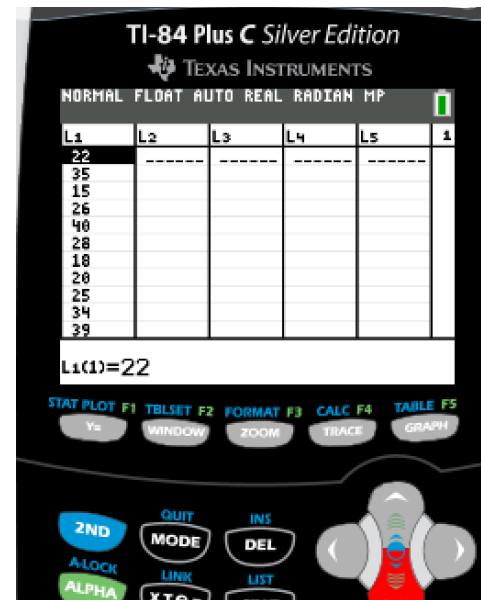


ENTER DATA INTO COLUMN (AKA “LIST”)

EXAMPLE: Number of Employees in various New York City restaurants

Type in each individual data value, enter, repeat until all data values are entered.

22	35	15	26
	40	28	
18	20	25	34
	39	42	
24	22	19	27
	22	34	
40	20	38	28

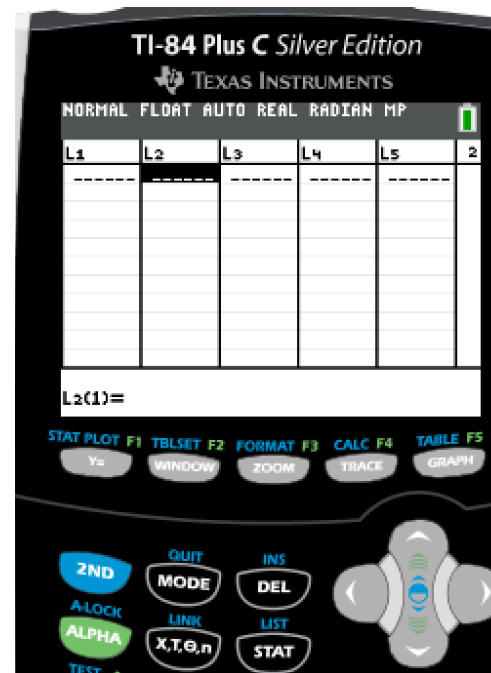


TWO VARIABLE DATA ENTRY

Go to



Choose Option 1: Edit

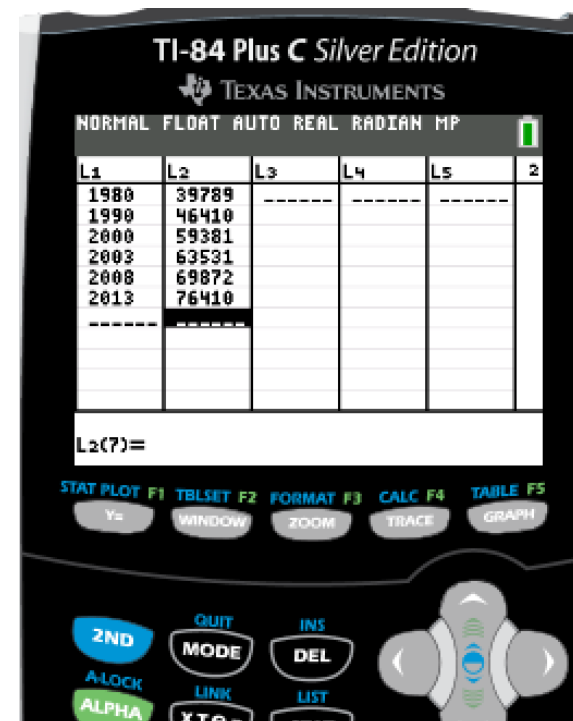


TWO VARIABLE DATA ENTRY

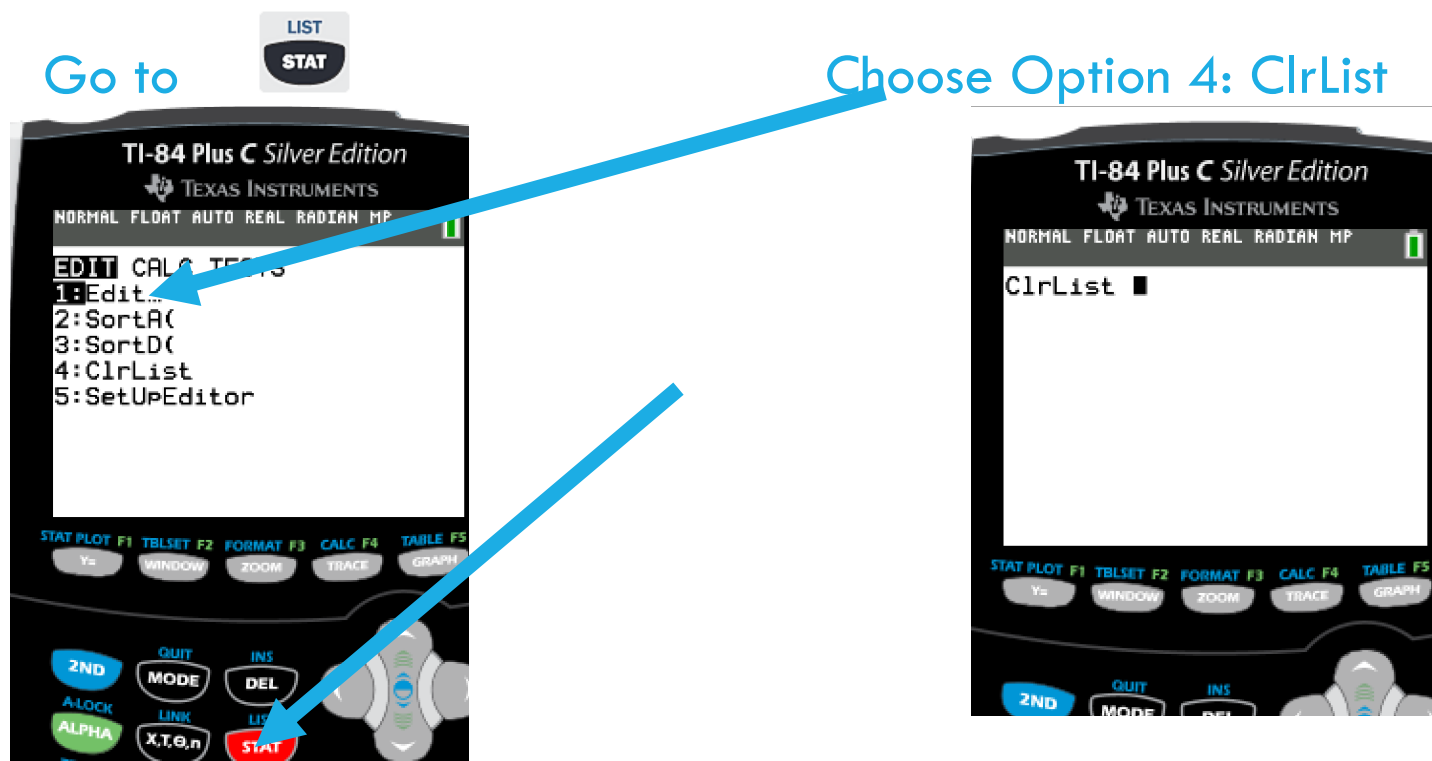
Enter Variable 1 (x) into first column (L1) and enter Variable 2 (y) into second column (L2)

EXAMPLE:

Year	Population within 20 minutes of Wal-Mart
1980	39789
1990	46410
2000	59381
2003	63531
2008	69872
2013	76410



CLEAR DATA FROM LISTS



Identify which list to clear
Click
Press



To identify multiple list at one time, separate list names by commas



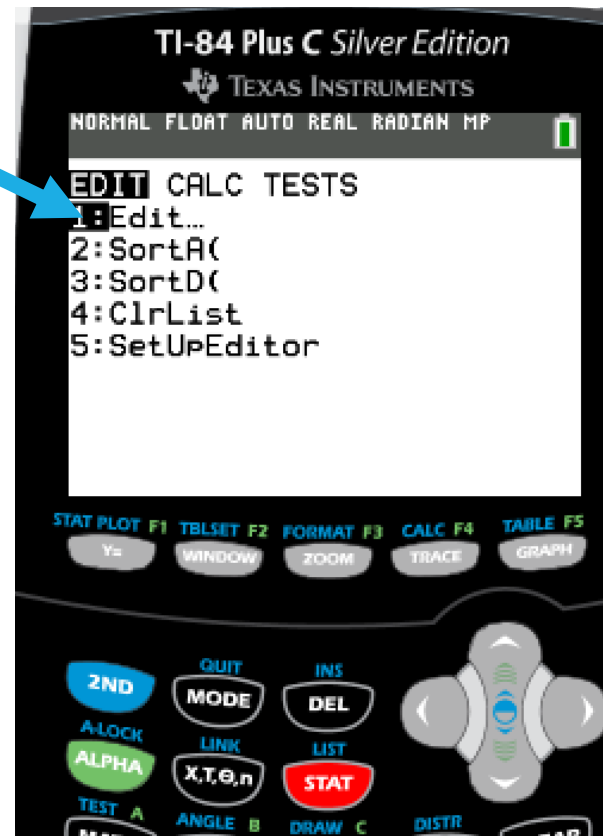
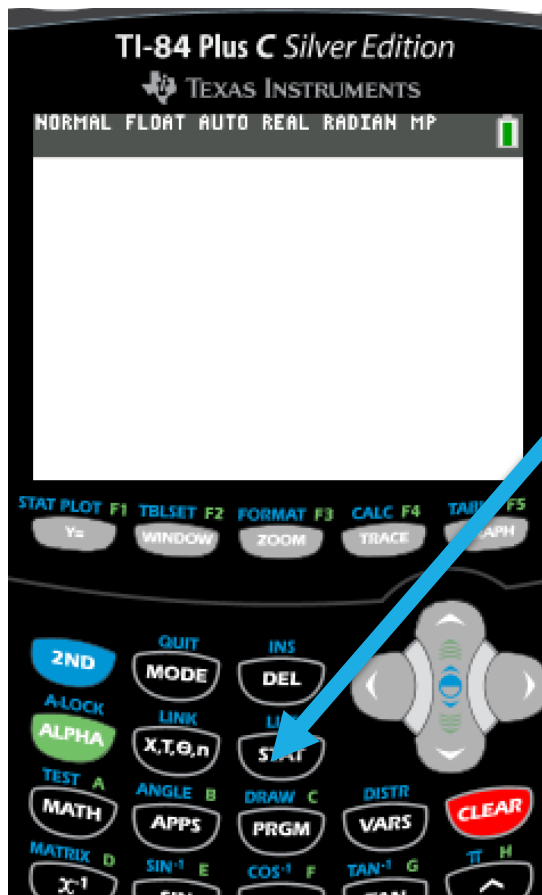
TO CREATE A HISTOGRAM USING A TI-83 PLUS OR EQUIVALENT CALCULATOR, YOU MUST CREATE A GROUPED FREQUENCY DISTRIBUTION FIRST.

EXAMPLE: Number of employees
at various Atlanta restaurants

22	35 40	15 28	26
18	20 39	25 42	34
24	22 22	19 34	27
40	20	38	28

Number of Employees	Class Midpoint	Frequency
10-19	14.5	3
20-29	24.5	11
30-39	34.5	5
40-49	44.5	3

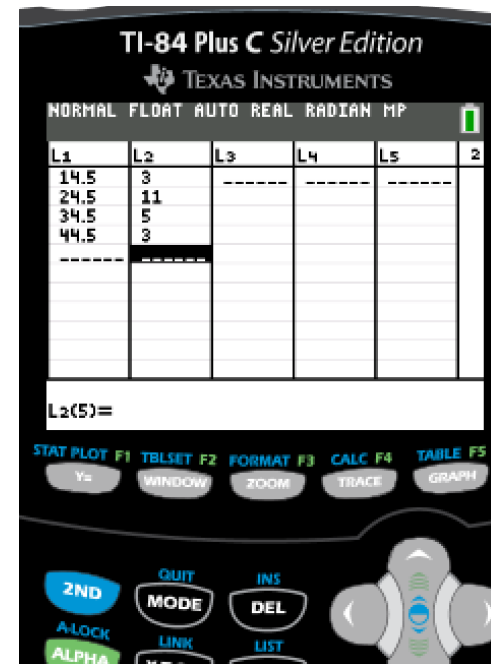
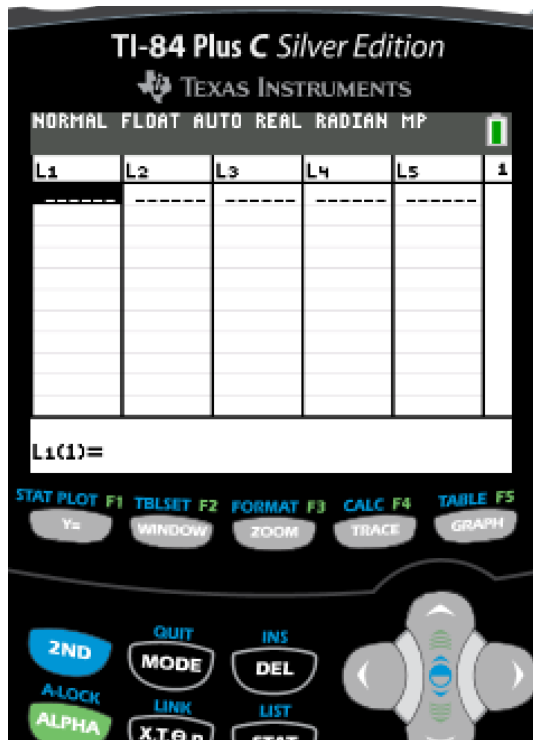
GO TO **STAT** MENU



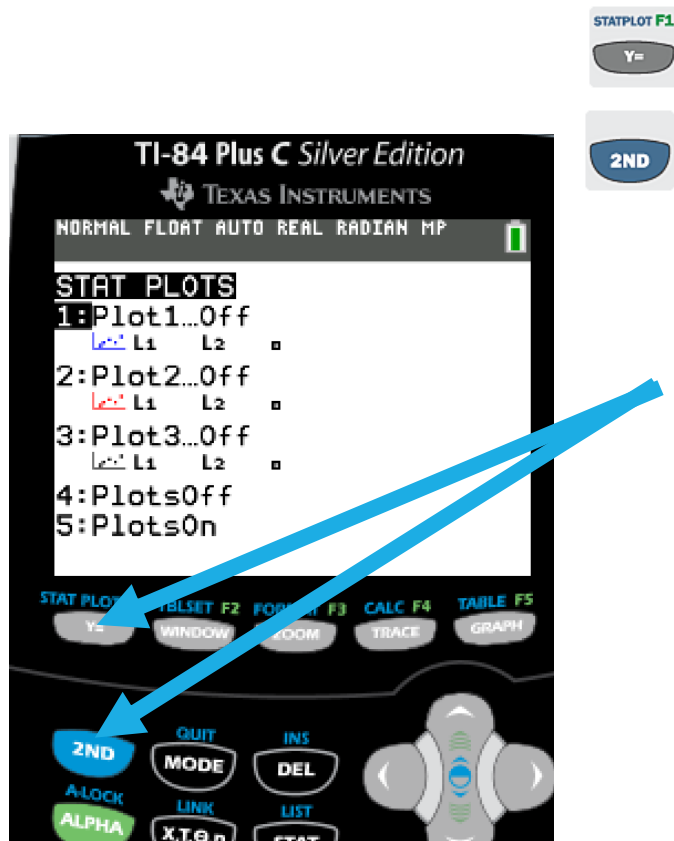
OPTION 1: EDIT

ENTER MIDPOINT FOR EACH CLASS
IN L1 & FREQUENCY IN L2

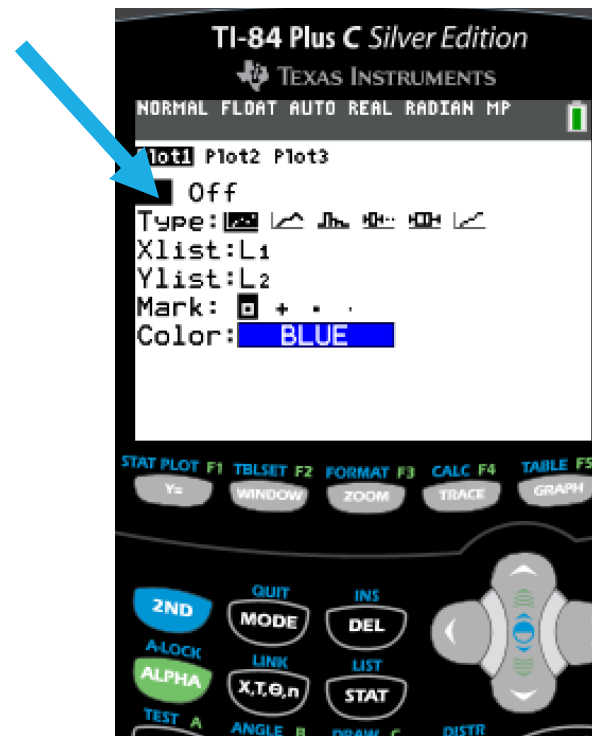
Example:



Go to STAT PLOT --Press



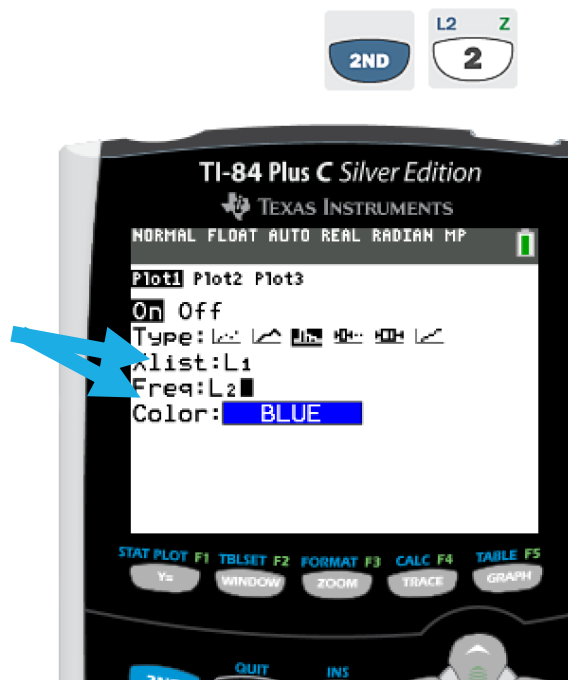
Turn on Plot 1 --choose option 1,
arrow to left highlight "On"



CONTINUE SETUP OF HISTOGRAM

Select type of graph —arrow down to “type”, arrow right 3 times

Indicate location of data---
Xlist (midpoint)= L1 (default)
Freq = L2 Press



SETUP WINDOW—SCALES FOR X- & Y- AXIS

Press



EXAMPLE:

GENERIC WINDOW

x_{min} = first lower class limit

X_{max} = last upper class limit

X_{scl} = class width

Y_{min} = 0

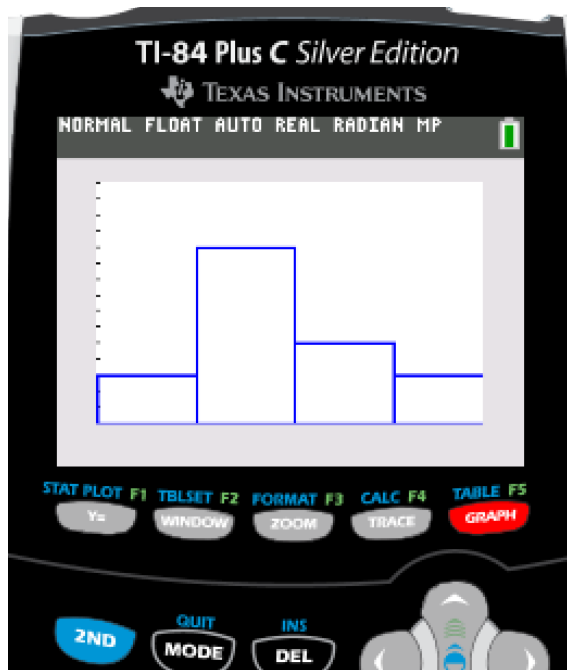
Y_{max} = at least high frequency

Y_{scl} 1



GRAPH

Press  to see histogram on screen



FREQUENCY POLYGONS

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. To plot the points, you need to find the midpoints of the classes.

For the beginning data point, you need a frequency of 0 and a lower midpoint. For the ending data point, you need a frequency of 0 and an upper bound above the higher midpoint.

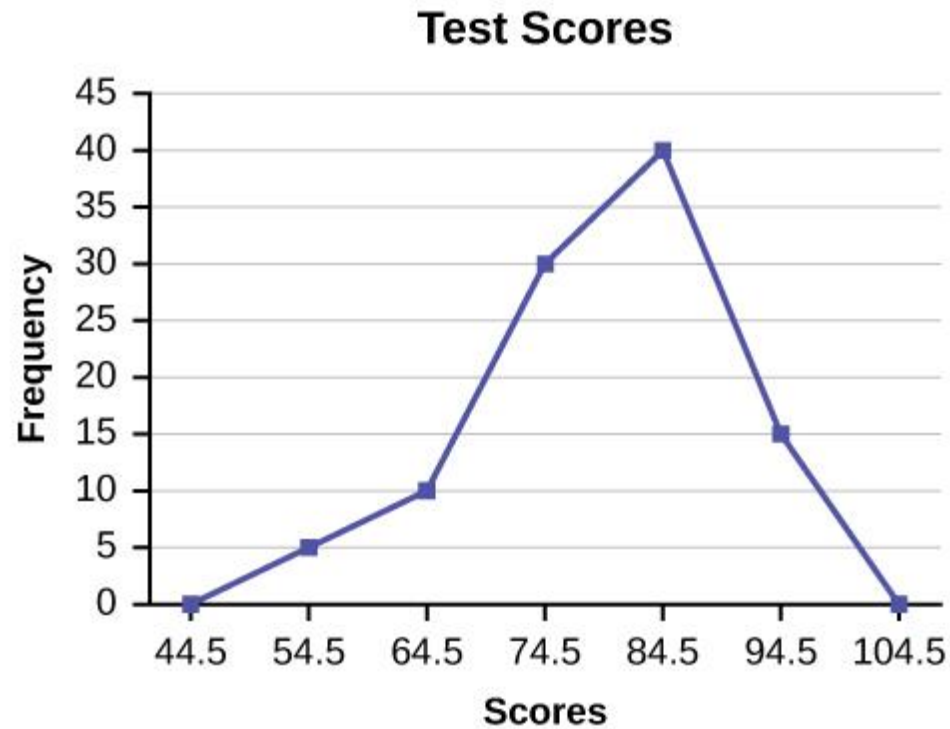
After all the points are plotted, draw line segments to connect them.

FREQUENCY POLYGON EXAMPLE

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Midpoint	Frequency	Cumulative Frequency
49.5	59.5	54.5	5	5
59.5	69.5	64.5	10	15
69.5	79.5	74.5	30	45
79.5	89.5	84.5	40	85
89.5	99.5	94.5	15	100

EXAMPLE OF FREQUENCY POLYGON



TIME SERIES GRAPH

To construct a time series graph, we must look at both pieces of our paired data set.

We start with a standard Cartesian coordinate system.

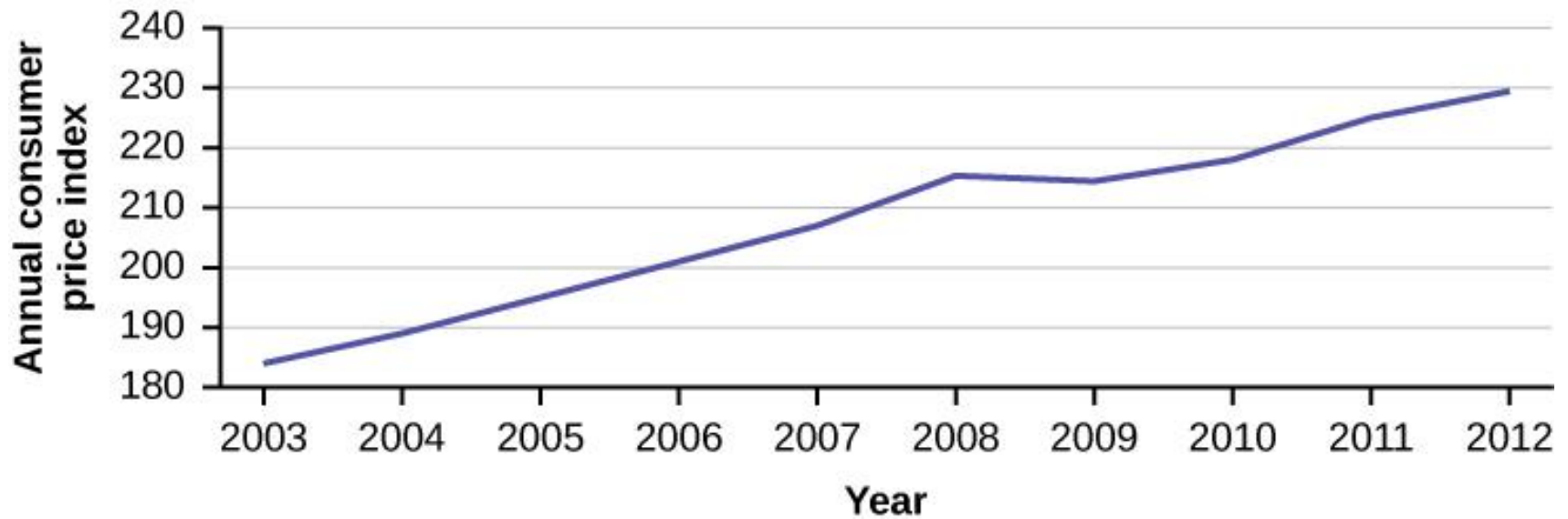
The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring.

By doing this, we make each point on the graph correspond to a date and a measured quantity.

The points on the graph are typically connected by straight lines in the order in which they occur.

TIME SERIES GRAPH

Annual CPI



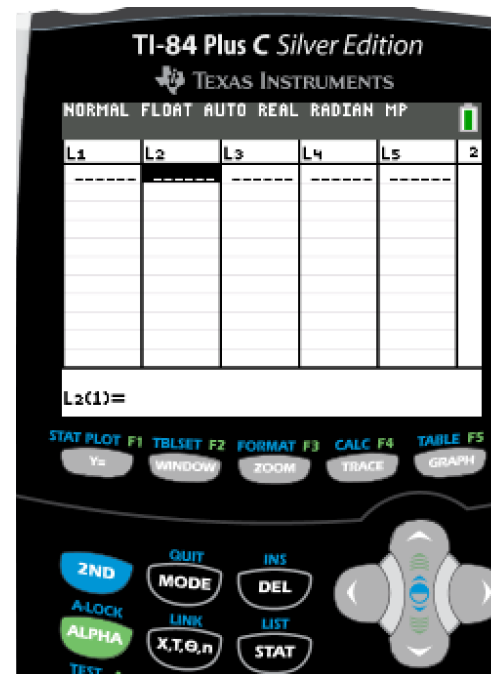
DESCRIPTIVE STATISTICS USING THE TI-83/84 CALCULATOR

DATA ENTRY

Go to



Choose Option 1: Edit

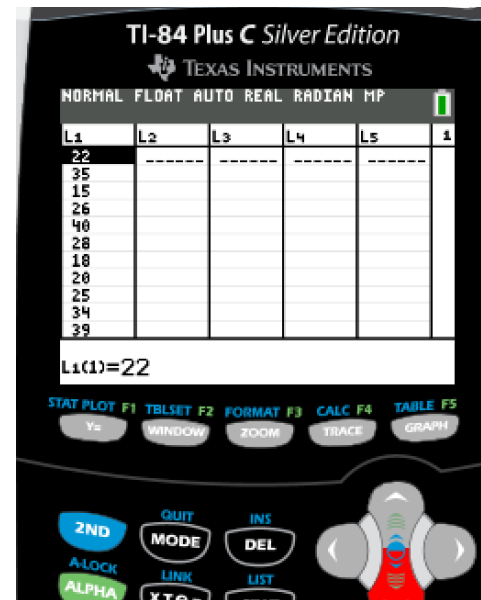


ENTER DATA INTO COLUMN (AKA “LIST”)

EXAMPLE: Number of
Employees in various New York
City restaurants

Type in each individual data
value, enter, repeat until all data
values are entered.

22	35	15	26
	40	28	
18	20	25	34
	39	42	
24	22	19	27
	22	34	
40	20	38	28



DESCRIPTIVE STATISTICS

Go to

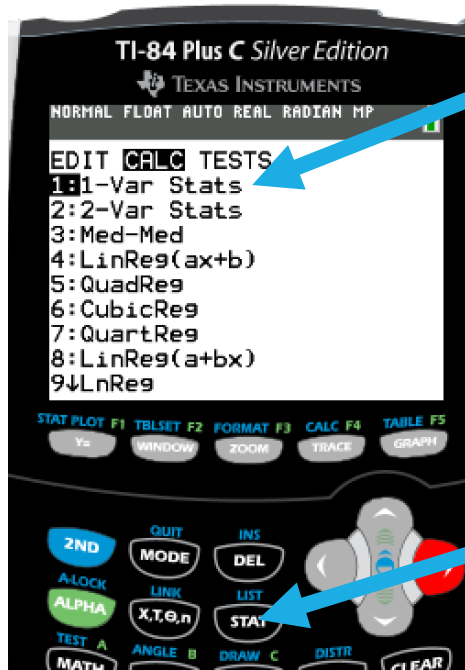


Arrow right



to "CALC"

Choose option 1: 1-Var Stats



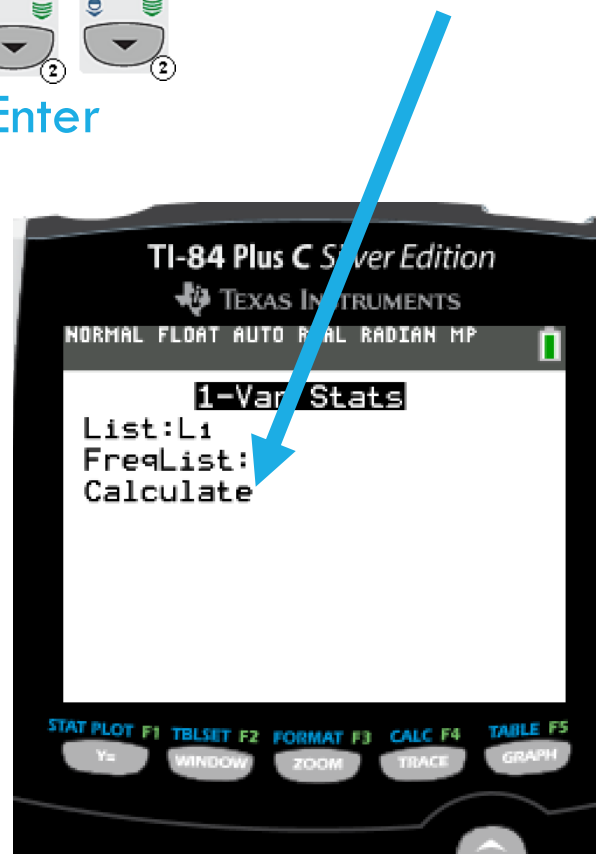
Indicate which list contains data
Click “number of list”



Arrow down twice to calculate



Enter



RESULTS ARE DISPLAYED IN THE WINDOW

\bar{x} = Sample mean

$\sum x$ = sum of the data values, x

$\sum x^2$ = sum of the square of the data values, x^2

s_x = Sample standard deviation

σ_x = Population standard deviation

n = Sample size (number of data values)

$\min X$ = minimum data value

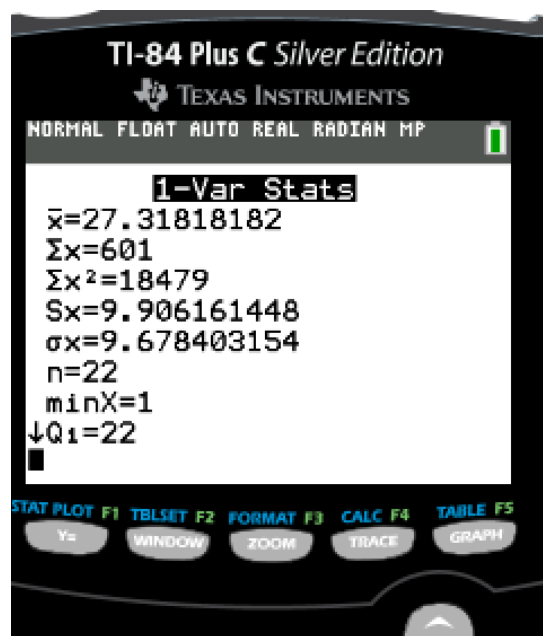
Q_1 = Quartile 1 (25th Percentile)

med = median (middle value) of the data values (50th Percentile)

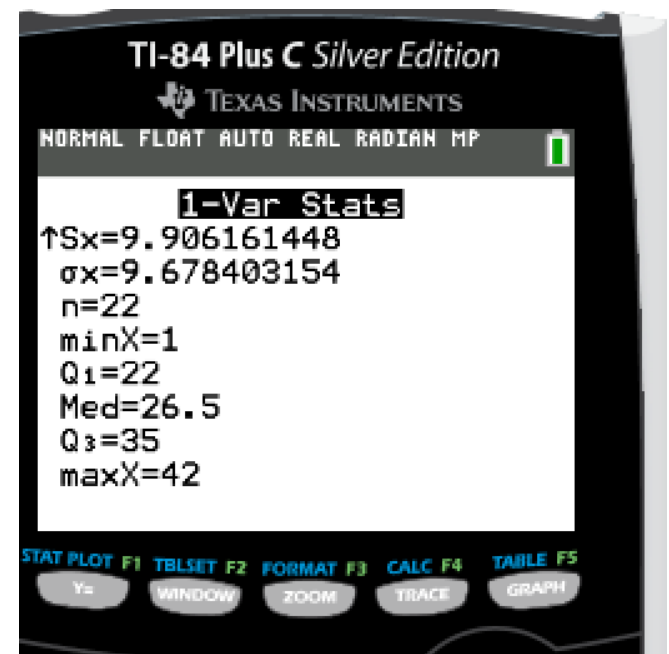
Q_3 = Quartile 3 (75th Percentile)

$\max X$ = maximum data value

EXAMPLE



Scroll down to see all the results



TI-83/84 CALCULATOR DOES NOT CALCULATE, BUT GIVES YOU THE INFORMATION SO YOU CAN DETERMINE THE FOLLOWING:

Mode = data value that occurs most often

Sample Variance = square of sample standard deviation = $(s_x)^2$

Population Variance = square of population standard deviation = $(\sigma_x)^2$

Range = maximum data value – minimum data value = $\max X - \min x$

Interquartile Range (IQR) = $Q_3 - Q_1$



SECTION 2.3

MEASURES OF THE LOCATION OF THE DATA

QUARTILES AND INTERQUARTILE RANGE

Quartiles are special percentiles.

The first quartile, Q_1 , is the same as the 25th percentile

The third quartile, Q_3 , is the same as the 75th percentile.

The median, M , is called both the second quartile and the 50th percentile

To calculate quartiles, the data must be ordered from smallest to largest.

Quartiles divide ordered data into quarters.

Interquartile range is found by subtracting the first quartile from the third quartile.

OUTLIERS

Outliers is a value that is considerably larger or considerably smaller than most of the values in a data set.

Method for finding Outliers:

Find the first quartile $Q1$, and the third quartile $Q3$.

Compute the interquartile range: $IQR = Q3 - Q1$.

Compute the outlier boundaries. These boundaries are the cutoff points for determining outliers:

Lower Outlier Boundary = $Q1 - 1.5(IQR)$

Upper Outlier Boundary = $Q3 + 1.5(IQR)$

Any data value that is less than the lower outlier boundary or greater than the upper outlier boundary is considered to be an outlier.

USING THE CALCULATOR TO SOLVE FOR QUARTILES

To find the quartiles, I will need to first look at the data below.

1,2,2,3,3,3,4,4,4,5,6,6,6,7,7

To use the calculator, follow these directions:

Press the STAT button

Press Enter (The first list on the far left is L1. If it is not, then you need to reset your calculator. Directions are the bottom of this sheet)

In L1, Put the following data going down (NOT ACROSS):

Press the STAT button

Press the Right arrow button (It moves the prompt from EDIT to CALC)

Press ENTER Twice

USING THE CALCULATOR TO SOLVE FOR QUARTILES

From the calculator results, we can determine the following information:

Q1 is the first quartile of the data is 3

Med is the median of the data is 4

Q3 is the third quartile of the data is 6

My IQR is $6 (Q3) - 3 (Q1) = 3$

There are no outliers because none of the data values are lower than -1.5 or greater than 10.5.

Lower Outlier Boundary = $3 - 1.5(3) = -1.5$

Upper Outlier Boundary = $6 + 1.5(3) = 10.5$

FINDING PERCENTILES IN DATA

Arrange the data in order from smallest to largest

Use the formula where **p** is the percentile and **n** is the number of values in the data

$$L = \frac{P}{100} (n + 1)$$

L means the location of the data value for that percentile

If L is a whole number, go to the location and get the actual value and the actual value to the right of it. Then find the average of the two actual values (not L). That is the value for that percentile.

If L is not a whole number, round it up to the next highest whole number. The *p*th percentile is the number in the position corresponding to the rounded-up value.

EXAMPLE OF PERCENTILE IN DATA

Find the 28th percentile of the data set

2,3,4,5,6,7,7,8,9,10,11

$$L = \frac{P}{100} (n + 1) = \frac{28}{100} (11 + 1) = 3.36 \text{ round up to 4 so 5 is the 28}^{\text{th}}$$

L means the location of the data value for that percentile

If L is a whole number, go to the location and get the actual value and the actual value to the right of it. Then find the average of the two actual values (not L). That is the value for that percentile.

If L is not a whole number, round it up to the next highest whole number. The p th percentile is the number in the position corresponding to the rounded-up value.

PROCEDURE FOR COMPUTING THE PERCENTILE CORRESPONDING TO A GIVEN DATA VALUE

Arrange the data in order from smallest to largest

x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile

y = the number of data values equal to the data value for which you want to find the percentile.

n = the total number of data.

$$\text{Percentile} = \frac{(x + 0.5y)}{(n)} (100)$$

Round the result to the nearest whole number

EXAMPLE OF COMPUTING THE PERCENTILE CORRESPONDING TO A GIVEN DATA VALUE

Find what percentile 9 is in the data set

2,3,4,5,6,7,7,8,9,10,11

$$\text{Percentile} = \frac{(x+0.5y)}{(n)} (100) = \frac{(8+0.5(1))}{(11)} (100)$$

The answer is 77.27 and round to 77

9 is at the 77th percentile

FIVE NUMBER SUMMARY

(Use the calculator to solve these problems)

The five number summary is the following numbers in a data set.

Minimum (smallest value in data)

First Quartile (25th percentile in data)

Median (50th percentile in data)

Third Quartile (75th percentile in data)

Maximum (largest value in data)

2.4 BOX PLOTS

BOX PLOTS OR (BOX-AND-WHISKER PLOTS)

A box plot is constructed from five values:

Minimum

First Quartile

Median

Third Quartile

Maximum

We use these values to compare how close other data values are to them. You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

BOX PLOTS

To construct a box plot, use a horizontal or vertical number line and a rectangular box.

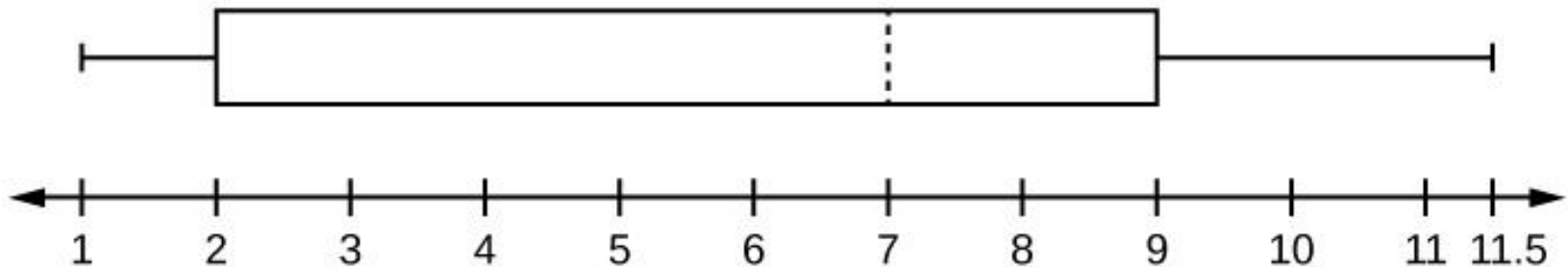
- The smallest and largest data values label the endpoints of the axis.
- The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately the middle 50 percent of the data fall inside the box.
- The "whiskers" extend from the ends of the box to the smallest and largest data values.
- The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both.

EXAMPLE OF BOX PLOT

Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.



EXAMPLE

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

Make two box plots (one for day and one for night)

EXAMPLE CONTINUED

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.

b. Find the smallest and largest values, the median, and the first and third quartile for the night class.

c. For each data set, what percentage of the data is between the smallest value and the first quartile?

the first quartile and the median?

the median and the third quartile?

the third quartile and the largest value?

What percentage of the data is between the first quartile and the largest value?

Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)?

2.5 MEASURES OF CENTER OF THE DATA

MEASURES OF CENTER

The "center" of a data set is also a way of describing location.

The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**.

To calculate the **mean** weight of 50 people, add the 50 weights together and divide by 50.

To find the **median** weight of the 50 people, order the data and find the number that splits the data into two equal parts.

The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

MEAN (AVERAGE)

Sample mean

The letter used to represent the sample mean is an x with a bar over it (pronounced “xbar”): \bar{x}

To get the sample mean, you add all your sample numbers together and divide by your sample size (**n**).

$$\bar{x} = \frac{\sum x}{n}$$

Population mean

The Greek letter μ (pronounced “mew”) represents the population mean

To get the population mean, you add all your population numbers together and divide by your population size (**N**).

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

EXAMPLE OF MEAN

Find the average age of students in Math 2200;

18, 18, 17, 16 , 19, 24, 22, 37, 19,
21, 20, 18

FIND THE MEAN USING THE CALCULATOR

To find the mean:

Press STAT

1:EDIT.

Put the data values into list L1.

Press STAT and

Arrow to CALC.

Press 1:1-VarStats.

Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

Look for \bar{x} for the sample and population mean

MEDIAN (M)

You can quickly find the location of the **median** by using the expression

$$\frac{(n+1)}{2}$$

The letter n is the total number of data values in the sample.

If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest).

If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered.

MEDIAN (M)

For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$.

The median is the 49th value in the ordered data.

If the total number of data values is 100, then $\frac{n+1}{2} = \frac{101+1}{2} = 50.5$.

The median occurs midway between the 50th and 51st values.

The location of the median and the value of the median are **not** the same.

EXAMPLE OF MEDIAN

Find the median age of students in Math 2200;

18, 18, 17, 16 , 19, 24, 22, 37, 19,
21, 20, 18

FIND THE MEDIAN USING THE CALCULATOR

To find the median:

Press STAT

1:EDIT.

Put the data values into list L1.

Press STAT and

Arrow to CALC.

Press 1:1-VarStats.

Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

Look for **Med** for the median

MEAN VERSUS MEDIAN

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the "center": the mean or the median?

MODE

Another measure of the center is the **mode**.

The **mode** is the most frequent value.

There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest.

A data set with two modes is called **bimodal**.

EXAMPLE OF MODE

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72;

72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

59 – 2

63 – 2

72 – 5 the mode is 72

84 - 3

THE LAW OF LARGE NUMBERS AND THE MEAN

The Law of Large Numbers says that if you take samples of larger and larger size from any population,

then the mean \bar{x} of the sample is very likely to get closer and closer to μ of the population

SAMPLING DISTRIBUTIONS AND STATISTIC OF A SAMPLING DISTRIBUTION

You can think of a **sampling distribution** as a relative frequency distribution with a great many samples.

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

CALCULATING THE MEAN OF GROUPED FREQUENCY TABLES

1.) Find the midpoint (or midrange) of each class:

$$MR = \frac{\text{highestvalue} + \text{lowestvalue}}{2}$$

$$2.) \text{ Mean of Frequency Table} = \frac{\sum(f * m)}{\sum(f)}$$

f= the frequency of the interval

m= the midpoint of the interval.

FIND THE BEST ESTIMATE OF THE CLASS MEAN.

A frequency table of Grade

Interval	Number of Students
----------	--------------------

50–56.5	1
---------	---

56.5–62.5	0
-----------	---

62.5–68.5	4
-----------	---

68.5–74.5	4
-----------	---

74.5–80.5	2
-----------	---

80.5–86.5	3
-----------	---

86.5–92.5	4
-----------	---

92.5–98.5	1
-----------	---

Find the midpoints for each class

50–56.5	53.25
---------	-------

56.5–62.5	59.5
-----------	------

62.5–68.5	65.5
-----------	------

68.5–74.5	71.5
-----------	------

74.5–80.5	77.5
-----------	------

80.5–86.5	83.5
-----------	------

86.5–92.5	89.5
-----------	------

92.5–98.5	95.5
-----------	------

FIND THE BEST ESTIMATE OF THE CLASS MEAN.

Calculate the sum of the product of each interval frequency and midpoint.

$$\sum (f*m) =$$

$$53.25(1)+59.5(0)+65.5(4)+71.5(4)+77.5(2)+83.5(3)+89.5(4)+95.5(1)=1460.25$$

$$\mu = \frac{\sum(f*m)}{\sum(f)} = 1460.25 / 19 = 76.86$$

USING THE CALCULATOR

You can use the calculator to solve for the mean of grouped frequency tables.

Press STAT and 1.

In L1, list the midpoints for the classes

In L2, list the frequencies for each of the classes. The frequency should match the class midpoint

Press STAT and scroll over to CALC.

Press 2: 2-Var STATS and ENTER again

The mean is listed. (\bar{x})

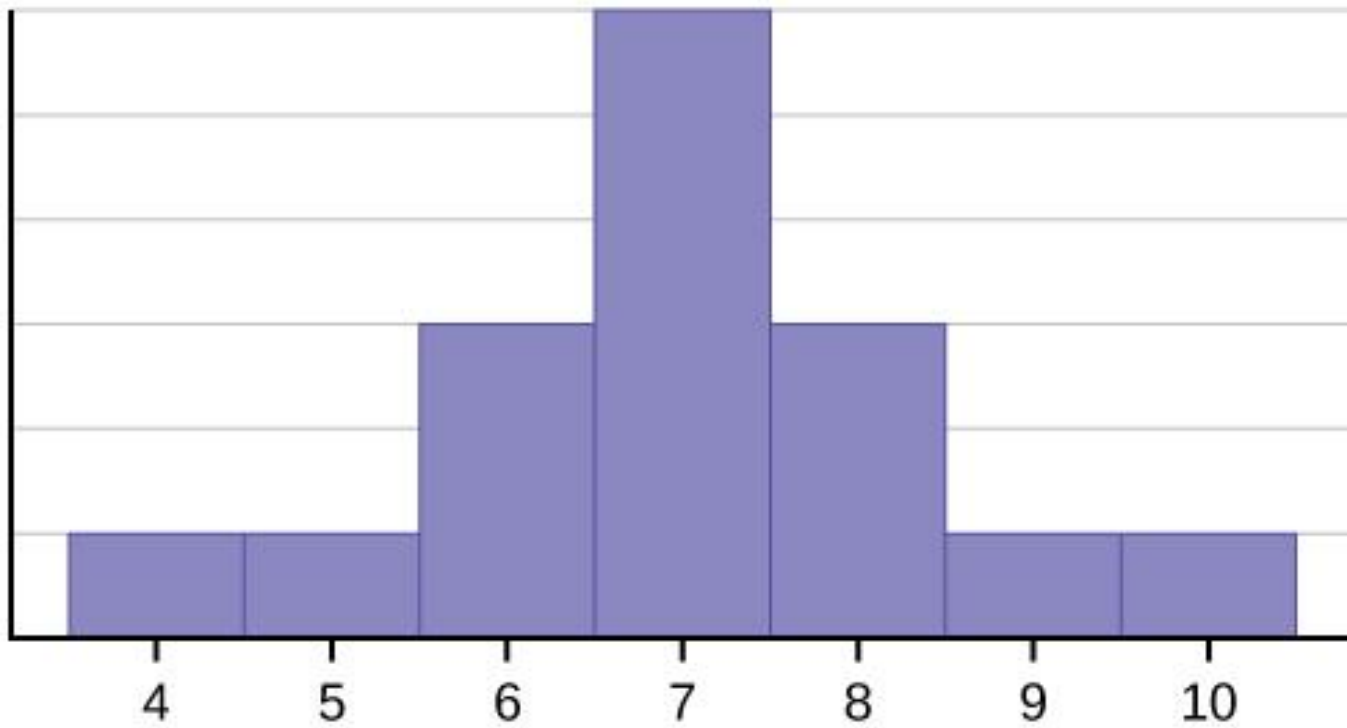
2.6 SKEWNESS AND THE MEAN, MEDIAN, AND MODE

SYMMETRIC SHAPE

The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other.

In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

SYMMETRIC HISTOGRAM



SKEWED TO THE LEFT

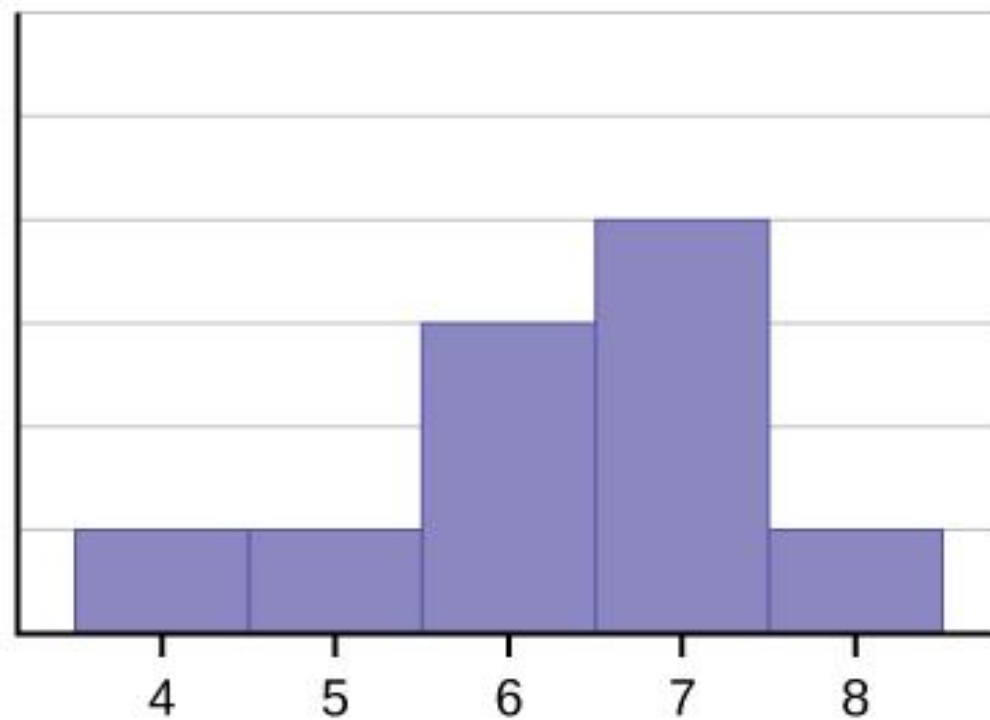
The histogram for the data:

4; 5; 6; 6; 6; 7; 7; 7; 7; 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called skewed to the left because it is pulled out to the left.

The mean is 6.3, the median is 6.5, and the mode is seven. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so. To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode.

SKEWED TO THE LEFT



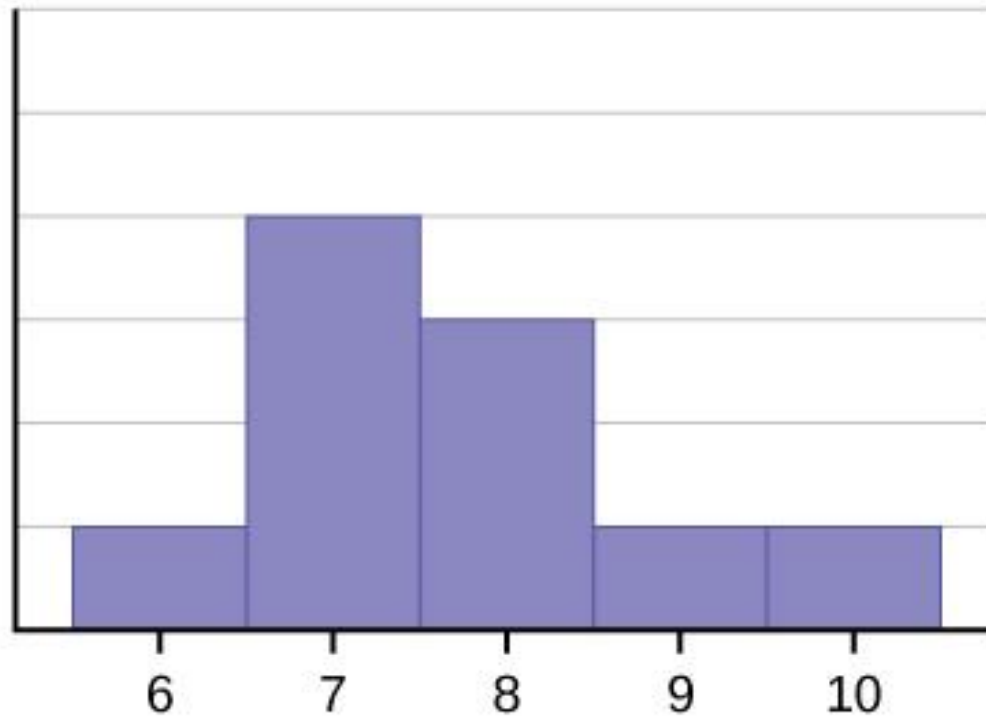
SKEWED TO THE RIGHT

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is skewed to the right.

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

SKEWED TO THE RIGHT



2.7 MEASURES OF THE SPREAD OF THE DATA

STANDARD DEVIATION

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- can be used to determine whether a particular data value is close to or far from the mean.

STANDARD DEVIATION

The standard deviation provides a measure of the overall variation in a data set. The standard deviation is always positive or zero.

The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread.

The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

EXAMPLE

Suppose that Rosa and Binh both shop at supermarket A.

Rosa waits at the checkout counter for seven minutes and Binh waits for one minute.

At supermarket A, the mean waiting time is five minutes and the standard deviation is two minutes.

So the mean is 5 and the standard deviation is 2

The standard deviation can be used to determine whether a data value is close to or far from the mean.

EXAMPLE OF ROSA

So the mean is 5 and the standard deviation is 2

Rosa waits for 7 minutes so I am going to compare it to the mean.

$$7 - 5 = 2$$

I know that 2 is the same as the standard deviation so 2 minutes is equal to 1 standard deviation.

Rosa's wait time of 7 minutes is 2 minutes longer than the average of 5 minutes.

Rosa's wait time of 7 minutes is 1 standard deviation above the average of 5 minutes..

EXAMPLE OF BINH

Again the mean is 5 and the standard deviation is 2

Binh waits for 1 minute

So $1 - 5 = -4$

1 is 4 minutes less than the average of 5; 4 minutes is equal to 2 standard deviations.

Binh's wait time of 1 minute is 4 minutes less than the average of 5 minutes.

Binh's wait time of 1 minute is 2 standard deviations below the average of 5 minutes.

LOOKING AT IT AGAIN

The mean is 5 and the standard deviation is 2

2 standard deviations above mean = $5 + (2 \times 2) = 9$ minutes

1 standard deviation above mean = $5 + (1 \times 2) = 7$ minutes

Mean = 5 minutes

1 standard deviation below mean = $5 - (1 \times 2) = 3$ minutes

2 standard deviations below mean = $5 - (2 \times 2) = 1$ minute

STANDARD DEVIATION

A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average.

Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule.

In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

STANDARD DEVIATION AND MEANS

The equation value = mean + (# of STDEVs)(standard deviation) can be expressed for a sample and for a population.

Sample: $x = \bar{x} + (\# \text{ of STDEV})(s)$

\bar{x} is the sample mean and s is the sample standard deviation.

Population: $x = \mu + (\# \text{ of STDEV})(\sigma)$

μ is the population mean and σ is the population standard deviation.

STANDARD DEVIATION

If x is a number, then the difference " $x - \text{mean}$ " is called its deviation.

In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation.

If the numbers belong to a population, in symbols a deviation is $x - \mu$.
For sample data, in symbols a deviation is $x - \bar{x}$

STANDARD DEVIATION

Sample Standard
Deviation

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

\bar{x} is the sample mean

n is the sample size

Population Standard
Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

μ is the population mean

N is the population

VARIANCE

Sample Variance

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

\bar{x} is the sample mean

n is the sample size

Population Variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

μ is the population mean

N is the population

STANDARD DEVIATION

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other.

The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

EXAMPLE OF STANDARD DEVIATION AND VARIANCE

Find the average age of students in Math 2200;

18, 18, 17, 16 , 19, 24, 22, 37, 19,
21, 20, 18

FIND THE STANDARD DEVIATION AND VARIANCE USING THE CALCULATOR

Press STAT

1:EDIT.

Put the data values into list L1.

Press STAT and

Arrow to CALC.

Press 1:1-VarStats.

Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

FIND THE STANDARD DEVIATION AND VARIANCE USING THE CALCULATOR

S_x = the sample standard deviation which is 5.577 or 5.6

To get the sample variance (**s²**) , you need to square the S_x value.

$$(5.6)^2 = 31.36 \text{ or } 31.4$$

σ_x = the population standard deviation which is 5.340 or 5.3

To get the population variance (**σ²**), you need to square the σ_x value.

$$(5.3)^2 = 28.09 \text{ or } 28.1$$

USING THE CALCULATOR FOR GROUPED FREQUENCY TABLES

You can use the calculator to solve for the standard deviation and variance of grouped frequency tables.

Press STAT and 1.

In L1, list the midpoints for the classes

In L2, list the frequencies for each of the classes. The frequency should match the class midpoint

Press STAT and scroll over to CALC.

Press 2: 2-Var STATS and ENTER again

S_x = the sample standard deviation

σ_x = the population standard deviation

The variance is found the same way as noted before.

COMPARING VALUES FROM DIFFERENT DATA SETS (Z-SCORES)

Number of standard deviations = $\frac{(\text{value} - \text{mean})}{\text{standard deviation}}$

Z scores for sample

$$X = \bar{x} + (z * s)$$

$$Z = \frac{(x - \bar{x})}{S}$$

\bar{x} is sample mean

S is sample standard deviation

X is a data value

Z is z-score

COMPARING VALUES FROM DIFFERENT DATA SETS (Z-SCORES)

Number of standard deviations = $\frac{(\text{value} - \text{mean})}{\text{standard deviation}}$

Z scores for populations

$$X = \mu + (z * \sigma)$$

$$Z = \frac{(x - \mu)}{\sigma}$$

μ is population mean

σ is population standard deviation

X is a data value

Z is z-score