Chapter 3

Displaying and Summarizing Quantitative Data

Homework

p68 6, 12, 14, 15, 21, 22, 23, 25,
27, 29, 31, 32, 35, 39, 40, 41

Objectives

- Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots.
- Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Dealing With a Lot of Numbers...

- Summarizing the data gives us a sense of the behavior of large sets (distributions) of quantitative data. So ...
    - ... the first thing a statistician (or any researcher) should do is to make a graphical summary of the data to get an overall view of the data distribution.
- Make a picher, make a picher, make a picher!
- Howsomever, we cannot use bar charts or pie charts for **quantitative** data, since those displays are for **categorical** variables.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histograms

- A histogram is much like a bar chart but with a significant difference. There are no gaps between bins in a histogram because the histogram displays continuity.
- To draw a histogram...
  - First, divide up the entire range of data values covered by the **quantitative** variable into **equal-width** intervals to be represented by bars we will call **bins**.
  - The intervals of the bins cover the entire **distribution** of the quantitative variable values.
- Shoot for **7 - 15 bins**. Too few and the picture is not revealing, just a few boxes. If you have too many bins the picture is too complex. We cannot see the forest for the trees.
  - The information you wish to communicate will help determine the appropriate number of bins.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

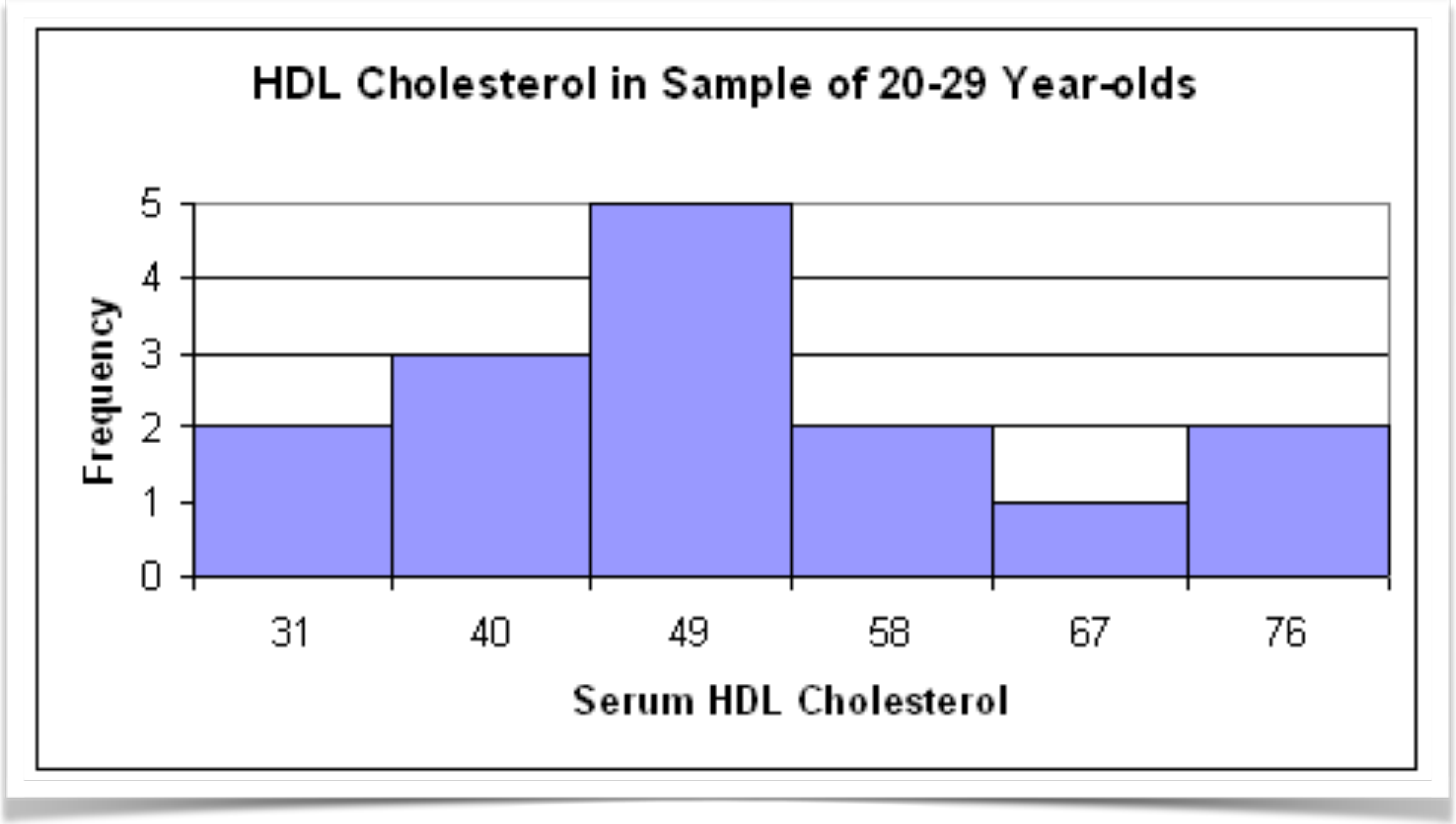
Histograms

- A **histogram** plots the bin counts as the heights of bars (same as a bar chart).
  - **Histograms** and bar charts are **NOT** the same. A bar chart has gaps between bins (categorical data), a **histogram** has no gaps between bins reflecting continuous, quantitative data.

- A histogram provides a graphical representation that illustrates the shape of your data distribution.

- Shown is a histogram of HDL (good) cholesterol counts. Higher values are better.

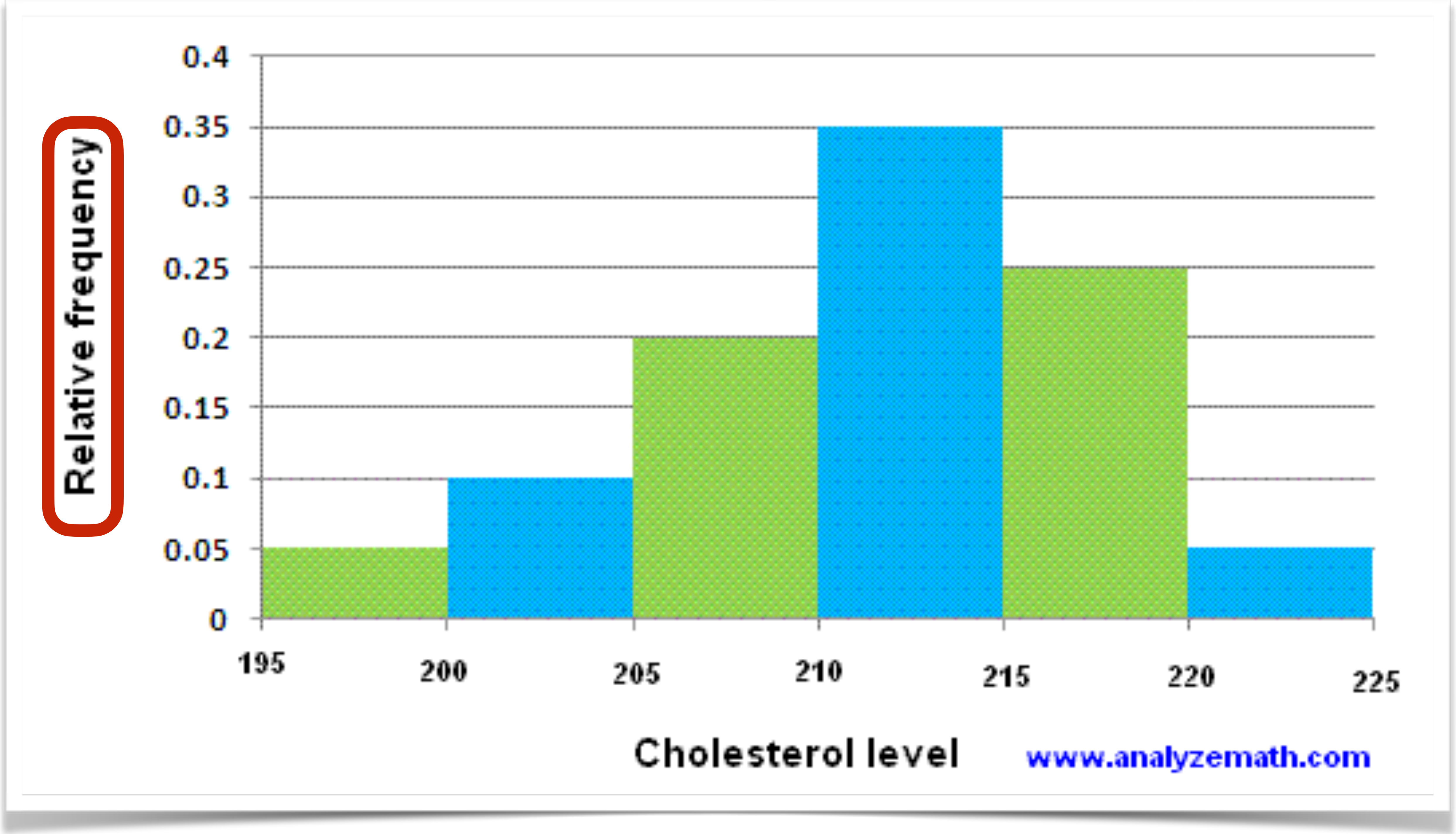
- The bins are labeled with integers but represent an interval. 31 = 30.5 - 31.5



Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histograms

- A **relative frequency histogram** displays the ***percentage*** of cases in each bin instead of the count.
  - The **relative frequency histogram** looks identical to the frequency histogram. The only change is in the vertical axis scale.
  - Here is a **relative frequency histogram** of overall cholesterol levels:
  - Only the one aspect changes.

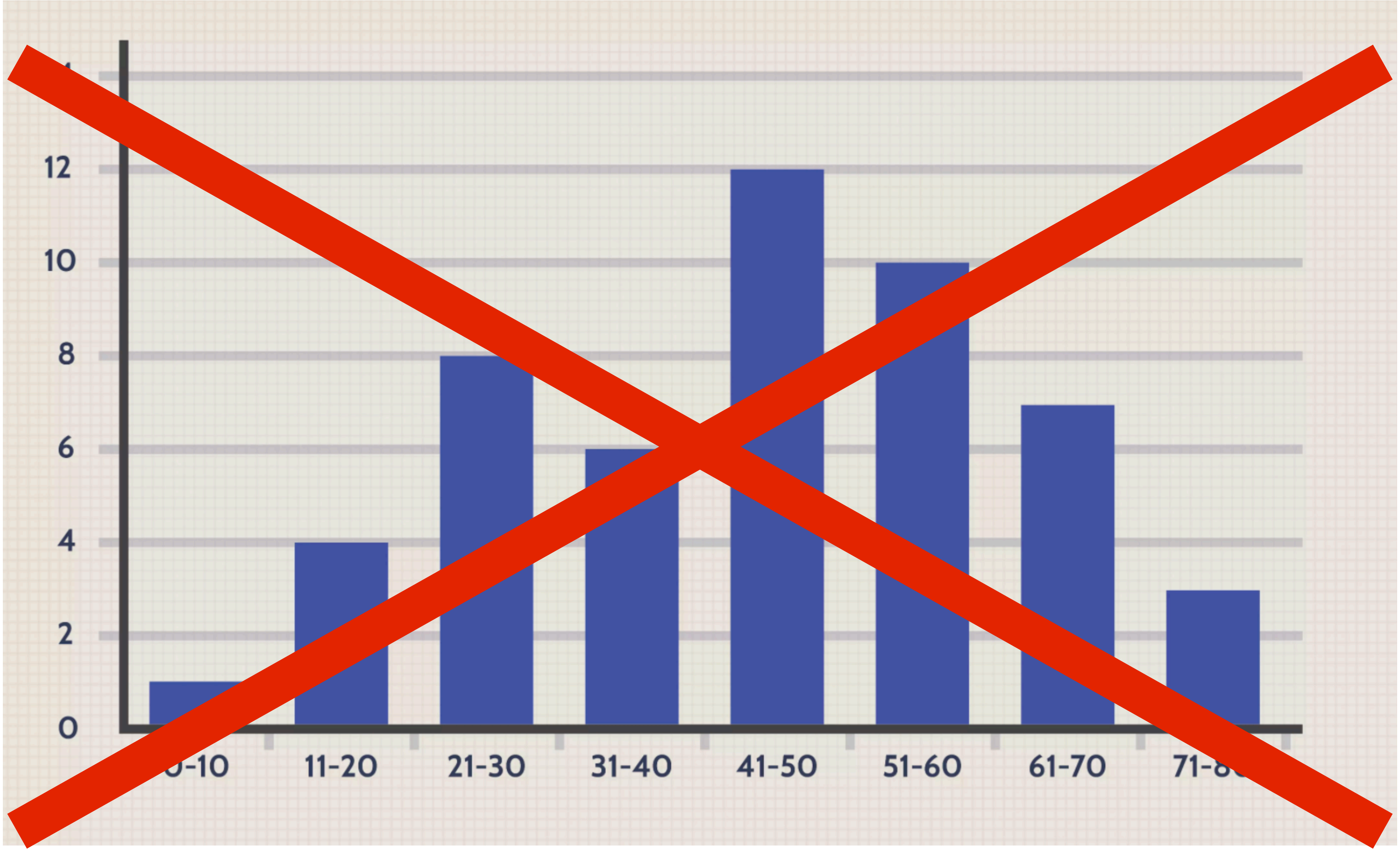


Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

- The histogram is a graph that displays the data by using bins (bars) of various heights (representing class frequencies).

- The histogram bins (bars) are contiguous (no gaps).
  - Gaps between bars = bar graph.



- Horizontal-axis: the lower boundaries or lower limits align with the sides of the bins, the midpoints of data classes align with center of bin.

- Vertical-axis: The height of the bar represents the frequency of the class.

Grouped Frequency Distribution

- A **Grouped Frequency Distribution** is used when the range of data values is large. The data is grouped into **classes that are more than a single unit in width**.

- **Class limits** represent the largest and smallest data values that an be included in the class. Class limits **are actual potential data values**.

- **Class boundaries** provide values that eliminate gaps between the classes in the frequency distribution. Class boundaries are one decimal place more accurate than the data. Class boundaries are **not potential data values**.

    - The **class boundaries** are typically one decimal place more accurate (average upper and lower limits of adjacent classes) and are **not** actual data values. The **class boundaries** are the result of the interval each datum represents.

Class Width

- The class width (width of class interval) can be calculated by subtracting
  - successive lower class limits (or boundaries)
  - successive upper class limits (or boundaries)
  - upper and lower class boundaries.

- The class midpoint X_m can be calculated by averaging upper and lower class limits (or boundaries).

$$\frac{\text{lower limit + upper limit}}{2} = \frac{10+14}{2} = 12$$

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

Rules for setting class width

1. No overlapping intervals (classes). The classes must be mutually exclusive and exhaustive.
   - One datum cannot belong in two classes and **all potential** data is included.

2. Class widths must be consistent.

3. The class width should be small enough to accurately portray the data but large enough to keep the number of classes manageable.
   - Readers of the data cannot assimilate too many classes, the forest gets lost in the trees.
   - The number of classes is best kept between 5 and 15. Fewer than 5 and the trends get lost, more than 15 and the information becomes lost. This rule is not cast in stone. I try to limit to between 7 and 12 depending on the data.

Histogram

Rules for setting class width

4. The whole idea of a frequency distribution is to provide an accurate and easily understood picture of the data as simply as possible.

5. If the class width is an odd number the midpoint of the class is an actual datum value. That can be useful for graphical representations of the data. Not a necessary condition but often useful because we like pretty pictures of our data.

Grouped Frequency Distribution

- **Procedure for constructing a grouped frequency distribution.**
  - Find the highest and lowest values in your data.
  - Find the range of the data.
  - Select the number of classes desired (7-12).
  - Find the width of each interval by dividing the range by the number of classes desired. If the result is not an integer, select the next largest integer.
  - Select a starting lower limit (usually the lowest value); add the class width to get successive lower limits.
  - Find the upper class limits.
  - Tally the data and record the class frequencies.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

- The following data represent the record high temperatures for each of the 50 states.

- Enter the data into a list on your calculator.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	109	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

STAT → 1:EDIT → Select List → "Enter first datum" → ENTER → Repeat to end of list → 2ND → QUIT

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	109	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

- The data represent the record high temperatures for each of the 50 states. Construct a grouped frequency distribution for the data using **7 classes**.

- **STEP 1** Determine the classes.

  - Find the class width by dividing the range by the desired number of classes 7.
  - Range = High – Low = 134 – 100 = 34
  - Width = Range/7 = 34/7 = 5 (next greater integer).

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	109	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

- We choose the lowest data value, 100, for the **initial class lower limit**.

Class Limits

100 - 104
105 - 109
110 - 114
115 - 119
120 - 124
125 - 129
130 - 134

- The subsequent lower class limits are found by adding the **class width** to the previous lower class limits.
- The **initial class upper limit** is one less than the succeeding class lower class limit.
- The subsequent upper class limits are found by adding the **class width** to the previous upper limit.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	109	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

- The class boundaries are midway between an upper class limit and a subsequent lower class limit.

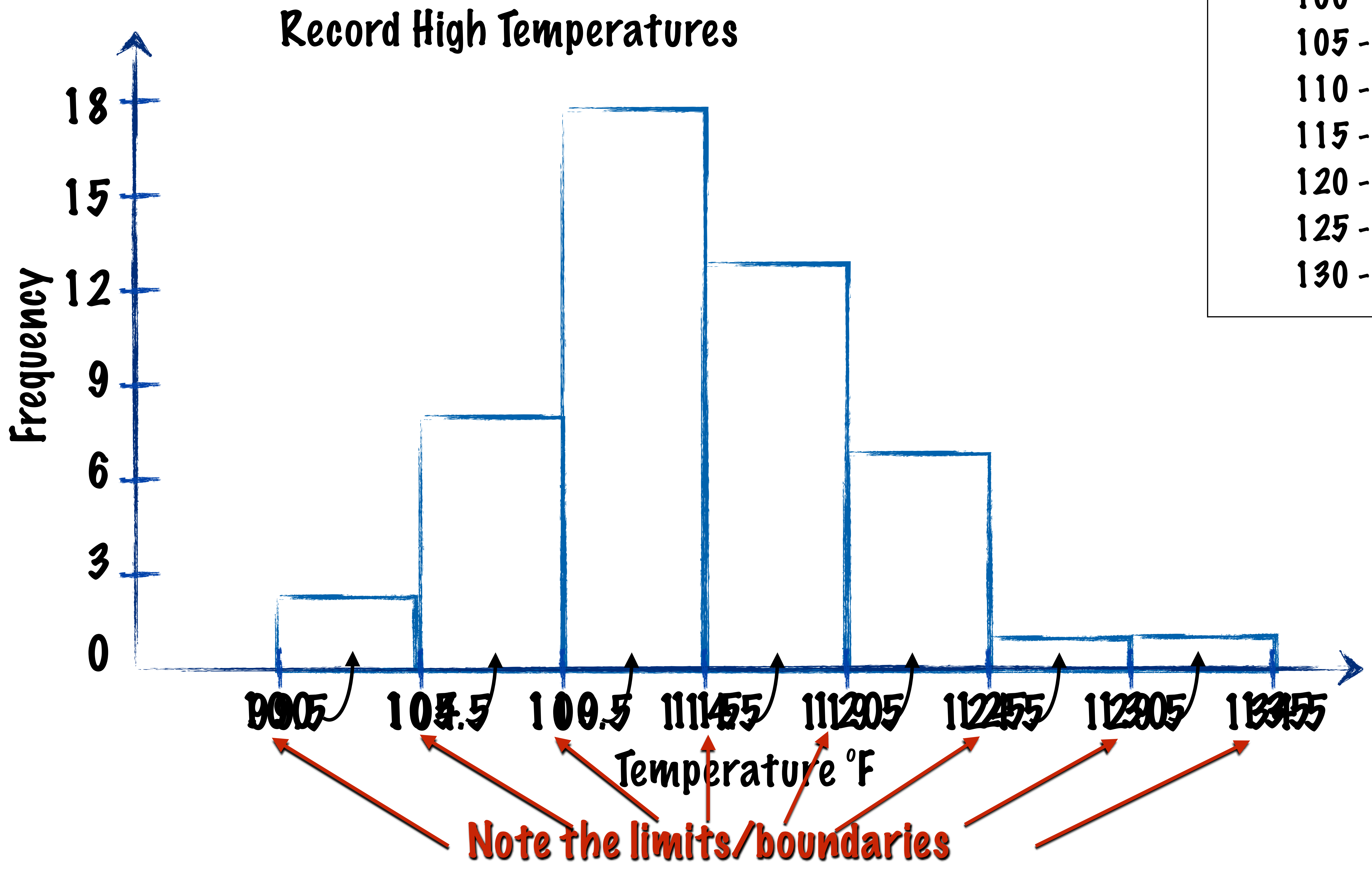
104, **104.5**, 105

- Find the frequency for each class.

Class Limits	Class Boundaries	Frequency (f)	Cumulative Frequency (cf)
100 - 104	99.5 - 104.5	2	2
105 - 109	104.5 - 109.5	8	10
110 - 114	109.5 - 114.5	18	28
115 - 119	114.5 - 119.5	13	41
120 - 124	119.5 - 124.5	7	48
125 - 129	124.5 - 129.5	1	49
130 - 134	129.5 - 134.5	1	50

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram



Class Limits	Class Boundaries	f
100 - 104	99.5 - 104.5	2
105 - 109	104.5 - 109.5	8
110 - 114	109.5 - 114.5	18
115 - 119	114.5 - 119.5	13
120 - 124	119.5 - 124.5	7
125 - 129	124.5 - 129.5	1
130 - 134	129.5 - 134.5	1

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Histogram

- Now, let us get real and use the calculator to draw the histogram. Using the data we entered into a list (the 50 state data), draw the same histogram we just created.

- To enter data into a list in your calculator …

STAT → 1:EDIT → Select List → "Enter first datum" → ENTER → Repeat to end of list → 2ND → QUIT

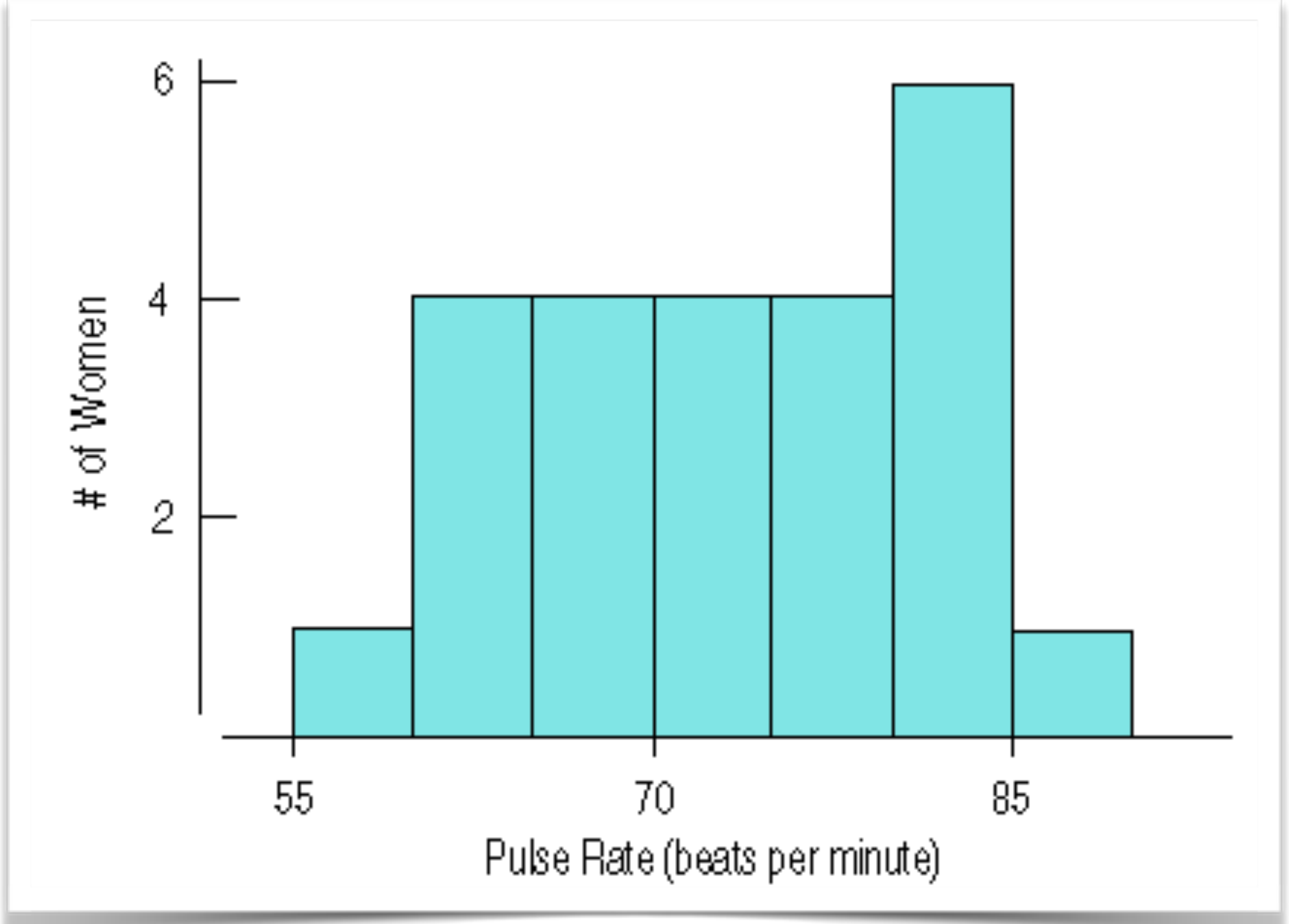
- To draw a histogram …

2nd (STAT PLOT) → y= → Enter → ON → TYPE: [histogram icon] → XList → 2nd → 1 (L_1) → Freq: 1 → 2nd → Quit → Zoom → 9

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Stem-and-Leaf Displays

- **Stem-and-leaf displays** show the distribution of a quantitative variable, like histograms do, while **preserving the individual values**.
  - Stem-and-leaf displays contain all the information found in a histogram and still, when properly created, satisfy the area principle and show a picture of the distribution.
- Compare the histogram and stem-and-leaf display for the pulse rates of 24 women at a health clinic. The stem-and-leaf displays the shape of the distribution, while retaining the actual data.



5	6
6	0444
6	8888
7	2222
7	6666
8	000044
8	8

Constructing a Stem-and-Leaf

- First, cut each data value into leading digits (the stems) and trailing digits (the leaves).
  - Use the stems to label the 'bins'.
  - Use **only one digit** for each leaf—either round or truncate the data values to one place after the stem.

Stems	Leaves
5	6
6	0444
6	8888
7	2222
7	6666
8	000044
8	8

- **Always include a "key"**

5 | 6 = 56

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Stem and Leaf Plot

- Create a stem-and-leaf plot (stem plot);

~~32~~, ~~18~~, ~~47~~, 65, 22, 33, 64, 44, 32, 15, 9, 16, 48, 77, 31, 25,
28, 55, 56, 12, 7, 10, 28, 22, 65, 47, 18, 32, 55, 15, 44

0	9 7
1	8 5 6 2 0 8 5
2	2 5 8 8 2
3	2 3 2 1 2
4	7 4 8 7 4
5	5 6 5
6	5 4 5
7	7

→

0	7 9
1	0 2 5 5 6 8 8
2	2 2 5 8 8
3	1 2 2 2 3
4	4 4 7 7 8
5	5 5 6
6	4 5 5
7	7

1 | 2 = 12

Be sure to include a key for the stemplot

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Stem and Leaf Plot

- Create a stem-and-leaf plot (stem plot);

~~12.3~~, ~~6.2~~, ~~12.4~~, 9.8, 12.5, 7.4, 6.4, 7.7, 8.5, 10.5, 9.1, 11.0

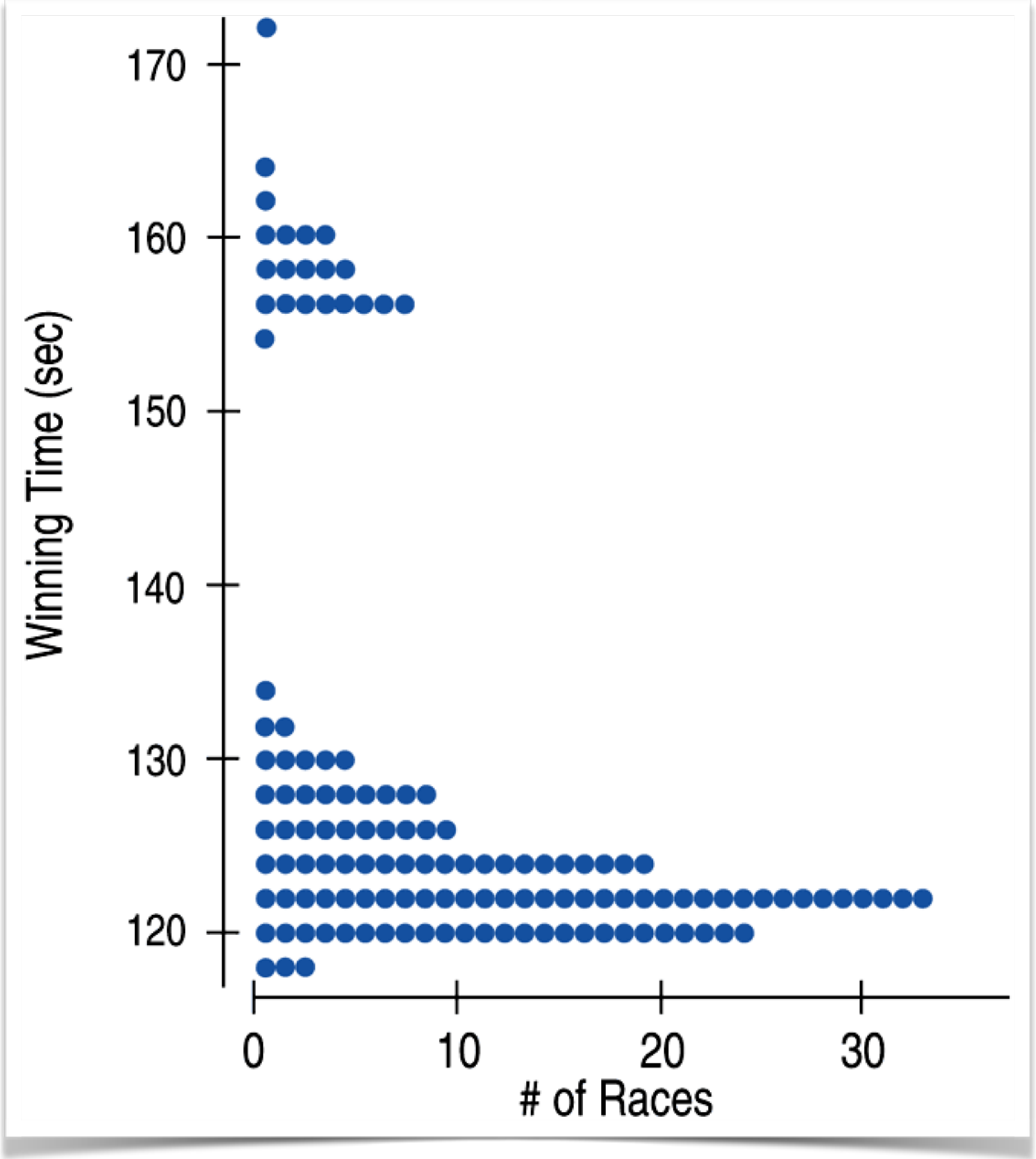
6	2 4
7	4 7
8	5
9	1 8
10	5
11	0
12	3 4 5

6 | 2 = 6.2

DO NOT FORGET to include a key for the stemplot

Dotplots

- A **dotplot** is a simple display. For each case in the data. simply place a dot along an axis.
  - The dotplot to the right shows Kentucky Derby winning times, plotting each race as its own dot.
  - You might see a dotplot displayed horizontally or vertically. My preference is a horizontal axis.



Think Before You Draw, Again

- Remember how we start with "Make a picher"?
  - Now that we have options for data displays, you need to think about which **type** of display to make. Different graphics emphasize different aspects of the story.
  - Before you create a graphic display of your data, be certain you match the appropriate type of display for your data and that the display illustrates the aspect of your data that you intended.
  - For quantitative data use a histogram, frequency polygon (line graph), stem-and-leaf, boxplot, or dot plot. For qualitative data you can use a bar chart, pie chart, or dot plot.

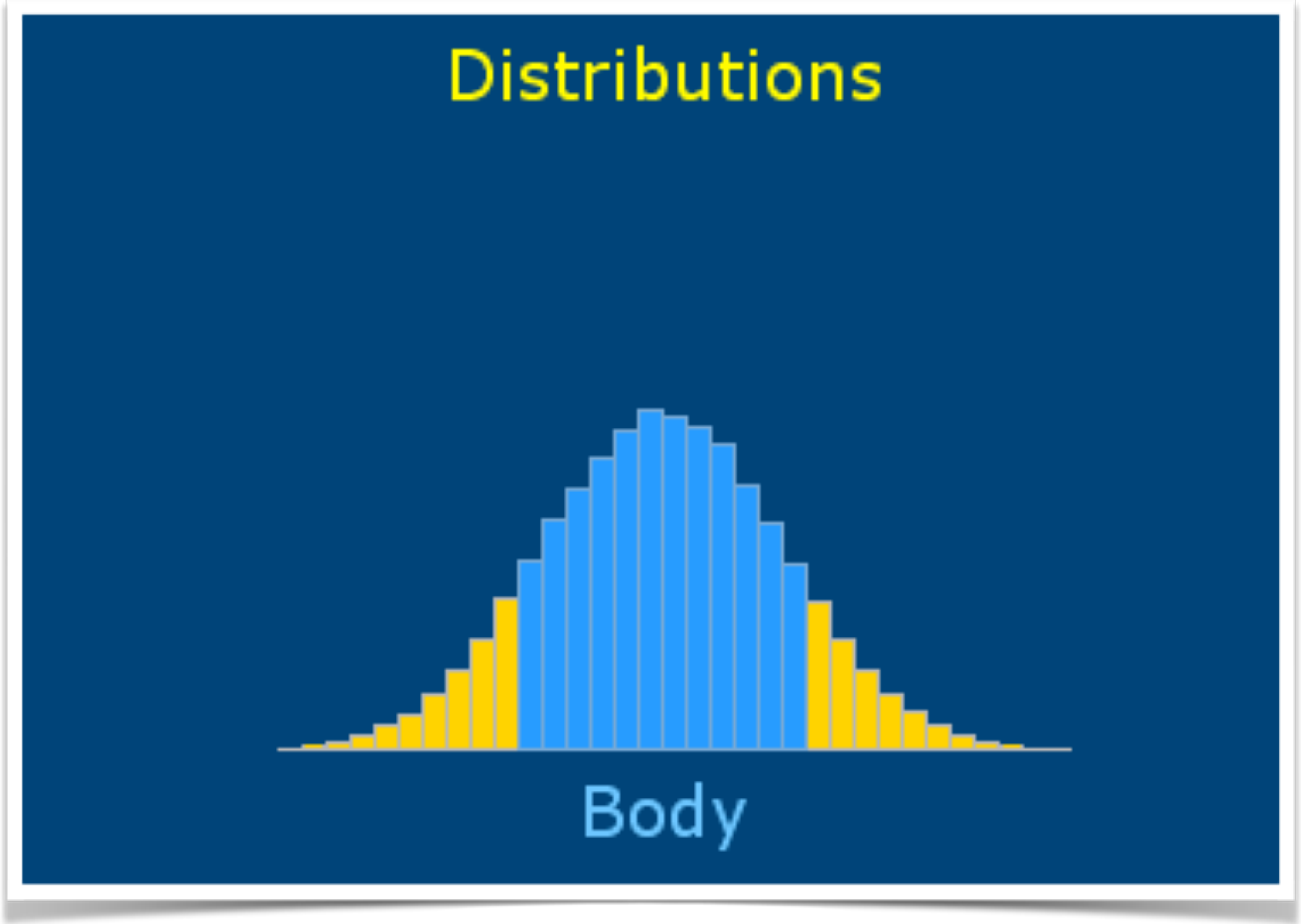
Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

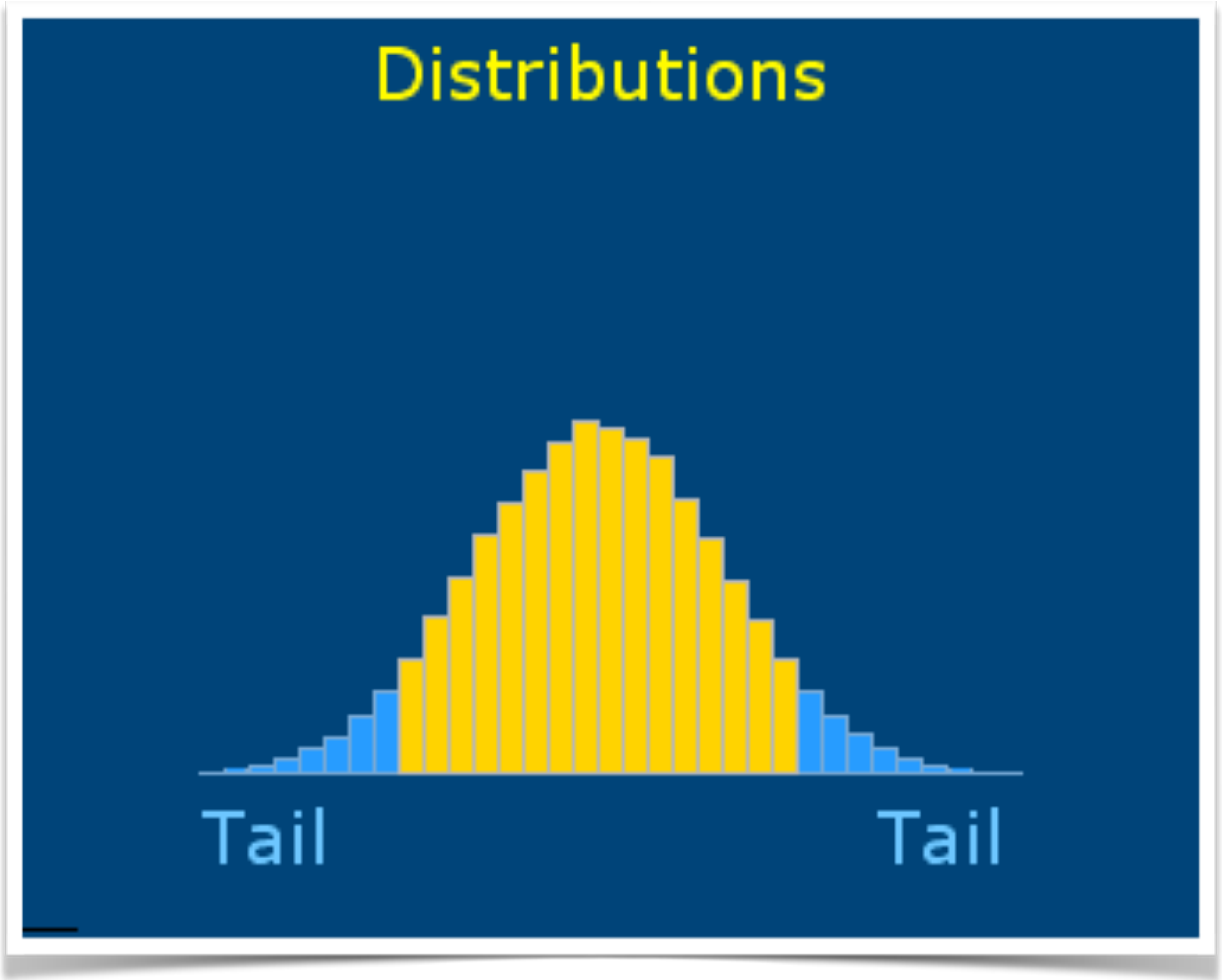
Shape, Center, Spread, Unusual Values

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Shape, Center, and Spread

- When describing a distribution, (and you will, often) you must describe three things: the **shape**, **center**, and **spread** of the distribution ...
  - Look at the body of the distribution and then examine the "tails".





Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Shape

- **Humps**, **symmetry**, and **unusual features**

1. Does the histogram have a single hump or several separated humps? (Body)

2. Is the histogram symmetric? Does the hump (body) tend to the center of the distribution of values or is it to one side? Are the tails similar in size and shape?

3. Do any unusual features stick out? Are there gaps in the data, are any data values off by themselves (outliers)? (Body and Tails)

Humps

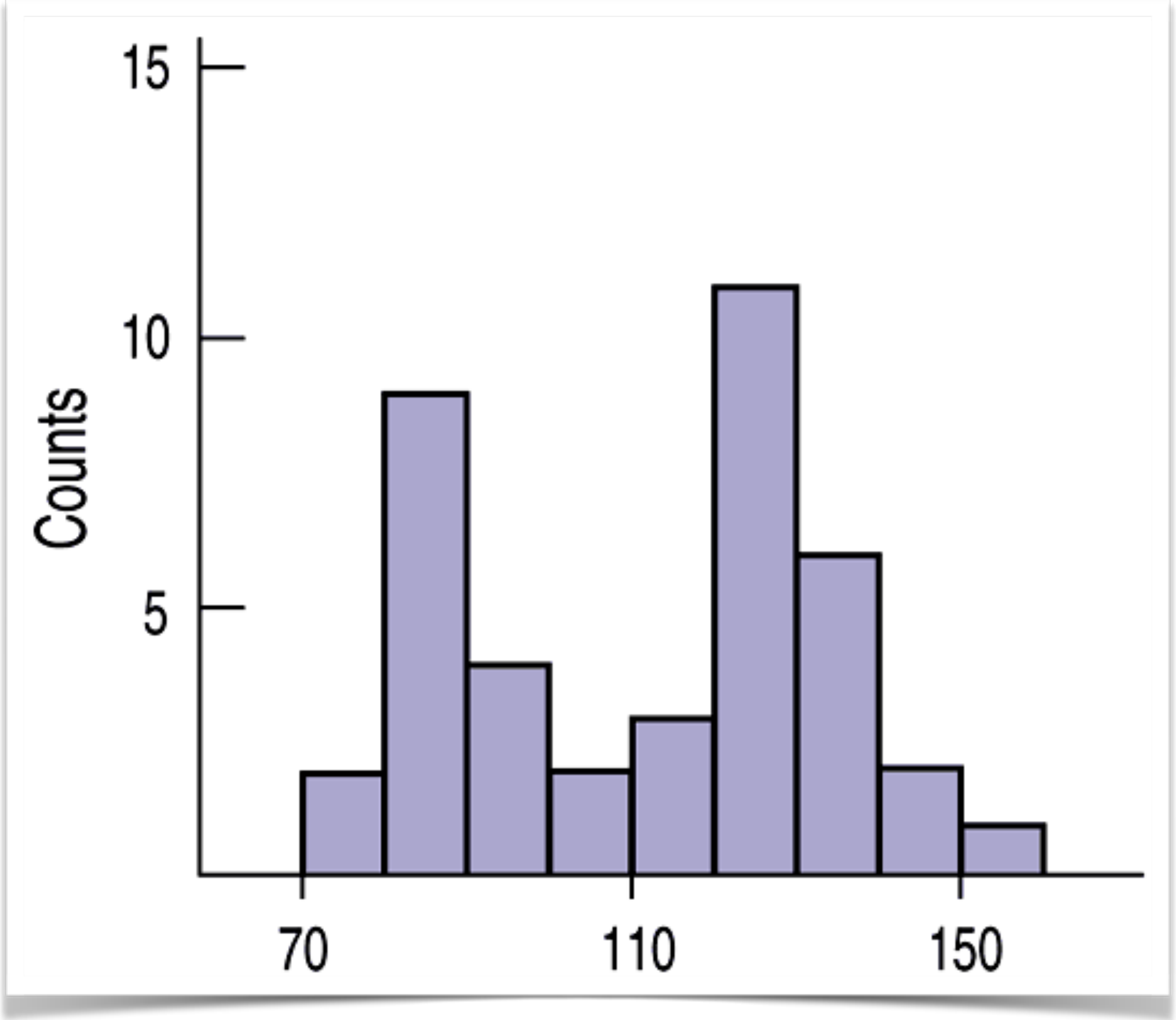
1. Does the histogram of the distribution have a single, central hump or multiple separated humps?

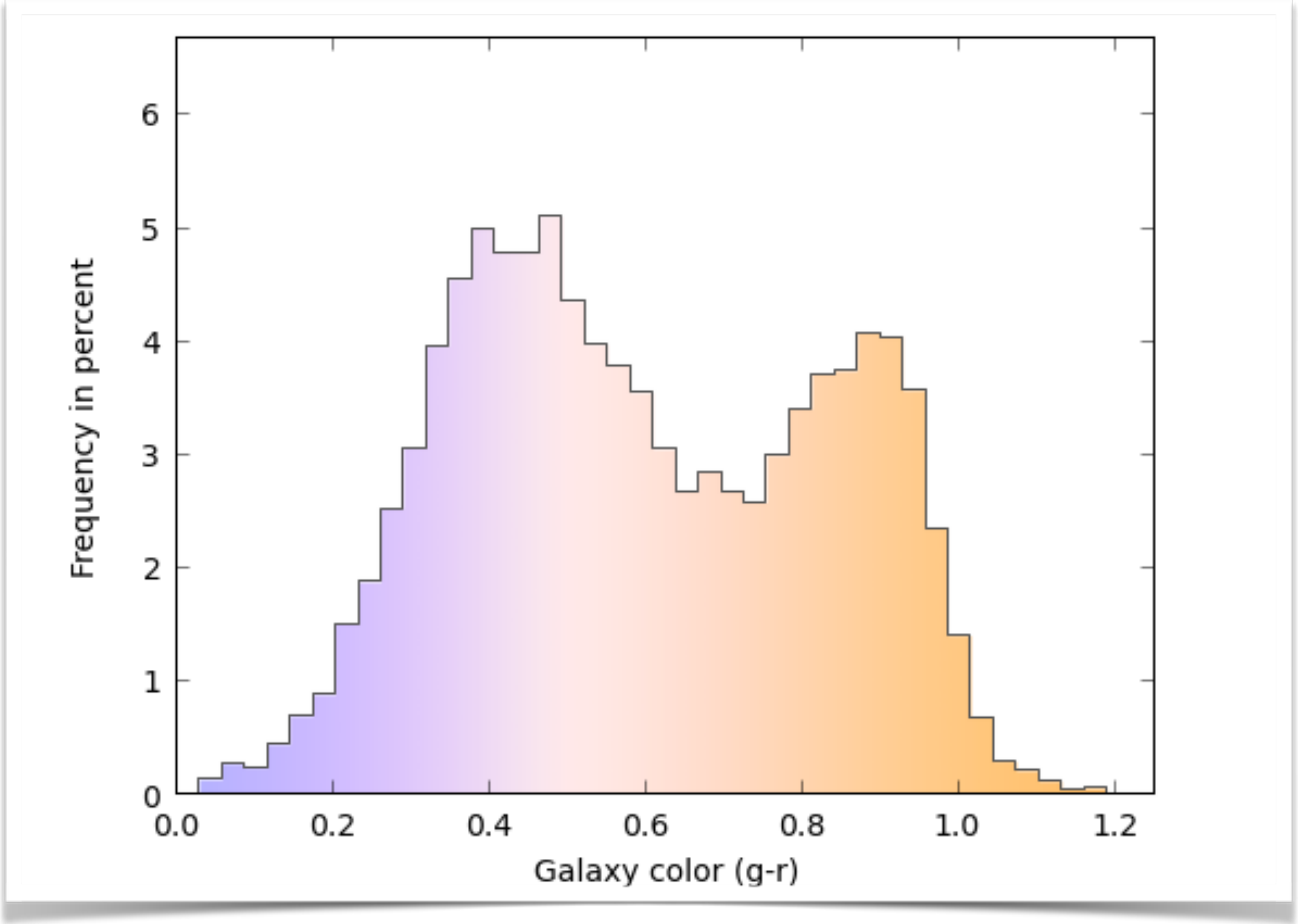
- Humps in a histogram are called **modes**.
  - This is a somewhat different definition of mode than the one you may have learned in an elementary math class.

- A distribution with one main hump (peak) in the graph is defined as **unimodal**; histograms with two main humps are **bimodal**; histograms with three or more humps are **multimodal**.

Humps

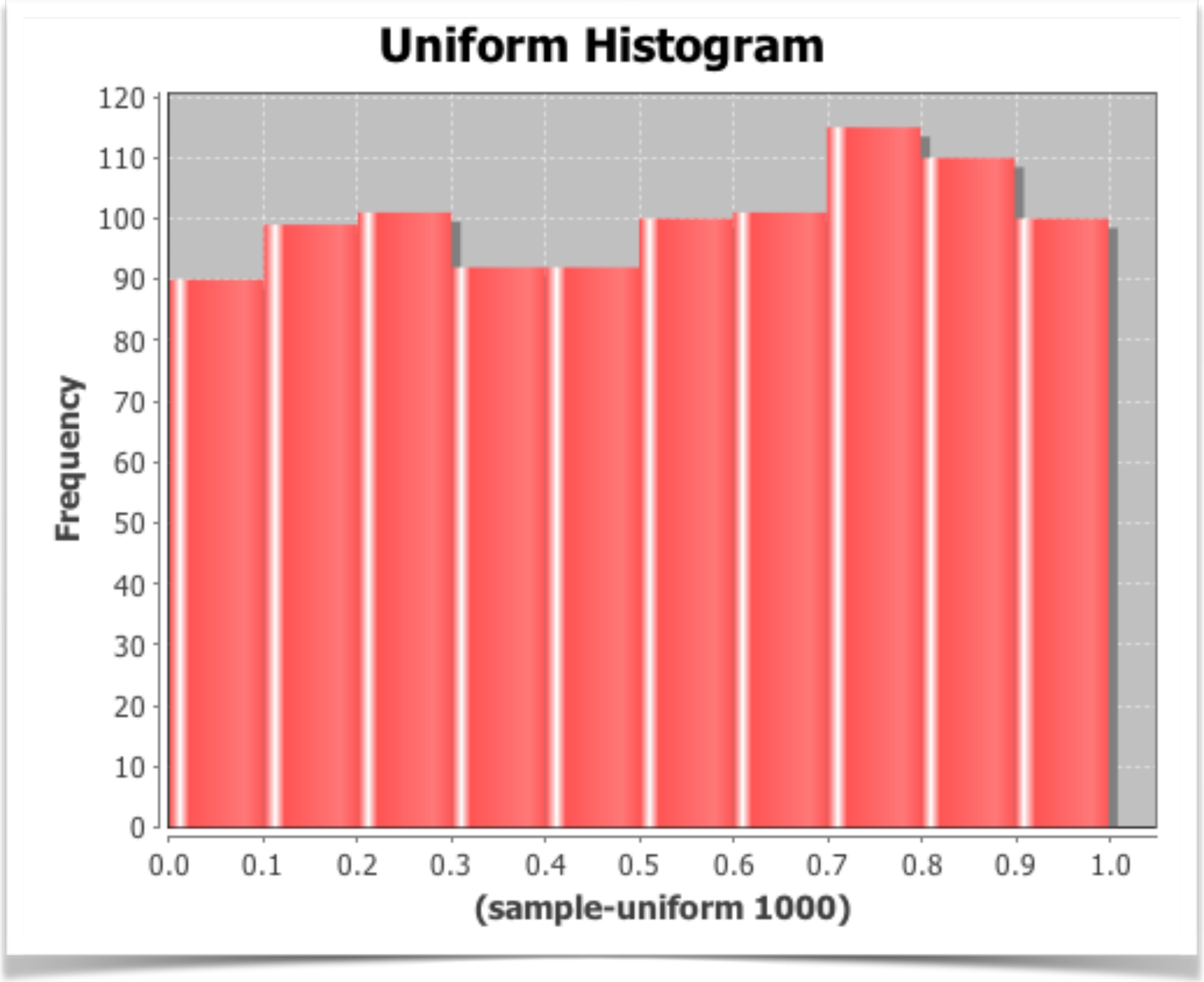
- These **bimodal** histograms have two apparent distinct peaks:

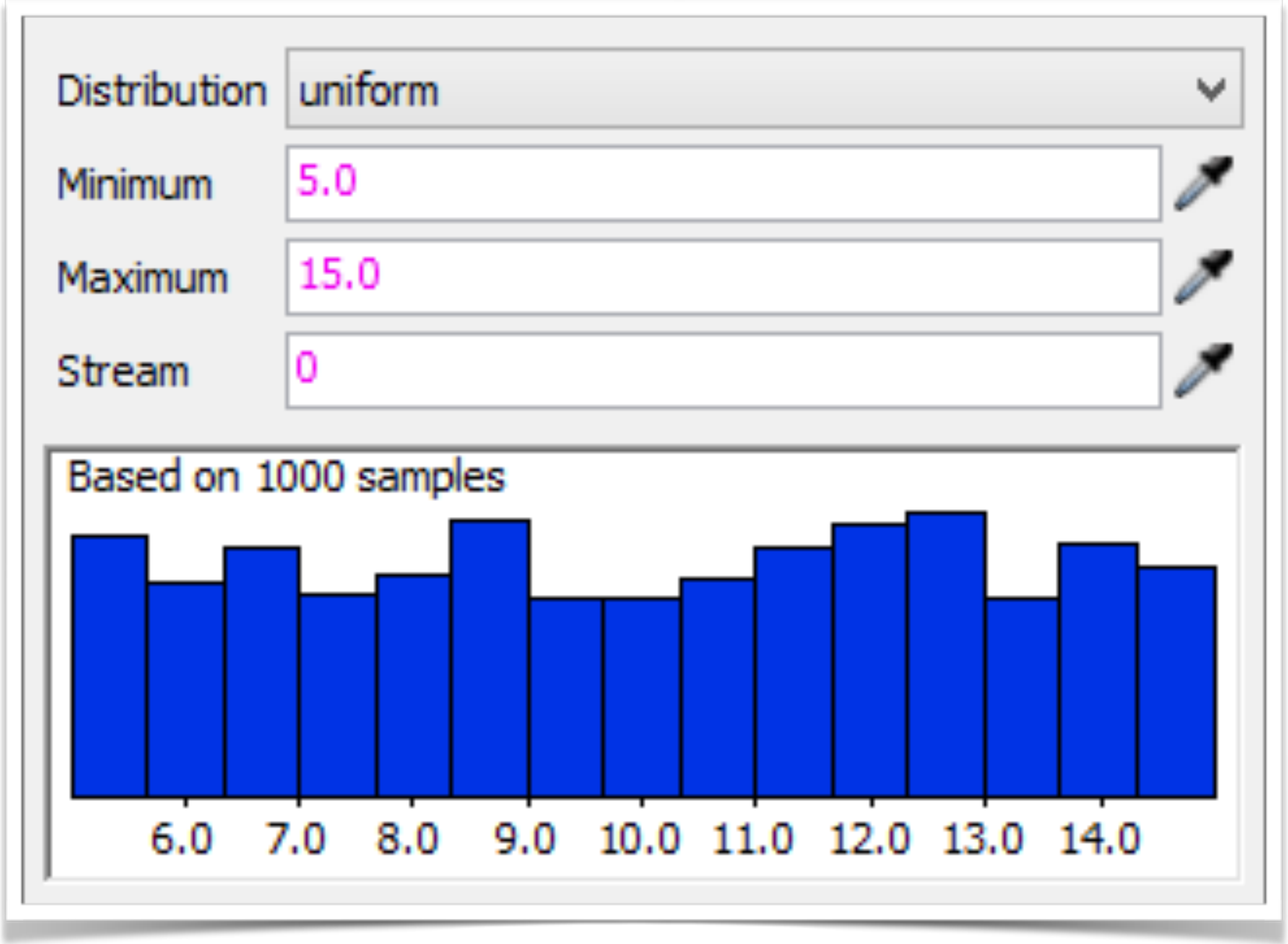




Humps

- A distribution that does not appear to have any significant peaks (mode) and in which all the bars are **approximately** the same height is called a **uniform** distribution.



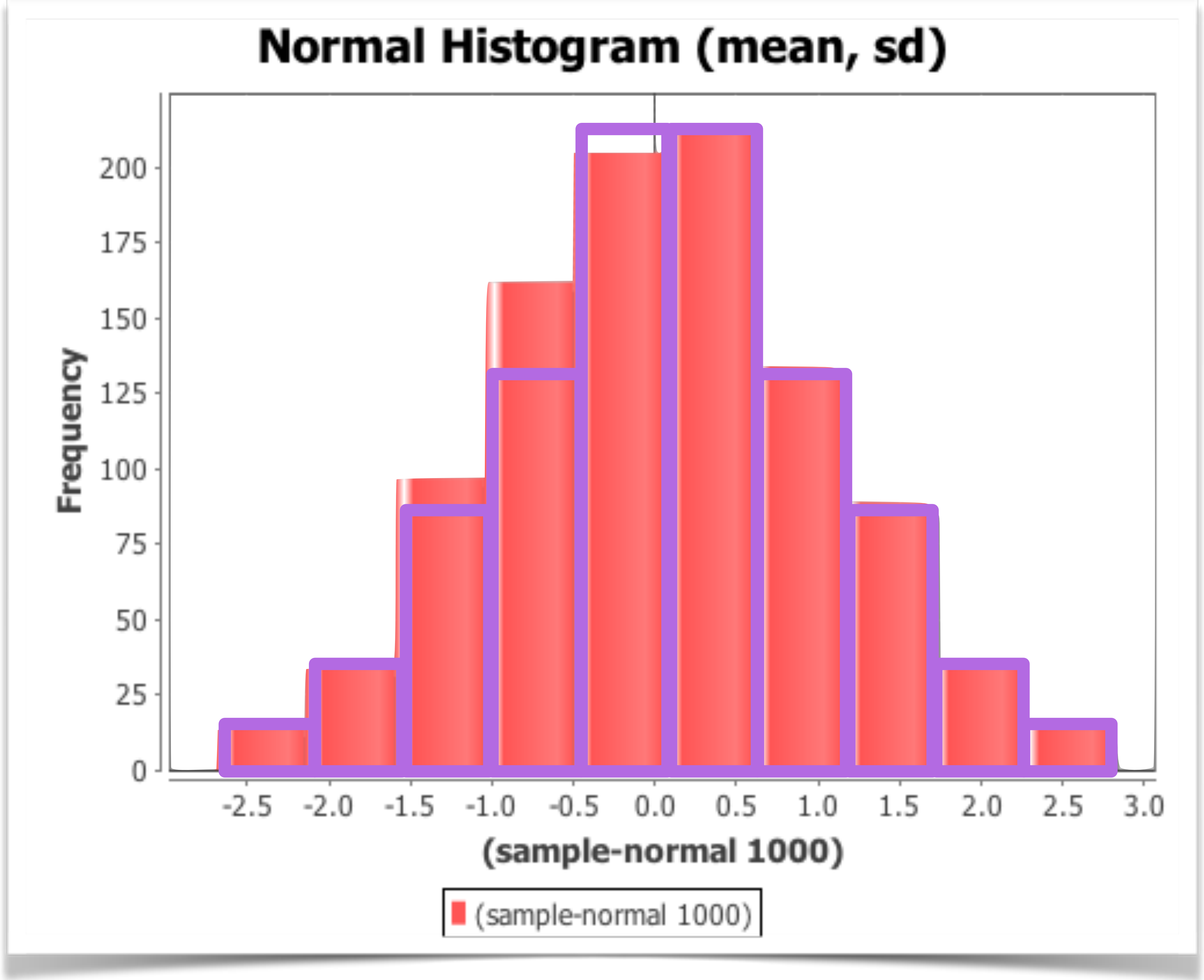


Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Symmetry

2. Is the histogram symmetric?

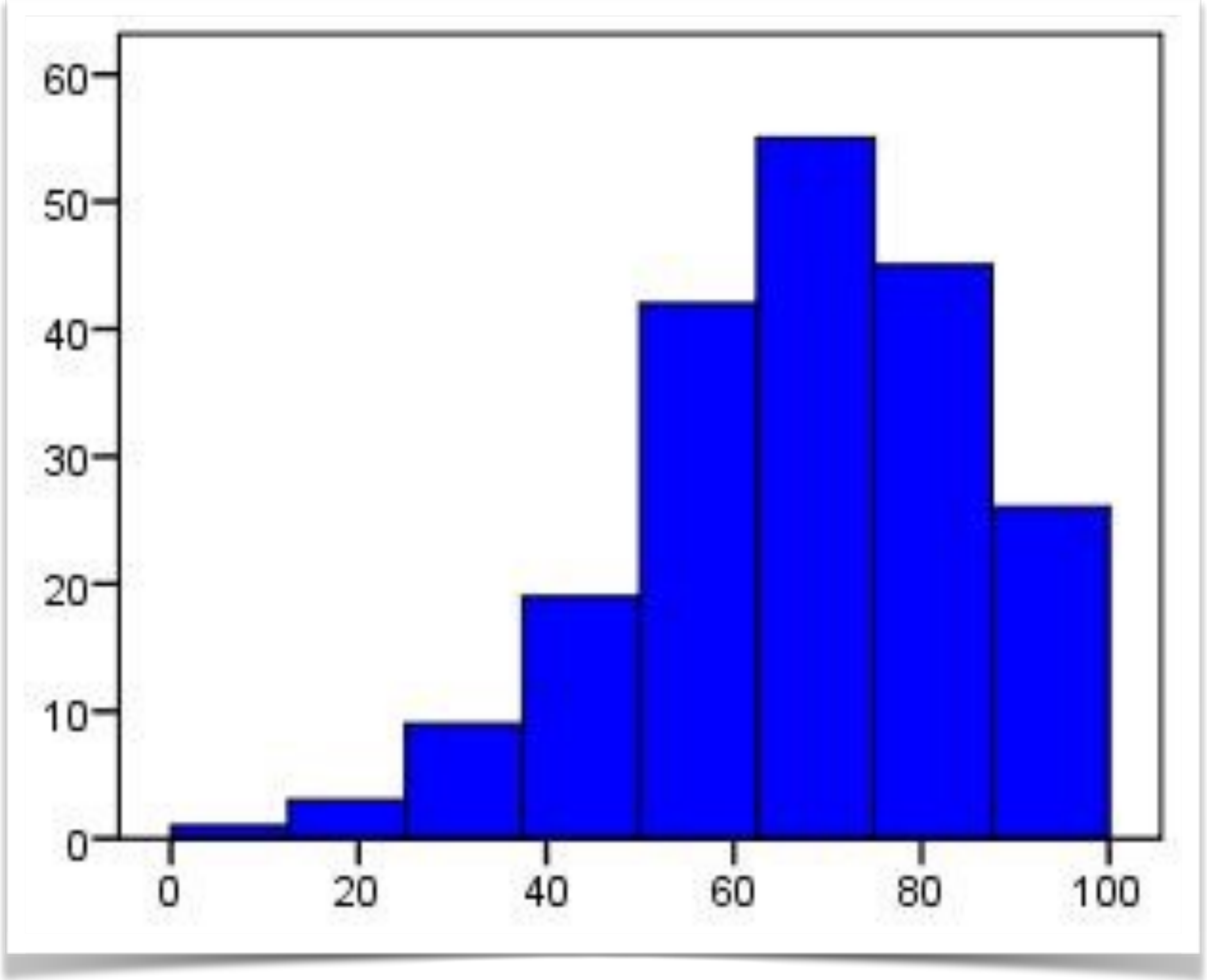
- If you can fold the histogram along a vertical line through the middle and have the halves match pretty closely, the histogram is relatively symmetric.



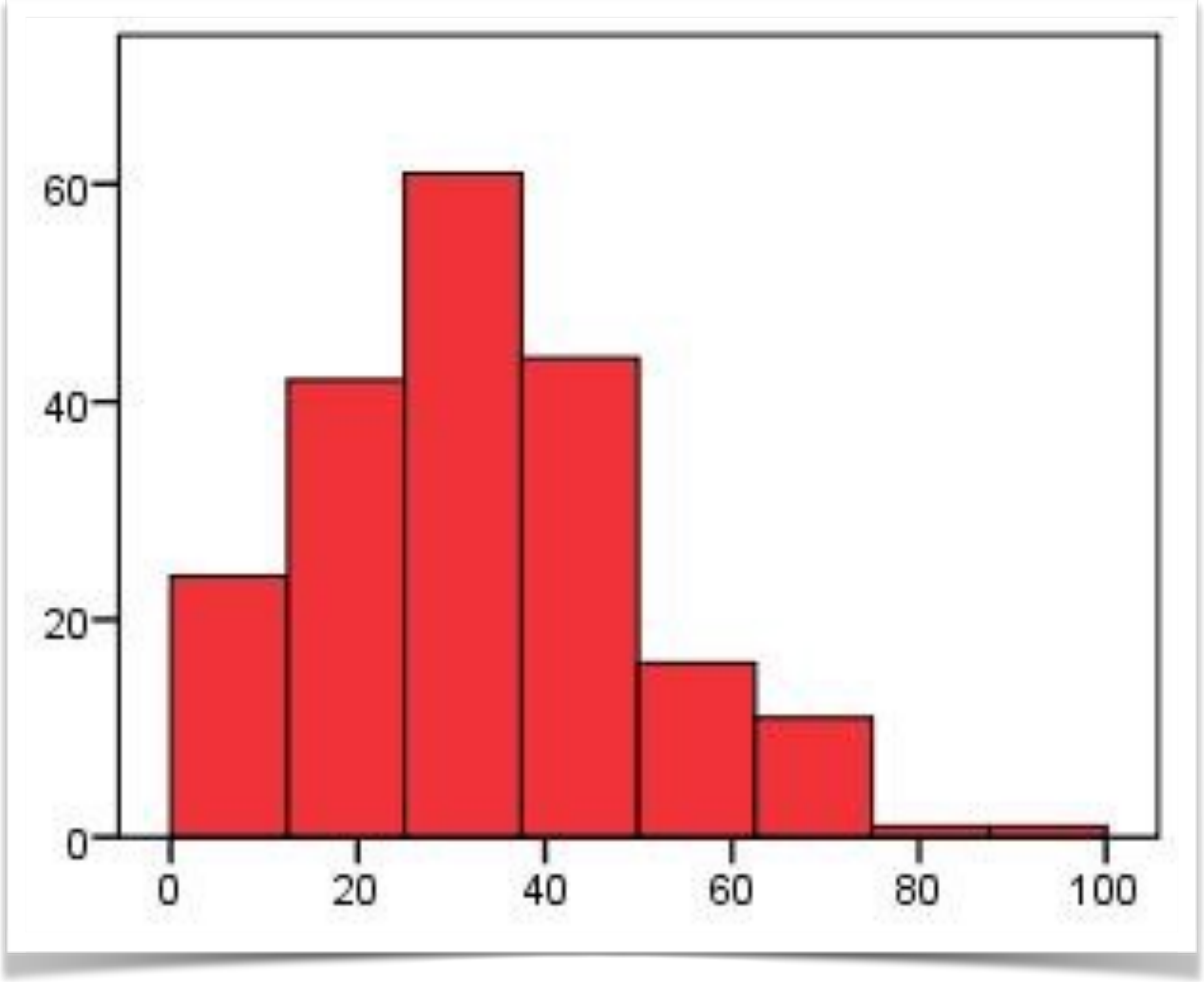
Symmetry

- The (usually) thinner ends of a distribution are called the **tails**. If the tail on one side stretches out farther than the other, the histogram is said to be **skewed** to the **side of the longer tail**.

Negatively (left) skewed



Positively (right) skewed



Anything Unusual (gaps, outliers)?

3. Do any unusual features stick out?

- Sometimes the unusual features indicate something of interest or importance.

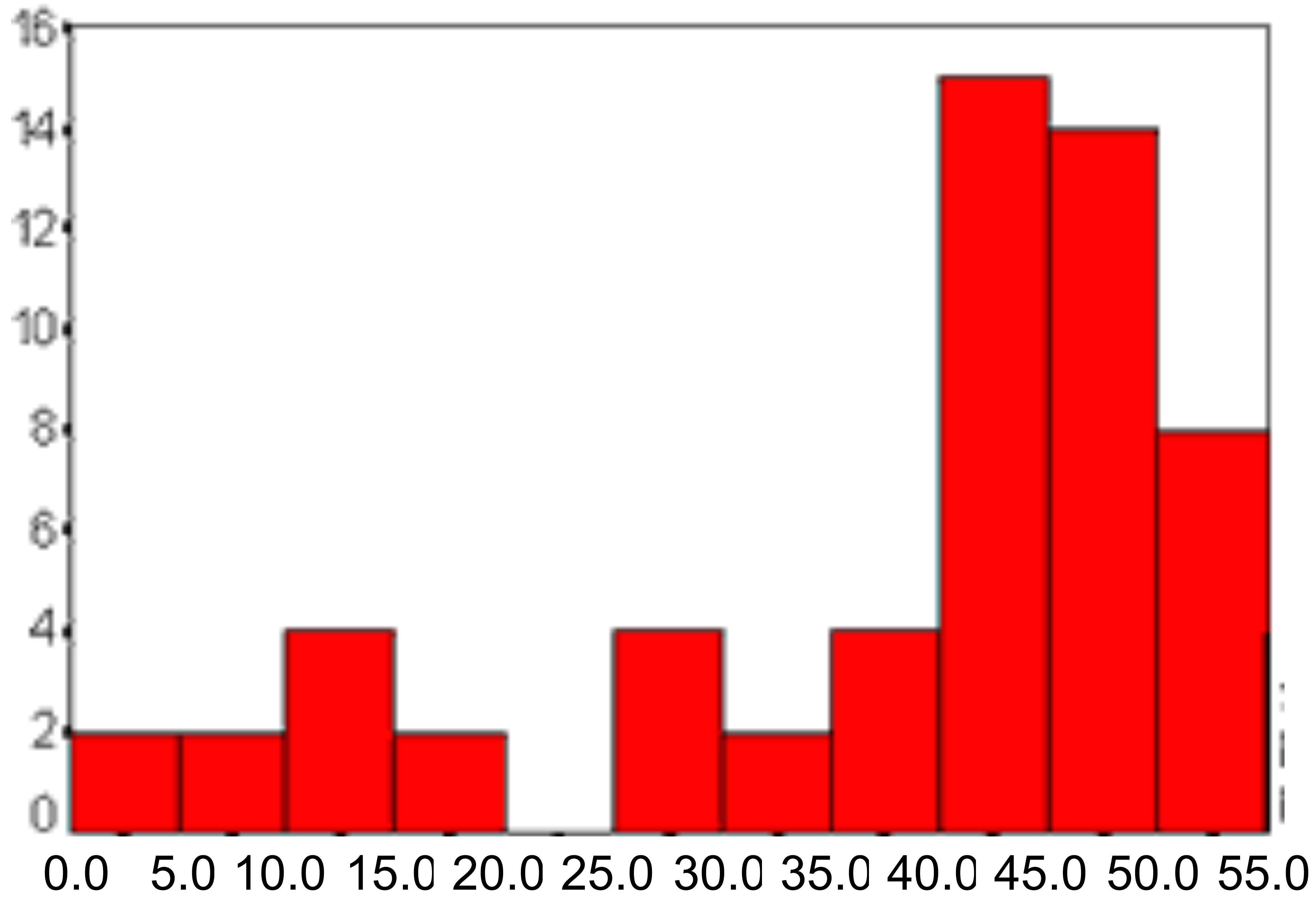
- You must always mention any loners, or **outliers**, that stand off away from the body of the distribution.

- Are there any **gaps** in the distribution? Always acknowledge gaps in the data. Gaps suggest potential multiple groupings.

Anything Unusual?

- The following negatively skewed histogram has a gap in the data and some potential outliers.

  - There is a peak in the data from about 40.0 to 50.0
  - There is a gap in the data between 20.0 to 25.0.
  - The gap in the data may suggest two distributions. Or perhaps it is a function of the low number of data values. Or, possibly, the low values are outliers. We do not speculate, we simply report the gap.



Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Center

The Center of the Distribution?

- If you had to pick a single value to describe a distribution of many data values, what value would you choose?
  - Most people would tell you the "average" value. But that begs the question; "what is the average value". It is relatively simple if your data is symmetric. The center of the histogram is easily determined. What if the distribution is not symmetric, but significantly skewed, or is multi-modal. Then what do you call the "center"?

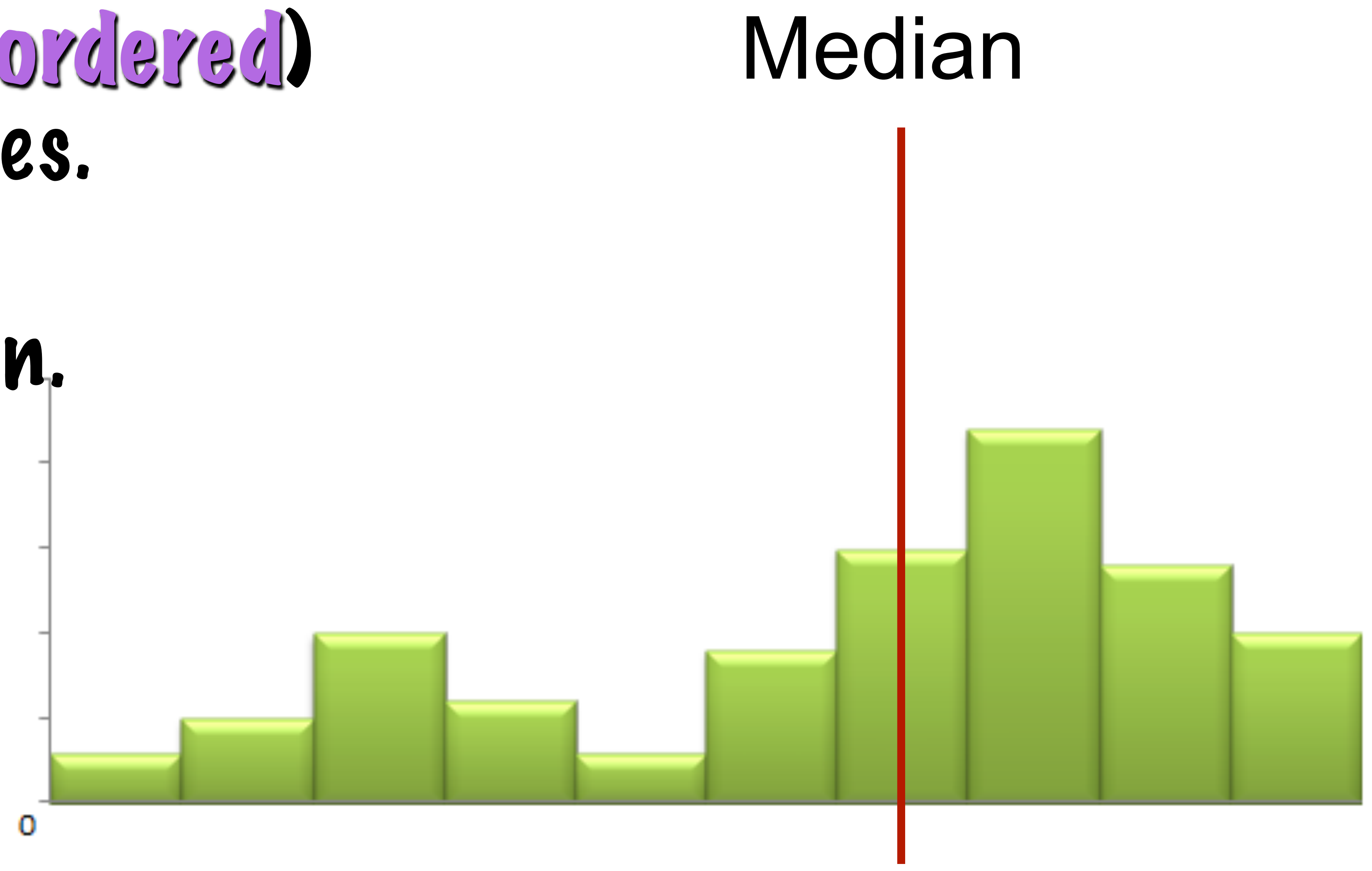
The Center of the Distribution?

- Where is the center?

- Here?

Median

- One measure of center is the median. The median is the value with **exactly half** the data values below it and **exactly half** the data values above it. In other words, the same number of observations fall below the median and above the median.
  - The median has the same units as the data.
  - It is the central value (**once the data values have been ordered**) that divides the histogram into two equal frequency halves.
  - The median is the $\frac{n+1}{2}$ th value in the **ordered** distribution.
  - The value of the median is that the median is resistant to changes in a few data values.



Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Spread

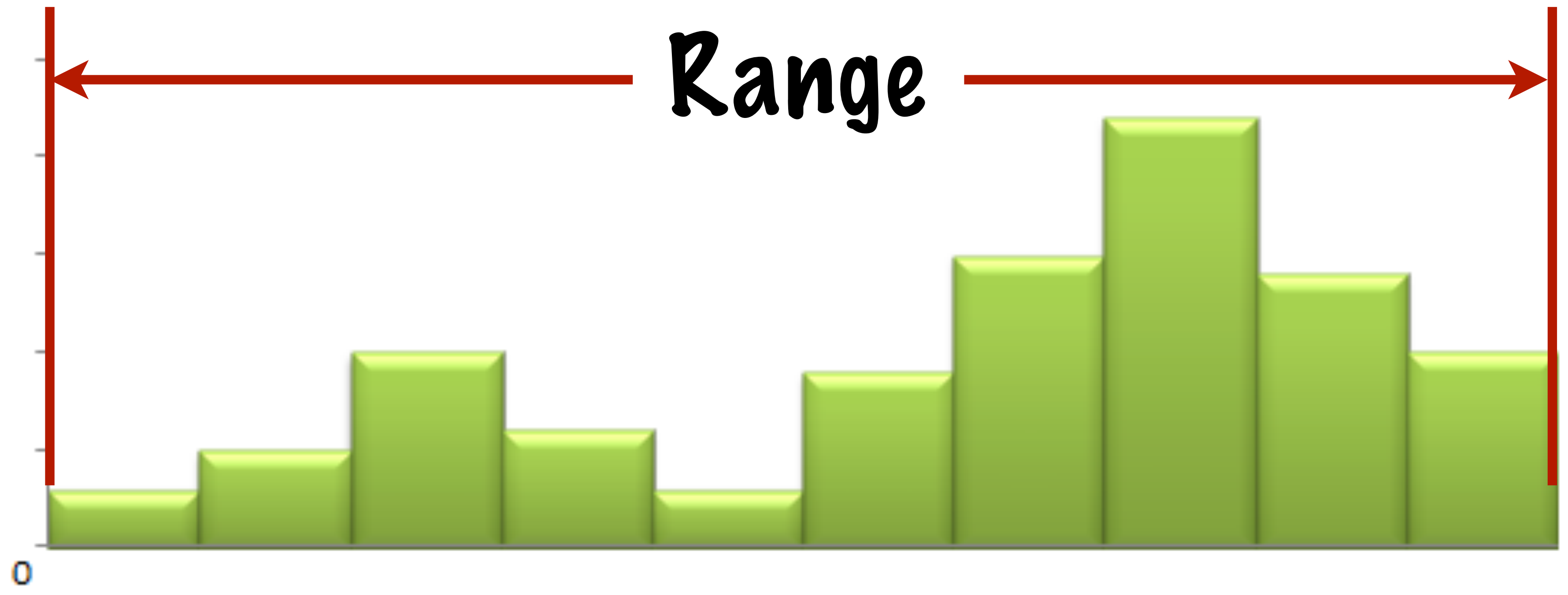
Spread

- Variation matters, in fact, statistics is about variation.
- Are the values of the distribution tightly clustered around the center or more spread out?
- No description of a data distribution is complete without a measure of spread. Always report a measure of spread along with the matching measure of center when describing a distribution.

Spread: Range

- One measure of spread is the familiar "**range**". The **range** of the data is the difference between the maximum and minimum values in the distribution.

Range = max – min



- A problem with the range is that a single extreme value will have a large effect on the range value and, thus, perhaps, not be a good representation of the data distribution.

- So, perhaps, we can find a better measure of variability when using median for our center ....

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Spread: The Interquartile Range

- The **interquartile range (IQR)** ignores data values at the extremes and uses only the **middle 50%** of the data.
  - To find the IQR, we first need to know what quartiles are...
- **Quartiles** divide the **ordered** data distribution into four equal (in frequency) sections.
  - One quarter of the data lies below the first quartile, Q_1
  - One quarter of the data lies above the third quartile, Q_3.
  - Q_1 and Q_3 border the middle 50% of the ordered data.
- The difference between the 3rd quartile and the 1st quartile is the **interquartile range (IQR)**.

IQR = third quartile – first quartile ($Q_3 - Q_1$)

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Spread: The Interquartile Range

- The first quartile (Q_1) is the median of the lower 50% of the data.
- The second quartile (Q_2) is the median of all the data.
- The third quartile (Q_3) is the median of the upper 50% of the data.

- Given the data set 5 52 28 47 50 30 42 12 49 56 find Q_1, Q_2, and Q_3

- First order the data. 5 12 28 30 42 | 47 49 50 52 56

28 ↑ Q_1 | ↑ Q_2 | 50 ↑ Q_3

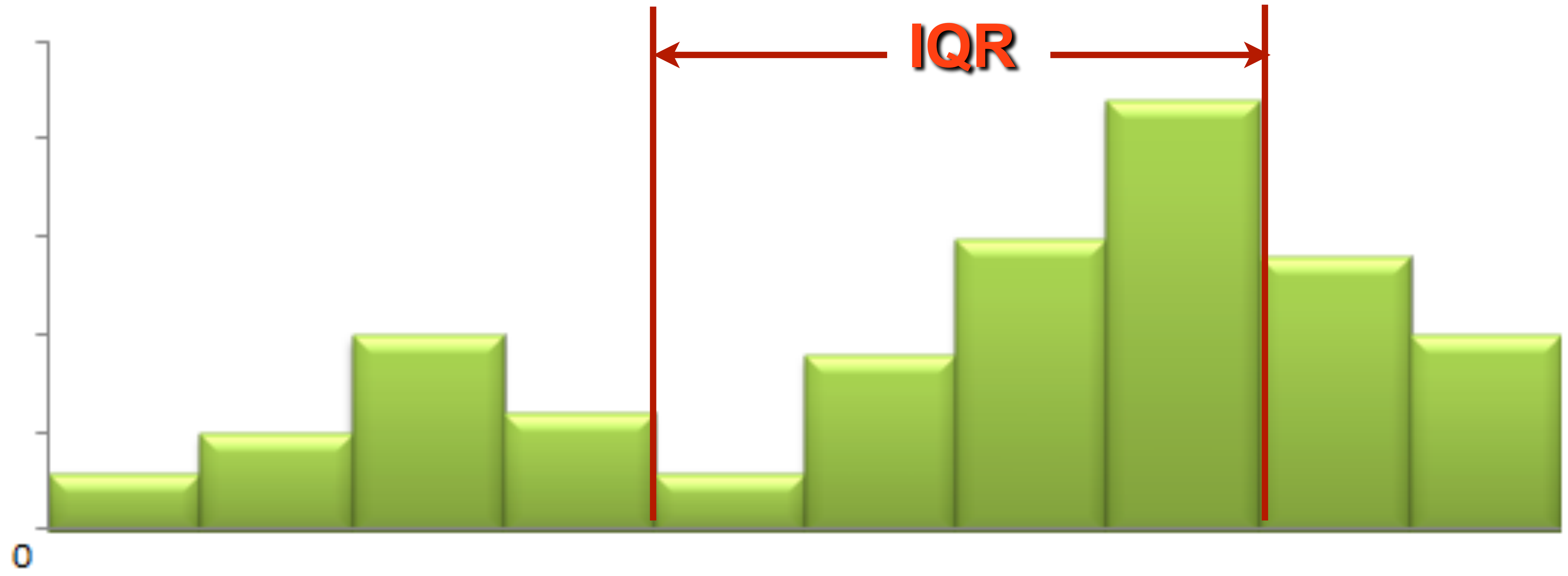
$$Q_2 = \frac{42+47}{2} = 44.5$$

IQR = $Q_3 - Q_1$ = 50 - 28 = 22

Spread: The Interquartile Range

- The first and third quartiles are the **25th** and **75th percentiles** of the data, so…

- The IQR contains the **middle 50%** of the values of the distribution, as shown:

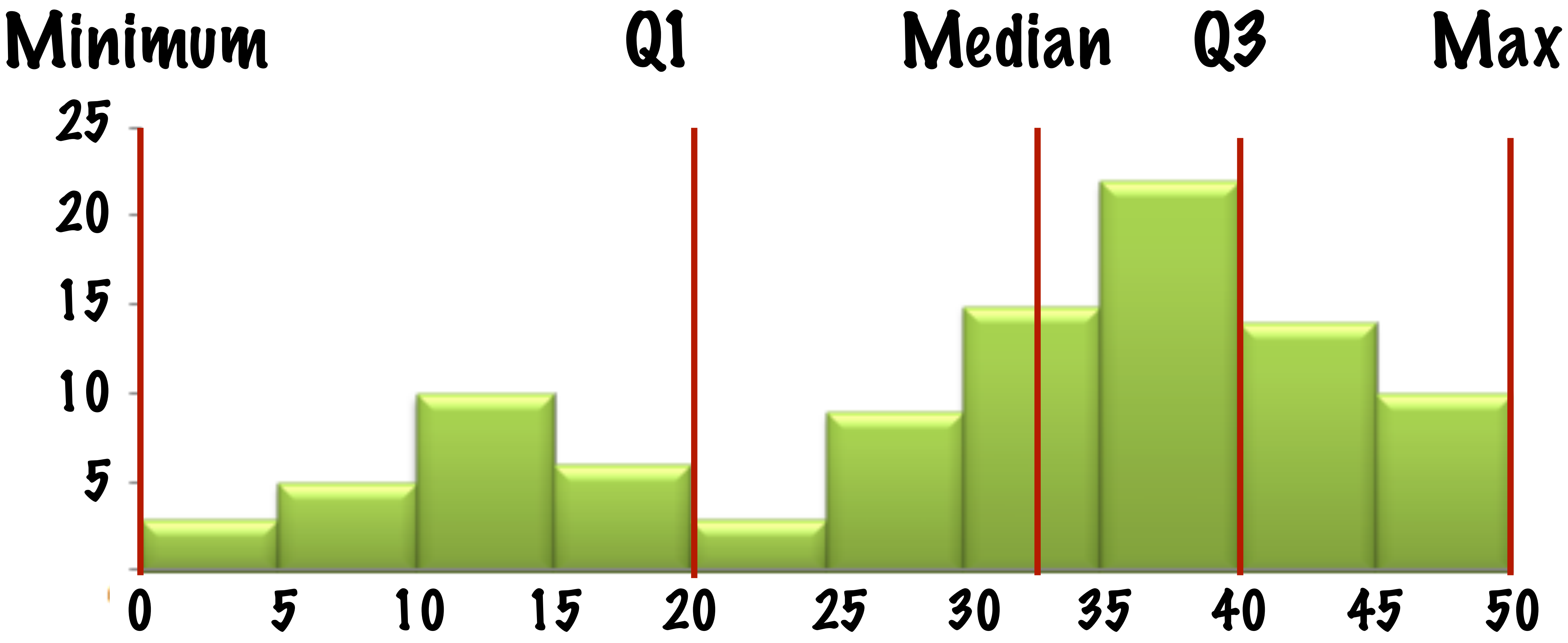


5-Number Summary

- The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum).

- The 5-number summary for the histogram we have been manipulating looks like this:

Maximum	49
Q3	40
Median	32.5
Q1	20
Minimum	0



The Mean

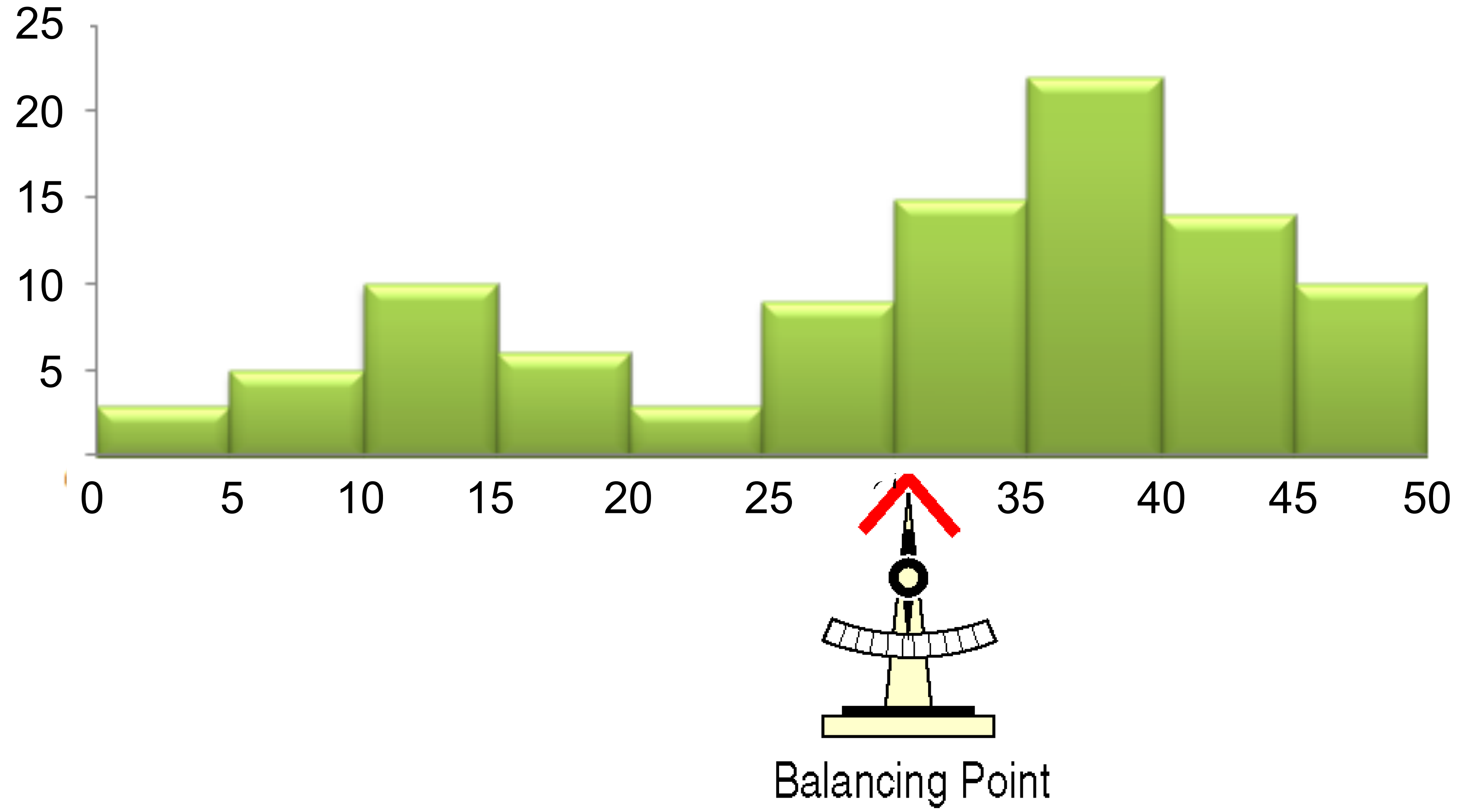
- When we have symmetric data, there is an alternative to the median.
  - We calculate the number that most people mean when they say "average" (arithmetic mean).
    - The Greek letter sigma denotes "sum" and for calculating mean we write:

$$\bar{y} = \frac{total}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- To find the mean, we add up all the values of the variable and divide by the number of data values, n.

The Mean

- The **mean** feels like the center because it is the point where the histogram balances:



Summarizing Symmetric Distributions -- The Midrange

- Another (though rarely used) measure of central tendency is the midrange.

- The midrange is the easiest value to find and its only redeeming feature is that it is quick and easy.

  - midrange = (max + min)/2
    - The mean of the maximum and minimum values

- This is the last time we will discuss the midrange.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Mean or Median?

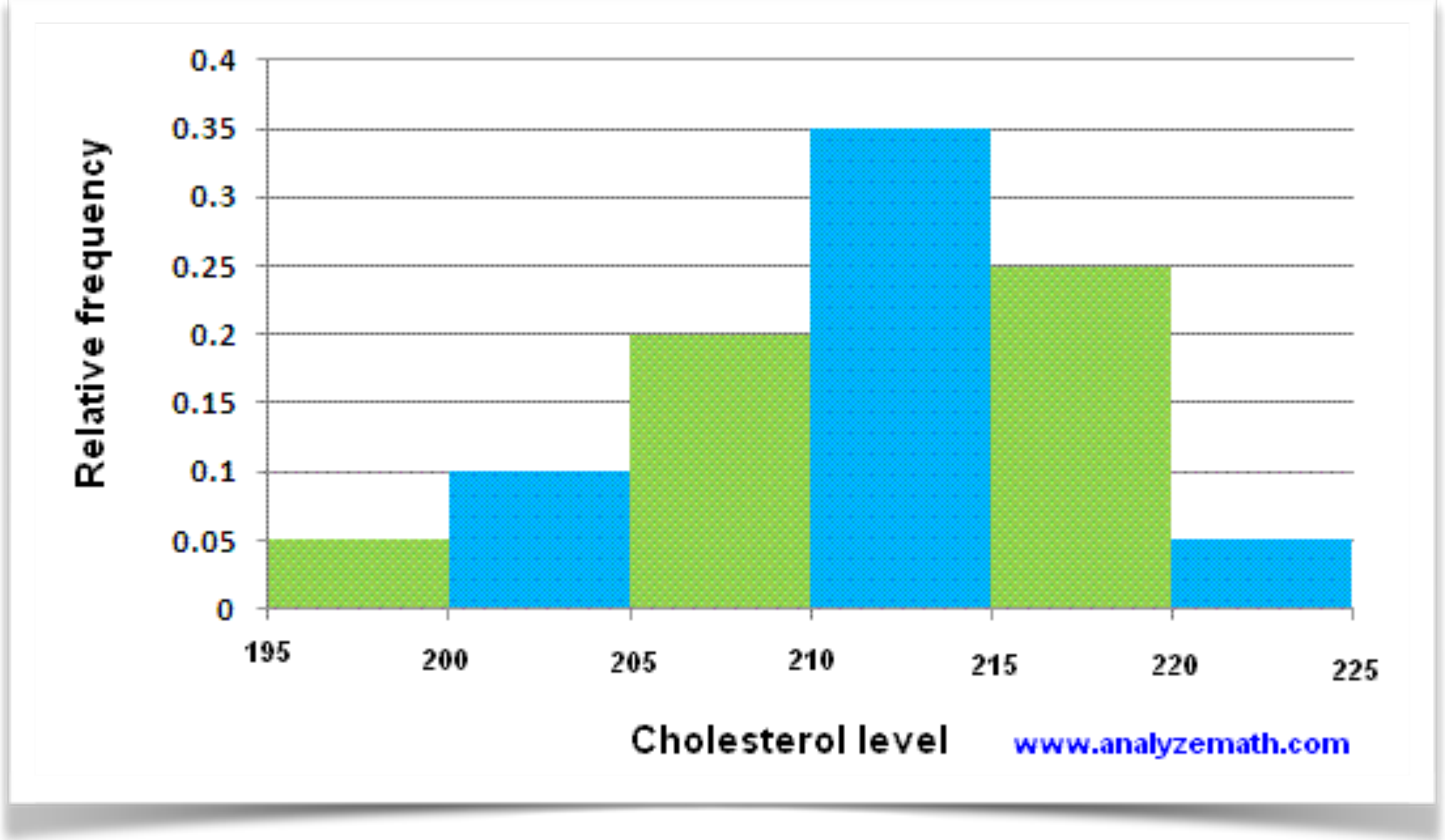
- Because the median considers only the location of values relative to the median, it is **resistant** to values that are extraordinarily large or small; it simply notes that they are "above" or "below" and ignores the actual size or value of the observation.

- To choose between the mean and median, start by looking at the data. If the histogram is **sufficiently symmetric and there are no significant gaps or outliers**, use the **mean**.

- If the histogram is skewed, multi-modal, has gaps or outliers, you are probably better off with the **median**.

Only Theory is Perfect

- Keep in mind that we are not looking for perfection. When choosing between mean and median look for "**sufficiently** unimodal and symmetric" when deciding to use the mean.
  - Do not be too critical of your data. If your data is significantly skewed, has significant gaps, or has significant outliers then you are probably better off using the median.
  - It is a judgement call, your judgement will improve as you become more familiar with the statistical methodology and what you are hoping to communicate.
- Acceptable for the **mean**



The Standard Deviation

- A measure of spread more common than IQR, and used when **mean** is the measure of center, is the **standard deviation.** The standard deviation uses how far each data value is from the mean.
  - A **deviation** is the distance that an observation is from the mean.
    - There is a huge problem with finding the "average deviation".
    - Adding all deviations together would total zero.
    - So we **square** each deviation and find an "average" of the **squared deviations**.
    - Hence, a "**standard deviation**".

The Standard Deviation

- The **variance**, notated by s^2 (for a sample), is found by summing the squared deviations and (kinda) averaging them:

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$$

- The variance plays a crucial role in statistics, but the fact that it is measured in squared units makes is problematic.

- To resolve the "squared" problem we simply take the square root of the variance to get ...

The Standard Deviation

- The **standard deviation, s;** the square root of the variance and measured in the same units as the original data.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(y - \bar{y}\right)^2}{n-1}}$$

The Standard Deviation

- To find variance we must first find a few values ...

  Step 1 - find the **mean** of the data.

  Step 2 - find the **deviation** for each datum value $(x - \overline{X})$.

  Step 3 - **square** each deviation. $(x - \overline{X})^2$.

  Step 4 - add all the squared deviations. This is the **Sum of Squared deviations** $\Sigma(x - \overline{X})^2$.

  (or **sum of squares** or **SS**).

  Step 5 - divide the sum of squared deviations by the number of data values (**minus 1 for a sample**) to find the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(y - \overline{y}\right)^2}{n - 1}}$$

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Calculating Standard Deviation

- 1. mean ($\bar{x}$)
- 2. deviations ($x - \bar{x}$)
- 2. squared deviations $(x - \bar{x})^2$
- 4. sum of squared deviations (ss) $\Sigma(x - \bar{x})^2$
- 5. variance

$$s^2 = \frac{\sum_{i=1}^{n}\left(y - \bar{y}\right)^2}{n-1}$$

- 6. standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(y - \bar{y}\right)^2}{n-1}}$$

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

The Standard Deviation

- For a population

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

x = datum value
μ = population mean
N = population size

- For a sample

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

x = datum value
$\bar{x}$ = sample mean
n = sample size

- When calculating the variance for a **sample**, the sum of squares is not divided by n (the sample size) but by **n - 1**. This value is known as the "**degrees of freedom**". The mean has been determined, therefore only **n - 1** values can change, the nth term is fixed by the mean.

The Standard Deviation

- The standard deviation is simply the square root of the variance.

- Population

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

- Sample

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

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

The Standard Deviation

 25 30 65 70 40 55 60 35 50 95 70

x
25
30
65
70
40
55
60
35
50
95
70
$\bar{X} = \frac{595}{11} = 54.1$

SS = 4341.91

$$s^2 = \frac{SS}{df} = \frac{\sum\left(X - \bar{X}\right)^2}{n-1}$$

$$= \frac{4341.91}{10} = 434.19$$

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

$$= \sqrt{434.19}$$

$$= 20.8372$$

Thinking About Variation

- Since Statistics is about variation, spread is an incredibly important fundamental concept of Statistics.
  - When the data values are tightly clustered around the center of the distribution, the IQR and standard deviation will be small.
  - When the data values are scattered far from the center, the IQR and standard deviation will be large.

Start by Drawing a Picture

- When communicating about quantitative variables, **start by making a histogram or stem-and-leaf display and describe the shape of the distribution.**

Shape, Center, and Spread

- **Humps, symmetry,** and **unusual features**
- Always report the **shape** of its distribution, along with a **center** and a **spread**.
  - If the shape is significantly **skewed**, report the **median** and **IQR**.
  - If the shape is sufficiently **symmetric**, report the **mean** and **standard deviation**.
- To be certain, it is certainly acceptable to report both mean and standard deviation as well as the median and IQR.

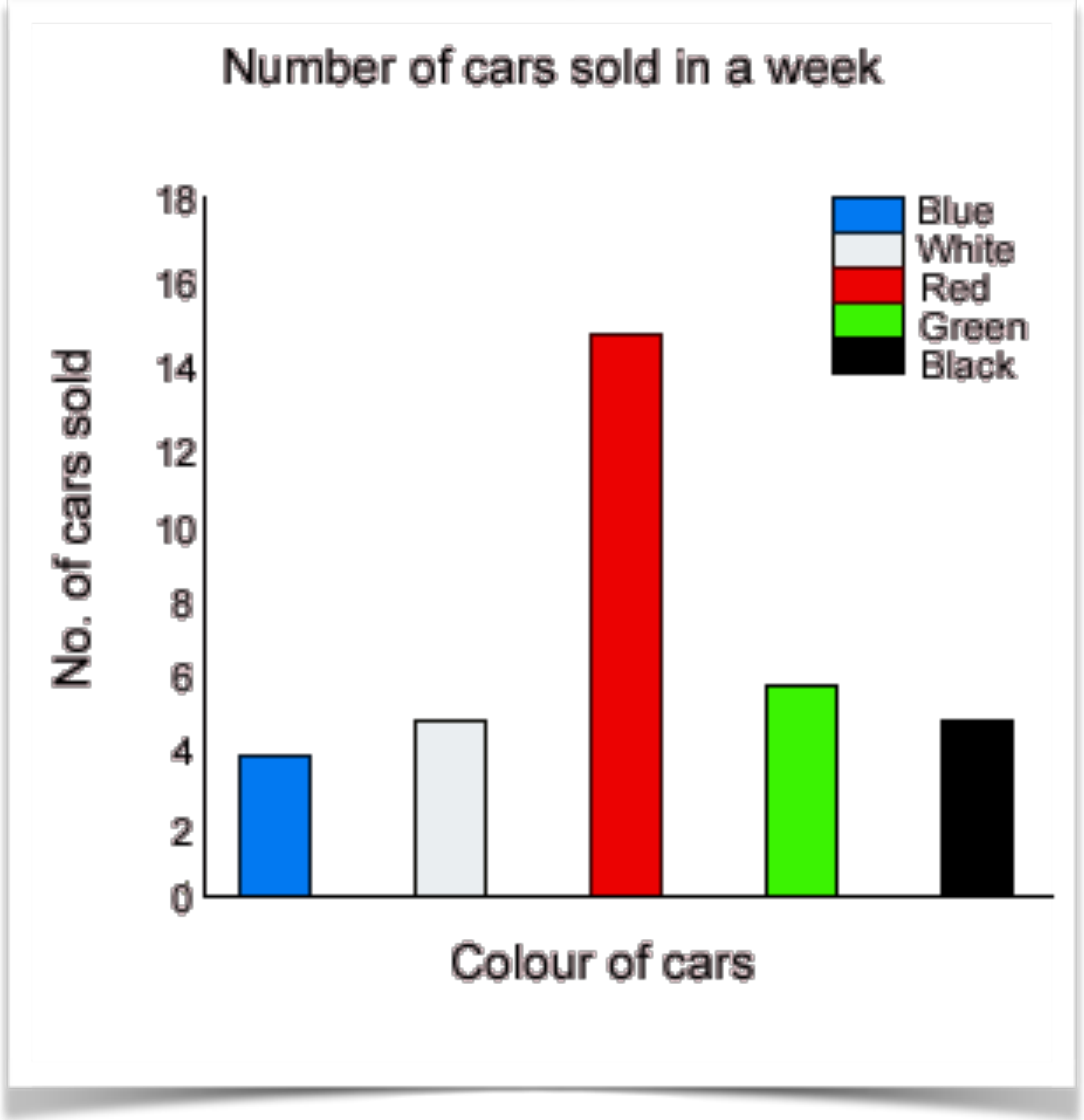
Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

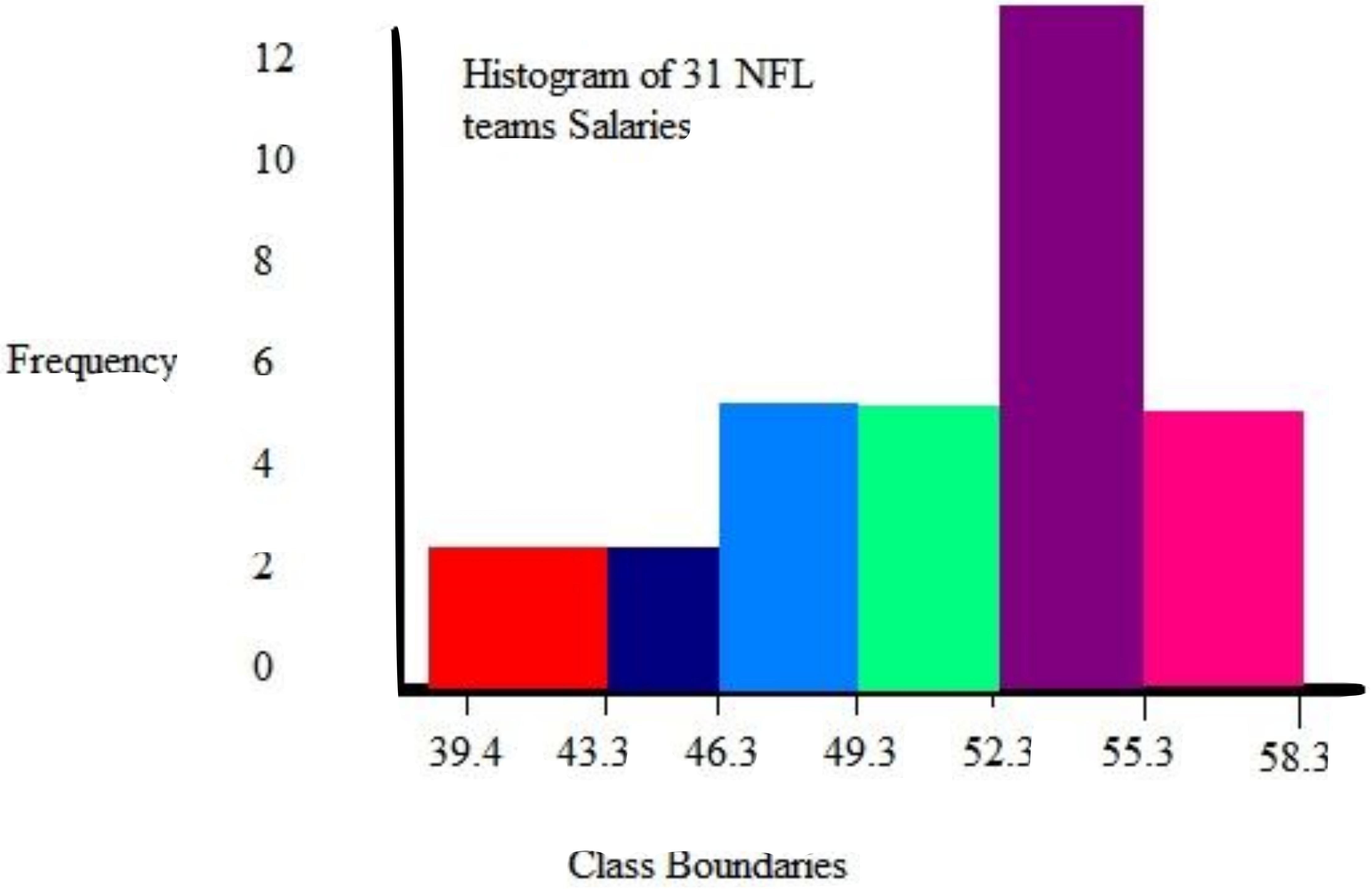
What About Unusual Features?

- If there are multiple modes, there may be an interesting reason. If you identify a reason for the separate modes, it may be a good idea to split the data into separate groups and examine the groups individually.
- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers **present** **and** with the outliers **removed**.
- Note: The median and IQR are "resistant", meaning not likely to be affected by the outliers.

Mistakes of Which to Be Aware

- Do not confuse bar charts with histograms. Bar charts are for categorical data, histograms are for quantitative, continuous data.
  - The bins of bar charts have no order so there is no describing shape, center, or spread.
  - The bins of histogram have order and we can describe shape, center, and spread.

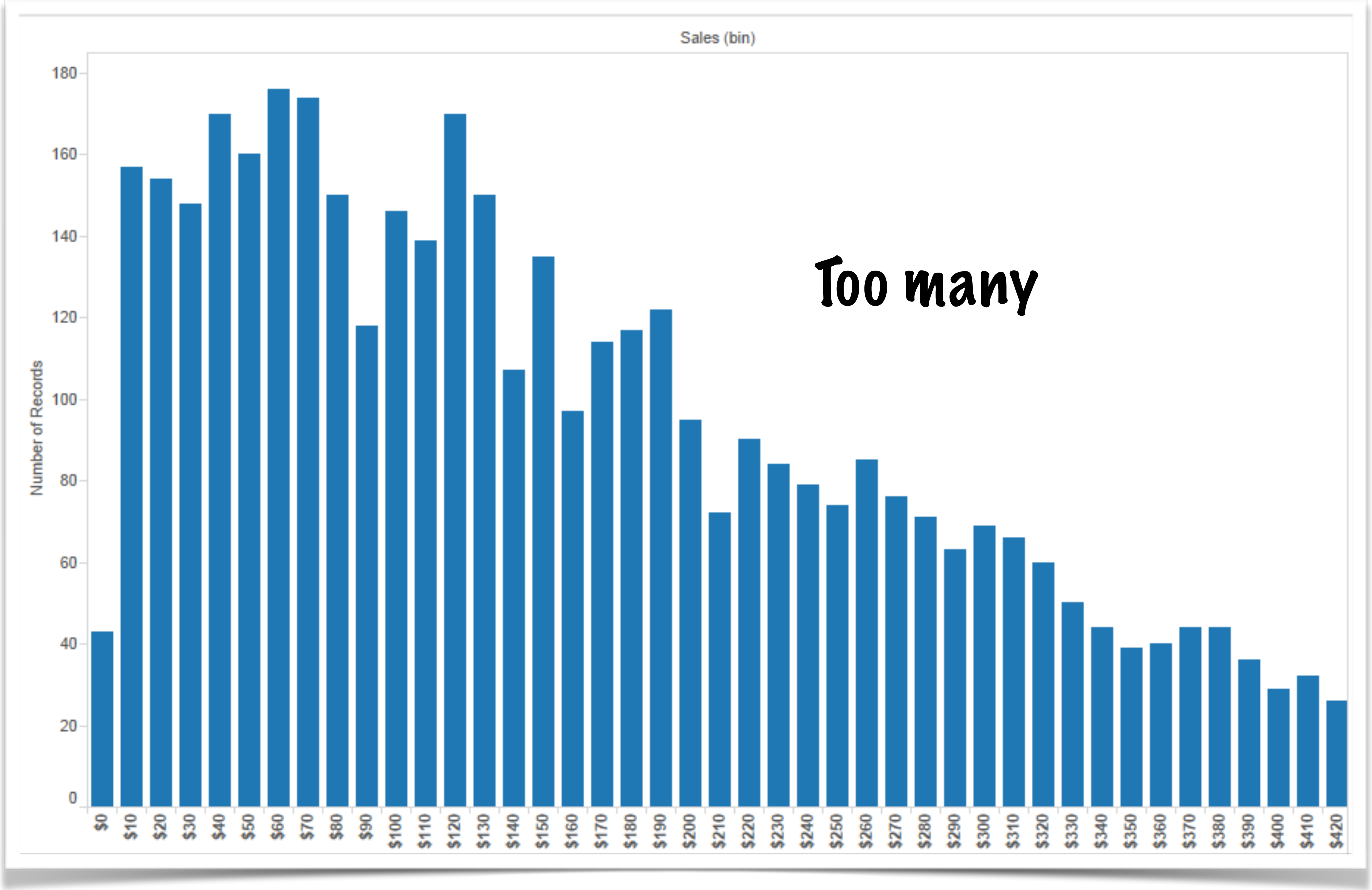




Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Mistakes of Which to Be Aware

- Choose a bin width appropriate to the data.
  - Changing the bin width changes the appearance of the histogram:

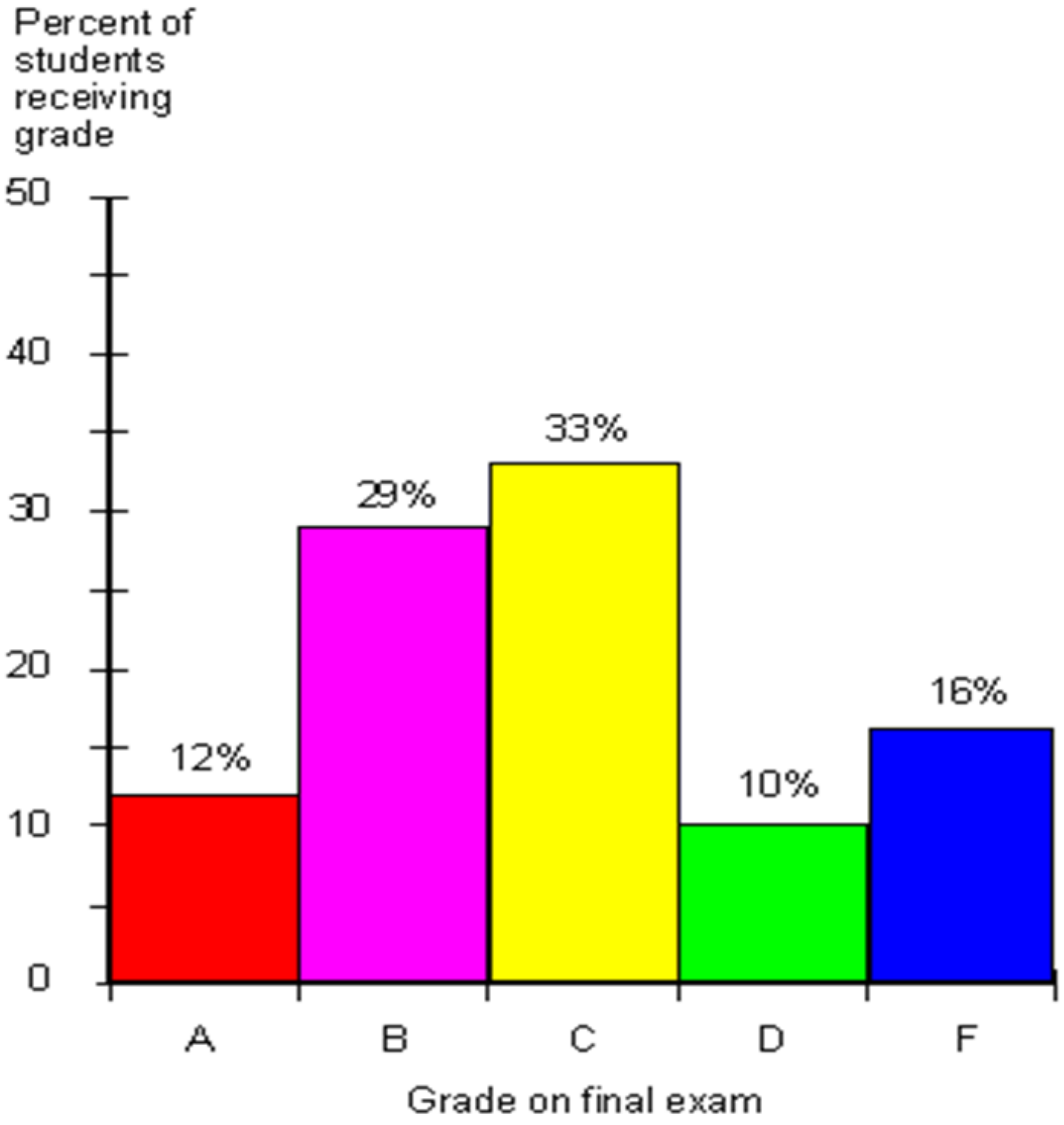


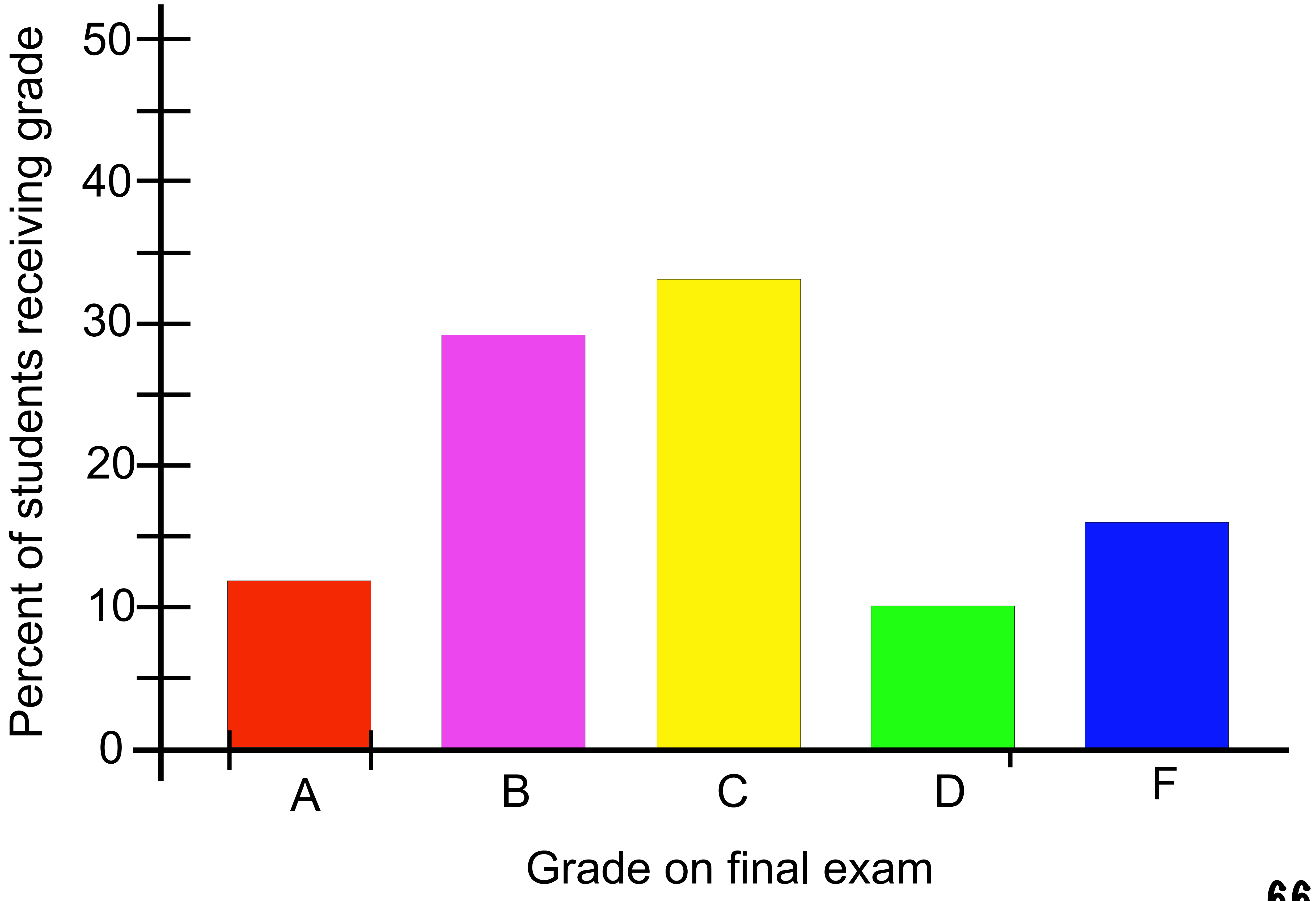
- Never, Never, Never

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

Mistakes of Which to Be Aware

- Do not use a histogram when you should use a bar chart.





Mistakes of Which to Be Aware

- Report data values to **4 decimal places**, for percentages use 2 decimals. **I will ding you** for reporting values to too few decimal places.

- Always draw a picture of some kind. As we go forward you will nearly always be drawing one or more graphs when reporting results.

- **Do not** round in the middle of a calculation, let the calculator remember. **I will ding you** if your values are too far off due to rounding error.

- Be **aware** of outliers. Report results with and without the outlier. You may want to use median, rather than mean.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

TI-84

- Enter the following data into a list.

88 56 80 60 76 72 68 80 64 80 84 64 68 72 80 76 72 76 84 76 72 68 68 64

STAT | 1:EDIT | Select List | "Enter first datum" | ENTER | Repeat to end of list | 2ND | QUIT

- To draw the histogram

STAT PLOT: 2nd | y= | Enter | ON | TYPE: (histogram) | XList | L_1: 2nd | 1 | Freq: 1 | 2nd | Quit | Zoom | 9

TI-84

- To change the appearance of the histogram
  - Window
  - Xmin = smallest x in window
  - Xmax = largest x in window
  - Xscl = bin width
  - Ymin = smallest Y, a negative value will lift the graph off the bottom of the screen.
  - Ymax = largest frequency you expect
  - Yscl = 1
  - Xres = 1
  - ΔX= let the calculator deal
  - **Trace**

Practice

- Here are some websites that will allow you to play with histograms.

http://www.shodor.org/interactivate/activities/Histogram/

http://www.amstat.org/publications/jse/v6n3/applets/histogram.html

http://statweb.calpoly.edu/chance/applets/Histogram.html

Bigfoot

- Record your shoe size on a dot plot on the board. Females use red, males use black or blue.
  - I am aware, just do it.

Continuous, Quantitative Data

- With your partner, measure your resting heart rate (pulse) by counting the number of beats in one minute.
  - Record your pulse on the board.
  - Create a grouped frequency distribution for the class.
  - Create a histogram of the class distribution.

Objectives: Students organize and describe distributions of data by using histograms, bar graphs, dotplots, stem-and-leaf displays, and box-and-whisker plots. Students will calculate the mean, variance, standard deviation, median, range and interquartile range of a distribution of data values.

M&M Mean and Standard Deviation

- Weigh a scoop of 50 m&ms. Remember to subtract the weight of the cup.
- Record the weight on the chart on the whiteboard.
- Copy the weights recorded onto your table and do the calculations.

What Does an M Weigh?

We are going to find the typical weight and how much variability there is in 50 m&m's. Scoop and weigh 50 m&ms and record the weight (4 decimals) on your table add the class results to your table. Complete the table, then calculate the mean and standard deviation of the weights.

Bag #	x	$x-\bar{x}$	$(x-\bar{x})^2$	Bag #	x	$x-\bar{x}$	$(x-\bar{x})^2$	Bag #	x	$x-\bar{x}$	$(x-\bar{x})^2$
1				14				27			
2				15				28			
3				16				29			
4				17				30			
5				18				31			
6				19				32			
7				20				33			
8				21				34			
9				22				35			
10				23				36			
11				24					$\sum_{i=1}^{n} x_i =$		$\sum_{i=1}^{n}(x_i-\bar{x})^2 =$
12				25							
13				26					$\bar{x} =$		