# STAT 22000 Lecture Slides
# Data Collection

Yibi Huang
Department of Statistics
University of Chicago

# Outline

The set of slides covers Section 1.3-1.4 of the text and more.

- Experiments (1.4)
- Observational Studies (1.3.4)
- Sampling (1.3)

# Experiments

# How to Establish Causality? (1)

General Question: How to test whether a new drug or a new treatment is effective?

- Applying the drug on one patient, if it works, is that suffice to prove the effectiveness of the drug?
- The patient may just recover for some other reason, not because of the drug. Need to test the drug on more people.
- **Principle 1 of experimental design** — *replication*

# How to Establish Causality? (2)

- E.g., want to prove Vitamin C prevent colds
- Study design: Ask 100 subjects taking a vitamin C pill everyday, and see if they get colds within the next 3 months.
- If only 10 of them get a cold with the next 3 month, does that suffice to prove the claim?
- Not flu season? Healthy subjects?
- Need another group of subjects, w/o taking Vitamin C, to compare with, called a *control group*
- **Principle 2 of experimental designs** — *control*

# How to Establish Causality? (3)

Basic method: **Comparison**

- Divide people (subjects) into 2 groups: *treatment group* and *control group*
- Give the new treatment to the treatment group but not to the control group (or give the control group another treatment);
- Compare: if the outcome is better on average in the treatment group, then the treatment is effective, but there are other concerns.

# How to Establish Causality? (4)

Key to the method: The 2 groups should be *as similar as possible*

- Ideally, if the two groups are identical except for being treated or not, the difference in the outcome must be due to the treatment. Causality can be established.
- If the two groups differ in other aspects, like one group is older than the other on average, then age can also cause the difference in the outcome, not sure if the new treatment is effective. Causality cannot be established.

A *confounder* is a factor that can also explain the difference in the outcomes of the treatment group and control group

- Also called: confounding variable, confounding factor, lurking variable

# Strategies to Combat Confounding

For example, in the Vitamin C study, age of the subjects is a confounder since older subjects might be more likely to get colds.

We can

- restrict the confounder (e.g. use only 18-35 year old subjects)
- balance the confounder (e.g. make the age distributions in the two groups similar)

However, there are too many confounders to balance, and many of them are unknown.

# Principle 3 of Experimental Designs — Randomization

*Randomization* is a simple way to ensure balance: which means assigning subjects to the treatment and control groups **randomly**.

- To avoid human factors, use coin tossinging/random number table/random number generator.
- By the law of large number, the treatment and control groups should be similar in all aspects
  — *Randomization is the (only) golden standard for causality*
- Extreme allocations are possible by randomization (like all healthier subjects in one group and weaker subjects in the other), but very unlikely.

# Example: Breast Cancer Screening[1]

- Since 1963, the Health Insurance Plan of Greater New York conducted large-scale screening trial for breast cancer. The goal is to see if screening programs speed up detection of breast cancer by enough to matter.
- Subjects: 62,000 women age 40 to 64, all members of the plan, were divided at random into two equal groups.
- In the treatment group, women were encouraged to come in for annual screening, including examination by a doctor and X-rays.
  - about 20,200 women did come in for the screening:
  - but 10,800 refused.
- The control group was offered usual health care.
- All the women were followed for 5 years.

[1] p.22 in *Statistics* 4ed, by David Freedman, Robert Pisani, and Roger Purves

# Example: Breast Cancer Screening

| | | Cause of Death in the first 5 years | | | |
|---|---|---|---|---|---|
| | | Breast Cancer | | Other Diseases | |
| Treatment Group | Size | Count | Percent | Count | Percent |
| Examined | 20,200 | 23 | 0.11 | 428 | 2.1 |
| Refused | 10,800 | 16 | 0.15 | 409 | 3.8 |
| Total | 31,000 | 39 | 0.13 | 837 | 2.7 |
| Control Group | 31,000 | 63 | 0.20 | 879 | 2.8 |

To see if the screening save lives from breast cancer, which two groups should we compare?

(a) Examined v.s. Refused
(b) Examined v.s. Control
(c) Examined v.s. (Refused + Control)
(d) Treatment (= Examined + Refused) v.s Control

# Placebo Effect & Blinding

After randomization, some confounders might be created during the progress of experiments . . .

- **Placebo effect**: Knowing being treated or not might be confounding, e.g., patients may feel better just by knowing they are treated
- **Single-blind**: keep the subjects from knowing whether they are in treatment or in control
- **Double-blind**: neither the subjects nor the researchers who interact with the patients know the allocation of subjects to the two groups
- To achieve blinding, subjects in the control group are given **placebos** like inert pills, sham surgery, or other deceive procedures similar to the treatment
- Sometimes blinding is impossible, or unethical

# Blocking – A More Sophisticated Design Technique


- Want to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- If some runners are known to be pro, and some are amateur, their performance may differ greatly.
- Randomization may divide pro and amateur runners roughly even between the two groups, but may not be exactly even.

# Blocking – A More Sophisticated Design Technique

As the pro/amateur status of runners is known, we should use this information in the allocation of runners to ensure that pro and amateur runners are split evenly between the two groups as follows.

- Divide runners to pro and amateur
- Randomly assign pro runners to gel and placebo groups
- Randomly assign amateur runners to gel and placebo groups
- Pro/amateur status is equally represented in the resulting treatment and control groups

$$\text{50 Runners} \begin{cases} \text{10 pro} & \begin{cases} \text{5 gel} \\ \text{5 placebo} \end{cases} \\ \text{40 amateur} & \begin{cases} \text{20 gel} \\ \text{20 placebo} \end{cases} \end{cases}$$

This is a experiment design that *block* on pro/amateur status of runners.

# Block Design in General

- In a randomized block design,
  - available subjects are first divided into groups called *blocks*, in a way that subjects in the same block are more similar than subjects in different blocks
  - subjects in each block are then split evenly to the treatment and the control group by randomization
- The more homogeneous subjects within a block and the more heterogeneous between blocks, the better a block design works.
- Block what you know, randomize the rest.

# Recap: Principles of Experimental Design

1. *Control:* Compare treatment of interest to a control group.
2. *Randomize* – powerful tool to remove confounders and golden standard to establish causality
3. *Replicate:* Within a study, replicate by collecting a sufficiently large sample to make sure the difference in the outcome is not due to chance. Or replicate the entire study.
4. *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

# Observational Studies

# Observational Studies

- We have said that randomized controlled experiments are the gold standard for determining cause-and-effect relationships
- However, such experiments are not always possible, ethical, or affordable
- A much simpler, more passive approach is to simply observe people's decisions and the consequences that seem to result from them, then attempt to link the two
- Such studies are called *observational studies*

# Smoking

- For example, smoking studies on human are observational – can't force one to take up smoking just to do experiments
- However, the idea of treatment (smokers) and control (nonsmokers) groups is still used, just as it was in controlled experiments
- The essential difference, however, is that the subject assigns themselves to the treatment/control group – the investigators just watch
- Because of this, confounding is possible
- Hundreds of studies have shown that smoking is *associated* with various diseases, but none can prove *causation*

# Example – Defibrillators in Hospitals v.s. in Casinos

## *Study finds hospitals slow to defibrillate*

### *Researchers say they're riskier than a casino in event of cardiac arrest.*


. . . Doctors already knew that more than half of those who suffer such attacks in airports and casinos survive. But a new study shows that only a third of victims in hospitals survive – primarily because patients do not receive life-saving defibrillation within the recommended two minutes.

Nearly 40% of hospital patients who received defibrillation within two minutes survived, compared with 22% of those for whom the response took longer, researchers reported in the *New England Journal of Medicine*. . . .

"It is probably fair to say that most patients assume – unfortunately, incorrectly – that a hospital would be the best place to survive a cardiac arrest," USC cardiologist Leslie Saxon wrote in an editorial accompanying the report.

People who suffer cardiac arrest in the middle of an airport or casino – where defibrillators are widely available – are typically noticed immediately, whereas a lone patient suffering an attack in a hospital room may not be noticed for much of the crucial window of opportunity during which defibrillation is most effective.

— News clip from *Los Angeles Times* in January 03, 2008

# Example – Defibrillators in Hospitals v.s. in Casinos (Cont'd)

- What are the treatment and control group compared in the news clip?
- Is the title of the news clip correct? Can you find other reasons why those having heart attack in hospital and receiving defibrillation were less likely to survive than those having it in a casinos or an airport, other than hospital being slow?

# Find the Confounders

- Study #1: Researchers found that students who eat breakfast tend to have better test scores than students who don't. They conclude that eating breakfast makes students better learners.

- Study #2: The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking. The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop.

# Find the Confounders

Study #3: Eating more chocolate produces more Nobel laureates?

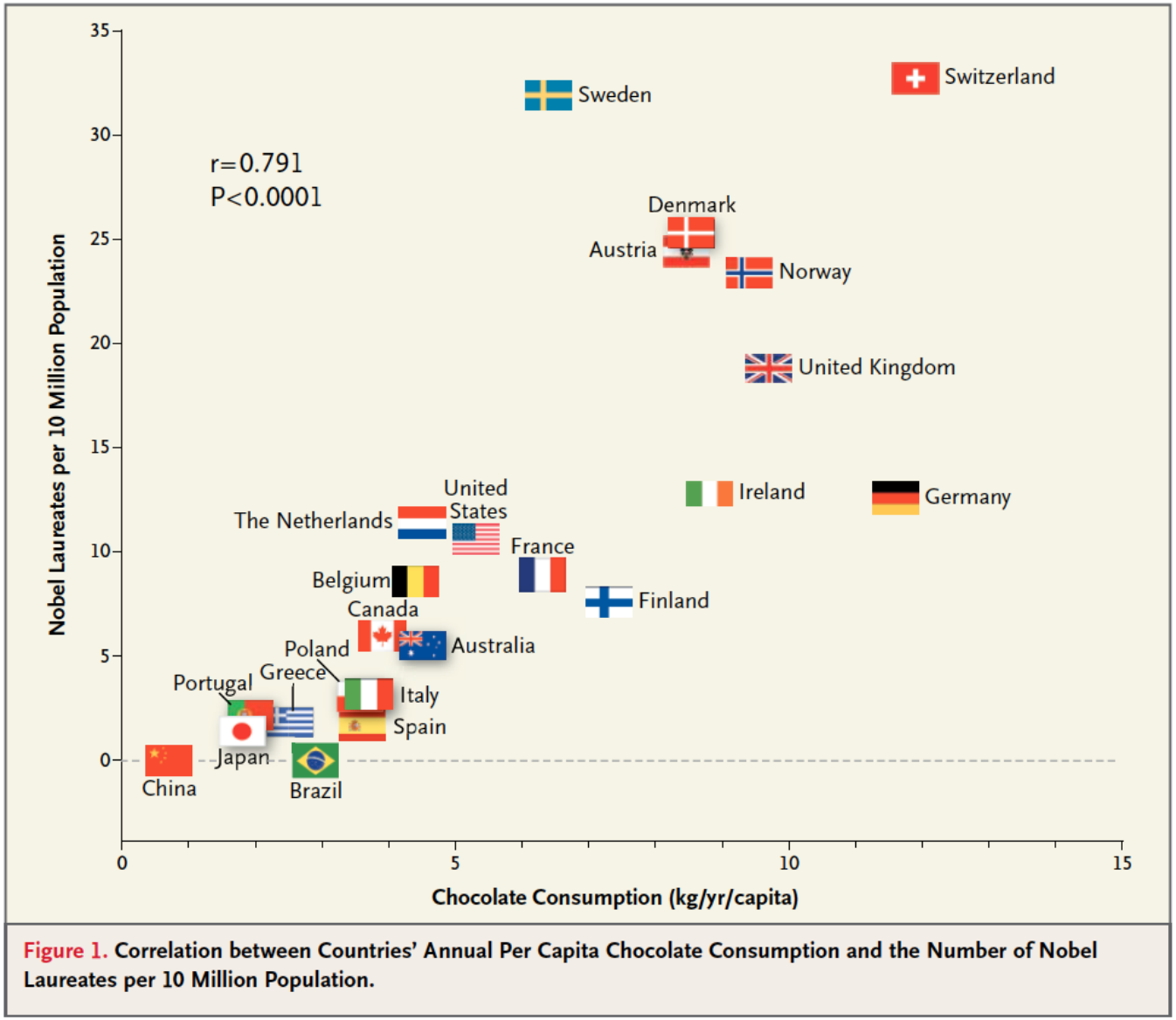



**Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Source: http://www.nejm.org/doi/full/10.1056/NEJMon1211064

# Controlling for Confounders

- However, just because confounding is possible in such studies does not mean that investigators are powerless to address it
- Instead, well-conducted observational studies make strong efforts to identify confounders and control for their effect
- There are many techniques for doing so; the most direct approach is to make comparisons separately for smaller and more homogeneous groups

# Controlling for Confounders (Cont'd)

- For example, studying the association between heart disease and smoking could be misleading, because men are more likely to have heart disease and also more likely to smoke
- A solution is to compare heart disease rates separately: compare male smokers to male nonsmokers, and the same for females
- Age is another common confounding factor that epidemiologists are often concerned with controlling for

# Example: Gender Gap in Income?

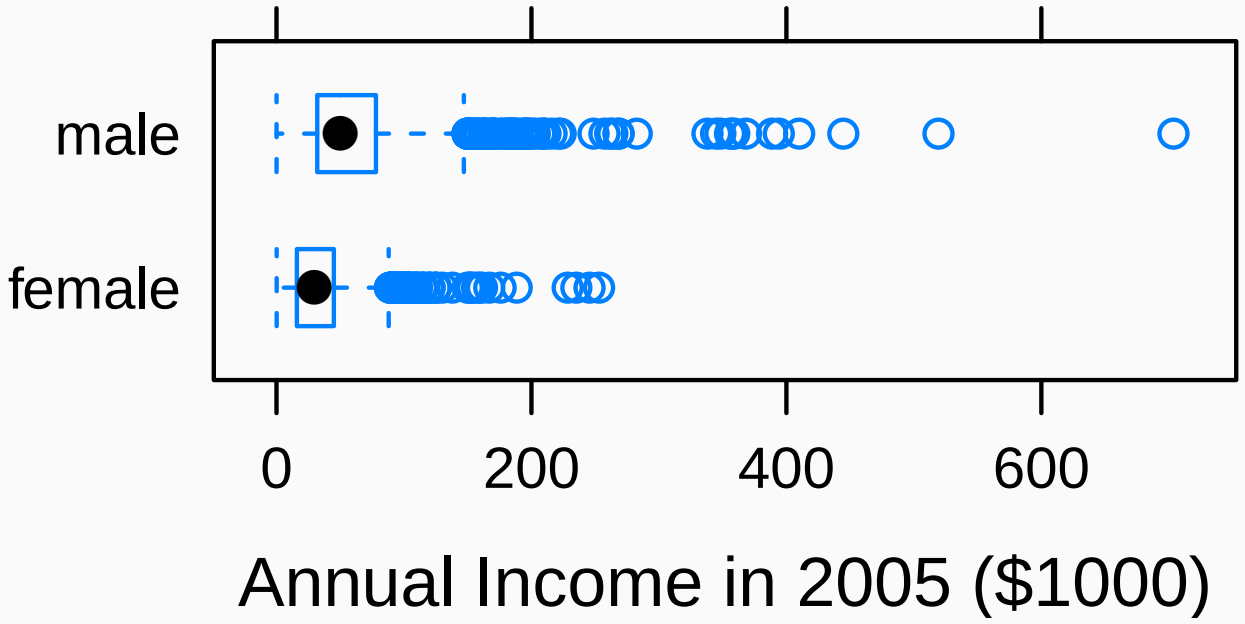
**Data**: 1306 American men and 1278 American women between the ages of 14 and 22 in 1979.

**Variables**

- *Gender*
- *AQFT*: intelligence test scores percentiles measured in 1981
- *Edu2006*: years of education achieved by time of interview in 2006
- *Income2005*: annual income in dollars in 2005

Original data come from the National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics https://www.bls.gov/nls/home.htm.

# Example: Gender Gap in Income?

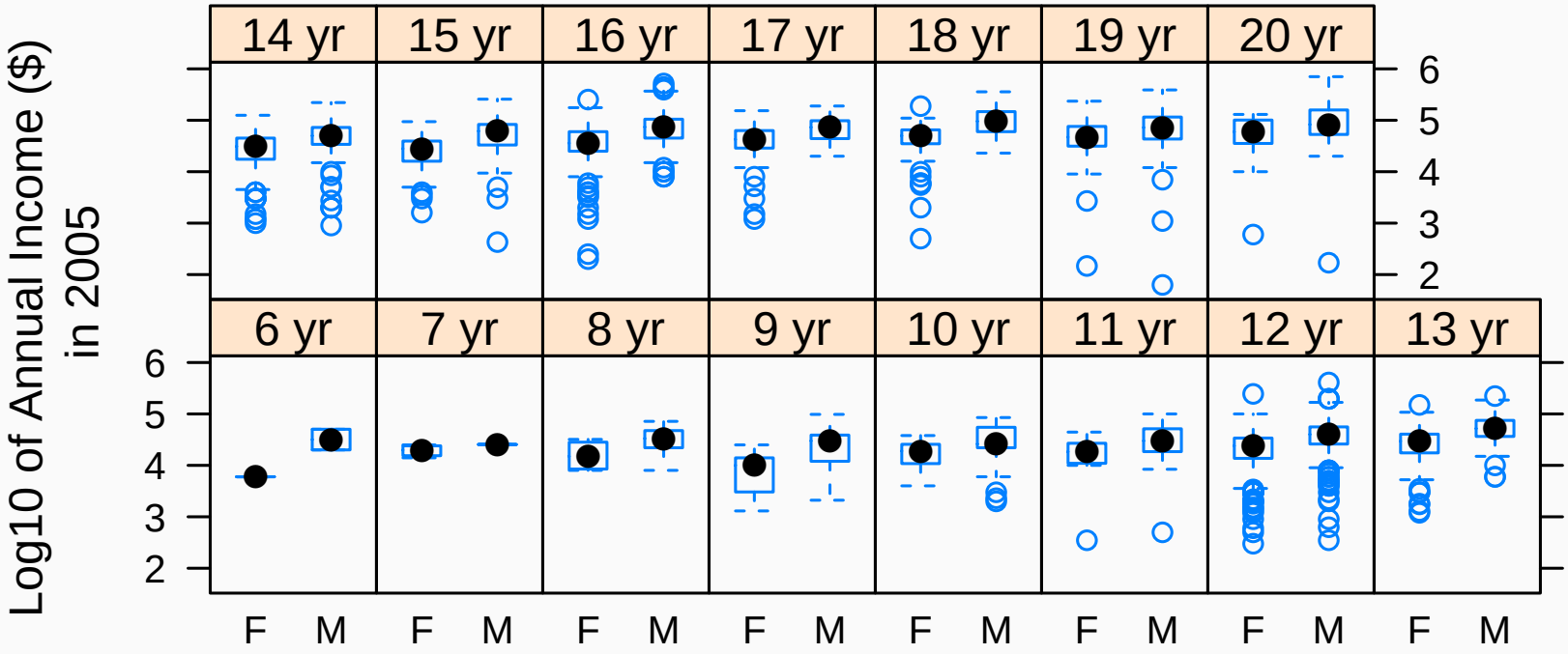



| Gender | min | Q1 | median | Q3 | max | mean | sd | n |
|---|---|---|---|---|---|---|---|---|
| female | 147 | 16000 | 29810.5 | 45000 | 253043 | 35210.68 | 28776.37 | 1278 |
| male | 63 | 32000 | 50000.0 | 78000 | 703637 | 63318.74 | 55861.07 | 1306 |

Do the data reflect males are paid more than females?

# Example: Gender Gap in Income, Adjusted for Education

- When talking about gender pay gap, one should compare the income of men and women with the same qualification.
- An indirect measurement of qualification is education level.
- One can compare the income of men and women with the same years of education.



# Example: Gender Gap in Income, Adjusted for Education

- Men earned more than women, even after adjusted for education level.
- The might be many confounding factors that can explain the gender gap in income, but the analysis in the previous slide rules out one of them — education level.
- There are methods that can rule out two, three or more confounding factors simultaneously. The more confounding factors we can rule out, the more convincing the conclusion.

# Example: Smoking and Longevity

A survey during 1972-74 recruited 1314 women in the United Kingdom and asked if they smoked. Twenty years later, a follow-up survey determined whether each woman was deceased or still alive.

| | Dead | Alive | Percentage Died |
|---|---|---|---|
| Smoker | 139 | 443 | $139/(139+443) \approx 23.9\%$ |
| Nonsmoker | 230 | 502 | $230/(230+502) \approx 31.4\%$ |

- Surprisingly, smokers had a higher survival rate than nonsmokers

# Example: Smoking and Longevity, Adjusted for Age

| Age in 1972 | Smoke? | Dead | Alive | Total | % Dead |
|---|---|---|---|---|---|
| 18-34 | Y | 5 | 174 | 179 | $5/179 \approx 2.8\%$ |
| | N | 6 | 213 | 219 | $6/219 \approx 2.7\%$ |
| 35-54 | Y | 41 | 198 | 239 | $41/239 \approx 17.2\%$ |
| | N | 19 | 180 | 199 | $19/199 \approx 9.5\%$ |
| 55-64 | Y | 51 | 64 | 115 | $51/115 \approx 44.3\%$ |
| | N | 40 | 81 | 121 | $40/121 \approx 33.1\%$ |
| 65+ | Y | 42 | 7 | 49 | $42/49 \approx 85.7\%$ |
| | N | 165 | 28 | 193 | $165/193 \approx 85.5\%$ |

In all age groups, smokers had a lower survival rate than nonsmokers.

How can smokers had a higher survival rate than nonsmokers when we combined all age groups?

# Simpson's Paradox

| Age in 1972 | Smoke? | Dead | Alive | Total | % Dead | % Smoke |
|---|---|---|---|---|---|---|
| 18-34 | Y | 5 | 174 | 179 | 2.8% | $\frac{179}{179+219} \approx 45.0\%$ |
| | N | 6 | 213 | 219 | 2.7% | |
| 38-54 | Y | 41 | 198 | 239 | 17.2% | $\frac{239}{239+199} \approx 54.6\%$ |
| | N | 19 | 180 | 199 | 9.5% | |
| 55-64 | Y | 51 | 64 | 115 | 44.3% | $\frac{115}{115+121} \approx 48.7\%$ |
| | N | 40 | 81 | 121 | 33.1% | |
| 65+ | Y | 42 | 7 | 49 | 85.7% | $\frac{49}{49+193} \approx 20.2\%$ |
| | N | 165 | 28 | 193 | 85.5% | |

- Old women (65+) were mostly dead 20 years later.
- There were fewer smokers among old women
- Smokers had a lower death rate because they were generally younger — age is a confounder here.
- *Simpson's paradox*: a trend observed in different groups of data can disappear or reverse when the groups are combined.

# Value of Observational Studies

- Carefully controlled and well-conducted observational studies can build up strong evidence for some "cause-and-effect" conclusions, though not conclusive.
- Observational studies have tremendous value as initial studies to build up support for larger, more resource-intensive controlled experiments
- Observational studies are a powerful and necessary tool
- However, they can be very *misleading* – identifying confounders is not always easy, and you can't control for everything
- Media often misinterpret association found in an scientific study as causation. Be cautious when reading news reports.

# Prospective vs. Retrospective Studies

- A *prospective* study identifies individuals and collects information as events unfold.
  - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- *Retrospective studies* collect data after events have taken place.
  - Example: Researchers reviewing past events in medical records.

# Sampling

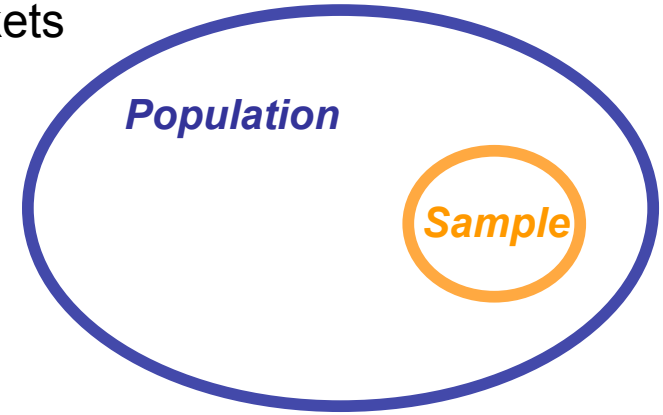
# Four Keywords in Sampling

- **Population**: The entire group of individuals in which we are interested but can't usually assess directly.

  Example: All humans, all working-age people in California, all crickets

- **Sample**: The part of the population we actually examine and for which we do have data.

  How well the sample represents the population depends on the sample design.



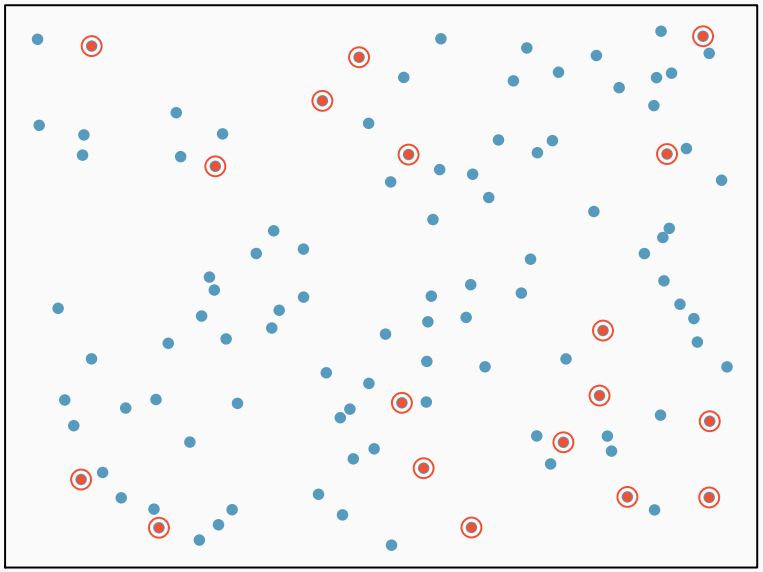

- A **parameter** is a number describing a characteristic of the population.

- A **statistic** is a number describing a characteristic of a sample.

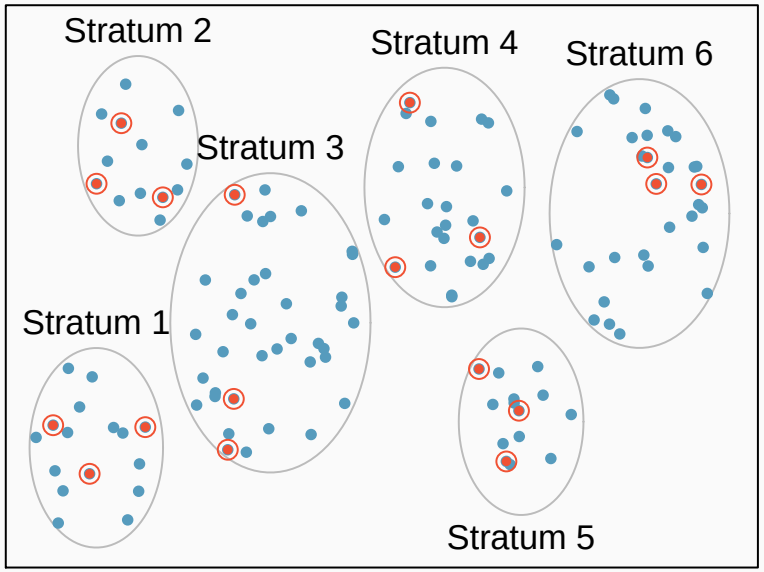
# Some Bad Sampling Methods

- *Convenience Sampling* — just sampling from those who are easily accessible
  - E.g. “Man on the street” survey (cheap, convenient, popular with TV “journalism”)
  - Problem: results may vary greatly with “when and where” the survey is done, lack of representation
- *Voluntary Response Sampling*
  - e.g., internet polls, call-in surveys
  - Only people visiting the website/watching the program will be sampled
  - People with strong opinions are more likely to participate

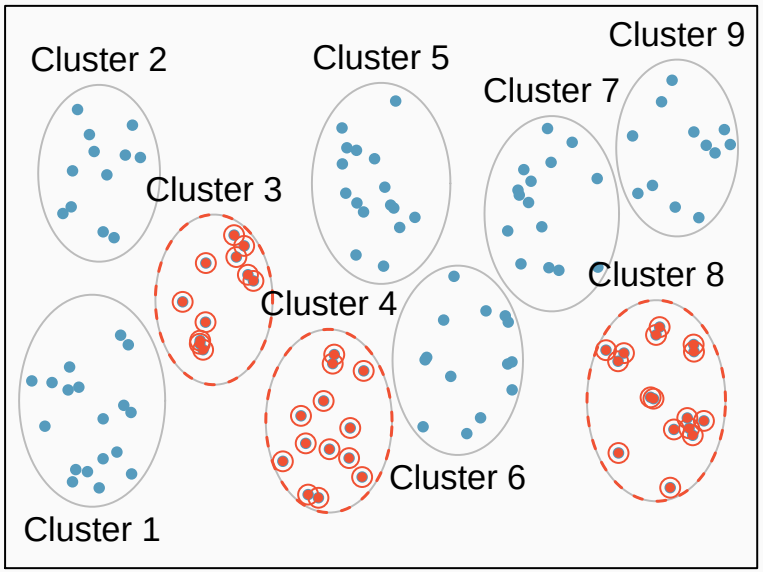
## Simple Random Sampling



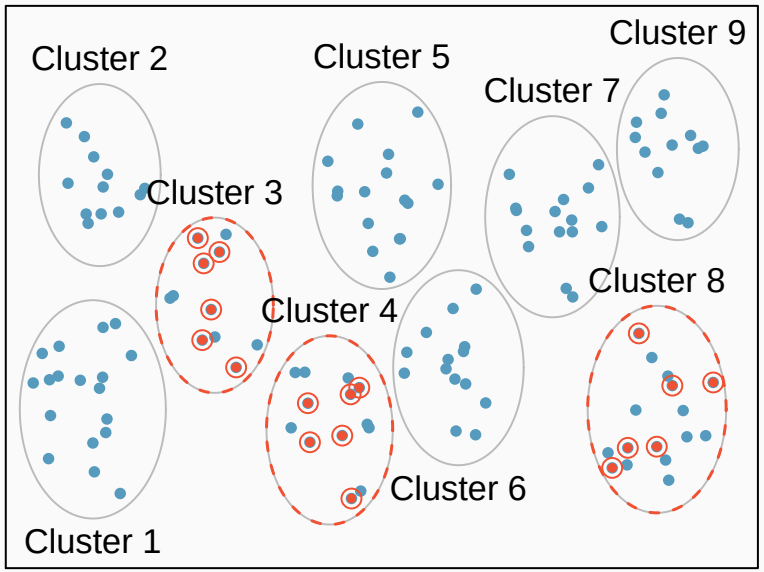
## Stratified Sampling



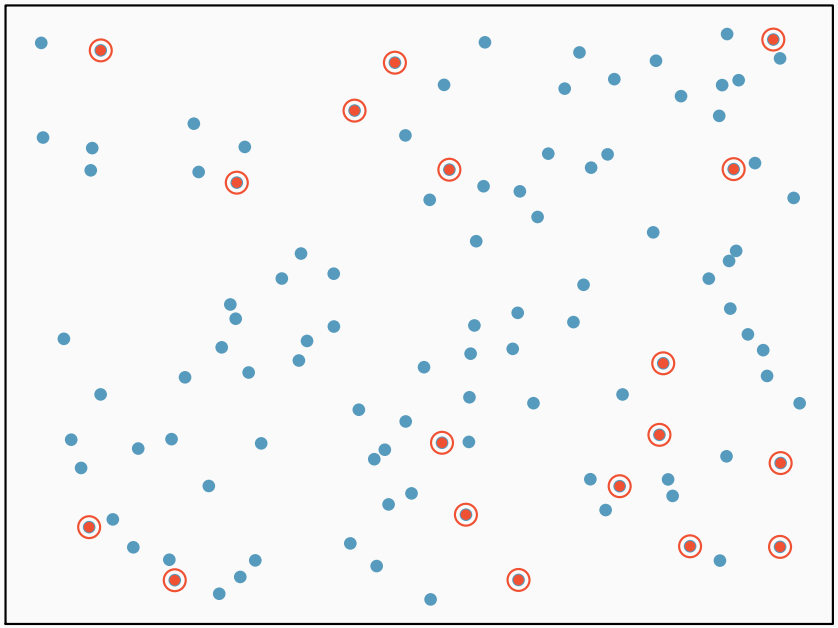

## Cluster Sampling




## Multistage Sampling



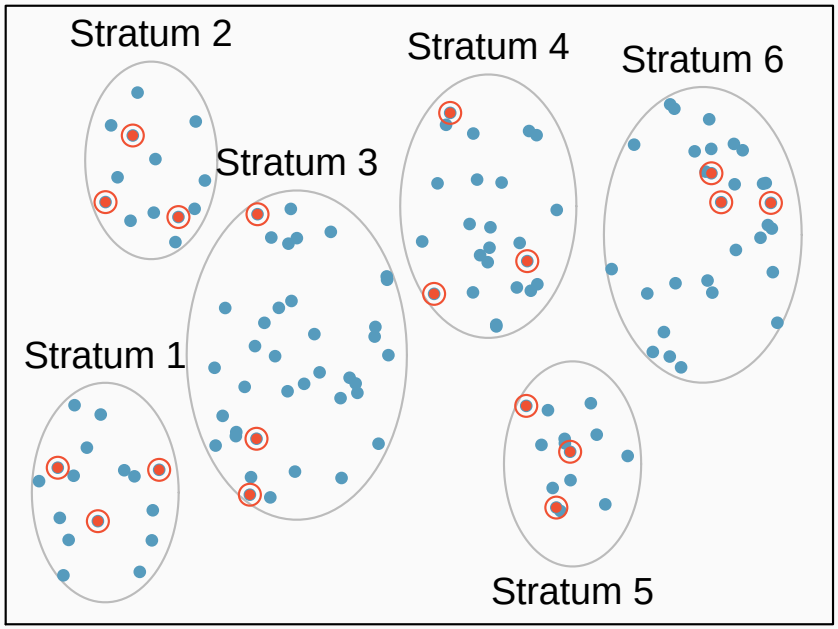
# Simple Random Sampling (SRS)



- Basic idea: put the names in a box, shake well, and make draws from the box.
- Need a list of names of all subjects in the population, called the *sampling frame*
- All subjects have the same chance to be chosen

- Pros: the makeup of the sample will mimic the makeup of the population (age/gender/race/income...), by the Law of Large Number.
- Cons: The need of a sampling frame makes it impractical for large populations

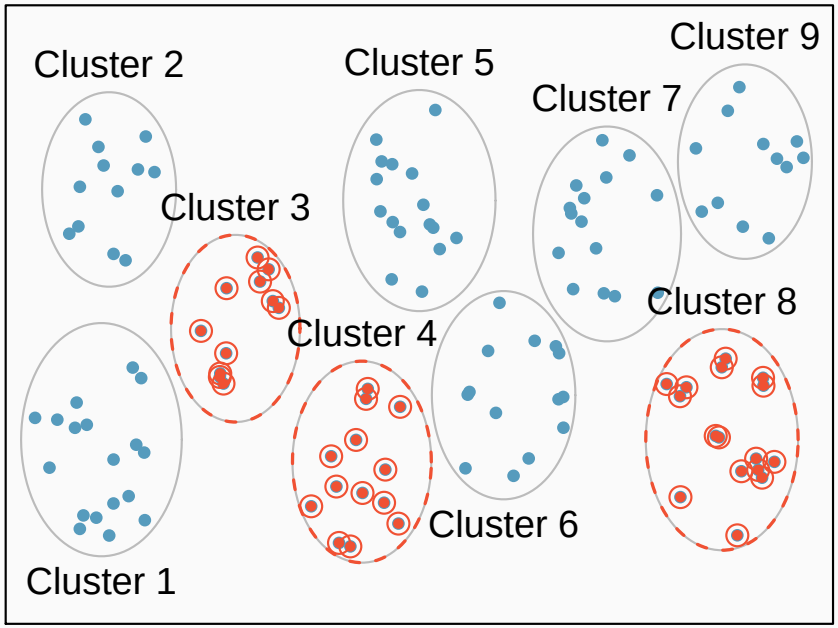
# Stratified Sampling




The population is divided into groups called **strata**, and then a separate simple random sample is chosen in each stratum.

- It works better when cases within a stratum are similar but there are large discrepancies between strata.
- Drawbacks: Need a sampling frame for each stratum — not practical for large populations.

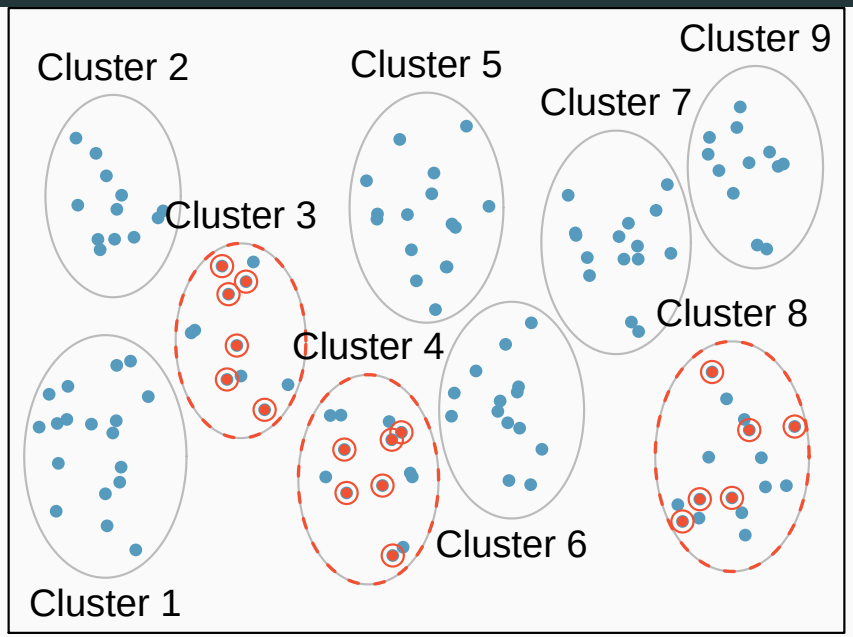
# Clustered Sampling




The population is divided into groups, called **clusters**. A sample of clusters is chosen. All subjects in the selected clusters are sampled.

- E.g., if Walmart wants to survey its employees, it may select a number of stores, and interview all employees in the selected stores. Here a cluster is a store.
- Cluster sampling works better when there is small cluster-to-cluster variation but large variation within clusters

# Multistage Sampling




- First stage: the population is divided into groups, called **clusters**, and a sample of groups is chosen.
- Second stage: the selected groups is further divided into subgroups, and a sample of subgroups is chosen in each selected group
- (Third stage: ...)
- (Fourth stage: ...)

# Multistage Sampling

Many nationwide surveys (like General Social Survey) use four-stage sampling.

- towns $\rightarrow$ wards $\rightarrow$ precincts $\rightarrow$ households

Advantage:

- selected subjects will all live in the selected towns, not scatter around nationwide, which can substantially lower the traveling cost of interviewers.
- no need to make sampling frame for unselected subgroups

# Problems in Sampling — Selection Bias

A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*.

- People with no permanent address are left out by mail survey
- About 1/3 of residential telephones are unlisted. Sampling phone numbers from white page would miss those unlisted numbers. Rich and poor are more likely to have unlisted numbers, so the telephone book tilts toward the middle class.
- Women are found to be more likely to pick up the phone than men. Telephone surveys often include more women than men.
- When a selection procedure is biased, taking a large sample does not help. This just repeats the same mistake on a larger scale.

# Example: The *Literary Digest* Poll

*Literary Digest*

- well-known magazine in U.S. from 1890 to 1936
- old issues at Regenstein
- had run presidential polls since 1920; always right
- bankrupt in 1938

The 1936 election

- 10 million postcard were sent (20% of voters in the country), about 2.4 million replied
- Names from phone lists, auto registrations, and club registers

| | FDR | Landon | Lemke | Sample Size |
|---|---|---|---|---|
| Literary Digest | 41% | 55% | 4% | 2.4 million |
| Gallup | 56% | ? | ? | 50,000 |
| Result | 61% | 37% | 2% | |

Why failed?

- Undercoverage: in 1936, poor people were less likely to have cars, phones or join clubs. They were under-represented
- Low response rate

# Problems in Sampling — Non-response bias

*Non-response bias* cause problems because non-respondents can be very different from respondents.

- Non-respondents may have long working hours, live alone, or not bother to respond, etc.
- E.g., in the past several decades, Gallup have found Republicans were more likely to respond than Democrats.
- When the response rate is low, one cannot take a new sample to replace those who don't respond.
- Cannot resolve by sampling more people to make up for the non-respondents, because the new sampled subjects are still the respondents, not the non-respondents.
- Must try to reach the non-respondents, by making more calls/visits, providing reward, etc.
- Always check the response rate. If low, the survey result might not be trust worthy.

# Problems in Sampling — Response bias

**Response bias** means the answers by respondents are influenced to some extent by the phrasing of the questions, and even the tone or attitude of the interviewer.

- Solution: interviewer control, proper design of questionnaires

# Can a Survey Result be Generalized to the Population?

Whether a *result is generalizable from a sample to the population* depends on whether the data came from a *random sample* from the population. Attempts to generalize to other populations may be biased, e.g.,

- For practical, ethical, and economic reasons, clinical trials usually involve only adults – children are excluded (only about 25% of drugs are subjected to pediatric studies)
- Physicians, however, are allowed to use any FDA-approved drug in any way that they think is beneficial, and aren't required to inform parents if the therapy hasn't been tested on children
- E.g., propofol is a sedative that has consistently proved safe in adults, was found causing higher death rates than other sedatives when applied on children.

# Recap: Random Assignment vs. Random Sampling

- Whether a *result is generalizable from data to a larger population* depends on whether the data came from a *random sample* from the population
- Whether a *cause-and-effect* relationship can be inferred depends on whether the subjects are *randomly assigned* to the treatments (and the control).