AP Statistics – Lesson Notes - Chapter 3: Displaying Categorical Data

'Make a picture'

A picture really is worth a thousand words. Although words and numbers convey details well, they are a construct of human invention. **Our brains are pre-wired to interpret visual information quickly**, so finding a way to represent data with a picture often reveals information that is difficult to see when looking at raw data.

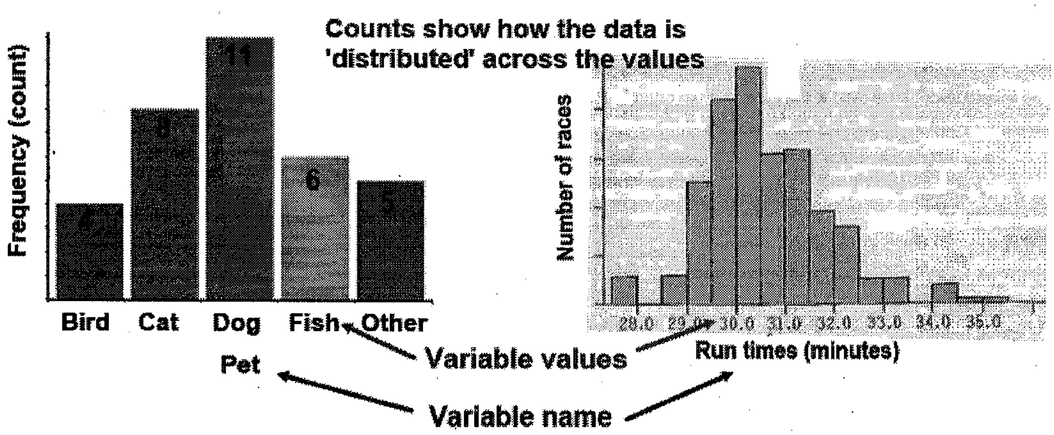
In Statistics, **we use different kinds of pictures to represent different kinds of data.**

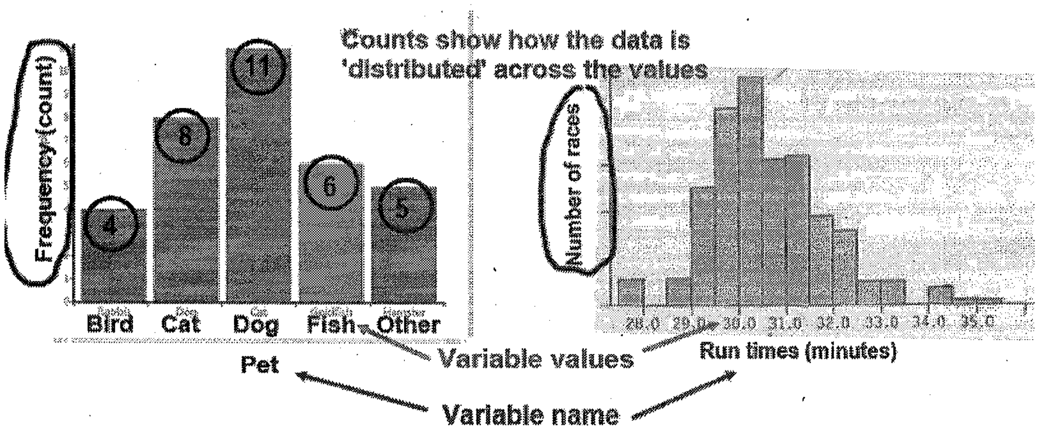
Categorical (Qualitative) vs. Numerical (Quantitative) data

There are two general types of data:

<u>Categorical (Qualitative) data</u> (Ch3)
- The variable's values are categories

<u>Numerical (Quantitative) data</u> (Ch4)
- The variable's values are numbers

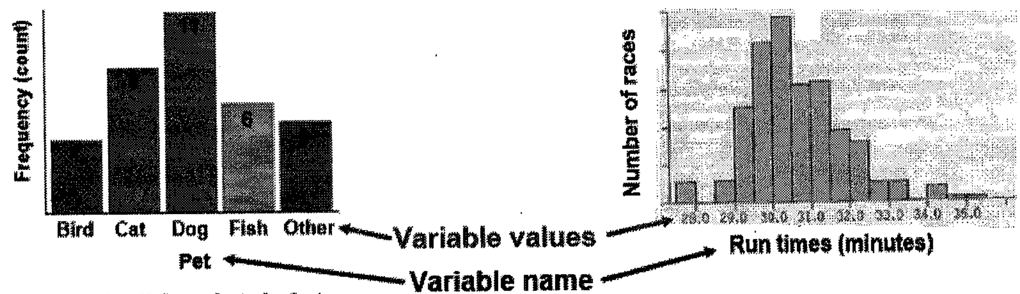




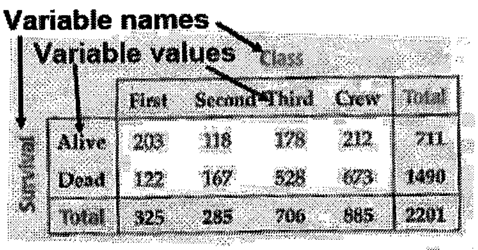
Important: The frequencies (counts) are <u>not data values</u>. They are showing how the data is distributed across the values of the variables.

Categorical (Qualitative) data (Ch3) | **Numerical Quantitative data** (Ch4)

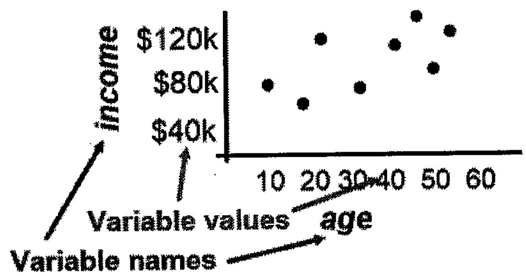
1 variable (univariate) data:



2 variable (bivariate) data:



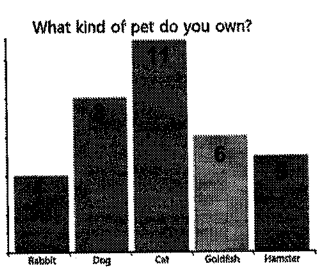
Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

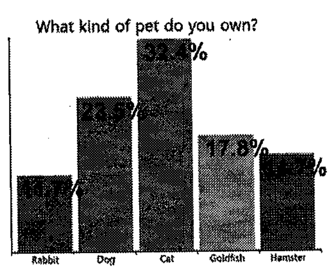


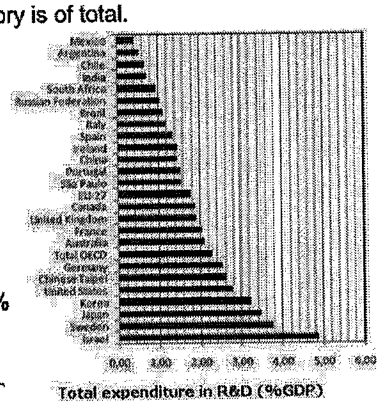
Displaying Categorical (Qualitative) data - Univariate

Bar Charts

- Can be vertical or horizontal.
- Height/length represents the amount in each category.
- Frequency charts: unit is frequency (count).
- Relative frequency charts: unit is percentage this category is of total.

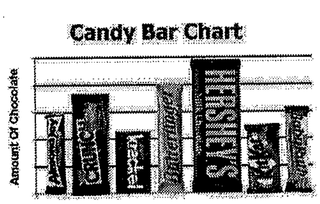




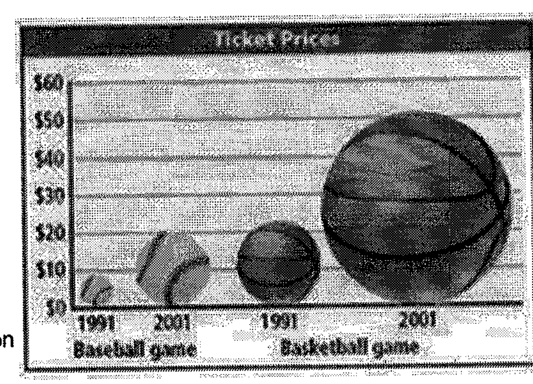


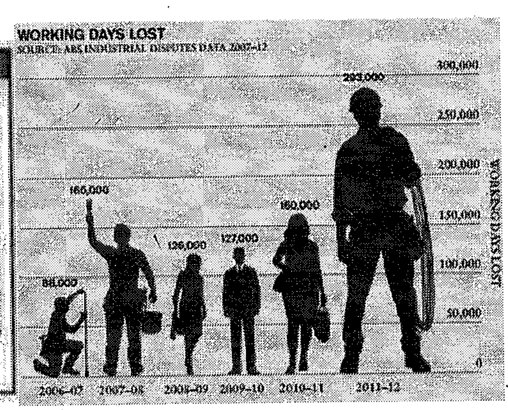
Bar Charts - ***The Area Principle***

The ***area*** occupied by a part of the graph must correspond to the magnitude of the value it represents.



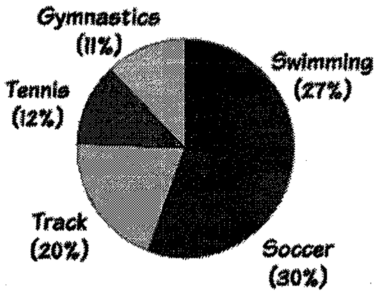
'Cute' charts can be okay, but do they convey information accurately?

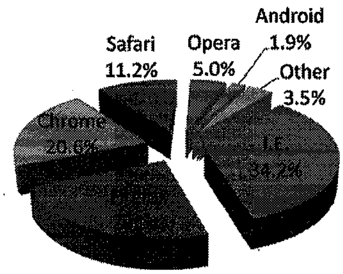




Pie Charts

- Areas of wedges represent percentage in category.
- Can be labeled as frequency/count or percentage.
- Sum of areas must total 100% of all categories.





Browser Usage on Wikimedia
October 2011

Displaying Categorical (Qualitative) data - Bivariate

Contingency Tables

- One variable in rows, one variable in columns.
- Each number in the table gives the output variable (count) for some combination (condition) of the input variables.

Example: Fate of people aboard the Titanic.

Variable	Categories
Class:	First, Second, Third, Crew
Survival:	Alive, Dead

		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

- Row and column totals are known as ***marginal distributions***.
- Marginal distributions show how the entire data set is distributed across one of the variables.

		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 Distribution of Survival:

Alive	Dead	
711	1490	/ 2201
(32%)	(68%)	

Marginal Distribution of Class:

First	Second	Third	Crew	
325	285	706	885	/ 2201
(15%)	(13%)	(32%)	(40%)	

Contingency Tables

Conditional Distributions

- Conditional distributions impose a condition by fixing the value of one of the variables, and then showing how that part of the data set is distributed across the other variable.

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

Conditional Distribution of Class, given they were Alive:

First	Second	Third	Crew	
203	118	178	212	/ 711
(29%)	(17%)	(25%)	(30%)	

Conditional Distribution of Class, given they were Dead:

First	Second	Third	Crew	
122	167	528	673	/1490
(8%)	(11%)	(35%)	(45%)	

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

Conditional Distribution of Survival, given they were First:

Alive	Dead	
203	122	/ 325
(62%)	(38%)	

Conditional Distribution of Survival, given they were Second:

Alive	Dead	
118	167	/ 285
(41%)	(59%)	

Conditional Distribution of Survival, given they were Third:

Alive	Dead	
178	528	/ 706
(25%)	(75%)	

Conditional Distribution of Survival, given they were Crew:

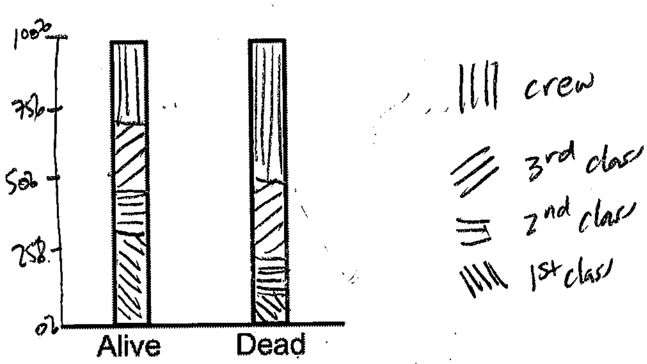
Alive	Dead	
212	673	/ 885
(24%)	(76%)	

Segmented Bar Graphs

- Since we are 'fixing' one variable (making it constant) in a conditional distribution, we effectively have only one other variable, so we can display that variable using a ***segmented bar graph***.

	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	
cumulative:	28.6	45.2	70.2	100	

	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	
cumulative:	8.2	19.4	54.8	100	

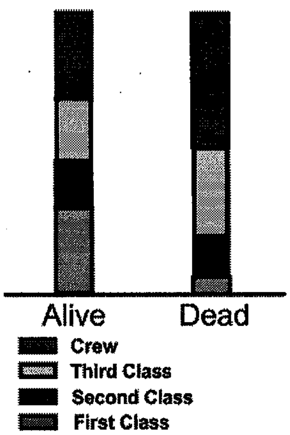


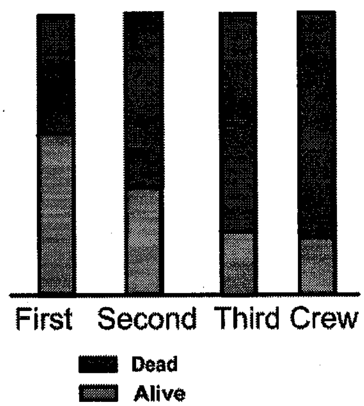
Are two variables dependent or independent?

If two variables are ***independent***, then the distribution of one variable should be the same regardless of the value of the other variable.

In the Titanic example, the variables 'survival' and 'class' are ***dependent*** because the distribution of one changes when the other variable changes. ~~You can~~ check this by examining conditional distributions of either variable against the other:

if distributions look significantly difference (% values more than 15% different then the variables depend on each other (are not independent)





Simpson's Paradox

Lurking Variable: A variable that affects data, but is not taken into account in a study.

What causes Simpson's Paradox:? A combination of a lurking variable and data from unequal sized groups being combined into a single data set.

Example: Admission into U.C. Berkeley

In 1973, admission records showed:

	applicants	admitted
men	2165	47%
women	849	31%

So...shame on U.C. Berkeley for admitting more men than women...right?

Well, what happens if we consider acceptance rate within each major separately...

major	men	women
1	511/825 62%	89/107 83%
2	352/560 63%	17/27 63%
3	137/407 34%	132/374 35%
4	22/373 6%	24/341 7%
all majors	1022/2165 47%	262/849 31%

For each individual major, as many (or more) women got accepted! But if you combine all majors together, a higher percentage of men got accepted overall. How can this happen?

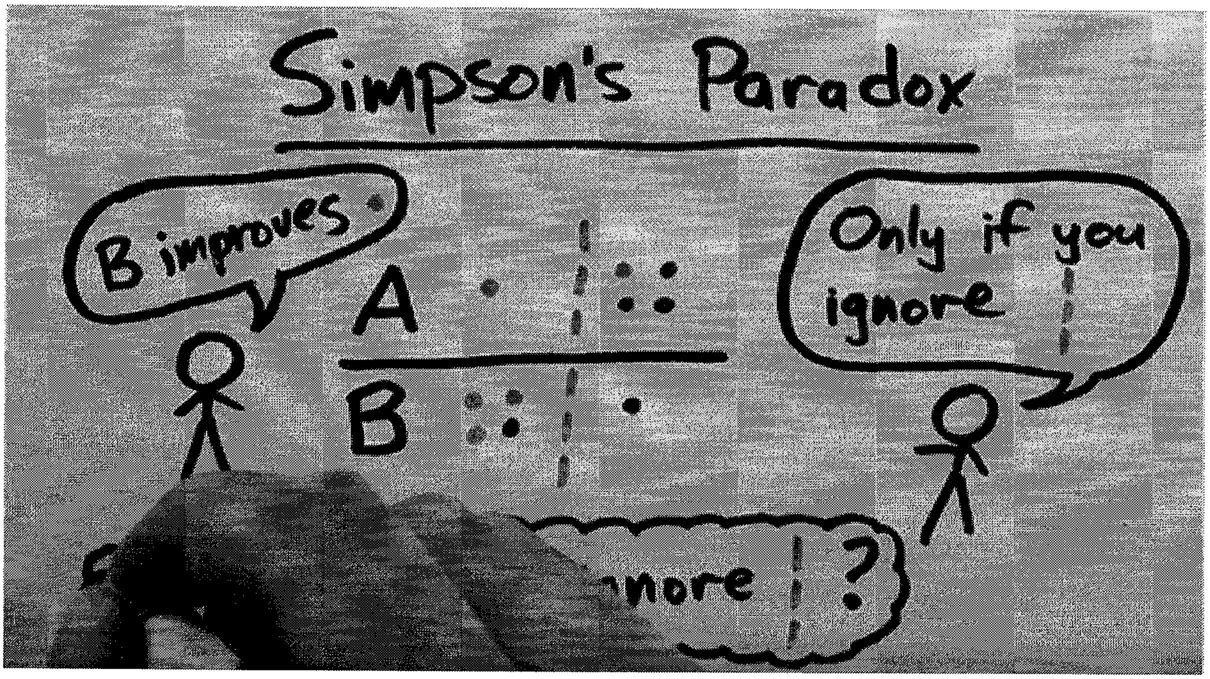
Look at the majors the men and women chose...

	major	men	women	
→ 38%	1	511/825 62%	89/107 83%	13%
→ 26%	2	352/560 63%	17/27 63%	3%
19%	3	137/407 34%	132/374 35%	44% ←
17%	4	22/373 6%	24/341 7%	40% ←
	all majors	1022/2165 47%	262/849 31%	

Men more often chose majors with higher acceptance rates and women chose majors with lower acceptance rates.

The lurking variable is the major. Not taking this into account by combining these unequal sized subgroups produces an erroneous conclusion.

Simpson's Paradox



https://www.youtube.com/watch?v=ebEkn-BiW5k

(Link included on www.mrfelling.com class page in the 'materials' section)