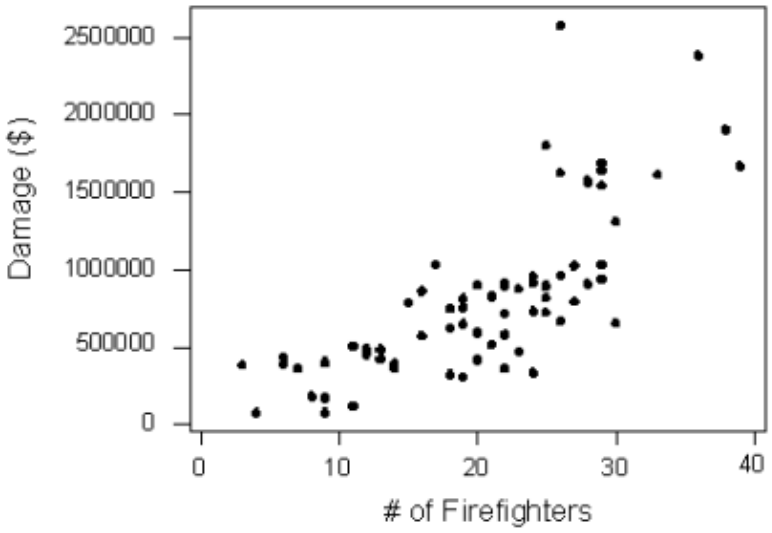
Lecture 5 –EDA Fundamentals Continued

Before moving on to how we gather/produce data, let's take a look at a special example…

Example 1: Fire Damage

The scatterplot below illustrates how the damage caused by a fire (Y) is related to the number of firefighters sent to the fire (X):



Lurking Variable →

a.

b.

What is one possible lurking variable in this case? What could happen if we included a lurking variable in our model?

Ex: Hospital Death Rates

Background: A government study collected data on the death rates in nearly 6000 hospitals in the US. These results were then challenged by researchers who said that the federal analyses failed to take into account the variation among hospitals in the severity of patients' illness when they were hospitalized. As a result, said the researchers, some hospitals were treated unfairly in the findings, which named hospitals with higher-than-expected death rates. What the researchers are saying is that when the federal study explored the relationship between the two variables: hospital and death rate, it should have also included in the study (or taken into account) the lurking variable – severity of illness.

Consider the following contingency table summarizing the data about the status of patients who were admitted to 2 hospitals in a certain city (Hospital A and Hospital B). Hospital is the explanatory, and the Patient's Status is the response.

Untitled

Hospital \ Patient's status	Died	Survived	Total
Hospital A	63	2037	2100
Hospital B	16	784	800
Total	79	2821	2900

Hospital \ Patient's status	Died	Survived	Total
Hospital A	3%	97%	100%
Hospital B	2%	98%	100%

The death rate was higher for Hospital A than for Hospital B.

Can we jump to the conclusion that this is true?

What could be a lurking variable in this situation?

Let's see what happens when we consider patient severity:

Hospital \ Patient's status	Died	Survived	Total
Hospital A	63	2037	**2100**
Hospital B	16	784	**800**
Total	79	2821	2900

Accounting for the lurking variable: "severity of illness"

Hospital \ Patient's status	Died	Survived	Total
Patients severely ill			
Hospital A	57	1443	**1500**
Hospital B	8	192	**200**
Total	65	1635	1700

Hospital \ Patient's status	Died	Survived	Total
Patients *not* severely ill			
Hospital A	6	594	**600**
Hospital B	8	592	**600**
Total	14	1186	1200

So, if we supplement our visual finding with the proper numerical summary, we will also include conditional percentages for both patients that are severely ill and those who are not:

Hospital \ Patient's status	Died	Survived	Total
Patients severely ill			
Hospital A	**3.8%**	96.2%	100%
Hospital B	**4.0%**	96.0%	100%

Hospital \ Patient's status	Died	Survived	Total
Patients *not* severely ill			
Hospital A	**1.0%**	99.0%	100%
Hospital B	**1.3%**	99.7%	100%

What do we notice?

a.

b.

This phenomenon is called **Simpson's (or Reversal) Paradox** (the direction of an association between 2 variables can change after including a third variable and analyzing the data at separate levels of that variable)

The point:

So, how can we conclude causation?

Some notes on Gathering/Producing Data

I. Sampling →

a. Terminology:

i. **Unbiased sampling method** → a sampling method that chooses a sample that is representative of the population

ii. **Biased Sampling method** → a sampling method that chooses a sample with one or more types of bias

iii. **Bias** → systematic tendency toward certain outcomes different from what would be observed in the population

iv. **Sampling bias** → bias due to the sample methodology

v. **Sampling frame** → the list of subjects in the population from which the sample is taken

Ex: In order to measure political inclination at the university, UConn researchers took a random sample of 500 from those students that are seniors.

What is the population?

What is the sampling frame?

What is the sample?

b. In order to obtain a sample representative of the population, the ***selection must be*** _____ (each person in the population must have an equal chance of being selected for the sample)

c. Different selection methods (through use of an example using the classroom as the population):

i. Simple Random Sampling –

ii. Stratified Sampling –

iii. Cluster Sampling –

iv. Multistage Sampling –

Example 1: Use a cluster sample of UConn by college, and then a stratified sample within each college where the stratum was gender

Example 2: To obtain a random sample of data scientists in the US, a researcher chooses 15 states at random (stage 1 = cluster). From each state, you choose at random 5 companies (stage 2 = cluster). Then, from each company, you randomly choose 3 data scientists (stage 3 = stratified).

d. **Under-coverage** → even though our sample is chosen randomly, it is biased because the sampling frame is not very close to the population (Ex: gender)

e. Biased Sampling Methods →

i. **Convenience Sampling** → selects whichever units of the population are easiest to reach (Ex: polling at the mall)

ii. **Voluntary Response Sampling** → consists of people who choose themselves by selecting to response to the general appeal (Ex: write-in, call-in, or online opinion polls)

→

f. **Response Bias** →

i. **Sensitive Questions**

1. Have you used illegal drugs?

2. Have you ever had same-sex sex?

→ Solution: Randomized Response

ii. **Wording of Questions (which would you choose?)**

1. *Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?*

2. *Should laws be passed to prohibit interest groups from contributing to campaigns or do groups have the right to contribute to the candidate they support?*

→ Ross Perot (3rd party candidate in 1992 presidential election) asked the first question.

a. 99% said yes by mail-in (useless)

b. then a survey firm asked the same question to a nationwide random sample (80% yes)

c. then, the firm asked the second question in another nationwide random sample and got 40% yes after realizing that the first question demanded a yes answer.

iii. **Order of Questions** (makes people think about something they wouldn't have otherwise considered, impacting their response to the next question → ***order is important***)

Consider the following questions:

1. To what extent do you think teenagers today worry about peer pressure related to drinking alcohol?

2. Name the top 5 pressures you think teenagers face today

What do you think will likely end up in the second response for almost every single person surveyed??

II. Study Design

a. **Experimental Study (Controlled Experiment) →**

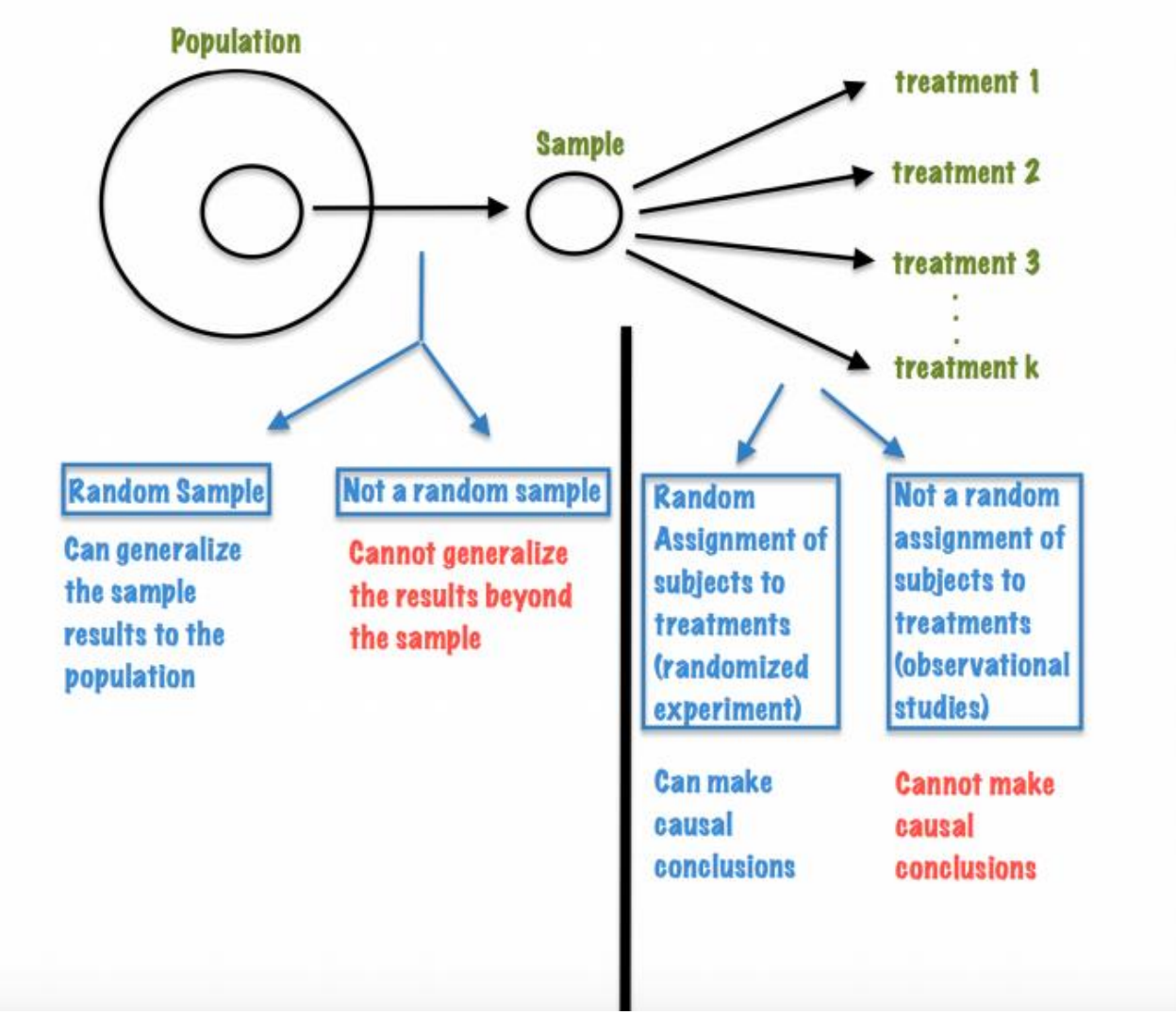
Example – drug effect on blood pressure

- ***i.*** Randomly assigning subjects ensures different treatment groups and the control group are as similar as possible
- ***ii.*** Randomized experimental study allows for causal conclusions since all lurking variables are ruled out/accounted for
- ***iii.*** **Blind Experimental Study →**
- ***iv.*** **Placebo Effect →**
- ***v.*** **Double blind →**

b. **Observational Study →**

- ***i.*** Can never conclude causation from an observational study!

To summarize (general structure of a study in which you compare several treatments):



EDA ENDS