ISTA 116: Statistical Foundations for the Information Age

Univariate Categorical Data

24 August 2011

Outline

1. Reminders/Announcements
2. Types of Data
3. Categorical Data
4. Example: Plurality vs. Approval Voting
5. Visualizing Categorical Data
   - Frequency Tables
   - Bar Plots
     - Misleading Barplots
     - Dot Charts
   - Pie Charts

- Web Quiz 1 Due Friday
- Lab Assignment 1 posted (due date extended to next Friday). If downloaded before today, get latest version!
  - LaTeXtemplate available to fill in
- Video of Monday's class up soon
  - Podcast available from iTunes U (`http://itunes.arizona.edu`)
  - Streaming video linked from d2l
- SISTA Student Survey at end of class today

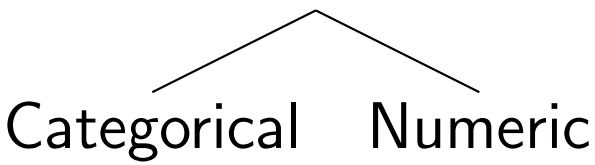
- The simplest type of data is **univariate**
  - One "outcome" per observation

- With two outcomes, we have **bivariate** data
- More than two, we usually just call it **multivariate**.
- Can further classify each outcome (or **variable**)

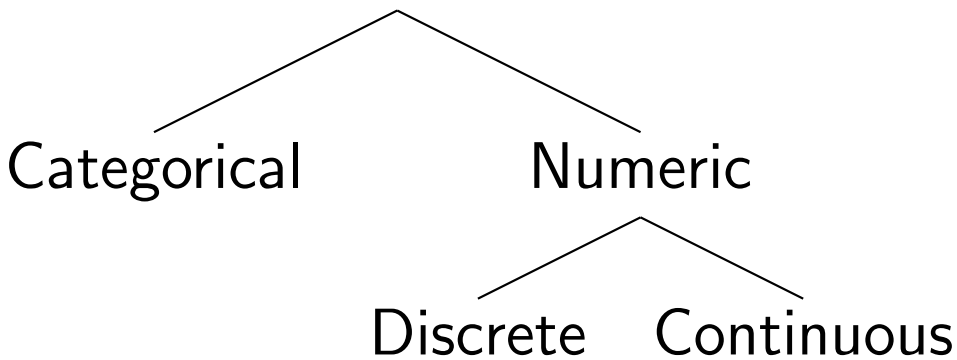
- Does the variable represent *qualitative* or *quantitative* information?

- Does the variable represent *qualitative*, or *quantitative* information?
  - Qualitative variables are known as **categorical**

- Does the variable represent *qualitative*, or *quantitative* information?
  - Qualitative variables are known as **categorical**
  - Quantitative variables are known as **numeric**



- Numeric variables can be further subdivided into **discrete** and **continuous**-valued varieties.



- Categorical variables include things like:
  - Sex (Levels: Male, Female)
  - Eye Color (Levels: Blue, Brown, Green, Other)
  - Home state (Levels: Alaska, Arkansas, Arizona, ...)
  - Others?

Key properties of *univariate* categorical data:

1. No underlying scale
   - No unique order
   - No "intermediate" values possible
2. Each observation must belong to only one category.
   - That is, categories are **mutually exclusive**, or **disjoint**.

Key properties of *univariate* categorical data:

1. No underlying scale
   - No unique order
   - No "intermediate values" possible
2. Each observation must belong to only one category.
   - That is, categories are **mutually exclusive**.
   - Should be able to put each observation on a line, with one column representing the variable.

Student #	**Home State**
1	Arizona
2	California
3	Arizona
4	Washington
...	...

Which of these are true univariate categorical variables?

- Major
- Dominant Hand
- Regular Church Attendance?
- Opinion about legalizing marijuana
- Letter grade
- Political identification
- Favorite hot beverage
- Which of these someone thinks are true univariate categorical variables

- Vote cast for president of the U.S. is a univariate, categorical variable.
- One important **statistic** resulting from this data is which candidate receives the most votes (technically, the **mode** of the data)
- However, with more than two candidates, this system violates the principle of "independence of irrelevant alternatives", i.e., that the group preference between A and B is influenced only by individual preferences between A and B; not, say, between B and C.

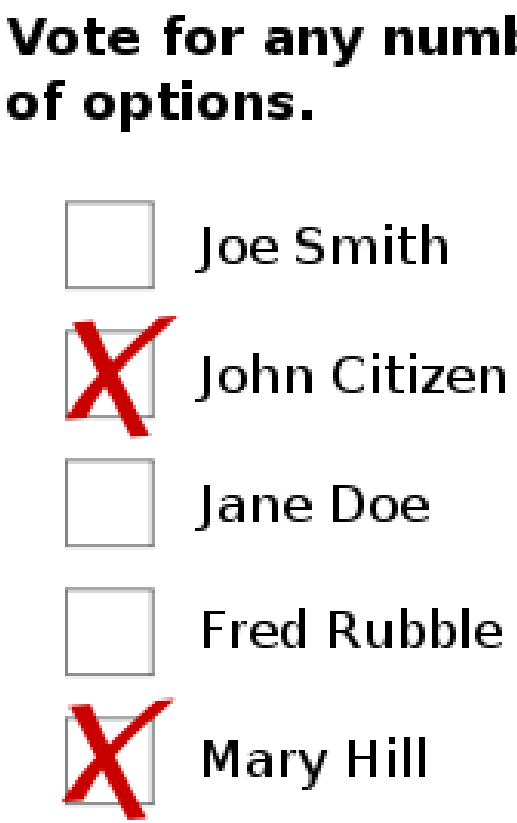
Here's an example vote:

Voter	First Choice	Second Choice	Third Choice
1	**Gore**	Nader	Bush
2	**Bush**	Gore	Nader
3	**Bush**	Gore	Nader
4	**Gore**	Bush	Nader
5	**Bush**	Nader	Gore
6	**Gore**	Nader	Bush
7	**Gore**	Nader	Bush
Winner	**Gore**		

But suppose voters 1 and 6 change the ordering of Gore and Nader, leaving Gore vs. Bush unchanged

Voter	First Choice	Second Choice	Third Choice
1	**Nader**	Gore	Bush
2	**Bush**	Gore	Nader
3	**Bush**	Gore	Nader
4	**Gore**	Bush	Nader
5	**Bush**	Nader	Gore
6	**Nader**	Gore	Bush
7	**Gore**	Nader	Bush
Winner	**Bush**		

An alternative system called "approval voting" fixes this by abandoning mutual exclusivity.



Something to think about: What new problems might this system introduce?

A Plurality of Americans Still Want New Stimulus

the Atlantic

contributors@theatlantic.com (Derek Thompson), On Thursday November 4, 2010, 10:30 am EDT

Given a list of priorities , a plurality of Americans say that a new stimulus bill is the most important thing the next Congress can do after the election. Weird?

Looking ahead, which of the following should be the highest priority for Congress after the election?

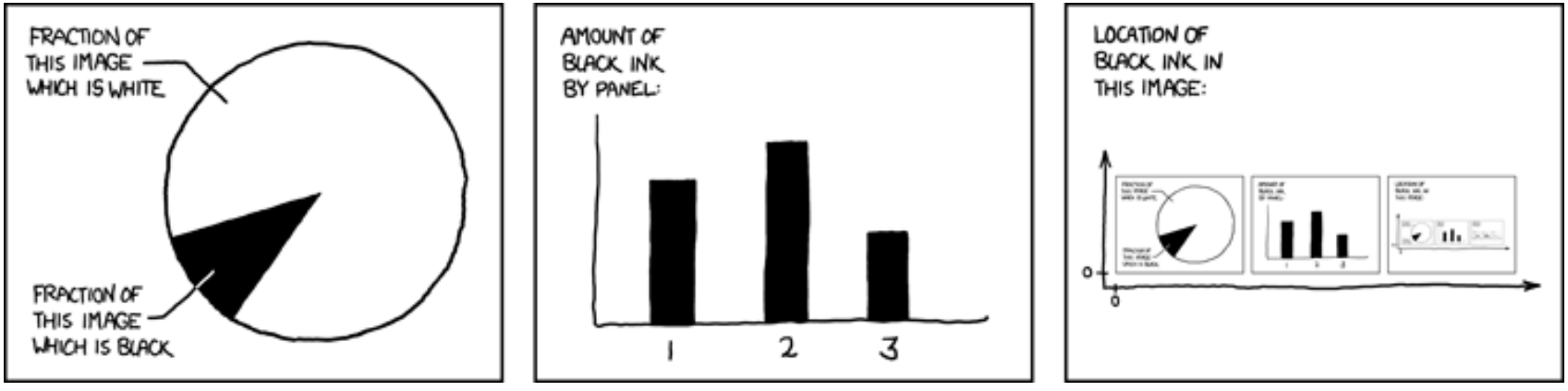
	% National adults	% Republicans	% Independents	% Democrats
Passing a new economic stimulus bill designed to create jobs	38	18	32	63
Cutting federal spending	24	29	28	15
Repealing the new healthcare law	23	36	23	12
Extending all federal income tax cuts enacted during the Bush administration	8	13	9	4

USA Today/Gallup, Oct. 28-31, 2010

GALLUP

Not so weird. Another way to put this chart in a sentence is: Americans would prefer the government focus on repealing Obamacare and cutting spending by a 47 to 38 margin over new stimulus. That's more or less in line with an election where Republicans took 56 percent of the House of Representatives.

Visualizing Categorical Data



Source: `http://www.xkcd.com/688`

- The only way to quantify truly categorical data is to count how often each value occurs.
- There are a few ways to display these counts.
- We'll talk about three:
  - Frequency Tables
  - Bar Plots
  - Pie Charts

- Raw data shows one observation per line
- When the number of categories is small compared to the number of observations, we can get all the info we need in a **frequency table**.
- Just display counts of each category.

Clear	**Partly Cloudy**	**Cloudy**
11	11	9

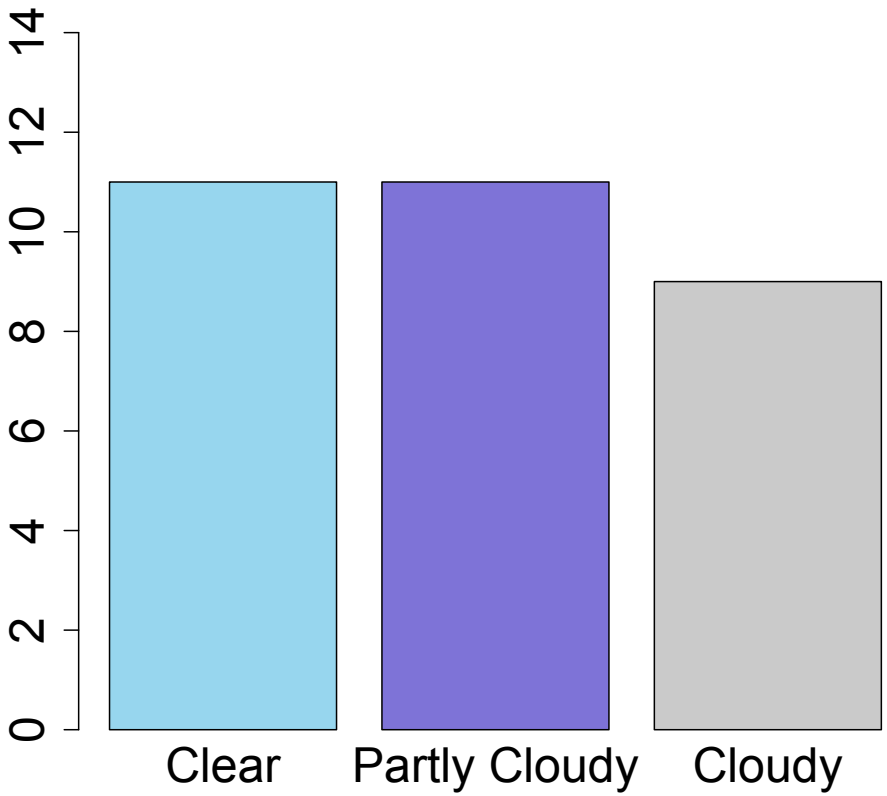
Table: Weather in Central Park in May

- We can convert counts into proportions by dividing each cell by the total **sample size**.
- These proportions are also called **relative frequencies**.
- Of course, to get percentages, multiply the relative frequencies by 100.

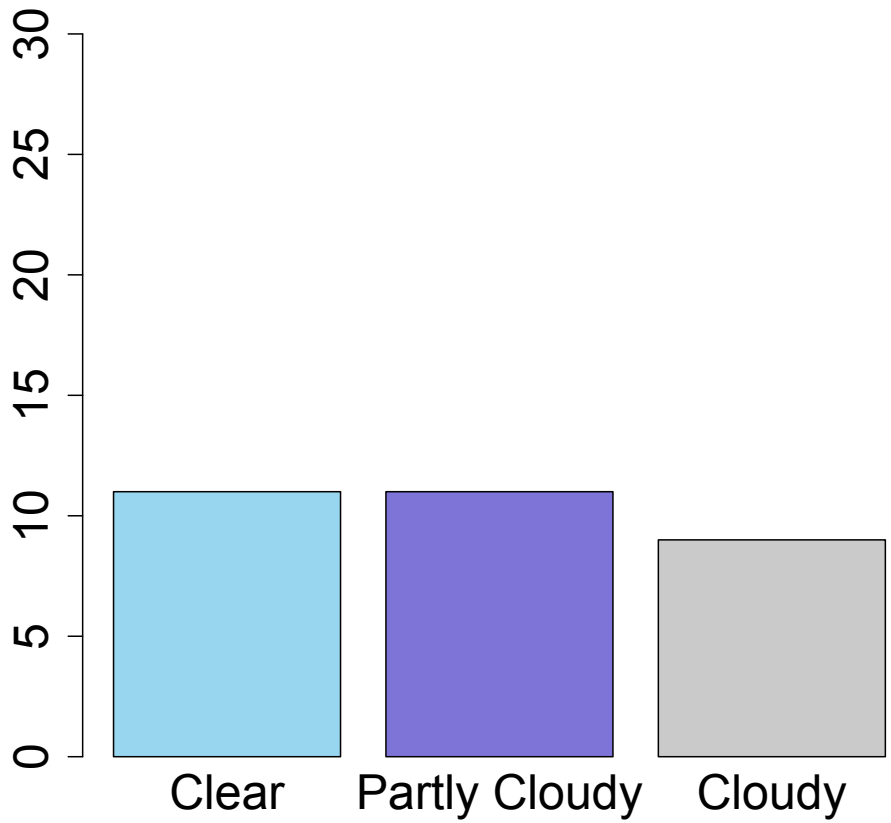
Clear	**Partly Cloudy**	**Cloudy**
0.355	0.355	0.290

Table: Weather in Central Park in May

- If we plot the frequencies or relative frequencies as bars (one per category), we get a **bar plot**.
- Notice the separation between bars, indicating separate categories.



- It's easy for the eye to judge differences by height, and ratios by area
- If proportion of the whole is of particular interest, might set the top of the y-axis to be the total.



- Because the eye is drawn to a ratio of areas, the bars should always start at zero!

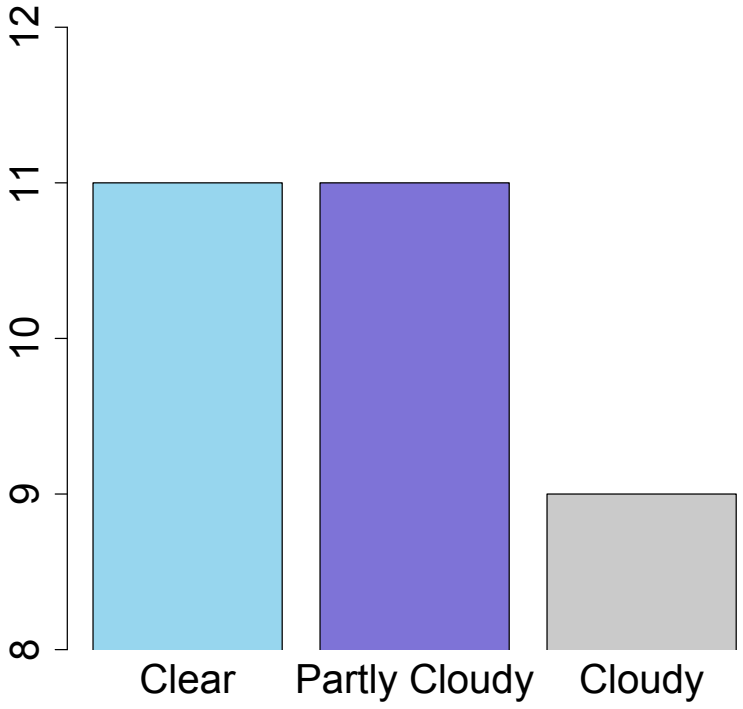


Figure: A deceptive bar plot

- To accentuate small differences without the false ratio information, can use a variation of the bar plot called a **dot chart**.

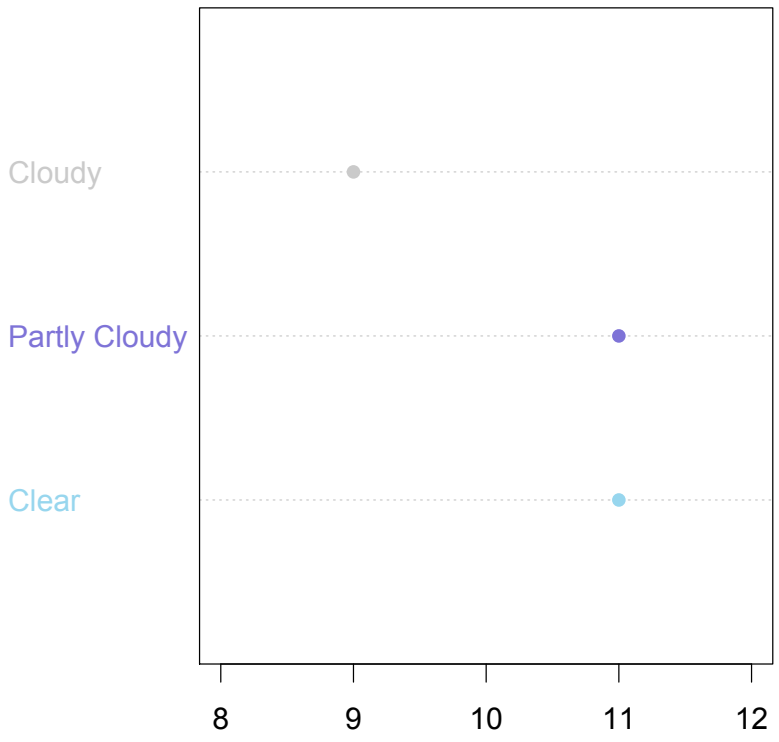
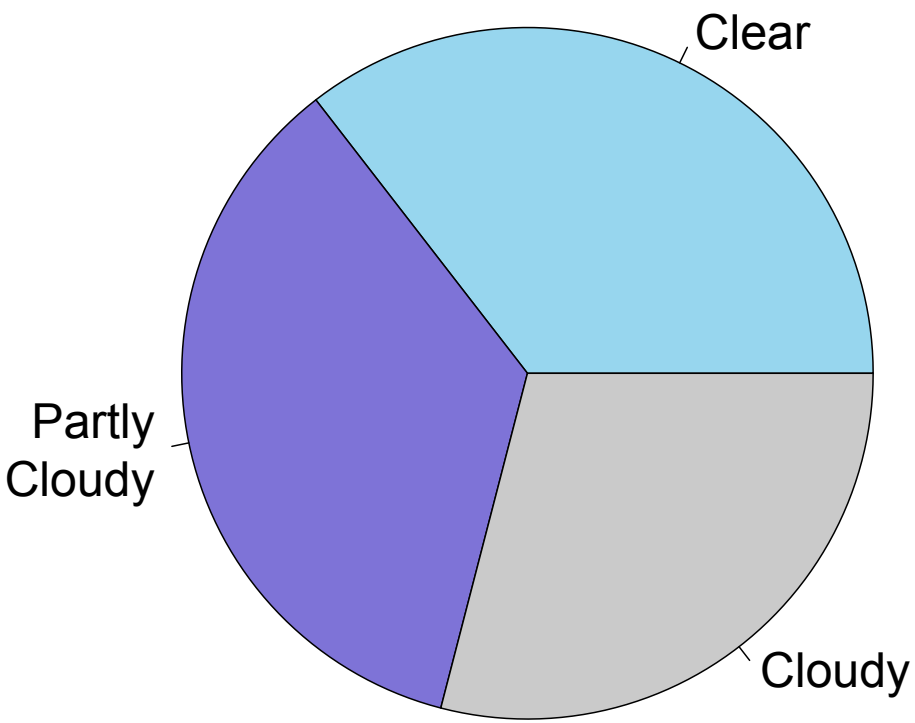


Figure: A Dot Chart of the Weather Data

- Another option is to represent proportions as angles
- Okay for judging each piece relative to the whole, not good for comparing "slices" (use a bar plot or dot chart).



- Next time: Numeric Data
- Reminder: Web Quiz Due Friday!