CHAPTER 2: Describing Distributions with Numbers

Basic Practice of Statistics

7th Edition

In Chapter 2 we cover …

- Measuring center: the mean
- Measuring center: the median
- Comparing the mean and the median
- Measuring variability: the quartiles
- The five-number summary and boxplots
- Spotting suspected outliers and the modified boxplot
- Measuring variability: the standard deviation
- Choosing measures of center and variability
- Examples of technology
- Organizing a statistical problem

Measuring center: the mean

The most common measure of center is the arithmetic average, or **mean.**

- To find the **mean,** $\bar{x}$ (pronounced "x-bar"), of a set of observations, add their values and divide by the number of observations. If the *n* observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is: $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- or, in more compact notation: $\bar{x} = \frac{1}{n}\sum x_i$
- Wait a second what is this ??? What is this notation ?

Example

- Here are the travel times in minutes of 10 randomly chosen Salt Lake City workers: 10, 30, 5, 25, 40, 20, 10, 15, 30, 20
- What is the mean ?

Problem

Five men in a room have a mean height of 70 inches. A tall man, 80 inches, enters the room. Now the mean height is:

a) 500 ÷ 6 inches.
b) 350 ÷ 6 inches.
c) 430 ÷ 6 inches.
d) 430 ÷ 5 inches.

Median (*M*)

- The **median**, ***M***, is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.
- At least half of the ***ordered*** values are less than or equal to the median value.
- At least half of the ***ordered*** values are greater than or equal to the median value.
- Location of the median: $L(M) = (n+1)/2$, where n = sample size.
- Example: If 25 data values are recorded, the Median would be the $(25+1)/2 = 13^{th}$ ordered value.

Median

- Example 1 data: 2 4 6

  Median (M) = 4

- Example 2 data: 2 4 6 8

  Median = 5 (ave. of 4 and 6)

- Example 3 data: 6 2 4

  Median $\neq$ 2

  (**order** the values: 2 4 6 , so Median = 4)

Example

- Here are the travel times in minutes of 10 randomly chosen Salt Lake City workers: 10, 30, 5, 25, 40, 20, 10, 15, 30, 20
- What is the median?

Problem

Find the median of the following nine numbers.

43 54 55 63 67 68 69 77 85

a) 65
b) 64
c) 67
d) 64.6

Problem

Consider the following data.

43 54 55 63 67 68 69 77 85

Suppose that the last value is actually 115 instead of 85. What effect would this new maximum have on the median of the data?

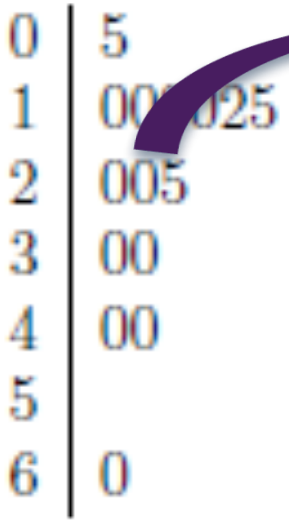
a) increase the value of the median
b) decrease the value of the median
c) no effect

Measuring center

Use the data below to calculate the mean and median of the commuting times (in minutes) of 15 randomly selected North Carolina workers.

30	20	10	40	25	10	20	60
15	40	5	30	12	10	10	

$$\bar{x} = \frac{30 + 20 + 10 + 40 + \ldots + 10 + 10}{15} = 22.5 \; minutes$$



Key: 1|5 represents a North Carolina worker who reported a 15-minute travel time to work.

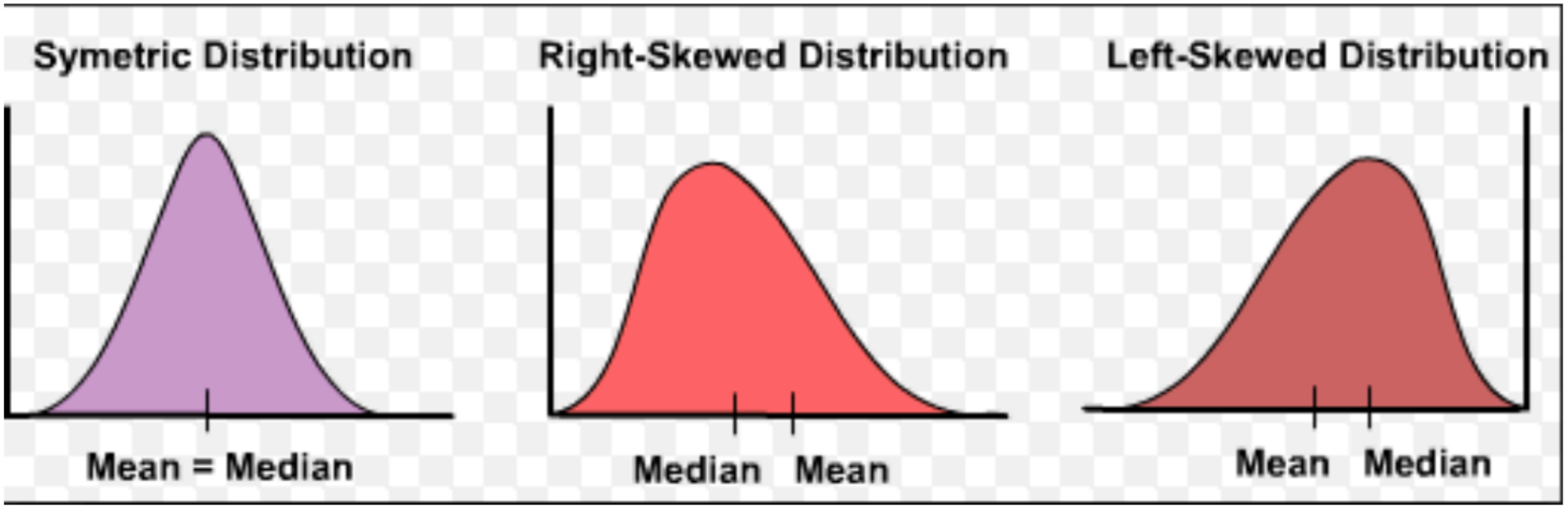
$$M = 20 \; minutes$$

Comparing the mean and the median

The mean and the median measure center in different ways, and both are useful.

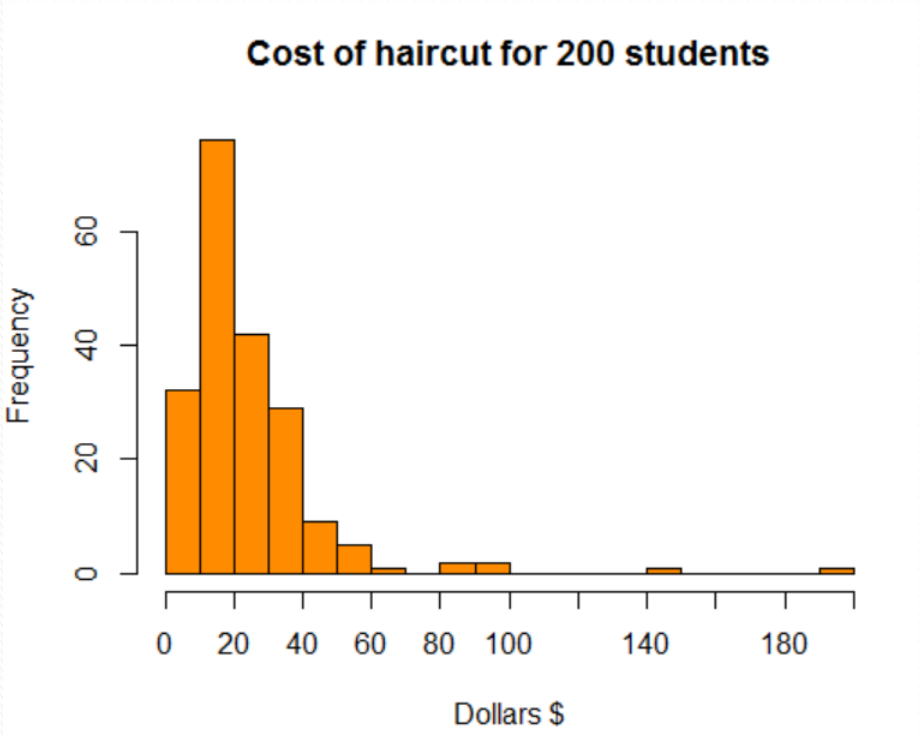
- The mean and the median of a roughly symmetric distribution are close together.
- If the distribution is exactly symmetric, the mean and the median are exactly the same.
- In a skewed distribution, the mean is usually farther out in the long tail than is the median.

Comparing the mean and the median



Problem

Based on the distribution of the data, approximately which values represent the mean and median cost of a haircut?



a) mean = \$19, median = \$24
b) mean = \$24, median = \$19

Question

A recent newspaper article in California said that the ***median*** price of single-family homes sold in the past year in the local area was $136,000 and the ***mean*** price was $149,160. Which do you think is more useful to someone considering the purchase of a home, the median or the mean?

Answer

Both! Average is affected by outliers while median is not. For example, if one house is extremely expensive, then the average will rise. The median would ignore that outlier.

Spread, or Variability

- If all values are the same, then they all equal to the mean. There is no variability.
- Variability exists when some values are different from (above or below) the mean.
- We will discuss the following measures of spread: range, quartiles, variance, and standard deviation

Range

- One way to measure spread is to give the smallest (*minimum*) and largest (*maximum*) values in the data set;

$$\text{Range} = \max - \min$$

- The range is strongly affected by outliers

(e.g. one house is extremely expensive and the rest all have the same price. The range is large while there is little variability!)

Measuring variability: quartiles

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and *a measure of spread*. We could look at the largest and smallest values (and we will!), but like the mean, they are (obviously) affected by extreme values—so we will examine other percentiles.

To calculate the quartiles:

- Arrange the observations in increasing order and locate the median *M*.
- The first quartile, Q_1, is the median of the observations located to the left of the median in the ordered list.
- The third quartile, Q_3, is the median of the observations located to the right of the median in the ordered list.

Quartiles

- Three numbers which divide the ordered data into four equal sized groups.
- Q_1 has 25% of the data below it.
- Q_2 has 50% of the data below it. (Median)
- Q_3 has 75% of the data below it.

Five-number summary

- The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.
- To get a quick summary of both center and spread, combine all five numbers.

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation—written in order from smallest to largest.

Minimum Q_1 M Q_3 Maximum

Weight Data: Sorted

100	124	148	170	185	215
101	125	150	170	185	220
106	127	150	172	186	260
106	128	152	175	187	
110	130	155	175	192	
110	130	157	180	194	
119	133	~~165~~	180	195	
120	135	165	180	203	
120	139	165	180	210	
123	140	170	185	212	

L(M)=(53+1)/2=27 L(Q1)=(26+1)/2=13.5

Weight Data: Quartiles

- $Q_1 = 127.5$
- $Q_2 = 165$ (Median)
- $Q_3 = 185$

Five-Number Summary

- minimum = 100
- Q_1 = 127.5
- M = 165
- Q_3 = 185
- maximum = 260

Interquartile Range (IQR)
= $Q_3 - Q_1$
= 57.5

IQR gives spread of middle 50% of the data

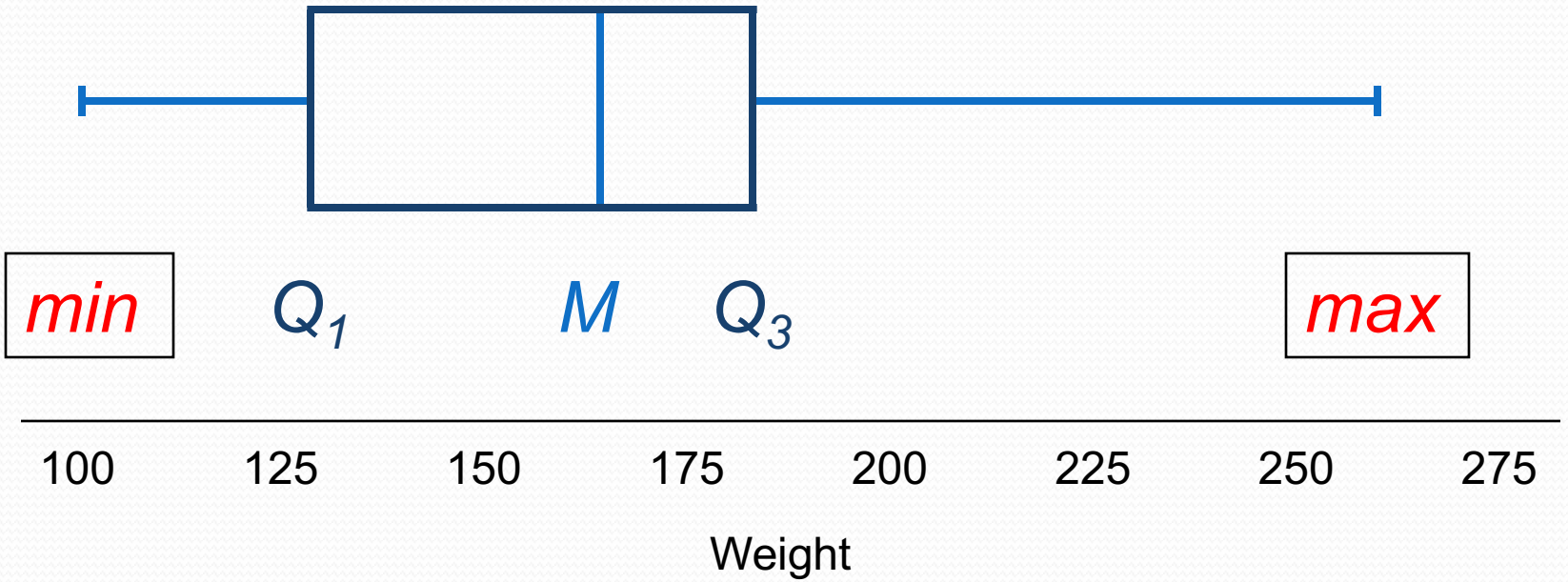
Boxplots

The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the boxplot.

How to Make a Boxplot

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

Weight Data: Boxplot

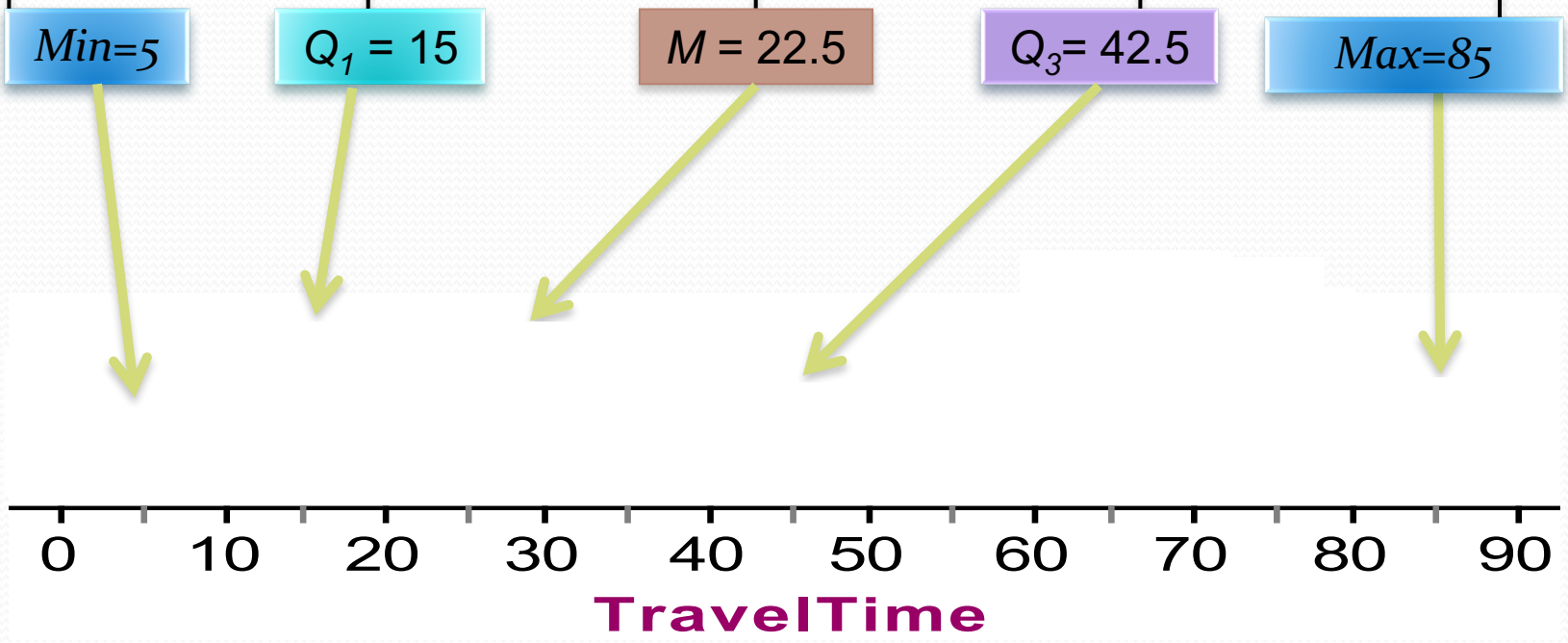


Five-number summary and boxplots

Consider a second travel times data set, these from New York. Find the five-number summary and construct a boxplot.

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85



Spotting suspected outliers and modified boxplots

- Having observed that the extremes (minimum and maximum) don't describe the spread of the majority of the data, we turn to the difference of the quartiles:

The interquartile range, or *IQR*, is the distance between the first and third quartiles

$$IQR = Q_3 - Q_1$$

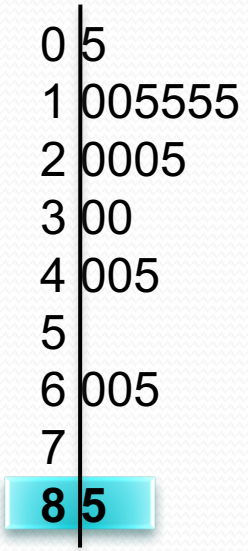
- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

The 1.5 × IQR Rule for Outliers

Call an observation a suspected outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile.

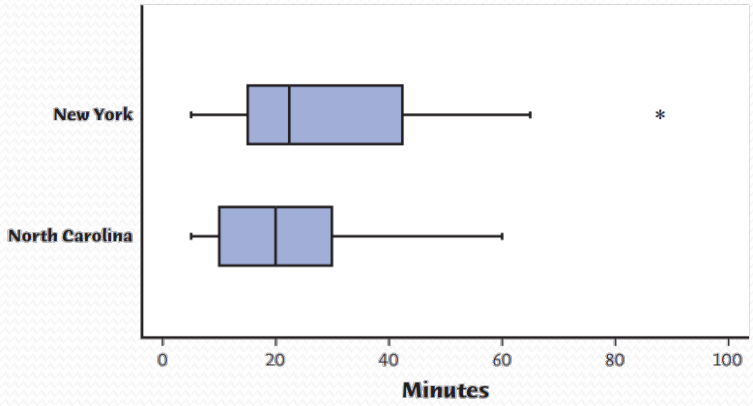
Spotting suspected outliers: example

- In the New York travel time data, Q_1 = 15 minutes, Q_3 = 42.5 minutes, and IQR = 27.5 minutes.
- For these data, 1.5 × IQR = 1.5(27.5) = 41.25
- Q_1 – 1.5 × IQR = 15 – 41.25 = –26.25
- Q_3+ 1.5 × IQR = 42.5 + 41.25 = 83.75



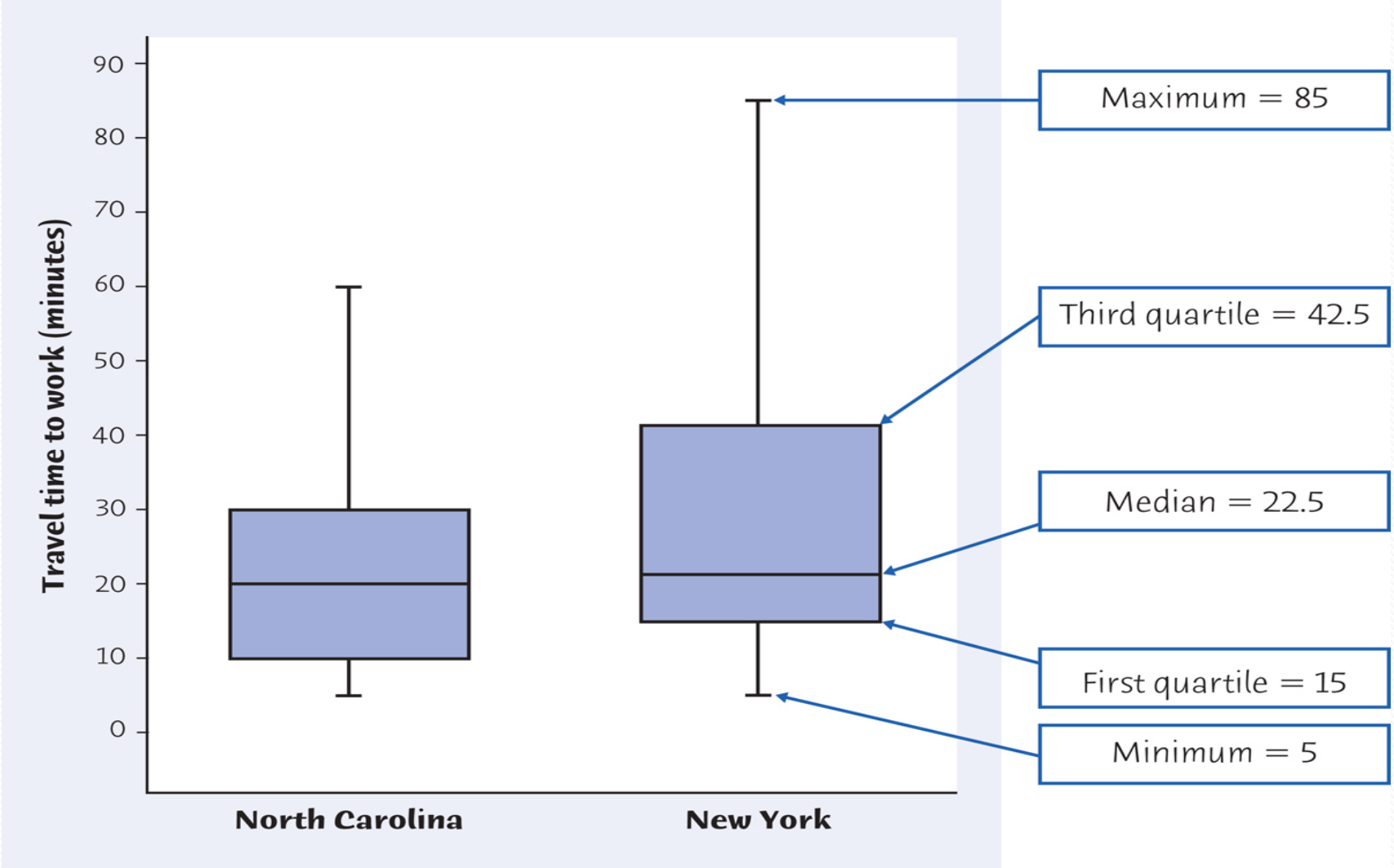
Spotting suspected outliers: example modified boxplot

- Any travel time shorter than –26.25 minutes or longer than 83.75 minutes is considered an outlier.
- So the maximum observation, 85 minutes, would be a suspected outlier.



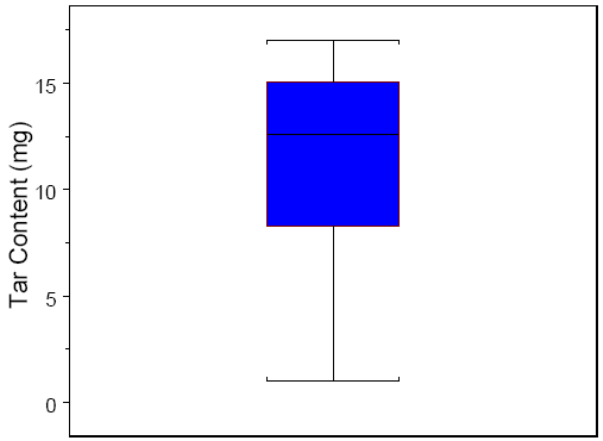
0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

- Note that North Carolina has no suspected outliers.


Travel time to work (minutes)
90
80
70
60
50
40
30
20
10
0
North Carolina
New York
Maximum = 85
Third quartile = 42.5
Median = 22.5
First quartile = 15
Minimum = 5

Problem

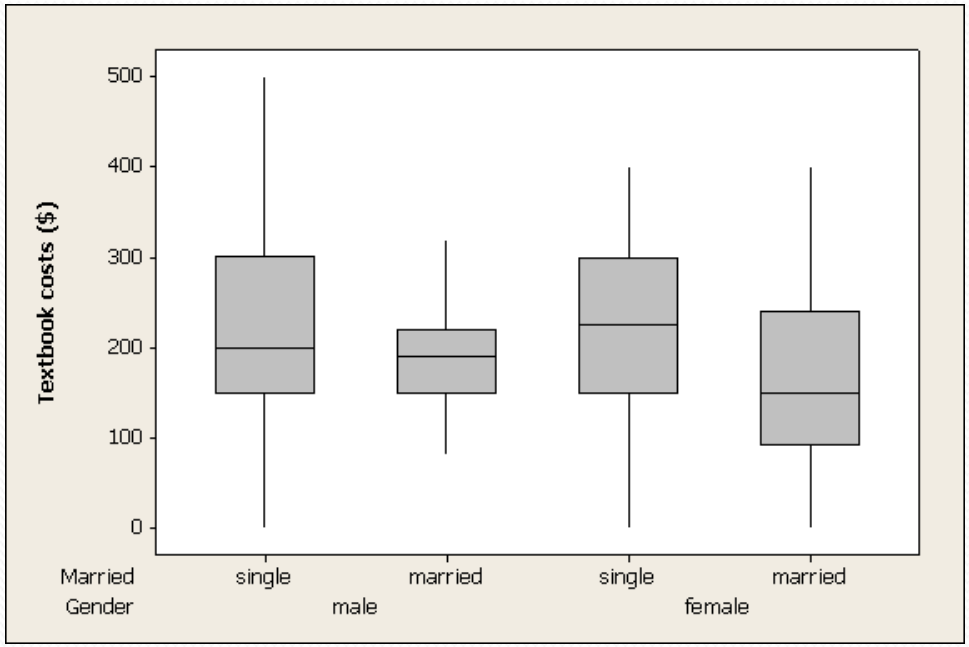
Below is a boxplot for the tar content of 25 different cigarettes. What is a plausible set of values for the five-number summary?



a) Min = 13, Q_1 = 10, Median = 12.6, Q_3 = 14, Max = 15
b) Min = 1, Q_1 = 8.5, Median = 12.6, Q_3 = 15, Max = 17
c) Min = 1, Q_1 = 8.5, Median = 11.5, Q_3 = 13, Max = 15
d) Min = 8.5, Q_1 = 10, Median = 11.5, Q_3 = 15, Max = 17

Problem

Which group has the largest spread?



a) married females
b) single females
c) married males
d) single males

Measuring variability : standard deviation

- The most common measure of spread looks at how far each observation is from the mean. This measure is called the standard deviation.

- The **variance**, s^2, of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2, x_3, ..., x_n$, is

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}$$

  Again, more briefly:

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

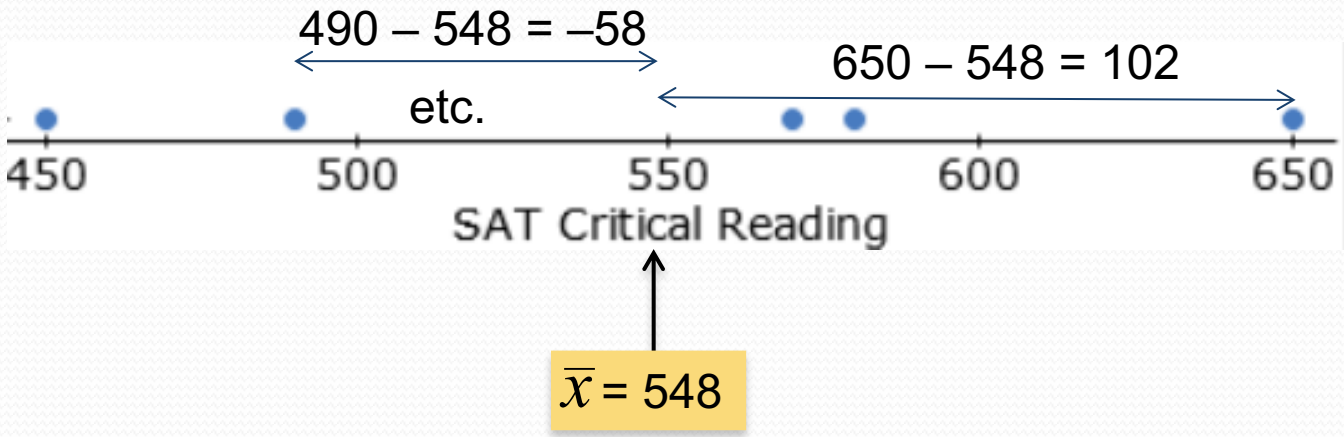
- The standard deviation, *s,* is the square root of the variance, s^2.

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

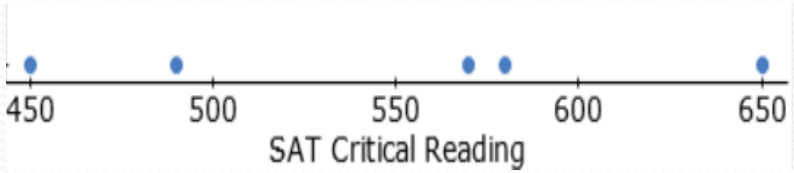
Calculating the Standard Deviation (1 of 2)

- EXAMPLE: Consider the following data on the SAT critical reading scores for 5 Georgia Southern University freshman in 2010.

1) Calculate the mean.
2) Calculate each *deviation.*
   deviation = observation – mean



Calculating the standard deviation (2 of 2)



3) Square each deviation.
4) Find the "average" squared deviation. Calculate the sum of the squared deviations divided by (n-1)…this is the **variance.**
5) Calculate the square root of the variance…this is the **standard deviation.**

x_i	$(x_i$-mean)	$(x_i$-mean)^2
650	650 – 548 = 102	$(102)^2 = 10404$
490	490 – 548 = –58	$(-58)^2 = 3364$
580	580 – 548 = 32	$(32)^2 = 1024$
450	450 – 548 = –98	$(-98)^2 = 9604$
570	570 – 548 = 22	$(22)^2 = 484$
	Sum = ?	**Sum = ?**

"Average" squared deviation = 24,880/(5 – 1) = 6220. This is the **variance.**

Standard deviation = square root of variance = $\sqrt{6220} = 78.87$

Example 2: Calculating the Standard Deviation

Metabolic rates of 7 men (cal./24hr.) :

1792 1666 1362 1614 1460 1867 1439

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7}$$

$$= \frac{11,200}{7}$$

$$= 1600$$

Example 2: Calculating the Standard Deviation

Observations	Deviations	Squared deviations
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1792	1792–1600 = 192	$(192)^2$ = 36,864
1666	1666 –1600 = 66	$(66)^2$ = 4,356
1362	1362 –1600 = -238	$(-238)^2$ = 56,644
1614	1614 –1600 = 14	$(14)^2$ = 196
1460	1460 –1600 = -140	$(-140)^2$ = 19,600
1867	1867 –1600 = 267	$(267)^2$ = 71,289
1439	1439 –1600 = -161	$(-161)^2$ = 25,921
	sum = 0	sum = 214,870

Example 2: Calculating the Standard Deviation

$$s^2 = \frac{214{,}870}{7-1} = 35{,}811.67$$

$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$

Example 3: Number of Books Read for Pleasure: Sorted

0	1	2	4	10	30
0	1	2	4	10	99
0	1	2	4	12	
0	1	3	5	13	
0	2	3	5	14	
0	2	3	5	14	
0	2	3	5	15	
0	2	4	5	15	
0	2	4	5	20	
1	2	4	6	20	

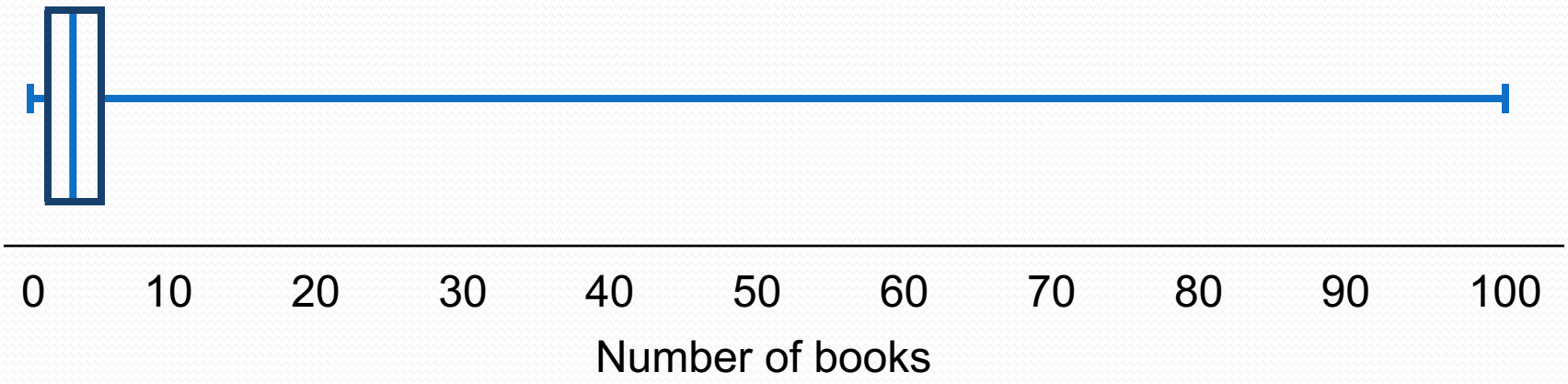
5.5+(5.5-1)x1.5=12.25

Five-Number Summary: Boxplot

Median = 3

interquartile range (iqr) = 5.5-1.0 = 4.5

range = 99-0 = 99



Mean = 7.06 s.d. = 14.43

Problem

Complete the calculations for the standard deviation of Mark McGwire's yearly home runs.

a) $3757.7/12$

b) $\sqrt{3757.7/12}$

c) $3757.7/\sqrt{12}$

d) $\sqrt{3757.7/11}$

Home Runs by Mark McGwire (1987 – 1998)			
Year	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1987	49	11.167	124.70
1988	32	-5.833	34.02
1989	33	-4.833	23.36
1990	39	1.167	1.36
1991	22	-15.833	250.68
1992	42	4.167	17.36
1993	9	-28.833	831.34
1994	9	-28.833	831.34
1995	39	1.167	1.36
1996	52	14.167	200.70
1997	58	20.167	406.71
1998	70	32.167	1034.72
Sum	454	0	3757.7

Properties of *s*

- $n - 1$ is called the degrees of freedom.
- s measures variability about the mean and should be used only when the mean is chosen as the measure of center.
- s is always zero or greater than zero. $s = 0$ only when there is no variability. This happens only when all observations have the same value. Otherwise, $s > 0$.
- As the observations become more variable about their mean, s gets larger.
- s has the same units of measurement as the original observations. For example, if you measure weight in kilograms, both the mean $\bar{x}$ and the standard deviation s are also in kilograms. This is one reason to prefer s to the variance s^2, which would be in squared kilograms.
- Like the mean $\bar{x}$, s is not resistant. A few outliers can make *s* very large.

Choosing measures of center and variability

We now have a choice between two descriptions for center and variability:

- mean and standard deviation
- median and interquartile range

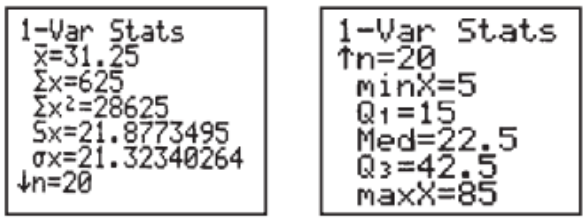
Choosing a Summary

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.

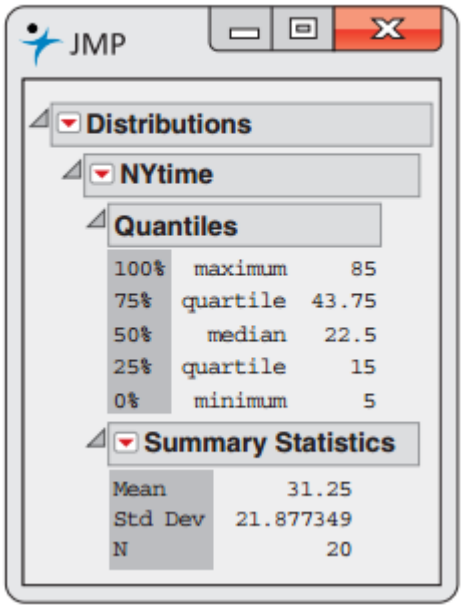
Examples of technology

- The displays below come from a Texas Instruments graphing calculator, JMP statistical software, and the Microsoft Excel spreadsheet program.
- Once you know what to look for, you can read output from any technological tool.

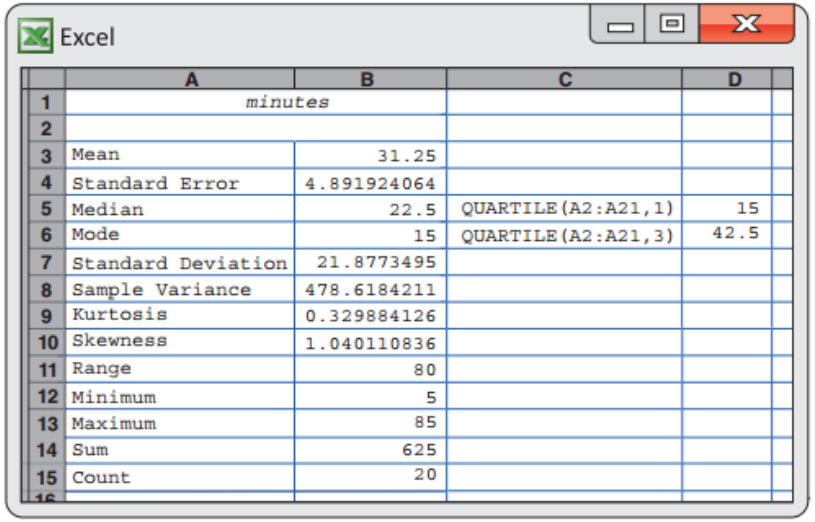
Texas Instruments Graphing Calculator



JMP Output



Microsoft Excel



	A	B	C	D
1	minutes			
2				
3	Mean	31.25		
4	Standard Error	4.891924064		
5	Median	22.5	QUARTILE(A2:A21,1)	15
6	Mode	15	QUARTILE(A2:A21,3)	42.5
7	Standard Deviation	21.8773495		
8	Sample Variance	478.6184211		
9	Kurtosis	0.329884126		
10	Skewness	1.040110836		
11	Range	80		
12	Minimum	5		
13	Maximum	85		
14	Sum	625		
15	Count	20		

Organizing a statistical problem

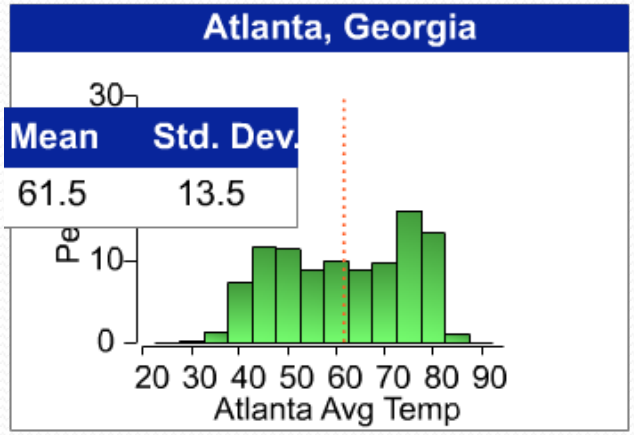
- As you learn more about statistics, you will be asked to solve more complex problems.
- Here is a four-step process you can follow.

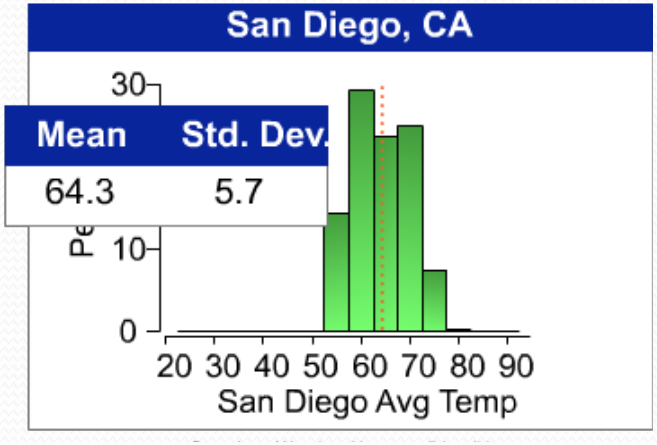
Organizing a Statistical Problem: A Four-Step Process

- **State:** What is the practical question, in the context of the real-world setting?
- **Plan:** What specific statistical operations does this problem call for?
- **Solve:** Make graphs and carry out calculations needed for the problem.
- **Conclude:** Give your practical conclusion in the setting of the real-world problem.

Problem

An individual requires a warm, stable climate. Which city is better?

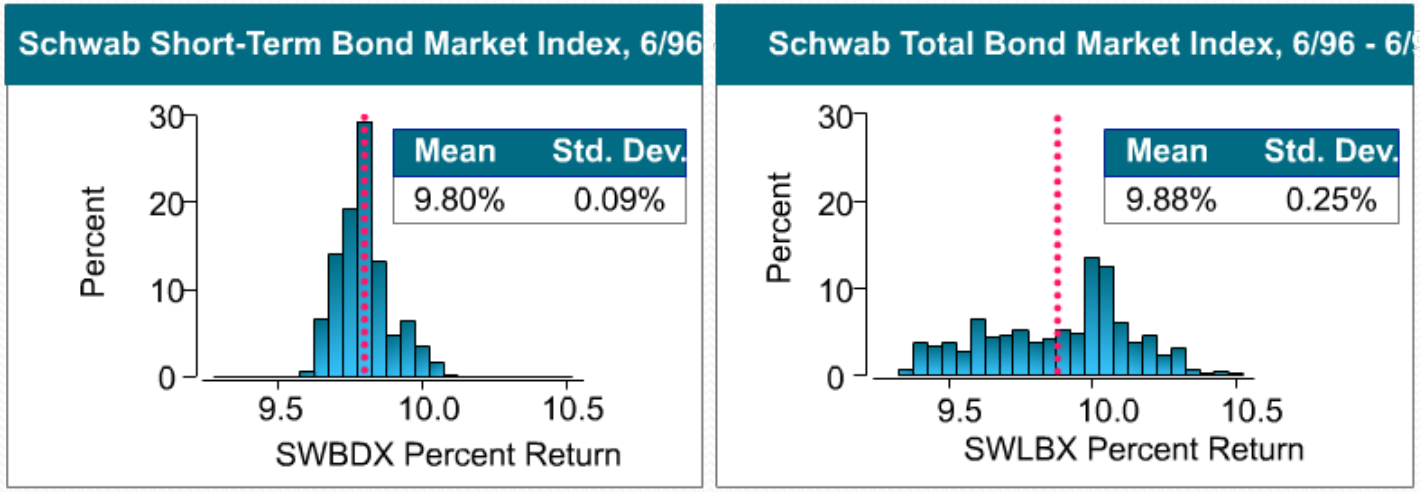




a) Atlanta, because the mean is lower.
b) San Diego, because the mean is higher.
c) Atlanta, because the standard deviation is higher.
d) San Diego, because the standard deviation is lower.

Problem

You have $100,000 to invest, and you don't like to take risks. Which mutual fund should you choose?



a) SWBDX, because the minimum is higher.
b) SWLBX, because the maximum is higher.
c) SWBDX, because the standard deviation is lower.
d) SWLBX, because the standard deviation is higher.