STA 2023

Module 2A

Organizing Data and Comparing Distributions (Part I)

Learning Objectives

Upon completing this module, you should be able to:

1. Classify variables and data as either qualitative or quantitative.
2. Distinguish between discrete and continuous variables and data.
3. Identify terms associated with the grouping of data.
4. Group data into a frequency distribution and a relative-frequency distribution.
5. construct a group-data table.
6. Draw a frequency histogram and a relative-frequency histogram.
7. Construct a dotplot.

Learning Objectives (Cont.)

8. Construct a stem-and-leaf diagram.
9. Draw pie chart and a bar graph.
10. Identify the shape and modality of the distribution of a data set.
11. Specify whether a a unimodal distribution is symmetric, right skewed, or left skewed.
12. Describe the relationship between sample distributions and the population distribution.
13. Identify and correct misleading graphs.

What Types of Variables do we have?

Variable: A characteristic that varies from one person or thing to another.

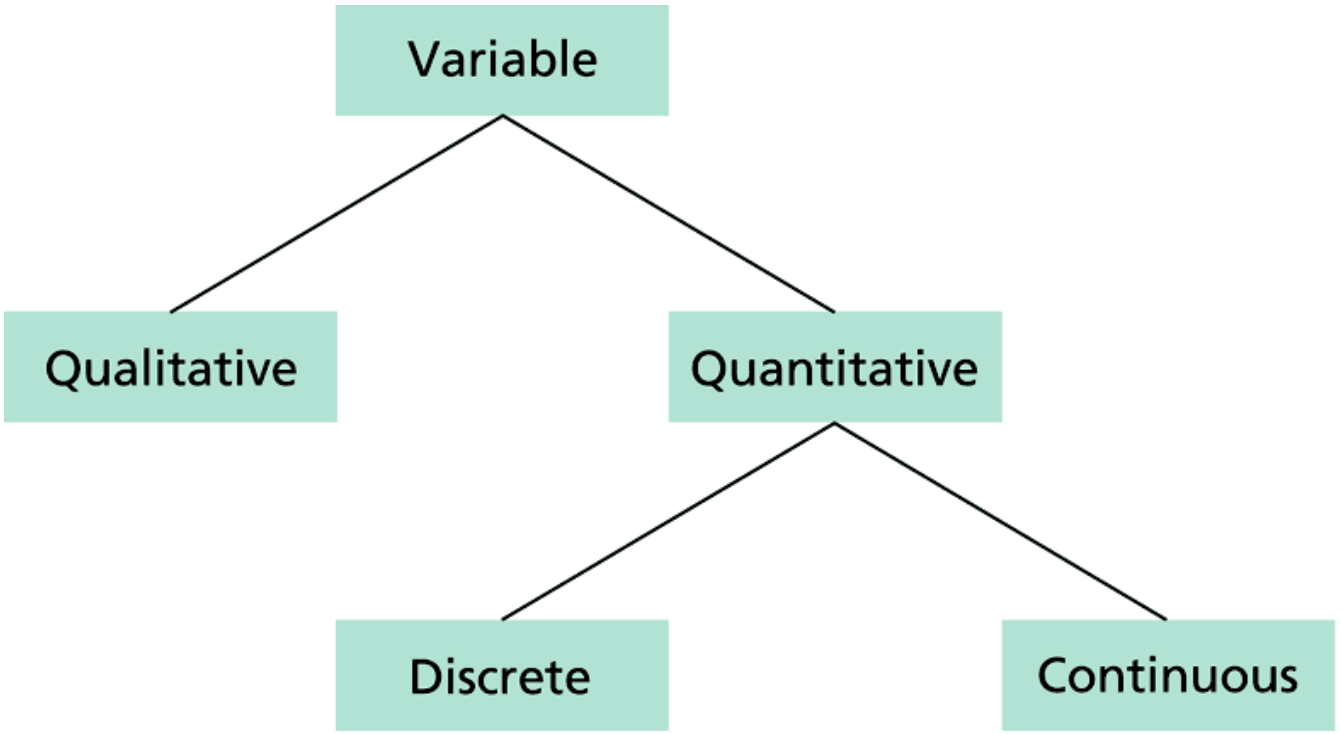
Qualitative variable: A nonnumerically valued variable.

Quantitative variable: A numerically valued variable.

Discrete variable: A quantitative variable whose possible values can be listed.

Continuous variable: A quantitative variable whose possible values form some interval of numbers.

Types of Variables



What are the Different Types of Data?

Data: Values of a variable.

Qualitative data: Values of a qualitative variable.

Quantitative data: Values of a quantitative variable.

Discrete data: Values of a discrete variable.

Continuous data: Values of a continuous variable.

Terms Used in Grouping

Classes: Categories for grouping data.

Frequency: The number of observations that fall in a class.

Frequency distribution: A listing of all classes and their frequencies.

Relative frequency: The ratio of the frequency of a class to the total number of observations.

Relative-frequency distribution: A listing of all classes and their relative frequencies.

Lower cutpoint: The smallest value that could go in a class.

Upper cutpoint: The smallest value that could go in the next higher class (equivalent to the lower cutpoint of the next higher class).

Midpoint: The middle of a class, found by averaging its cutpoints.

Width: The difference between the cutpoints of a class.

Why Grouping Data?

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

It makes the data much easier to read and understand.

Days to maturity	Tally	Number of investments
$30 \leq 40$	|||	3
$40 \leq 50$	|	1
$50 \leq 60$	𝍸 |||	8
$60 \leq 70$	𝍸 𝍸	10
$70 \leq 80$	𝍸 ||	7
$80 \leq 90$	𝍸 ||	7
$90 \leq 100$	||||	4
		40

How to construct a Group-Data Table?

Days to maturity	Frequency	Relative frequency	Midpoint
$30 \leq 40$	3	0.075	35
$40 \leq 50$	1	0.025	45
$50 \leq 60$	8	0.200	55
$60 \leq 70$	10	0.250	65
$70 \leq 80$	7	0.175	75
$80 \leq 90$	7	0.175	85
$90 \leq 100$	4	0.100	95
	40	1.000	

What is a Frequency Table?

- We can "pile" the data by counting the number of data values in each category of interest.
- We can organize these counts into a frequency table, which records the totals and the category names.

Class	Count
First	325
Second	285
Third	706
Crew	885

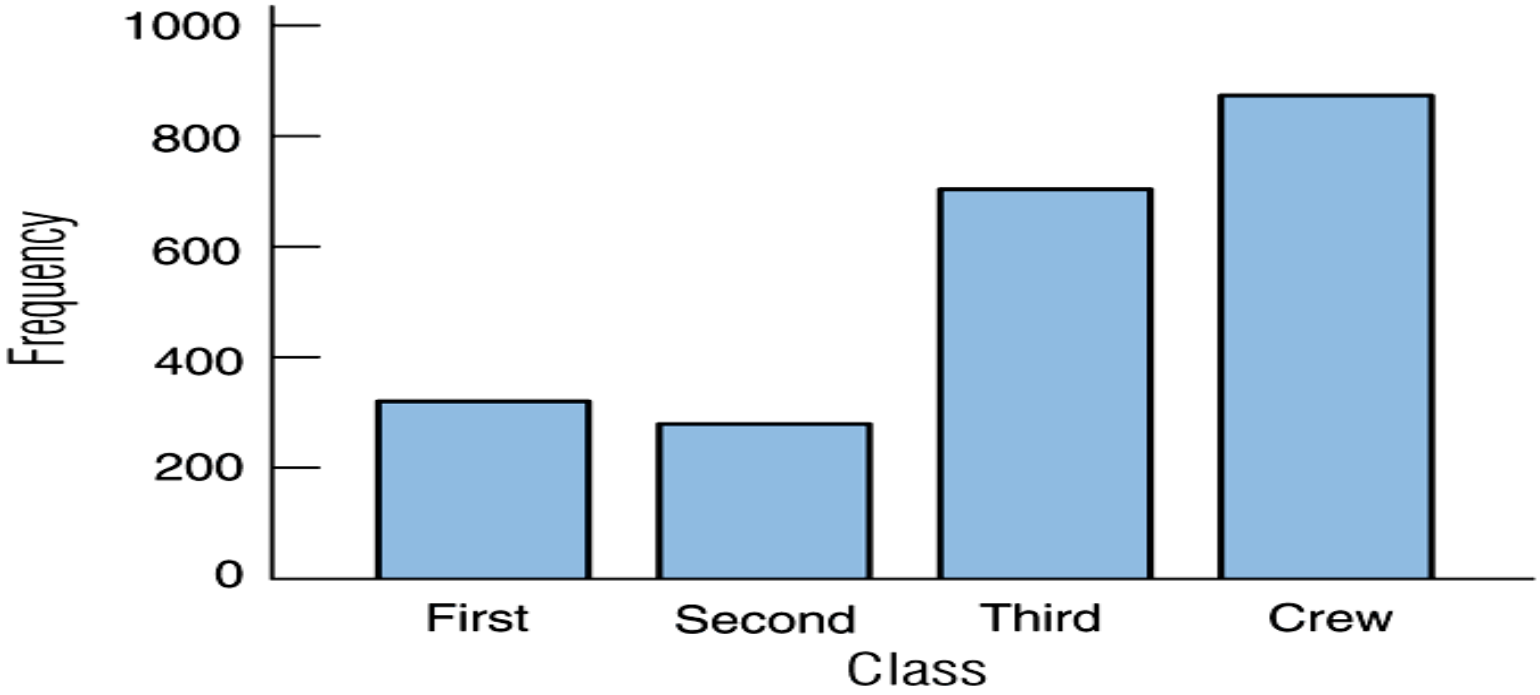
What is a Relative Frequency Table?

- A relative frequency table is similar, but gives the percentages (instead of counts) for each category.

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

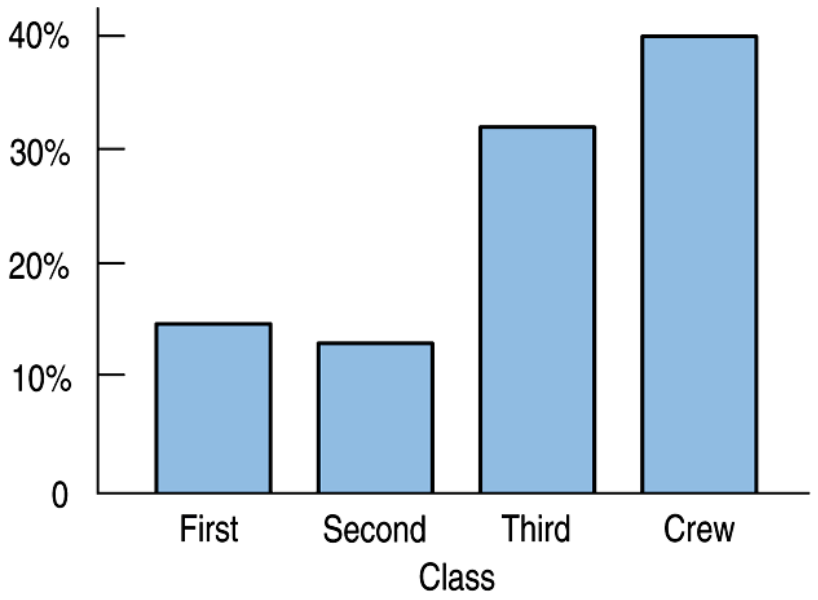
What is a Bar Chart?

A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.



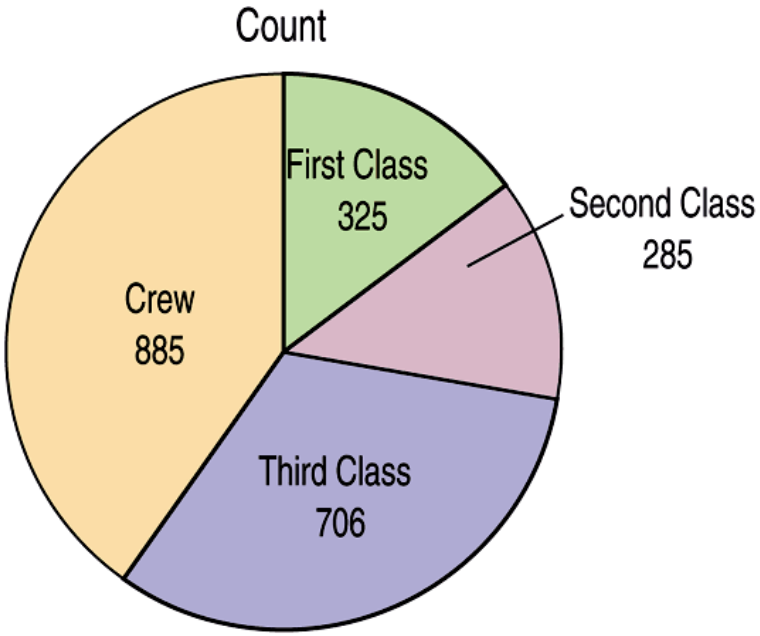
What is a Relative Frequency Bar Chart?

- A relative frequency bar chart displays the relative *proportion* of counts for each category.
- Replacing counts with percentages.



What is a Pie Chart?

- When you are interested in parts of the whole, a pie chart might be your display of choice.
- Pie charts show the whole group of cases as a circle.
- They slice the circle into pieces whose size is proportional to the fraction of the whole in each category.



What is a Contingency Table?

- A contingency table allows us to look at two categorical variables together.
- It shows how individuals are distributed along each variable, contingent on the value of the other variable.
  - Example: we can examine the class of ticket and whether a person survived the *Titanic*:

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

What is a Marginal Distribution?

- The margins of the table, both on the right and on the bottom, give totals and the frequency distributions for each of the variables.
- Each frequency distribution is called a marginal distribution of its respective variable.

Contingency Tables (cont.)

- Each cell of the table gives the count for a combination of values of the two values.
  - For example, the second cell in the crew column tells us that 673 crew members died when the *Titanic* sunk.

Survival	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

What is a Conditional Distribution?

- A conditional distribution shows the distribution of one variable for just the individuals who satisfy some condition on another variable.
  - The following is the conditional distribution of ticket *Class*, conditional on having survived:

	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%

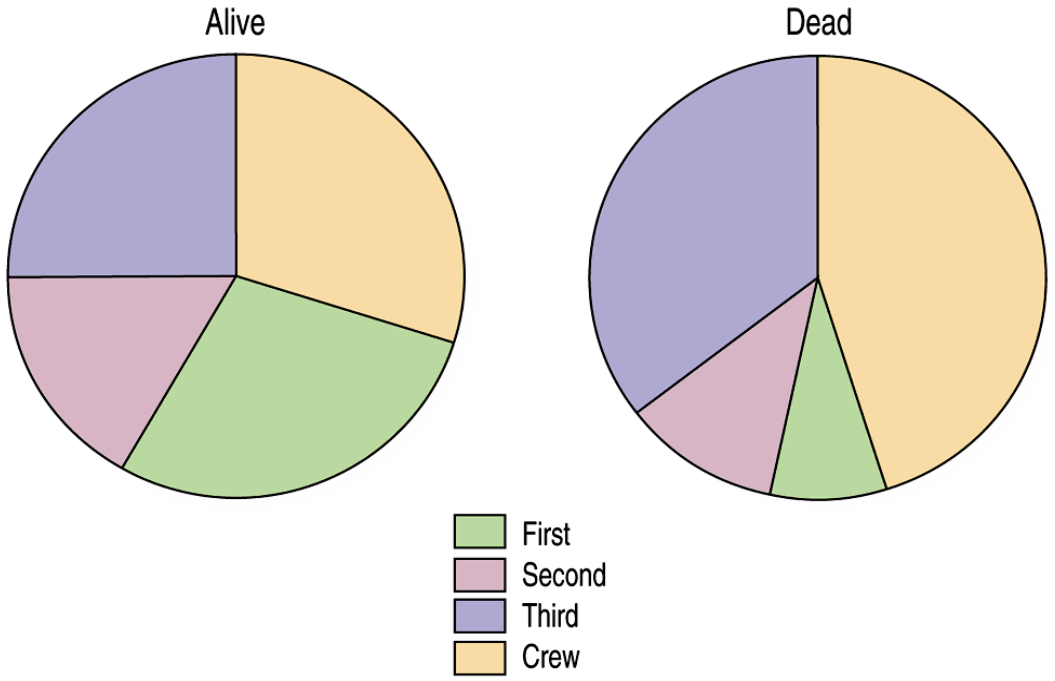
Conditional Distributions (cont.)

– The following is the conditional distribution of ticket *Class*, conditional on having perished:

	Class				
	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%

Conditional Distributions (cont.)

- The conditional distributions tell us that there is a difference in class for those who survived and those who perished.
- This is better shown with pie charts of the two distributions:

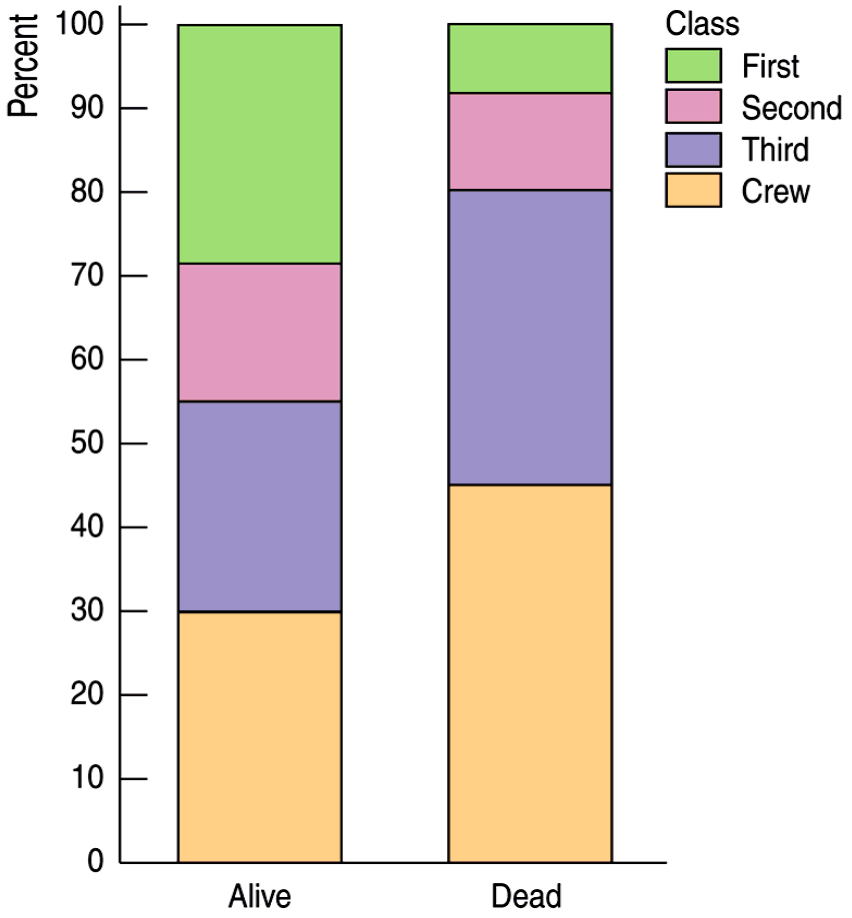


Conditional Distributions (cont.)

- We see that the distribution of *Class* for the survivors is different from that of the nonsurvivors.
- This leads us to believe that *Class* and *Survival* are associated, that they are not independent.
- The variables would be considered independent when the distribution of one variable in a contingency table is the same for all categories of the other variable.

What is a Segmented Bar Chart?

- A segmented bar chart displays the same information as a pie chart, but in the form of bars instead of circles.
- Here is the segmented bar chart for ticket *Class* by *Survival* status:



Dealing With a Lot of Numbers…

- Summarizing the data will help us when we look at large sets of quantitative data.
- Without summaries of the data, it's hard to grasp what the data tell us.
- The best thing to do is to make a picture…
- We can't use bar charts or pie charts for quantitative data, since those displays are for categorical variables.

Two types of Histograms

Frequency histogram: A graph that displays the classes on the horizontal axis and the frequencies of the classes on the vertical axis. The frequency of each class is represented by a vertical bar whose height is equal to the frequency of the class.

Relative-frequency histogram: A graph that displays the classes on the horizontal axis and the relative frequencies of the classes on the vertical axis. The relative frequency of each class is represented by a vertical bar whose height is equal to the relative frequency of the class.

How to make a Histogram?

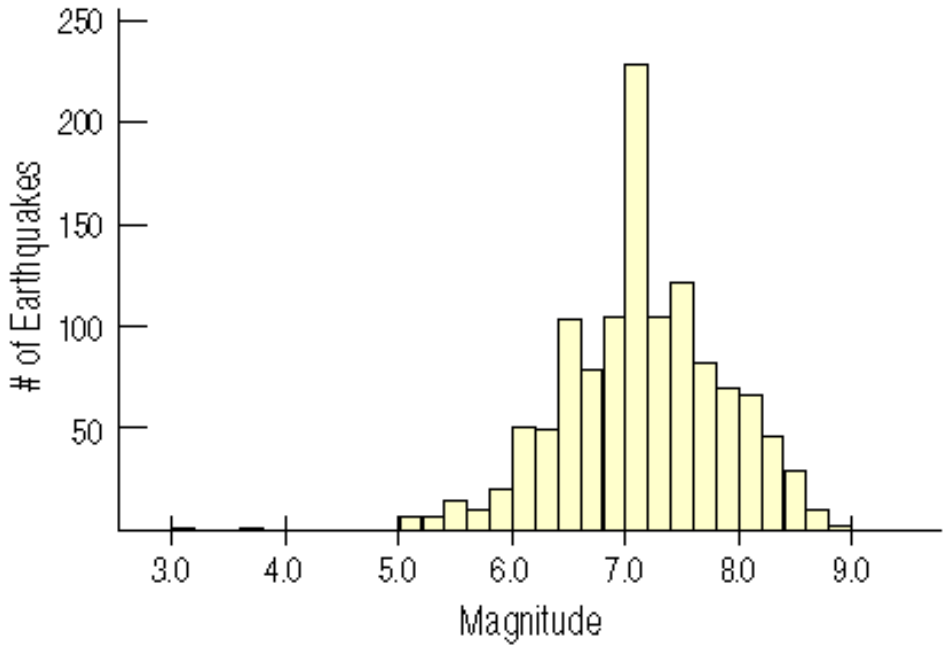
First, slice up the entire span of values covered by the quantitative variable into equal-width piles called bins.

The bins and the counts in each bin give the distribution of the quantitative variable.

Example of a Histogram

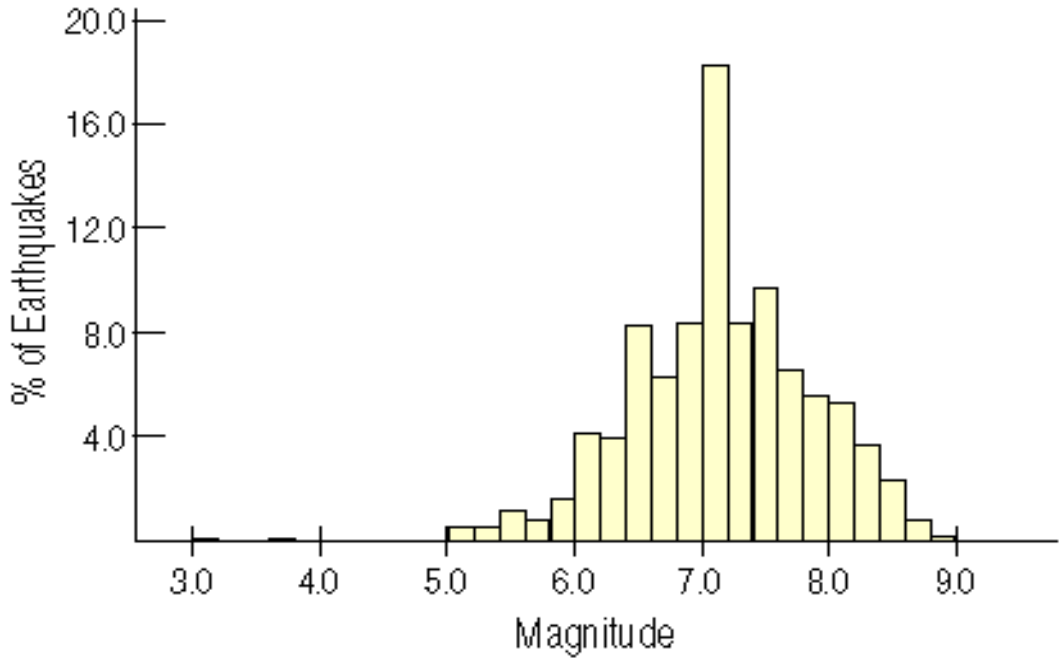
- A histogram plots the bin counts as the heights of bars (like a bar chart).

- Here is a histogram of earthquake magnitudes



Relative Frequency Histogram

A relative frequency histogram displays the *percentage* of cases in each bin instead of the count.

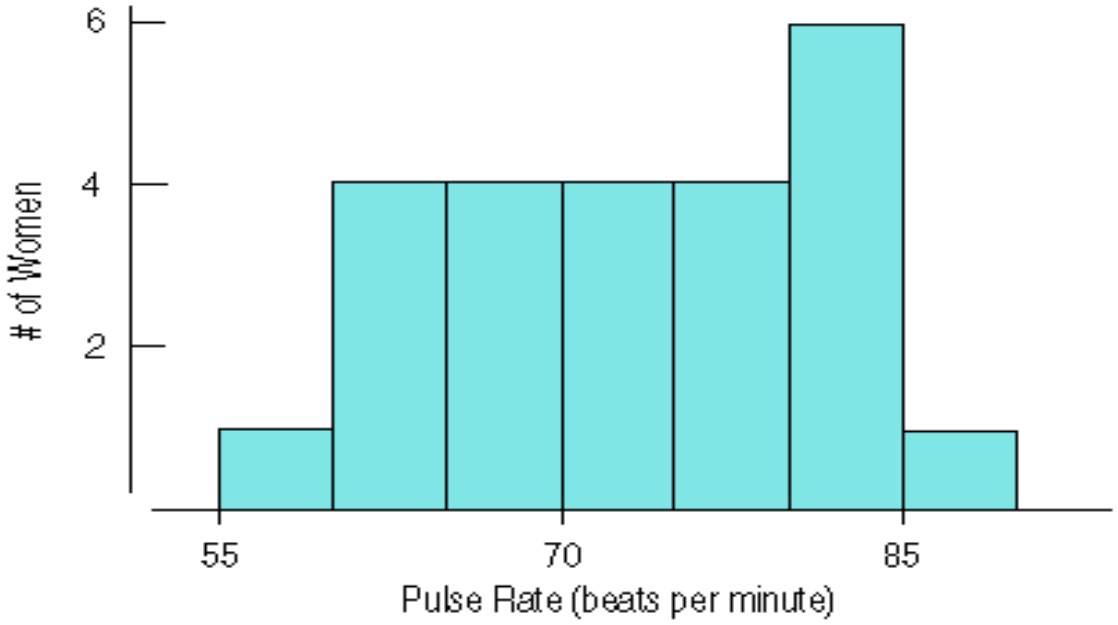


Stem-and-Leaf Displays

- 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, when carefully drawn, satisfy the area principle and show the distribution.

Stem-and-Leaf Example

Compare the histogram and stem-and-leaf display for the pulse rates of 24 women at a health clinic. Which graphical display do *you* prefer?



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

How to Construct a Stem-and-Leaf Display?

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

Another Example of Stem-and-Leaf Display

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

(a)

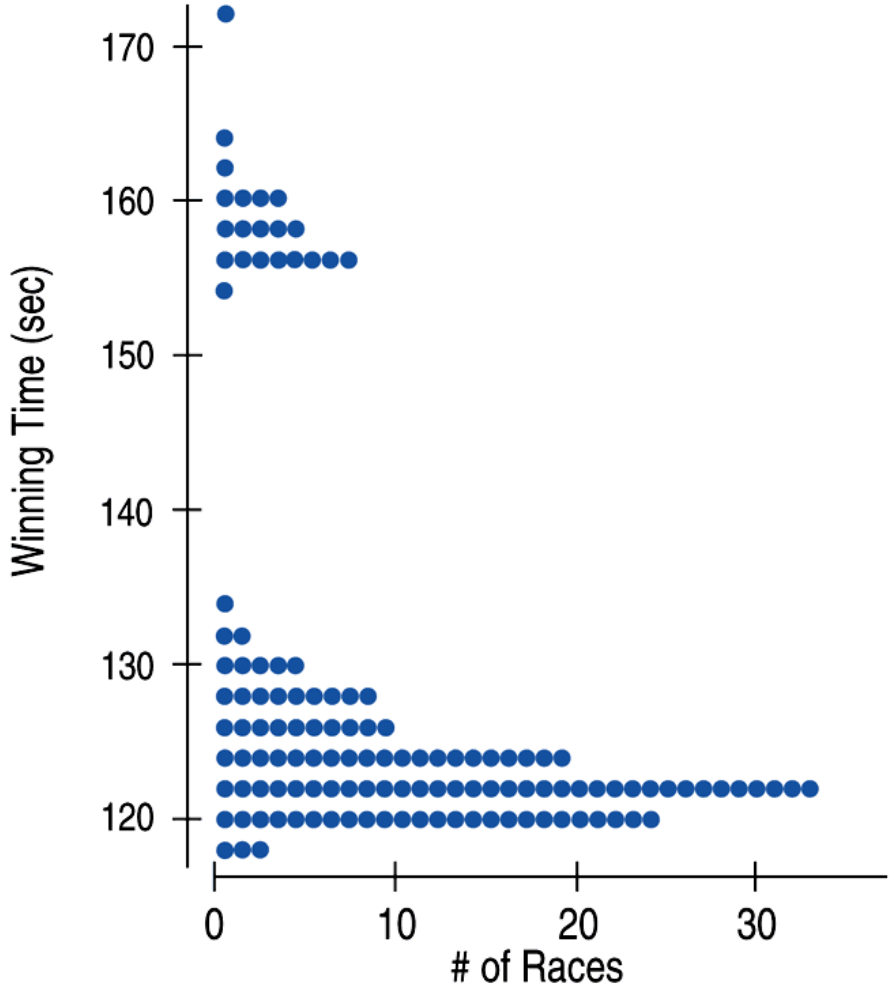
Stems	Leaves
3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 8 9
7	0 0 0 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

(b)

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

What is a Dotplot?

- A dotplot is a simple display. It just places a dot along an axis for each case in the data.
- 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.



Think Before You Draw

- Remember the "Make a picture" rule?
- Now that we have options for data displays, you need to *Think* carefully about which type of display to make.
- Before making a stem-and-leaf display, a histogram, or a dotplot, check the
  - Quantitative Data Condition: The data are values of a quantitative variable whose units are known.

Distribution of a Data Set

The **distribution of a data set** is a table, graph, or formula that provides the values of the observations and how often they occur.

A distribution is unimodal if it has one peak; bimodal if it has two peaks; and multimodal if it has three of more peaks.

A distribution that can be divided into two pieces that are mirror images of one another is called symmetric.

The Sample Distributions

For a simple random sample, the sample distribution approximates the population distribution (i.e., the distribution of the variable under consideration). The larger the sample size, the better the approximation tends to be.

Distribution of a Variable

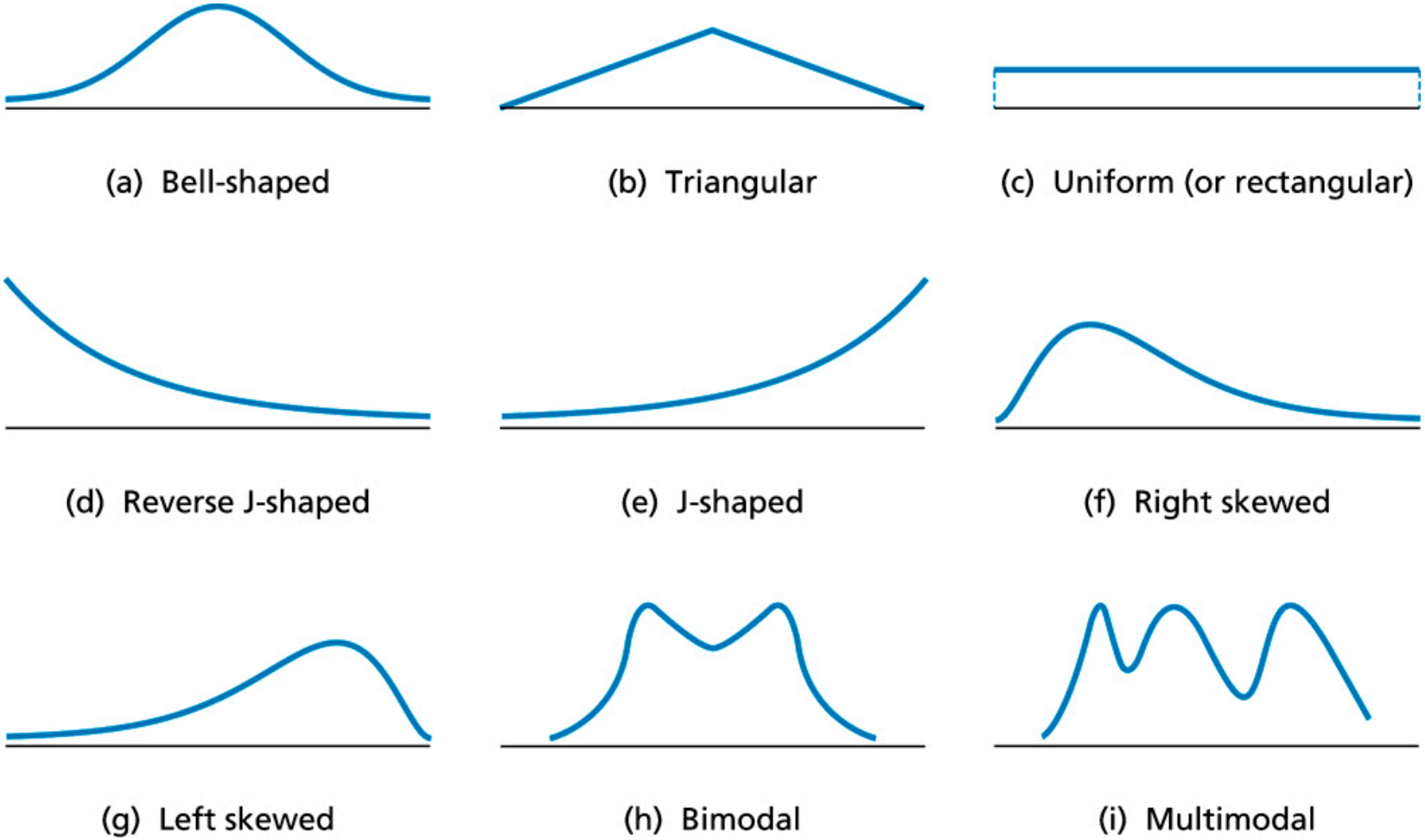
The distribution of population data is called the **population distribution,** or the **distribution of the variable.**

The distribution of sample data is called a **sample distribution.**

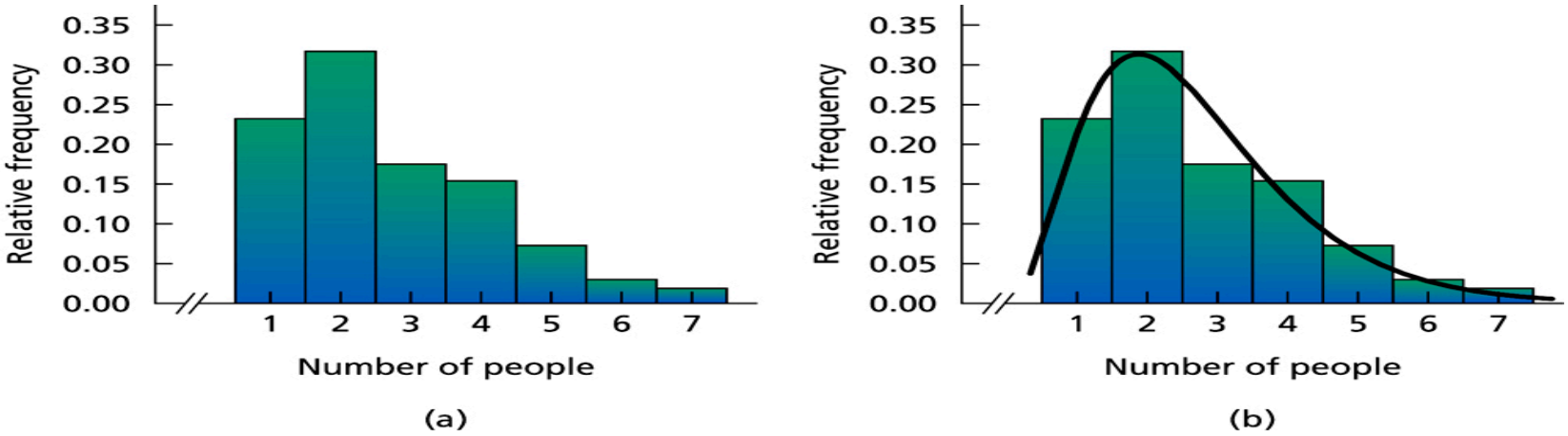
A bell-shaped, triangular, and uniform distributions are symmetric.

A unimodal distribution that is not symmetric is etiher right skewed (its “right tail” is longer) or left skewed (its “left tail” is longer.)

The Common Distribution Shapes



Relative Frequency Histogram and Distribution Shape

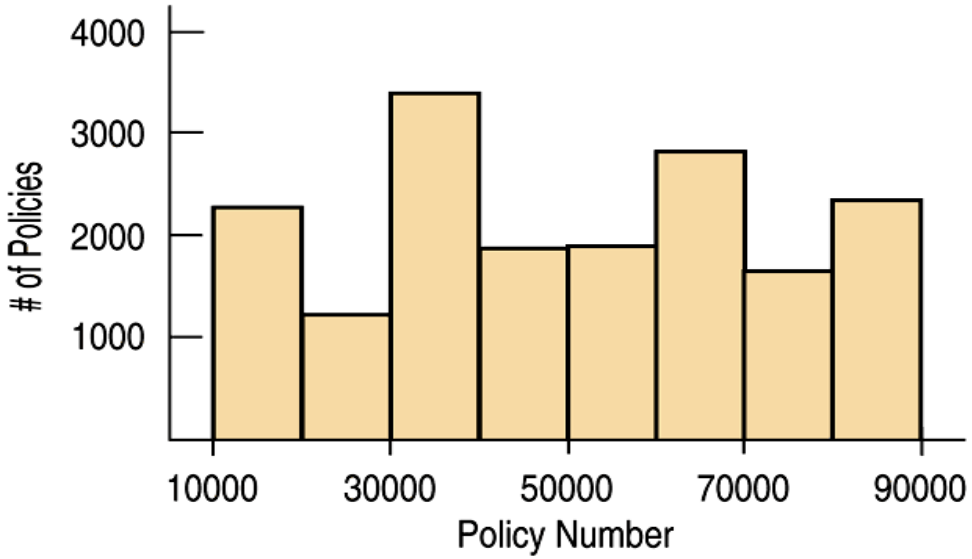


To identify the distribution shape in (a), we can draw a smooth curve through the histogram as in (b). Then, by referring back to the common distribution shapes, we can see that the distribution is right skewed.

What Can Go Wrong?

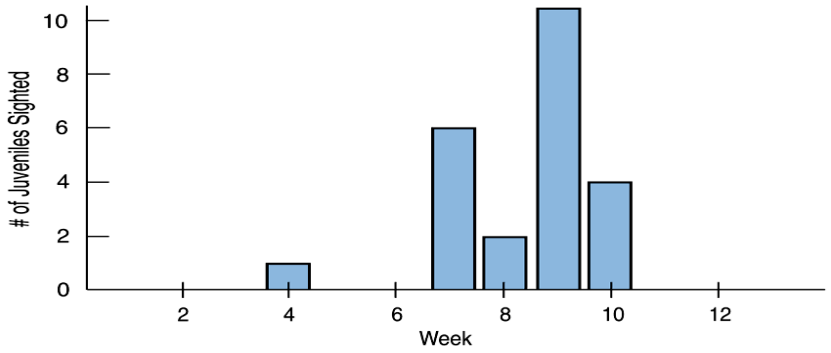
- Don't make a histogram of a categorical variable—bar charts or pie charts should be used for categorical data.

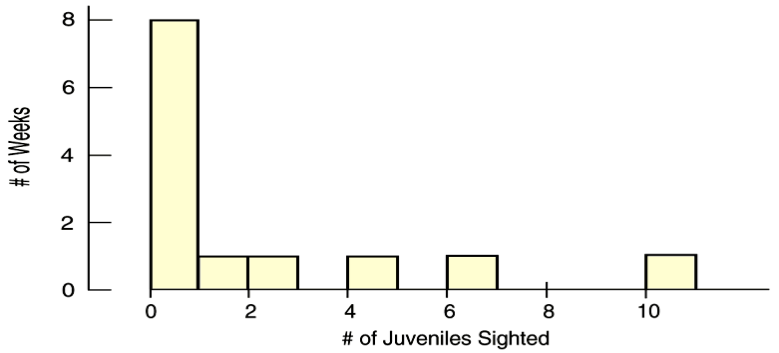
- Don't look for shape, center, and spread of a bar chart.



What Can Go Wrong? (Cont.)

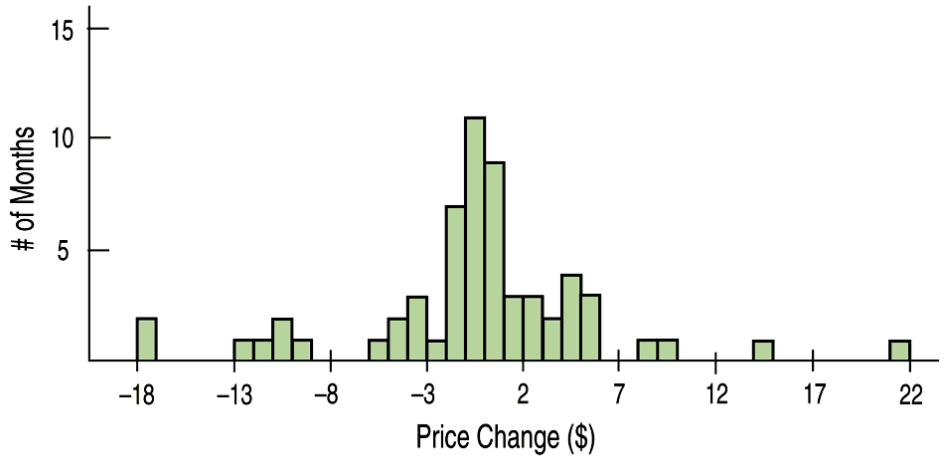
- Don't use bars in every display—save them for histograms and bar charts.
- Below is a badly drawn plot and the proper histogram for the number of eagles sighted in a collection of weeks:

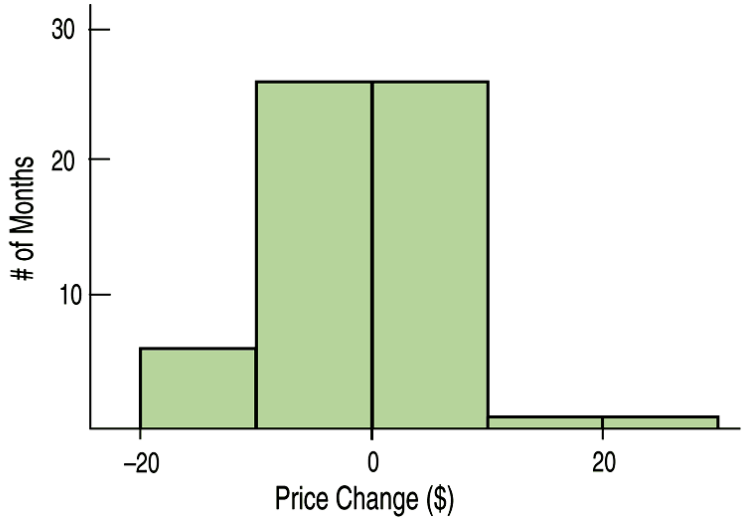




What Can Go Wrong? (Cont.)

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





What Can Go Wrong? (Cont.)

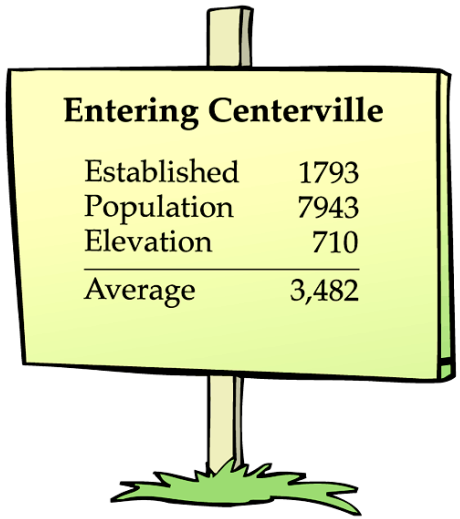
- Don't confuse similar-sounding percentages—pay particular attention to the wording of the context.
- Don't forget to look at the variables separately too—examine the marginal distributions, since it is important to know how many cases are in each category.

What Can Go Wrong? (Cont.)

- Be sure to use enough individuals!
  - Do not make a report like "We found that 66.67% of the rats improved their performance with training. The other rat died."

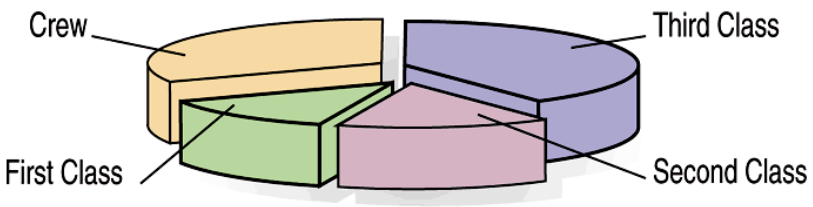
What Can Go Wrong? (Cont.)

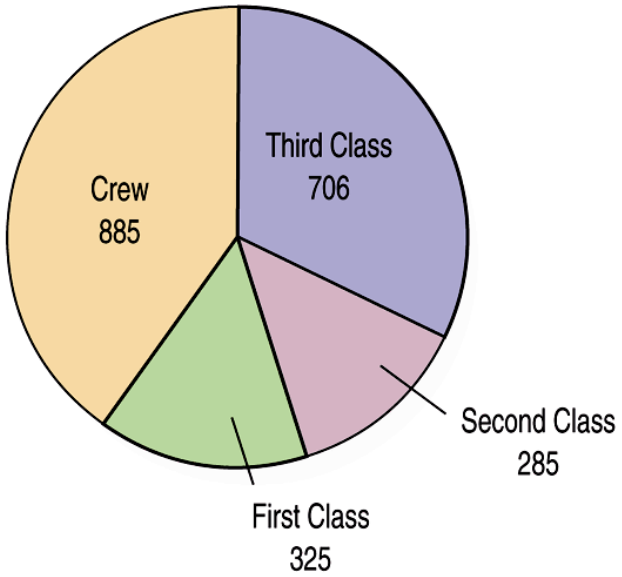
- Don't overstate your case—don't claim something you can't.
- Don't use unfair or silly averages—this could lead to Simpson's Paradox, so be careful when you average one variable across different levels of a second variable.



What Can Go Wrong? (Cont.)

- Don't violate the area principle.

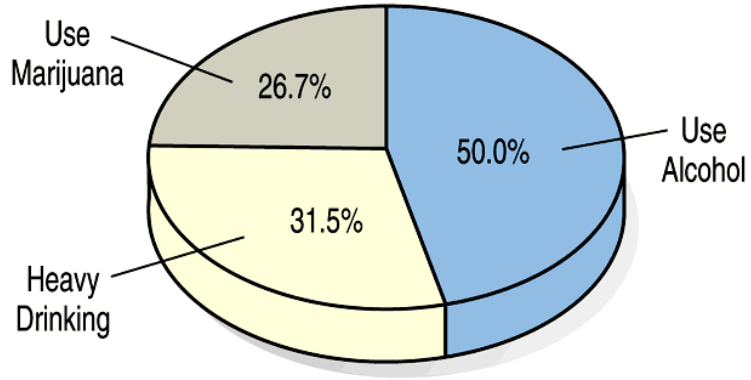




  - While some people might like the pie chart on the left better, it is harder to compare fractions of the whole, which a well-done pie chart does.

What Can Go Wrong? (Cont.)

- Keep it honest—make sure your display shows what it says it shows.



  - This plot of the percentage of high-school students who engage in specified dangerous behaviors has a problem. Can you see it?

What Can Go Wrong?

- Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.

What have we learned?

We have learned to:

1. Classify variables and data as either qualitative or quantitative.
2. Distinguish between discrete and continuous variables and data.
3. Identify terms associated with the grouping of data.
4. Group data into a frequency distribution and a relative-frequency distribution.
5. construct a group-data table.
6. Draw a frequency histogram and a relative-frequency histogram.
7. Construct a dotplot

What have we learned? (cont.)

8. Construct a stem-and-leaf diagram.
9. Draw pie chart and a bar graph.
10. Identify the shape and modality of the distribution of a data set.
11. Specify whether a a unimodal distribution is symmetric, right skewed, or left skewed.
12. Describe the relationship between sample distributions and the population distribution.
13. Identify and correct misleading graphs.

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition