Chapter 3:
Displaying and Describing Categorical Data

Obj - SWBAT recognize when a variable is categorical and choose an appropriate display for it, as well as understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

The Three Rules of Data Analysis

- The three rules of data analysis won't be difficult to remember:
  - > Make a picture—things may be revealed that are not obvious in the raw data. These will be things to think about.
  - > Make a picture—important features of and patterns in the data will show up. You may also see things that you did not expect.
  - > Make a picture—the best way to tell others about your data is with a well-chosen picture.

Frequency Tables: Making Piles

- 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

Frequency Tables: Making Piles (cont.)

Class	Count
First	325
Second	285
Third	706
Crew	885

Class	[redacted]
First	14.77
Second	12.95
Third	32.08
Crew	40.21

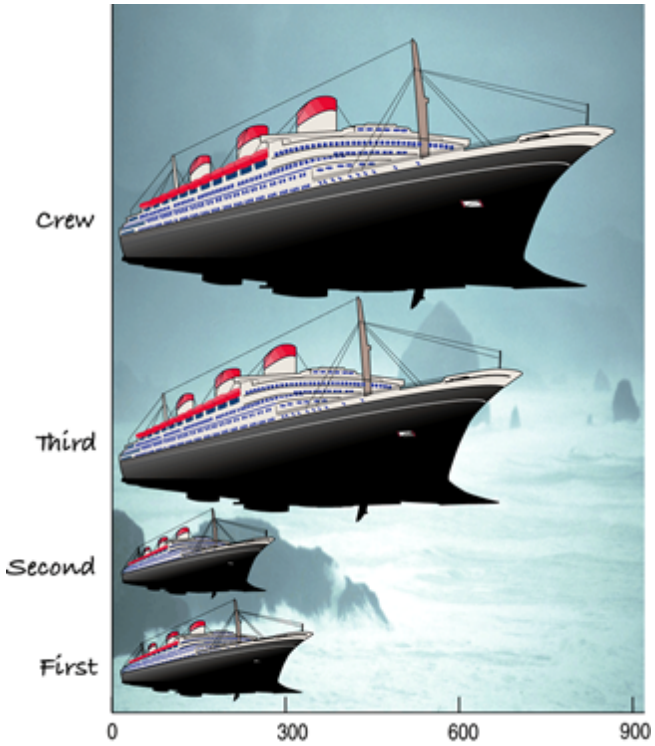
Distribution: the distribution of a variable gives:

- the possible values of the variable and
- the relative frequency of each value.

Frequency Tables: Making Piles (cont.)

- Both types of tables show how cases are distributed across the categories.
- They describe the distribution of a categorical variable because they name the possible categories and tell how frequently each occurs.

What's Wrong With This Picture?



The Area Principle

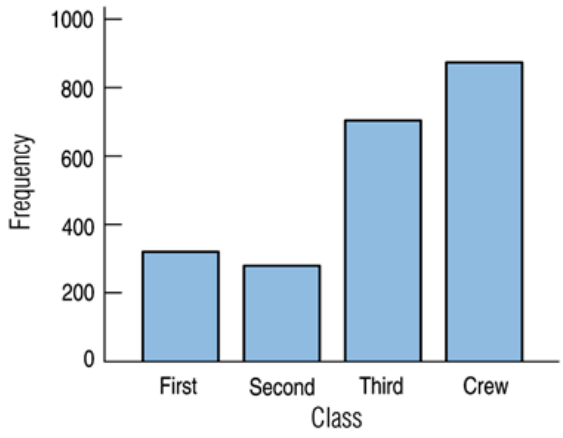
- The ship display makes it look like most of the people on the Titanic were crew members, with a few passengers along for the ride.
- When we look at each ship, we see the area taken up by the ship, instead of the length of the ship.
- The ship display violates the area principle:
- The area occupied by a part of the graph should correspond to the magnitude of the value it represents.

Area Principle: in a statistical display, each data value should be represented by the same amount of area.

Bar Charts

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

A bar chart stays true to the area principle. Thus, a better display for the ship data is:

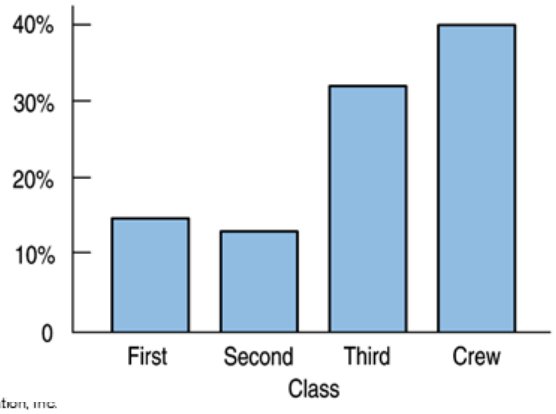


Bar Charts (cont.)

A relative frequency bar chart displays the relative proportion of cour each category.

A relative frequency bar chart also stays true to the area principle.

Replacing counts with percentages in the ship data:

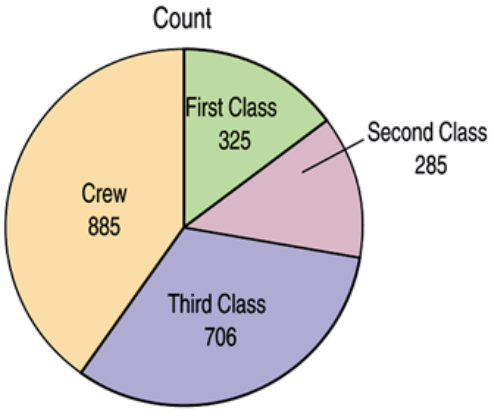


Pie Charts

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.



Before you construct a bar chart or a pie chart, you must check the...

Categorical Data Condition: the data are counts or percentages of individuals in categories.

This statistical display is not appropriate for quantitative data!

Contingency Tables

- 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

Contingency Tables (cont.)

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.

The marginal distribution of Survival is:

		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

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.

		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

Conditional Distributions

- 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%

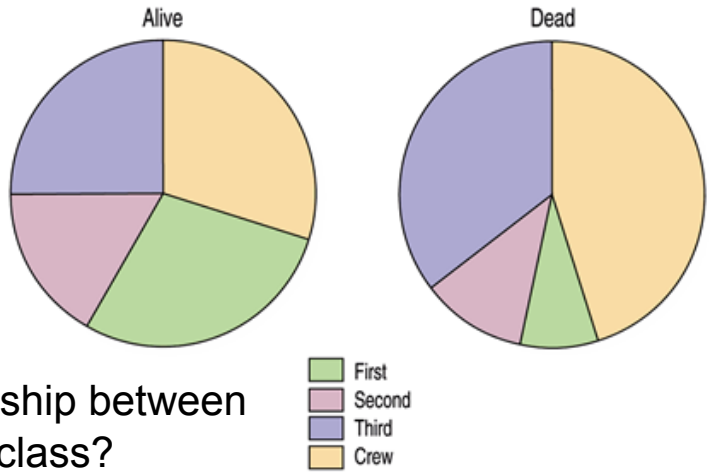
Conditional Distributions (cont.)

- 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:



- Is there a relationship between survival rate and class?

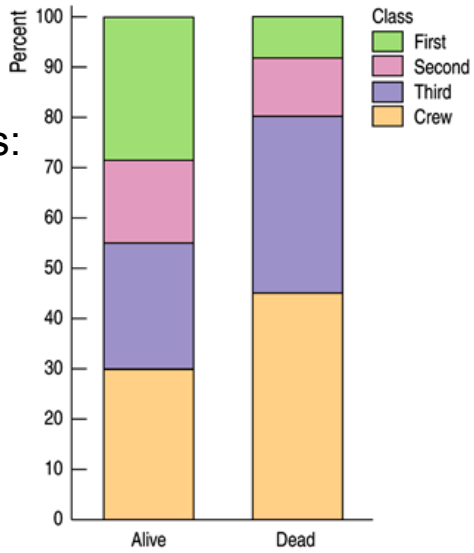
Independence: variables are said to be independent if the conditional distribution of one variable is the same for each category of the other.

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.

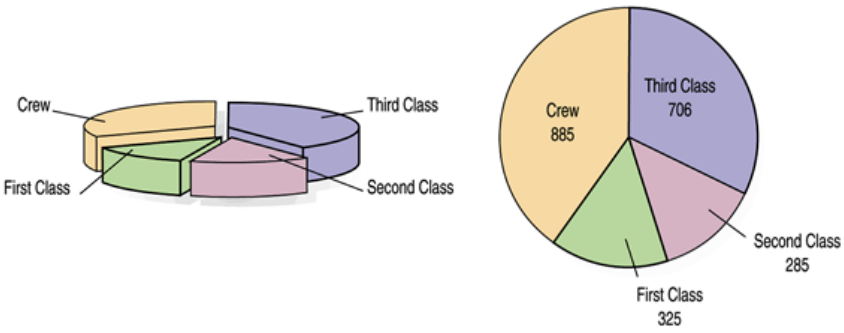
Segmented Bar Charts

- A segmented bar chart displays the same information as a pie chart, but in the form of bars instead of circles.
- Each bar is treated as the "whole" and is divided proportionally into segments corresponding to the percentage in each group.
- Here is the segmented bar chart for ticket Class by Survival status:



What Can Go Wrong?

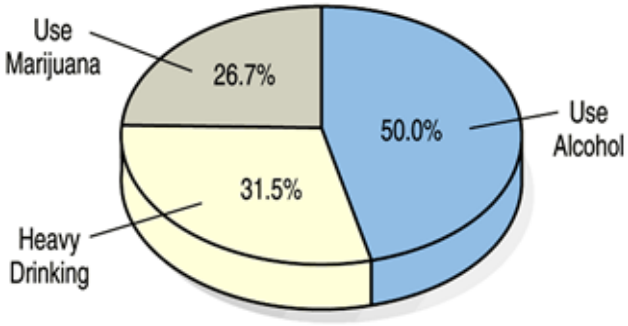
- Don't violate the area principle.



- 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? (cont.)

Don't confuse similar-sounding percentages—pay particular attention to the wording of the context.

		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**

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.

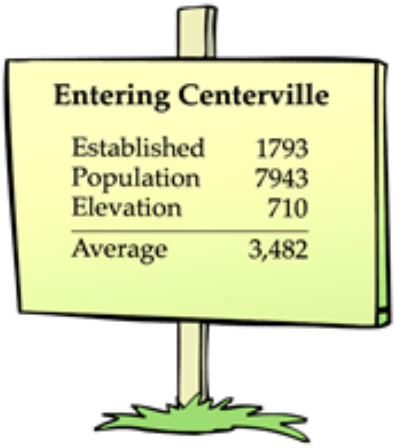
What Can Go Wrong? (cont.)

- Be sure to use enough individuals: especially when you consider percentages.
- Do not overstate your case. It is rare for two variables to be *completely* independent.

What Can Go Wrong? (cont.)

- 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.

Simpson's Paradox: when averages are taken across different groups, they can appear to contradict the overall averages.



What have we learned?

- We can summarize categorical data by counting the number of cases in each category (expressing these as counts or percents).
- We can display the distribution in a bar chart or pie chart.
- And, we can examine two-way tables called contingency tables, examining marginal and/or conditional distributions of the variables.

Ex. 1: Portable MP3 music players, such as the iPod, are popular - but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to a survey of 1112 randomly selected people.

Age group (years)	% owning MP3 player
12 to 17	54
18 to 24	30
25 to 34	30
35 to 54	13
55 and older	5

- Make a well-labeled bar graph to display the data. Describe what you see.

- Would it be appropriate to make a pie chart for these data? Why or why not?

A survey of 4826 randomly selected young adults (aged 19 to 25) asked, "what do you think are the chances you will have much more than a middle-class income at age 30?" The table below shows the responses, omitting a few people who refused to respond or who said they were already rich.

Opinion	Female	Male	Total
almost no chance	96	98	194
some chance but probably not	426	286	712
a 50-50 chance	696	720	1416
a good chance	663	758	1421
almost certain	486	597	1083
Total	2367	2459	4826

- Use the data in the two-way table to calculate the marginal distribution (in percents) of opinions. Create a graph to display the marginal distribution and describe what you see.

A sample of 200 children from the United Kingdom aged 9-17 was selected. The gender of each student was recorded along with which superpower they would most like to have in the two-way table below.

Superpower	Female	Male	Total
invisibility	17	13	30
superstrength	3	17	20
telepathy	39	5	44
fly	36	18	54
freeze time	20	32	52
Total	115	85	200

- Use the data in the table to calculate the marginal distribution in percents of superpower preferences. Make a graph to display the marginal distribution. Describe what you see.

- Calculate the conditional distribution of responses for the males and females.

Superpower	% of females	% of males
invisibility		
superstrength		
telepathy		
fly		
freeze time		
Total		