Stat 145

Chapter 2

DESCRIBING DATA – USING GRAPHS

DESCRIBING DATA – USING GRAPHS

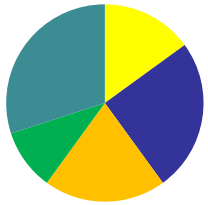
GRAPHS ARE USED TO DESCRIBE:

1. The shape or distribution of the data
   - Are there an equal number of scores from low to high?
   - Or are the scores clustered at the low or high end, or more towards the middle?
2. A central point in the data set where most of the scores are clustered around
3. The correlation of two or more variables
   - As one variable changes, how much change can be predicted in a second variable?

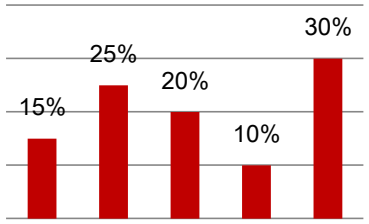
DESCRIBING DATA – USING GRAPHS

FIVE COMMON GRAPHS USED TO DESCRIBE DATA ARE:

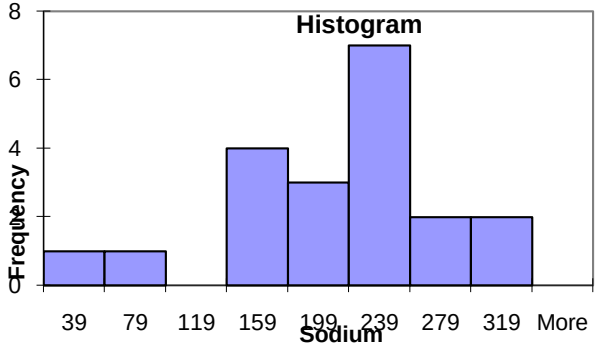
1. Pie Charts



2. Bar Graphs



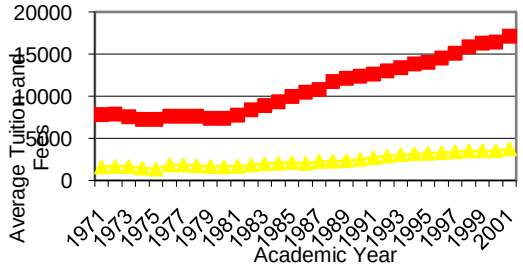
3. Histograms



4. Stem Plots

23	0
24	0
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0236

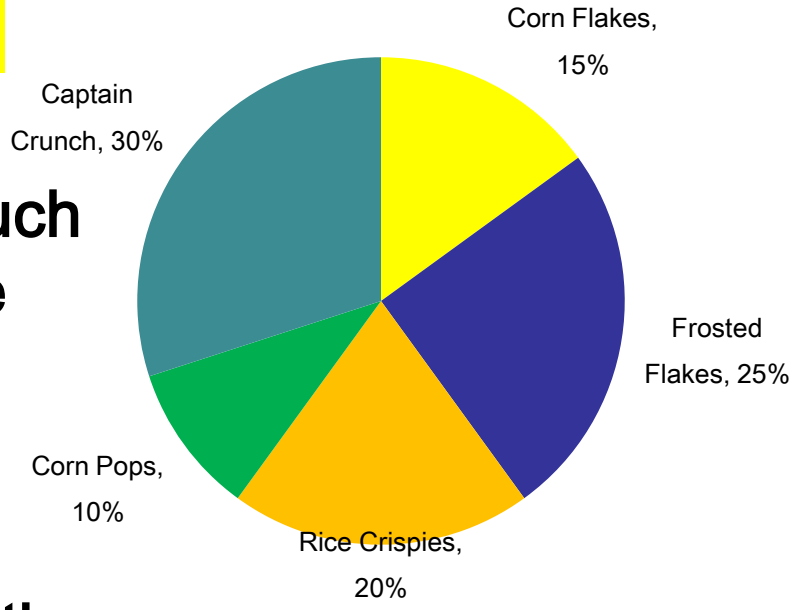
5. Time Plots



Chapter 2

DESCRIBING DATA – PIE CHARTS

PIE CHARTS ARE USED TO DESCRIBE:



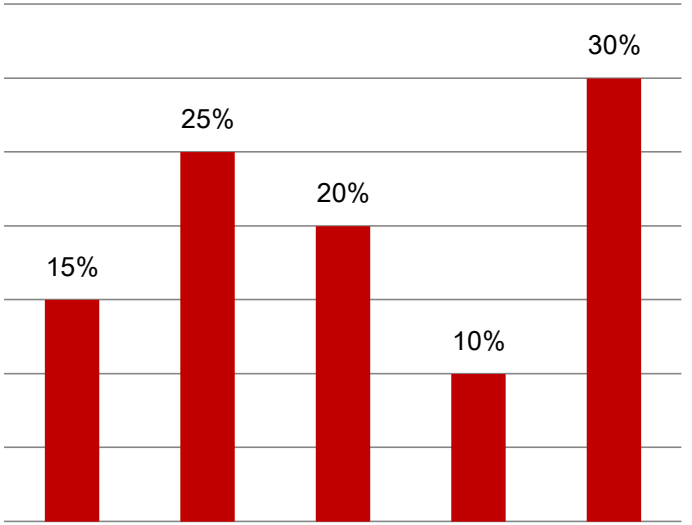
1. **Qualitative or Categorical data** such as political party, gender, favorite brand of cereal, year in school

2. Each piece of the pie represents the **frequency**, **proportion** or **percentage** observed for each group or category

Watch a tutorial on how to create a pie chart in Excel

DESCRIBING DATA – BAR GRAPHS

BAR GRAPHS ARE USED TO DESCRIBE:



1. **Qualitative or Categorical** data such as political party, gender, favorite brand of cereal, year in school

2. The height of each bar represents the **frequency**, **proportion** or **percentage** observed for each group or category

Watch a tutorial on how to create a bar graph in Excel

Chapter 2

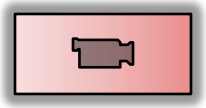
DESCRIBING DATA – HISTOGRAMS

Histograms are used to describe:

1. Quantitative data such as height in inches, strength measured in pounds lifted, calories burned, amount of product yielded measured in ounces

2. A continuous variable that has been divided into equal intervals

3. The height of each bar represents the frequency of each interval observed for each group or category



Watch a tutorial on how to create a histogram in Excel

DESCRIBING DATA – HISTOGRAMS

How man intervals (i.e., bars) should a histogram have?

1. As a general rule, it is recommended that histograms have 5 to 15 intervals
   - Typically, the larger the range, the more intervals needed
   - The width of the intervals are created by
     - First sorting the data from lowest to highest
     - Then dividing the range by the number of desired intervals and
     - Then rounding to the unit that makes most sense
   - It's a good idea to have intervals that are in increments of 5 and 10 units
   - Try to avoid interval increments that are in decimal numbers such as 4.5 or .25 or 7.75

DESCRIBING DATA – HISTOGRAMS

How to calculate the interval width for a histogram:

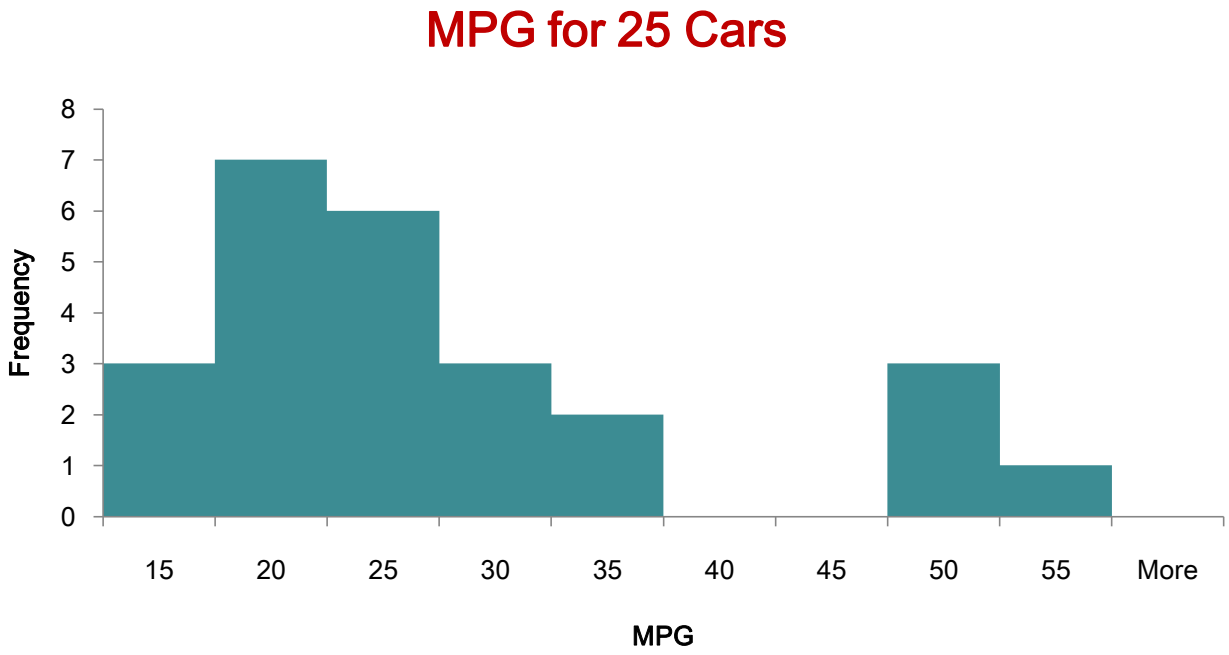
EXAMPLE: Suppose 50 cars are measured for fuel efficiency. The car with the best gas mileage got 52 MPG versus the car with the worst gas mileage got just 12 MPG.

- First, the data must be sorted from lowest to highest
- Next, the range is 52 – 12 = 40
- Let's say we want seven intervals
- To get the interval width, divide the range by the desired number of intervals:
  - 40 / 7 = 5.71
- Round to the unit that makes most sense
  - Although it may be tempting to round to 6, it's better to round to 5's and 10's, so we'll round to 5
- Each interval will be 5 MPG wide

DESCRIBING DATA – HISTOGRAMS

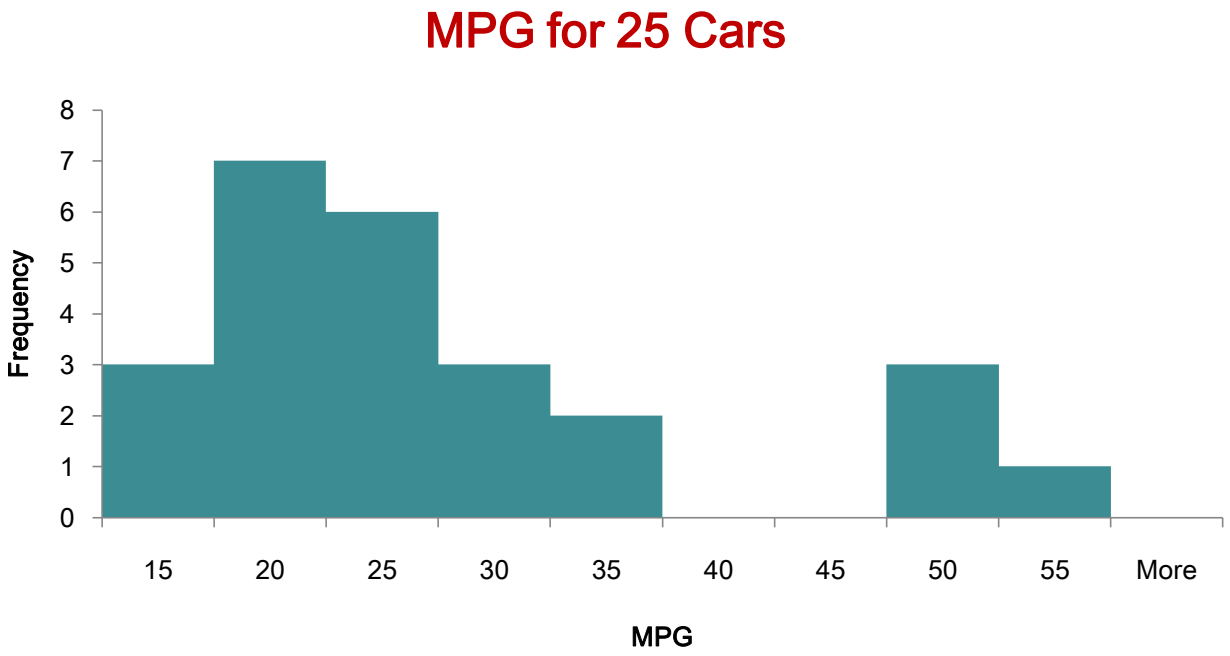
How to calculate the interval width for a histogram:

- Because we rounded the interval widths to 5 MPG, we'll actually need nine intervals, instead of seven, to capture all the data
  - This is not uncommon since we often have to round the original number calculated for the interval widths



DESCRIBING DATA – HISTOGRAMS

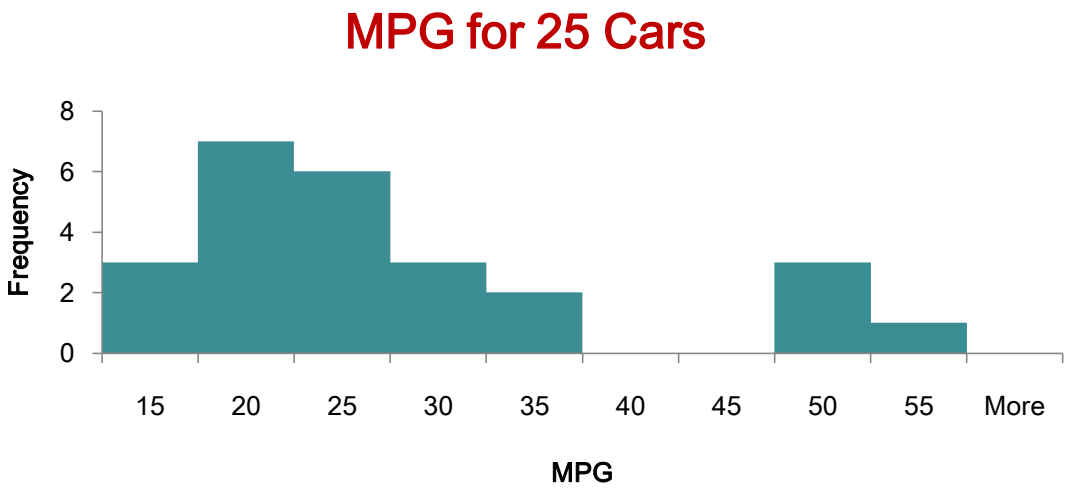
How to calculate the interval width for a histogram:



- We started the first interval at 15 mpg since the lowest mpg measured was 12. There were no cars with mpg below 10, so no intervals were needed
- By default Excel puts “More” at the end of each histogram even if there are no other data past the last interval

DESCRIBING DATA – HISTOGRAMS

How to calculate the interval width for a histogram:



- With the exception of the first and last intervals, all intervals must be the same width
  - The first and last intervals may be different depending on where the researcher begins the scale and the size and number of outliers that exist

Watch a tutorial on how to create a histogram in Excel

DESCRIBING DATA – HISTOGRAMS

What does the shape of a histogram tell us?

1. The shape of a histogram describes how the scores are distributed from low to high

2. Where the scores are clustered or massed
   - Taller Bars in the histogram indicate more data points are clustered around that point

3. Whether the shape of the histogram is normal, skewed, or some other shape

DESCRIBING DATA – HISTOGRAMS

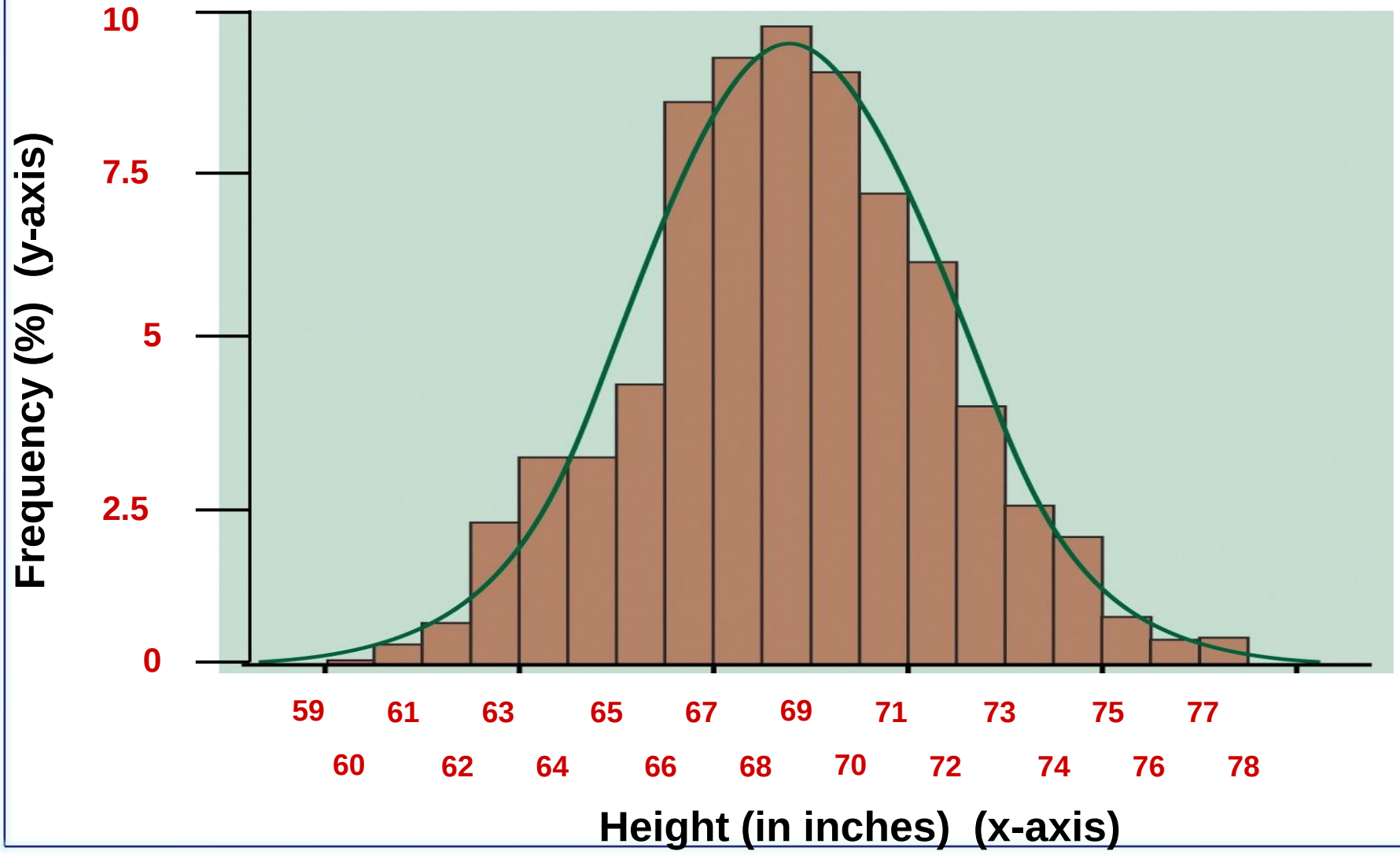
Normal or *bell-shaped* histogram

1. A normal or bell-shaped histogram is where scores are evenly distributed above and below a central point

2. Variables that naturally occur are typically normal in shape
   - EXAMPLES: height, amount of time taken to complete an exam, average temperature of winter for last 100 years
3. A special normal histogram is the symmetrical histogram (it can be folded onto itself perfectly)

DESCRIBING DATA – HISTOGRAMS

Normal or *bell-shaped* histogram



DESCRIBING DATA – HISTOGRAMS

Skewed histograms

1. A skewed histogram is where scores are more heavily clustered on the lower or higher end of the scale

2. Human created variables often have a skewed shape
   - EXAMPLES: Academic test scores, customer satisfaction survey scores, annual income, concert ticket prices, home prices

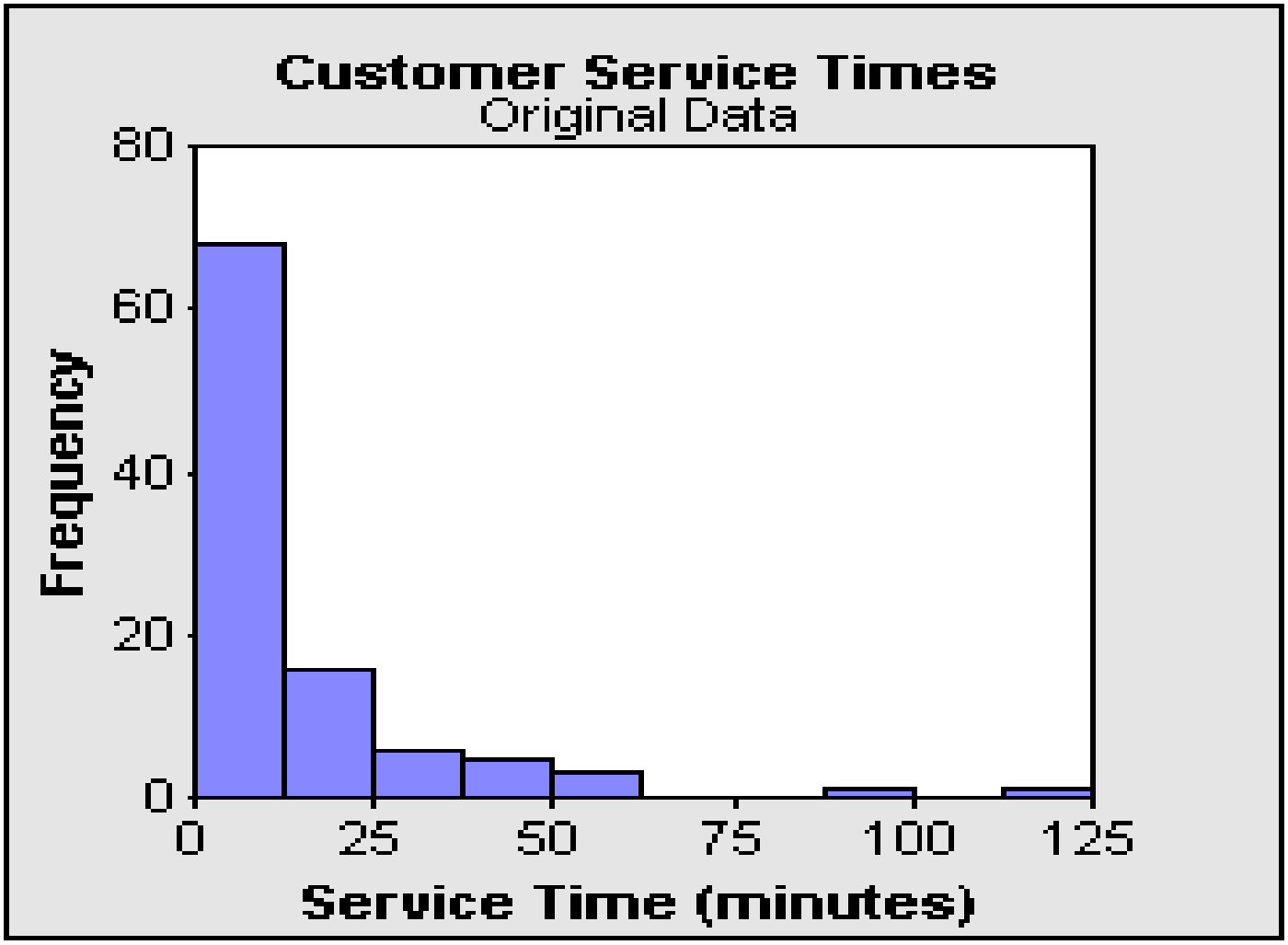
DESCRIBING DATA – HISTOGRAMS

Types of skewness

1. Positive skew histograms
   - A positive skew is where the scores are clustered on the low (or left) end of the scale and the tail points to the right
   - EXAMPLES: Customer satisfaction survey scores, annual income, home prices

DESCRIBING DATA – HISTOGRAMS

Positive skew histogram (tail points to the right)

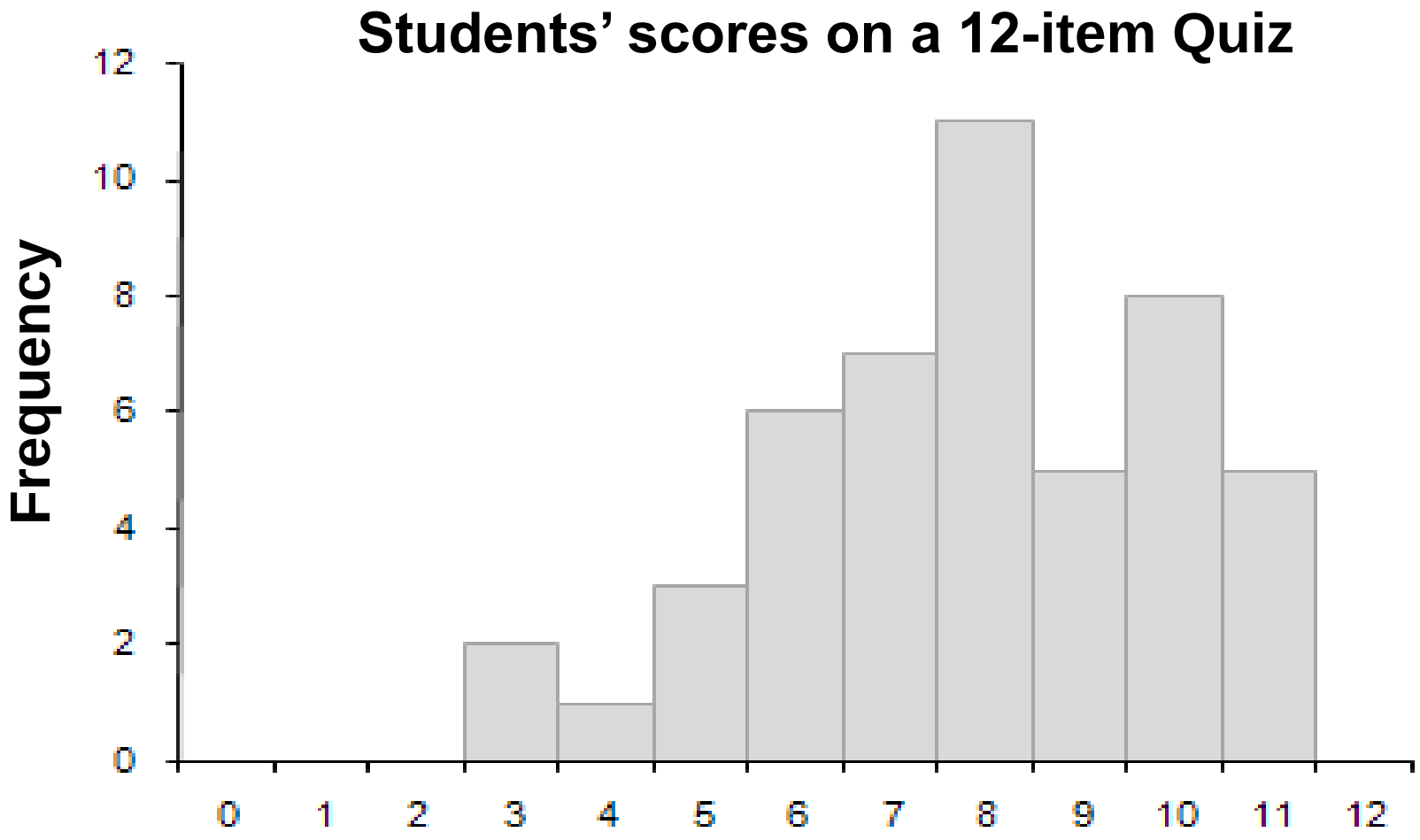


Types of skewness

1. Negative skew histograms
   - A negative skew is where the scores are clustered on the high (or right) end of the scale and the tail points to the left
   - EXAMPLES: Academic test scores, Number of prescriptions written since 1940, Annual federal deficit, Average size of hard disks in new computers since 1985

DESCRIBING DATA – HISTOGRAMS

Negative skew histogram (tail points to the left)



DESCRIBING DATA – STEM PLOTS

Stem plots are used to describe:

1. Quantitative data

2. Datasets that are small (e.g., N < 25)

3. Each observation measured (every score is represented in a stem plot)

DESCRIBING DATA – STEM PLOTS

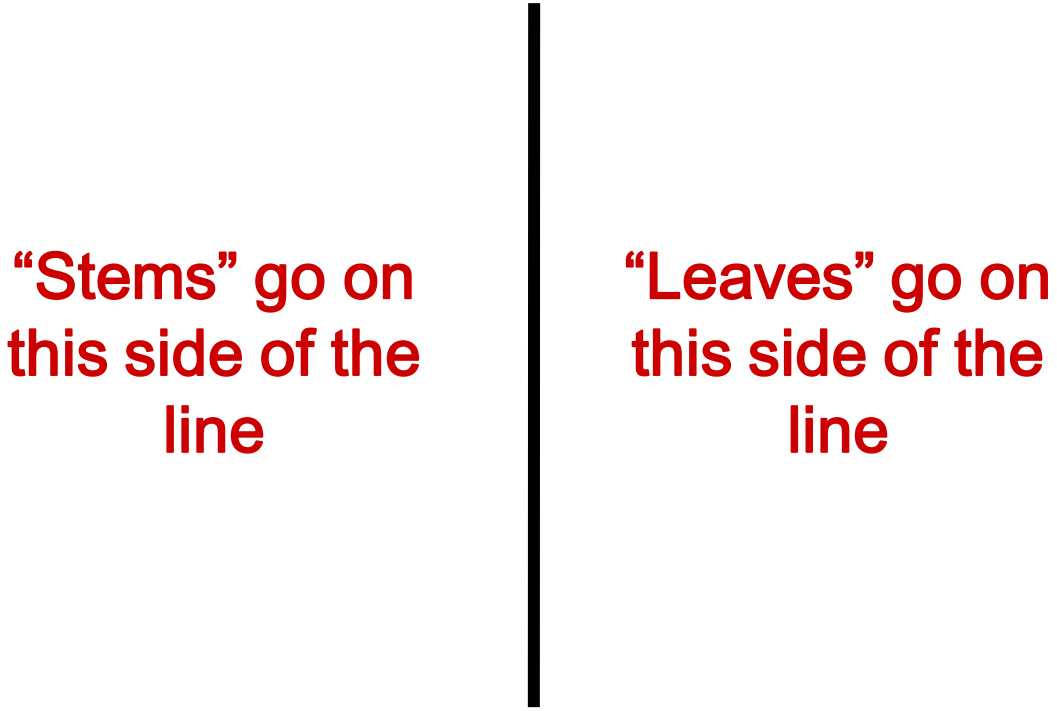
Stem plots are used to describe:

- A stem plot consists of a stems and leaves
  - Stems are intervals like in a histogram
    - Like intervals in a histogram, the width of each stem must be equal
    - Are typically rounded to a meaningful unit (e.g., whole numbers in increments of one, five, or ten)
  - Leaves are each observation or data point measured
    - Each value is rounded to some chosen value
      - The first digit of the rounded number is listed as a leaf in the stem plot

DESCRIBING DATA – STEM PLOTS

Stem plots:

- A stem plot consists of a stems and leaves

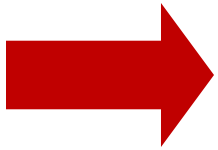


DESCRIBING DATA – STEM PLOTS

Stem plots:

- **EXAMPLE:** The load strength (i.e., lbs per square inch) of twenty pieces of wood are recorded. The numbers listed below indicate the pressures when the pieces of wood failed and broke.

Load	
23040	32030
24050	32320
26520	32340
28730	32590
30170	32700
30460	32720
30930	33020
31300	33190
31860	33280
31920	33650



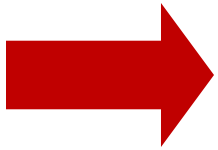
23	0
24	1
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0237

DESCRIBING DATA – STEM PLOTS

Stem plots:

- The stems are in increments of 1000 lbs starting at 23000
- The leaves are rounded to the nearest hundred pounds

Load	
23040	32030
24050	32320
26520	32340
28730	32590
30170	32700
30460	32720
30930	33020
31300	33190
31860	33280
31920	33650



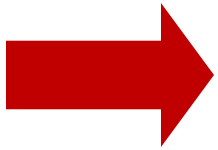
23	0
24	1
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0237

DESCRIBING DATA – STEM PLOTS

Stem plots:

- **EXAMPLE:** The value 26520 is rounded to 26500 and is highlighted below:

Load	
23040	32030
24050	32320
26520	32340
28730	32590
30170	32700
30460	32720
30930	33020
31300	33190
31860	33280
31920	33650



23	0
24	1
25	
26	5
27	
28	7
29	
30	259
31	399
32	033677
33	0237

Chapter 2

DESCRIBING DATA – TIME PLOTS

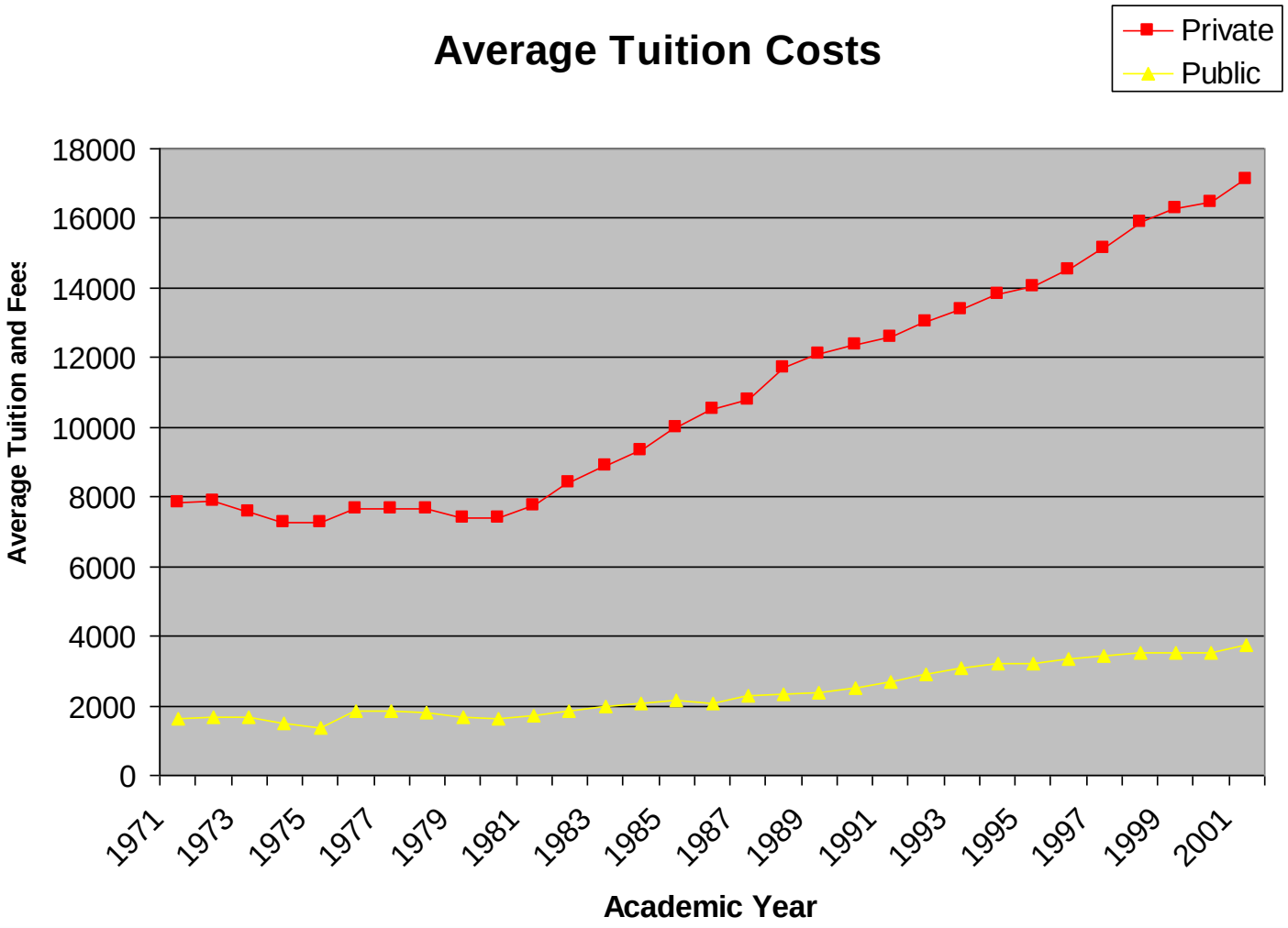
Time plots :

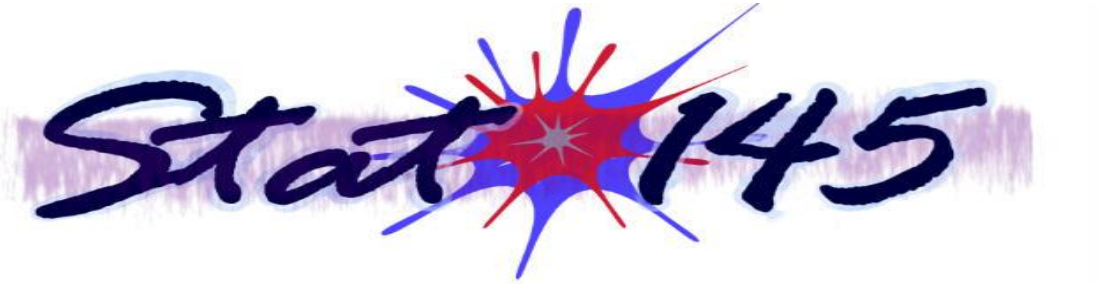
1. Time plots are used to describe changes in quantitative data over a specified period of time
   - EXAMPLES: Sales of a product, enrollment at UNM, amount of rainfall
2. The dependent variable is represented on the y-axis and time is plotted along the x-axis
3. The independent variable is represented by the plotted lines in the graph
4. Each plotted point represents an observation measured at a particular point in time
   - Connecting the points creates a timeline

DESCRIBING DATA – TIME PLOTS

Time plots :

❶ **EXAMPLE:** Changes in tuition costs from 1971 to 2001





End of Chapter 2 – Part 1