Bivariate Relationships

- ***Two quantitative variables***
  - Scatter plot
  - Side by side stem and leaf plots
- ***Two qualitative variables***
  - Tables
  - Bar charts
- ***One quantitative and one qualitative variable***
  - Side by side box plots
  - Bar chart

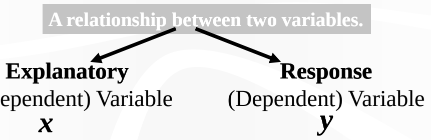
Scatterplots

- The association between two quantitative variables can be shown on one graph by plotting data points as ordered pairs on axes. Such a graph is called a ***scatterplot***.
- If it seems that one variable is a response to the other, then plot that variable on the *y*-axis. It is called the ***response variable***.
  - The *x*-axis then has the ***explanatory variable***.

Response and explanatory variables

- ***Response variable—*** the variable which we intend to model.
  - we intend to explain through statistical modeling
- ***Explanatory variable—*** the variable or variables which may be used to model the response variable
  - values may be related to the response variable

Two quantitative variables



Hours of Training	Number of Accidents
Shoe Size	Height
Cigarettes smoked per day	Lung Capacity
Score on SAT	Grade Point Average
Height	IQ

What type of relationship exists between the two variables and is the association significant?

Describing the Association

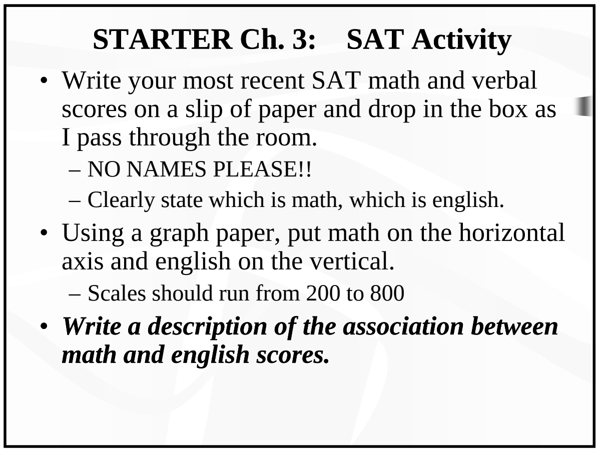
- Which variable should go on the *x*-axis?
  - Do cold days cause gas usage, or does gas usage cause cold days(!)?
  - *Since cold days cause gas usage, degree-days is the* ***explanatory variable*** *and goes on the x-axis.*
    - ➢*Gas usage responds to degree-days, so it is the* ***response variable*** *and goes on the y-axis.*

STARTER Ch. 3: SAT Activity

- Write your most recent SAT math and verbal scores on a slip of paper and drop in the box as I pass through the room.
  - NO NAMES PLEASE!!
  - Clearly state which is math, which is verbal.

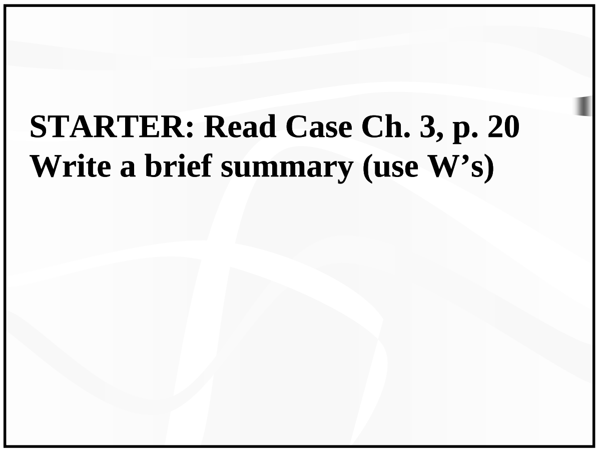
AGENDA

- **HW Chapter 3, 1-35 odds**
- **BRING Graph papers, colored pencils**

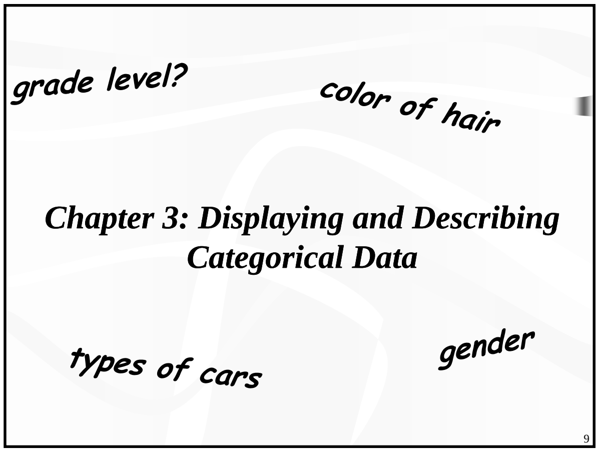


STARTER Ch. 3: SAT Activity

- Write your most recent SAT math and verbal scores on a slip of paper and drop in the box as I pass through the room.
  - – NO NAMES PLEASE!!
  - – Clearly state which is math, which is english.
- Using a graph paper, put math on the horizontal axis and english on the vertical.
  - – Scales should run from 200 to 800
- ***Write a description of the association between math and english scores.***



STARTER: Read Case Ch. 3, p. 20
Write a brief summary (use W's)



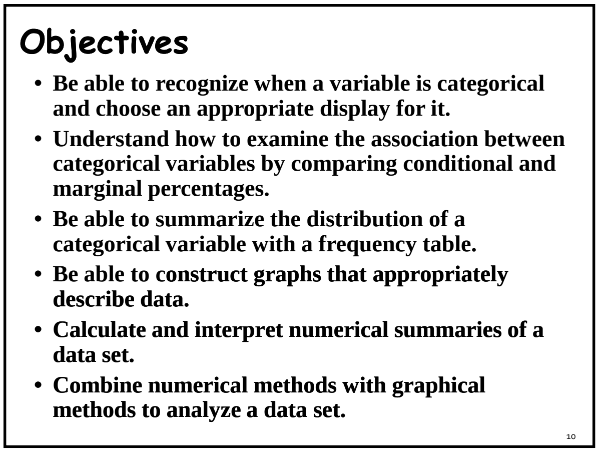
grade level?

color of hair

Chapter 3: Displaying and Describing Categorical Data

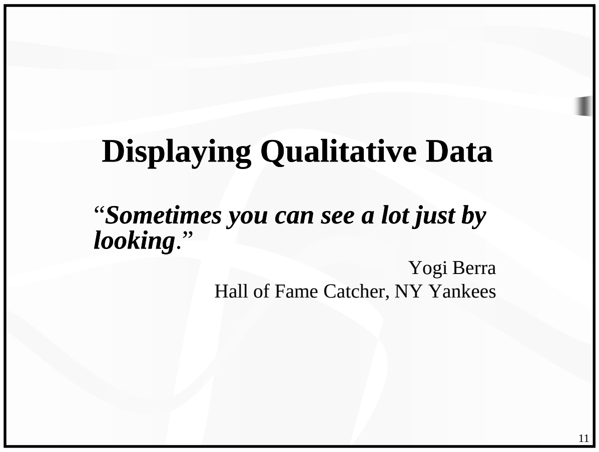
types of cars

gender



Objectives

- **Be able to recognize when a variable is categorical and choose an appropriate display for it.**
- **Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.**
- **Be able to summarize the distribution of a categorical variable with a frequency table.**
- **Be able to construct graphs that appropriately describe data.**
- **Calculate and interpret numerical summaries of a data set.**
- **Combine numerical methods with graphical methods to analyze a data set.**



Displaying Qualitative Data

"***Sometimes you can see a lot just by looking***."

Yogi Berra
Hall of Fame Catcher, NY Yankees

The Three Rules of Data Analysis

Make a Picture

Make a Picture

Make a Picture

The Three Rules of Data Analysis

The three rules of data analysis won't be difficult to remember:

1. **Make a picture** — things may be revealed that are not obvious in the raw data. These will be things to ***think*** about.
2. **Make a picture** — important features of and patterns in the data will ***show*** up. You may also see things that you did not expect.
3. **Make a picture** — the best way to ***tell*** others about your data is with a well-chosen picture.

Launched: 31st May 1911
Builders: Harland and Wolff, Belfast
Port of Registry: Liverpool

Passengers Lost: 818 (62%)
Crew Lost: 684 (77%)
Total Lost: 1,502 (68%)

Contingency Tables

A **contingency table** is used to organize multiple variables.

Ex: Contingency Table of Titanic passengers

		Class of Passenger				
		1st	2nd	3rd	Crew	**Total**
Survival	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	325	285	706	885	**2,201**

Ways to present categorical data

- You've seen data represented in newspapers, magazines, online. How do you normally see it?
  - ➢ Tables (frequency tables)
  - ➢ Bar charts
  - ➢ Pie charts
  - ➢ Line graphs
  - ➢ Contingency tables

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

Relative Frequency Tables

- **Percentages (proportions)** instead of **counts**.

Class	Count		%
First	325	325/2,201	14.77
Second	285	285/2,201	12.95
Third	706	706/2,201	32.08
Crew	885	885/2,201	40.21

TOTAL 2,201

325/2,201
285/2,201
706/2,201
885/2,201

Both describe the **distribution** of a categorical variable.

Distribution:
name of categories and how frequently each occurs

Class	Count
First	325
Second	285
Third	706
Crew	885

Frequency distribution

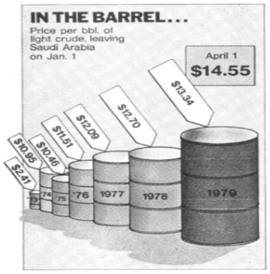
Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Relative frequency distribution

The “Area Principle”

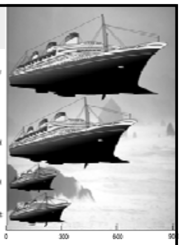
The **Area Principle** says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents.

1958 — EISENHOWER: $1.00
1963 — KENNEDY: 96¢
1968 — JOHNSON: 83¢
1973 — NIXON: 64¢
1978 — CARTER: 44¢
Purchasing Power of the Diminishing Dollar



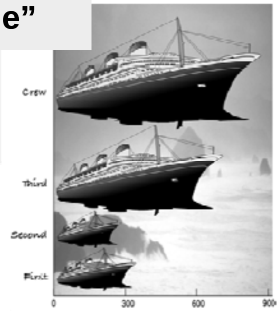
The “Area Principle”

- 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.



this is a *Violation* of the “Area Principle”

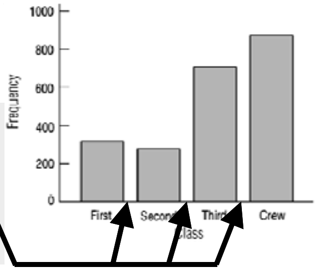
First	325
Second	285
Third	706
Crew	885



When we look at each ship, we see the **area** taken up by the ship, instead of the **length** of the ship.

Bar Charts

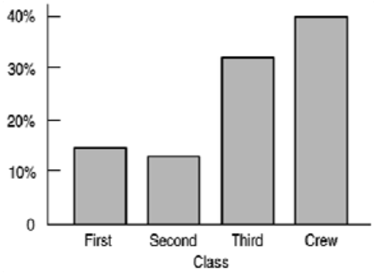
- **A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.**
- **A bar chart stays true to the *area* principle.**



For bar charts (with categorical data), be sure to leave **spaces between the bars!!!**

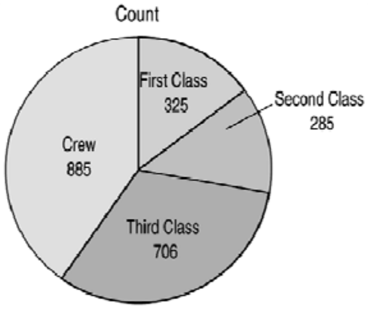
Bar Charts

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



Pie Charts

When you are interested in parts of the whole, a **pie chart** might be your display of choice.



Some questions…

PET ACTIVITY: Please put a single tally mark on the board to classify yourself by gender and type of pet you own.

What percentage of our class is male?

What percentage of our class has a dog only?

What percentage of our class does NOT have a cat or dog?

Some questions…

1) What percentage of our class is male?
2) What percentage of our class has a dog only?
3) What percentage of our class does NOT have a cat or dog?
4) What percentage of the males have a cat only?
5) What percentage of dog (only) owners are female?
6) What percentage of our class are female cat (only) owners?
7) If you have both a dog and a cat, what is the percent chance that you will be male?

More Questions

What percentage of the males have a cat only?

What percentage of dog (only) owners are female?

What percentage of our class are female cat (only) owners?

If you have both a dog and a cat, what is the percent chance that you will be male?

Marginal Distributions

A distribution of **one of the variables** in a contingency table is its **marginal distribution**.

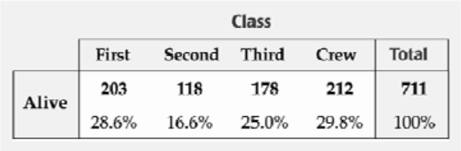
Example:

a) For our data, what is the marginal distribution of gender?

b) For our data, what is the marginal distribution of pets?

Conditional Distributions

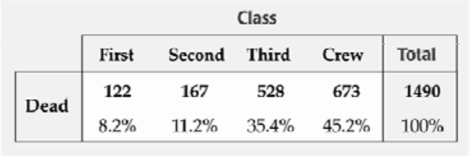
- A **conditional distribution** shows the distribution of one variable for just the individuals who satisfy some condition on another variable.
  - The following is the conditional distribution of ticket *Class*, conditional on having survived:



	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.)

- 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.

Conditional Distributions

At times, we may want to limit our “Who” and look at only a specific variable value for that “Who” only

A distribution of one variable for only those individuals **satisfying some condition** of the other variable is a **conditional distribution**.

Conditional Distributions

A distribution of one variable for only those individuals **satisfying some condition** of the other variable is a **conditional distribution**.

a) What is the conditional distribution of pets for males?

b) What is the conditional distribution of pets for females?

Conditional Distribution

How do these differ:

- Conditional Distribution of pet for each gender

- Conditional Distribution of gender for each pet

Independence

In a contingency table, when the distribution of one variable is the same for all categories of another, we say the variables are **independent**.

- ❖ Look at the conditional distributions of the table
  - ❖ If the distributions are similar, we can say the variables are **independent**.
  - ❖ If the distributions are different, we can say the variables are **dependent**.

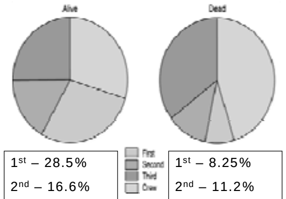
Segmented Bar Charts

An alternative to a Pie Chart, a Segmented Bar Chart divides up bars instead of circles.

Each bar is treated as a "whole" (100%) and is **divided proportionally** into segments corresponding to percentages in each group.

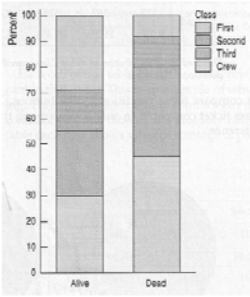
Segmented Bar Charts are great visual displays for seeing if distributions are alike or different in order to decide on independence.

Comparing the Graphs



1st – 28.5%
2nd – 16.6%
3rd – 25.1%
Crew – 29.9%

1st – 8.25%
2nd – 11.2%
3rd – 35.4%
Crew – 45.1%



back to the Titanic…

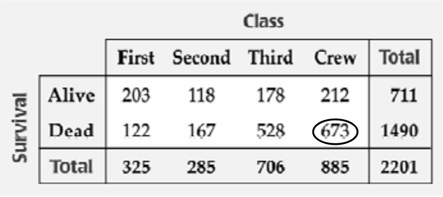
A **contingency table** allows us to look at two categorical variables together.

		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**

marginal distributions

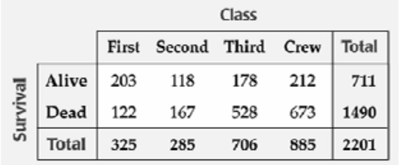
A **contingency table** allows us to look at two categorical variables together.

- 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

43



		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 percent of the people on the Titanic died?

1490/2201 = 67.7%

- What percent of the people were surviving crew?

212/2201 = 9.6%

- *What percent of the survivors were First class?

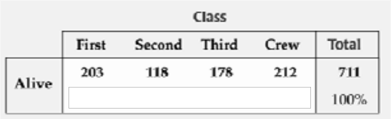
203/711 = 28.6%

- *What percent of First class survived?

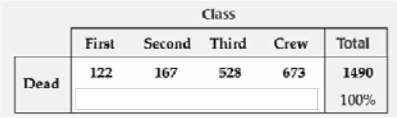
203/325 = 62.5%

44

- A **conditional distribution** shows the distribution of one variable for just the individuals who satisfy some condition on another variable.



	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
					100%



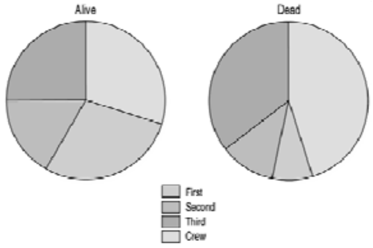
	Class				
	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
					100%

45

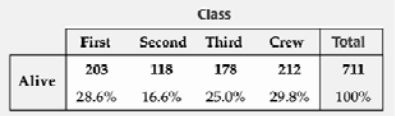
Conditional Distributions

- 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:
- Pie charts

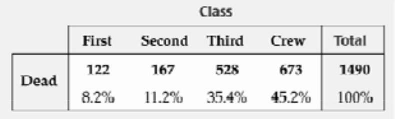
of the two distributions:



46



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



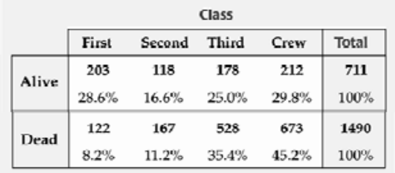
	Class				
	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%

We see that the **distribution of Class** for the **survivors** is **different** from that of the **non-survivors**…

so class and survival are associated

(they are dependent).

47



	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%

The variables would be considered **independent** if the distribution of one variable were the **same for all categories of the other variable.**

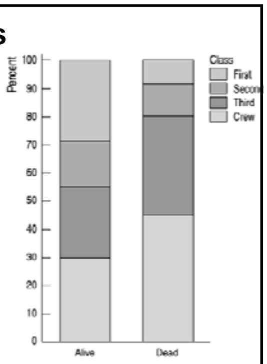
independent = no association

dependent = association

48

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 o the percentage in each group.
- Here is the segmented bar chart for ticket *Class* by *Survival* status:



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

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

The distributions for each gender are the same, so gender is **independent** of level of education.
(no association)

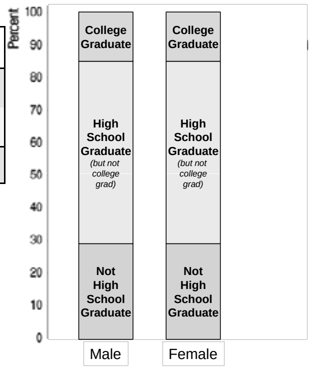
Gender	Not High School Graduate	High School Graduate*	College Graduate	Total
Male	318	603	165	**1086** *100%*
Female	212	402	110	**724** *100%*
Total	**530**	**1005**	**325**	**1800** *100%*

*and **not** a college graduate

Level of Education by Gender

	Not High School Graduate	High School Graduate*	College Graduate	Total
Male	318 *29.3%*	603 *55.5%*	165 *15.2%*	**1086** *100%*
Female	212 *29.3%*	402 *55.5%*	110 *15.2%*	**724** *100%*
Total	**530** *29.3%*	**1005** *55.5%*	**325** *15.2%*	**1800** *100%*

Gender is
independent
of level of education
(no association)



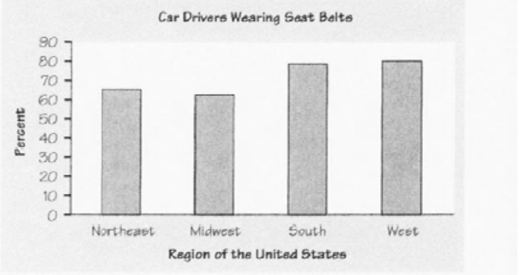


FIGURE 1.2 A bar graph showing the percentage of drivers who wear their seat belts in each of four U.S. regions.

In which region do the greatest number of people wear seatbelts?

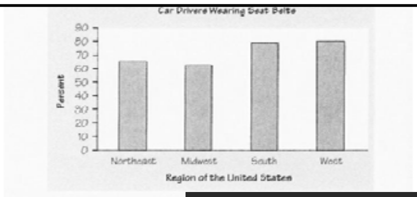


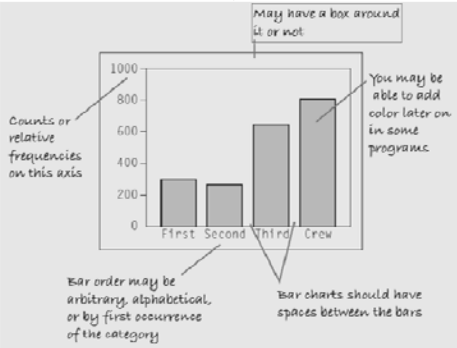
FIGURE 1.2 A bar graph showing the perce

- Overall, the bar chart s
of the country have mo
drivers wearing seat be

Note: we are using the word "proportion" (or "percentage")… …NOT the word "number"

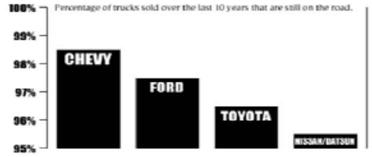
- The **Midwest** has the **smallest proportion** of car drivers wearing seat belts (about 62%) where the **South and West** have the **largest proportion** (about 78- 80%).

Displaying Categorical Data on the Computer



GRAPHS

Reproduced from the back cover of *Time* magazine, 140:15 (October 12, 1992)



More than 98% of all Chevy trucks sold in the last 10 years are still on the road.

Over the years, no other truck is that dependable, foreign or domestic.

"Best full-size pickup in initial quality."–J.D. Power and Associates

"Best truck line overall in sales satisfaction."–J.D. Power and Associates

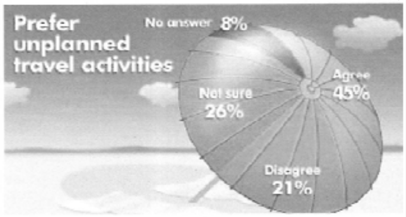
No deductible. 3-year/36,000-mile Bumper to Bumper Plus Warranty

Chevrolet. The Most Dependable, Longest-Lasting Trucks.

LIKE A ROCK

MORE GRAPHS…

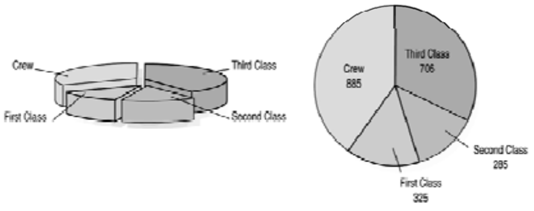
More than four in 10 adults say they would rather keep their schedule free when they travel, vs. organizing their days with planned activities.



By Anne R. Carey and Gia Kereselidze, USA TODAY
Source: Simmons Market Research

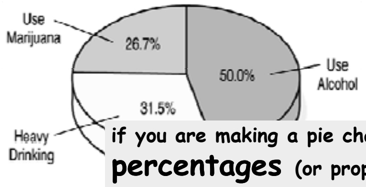
What Can Go Wrong?

- 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?

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



if you are making a pie chart with **percentages** (or proportions), **make sure the percentages add up to 100%!!!**

What Can Go Wrong?

- Be sure to use enough individuals!
  - Do not make a report like "We found that

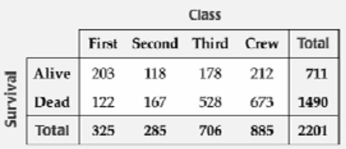
    66.67% of the rats improved their performance with training.

    The other rat died."

What Can Go Wrong?

These percentages sound similar but are different:

- The percentage of the passengers who were both in first class and survived. This would be 203/2201, or 9.4%
- The percentage of the first-class passengers who survived. This would be 203/325, or 62.5%.
- The percentage of the survivors who were in first class. This is 203/711, or 28.6%.

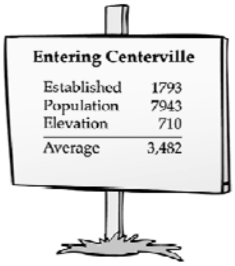


		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Pay attention to the **WHO** implicitly defined by the phrase. Often, there is a restriction to a smaller group.

What Can Go Wrong?

- Don’t use unfair or silly averages ~



When using averages of proportions across several different groups, it’s important to make sure that the groups really are comparable!

61

What Can Go Wrong?

- Don’t forget to look at the variables separately .

When you make a contingency able or display conditional distribution, be sure you also examine the marginal distributions. It is important to know **how many cases** are in each category.

62

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.**

63

we need **data** for next time!
(average hair length)

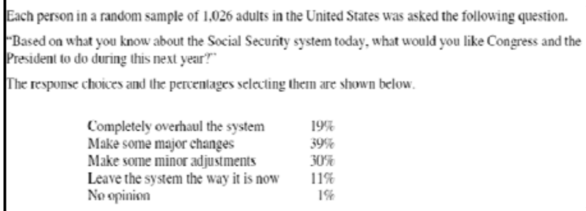
64

Assignment

Chapter 3	**Lesson**: Categorical Data	**Read**: Chapter 4	**Problems**: 1 – 35 (odds) p. 37-42

65

CW Ch. 3 **2002B #4**



Each person in a random sample of 1,026 adults in the United States was asked the following question.

“Based on what you know about the Social Security system today, what would you like Congress and the President to do during this next year?”

The response choices and the percentages selecting them are shown below.

Completely overhaul the system	19%
Make some major changes	39%
Make some minor adjustments	30%
Leave the system the way it is now	11%
No opinion	1%

- Is this data categorical or quantitative?
- Sketch two graphs of this data. Make one a bar chart and the other a pie chart. What are pros/cons of each graph?

66