Lecture 1: Chapters 1, 2 Introduction, Sampling

- Variable Types and Roles
- Summarizing Variables
- 4 Processes of Statistics
- Data Production; Sampling

Example: *What Statistics Is All About*

- **Background**: Statistics teacher has a large collection of articles and reports of a statistical nature.
- **Question:** How to classify them?
- **Background:** Statistics students are faced with a collection of exam problems at the end of the semester.
- **Question:** How to choose the right procedures to solve them?

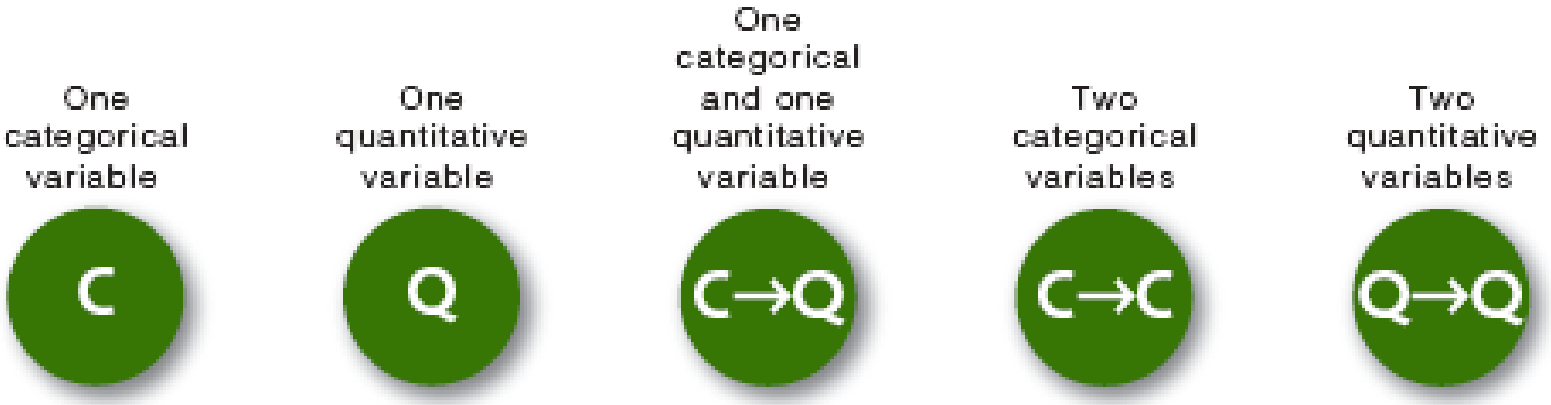
Example: *What Statistics Is All About*

- **Response (to both questions):** Statistics is all about…

Looking Ahead: *Identifying what kind of variables are involved is the key to classifying statistics problems and choosing the right solution tool.*

The Five Variable Situations

- When studying relationships between two variables, we often think of one as explanatory and the other as response.
- Depending on the variables' types and roles, we consider five possible situations.



Example: *Identifying Types of Variables*

- **Background**: Consider these headlines…
  - *Dark chocolate might reduce blood pressure*
  - *Half of moms unaware of children having sex*
  - *Vampire bat saliva researched for stroke*
- **Question:** What type of variable(s) does each article involve?
- **Response:**
  - Dark chocolate or not is
    blood pressure is
  - Being aware or not of children having sex is
  - Bat saliva or not is
    stroke recovery is probably

Example: *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Individual teenagers were surveyed about drug use.

Teenager	Marijuana?	Harder Drugs?
#1	Yes	Yes
#2	No	No
#3	No	No
#4	Yes	No
. . .	. . .	. . .

- **Question:** What type of variable(s) does this involve?
- **Response:**
  - marijuana or not is
  - harder drugs or not is

Example: *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Percentages of teenagers using marijuana or hard drugs are recorded for a sample of countries.

Country	% Marijuana	% Harder Drugs
#1	22	4
#2	37	16
#3	7	3
#4	23	14
. . .	. . .	. . .

- **Question:** What type of variable(s) does this involve?
- **Response:**
  - percentage using marijuana is
  - percentage using harder drugs is

Example: *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Percentages of teenagers using marijuana or hard drugs are recorded for a sample of countries.

Country	% Marijuana	% Harder Drugs
#1	22	4
#2	37	16
#3	7	3
#4	23	14
. . .	. . .	. . .

- **Question:** What type of variable(s) does this involve?
- **Response:** (another perspective)
  - type of drug (marijuana or harder drugs) is
  - % using the drugs is

Example: *Quantitative Variable Giving Rise to Categorical Variable*

- **Background**: Researchers studied effects of dental X-rays during pregnancy.
  - *First approach:* X-rays or not; baby's weight
  - *Second approach:* X-rays or not; classify baby's wt. as at least 6 lbs. (considered normal) or below 6 lbs.
- **Question:** What type of variable(s) does each approach involve?
- **Response**:
  - X-rays or not is _____ baby's weight is
  - X-rays or not is

    baby's wt. at least 6 lbs. or below 6 lbs. is

Definitions

- **Data**: recorded values of categorical or quantitative variables
- **Statistics:** science concerned with
  - gathering data about a group of individuals
  - displaying and summarizing the data
  - using info from data to draw conclusions about larger group

(All these skills are essential in both academic and professional settings.)

Summarizing Data

- **Categorical** data:
  - **Count:** number of individuals in a category
  - **Proportion:** count in category divided by total number of individuals considered
  - **Percentage:** proportion as decimal × 100%
- **Quantitative** data: **mean** is sum of values divided by total number of values

Example: *Summarizing Variables*

- **Background**: *"...1.9% of students nationwide got special accommodations for SAT...At 20 prominent NE private schools, nearly 1 in 10 received special treatment..."*
- **Question:** What type of variable is involved, and how is it summarized?
- **Response**: special accommodations for SAT is , summarized with or

Hint: think about who or what are the individuals. What information is recorded for each of them?

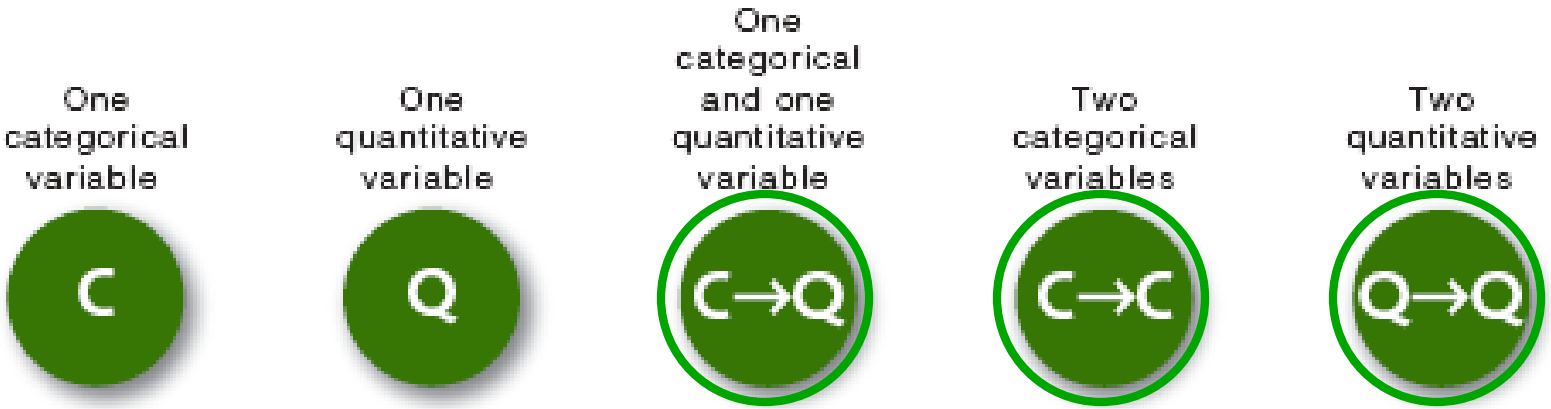
Example: *Summarizing Variables*

- **Background**: *"...On average, a white man with a college diploma earned $65,000 in 2001. Similarly educated white women made 40% less; black and Hispanic men earned 30% less..."*
- **Question:** What type of variable is considered for each demographic group, and how is it summarized?
- **Response**: Earnings is _____ summarize with _____

A Closer Look: *When comparing quantitative values for two or more categorical groups, we sometimes quantify the difference by reporting what percentage higher or lower one mean is compared to the other.*

Roles of Variables

When studying relationships between two variables, we often think of one as explanatory and the other as response.



Example: *Identifying Types and Roles*

- **Background:** Consider these headlines---
  - *Men twice as likely as women to be hit by lightning*
  - *Do Oscar winners live longer than less successful peers?*
- **Questions:** What types of variables are involved?
  For relationships, what roles do the variables play?
- **Responses:**
  - Gender is _____ and
    Hit by lightning or not is _____ and
  - Winning an Oscar or not is
    _____ and
    Life span is _____ and

Example: *More Identifying Types and Roles*

- **Background:** Consider these headlines---
  - *35% of returning troops seek mental health aid*
  - *Smaller, hungrier mice*
  - *County's average weekly wages at $811, better than U.S. average*
- **Questions:** What types of variables are involved?

  For relationships, what roles do the variables play?
- **Responses:**
  - Seeking mental health aid or not is
  - Size is _____ and

    Appetite is _____ and
  - Wages are

Definitions

- A **random** occurrence is one that happens by chance alone, and not according to a preference or an attempted influence.
- **Probability:** formal study of the chance of occurring in a random situation.
- **Statistical Inference**: drawing conclusions about population based on sample.

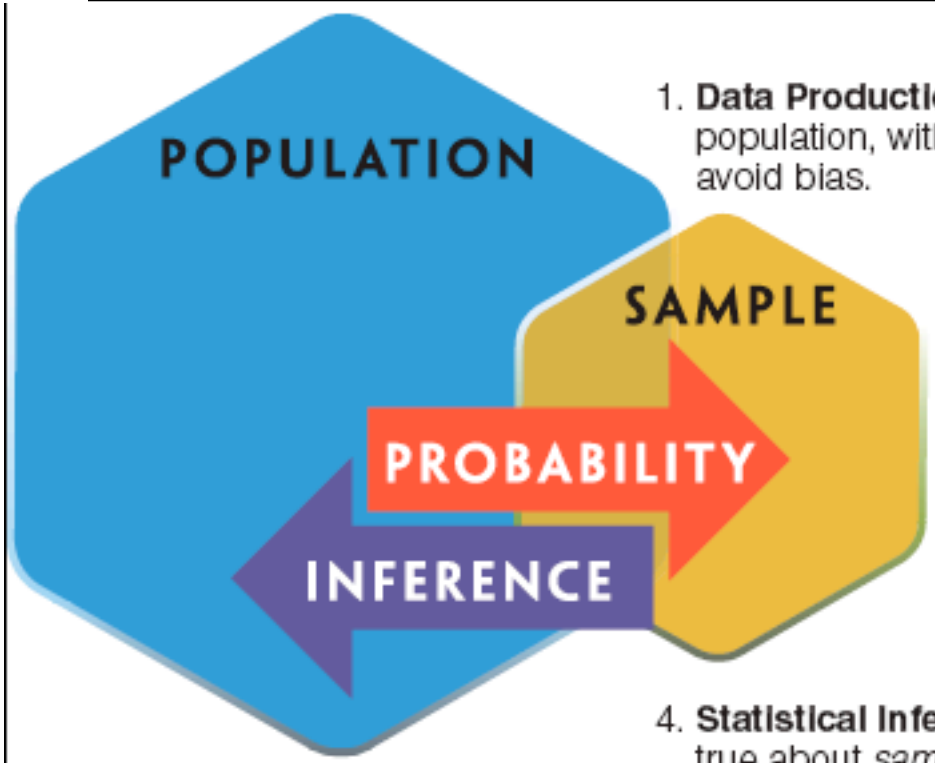
Looking Ahead: *Probability and Inference are linked through their roles in the 4-stage process of Statistics.*

Statistics as Four-Stage Process

- **Data Production**
- **Displaying and Summarizing**
- **Probability**
- **Statistical Inference**

Looking Ahead: *Besides the word "probability", a Probability statement may use the word "chance" or "likelihood" (the only synonyms available).*

Four Processes of Statistics



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?*

Data Production

- Use a good **sampling design** to get an unbiased sample so we can ultimately generalize from sample to population (Part 4)
- Create a good **study design** so what we learn is unbiased summary of what's true about the variables in our sample (Part 2)

Definition

- **Bias:** tendency of an estimate to deviate in one direction from a true value

Some sources of bias:

- selection bias: due to unrepresentative sample, rather than to flawed study design
- sampling frame doesn't match population
- self-selected (volunteer) sample
- haphazard sample
- convenience sample
- non-response

Example: *Bias in Sampling*

- **Background**: Professor seeks opinions of 6 from 80 class members about textbook…

1. *Have students raise hand if they'd like to give an opinion*
2. *Sample the next 6 students coming to office hours*
3. *Pick 6 names "off the top of his head"*

- **Questions:** Is each sampling method biased? If so, how?
- **Responses:**

1.

2.

3.

Example: *More Bias in Sampling*

- **Background**: Professor seeks opinions of 6 from 80 class members about textbook…

1. *Assign each student in classroom a number (1, 2, 3, …), then use software to select 6 at random…*
2. *Take a random sample from the roster of students enrolled; mail them anonymous questionnaire…*

- **Questions:** Is each sampling method biased? If so, how?
- **Responses:**

1.

2.

Definitions

- **Probability sampling plan** incorporates randomness in the selection process so rules of probability apply.
- **Simple random sample** is taken at random and without replacement.
- **Stratified random sample** takes separate random samples from groups of similar individuals (strata) within the population.

Definitions

- **Cluster sample** selects small groups (clusters) at random from within the population (all units in each cluster included).
- **Multistage sample** stratifies in stages, randomly sampling from groups that are successively more specific.
- **Systematic sampling plan** uses methodical but non-random approach (select individuals at regularly spaced intervals on a list).

Lecture Summary *(Introduction, Sampling)*

- **Variables**
  - Categorical or quantitative
  - Explanatory or response
- **Summaries**
  - **Categorical:** count, proportion, percentage
  - **Quantitative:** mean
- **4 Processes:** Data Production, Displaying and Summarizing, Probability, Inference
- **Data Production:** need unbiased sampling and unbiased study design
- **Types of Bias**
- **Types of Samples**