Sampling and Study Design

Ryan Miller

Outline

1. Samples and populations
2. Observational designs
3. Confounding variables
4. Randomized experiments

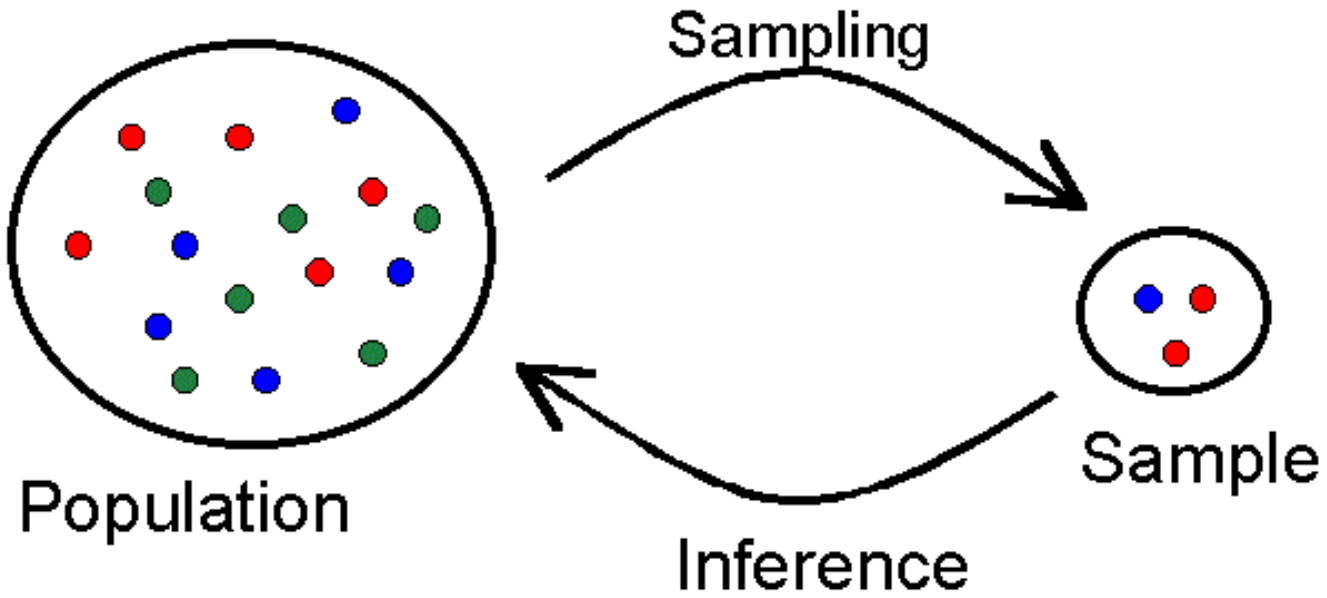
Introduction

Suppose a biologist wants to learn about the species of fish that reside within a particular lake

1) Do they need to capture and study *every* fish in this lake in order to achieve their goal?
2) What trade-offs are involved in collecting data on only *some* of the fish rather than *all* of them?

Sampling from a population

- Although desirable, seldom do we have data on *all* of the cases we are interested in
  - Instead, data is typically a **sample** of cases from a broader **population**



Note: We'll denote the number of cases in our sample as n and the size of the population as N (which is often unknown)

Sampling from a population

- **Inference** addresses the statistical question: "how reliably will trends in a sample reflect what is true of the population?"

Sampling from a population

- **Inference** addresses the statistical question: "how reliably will trends in a sample reflect what is true of the population?"
  - For example, if two variables, X and Y, have a correlation of $r = 0.71$ in a sample of size $n = 100$, how do you think these variables are related in the population?

Sampling from a population

- **Inference** addresses the statistical question: "how reliably will trends in a sample reflect what is true of the population?"
  - For example, if two variables, X and Y, have a correlation of $r = 0.71$ in a sample of size $n = 100$, how do you think these variables are related in the population?
- As a starting point, we might use the sample correlation, r, as an **estimate** of the population-level correlation, the **population parameter** denoted ρ
  - That is, you'd expect the correlation in the population to be *near* 0.71, assuming the sample data are *representative* of the cases in the population

Notation for estimates and population parameters

Statisticians use notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample):

	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

Two sources of sampling error

There are two main reasons why trends observed in the sample data might differ from those in the population:

1) **Sampling Bias** - a systematic flaw in the way cases were selected that leads to certain types of cases being disproportionately represented in the sample data

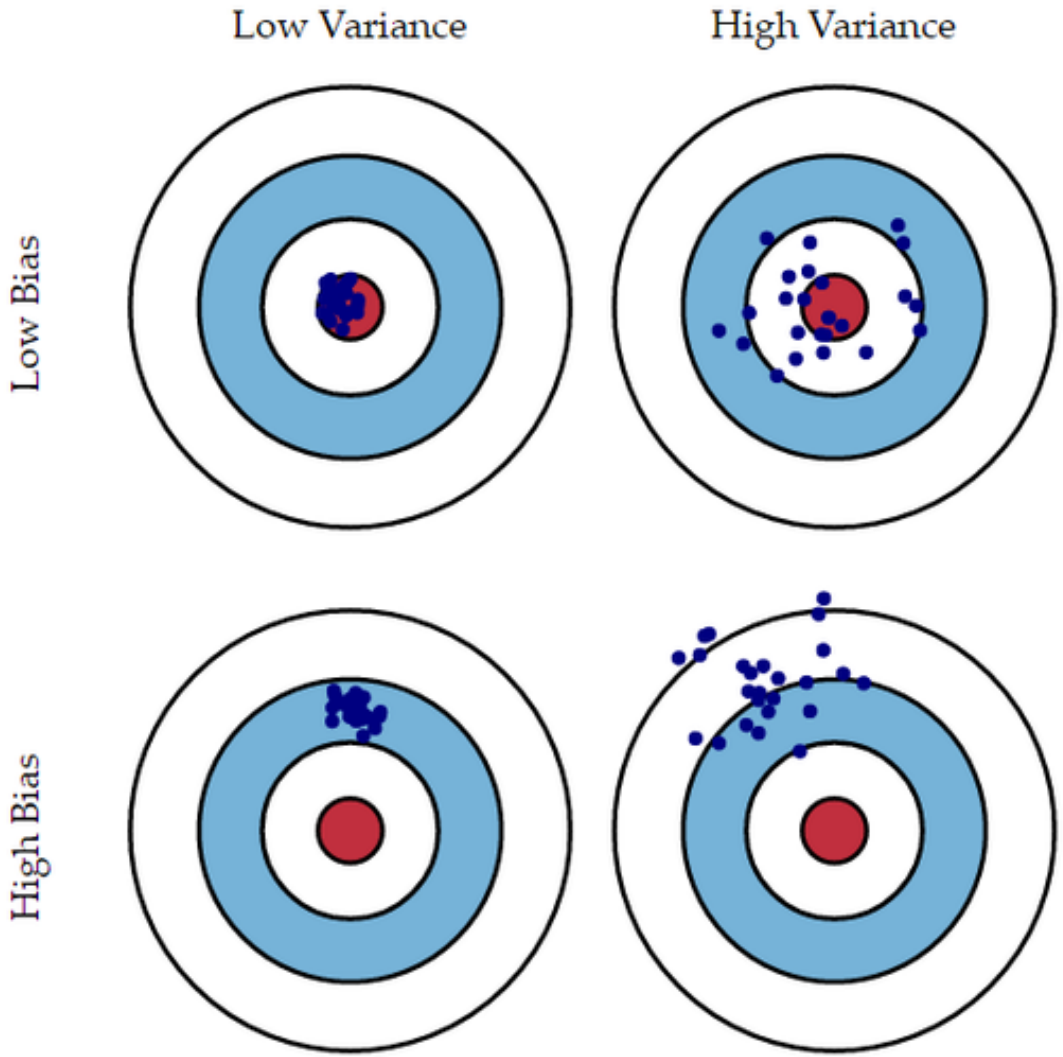
Two sources of sampling error

There are two main reasons why trends observed in the sample data might differ from those in the population:

1) **Sampling Bias** - a systematic flaw in the way cases were selected that leads to certain types of cases being disproportionately represented in the sample data
2) **Sampling Variability** - since a sample doesn't include all of the population, any individual sample might differ from the population due to *random chance* (ie: "the luck of the draw")

Sampling error

Four possible types of sampling procedures:



Remarks

- A larger sample size will *decrease* sampling variability, but it *will not* alleviate sampling bias

Remarks

- A larger sample size will *decrease* sampling variability, but it *will not* alleviate sampling bias
- Sampling procedures with high variance might seem problematic, but statisticians have developed tools (rooted in probability theory) to facilitate decision making in the face of this uncertainty

Practice

Shown below is the text of the Gettysburg Address, your goal is to accurately estimate the document's average word length:

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.

It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

To obtain an estimate, take your own sample of 5 words (trying to be *representative*).

Practice

Now, answer the following:

1) What is the *population* and the *sample*?
2) What is the *population parameter* and the corresponding *sample estimate*?
3) Which quadrant of the bias/variance matrix ("two sources of error" slide) do you think your sampling procedure belongs to?

Practice (solution)

1) The population is all words in the Gettysburg Address (with the individual words being cases within this population). The sample is the 5 words you selected.
2) The population parameter is the average word length for the full address. The sample estimate is the average of the 5 words you selected.
3) Your procedure was likely biased and also high variance.

Common sampling methods

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)

Common sampling methods

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)
- **Simple random sampling** - randomly select cases from the target population
  - Pros: eliminates sampling bias
  - Cons: can be difficult to execute

Common sampling methods

- **Convenience sampling** - select all cases from the target population that are easily accessible
  - Pros: data is easy to collect
  - Cons: high potential for sampling bias (though not guaranteed)
- **Simple random sampling** - randomly select cases from the target population
  - Pros: eliminates sampling bias
  - Cons: can be difficult to execute
- **Stratified or cluster random sampling** - randomly select cases separately from different population segments, potentially using different strategies for each segment
  - Pros: low potential for sampling bias, more flexibility than simple random sampling
  - Cons: data analysis becomes complicated (sampling weights, etc.)

Study design

Sampling, which focuses on how cases were obtained, is only one piece of the broader area of *study design*:

1) **Experimental study**, the explanatory variable is manipulated by the researcher
2) **Observational study**, explanatory and response variables are observed as they naturally occur

Study design

Sampling, which focuses on how cases were obtained, is only one piece of the broader area of *study design*:

1) **Experimental study**, the explanatory variable is manipulated by the researcher
2) **Observational study**, explanatory and response variables are observed as they naturally occur

- Consider research exploring the association between diet and health outcome
  - In an experimental study, participants might be assigned to specific diets
  - In an observational study, participants might be surveyed about their current diets

Two types of observational studies

- **Retrospective** observational studies will identify cases and collect data on things that have already happened
  - For example, surveying participants about their childhood experiences and current happiness

Two types of observational studies

- **Retrospective** observational studies will identify cases and collect data on things that have already happened
  - For example, surveying participants about their childhood experiences and current happiness
- **Prospective** (or cohort) observational studies will identify cases and then follow them forward in time until the outcome of interest is observed
  - For example, identifying a cohort of young people and following them to see who develops heart disease

Two types of observational studies

- **Retrospective** observational studies will identify cases and collect data on things that have already happened
  - For example, surveying participants about their childhood experiences and current happiness
- **Prospective** (or cohort) observational studies will identify cases and then follow them forward in time until the outcome of interest is observed
  - For example, identifying a cohort of young people and following them to see who develops heart disease
  - Prospective studies are generally considered to be *stronger evidence* than retrospective studies

Practice

In 1980, researchers at the University of Chicago collected data on all murders that took place during a felony in the state of Florida between 1972 and 1977. The researchers were interested in racial bias in the administration of the death penalty following the Civil Rights Act.

```
dp <- read.csv("https://remiller1450.github.io/data/DeathPenaltySentencing.csv")
```

1) What are the *sample* and the *population* in this study? What type of sampling procedure was used?
2) How would you describe the type of study design used by these researchers?
3) Using R, find the proportion of black offenders and the proportion of white offenders who were sentenced to death. Do you see evidence of racial bias?

Practice (solution)

1) The population are all individuals facing the death penalty. The sample consists of murder suspects from Florida between 1972 and 1977 who were facing the death penalty. This is a convenience sample.
2) This is a retrospective observational study, as the verdicts were already determined when the data were collected and the explanatory variable (offender's race) was not manipulated by the researchers.
3) 21.1% of black offenders and 23.2% of white offenders were sentenced to death, which does not seem to indicate any racial bias.

Race of victim

The results shown below **stratify** by race of the victim, does this added step change your conclusion?

```
dp_wv <- subset(dp, VictimRace == "white")
table(dp_wv$OffenderRace, dp_wv$DeathPenalty)
```

```
##
##          death not
##   black     37  41
##   white     46 144
```

```
dp_bv <- subset(dp, VictimRace != "white")
table(dp_bv$OffenderRace, dp_bv$DeathPenalty)
```

```
##
##          death not
##   black      1 101
##   white      0   8
```

Confounding variables

- A **confounding variable** is one that is associated with *both* the explanatory and the response variable in an analysis

Confounding variables

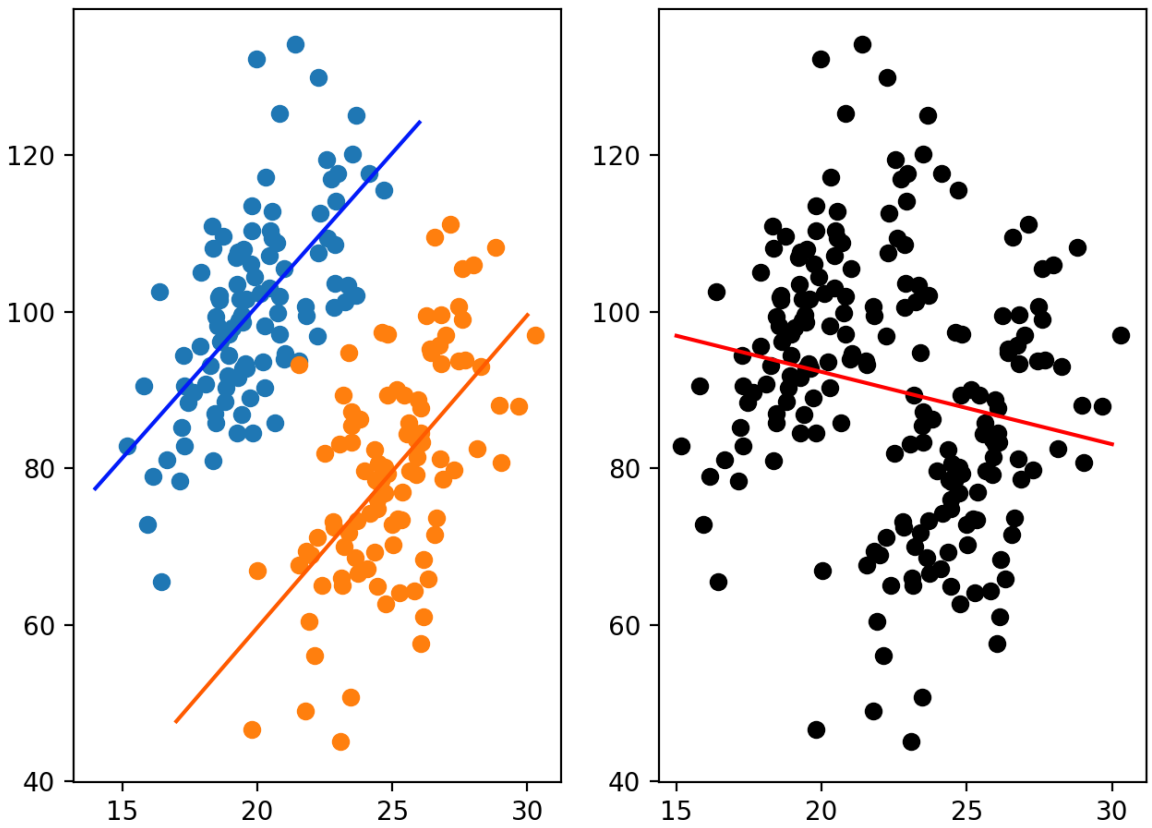
- A **confounding variable** is one that is associated with *both* the explanatory and the response variable in an analysis
  - A confounding variable will *obscure* the true association between explanatory and response variables
  - Stratification is one method for *controlling* for a confounding variable

Confounding variables

- A **confounding variable** is one that is associated with *both* the explanatory and the response variable in an analysis
  - A confounding variable will *obscure* the true association between explanatory and response variables
  - Stratification is one method for *controlling* for a confounding variable
- *All* confounding variables must be controlled for in order to make a reliable claim of **causation**
  - The Bradford Hill criteria provide a framework for determining causation using observational data

Simpson's paradox

Simpson's paradox occurs when the the impact of a confounding variable is so drastic that it *reserves* trend that was observed prior to stratification



Simpson's paradox vs. the ecological fallacy

The figure on the previous slide looks similar to one we've seen when discussing the ecological fallacy

- Simpson's paradox stems from ignoring an important confounding variable
- The ecological fallacy stems from inappropriately aggregating the data, which could involve a categorical grouping variable

Practice

In the early 1970s, administrators at UC-Berkeley grew concerned that females were being discriminated against in admissions to the university's graduate programs. In 1973, 44% of the 8442 male applicants were admitted, while only 35% of the 4321 female applicants were admitted.

Practice

In the early 1970s, administrators at UC-Berkeley grew concerned that females were being discriminated against in admissions to the university's graduate programs. In 1973, 44% of the 8442 male applicants were admitted, while only 35% of the 4321 female applicants were admitted.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	***825***	62%	108	**82%**
B	585	63%	***560***	63%	25	**68%**
C	918	35%	325	**37%**	***593***	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	***393***	24%
F	714	6%	373	6%	341	**7%**

1) Identify the explanatory, response, and confounding variable in the stratified table presented above
2) Explain *why* the overall admission rates are so different than those within each department

Practice (solution)

1) Sex is the explanatory variable, admission is the response variable, and program/department is the confounding variable.
2) Men more frequently applied to less competitive programs (such as A and B) that admitted both men and women at high rates, thereby "inflating" their overall admissions rate relative to that of women.

Balance

The underlying issue created by confounding variables can be described as **imbalanced groups**

- In the death penalty example, offenders were more likely to victimize their own race (with crimes against whites being punished more harshly)
  - This led to the groups of white offenders and black offenders being *systematically different* in an *important way* (victims race)

Balance

The underlying issue created by confounding variables can be described as **imbalanced groups**

- In the death penalty example, offenders were more likely to victimize their own race (with crimes against whites being punished more harshly)
  - This led to the groups of white offenders and black offenders being *systematically different* in an *important way* (victims race)
- In the UC-Berkeley example, men were more likely to apply to less competitive programs (which had higher overall admissions rates)
  - This led to the groups of male and female applicants being *systematically different* in an *important way* (department)

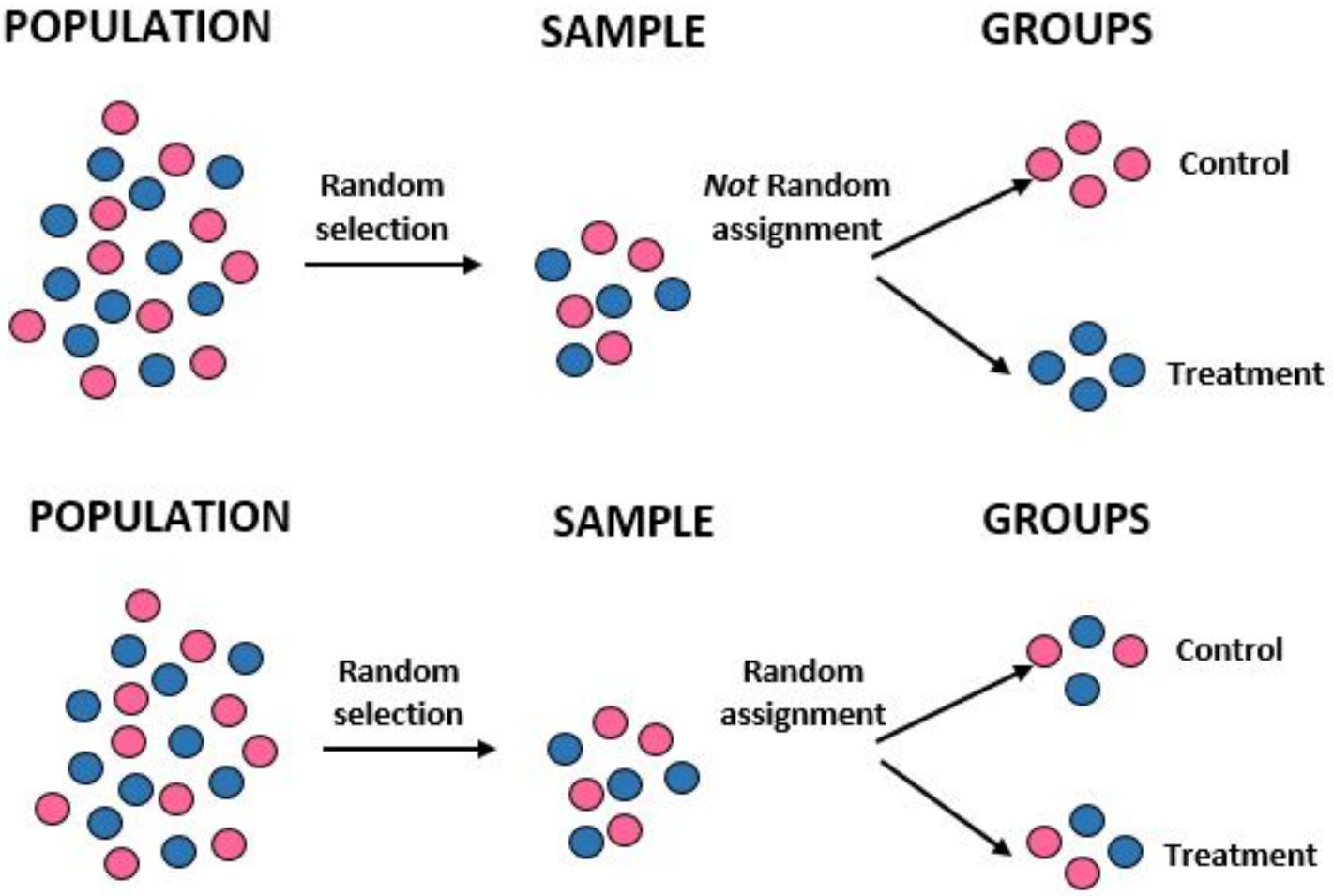
Obtaining "balanced" groups

- **Stratification** is a way of *forcing* balance
  - For example, when we compare male and female admissions within department A, these males and females are identical in terms of department

Obtaining "balanced" groups

- **Stratification** is a way of *forcing* balance
  - For example, when we compare male and female admissions within department A, these males and females are identical in terms of department
- **Random assignment** of the explanatory variable is another way to achieve balanced groups
  - On average, both groups (defined by the explanatory variable) will have the *same distribution* for *any* potential confounding variables
  - Random assignment is only possible in an experimental design

Random assignment



Practice

Suppose we want to know: "Is arthroscopic surgery is effective in treating arthritis of the knee?" Describe both an *observational study* and a *randomized experiment* that you could conduct to answer this question. Be sure to address the following during your discussion:

1. How costly will it be for the researchers to collect data with each design?
2. Are there any feasibility problems or ethical issues with each design?

Sham knee surgery

In the 1990s a study was conducted in 10 men with arthritic knees that were scheduled for surgery. They were all treated identically expect for one key distinction: only half of them actually got surgery! Once each subject was in the operating room and anesthetized, the surgeon looked at a randomly generated code indicating whether he should do the full surgery or just make three small incisions in the knee and stitch up the patient to leave a scar. All patients received the same post-operative care, rehabilitation, and were later evaluated by staff who didn't know whether they had actually received the surgery or not. The result? Both the sham knee surgery and the real knee surgery showed indistinguishable levels of improvement

Source: https://www.nytimes.com/2000/01/09/magazine/the-placebo-prescription.html

Vocabulary for randomized experiments

The Sham Knee Surgery example illustrates several important aspects of a well-designed experiment that we've yet to discuss:

- **Control Group** - Some patients were randomly assigned not to receive the knee surgery, providing a comparison group that is, on average, balanced with surgery group in all baseline characteristics
- **Placebo** - Patients in the control group received a fake surgery
- **Blinding** - Using a placebo is not helpful if patients know which group they're in. Similarly, the staff interacting with the patients might treat them differently if they knew the patient's group
  - **Single-blind** - the participants don't know the treatment assignments
  - **Double-blind** - the participants *and* everyone interacting with the participants don't know the treatment assignments

A few other sources of bias

1. **Non-response Bias** - Subjects who are recruited but decline to participate in a study differ in important ways from those who do participate or respond
2. **Non-ignorable Missing Data** - Subjects who are excluded from analysis due to missing data differ in important ways from those with complete data
3. **Social Desirability Bias** - Respondents tend to answer questions in ways that portray themselves in a positive light - Link
4. **Interviewer Bias** - The interviewer causes subjects to the behave differently than they otherwise would

Study design and statistical inference

- The overarching goal of a statistician is to *rule out* possible explanations for an observed association

Study design and statistical inference

- The overarching goal of a statistician is to *rule out* possible explanations for an observed association
- The following explanations can be addressed by study design:
  - **Sampling bias** - Simple random sampling
  - **Confounding variables** - Random assignment, stratification
  - **Other sources of bias** - Placebo, double-blinding, etc.

Study design and statistical inference

- The overarching goal of a statistician is to *rule out* possible explanations for an observed association
- The following explanations can be addressed by study design:
  - **Sampling bias** - Simple random sampling
  - **Confounding variables** - Random assignment, stratification
  - **Other sources of bias** - Placebo, double-blinding, etc.
- Ideally, we can use careful study design to reduce the viable explanations to either **random chance** or a **real relationship**

Summary

1. Samples and populations
   - Sampling bias and sampling variable are two reasons for trends observed in sample data not reflecting the truth about a population
2. Observational designs
   - Study design is an additional cause for concern when trying to infer causation
3. Confounding variables
   - In observational designs, confounding variables can obscure the underlying relationship between explanatory and response variables
4. Randomized experiments
   - Random assignment of the explanatory variable eliminates the possibility confounding variables
   - Other measures should be taken to prevent bias (ie: placebo, blinding, etc.)