Topic 3

Gathering Data

Overview

Questions to think about

- What determines the extent to which conclusions can be generalized?

- How important are issues such as the sampling method and the amount of non-response?

- When is it appropriate to make conclusion about causes and effects?

Review

- *Population*:
  All the subjects of interest
- *Sample*:
  Subset of the population - data are collected on the sample

- *Response variable (Y)*:
  measures the outcome of interest
- *Explanatory variable (X)*:
  explains the response variable

Types of Studies

- Sample Survey: A study that asks questions / take measurements of the subjects in a sample drawn from the population randomly in the hope of learning something about the entire population. A census is a survey that attempts to count the number of people in the population and to measure certain characteristics about them.
- Experiments: An experiment is conducted by assigning subjects to certain experimental conditions (treatments) and then observing outcomes on the response variable.
- Observational Studies: A study that observes values of the response variable and explanatory variables for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).

Example: Does Drug Testing Reduce Students' Drug Use?

- **Headline: "Student Drug Testing Not Effective in Reducing Drug Use"**
- **Facts about the study:**
  - **76,000 students nationwide**
  - **Schools selected for the study included**
    schools that tested for drugs and
    schools that did not test for drugs
  - **Each student filled out a questionnaire asking about his/her drug use**

Example: Does Drug Testing Reduce Students' Drug Use?

Conditional Proportions on Drug Use			
Drug Tests?	**Drug Use** Yes	No	***n***
Yes	0.37	0.63	5653
No	0.36	0.64	17,437

Conclusion:

Drug use was similar in schools that tested for drugs and schools that did not test for drugs

Example: Does Drug Testing Reduce Students' Drug Use?

- **What were the response and explanatory variables?**

  Response varb :
  Explanatory varb :

- **Was this an observational study or an experiment?**

Topic 3

Gathering Data

Sample Survey

Setting Up a Sample Survey

- Step 1:
  Identify the Population

- Step 2:
  Compile a list of subjects in the population from which the sample will be taken. This is called the *sampling frame*.

- Step 3:
  Specify a method for selecting subjects from the sampling frame. This is called the *sampling design*.

  Good sampling designs employ randomization.

Random Sampling

- Randomizing protects us by giving us a representative / typical sample over effects we were unaware of.
- Best way of obtaining a representative sample
- Why not match the sample to the population?

Simple Random Sampling (SRS)

- A *simple random sample* of 'n' subjects from a population is one in which

  each possible sample of that size

  has the same chance of being selected

- Example
  Two will be chosen from these five members:

  President (P), Vice-President (V), Secretary (S), Treasurer (T) & Activity Coordinator (A)

- The possible samples are:

  (P,V) (P,S) (P,T) (P,A) (V,S)
  (V,T) (V,A) (S,T) (S,A) (T,A)

Selecting a Simple Random Sample

- Use a Random Number Table

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	10480	15011	01536	02011	81647	91646	69179	14194
2	22368	46573	25595	85393	30995	89198	27982	53402
3	24130	48360	22527	97265	76393	64809	15179	24830
4	42167	93093	06243	61680	07856	16376	39440	53537
5	37570	39975	81837	16656	06121	91782	60468	81305

- Use a Random Number Generator (Minitab)

Using Random Numbers to select a SRS

- To select a simple random sample:
  - number the subjects in the sampling frame using numbers of the same length (number of digits).
  - select numbers of that length from a table of random numbers or using a random number generator.
  - include in the sample those subjects having numbers equal to the random numbers selected.

Cluster Random Sampling

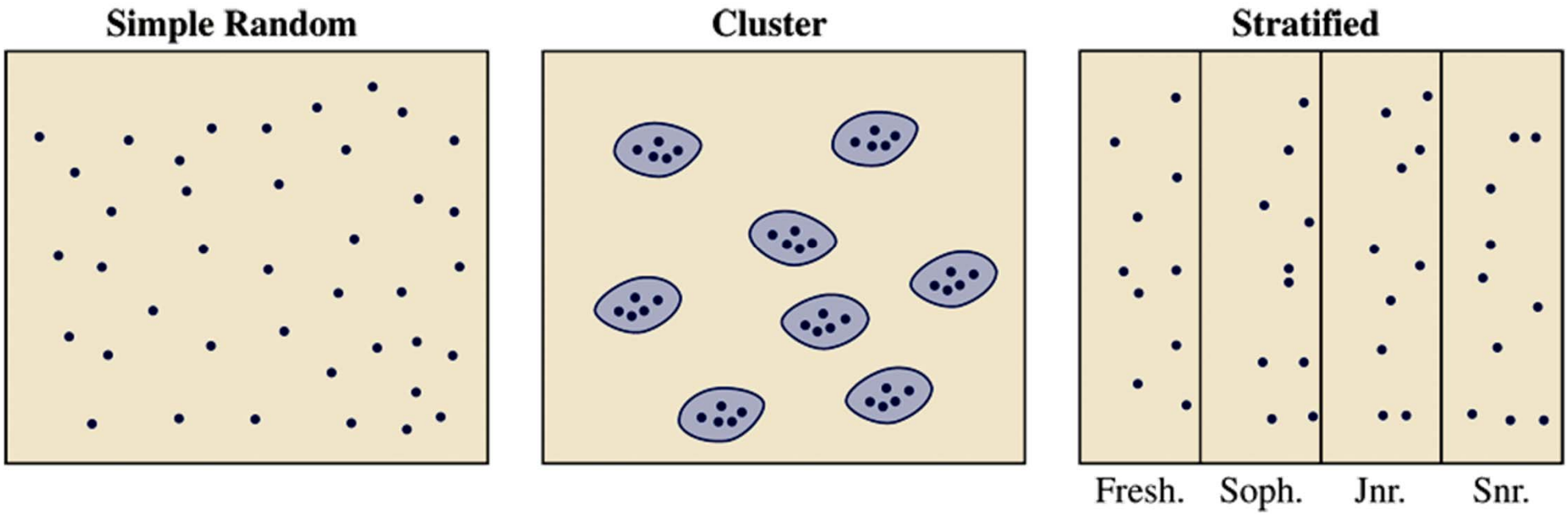
- Divide the population into many *clusters*
  (In a cluster, the subjects are heterogeneous).
- Select a simple random sample of the clusters.
- Use the subjects in those clusters as the sample.

Stratified Random Sampling

- Divide the population into separate groups, called *strata* (In a stratum, the subjects are homogeneous).
- Select a simple random sample from each stratum.
- Combine the samples from all the strata to form complete sample.

all groups represented

Comparing Random Sampling Methods



Comparing Random Sampling Methods

Method	Description	Advantages
Simple random sample	Each possible sample is equally likely	Sample tends to be a good reflection of the population
Cluster random sample	Identify clusters of subjects, take simple random sample of the clusters	Do not need a sampling frame of subjects, less expensive to implement
Stratified random sample	Divide population into groups (strata), take simple random sample from each stratum	Ensures enough subjects in each group that you want to compare

Disadvantage

Cluster : usually need a larger sample size than a SRS in order to achieve a particular margin of error.

Stratified : you must have a sampling frame and know the stratum into which each subject belongs.

Methods of Collecting Data in Sample Surveys

- Personal Interview
- Telephone Interview
- Self-administered Questionnaire

How Accurate Are Results from Surveys with Random Sampling?

- Sample surveys are commonly used to estimate population parameters.
- These estimates include a *margin of error* which tells us how well the sample estimate predicts the parameter.
- When a SRS of *n* subjects is used, the margin of error is approximately: $\frac{1}{\sqrt{n}} \times 100\%$
- Example: A survey result states:

  "The *margin of error* is plus or minus 3 percentage points"
- This means:

  "It is very likely that the reported sample percentage is no more than 3% lower or 3% higher than the population percentage.

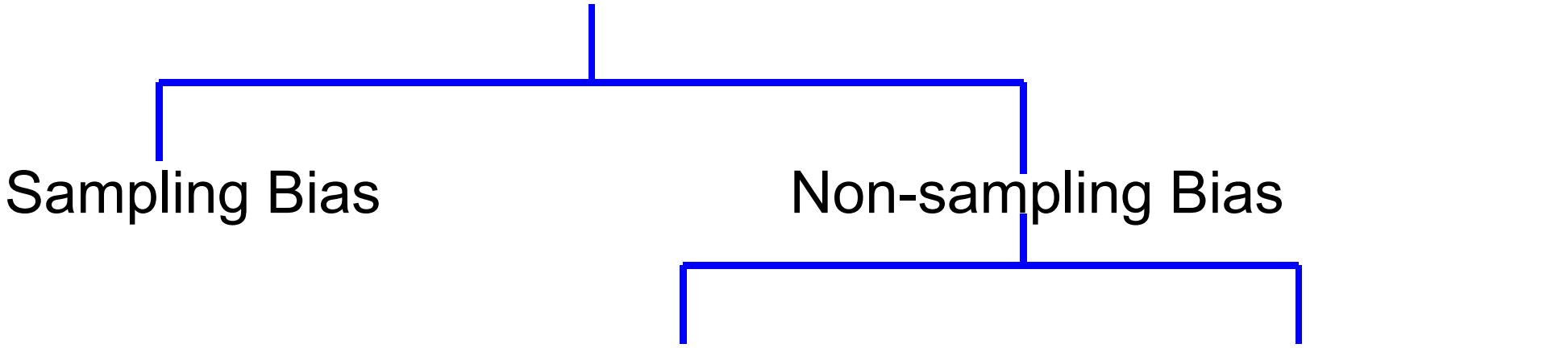
ok not ok

Error vs. Bias

- Sampling error is to be expected. It is caused by taking sample.
- A variety of problems can cause responses from a sample to tend to favor some parts of the population over others (BIASED).
- Bias is something we strive to avoid.

  It means our sample is out of step with the population, and that will surely lead to mistakes.

Types of Bias in Sample Surveys



- Subject gives incorrect response
- Misleading questions

- Cannot be reached
- Refuse to participate

- Under coverage
- Nonrandom sample

Convenience Sample

- A sample that is easy to obtain
  - Not a random sample
  - At a convenient time and location of the interview / judgment of the interviewer about whom to interview.
  - Unlikely to be representative of the population
  - Severe biases may result
  - *Volunteer Sample:* most common form of convenience sample

Warning:

A Large Sample Does Not Guarantee An Unbiased Sample

Question:
Which of the two SRS below is preferred?

N= population size and

n = sample size

a. N = 200	n = 100	50% sampled
b. N = 1,000,000	n = 1000	0.1% sampled

<10%, else violated independent assumption

Topic 3

Gathering Data

Experiment

Experiment

- Subject = *experimental unit*
- Experimental condition = *treatment*

We observe the outcome on the *response variable*

Investigate the *association*

– how the treatment affects the response

Elements of a Good Experiment - Control

- Primary treatment of interest
- Secondary treatment / control group

  (actual treatment / placebo)

Comparing the primary treatment results to the secondary treatment results help to analyze the effectiveness of the primary treatment.

Is the treatment group better, worse, or no different than the control group?

Example:
400 volunteers are asked to quit smoking and each start taking an antidepressant. In 1 year, how many have relapsed?
Without a control group (individuals who are not on the antidepressant), it is not possible to gauge the effectiveness of the antidepressant.

Elements of a Good Experiment - Control

- A ***placebo*** is a dummy treatment, a treatment known to have no effect, i.e. sugar pill.
  Many subjects respond favorably to any treatment, even a placebo.

- A **control group** typically receives a placebo. It is administered so that all groups experience the same conditions

- ***Placebo effect*** = Response to a dummy treatment. The placebo effect is an improvement in health due not to any treatment but only to the patient's belief that he or she will improve.

- By comparing with a placebo, we can be sure that the observed effect of a treatment is not due simply to the placebo effect.

Elements of a Good Experiment - Random

- It is important to randomly assign subjects to the treatments.

- In doing so, we
  - Eliminates bias that may result from the researcher assigning the subjects.
  - Balance the groups on variables known to affect the response.
  - Balance the groups on lurking variables that may be unknown to the researcher.

Elements of a Good Experiment - Blinding

- Ideally, subjects are unaware, or *blind*, to the treatment they are receiving.
- Placebo is used to blind the subjects.
- If an experiment is conducted in such a way that neither the subjects nor the investigators working with them know which treatment each subject is receiving, then the experiment is *double-blinded.*
- A double-blinded experiment controls response bias from the subject and experimenter.

Elements of a Good Experiment

- Comparison / Control
- Random
- Blinding
- Use enough subjects / Replication
  - Allows us to attribute observed effects to the treatments rather than ordinary variability.
  - Repeat studies to increase confidence in the conclusions.

Matched Pairs Design

- In a matched pairs design, the subjects receiving the two treatments are somehow matched (same person, husband/wife, two plots in the same field, etc.).

- Randomly assign the two treatments to the two matched subjects or

- Randomize the order of applying the treatments.

- The number of replicates equals the number of pairs.

- Helps to reduce effects of lurking variables.

Advantages of Experiments

- An **experiment** reduces the potential for *lurking variables* to affect the result. Thus, an experiment gives the researcher more control over outside influences.

- Only an experiment can establish **cause and effect**. Observational studies can not.

- **Experiments** are not always possible due to ethical reasons, time considerations and other factors.

Define Statistical Significance

- If an experiment (or other study) finds a difference in two (or more) groups, is this difference really **important**?

- If the observed difference is larger than what would be expected just by chance, then it is labeled **statistically significant**.

- Rather than relying solely on the label of *statistical significance*, also look at the actual results to determine if they are *practically significant*.

Generalizing Results

- Recall that the goal of experimentation is to analyze the association between the treatment and the response for the *population*, not just the sample.

- However, care should be taken to generalize the results of a study *only* to the population that is *represented by the study.*

Example: Reducing high blood pressure

A pharmaceutical company has developed a new drug for treating high blood pressure.

They would like to compare its effects to those of the most popular drug currently on the market.

Two hundred volunteers with a history of high blood pressure and who are currently not on medication are recruited to participate in a study.

- Identify the following

  experiment units

  treatments

  response variable

  explanatory variable

Example: Reducing high blood pressure

- How should the researchers assign the subjects to the two treatment groups?

Example: Reducing high blood pressure

- What else is important in designing this experiment?

Topic 3

Gathering Data

- Observational Studies

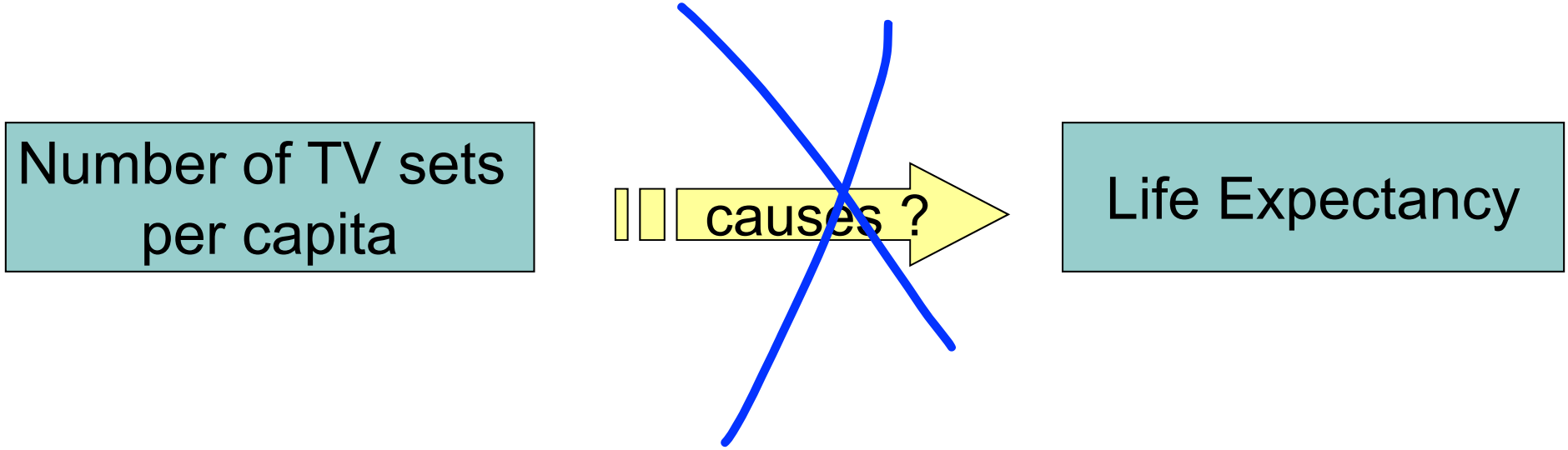
Experiments vs. Observational Studies

- When the goal of a study is to establish *cause and effect*, an experiment is needed
- An observational study can spot *associations* between variables, but cannot establish cause and effect.

  There is possibility of lurking or confounding variables.
- An observational study is a practical way of answering questions that do not involve trying to establish causality.
- It can yield useful information when an experiment is not practical (time constraints, ethical issues,..)

Lurking Variables in Observational Studies

There is a positive correlation between the number of TV sets per capita and the life expectancy.



Types of Observational Studies

An observational study can yield useful information when an experiment is not practical.

Types of observational studies:

- Sample Survey: attempts to take a cross section of a population at the current time.
- *Retrospective* study: looks into the past.
- *Prospective* study: follows its subjects into the future.

Causation can never be definitively established with an observational study, but well designed studies can provide supporting evidence for the researcher's beliefs.

Retrospective Case-Control Study

- A *case-control study* is a retrospective observational study in which subjects who have a response outcome of interest (the cases) and subjects who have the other response outcome (the controls) are compared on an explanatory variable
- *'Matching'* is used in creating control group for comparison.
- Similar subjects (in terms of gender, age, education, …) are selected for the control group. So that two groups (case & control) are as similar as possible.
- Matching reduced unwanted variation.

Example: Case-Control Study

- Response outcome of interest: Lung cancer
  - The *cases* have lung cancer
  - The *controls* did not have lung cancer

- The two groups were compared on the explanatory variable: Whether the subject had been a smoker

  The cases had lung cancer. The controls did not. The retrospective aspect refers to studying whether subjects had been smokers in the past

Example: Case-Control Study

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	**709**	**709**

Compare the proportion of smokers for the lung cancer cases and the controls.

Example: Case-Control Study

- Are subjects matched according to all possible lurking variables? no
- Can we prove a causal link between smoking and lung cancer? no
- The conditional proportions are meaningless for case-control study and thus could not be generalized to the entire population. 51%of smokers had lung cancer
- Why? they are not randomly drawn from the population, thus not representative of the entire population. It can't be generalized to the entire population

Example: Prospective Study

Nurses' Health Study:

- Began in 1976 with 121,700 female nurses aged 30 to 55; questionnaires are filled out every two years.

- Purpose was to explore the relationships among diet, hormonal factors, smoking habits and exercise habits and the risk of coronary heart disease, pulmonary disease and stroke.

- Nurses are followed into the future to determine whether they eventually develop an outcome such as lung cancer and whether certain explanatory variables are associated with it.

Example: Type of study

Among a group of married women who were tracked for ten years, those who worked full time were more likely to divorce than those who did not work full time.

a. Sample survey

b. Randomized experiment

c. Retrospective observational study

d. Prospective observational study