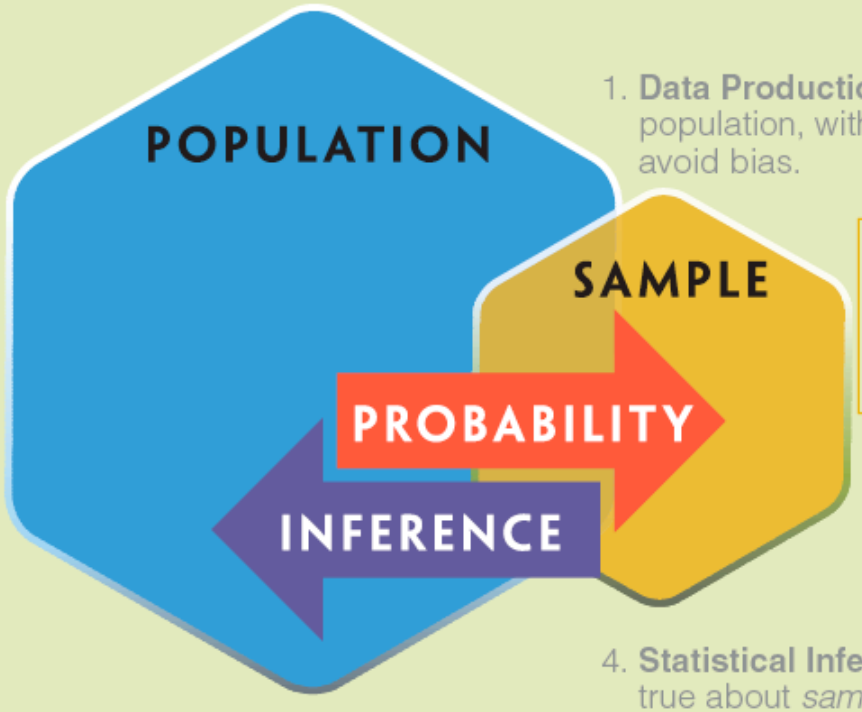
Lecture 5: Chapter 4, Section 1
Single Variables
(Focus on Categorical Variables)

- Displays and Summaries
- Data Production Issues
- Looking Ahead to Inference
- Details about Displays and Summaries

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing
    - Single variables: 1 categorical, 1 quantitative
    - Relationships between 2 variables
  - Probability
  - Statistical Inference

Focus on Displaying and Summarizing



1. **Data Production:** Take sample data from the population, with sampling and study designs that avoid bias.

2. **Displaying and Summarizing:** Use appropriate displays and summaries of the sample data, according to variable types and roles.

3. **Probability:** Assume we know what's true for the *population;* how should **random** *samples* behave?

4. **Statistical Inference:** Assume we only know what's true about *sampled* values of a single variable or relationship; what can we **infer** about the larger *population?*

Handling Single Categorical Variables

- **Display:**
  - Pie chart
  - Bar graph
- **Summary:**
  - Count
  - Percent
  - Proportion

Definitions and Notation

- **Statistic:** number summarizing sample
- **Parameter:** number summarizing population

- $\hat{p}$: sample proportion (a statistic) [“p-hat”]
- p: population proportion (a parameter)

Example: *Issues to Consider*

- **Background**: 246 of 446 students at a certain university had eaten breakfast on survey day.
- **Questions:**
  - Are intro stat students representative of all students at that university?
  - Would they respond without bias?
- **Responses:**
  -
  -

***Looking Back**: these are data production issues.*

Example: *More Issues to Consider*

- **Background**: 246 of 446 students at a certain university had eaten breakfast on survey day.
- **Questions:**
  - How do we display and summarize the info?
  - Can we conclude that a majority of *all* students at that university eat breakfast?
- **Responses:**
  - Display:

  Summary:
  -

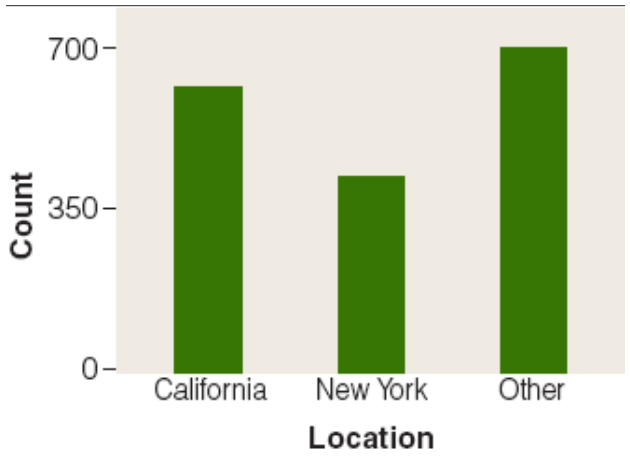
***Looking Ahead**: This would be statistical inference.*

Example: *Statistics vs. Parameters*

- **Background**: 246 of 446 students at a certain university had eaten breakfast on survey day.
- **Questions:**
  - Is 246/446=0.55 a statistic or a parameter? How do we denote it?
  - Is the proportion of all students eating breakfast a statistic or a parameter? How do we denote it?
- **Responses:**
  - 246/446=0.55 is a _____ denoted _____
  - Proportion of all students eating breakfast is a _____ denoted _____

Example: *Summary Issues*

- **Background**: Location (state) for all 1,696 TV series in 2004 with known settings:
  - 601 in California (601/1696=0.35)
  - 412 in New York (412/1696=0.24)
  - 683 in other states (683/1696=0.40)



- **Questions:**
  - 0.35+0.24+0.40=0.99→mistake?
  - Why is it not appropriate to use this info to draw conclusions about a larger population in 2004?
- **Responses:**
  -
  -

Example: *Notation*

- **Background**: In study of 20 antarctic prions (birds), 17 correctly chose the one of two bags that had contained their mate.
- **Questions:** How do we denote sample and population proportions? Are they statistics or parameters?
- **Responses:**
  - sample proportion is a
  - population proportion is a

Definitions

- **Mode:** most common value
- **Majority:** more common of two possible values (same as mode)
- **Minority:** less common of two possible values

Example: *Role of Sample Size*

- **Background**: In study of 20 antarctic prions (birds), 17 correctly chose the one of two bags that had contained their mate.
- **Question:** Would we be more convinced that a *majority* of all prions would choose correctly, if 170 out of 200 were correct?
- **Response:**

Example: *Sampling Design*

- **Background**: In study of 20 antarctic prions (birds), 17 correctly chose the one of two bags that had contained their mate.
- **Question:** Is the sample biased?
- **Response:**

Example: *Study Design*

- **Background**: Antarctic prions presented with Y-shaped maze, a bag at the end of each arm. One bag had contained mate, the other not.
- **Question:**
  - What were researchers attempting to show?
- **Response:**

Example: *Study Design*

- **Background**: Antarctic prions presented with Y-shaped maze, a bag at the end of each arm. One bag had contained mate, the other not.
- **Question:**
  - Why use bags and not birds themselves?
- **Response:**

Example: *Study Design*

- **Background**: Antarctic prions presented with Y-shaped maze, a bag at the end of each arm. One bag had contained mate, the other not.
- **Question:**
  - Why *"had"* contained (bird no longer in bag)?
- **Response:**

Example: *Study Design*

- **Background**: Antarctic prions presented with Y-shaped maze, a bag at the end of each arm. One bag had contained mate, the other not.
- **Question:**
  - OK to always place correct bag on right?
- **Response:**

Example: *Study Design*

- **Background**: Antarctic prions presented with Y-shaped maze, a bag at the end of each arm. One bag had contained mate, the other not.
- **Question:** Should the other be just any empty bag?
- **Response:**

Looking Ahead: *Researchers were careful to avoid bias in their study design. A success rate of 85% is impressive but we need inference methods to quantify claims that prions in general can recognize their mate by smell.*

Example: *Proportions in Three Categories*

- **Background**: Student wondered if she should resist changing answers in multiple choice tests. "Ask Marilyn" replied:
  - 50% of changes go from wrong to right
  - 25% of changes go from right to wrong
  - 25% of changes go from wrong to wrong
- **Question:** How to display information?
- **Response:**

Definition

- **Bar graph:** shows counts, percents, or proportions in various categories (marked on horizontal axis) with bars of corresponding heights

Example: *Bar Graph*

- **Background**: Instructor can survey students to find proportion in each year (1^{st}, 2^{nd}, 3^{rd}, 4^{th}, Other).
- **Questions:**
  - How to display the information?
  - What to look for in display?
- **Responses:**
  - Construct
    - on horizontal axis
    - graphed vertically
  - Look for
    compare

Example: *Overlapping Categories*

- **Background**: Report by ResumeDoctor.com on over 160,000 resumes:
  - 13% said applicant had "communication skills"
  - 7% said applicant was a "team player"
- **Question:** Can we conclude that 20% claimed communication skills or team player?
- **Response:**

Example: *Proportion from Raw Data*

- **Background**: Harvard study claimed 44% of college students are binge drinkers. Agree on survey design and have students self-report: on one occasion in past month, alcoholic drinks more than 5 (males) or 4 (females)? *Or use these data:*

yes	no	yes	no	no	yes
no	yes	yes	no	yes	no
yes	yes	no	no	yes	yes
yes	no	yes	yes	no	no
yes	no	yes	yes	yes	yes
no	no	yes	no	yes	no
no	yes	no	no	yes	no
no	no	no	no	yes	yes
yes	no	no	no	no	no
no	no	no	no	no	no
no	yes	yes	no	no	yes

- **Question:** Are data consistent with claim of 44%?
- **Response:**

Lecture Summary *(Categorical Variables)*

- **Display:** pie chart, bar graph
- **Summarize:** count, percent, proportion
- **Sampling:** data unbiased (representative)?
- **Design:** produced unbiased summary of data?
- **Inference:** will we ultimately draw conclusion about population based on sample?
- **Mode, Majority:** most common values
- **Larger samples:** provide more info
- **Other issues:** Two or more possibilities? Categories overlap? How to handle raw data?