Chapter 3 – Descriptive Statistics
Numerical Summaries

Section 3.1 Measures of Central Tendency

Please Note: The Mean, Median, Variance, Standard Deviation, and 5-number summary will be computed using the calculator TI 83/84.

1. Mean (Also called the Arithmetic Mean)

The mean of a data set is the sum of the observations divided by the number of observations.

If the data are $x_1, x_2, x_3, \ldots, x_n$, then Mean = $\frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$

Two Notations for the mean:(a) Sample mean: $\bar{x}$ (read as x-bar)

(b) Population Mean: μ ("Mu")

Thus $\bar{x} = \frac{\sum x}{n}$ where n = # of items in the sample data, and

$\mu = \frac{\sum x}{N}$ where N = size of the population.

Note: Σ (sigma) is a Greek symbol that signifies summation.

Example 1: Find the mean for this sample data:

2, 3, 6, 7, 7, 8, 9, 9, 9, 10

Solution: $\bar{x} = \frac{\sum x}{n} = \frac{2+3+6+7+7+8+9+9+9+10}{10}$ = 70/10 = 7

Example 2: A sample of five families in Cucumber Town, Iowa showed the following annual family incomes:

$17,500, $23,000, $24,000, $26,000, $320,000

Find the mean for this data.

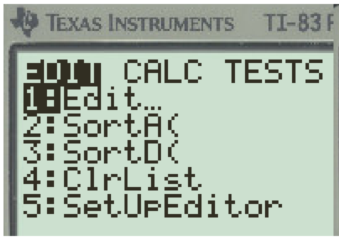
$\bar{x} = \frac{\sum x}{n} = \frac{17500 + 23000 + 24000 + 26000 + 320000}{5}$ = 410500/5 = $82,100

Extreme Value/Outlier: a data value that is too large or too small as compared to most of the data values. Note: The Mean is influenced by outliers.

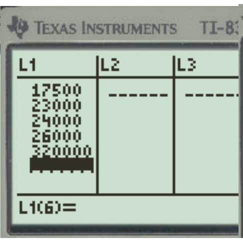
2. **Median** (The median is the middle value of the data when the data has been arranged in ascending/descending order.)

In Example 2 let us use the calculator to find the mean and median.

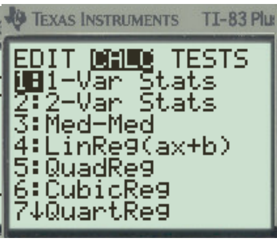
Using TI-83/84: Stat, EDIT, Choose 1 and enter the data in L1, then Stat and move the arrow to the right to CALC, then choose 1: 1-Var Stats, then hit Enter and do 2nd and 1 to choose L1, hit Enter for the results, scroll down to get the median and 5-number summary. In this example, we only need the median.



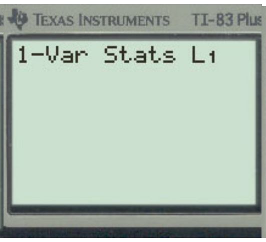
```
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```



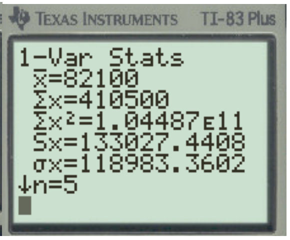
```
L1      L2      L3
17500   ------  ------
23000
24000
26000
320000

L1(6)=
```



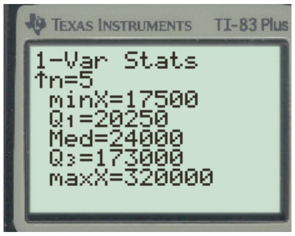
```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```



```
1-Var Stats L1
```



```
1-Var Stats
 x̄=82100
 Σx=410500
 Σx²=1.04487E11
 Sx=133027.4408
 σx=118983.3602
↓n=5
```



```
1-Var Stats
↑n=5
 minX=17500
 Q1=20250
 Med=24000
 Q3=173000
 maxX=320000
```

The calculator does not distinguish between the population mean μ and the sample mean $\bar{X}$ (every mean, population, or sample, is listed as $\bar{X}$). Since in our example the data is for a sample the mean is $\bar{X} = \$82100$. The median is $24000

Example 3: Find the median for the data set 1 and data set 2.

Data Set 1: 7, 2, 8, 5, 9, 4, 7, 8, 6
Data Set 2: 7, 2, 8, 5, 9, 4, 8, 8

Solution: **The median for data set 1 is 7**
The median for the data set 2 is 7.5

Example 4: Find the median for the data in Example 2.

Solution: From the TI 83/84 the **Median = $24,000.**

Note: The median is not affected by extreme values. Thus in the presence of extreme values, the median may be a better indicator of the center. In Example 2 the Median = $24000 is a better measure of the center. The sample mean $\bar{X} = \$82100$ is influenced by an extreme value(outlier) of $320000, there for the median is a better measurement for the center.

3. Mode

The most frequently occurring data value in a set of data is called the mode. That is, the mode is the value that occurs with the greatest frequency.

Example 5. Find the mode for the given data:

2, 3, 3, 2, 2, 8, 7, 8, 7, 9, 8, 8

Solution: **Mode = 8**

Example 6. Find the mode for the given data:

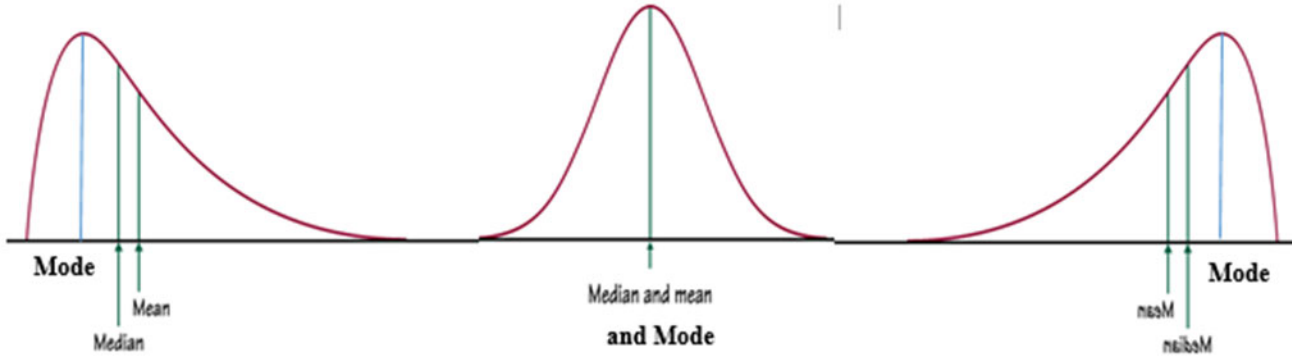
2, 3, 3, 2, 2, 8, 7, 8, 7, 9, 8, 8, 2

Solution: **Mode = 2 and 8.**

Note: Such a distribution is called **bimodal**.

Note: Mode can be used to summarize qualitative variables.

Discuss the shapes of the distribution on page 84.



Homework-Section 3.1 Online - MyStatLab

Section 3.2 Measures of Dispersion (Sample Standard Deviation)

Range = Largest Value – Smallest Value

Example 1: Given the two data sets below, find the range, mean, mode, and median.

Data Set 1: 99, 91, 84, 84, 80, 80, 80, 76, 76, 69, 61
Data Set 2: 99, 80, 80, 80, 80, 80, 80, 80, 80, 80, 61

Soln: For all of the data sets, **Range = 99 – 61 = 38** and **Mean=Median= 80**

Note: The range is based on only two of the items in the data set and thus is influenced too much by extreme values.

Variance: Average Squared Deviation from the Mean

Population Variance $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \dfrac{\sum_{i=1}^{N}x_i^2 - \dfrac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}}{N}$, (N population size).

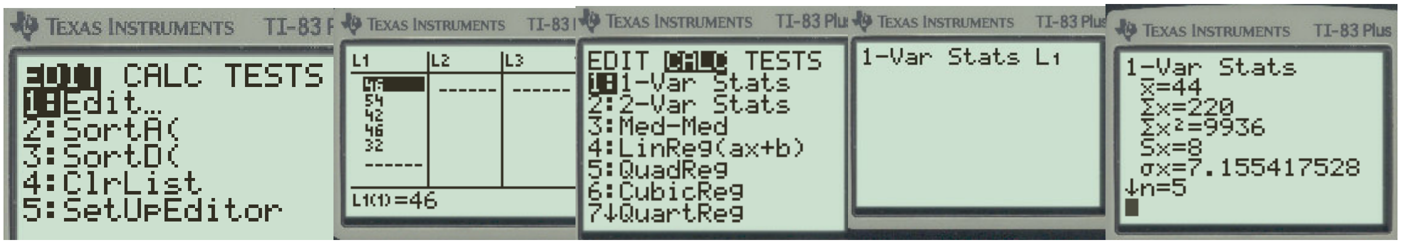
Sample Variance $s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1} = \dfrac{\sum_{i=1}^{n}x_i^2 - \dfrac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}$, (n sample size).

Standard Deviation = $\sqrt{Variance}$
Sample Standard Deviation = s = $\sqrt{s^2}$
Population Standard Deviation = σ = $\sqrt{\sigma^2}$

If the computations are done by hand, first we compute the variance and then take the square root of the variance to get the standard deviation, for the population or sample.
If using the calculator TI-83/84, first we get the standard deviation and then we square the standard deviation to get the variance, for a population or a sample.

The following example will help explain.
Given the data 46, 54, 42, 46, 32. The mean (μ) for this data is 44.
Please note this is a population.

Using TI-83/84: Stat, EDIT, Choose 1 and enter the data in L1, then Stat and move the arrow to the right to CALC, then choose 1: 1-Var Stats, then hit Enter and do 2nd and 1 to choose L1, hit Enter for the results.



```
L1      L2      L3
46      ------  ------
54
42
46
32
------
L1(1)=46
```

```
1-Var Stats L1
```

```
1-Var Stats
 x̄=44
 Σx=220
 Σx²=9936
 Sx=8
 σx=7.155417528
↓n=5
```

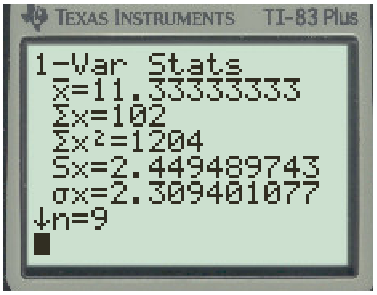
The calculator does not distinguish between the population mean μ and the sample mean $\bar{X}$ (every mean, population, or sample, is listed as $\bar{X}$). Since in our example the data is for a population the mean is $\mu = 44$.

The calculator does distinguish between the population standard deviation σ and the sample standard deviation S. It shows them as follows: population standard deviation σ_x and sample standard deviation S_x. Since in our example the data is for a population, the standard deviation is $\sigma_x = 7.1554$ and the variance is $\sigma_X^2 = (7.1554)^2 =$ 51.2. (rounded to the first decimal). Please note: σ_X means the standard deviation of the random variable X and is the same as σ .

Please note also, from the calculator, we get first σ =7.1554 and then we square it to get the variance, $\sigma^2 = (7.1554)^2 =$ 51.2. (rounded to the first decimal)

Example 2: Given the sample data 9, 11, 16, 14, 12, 12, 10, 9, 9 find the mean and standard deviation.

Using TI-83/84: Stat, EDIT, Choose 1 and enter the data in L1, then Stat and move the arrow to the right to CALC, then choose 1: 1-Var Stats, then hit Enter and do 2nd and 1 to choose L1, hit Enter for the results.

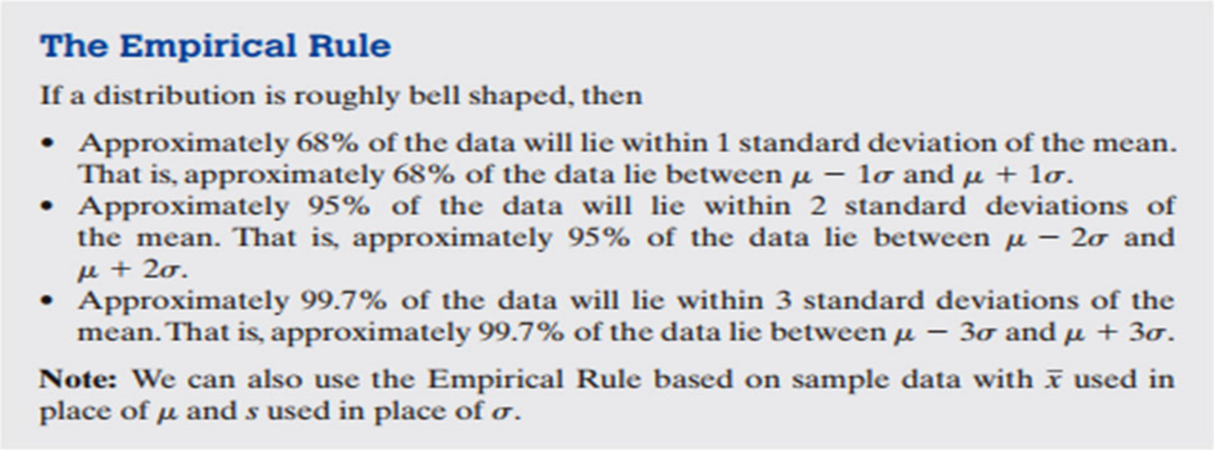


```
1-Var Stats
 x̄=11.33333333
 Σx=102
 Σx²=1204
 Sx=2.449489743
 σx=2.309401077
↓n=9
```

Since in our example the data is for a sample the mean is $\bar{X} = 11.33$ and the sample standard deviation is $S = 2.4495$.

Sample variance is $S^2 = (2.4495)^2 = 6$. (rounded to an integer)

The Empirical Rule (For Bell Shaped Distributions)



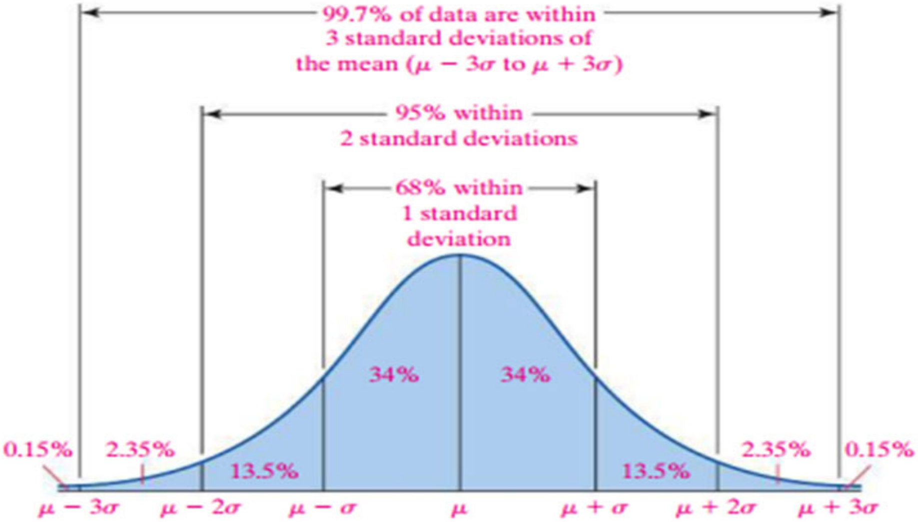
The Empirical Rule

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

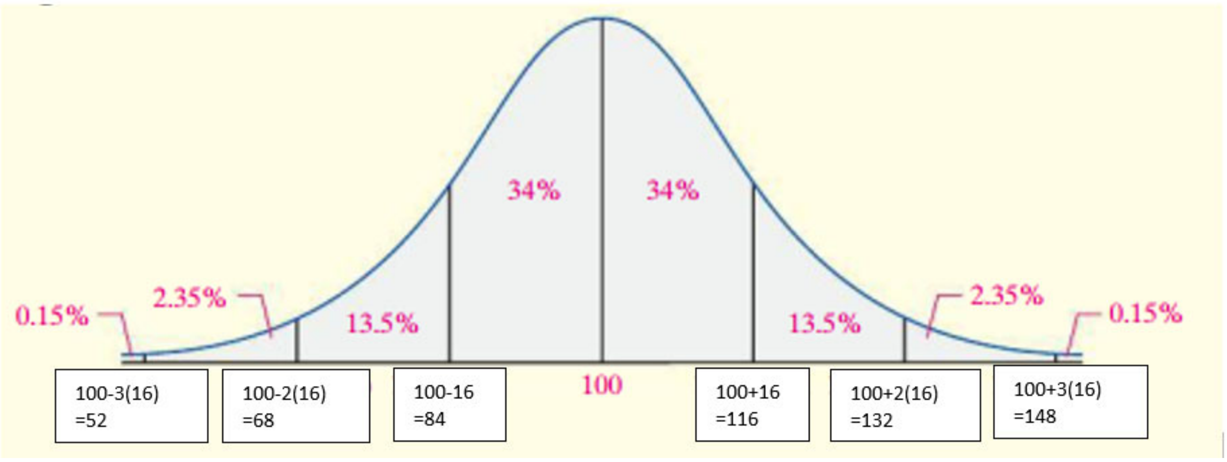
Note: We can also use the Empirical Rule based on sample data with $\bar{x}$ used in place of μ and s used in place of σ.

Figure 13 illustrates the Empirical Rule.


99.7% of data are within 3 standard deviations of the mean ($\mu - 3\sigma$ to $\mu + 3\sigma$)
95% within 2 standard deviations
68% within 1 standard deviation
34%
34%
0.15%
2.35%
13.5%
13.5%
2.35%
0.15%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Example Using the Empirical Rule

The mean IQ score of students enrolled at a certain University is 100 and the standard deviation is 16. We draw a bell-shaped curve, with $\bar{X} = 100$ s =16 to help us answer the following question.



(a) Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.
According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean [that is, greater than or equal to 100-3(16) = 52 and less than or equal to 100+3(16) = 148). 99.7% of the students have IQ between 52 and 148.

(b) Determine the percentage of students who have IQ scores between 68 and 132 according to the Empirical Rule.
Since 68 percent is exactly 2 standard deviations below the mean [100-2(16) =68] and 132 is exactly 2 standard deviations above the mean [100+2(16) = 132], the Empirical Rule tells us that approximately 95% of the IQ scores lie between 68 and 132.

(c) Determine the percentage of students who have IQ scores between 84 and 132 according to the Empirical Rule.
Since between 84 and 100 percent is exactly 34% and between 100 and 132 is 34% +13.5%=47.5%, the total percentage that has IQ scores between 84 and 132 is, [34%+47.5% = 81.5%], 81.5%.

(d) Determine the percentage of students who have IQ scores less than 84 and greater than 116.
Since between 84 and 116 is exactly 68%, the percentage outside this range is 100% - 68% = 32%. Another way to compute it: less than 84% is [13.5%+2.35%+0.15% = 16%] and greater than 116% is [13.5%+2.35%+0.15% = 16%], the total percentage that has IQ scores less than 84 and greater than 116, [16%+16% = 32%], is 32%.

(e) According to the Empirical Rule, what percentage of students have IQ scores above 132?
Based on the Figure above, approximately 2.35%+0.15%=2.5% of students have IQ scores above 132.

CHEBYSHEV'S THEOREM

Based on $\left(1-\frac{1}{K^2}\right)\%$ where K = number of standard deviations.

At least 75% of the items must lie within two standard deviations of the mean;

$$100\left(1-\frac{1}{2^2}\right)\%=100\left(1-\frac{1}{4}\right)\%=100(1-.25)\%=100(0.75)\%=75\%$$

At least 88.89% of the items must lie within three standard deviations of the mean;

$$100\left(1-\frac{1}{3^2}\right)\%=100\left(1-\frac{1}{9}\right)\%=100(1-.111)\%=100(0.889)\%=88.9\%$$

At least 93.75% of the items must lie within four standard deviations of the mean.

$$100\left(1-\frac{1}{4^2}\right)\%=100\left(1-\frac{1}{16}\right)\%=100(1-0.0625)\%=100(0.9375)\%=93.75\%$$

The **CHEBYSHEV'S THEOREM** applies to Non-bell shape distributions

Example Using Chebyshev's THEOREM

The average age of college students at graduation is 28 years with a standard deviation of 2. Answer the following questions.

(a) What percentage of the students graduate between the ages of 24 and 32 years old?
Since 24 is 2 standard deviations below the mean, [28-2(2)=28-4=24], and 32 is 2 standard deviations above the mean, [28+2(2)=28+4=32], the percentage of students that graduate between the ages 24 and 32 years old is 75%.

(b) What percentage of the students graduate between the ages of 20 and 36 years old?
Since 20 is 4 standard deviations below the mean, [28-4(2)=28-8=20], and 32 is 4 standard deviations above the mean, [28+4(2)=28+8=36], the percentage of students that graduate between the ages 20 and 36 years old is 93.75%.

(c) What is the age range such that 88.9% of the students graduating have an age in this range? What is the range? Since 88.9% is within 3 standard deviations, [28-3(2)=28-6=22 and 28+3(2)=28+6=34] the range is 22 to 34 years old.

Homework-Section 3.2 Online - MyStatLab

Section 3.4 Measures of Position and Outliers

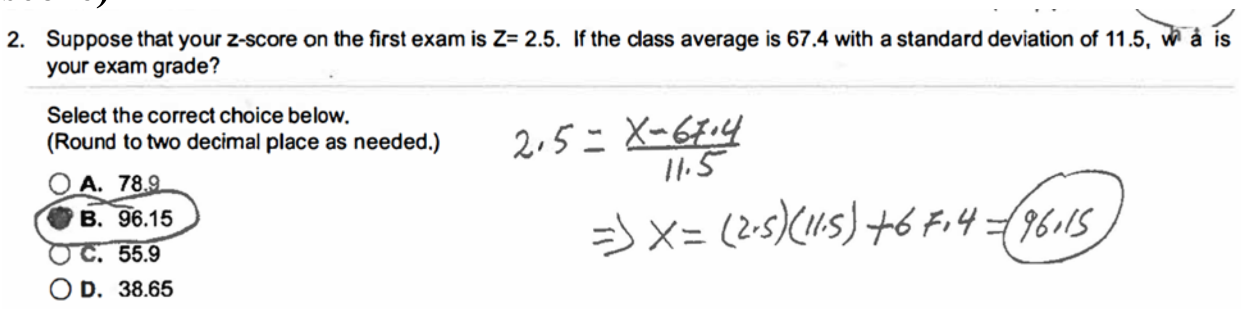
Z-score = $\dfrac{x - \overline{X}}{s}$ where s is the sample s.d.

Z-score = $\dfrac{x - \mu}{\sigma}$ where σ is the population s.d.

The Z-score for any data item is referred to as its standardized value. It can be interpreted as a measure of the relative location of an item in the data.

Example 5: If the Z-score of a data value is Z=2, the data value is 2 standard deviations above the sample mean.

Homework problem 2 (Chapters 3.4 and 3.5-Find X from the Z-score)



2. Suppose that your z-score on the first exam is Z= 2.5. If the class average is 67.4 with a standard deviation of 11.5, what is your exam grade?

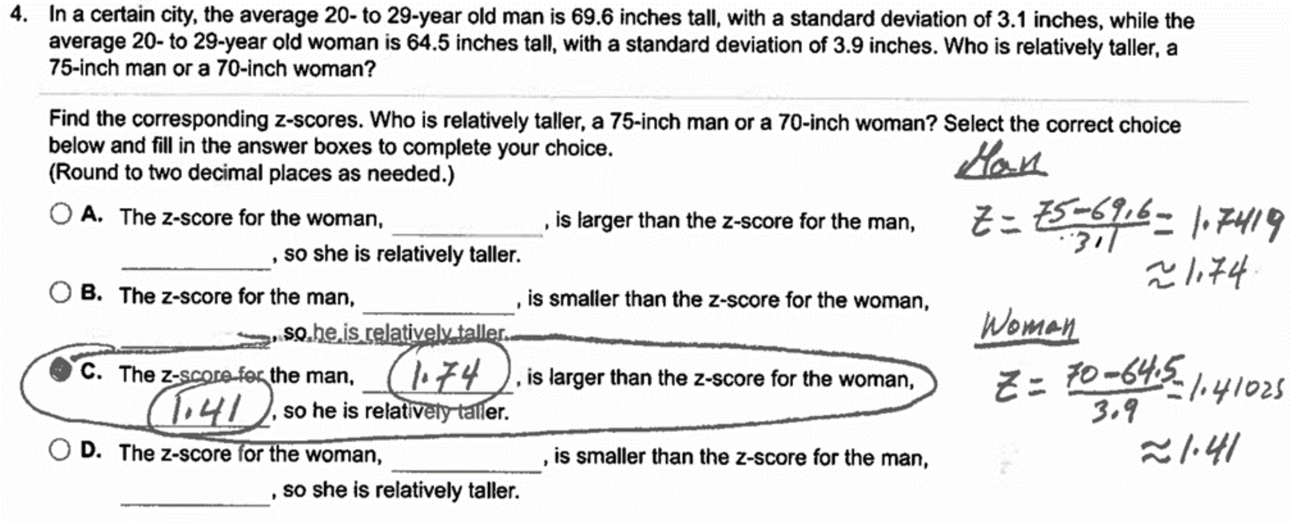
Select the correct choice below.
(Round to two decimal place as needed.)

○ A. 78.9
● B. 96.15
○ C. 55.9
○ D. 38.65

$2.5 = \dfrac{X - 67.4}{11.5}$

$\Rightarrow X = (2.5)(11.5) + 67.4 = \boxed{96.15}$

Homework problem 4 (Chapters 3.4 and 3.5) (Comparing Z-scores)



4. In a certain city, the average 20- to 29-year old man is 69.6 inches tall, with a standard deviation of 3.1 inches, while the average 20- to 29-year old woman is 64.5 inches tall, with a standard deviation of 3.9 inches. Who is relatively taller, a 75-inch man or a 70-inch woman?

Find the corresponding z-scores. Who is relatively taller, a 75-inch man or a 70-inch woman? Select the correct choice below and fill in the answer boxes to complete your choice.
(Round to two decimal places as needed.)

○ A. The z-score for the woman, ______, is larger than the z-score for the man, ______, so she is relatively taller.

○ B. The z-score for the man, ______, is smaller than the z-score for the woman, ______, so he is relatively taller.

● C. The z-score for the man, 1.74, is larger than the z-score for the woman, 1.41, so he is relatively taller.

○ D. The z-score for the woman, ______, is smaller than the z-score for the man, ______, so she is relatively taller.

Man

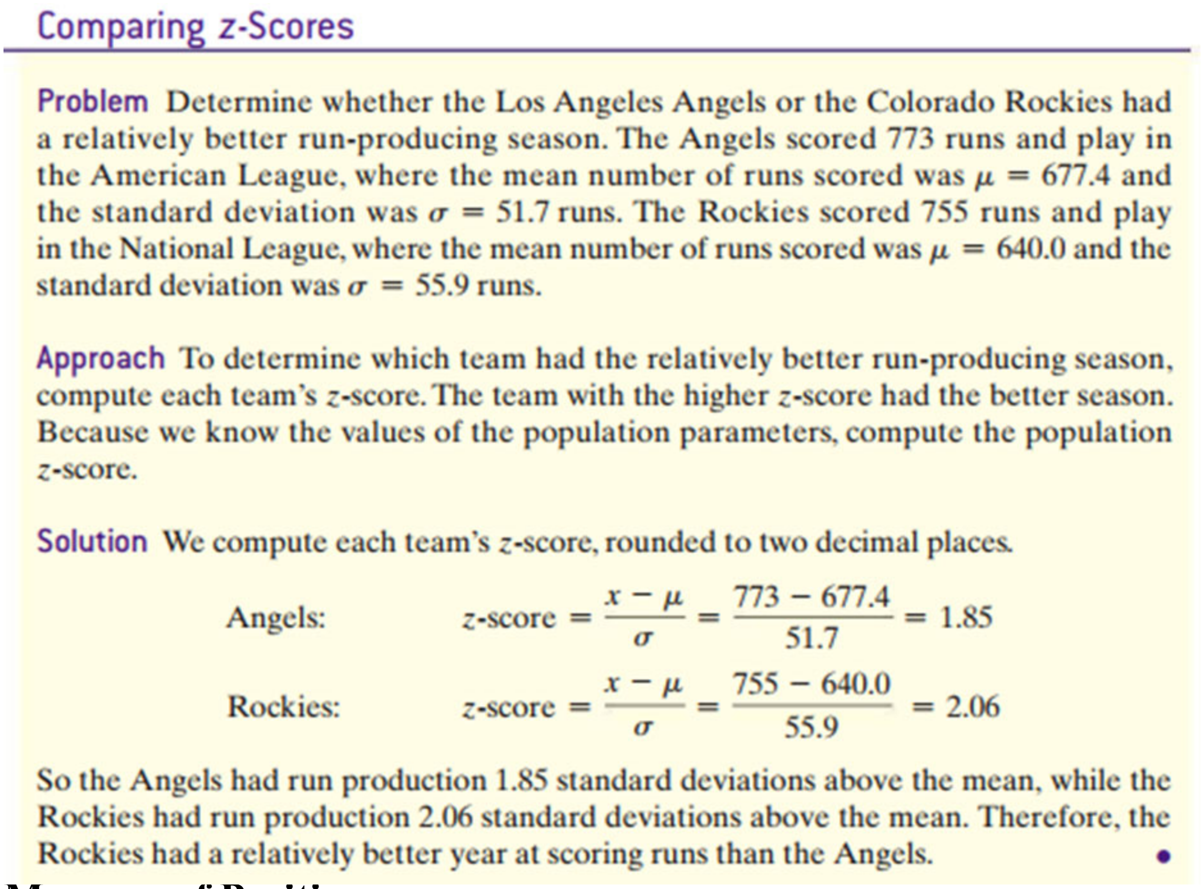
$Z = \dfrac{75 - 69.6}{3.1} = 1.7419$

≈ 1.74

Woman

$Z = \dfrac{70 - 64.5}{3.9} = 1.41025$

≈ 1.41



Comparing z-Scores

Problem Determine whether the Los Angeles Angels or the Colorado Rockies had a relatively better run-producing season. The Angels scored 773 runs and play in the American League, where the mean number of runs scored was $\mu = 677.4$ and the standard deviation was $\sigma = 51.7$ runs. The Rockies scored 755 runs and play in the National League, where the mean number of runs scored was $\mu = 640.0$ and the standard deviation was $\sigma = 55.9$ runs.

Approach To determine which team had the relatively better run-producing season, compute each team's z-score. The team with the higher z-score had the better season. Because we know the values of the population parameters, compute the population z-score.

Solution We compute each team's z-score, rounded to two decimal places.

$$\text{Angels:} \qquad z\text{-score} = \frac{x-\mu}{\sigma} = \frac{773-677.4}{51.7} = 1.85$$

$$\text{Rockies:} \qquad z\text{-score} = \frac{x-\mu}{\sigma} = \frac{755-640.0}{55.9} = 2.06$$

So the Angels had run production 1.85 standard deviations above the mean, while the Rockies had run production 2.06 standard deviations above the mean. Therefore, the Rockies had a relatively better year at scoring runs than the Angels. •

Measure of Position

Interpret Percentiles: Recall that the median divides the lower 50% of a set of data from the upper 50%. The median is a special case of a general concept called the percentile.

Definition: The **kth** percentile, denoted P_k by a set of data is a value such that k percent of the observations are less than or equal to the value.

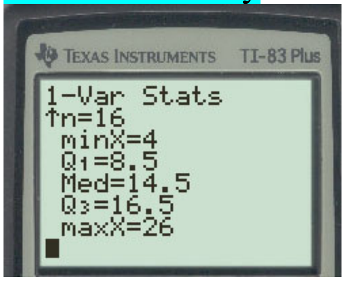
Quartiles: It is often desired to divide a data set into four parts with each part containing one-fourth(25%) of the data.

Q_1(First Quartile)= 25% percentile
Q_2(Second Quartile)=50% percentile
Q_3(Third Quartile)= 75% percentile

Example 2: Given the data below, find Q_1=25th, Q_2=50th, and Q_3=75th percentiles using the TI 83/84 calculator.

26, 4, 5, 20, 6, 12, 15, 15, 15, 8, 9, 10, 14, 18, 16, 17

Using TI-83/84: Stat, EDIT, Choose 1 and enter the data in L1, then Stat and move the arrow to the right to CALC, then choose 1: 1-Var Stats, then hit Enter and do 2nd and 1 to choose L1, hit Enter for the results, scroll down to get the median and 5-number summary.



```
1-Var Stats
↑n=16
 minX=4
 Q1=8.5
 Med=14.5
 Q3=16.5
 maxX=26
```

For the data given in **Example 2**, find the first, second, and third quartiles.

Soln. Q_1 or $P_{25} = 8.5$, Q_2 or $P_{50} = 14.5$, and Q_3 or $P_{75} = 16.5$

Interpretation: 25% of the values lie at 8.5 or below. 50% of the values lie at 14.5 or below. 75% of the values lie at 16.5 or below.

Interpret a Percentile

Problem Jennifer just received the results of her SAT exam. Her SAT Mathematics score of 600 is in the 74th percentile. What does this mean?

Approach The *k*th percentile of an observation means that *k* percent of the observations are less than or equal to the observation.

Interpretation A percentile rank of 74% means that 74% of SAT Mathematics scores are less than or equal to 600 and 26% of the scores are greater. So 26% of the students who took the exam scored better than Jennifer. ●

Homework-Section 3.4 and 3.5 Online - MyStatLab

Section 3.5 The Five Number Summary and Boxplots

The Interquartile Range (IQR): $\mathbf{IQR = Q_3 - Q_1}$

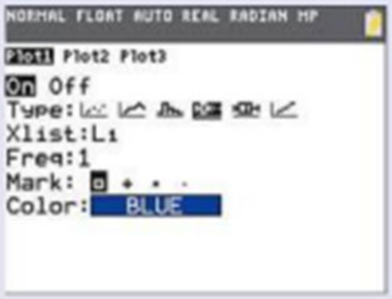
Note: The IQR gives the range of the middle 50% of the observations.

The Five-Number Summary

The five-number summary of a data set: Min, Q_1, Q_2, Q_3, and Max. The IQR together with the 5-number summary is used to build a Boxplot to detect outliers. We will use the TI 83/84 to build a Boxplot.

TI-83/84 Plus

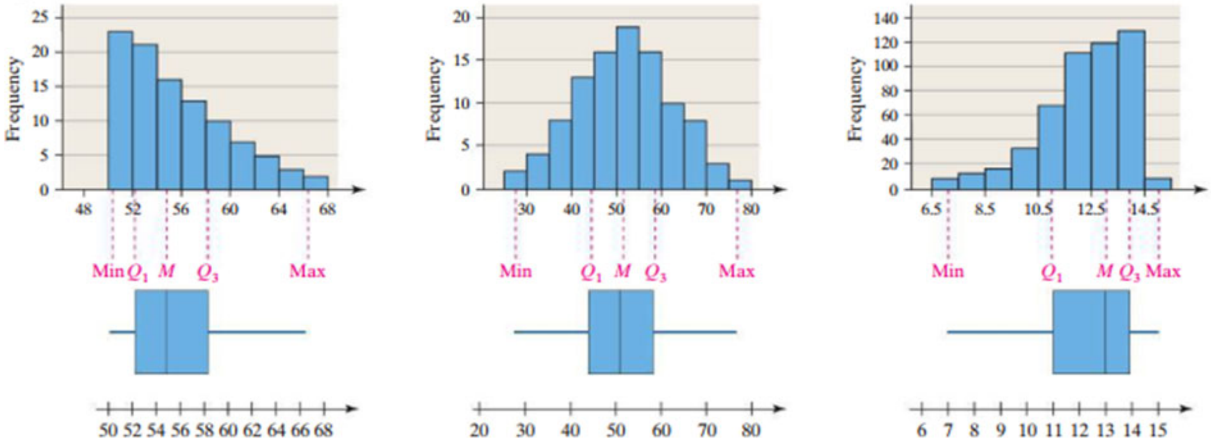
1. Enter the raw data into `L1`.
2. Press `2nd` `Y=` and select `1:Plot 1.`
3. Turn the plots ON. Use the cursor to highlight the modified boxplot icon. Your screen should appear as follows:



```
NORMAL FLOAT AUTO REAL RADIAN MP
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: □ + · ·
Color: BLUE
```

4. Press ZOOM and select `9:ZoomStat.`

Figure 22



(a) Skewed right **(b)** Symmetric **(c)** Skewed left

From the Boxplot we can also decide the shape of the data set; skewed left, right, or symmetric?

Note: When the Median is close to Q_1 the distribution is Skewed Right.
When the Median is closed to Q_3 the distribution is Skewed Left.
When the Median is in the middle of the Box the distribution is Symmetric.

Homework problem 7 (Chapters 3.4 and 3.5) ________ ________

7. The data represent the age of world leaders on their day of inauguration. Find the five-number summary, and construct a boxplot for the data. Comment on the shape of the distribution.

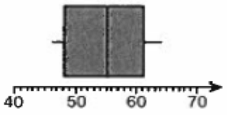
48	50	48	55
46	47	49	64
56	55	61	62
58	52	63	

The five-number summary is 46, 48, 55, 61, 64.

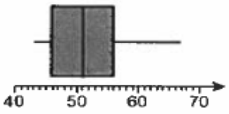
Enter the data in TI 83/84 and build a boxplot.

Choose the correct boxplot of the data below.

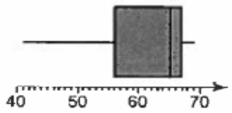
● A.



○ B.



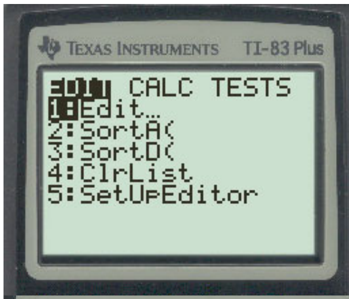
○ C.



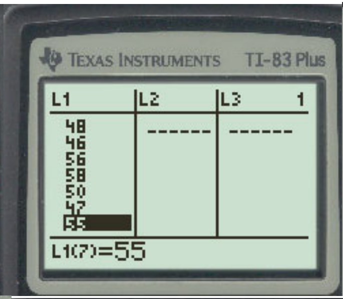
Choose the correct description of the shape of the distribution.

● A. The distribution is roughly symmetric.
○ B. The distribution is skewed to the right.
○ C. The distribution is skewed to the left.
○ D. The shape of the distribution cannot be determined from the boxplot.

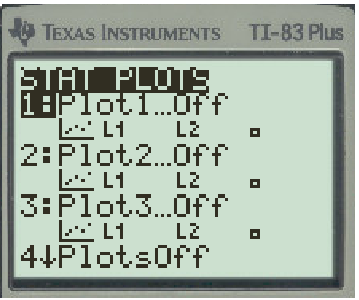
Stat, 1:Edit, and then enter the data in L1.
2nd and Y to access the STATPLOT, Choose 1 for Plot 1, put the cursor on "ON" and enter to turn on Plot 1, move the arrow down, and choose the 1st boxplot on the second row (newer calculators the graphs are in one row), then move the cursor to the XList: and do 2nd and 1 to choose L1, Freq: 1, Mark: chose a symbol to represent the outliers if there are any, then hit Zoom and chose #9, the hit Trace and move the arrow to the right to identify the 5-number summary and outliers.



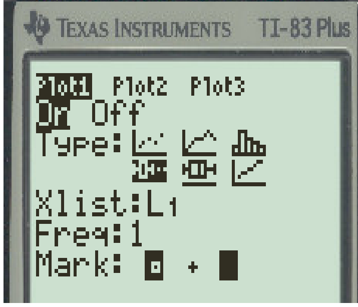
```
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```



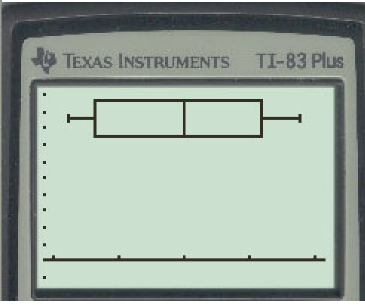
```
L1    |L2     |L3    1
48    |------ |------
46
56
58
50
47
55
L1(7)=55
```

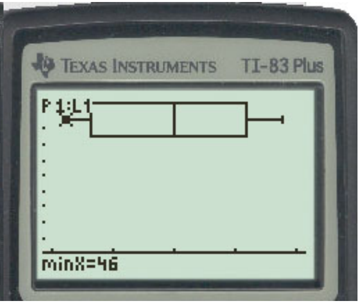


```
STAT PLOTS
1:Plot1…Off
   L1   L2   □
2:Plot2…Off
   L1   L2   □
3:Plot3…Off
   L1   L2   □
4↓PlotsOff
```



```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: □ + ▪
```





```
P1:L1
minX=46
```

Homework-Section 3.4 and 3.5 Online - MyStatLab