The Basic Practice of Statistics Ninth Edition

David S. Moore William I. Notz

Chapter 2
Describing Distributions with Numbers

Lecture Slides

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

Measuring center: the mean (or average)

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

Joy of Stats – Averages

- https://www.youtube.com/watch?v=hUGUWr-TjR8

Measuring center: the median

Because the mean cannot resist the influence of extreme observations, it is not a **resistant (or robust) measure** of center.

Another common measure of center is the **median.**

- 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.
- To find the median of a distribution:

1. Arrange all observations from smallest to largest.
2. If the number of observations *n* is odd, the median *M* is the center observation in the ordered list. If the number of observations *n* is even, the median *M* is the average of the two center observations in the ordered list.
3. You can always locate the median in the ordered list of observations by counting up (n + 1)/2 observations from the start of the list.

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.

20	35	8	70	5	15	25	30
40	35	10	12	40	15	20	

Mean:

Median (**odd** number of observations):

5 8 10 12 15 15 20 20 25 30 35 35 40 40 70

$M = 20$

Measuring center

Use the data below to calculate the mean and median of the commuting times (in minutes) of **16** randomly selected North Carolina workers.

20	35	8	70	5	15	25	30
40	35	10	12	40	15	20	28

Median (**even** number of observations):

5 8 10 12 15 15 20 20 25 28 30 35 35 40 40 70

$M = (20 + 25) / 2 = 22.5$

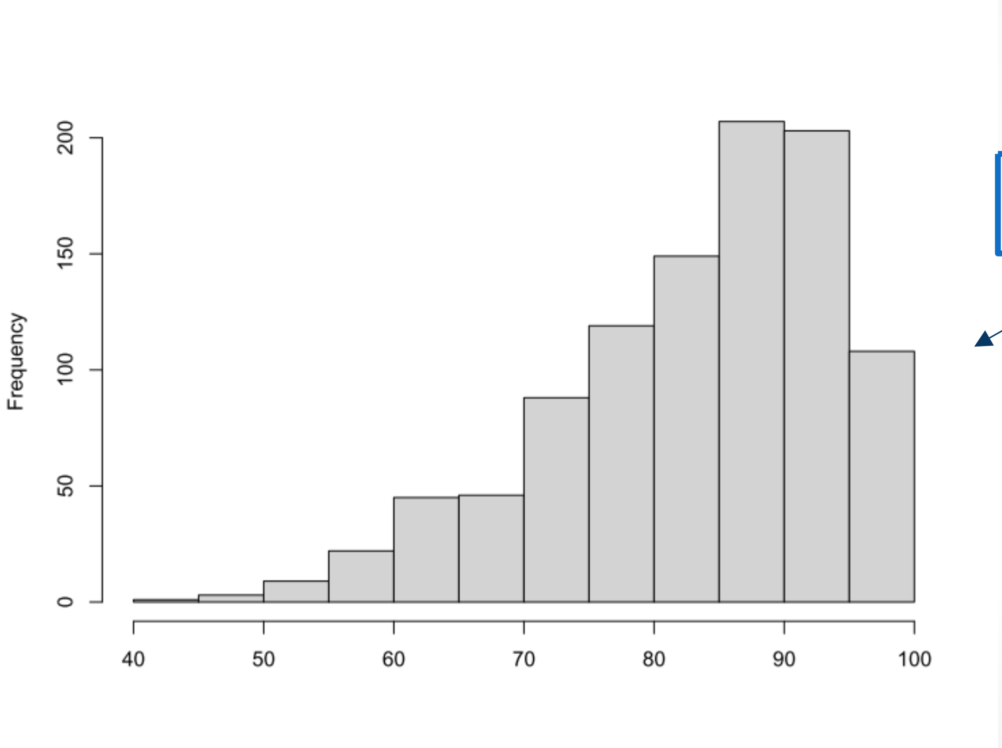
Comparing the mean and the median

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

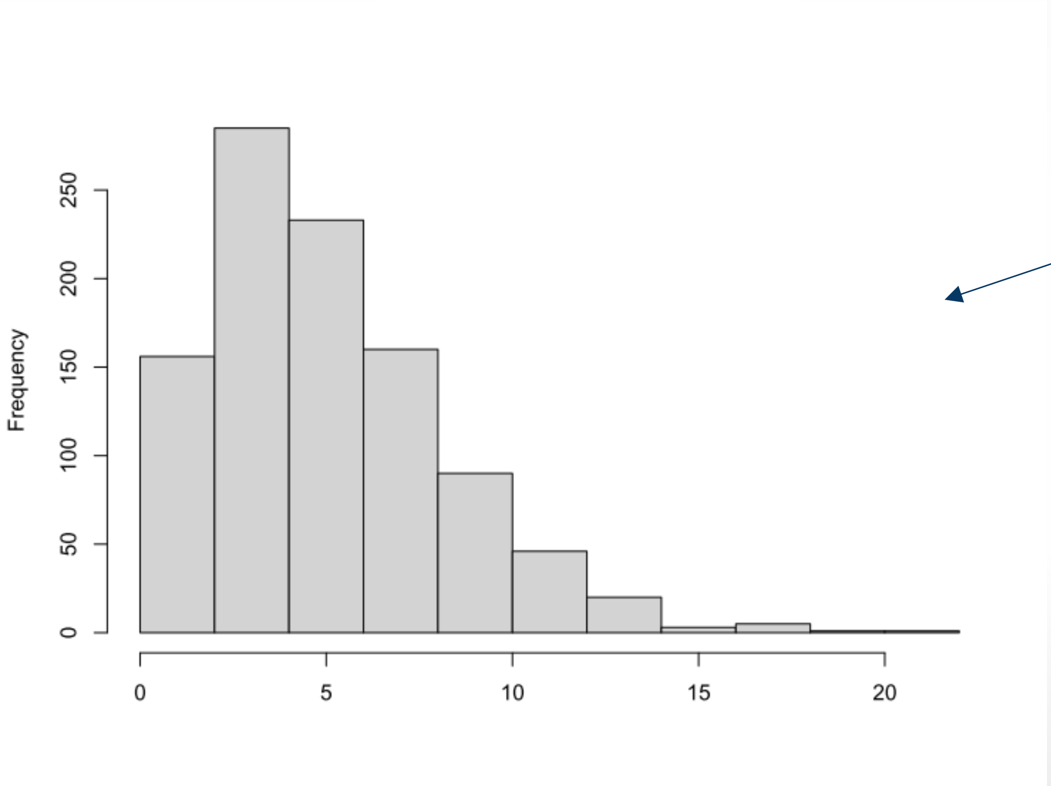
Example: skewed to the left distribution



Mean = 83.09326 < Median = 85.72406

Comparing the mean and the median

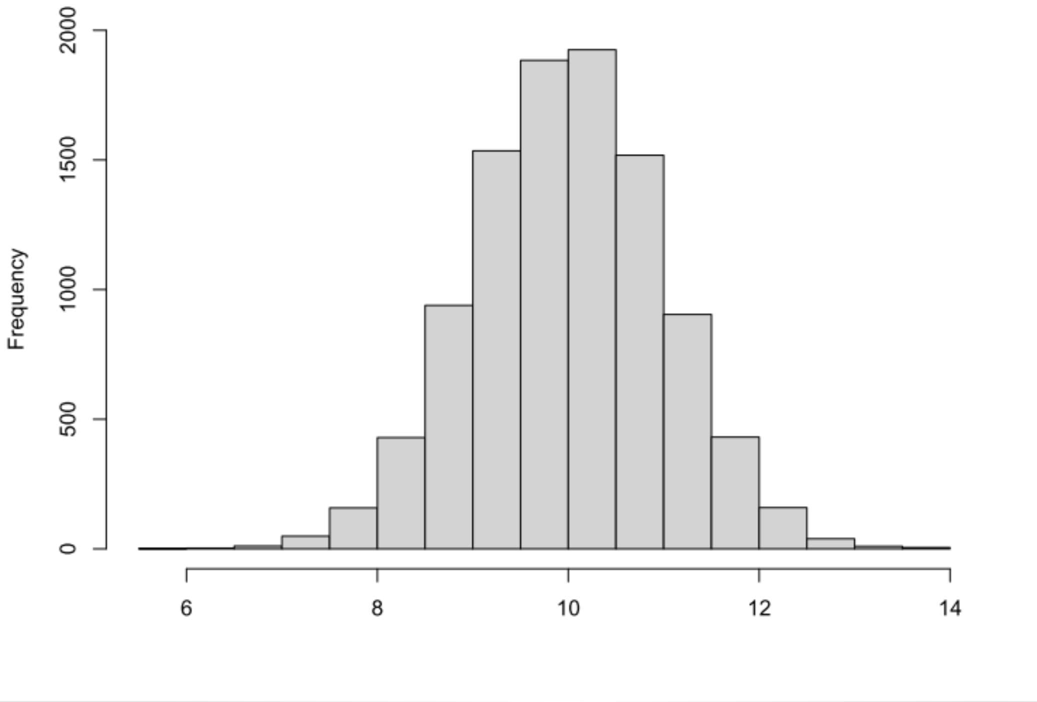
Example: skewed to the right distribution



Mean = 5.026254 > Median = 4.450805

Comparing the mean and the median

Example: symmetric distribution



Median and mean approx. 10

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 quartiles/percentiles.

Measuring variability: quartiles

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.

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

The five-number summary

Minimum = 1 Q_1 = ? M = ? Q_3 = ? Maximum=29

1, 2, 5, 9, 12, 15, 17, 21, 23, 25, 29 **n = 11** data points

Ordered observations

The five-number summary

Minimum = 1 Q_1 = ? M = 15 Q_3 = ?   Maximum=29

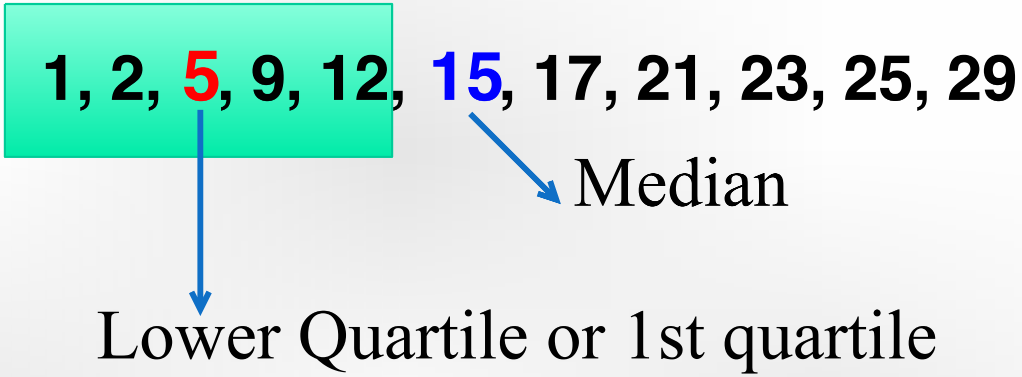
Odd number

1, 2, 5, 9, 12, 15, 17, 21, 23, 25, 29   **n = 11** data points

Median = value at position (n+1)÷2

The five-number summary

Minimum = 1 $Q_1 = 5$ $M = 15$ $Q_3 = ?$ Maximum=29



1, 2, 5, 9, 12, 15, 17, 21, 23, 25, 29

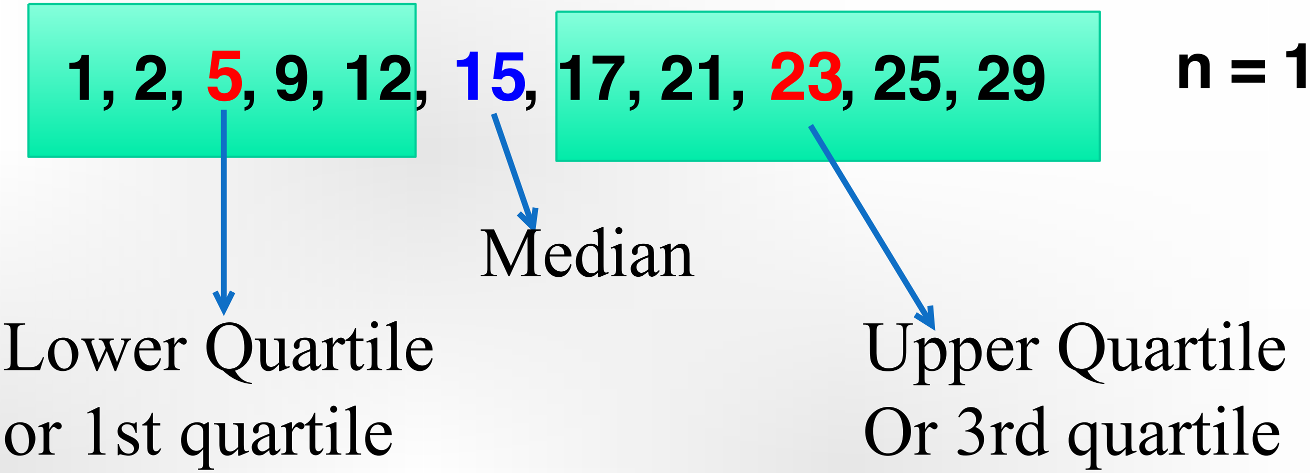
Median

Lower Quartile or 1st quartile

n = 11 data points

The five-number summary

Minimum = 1 Q_1 = 5 M = 15 Q_3 = 23   Maximum=29



1, 2, <span>5</span>, 9, 12, 15, 17, 21, 23, 25, 29

n = 11 data points

Median

Lower Quartile
or 1st quartile

Upper Quartile
Or 3rd quartile

The five-number summary

Minimum = 1 Q_1 = ? M = 13.5 Q_3 = ?   Maximum=25

1, 2, 5, 9, 12, 15, 17, 21, 23, 25

Even number

n = 10 data points

position (n+1)÷2 = 5.5, so the median is the average value between the values in positions 5 and 6 ?   **Median** = 12+15/2 = **13.**

The five-number summary

Minimum = 1 $Q_1 = 5$ $M = 13.5$ $Q_3 = ?$ Maximum=25

1, 2, 5, 9, 12, 15, 17, 21, 23, 25

n = 10 data points

position (5+1)÷2 = 3, so the lower quartile is the value of position 3 ? **Lower quartile = 5**

The five-number summary

Minimum = 1 Q_1 = 5 M = 13.5 Q_3 = 21 Maximum=25

1, 2, 5, 9, 12, 15, 17, 21, 23, 25

n = 10 data points

⍰ **Upper quartile = 21**

Example

78,95,60,93,55,84,76,92,62,83,80,90,64,75,79,32,75,64,98,73,88,61,82,68,79,78,80,85

Create a five-number summary for these 28 scores.

Example

78,95,60,93,55,84,76,92,62,83,80,90,64,75,79,32,75,64,
98,73,88,61,82,68,79,78,80,85

Create a five-number summary for these 28 scores.

32 55 60 61 62 64 64 68 73 75 75 76 78 78 79 79 80 80 82 83 84 85 88 90 92 93 95 98

Position (28+1)/2 =14.5, so take average between positions 14 and 15

Minimum = 32 Q_1 M = 78.5 Q_3 Maximum = 98

Example

78,95,60,93,55,84,76,92,62,83,80,90,64,75,79,32,75,64, 98,73,88,61,82,68,79,78,80,85

Create a five-number summary for these scores.

32 55 60 61 62 64 64 68 73 75 75 76 78 78 79 79 80 80 82 83 84 85 88 90 92 93 95 98

Position (14+1)/2 =7.5, so take average between positions 7 and 8 ? **(64+68)/2 =66**

Minimum = 32 Q_1 = 66 M = 78.5 Q_3 Maximum = 98

Example

78,95,60,93,55,84,76,92,62,83,80,90,64,75,79,32,75,64,
98,73,88,61,82,68,79,78,80,85

Create a five-number summary for these scores.

32 55 60 61 62 64 64 68 73 75 75 76 78 78 79 79 80 80 82 83 84 85 88 90 92 93 95 98

Minimum = 32 Q_1= 66 M = 78.5 Q_3 = 84.5 Maximum = 98

Spotting suspected outliers

- 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.

Boxplots (modified boxplots*)

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

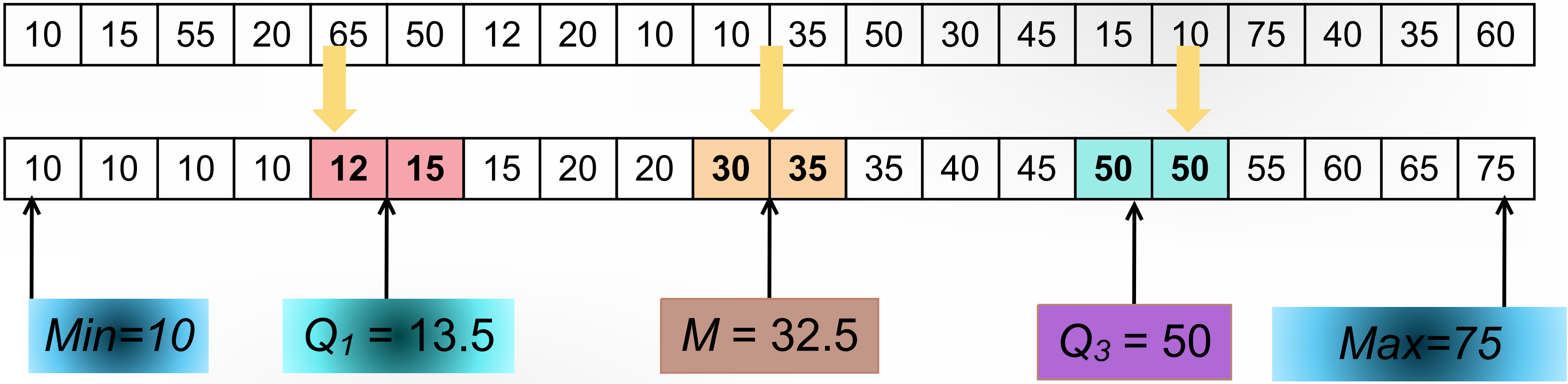
How to Make a Boxplot:

1. A central box spans the quartiles Q_1 and Q_3.
2. A line in the box marks the median M.
3. Compute IQR (interquartile range) = distance between quartiles.
4. Compute 1.5 x IQR; outlier is any value more than this distance from closest quartile.
5. Draw line (whisker) from each end of box extending to farthest data value that is not an outlier. (If no outlier, then to min and max.)
6. Draw asterisks to indicate the outliers.

**In the textbook, they present boxplots and then modified boxplots, here I am focusing only on the modified version, which is the one used by most software packages.*

Five-number summary and boxplots (1 of 2)

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



10	15	55	20	65	50	12	20	10	10	35	50	30	45	15	10	75	40	35	60

10	10	10	10	**12**	**15**	15	20	20	**30**	**35**	35	40	45	**50**	**50**	55	60	65	75

Min=*10*

Q_1 = 13.5

M = 32.5

Q_3 = 50

Max=*75*

Spotting suspected outliers: example

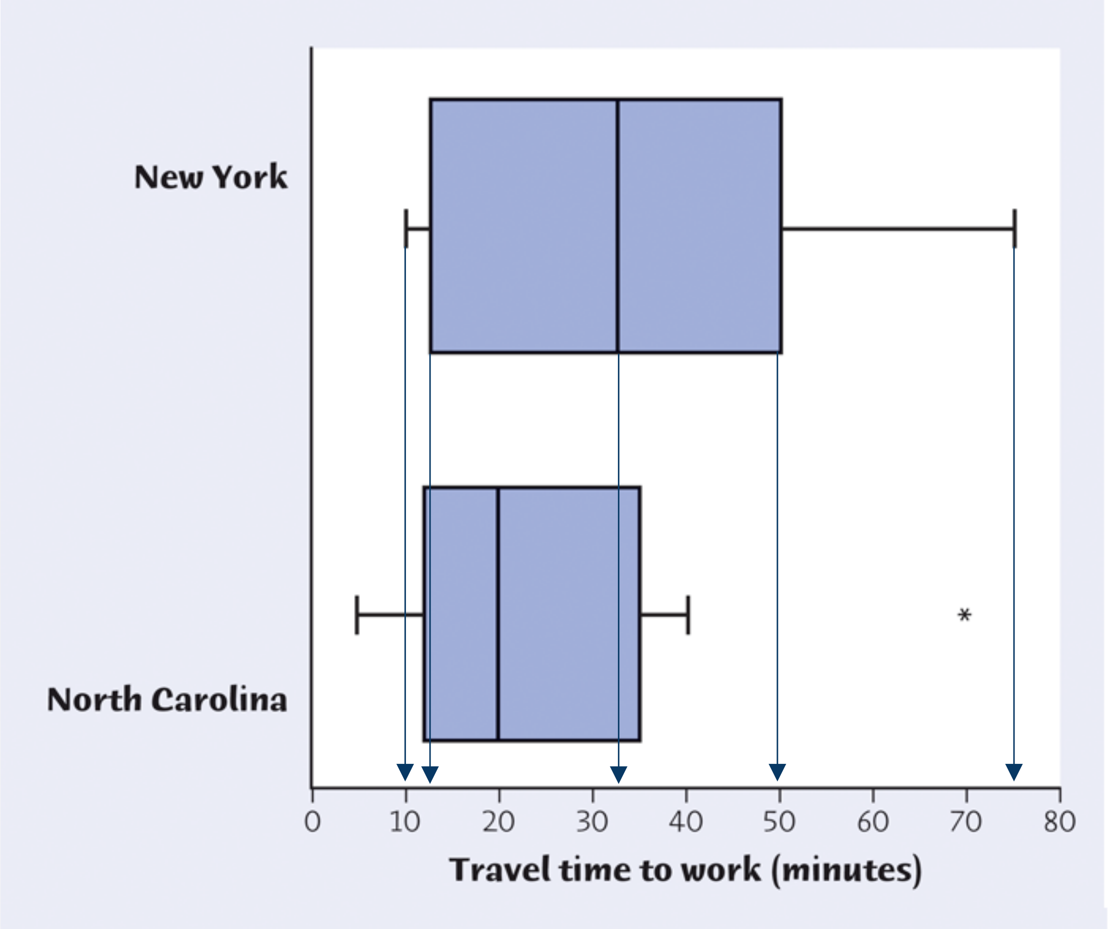
- In the New York travel time data, Q_1 = 13.5 minutes, Q_3 = 50 minutes, and IQR = 36.5 minutes.
- For these data, 1.5 ⍰ IQR = 1.5(36.5) = 54.75
- Q_1 – 1.5 ⍰ IQR = 13.5 – 54.75 = –41.25
- Q_3+ 1.5 ⍰ IQR = 50 + 54.75 = 104.75
- Any travel time shorter than –41.25 minutes or longer than 104.75 minutes is considered an outlier.
- So none of the observations would be a suspected outliers.

0	
1	0000255
2	00
3	055
4	05
5	005
6	05
7	5

Five-number summary and boxplots (2 of 2)

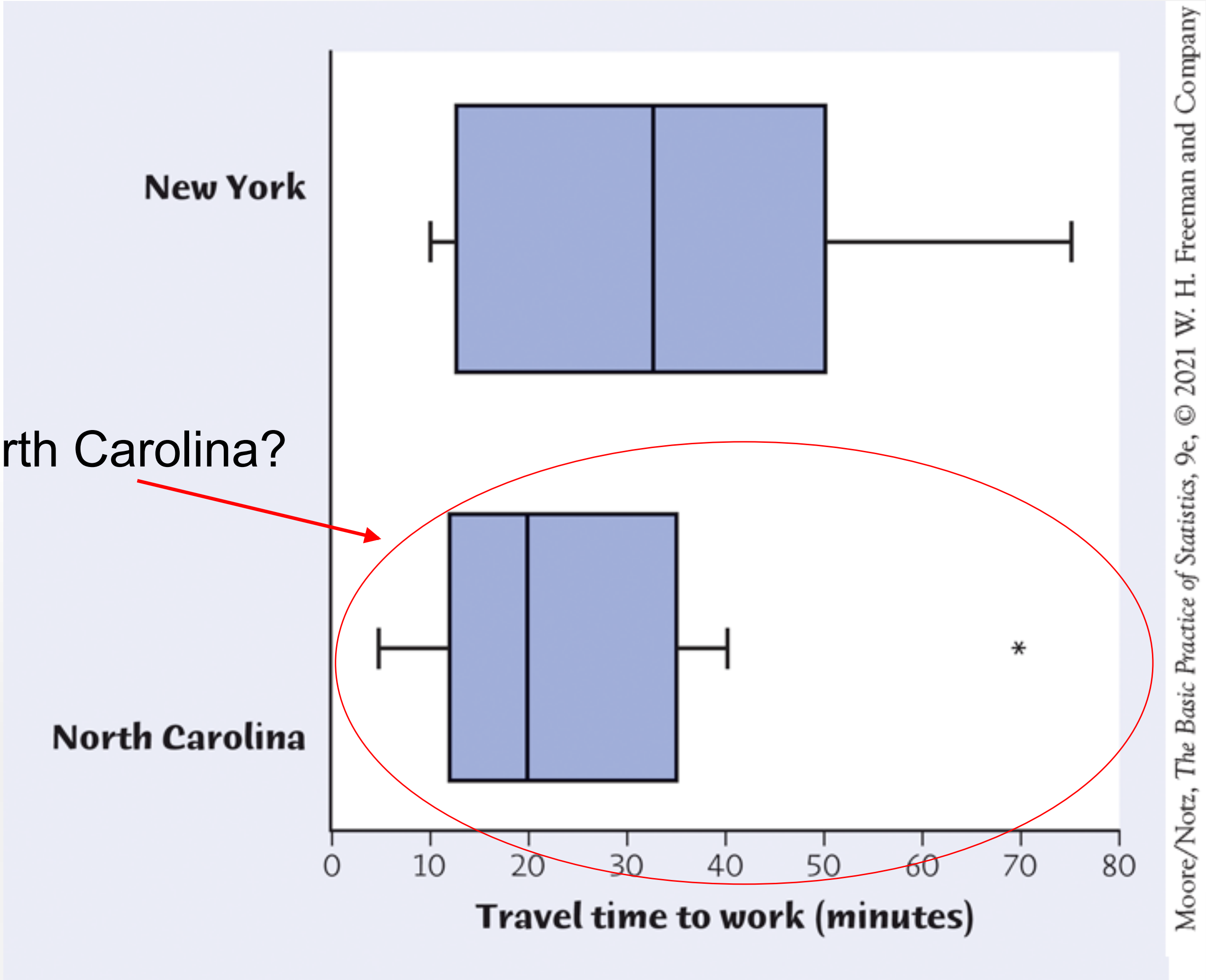
Consider a second travel times data set, these from **New York**. Find the five-number summary and construct a boxplot.

Min=10 Q_1 = 13.5 *M* = 32.5 Q_3 = 50 *Max*=75



Moore/Notz, *The Basic Practice of Statistics*, 9e, © 2021 W. H. Freeman and Company

What about North Carolina?



Moore/Notz, *The Basic Practice of Statistics*, 9e, © 2021 W. H. Freeman and Company

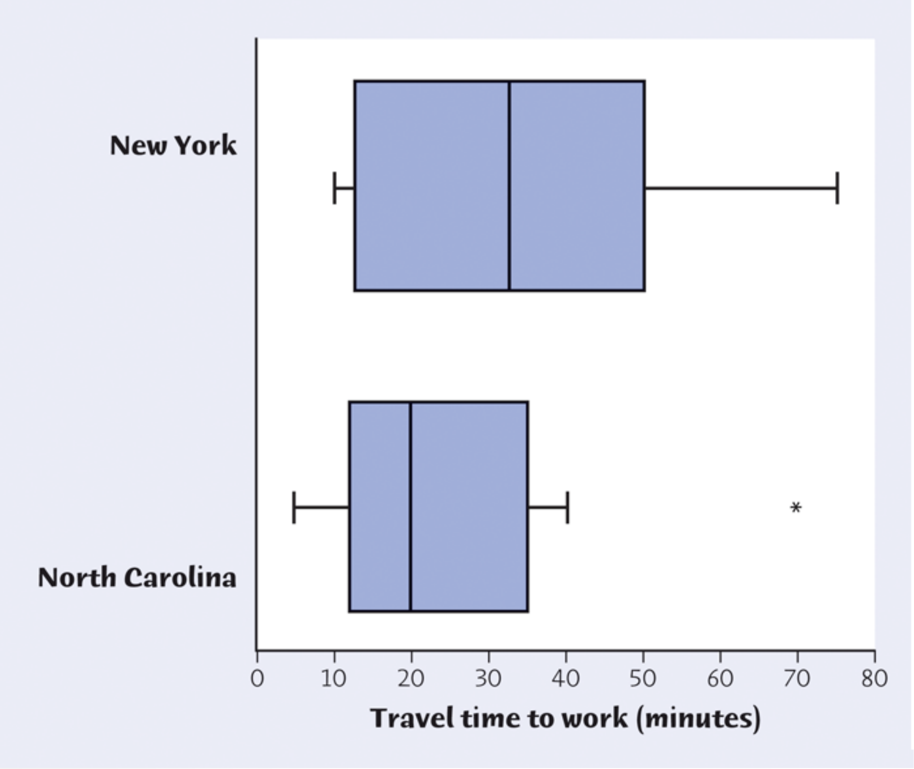
Spotting suspected outliers: example

- Note that North Carolina has a suspected outlier and that the upper whisker extends only to the greatest observation smaller than the 69.5 minutes (= 35 + 1.5*23 = Q3 + 1.5 x IQR) cutoff, *not to the cutoff itself*.

- The suspected outlier corresponds to only one observation of 70 minutes.

- The next-longest travel times are 40 minutes.

- See data on page 48 of the textbook.



Moore/Notz, *The Basic Practice of Statistics*, 9e, © 2021 W. H. Freeman and Company

Measuring variability: standard deviation

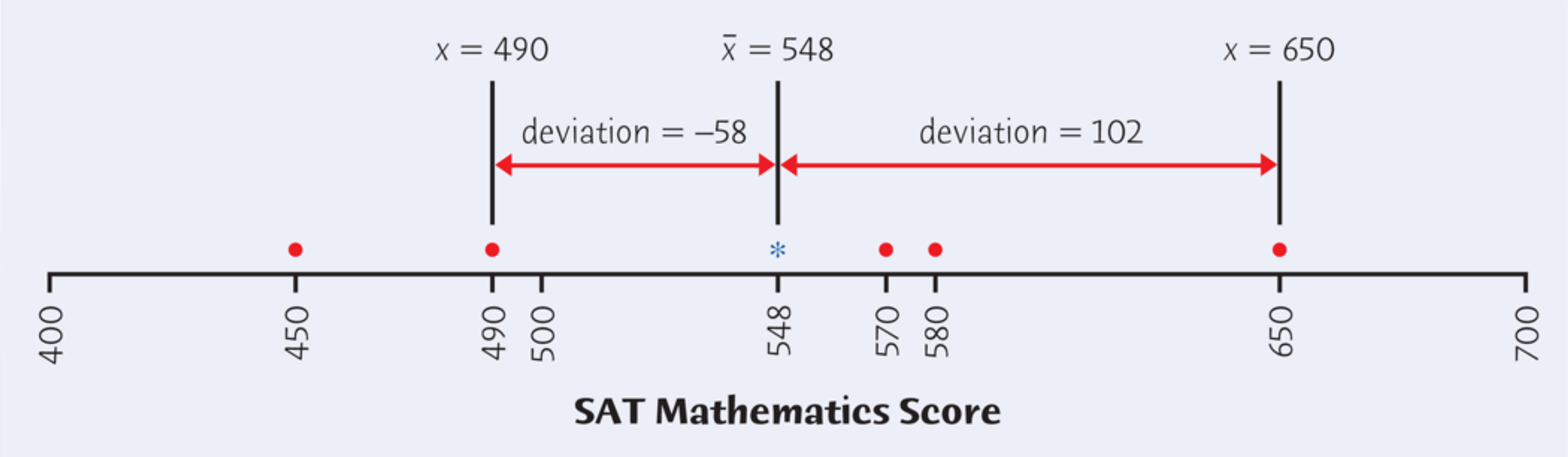
The most common measure of spread looks at how far each observation is from the mean. This measure is called the standard deviation.

Calculating the Standard Deviation (1 of 2)

- EXAMPLE: Consider the following data on the SAT mathematics scores for 5 Georgia Southern University freshman in 2015.

1) Calculate the mean.
2) Calculate each *deviation.*

   deviation = observation – mean



Moore/Notz, *The Basic Practice of Statistics,* 9e, © 2021 W. H. Freeman and Company

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**
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	$(22)^2 = 484$
650	650 – 548 = 102	$(102)^2 = 10404$
		Sum = 24,880

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

Standard deviation = square root of variance =

Properties of *s*

- is called the degrees of freedom.

- measures variability about the mean and should be used only when the mean is chosen as the measure of center.

- is always zero or greater than zero. only when there is no variability. This happens only when all observations have the same value. Otherwise, .

- As the observations become more variable about their mean, gets larger.

Properties of *s* (continued)

- has the same units of measurement as the original observations. For example, if you measure weight in kilograms, both the mean and the standard deviation are also in kilograms. This is one reason to prefer to the variance , which would be in squared kilograms.

- Like the mean , is not resistant (robust). 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:

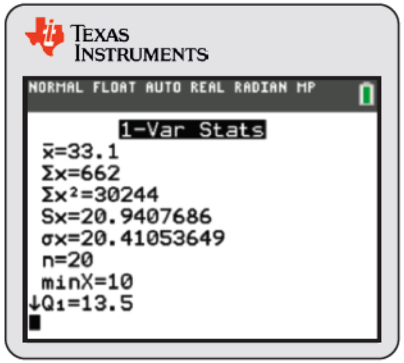
- mean and standard deviation
- median and interquartile range

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 and *s* only for reasonably symmetric distributions that are free of outliers.

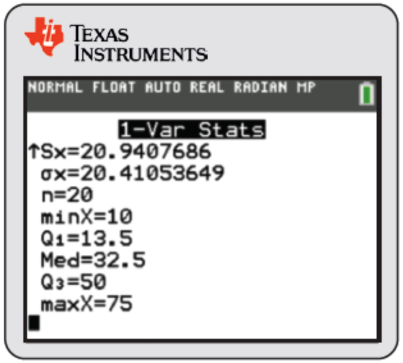
Examples of technology

- 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.



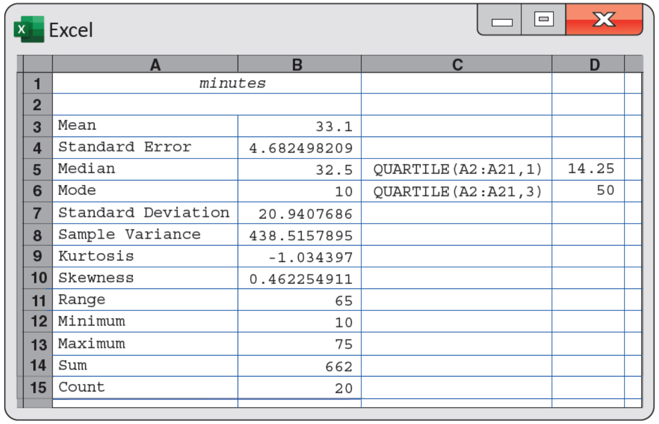
Texas Instruments

```
NORMAL FLOAT AUTO REAL RADIAN MP
          1-Var Stats
 x̄=33.1
 Σx=662
 Σx²=30244
 Sx=20.9407686
 σx=20.41053649
 n=20
 minX=10
↓Q₁=13.5
```



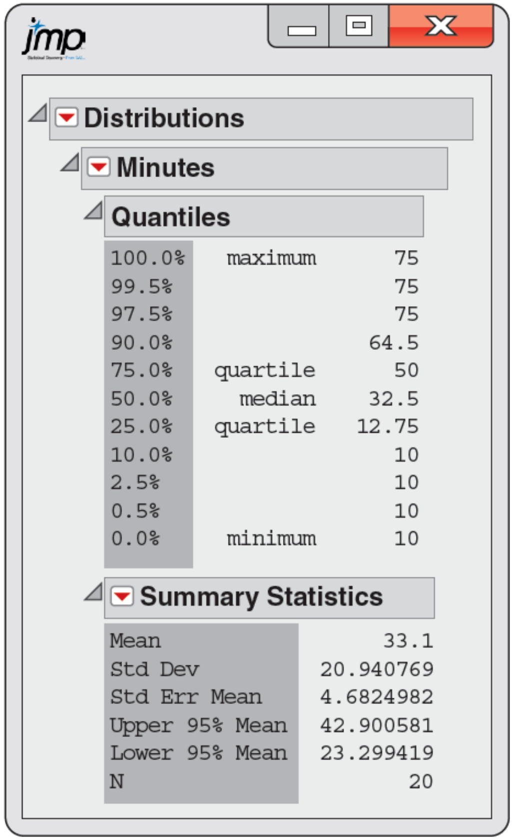
Texas Instruments

```
NORMAL FLOAT AUTO REAL RADIAN MP
          1-Var Stats
↑Sx=20.9407686
 σx=20.41053649
 n=20
 minX=10
 Q₁=13.5
 Med=32.5
 Q₃=50
 maxX=75
```



Excel

	A	B	C	D
1	*minutes*			
2				
3	Mean	33.1		
4	Standard Error	4.682498209		
5	Median	32.5	QUARTILE(A2:A21,1)	14.25
6	Mode	10	QUARTILE(A2:A21,3)	50
7	Standard Deviation	20.9407686		
8	Sample Variance	438.5157895		
9	Kurtosis	-1.034397		
10	Skewness	0.462254911		
11	Range	65		
12	Minimum	10		
13	Maximum	75		
14	Sum	662		
15	Count	20		



jmp

Distributions

Minutes

Quantiles

```
100.0%   maximum     75
 99.5%               75
 97.5%               75
 90.0%             64.5
 75.0%  quartile     50
 50.0%    median   32.5
 25.0%  quartile  12.75
 10.0%               10
  2.5%               10
  0.5%               10
  0.0%   minimum     10
```

Summary Statistics

```
Mean                 33.1
Std Dev         20.940769
Std Err Mean    4.6824982
Upper 95% Mean  42.900581
Lower 95% Mean  23.299419
N                      20
```

Organizing a statistical problem

- As you learn more about statistics and data science, 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.