


Data Displays

Chapter 2


Data Distribution

- ▶ The distribution of a variable tells us what values it takes and how often it take these values.
 - ▶ Can be visualized as the shape of a table or graph.
 - ▶ Distributions enable us to examine the variation of the data.
- 

Descriptive Statistics

- ▶ Key Characteristics of a distribution “SOCS”
 - Shape (Distribution)
 - Outliers
 - Center
 - Spread (Variation)
- ▶ When examining a distribution visually, we want to mainly focus on:
 - Shape (including deviations of the overall pattern)
 - Any unusual values (Outliers)
- ▶ We describe numerically:
 - Typical Value (center)
 - Variability (spread)

Visualizing Data: Goals

- ▶ Using a picture to display the data will help us see patterns.
 - ▶ Different visual representations capture different aspects in the data
 - ▶ Some methods will be more appropriate for certain types of data
 - ▶ The picture must:
 - Record the data values.
 - Indicate the frequency (count) of the data values.
- 

Basic Data Visualization tools

- ▶ Stem-and-leaf Plots
- ▶ Boxplots
- ▶ Histograms
- ▶ Time Series Plots

- ▶ Later in the semester we will see:
 - Q-Q (Normal) plots
 - Scatter plots

Stem and leaf Diagrams


- ▶ Steps to construct a stem-and-leaf diagram:
 - 1) Divide each number (x_i) into two parts: a **stem**, consisting of the leading digits, and a **leaf**, consisting of the remaining digit, e.g., 123.
 - 2) List the stem values in a vertical column (no skips).
 - 3) Record the leaf for each observation beside its stem.
 - 4) Write the units for the stems and leaves on the display.

Example: For the number 173, the stem would be “17” and the leaf would be “3”



Example

Instructions:

- ▶ Open or paste the data file
 - ▶ Click the Graph box and select Stem Plot.
 - ▶ Click on “High Temp. in F” to enter it into the graph box to the right.
 - ▶ Edit the title and labels if needed.
 - ▶ Finally, click Compute!
- 

Stem Plot

Stem-and-Leaf Display: High Temp in F

Stem-and-leaf of High Temp in F N = 50

3	10	004
11	10	56677899
(16)	11	0000122223334444
23	11	5556777888889
10	12	0000112
3	12	58
1	13	4

Stem Plot Information

- ▶ Notice the counts on the left side these represent the count from either end to the middle value.
- ▶ When the observed values have too many digits, **round** the numbers before making a stem plot.
- ▶ When plotting a moderate number of observations, you can **split** each stem.
 - With our Temperature example, Minitab uses this method.

Splitting stems

- ▶ The purpose of the stem-and-leaf is to describe the data distribution graphically.
- ▶ If the data are too clustered, we can split and have multiple stems, thereby increasing the number of stems.
 - Split 2 for 1:
 - Lower stem for leaves 0, 1, 2, 3, 4
 - Upper stem for leaves 5, 6, 7, 8, 9

Stem Plot Advantages / Disadvantages

- ▶ *Advantages:*
 - Quick way to sort data.
 - Can recover the actual data.
- ▶ *Disadvantages:*
 - Best used with small data sets.
 - **Histogram** more flexible in choice of stems/bins.

Measure of Center: Median

- ▶ Median – Middle value, halfway point, or the value exceeded by half the readings.
 - If the sample size, n , is odd, the median is the middle ordered data value
 - If n is even, the median is the mean of the two middle ordered data values

Median Calculation

- ▶ Use the following formula to determine the median's location.

$$L(M) = \frac{n + 1}{2}$$

- ▶ $L(M)$ stands for the location of the median, it is not the median itself.
- ▶ In our Example $n=50$ (even):
 - $L(M) \frac{50+1}{2} = 25.5$
 - Locate observations 25&26, then split the difference:
 - $M = 114$

Quartiles

- ▶ The three quartiles partition the data into four equally sized counts or segments.
 - 25% of the data is less than q_1 .
 - 50% of the data is less than q_2 , aka the median.
 - 75% of the data is less than q_3 .

Two Ways to find Quartiles:

1. Treat as “Medians” of respective halves (OK)
2. Calculated as $Index = f(n+1)$ (More precise)
 - $Index (I)$ is the I^{th} item (interpolated) of the sorted data list.
 - f is the fraction associated with the quartile.
 - n is the sample size.

Quartiles Example

- ▶ In our High Temps Example. The $M=114$ and divided the data into two halves of $n=25$.

$$L(M) \frac{25 + 1}{2} = 13$$

- ▶ Consider another example with $n=80$

- PSI Data in Canvas
- Treat as “Medians”
- More precise method

=>

$n = 80$		Value of indexed item		
f	$Index$	l^{th}	$(l+1)^{th}$	quartile
0.25	20.25	143	144	143.25
0.50	40.50	160	163	161.50
0.75	60.75	181	181	181.00

Five Number Summary

- ▶ This summary displays the distribution of the data using specific values.

Minimum Q_1 Median Q_3 Maximum

- ▶ Can be easily obtained using Minitab...
 - Stat → Basic Statistics → Display descriptive Statistics
 - Click the Statistics button to tell Minitab what you want

High Temperature Example

Use the “High Temperature” Data provided in Canvas

<u>Minimum</u>	<u>Q1</u>	<u>Median</u>	<u>Q3</u>	<u>Maximum</u>
100	110	114	118	134

- ▶ Even though the quartiles would be what we expected in this example, the quartiles may be calculated a bit differently in certain places.

Measure of Spread: Interquartile Range

- ▶ The interquartile range (*IQR*) is the difference between the third and first quartiles.

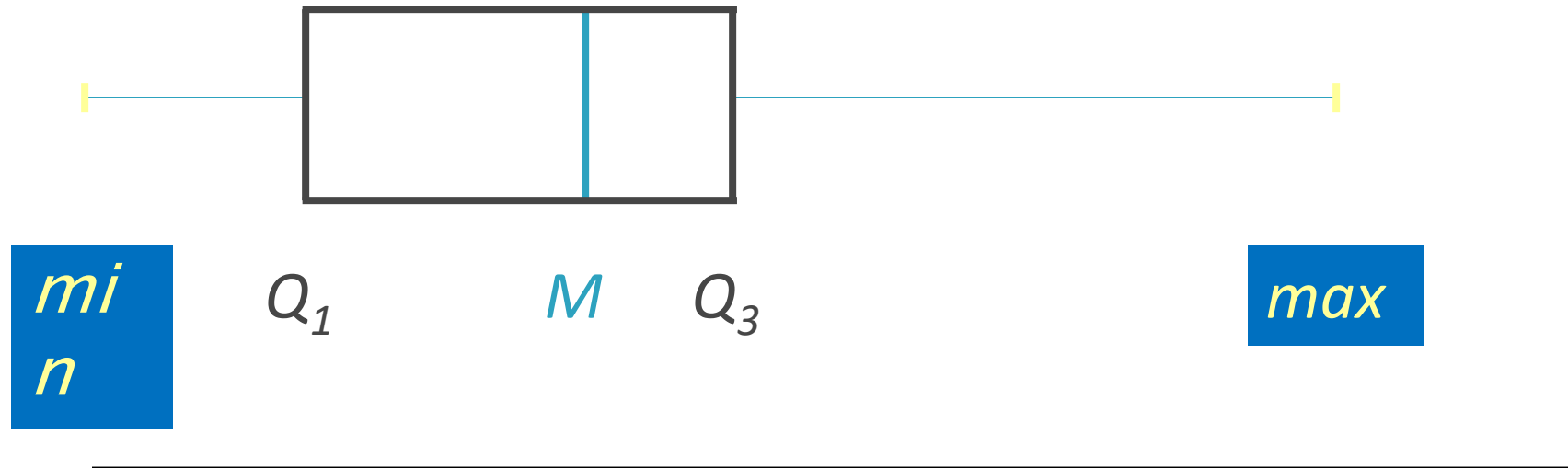
$$IQR = Q_3 - Q_1$$

- ▶ The IQR is measure of variation that is not affected by skewness and outliers.
- ▶ This is what we will use to describe skew distributions along with the Median

Boxplot

- ▶ A **boxplot** is a graphical representation of the five-number summary
- ▶ Central box spans Q_1 and Q_3 .
- ▶ A line in the box marks the median M (Q_2).
- ▶ Lines (or whiskers) extend from the box out to the minimum and maximum when there are no outliers in the data set.

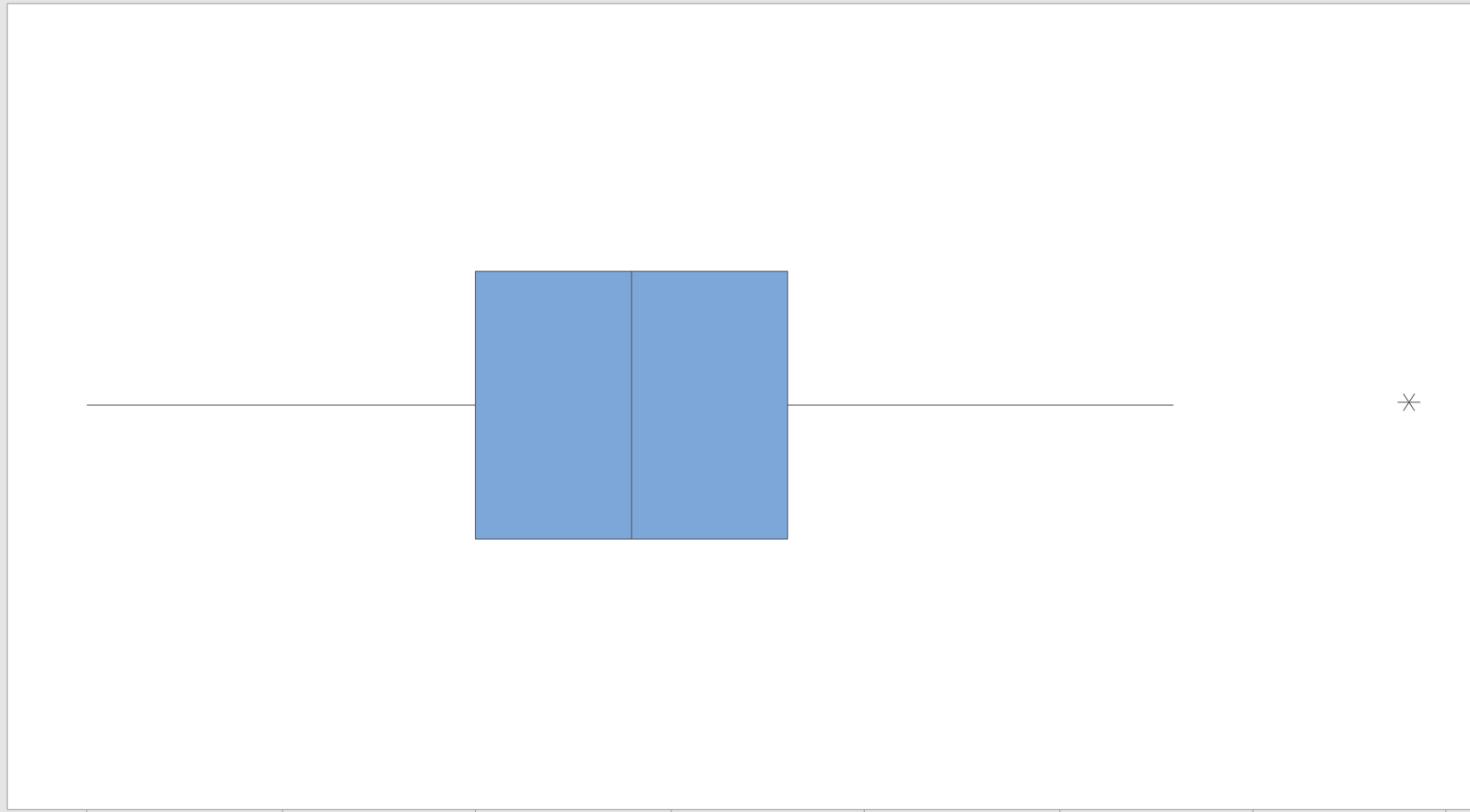
Boxplot of Data



Boxplot for High Temperatures

- ▶ Graph → Boxplot.
- ▶ Click High Temperature into the ‘graph variables’ box.
- ▶ Click the “Scale button” and the ‘Transpose value and category’ box. (This is optional but I prefer horizontal Boxplots...)


Boxplot of High Temp in F



100 105 110 115 120 125 130 135

High Temp in F

Boxplots with Outliers

- ▶ Minitab can create Boxplots for us.
 - ▶ If Minitab believes there is an **outlier** present, it will star it on the graph.
 - ▶ Then, the whiskers will only extend to the next highest or lowest value in the data set not calculated as an outlier.
- 

Outliers

Outliers are extreme observations in the data. They are values that are significantly too high or too low, based on the spread of the data.

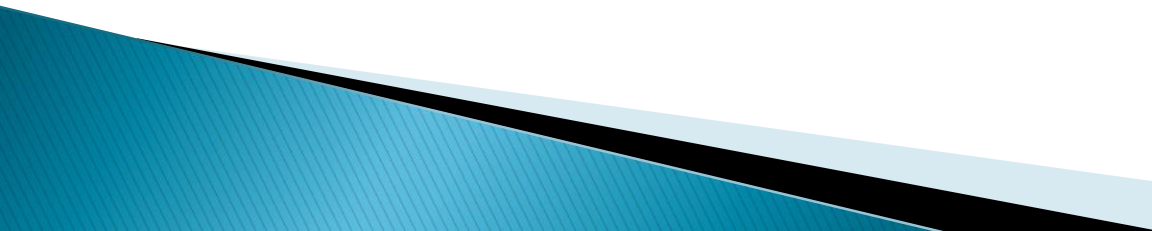
- ▶ Outliers should be identified and investigated.
- ▶ Outliers could be:
 - Chance occurrences
 - Measurement errors
 - Data entry errors
 - Sampling errors
- ▶ Outliers are not necessarily invalid data.

Check for Outliers

Fence Rule for checking for outliers using the quartiles:

- ▶ Calculate lower and upper fences:
 - Lower fence = $LF = Q_1 - (1.5 \times IQR)$
 - Upper fence = $UF = Q_3 + (1.5 \times IQR)$
- ▶ Values less than the lower fence or greater than the upper fence could be considered possible outliers.
- ▶ Minitab uses this method to star outliers.

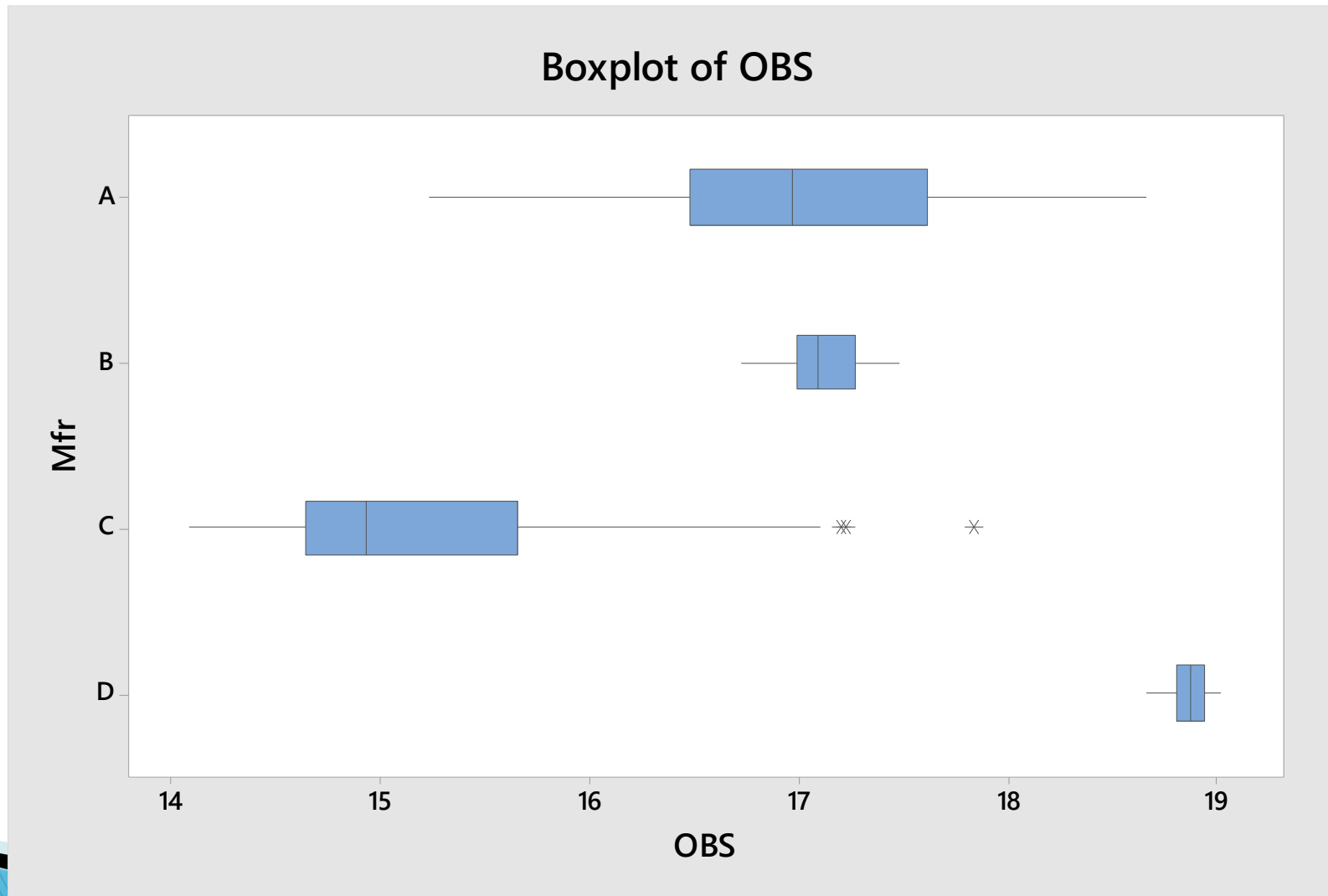
Check Fences

- ▶ Lower Fence = $110 - (1.5 \times 8) = 98$
 - ▶ Upper Fence = $118 + (1.5 \times 8) = 130$
 - ▶ 134 is a possible outlier since it's larger than 130.
 - ▶ Though it is considered a possible outlier, it is not a mistake and will remain in the data set.
- 

Comparative Boxplots

- ▶ Suppose there are four (A,B,C, and D) manufacturers producing stainless steel rods that are specified to be 17mm in diameter. Imagine you were picking which manufacturer you want to use.
 - Lets make a Boxplot to compare the Groups:
- ▶ Graph -> Boxplot -> choose with “Multiple Y’s”
- ▶ Click in the columns of data you want to compare .
- ▶ Click Scale Button and choose “transpose”

Comparative Boxplots



Histograms

- ▶ A graph for Quantitative data
- ▶ Steps:
 - Group data into bins (or classes) .
 - Create bars above each bin that represent the **frequency** (or **relative frequency**) in each bin.

Frequency  Count

Relative Frequency  Percentage (or proportion)

Bin or Class Information

Definitions:

▶ **Lower and Upper Class Limits**

For the class 30 – 39:

- 30 is the lower class limit
- 39 is the upper class limit

▶ **Bin Width:** Difference between consecutive lower class limits

- For the class 30 – 39, the class width = $40 - 30 = 10$.
- (All bins must have the same widths.)
- The class width is NOT the difference between upper and lower class limits for a single class.
- The class 30 – 39 years old actually is 30 years to 39 years 364 days old ... or 30 years to just less than 40 years old.
- The class width is 10 years, all adults in their 30's.

▶ **Class Midpoint:** The values in the middle of the classes. Often found by adding the lower limit and upper limit, then dividing by 2.

- Ex. For the class 30 – 39, the class midpoint = $(30 + 39)/2$

For ages of adults 20 – 69, a possible set of bins is:

20 – 29

30 – 39

40 – 49

50 – 59

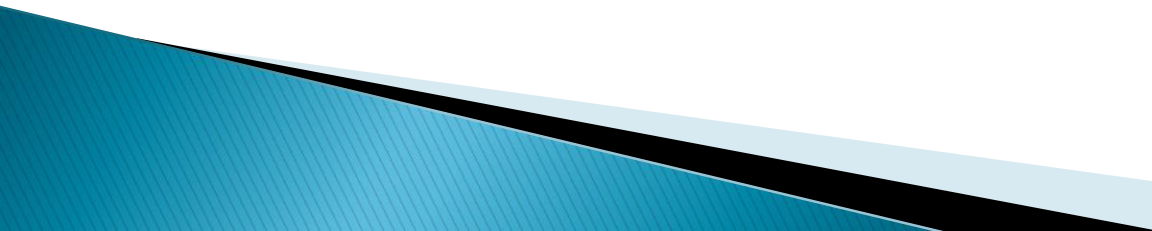
60 – 69

Good Practices Concerning Binning

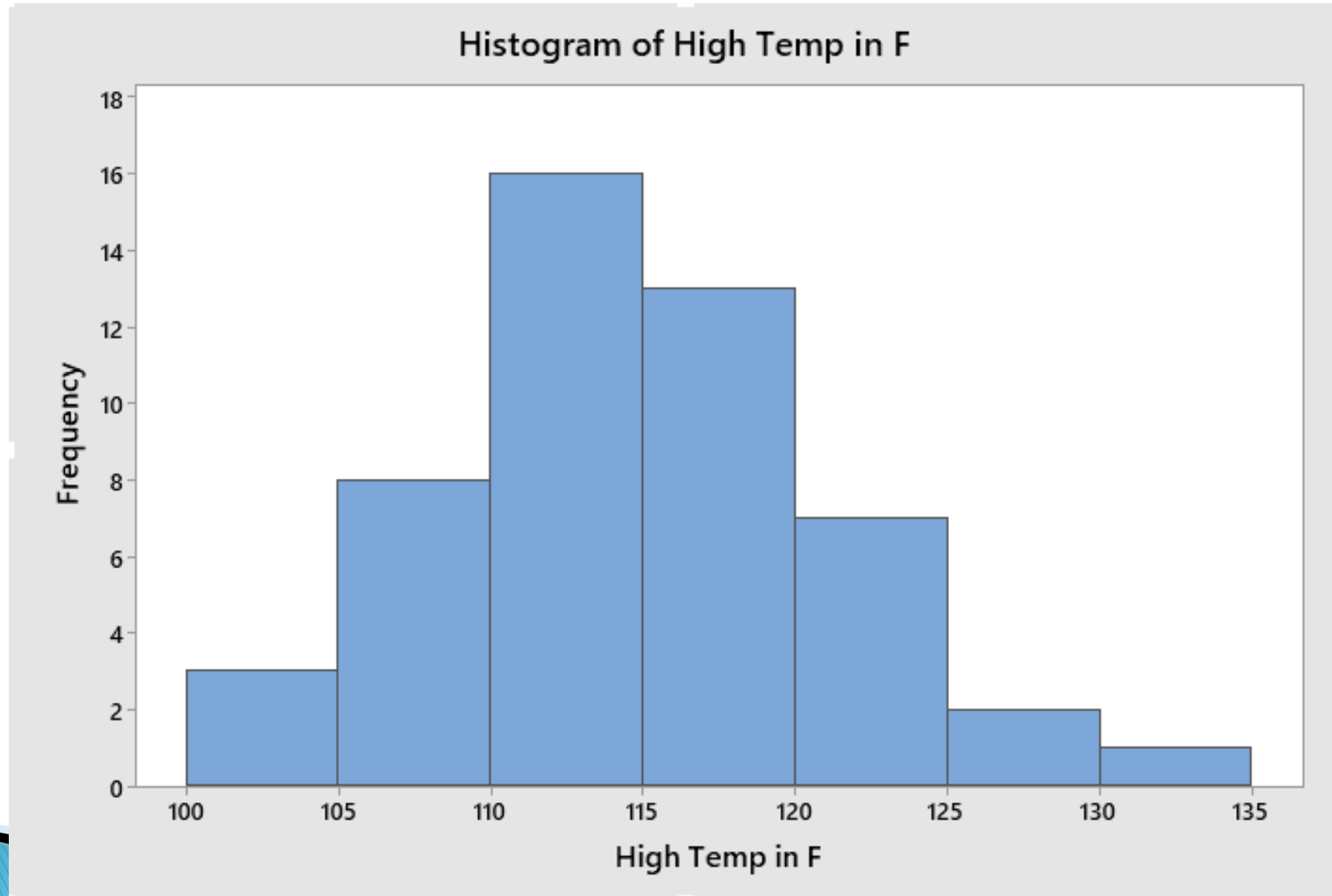
- ▶ There is no “perfect” way of selecting bins. But, the bins must:
 - Cover the range of the data.
 - Have the same width
 - Not overlap.
 - Not have any gaps between them.
 - Should be “reasonable” numbers.
- ▶ How many bins should we use?
 - Too few or too many bins spread the data out too far, making it difficult to detect patterns (5–20 bins?).
 - A good place to start is the square root of your number of observations

High Temperature Histogram Example

Instructions:

- ▶ Open or paste the data file
 - ▶ Click the Graph box and select Histogram.
 - ▶ Click on “High Temp. in F” to enter it into the graph box to the right.
 - ▶ Edit the title and labels.
 - ▶ Finally, click Compute!
- 

Histogram




Histogram

Advantages / Disadvantages

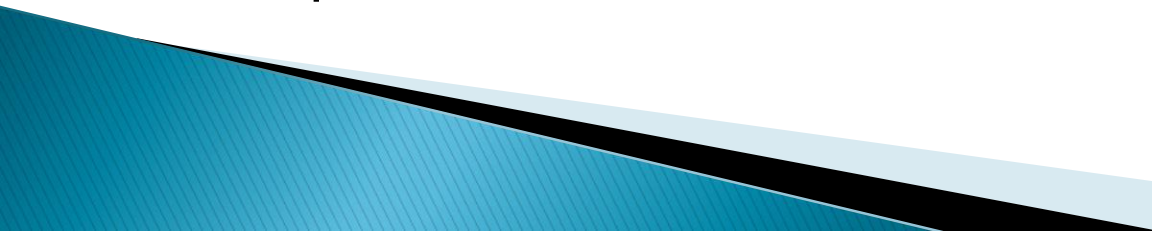
▶ Advantages

- Good for large data sets
- Gives a good picture of the general *shape* of the data


▶ Disadvantages

- Individual data values are not visible (lost)
 - Distribution shape affected by change in bin width
- 

Summarizing Numerical Distributions

- ▶ When examining a distribution visually, we want to mainly focus on:
 - Shape (including deviations of the overall pattern)
 - Any unusual values (Outliers)
 - ▶ We put less focus on but still need to note:
 - Typical Value (center)
 - Variability (spread)
 - ▶ “SOCS”
 - **Shape (Distribution)**
 - Outliers
 - Center
 - Spread (Variation)
- 

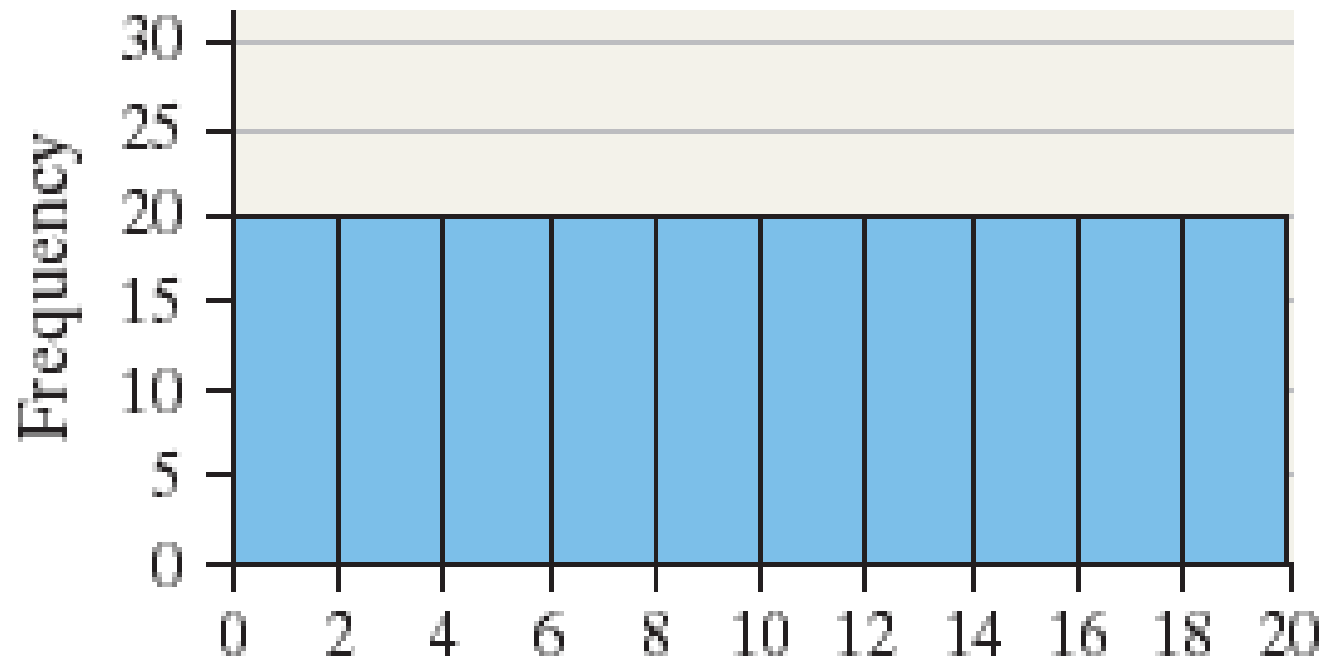
Characteristics of Shape

- ▶ Is the distribution symmetric or skewed?
 - ▶ How many mounds (or peaks) appear?
- 

Shape: Uniform

A variable has a uniform distribution when:

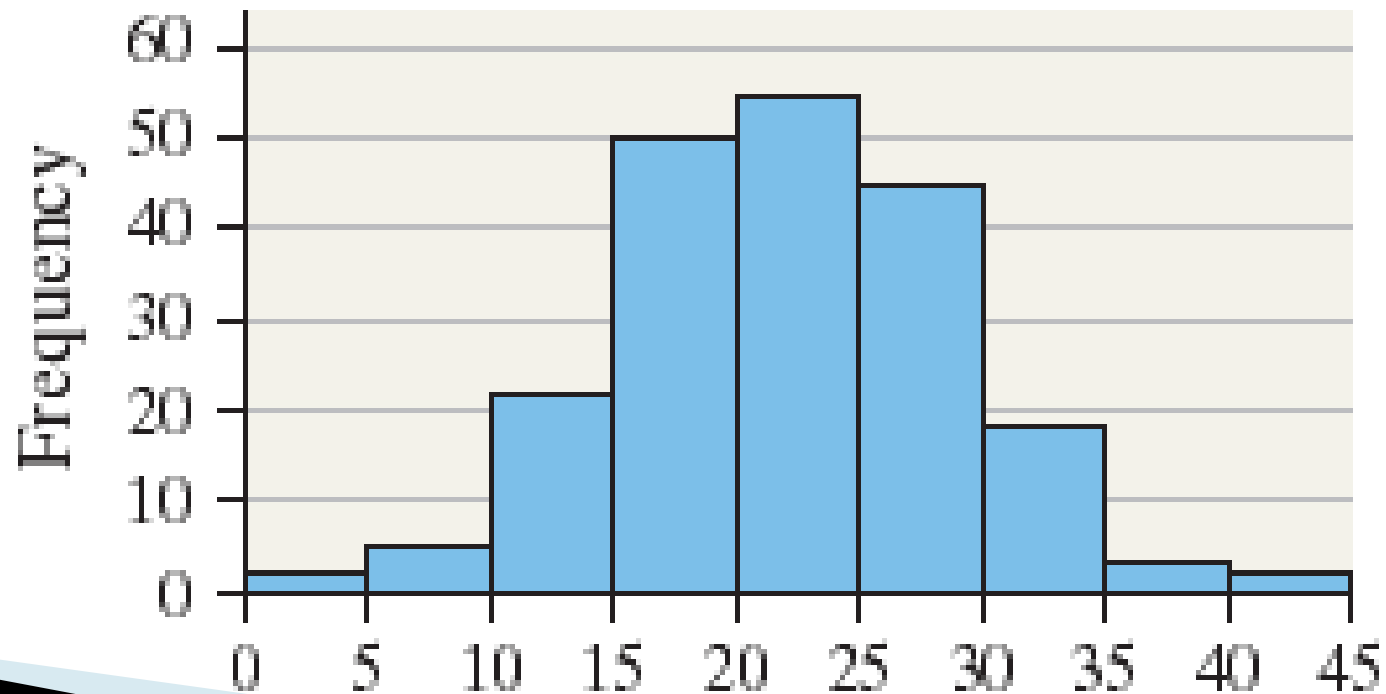
- Each of the values tends to occur with the same frequency.
- The histogram looks flat.



Shape: Bell-Shaped

A variable has a bell-shaped distribution when:

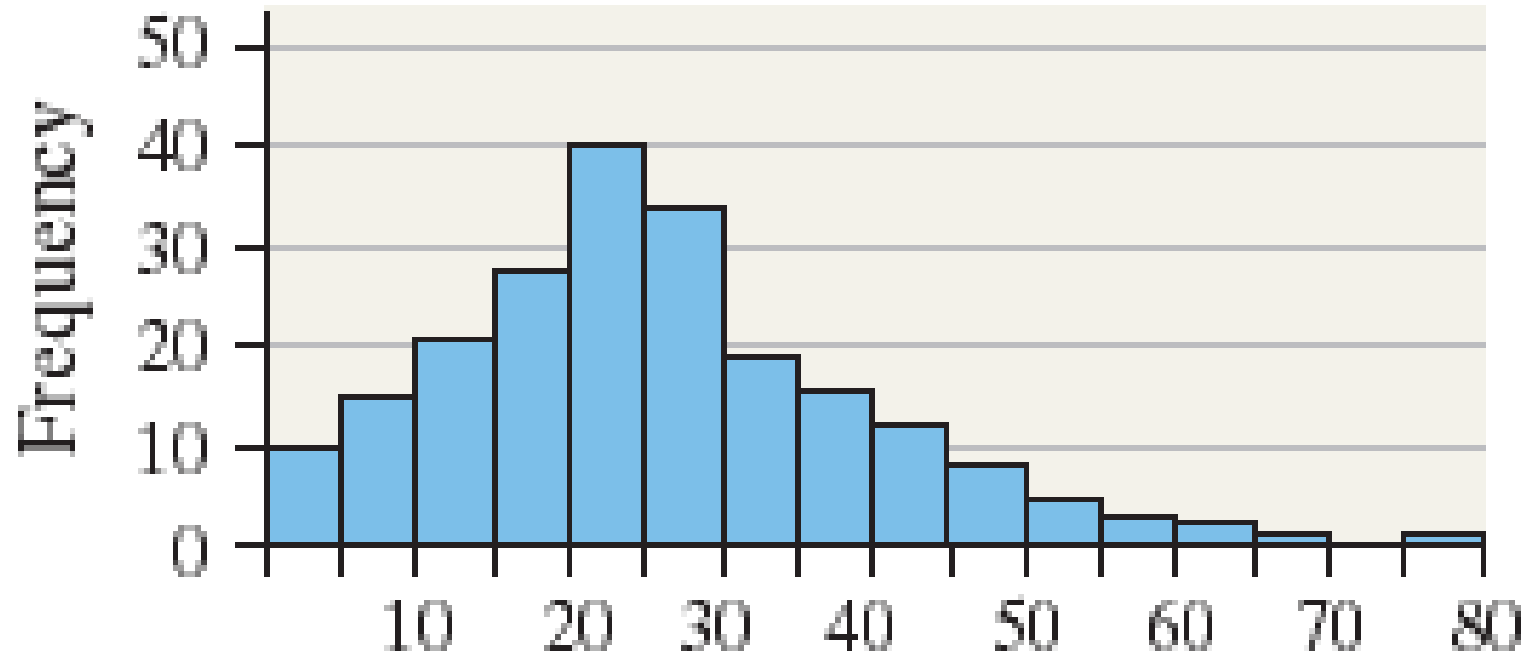
- Most of the values fall in the middle.
- The frequencies tail off to the left and to the right.
- It is symmetric (i.e., left half mirror image of right half).



Right-Skewed Distribution

A variable has a right-skewed distribution when:

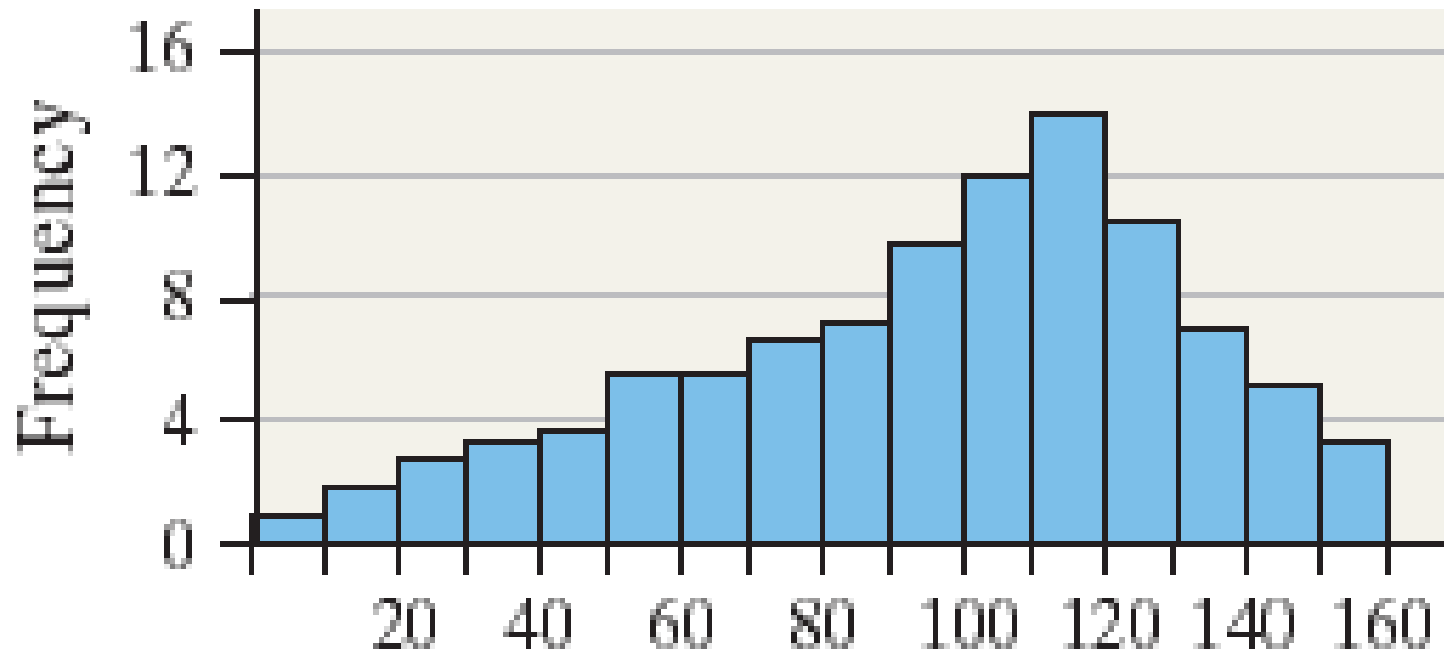
- The “tail” extends out to the right.
- A few large values skew the distribution.



Left-Skewed Distribution

A variable has a left-skewed distribution when:

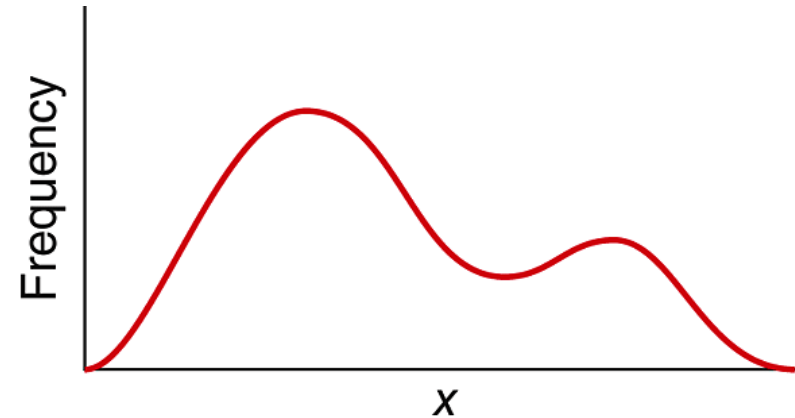
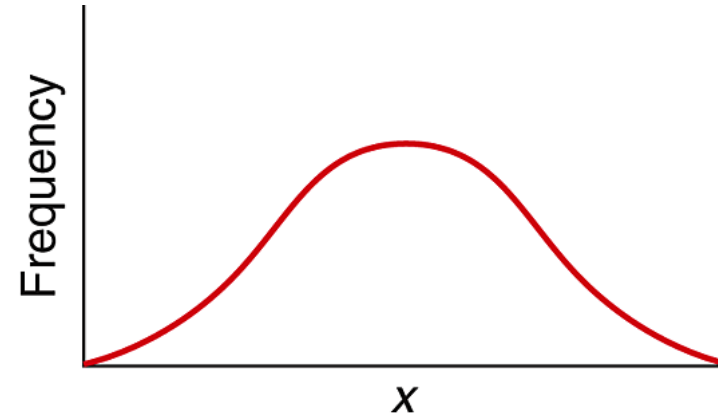
- The “tail” extends out to the left.
- A few small values skew the distribution.



Shape: Modality

Classify data by how many mounds are present:

- ▶ Unimodal: One main mound
- ▶ Bimodal
 - Two main mounds
- ▶ Multimodal
 - More than two main mounds




Notes on Modality

- ▶ Mounds can be different heights.
- ▶ Bimodal and multimodal data may indicate existence of different groups within the data.
- ▶ In this case, it may be preferable to separate the data into two groups and provide separate graphs for each group.
- ▶ Examples:
 - Men and women's heights
 - Afternoon and evening sales at a restaurant

Shapes of Distributions

Some common shapes of distributions are:

- ▶ Symmetric
 - Uniform
 - bell shaped
 - other symmetric shapes
 - ▶ Asymmetric
 - right skewed
 - left skewed
 - ▶ Unimodal, bimodal
- 

Shape: Examples

What shape would you expect to see in a histogram of the following data sets?

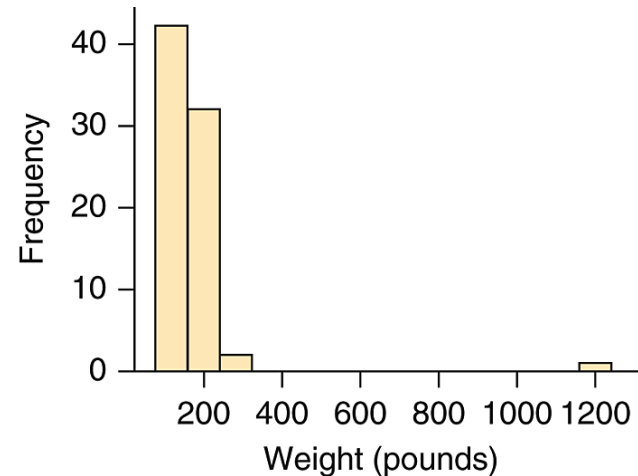
- ▶ GPA of college students
 - *Skewed left*
- ▶ SAT scores
 - *Symmetric (Unimodal)*
- ▶ Last digit of Social Security numbers for a random sample of students
 - *Symmetric (Uniform)*
- ▶ Income of USA residents
 - *Skewed right*

Important Characteristics of Data

- ▶ “SOCS”
 - Shape (Distribution)
 - Outliers
 - Center
 - Spread (Variation)

Outliers

- ▶ Potential Outliers are:
 - Extremely large or small values
 - Do not fit the pattern of the rest of the data and
 - May be apparent visually but subject to opinion.




- ▶ If you see extremely large or small values:
 - Report the values.
 - Realize they could be sources of error (typos, etc.).
 - Genuine outliers are unusually interesting data values

Important Characteristics of Data

- ▶ “SOCS”
 - Shape (Distribution)
 - Outliers
 - Center
 - Spread (Variation)

Measures of Center

- ▶ Center of the data:
 - Numeric values that represent the most “**typical value**” of a quantitative variable.
 - Also called “central tendency.”
 - ▶ Two ways to think about center
 - Balancing point (Mean)
 - Halfway point (Median)
 - ▶ We may prefer one versus the other depending on the **shape** of the distribution
- 

Measure of Center: Mean

- ▶ Can be thought of as the “balancing point of the distribution”
- ▶ Technically the Arithmetic Mean, also called the average.
- ▶ There are other ways to calculate averages, but only one way to calculate the arithmetic mean.
 - E.g. weighted average to find your GPA
 - “Trimmed” Mean



Population Mean Formula

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

The population mean is a parameter

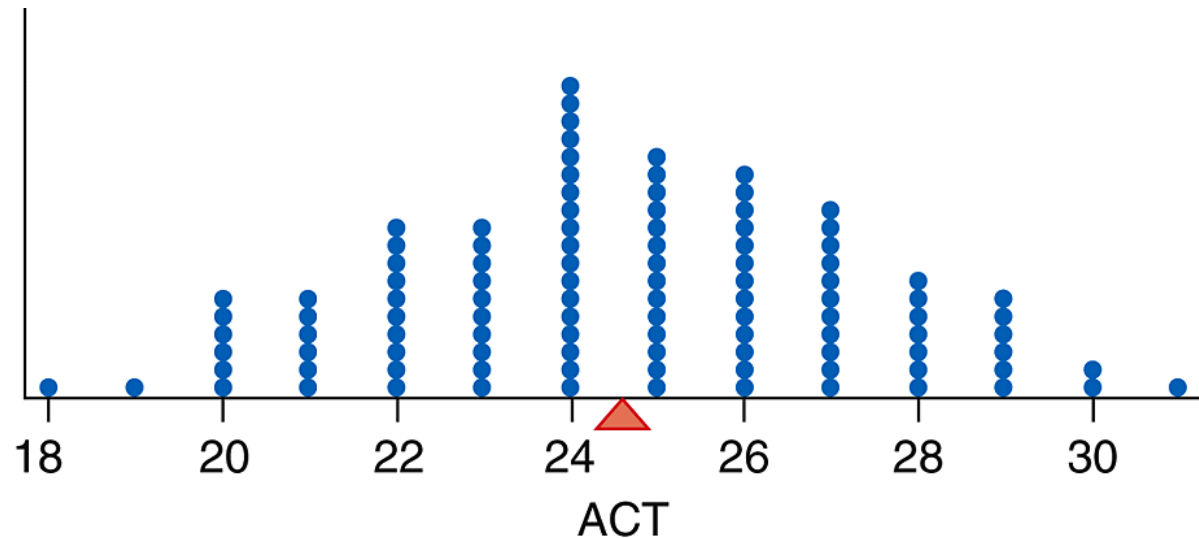
Sample Mean Formula

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The sample mean is a statistic

The Mean: Symmetric Distributions

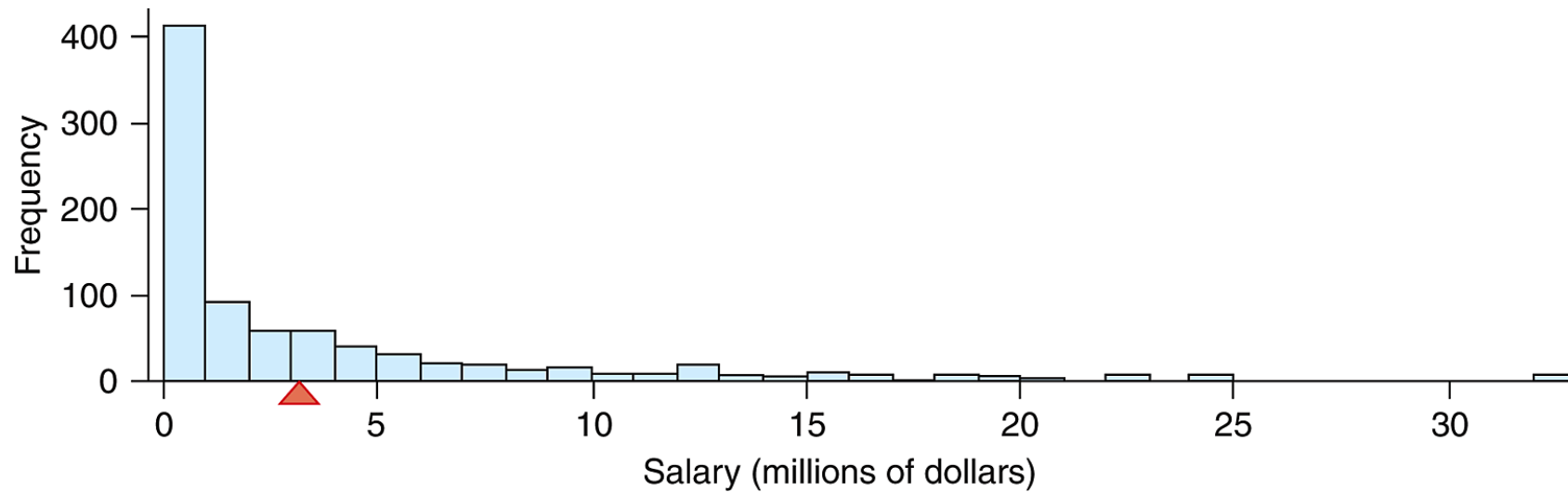
This dot plot shows the distribution of ACT scores for a sample of statistics students.



For symmetric distributions, the mean is a good representation of a “typical value” of the data set.


The Mean: Skewed Distributions

This distribution shows the salaries of professional baseball players in 2010.



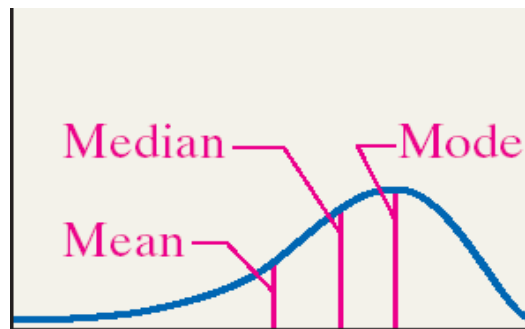
Do you think the mean is a good representation of the “typical” baseball salary for that year?

Properties of the Mean

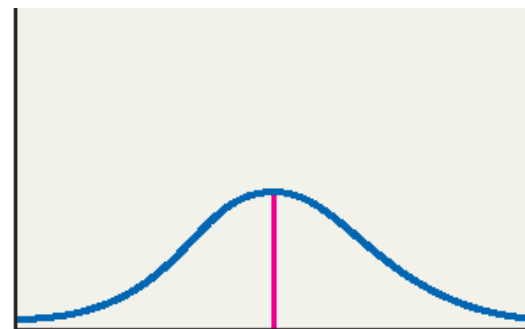
- ▶ The mean is sensitive to the influence of a few extreme observations.
 - ▶ Also, skewed distributions (with or without outliers) can pull the mean towards its long tail.
 - ▶ This is another reason why the mean is not always a great measure of center for skewed data sets.
- 

Comparing Mean and Median

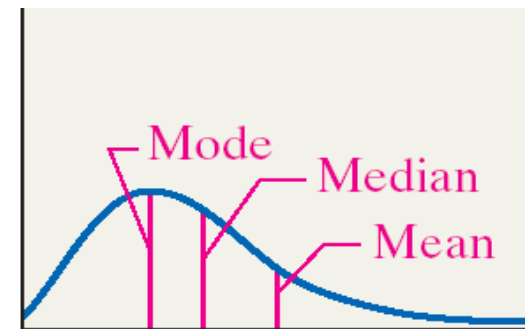
- ▶ In a skewed distribution, the mean gets ‘pulled’ in the direction of the few extreme values (the mean works as a point of balance).
- ▶ The difference between the mean and the median may provide clues about a distribution’s shape.
 - **Symmetric** – mean will usually be close to the median.
 - **Skewed left** – mean will usually be smaller than the median.
 - **Skewed right** – mean will usually be larger than the median



(a) Skewed Left
Mean < Median



Mean = Median = Mode
(b) Symmetric
Mean = Median



(c) Skewed Right
Mean > Median

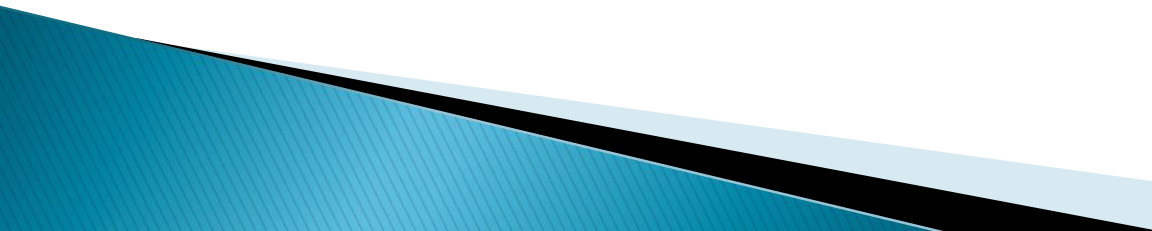
“Best” Measure of Center

- ▶ We want our measure of center to describe the most “Typical Value” of a distribution
- ▶ We will use the mean to do this when it is symmetric
- ▶ We should use the median if it is skewed

Important Characteristics of Data

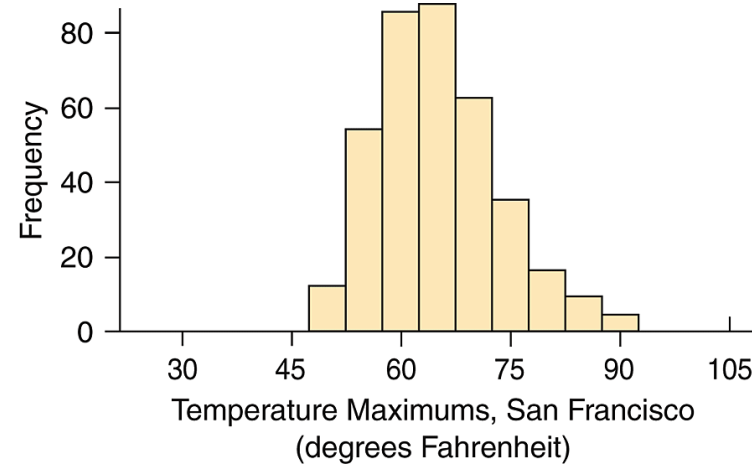
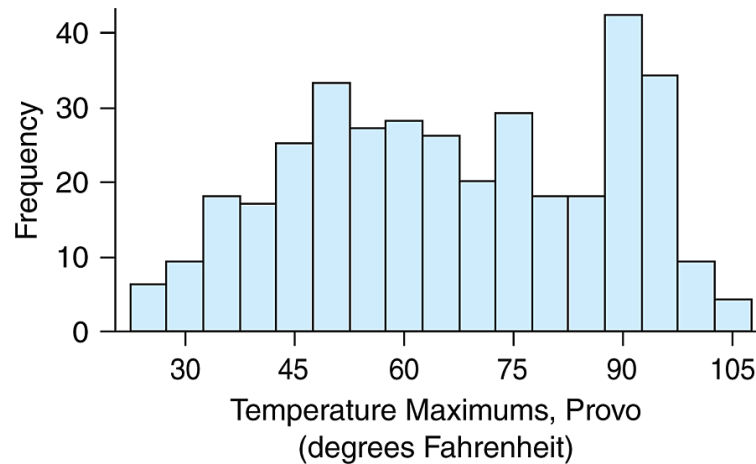
- ▶ “SOCS”
 - Shape (Distribution)
 - Outliers
 - Center
 - Spread (Variation)

Measures of Spread (Variation)

- ▶ The variability in a distribution can be measured by the horizontal spread.
 - ▶ We want to know if most of the data is near the center or far from it.
 - ▶ So far we have:
 - Range
 - IQR (Typically used along with the Median for skewed Distributions)
 - ▶ Variation **supplements** our understanding of the center
- 

Measuring the Spread: Example

The following histograms record the daily high temperatures in degrees Fahrenheit over one recent year at two locations:



- ▶ Both distributions have roughly the same shape and center
- ▶ However, the spread in both distributions is very different:
 - Provo: More spread out (data values farther from center)
 - SF: Less spread out (data values closer to the center)

Quantifying the Spread

- ▶ Numbers that measure how far away the typical observation is from the mean (center)
- ▶ Variance – The average of the squares of the distance each value is from the mean.
- ▶ Standard Deviation
 - The square root of the variance.
 - Think of the standard deviation as the Average distance of the observations from their mean.

Standard Deviation Formulas

- ▶ Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

- ▶ Sample Standard Deviation

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

Standard Deviation Formula

The formula for the standard deviation is:

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

1. Deviation (or distance) of observation, x , from the mean.

2. Square to make positive.

3. Add all squared deviations

4. Divide by 1 less than the sample size. Think of this as averaging the squared deviations.

5. Take square root to restore original units.

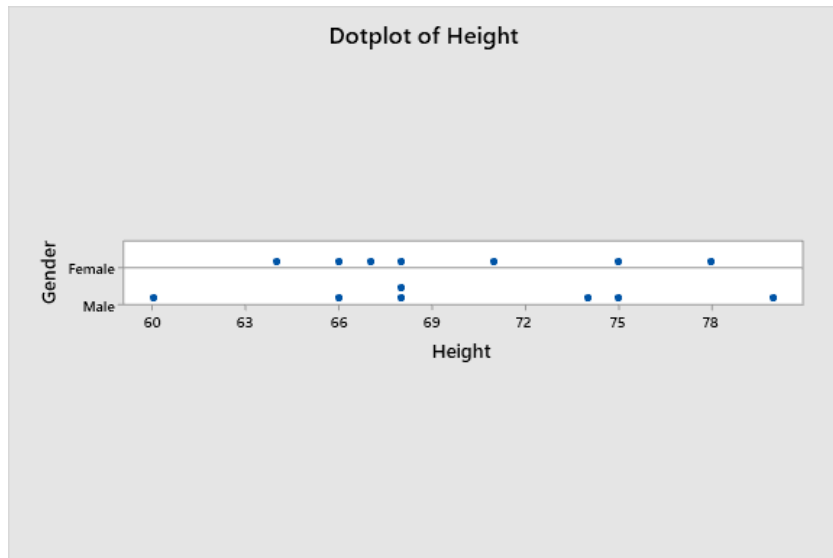
SD calculation Example

- ▶ Consider a random sample of seven female and seven male students from a class.
- ▶ We want to find the mean and standard deviation of their heights.
- ▶ Practice calculating these by hand, and then check yourself using technology

Gender	Age	Height
Female	29	78
Female	22	66
Female	19	64
Female	33	68
Female	35	75
Female	56	67
Female	25	71
Male	20	80
Male	19	74
Male	19	66
Male	23	68
Male	20	75
Male	23	68
Male	23	60

Gender Comparison

- ▶ Male Height standard deviation is 6.644 while Female Height standard deviation is 5.080



Statistics

Variable	Gender	Total Count	Mean	StDev	Variance
Height	Female	7	69.86	5.08	25.81
	Male	7	70.14	6.64	44.14

- ▶ Since the standard deviation of the male height is bigger, we can say that there is more variation in the Male sample data.
 - We can see on a **dotplot** that the male heights are more spread out.

Standard Deviation Properties and Context

- ▶ The mean is the context of the Standard deviation
- ▶ Standard deviations mean different things to different means
 - We should only compare standard deviations if the data sets' means are similar.
- ▶ $s = 0$ only when all observations have the same value and there is **no spread**. Otherwise, $s > 0$.
- ▶ s is affected by outliers and skewness.
- ▶ s has the same units of measurement as the original observations.

Time Series Plots

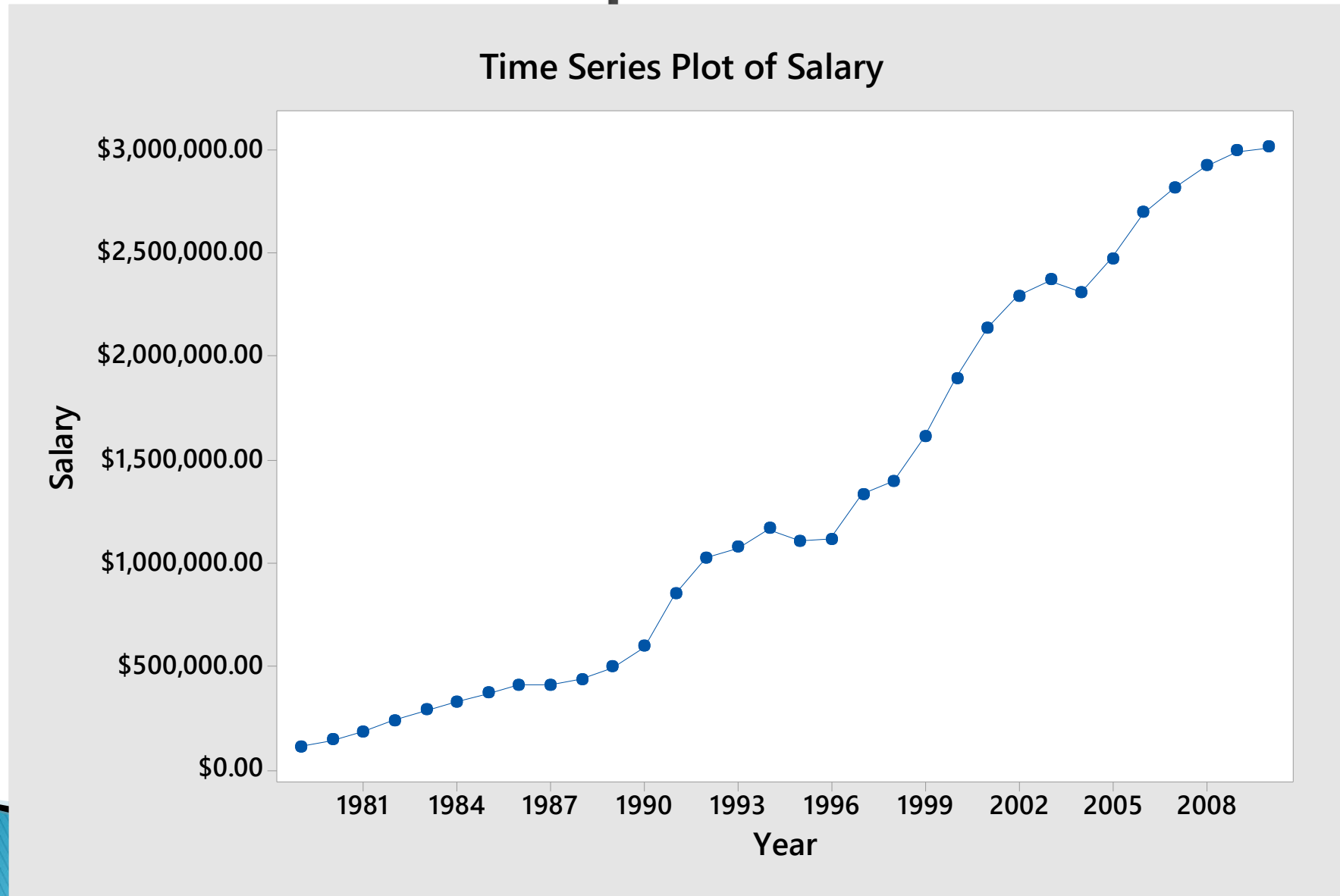
- ▶ A time series plot shows the data value, or statistic, on the vertical axis with time on the horizontal axis.
- ▶ A time series plot reveals trends, cycles or other time-oriented behavior that could not be otherwise seen in the data.



Time Series Example

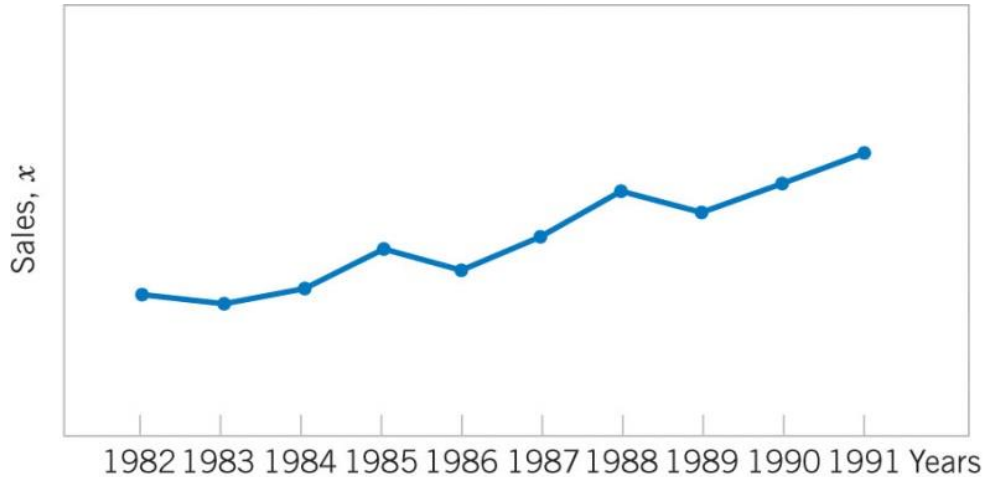
- ▶ Using Baseball Salary Data
- ▶ Graph -> Time Series Plot
- ▶ Select the variable of interest (in our case salary)
- ▶ Look at data and Choose Time/Scale
 - We have 2 choices
 - If we have the associated years then click Stamp and choose your year variable
 - If you do not have the data for year choose calendar -> year and then tell which year to start (in our case 1979)

Time Series Example

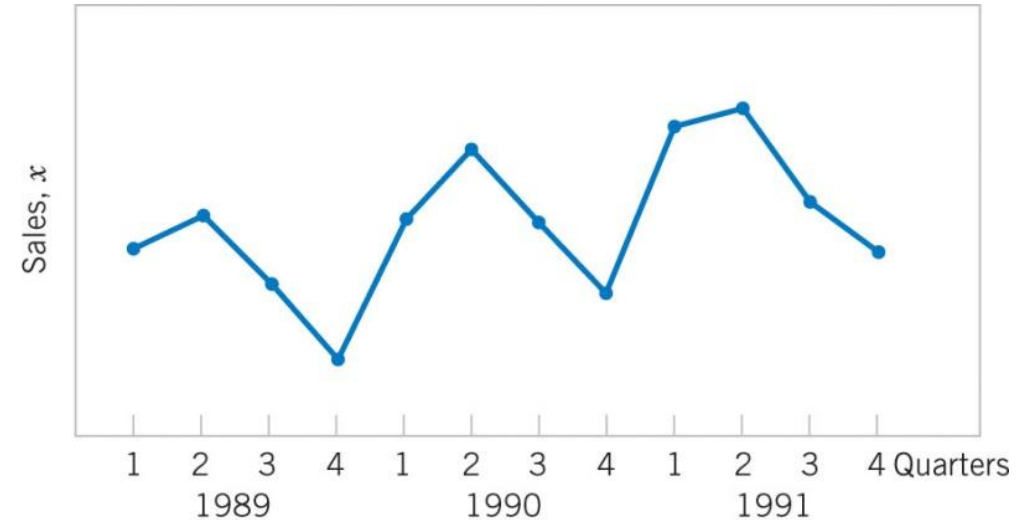


Time Series Plots (cont...)

- ▶ We can look at trend over any increment of time. Sometimes, changing the time increment can change the pattern



(a)



(b)

Company sales by year (a) & by quarter (b). The annual time interval masks cyclical quarterly variation, but shows consistent progress.

Summary of Measures

- ▶ We should consider a few things when deciding what measures best describe our data:
 - The shape of the Distribution
 - Presence of extreme values/possible outliers
- ▶ For Symmetric Distributions we prefer:
 - Mean
 - S.D.
- ▶ For Skewed Distributions we prefer:
 - Median
 - IQR
 - (Or just Five Number Summary)