Elementary Statistics – Sampling and Random Experiments

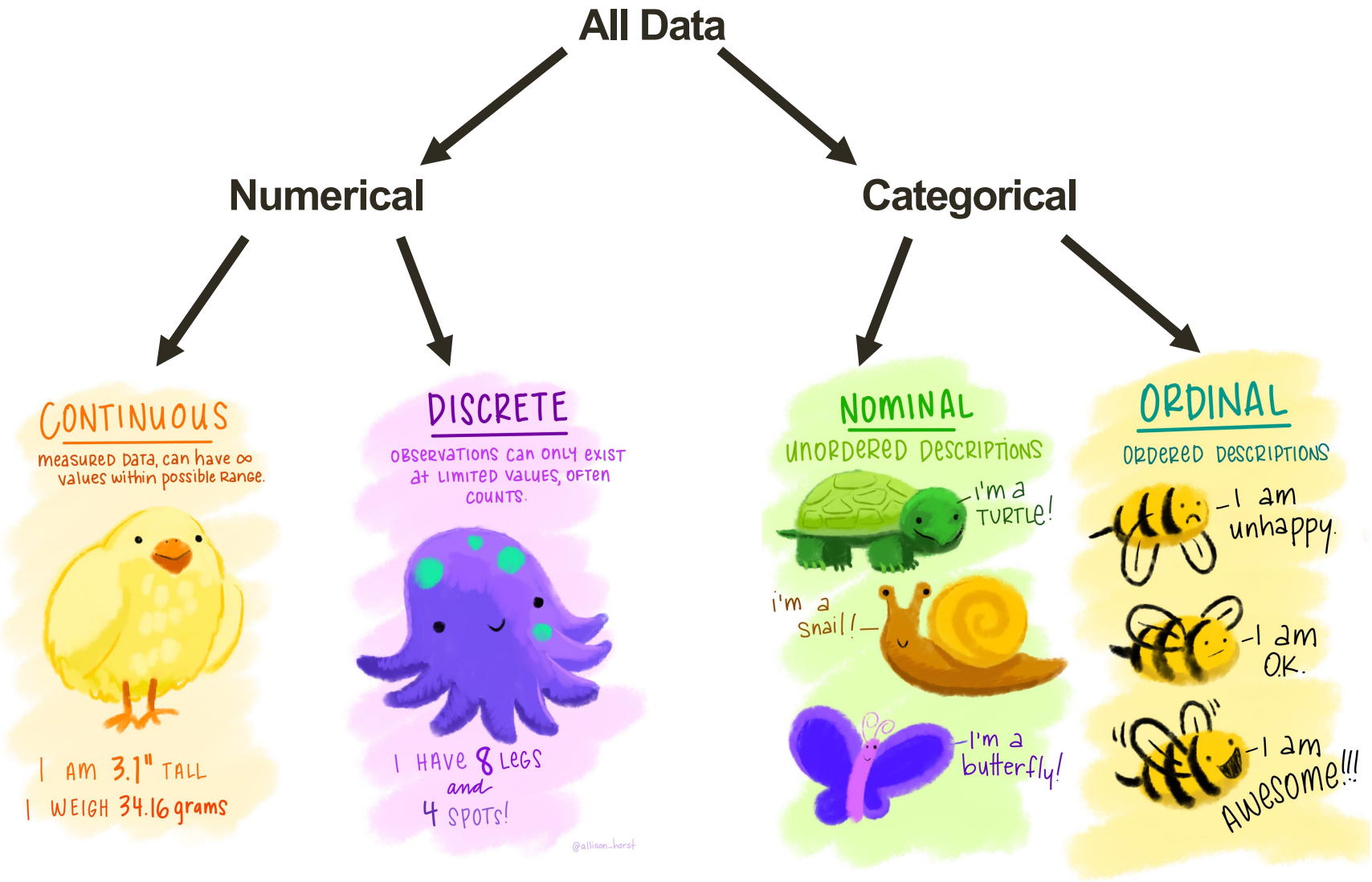
Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Wrap up Data
- Sampling
  - Why
  - Techniques
- Randomized Experiments
  - Causality
  - Confounding

Data: Variable Taxonomy



Exercise

Data Collection:

Form a group of three and take turns answering the following questions:

What is your name?

What is your class year?

What is your hometown?

Do you have siblings (1 = Yes, 0 = No)?

What is the furthest away from Westfield (in miles) that you were over winter break?

While one group member answers, the other two will write down the answers. Then open the data collection Google spreadsheet (http://bit.ly/48ceWSa) and start entering the data!

Exercise

Identifying Variable Types:

Take a couple minutes on your own to answer the following:

- What is the observational unit in our dataset?
- For each variable in the spreadsheet, is it discrete, continuous, ordinal, or nominal?

Exercise

Identifying Variable Types:

Take a couple minutes on your own to answer the following:

- What is the observational unit in our dataset?
- For each variable in the spreadsheet, is it discrete, continuous, ordinal, or nominal?

Discuss your answers with your neighbor!

Exercise

What type of variable is "Class Year"?

- Discrete
- Continuous
- Ordinal
- Nominal

Exercise

What type of variable is “Siblings”?

- Discrete
- Continuous
- Ordinal
- Nominal

Exercise

What type of variable is "Distance from Westfield over Break" (Miles-Away)?

- Discrete
- Continuous
- Ordinal
- Nominal

Recap: Big Picture

Although our statistical questions are framed in terms of populations and parameters, what we have at our disposal to *answer* those questions is often only a sample

Population		Sample (a.k.a Data)
all likely US voters	→	2000 individuals in a snap political poll
all French paintings from the 1800s	→	French paintings from the 1800s in the Louvre
humanity	→	43,448 individuals in a COVID vaccine trial

Recap: Big Picture

Although our statistical questions are framed in terms of populations and parameters, what we have at our disposal to *answer* those questions is often only a sample

Population		Sample (a.k.a Data)
all likely US voters	→	2000 individuals in a snap political poll
all French paintings from the 1800s	→	French paintings from the 1800s in the Louvre
humanity	→	43,448 individuals in a COVID vaccine trial

Sometimes we actually are able to collect data on our entire population: this is called a census! In that case, we can calculate the value of the parameter directly. . .

. . . But this happens only rarely and with known issues.

Problems with a Census

Consider the US census, which is supposed to capture data on *every person* in the US.

Do you know of any issues with the US census?
Brainstorm with a buddy.

Sometimes we actually are able to collect data on our entire population: this is called a census! In that case, we can calculate the value of the parameter directly. . .

. . . But this happens only rarely and with known issues.

Problems with a Census

Consider the US census, which is supposed to capture data on *every person* in the US.

Issues with the US census:

- Does not reach all subpopulations equally
  - undercounts Hispanic, Black, and Native American residents
  - overcounts white and Asian American residents
  - undocumented residents
  - undercounts residents living in poor urban areas
  - undercounts unhoused people
- Does not collect data in a representative way
  - sex (only binary options, no intersex option)
  - gender is not collected
  - race includes no multiracial option

Problems with a Census

So while a census is nice in theory…

- It can be difficult to complete. There always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*

- Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.

Approaches to Data Collection

Anecdotal evidence: evidence based on personal observation, collected in a non-systematic manner and often involving only a few cases

- Ex: asking a few of your friends for their opinions

Convenience sample: individuals who are easier to contact or reach are more likely to be included in the sample

- Ex: conducting a school-wide survey by stopping WSU students as they exit the Dining Commons

Random (probability) sample: every individual in the population has a known, non-zero probability of being included in the sample and is randomly selected according to this probability

- Ex: conducting that same survey by randomly selecting names from the school directory

Approaches to Data Collection

Anecdotal evidence: evidence based on personal observation, collected in a non-systematic manner and often involving only a few cases

- Ex: asking a few of your friends for their opinions

Convenience sample: individuals who are easier to contact or reach are more likely to be included in the sample

- Ex: conducting a school-wide survey by stopping WSU students as they exit the Dining Commons

Random (probability) sample: every individual in the population has a known, non-zero probability of being included in the sample and is randomly selected according to this probability

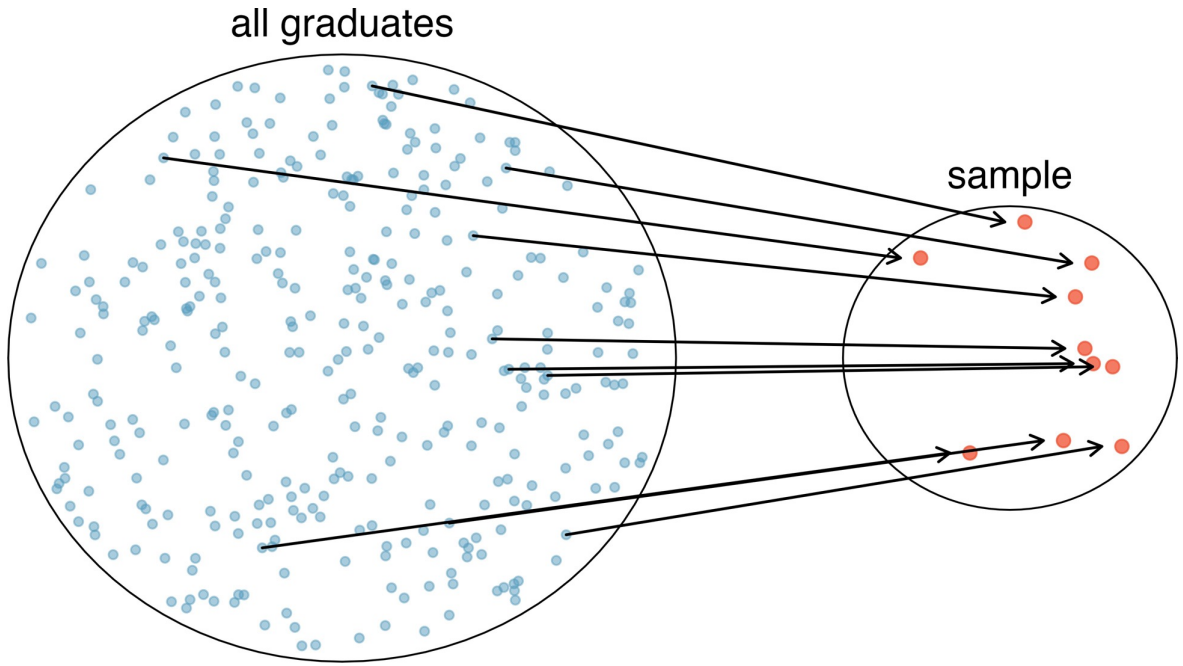
- Ex: conducting that same survey by randomly selecting names from the school directory

Which approach will produce the best data and insight into the most popular majors at WSU? Why?

Simple Random Sampling

Analogous to using a raffle to select observations

- All members in the population have an equal chance of being included in the final sample
- No implied connection between the observations in our sample

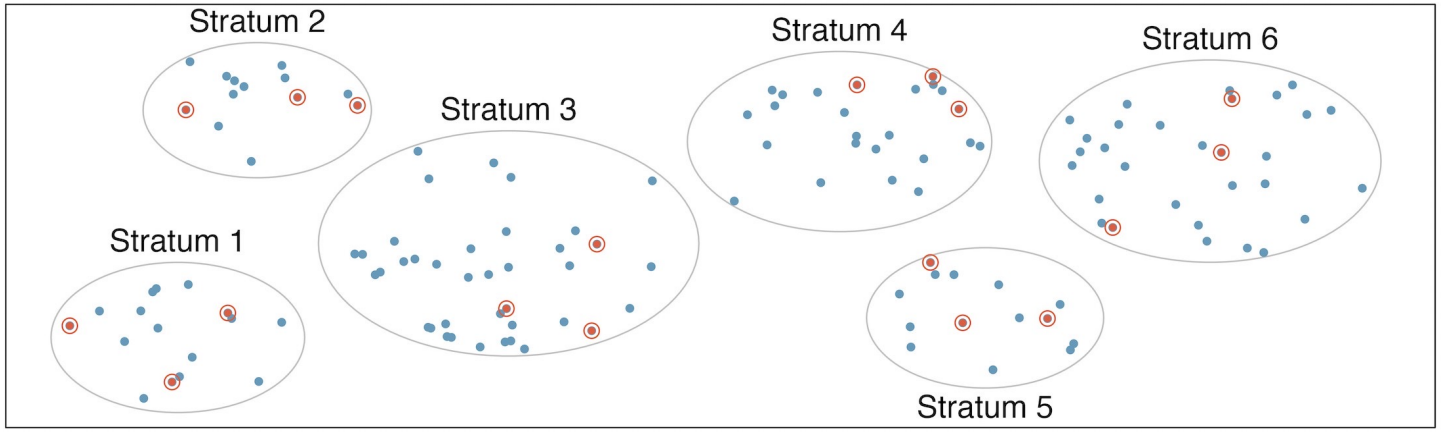


Stratified Random Sampling

Divide-and-conquer sampling strategy that:

- Divides the population into “strata” that group similar observations together
- Conducts a simple random sample within each stratum

Helps our sample to be more balanced with respect to the grouping/ stratification variable

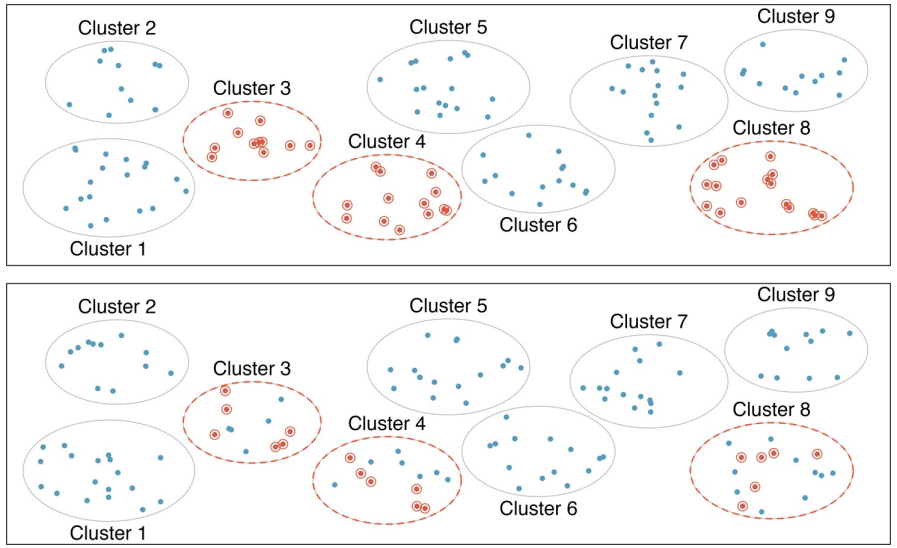


Clustered and Multistage Sampling

Our population may naturally be grouped into clusters (e.g., families, schools, communities)

- Cluster sampling: Randomly select clusters and include all members of those clusters
- Multistage sampling: Randomly select clusters and include a simple random sample of the cluster members

Often pursued for logistical convenience or economic reasons



What are the pitfalls of non-representative sampling?

Sampling bias refers to systematic error due to non-random sampling, in which some members of the target population are much more likely to be included in the sample than others *in ways that we did not intend*

What are the pitfalls of non-representative sampling?

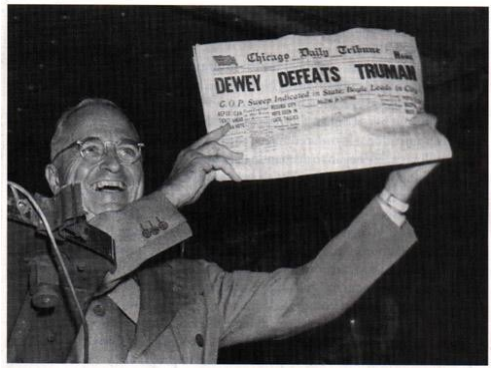
Sampling bias refers to systematic error due to non-random sampling, in which some members of the target population are much more likely to be included in the sample than others *in ways that we did not intend*

1948 Presidential Election

On the night of the election, the Chicago Tribune mistakenly printed a headline declaring that Dewey had defeated Truman

Conclusion was based on the results of a telephone survey

- Individuals with telephones tended to be wealthier and have stable home addresses



What are the pitfalls of non-representative sampling?

Sampling bias refers to systematic error due to non-random sampling, in which some members of the target population are much more likely to be included in the sample than others *in ways that we did not intend*

Visual Bias in Art Museums

Art mediums, political agendas, and both artist and art curator identities inform what art is deemed "worthy" of preserving and displaying

- Museum collections still vastly overrepresent artists who are white, Western, and male

The Horse Fair, Rosa Bonheur, 1852–55.

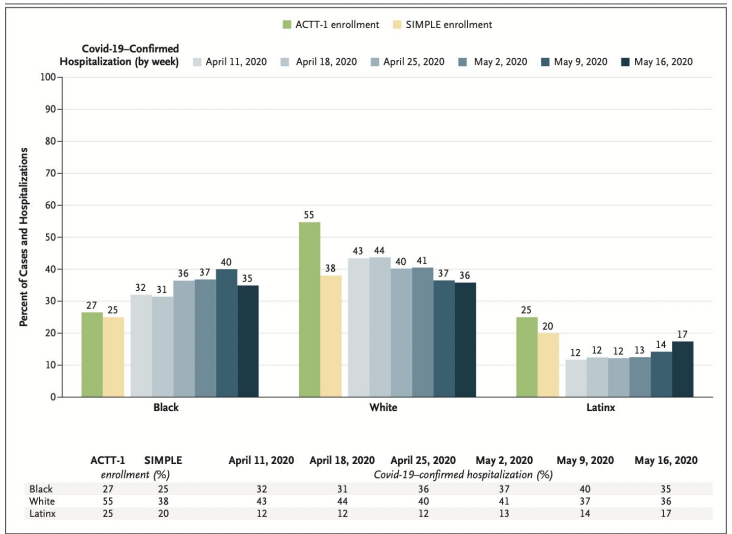
What are the pitfalls of non-representative sampling?

Sampling bias refers to systematic error due to non-random sampling, in which some members of the target population are much more likely to be included in the sample than others *in ways that we did not intend*

Representation in COVID-19 Studies

Historic mistreatment of people of color as well as ongoing racism and discrimination have bred distrust of the medical field, particularly within racially minoritized communities

Other barriers to participation: language barriers, health literacy, time/cost



	ACTT-1 enrollment (%)	SIMPLE enrollment (%)	April 11, 2020 Covid-19–confirmed hospitalization (%)	April 18, 2020 Covid-19–confirmed hospitalization (%)	April 25, 2020 Covid-19–confirmed hospitalization (%)	May 2, 2020 Covid-19–confirmed hospitalization (%)	May 9, 2020 Covid-19–confirmed hospitalization (%)	May 16, 2020 Covid-19–confirmed hospitalization (%)
Black	27	25	32	31	36	37	40	35
White	55	38	43	44	40	41	37	36
Latinx	25	20	12	12	12	13	14	17

Figure 1. Covid-19–Associated Hospitalizations According to Race and Ethnicity throughout the United States as Compared with Enrollment in the ACTT-1 and SIMPLE Trials.

Data are from the CDC COVIDView[2] weekly surveillance reports of laboratory-confirmed Covid-19–associated hospitalizations with available information on race and ethnicity through COVID-NET (blue bars) for a time frame comparable to the enrollment periods for ACTT-1 (green bars) and the SIMPLE trials (yellow bars).

From the *New England Journal of Medicine*.

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures.

Form 8 groups, each will be assigned a scenario. For your scenario, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context. *Be prepared to share with the class!*

a) Randomly sample 200 households from the city.
b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.
c) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.
d) Sample the 200 households closest to the city council offices.

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures.

a) Randomly sample 200 households from the city.

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures.

b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures.

c) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures.

d) Sample the 200 households closest to the city council offices.

Big Picture

Many research questions are, fundamentally, questions about the relationship between two or more variables of interest:

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D., Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D., Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D., Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

CONCLUSIONS

A two-dose regimen of BNT162b2 conferred 95% protection against Covid-19 in persons 16 years of age or older. Safety over a median of 2 months was similar to that of other viral vaccines. (Funded by BioNTech and Pfizer; ClinicalTrials.gov number, NCT04368728.)

Big Picture

Many research questions are, fundamentally, questions about the relationship between two or more variables of interest:

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014

The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

Big Picture

Many research questions are, fundamentally, questions about the relationship between two or more variables of interest:

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014 The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

How can we design a research study to answer these sorts of questions? And what sorts of conclusions can we draw about these relationships?

Types of Study Designs

We often label these variables according to their hypothesized relationship with one another. . .

Response variable: the measured outcome of interest

Explanatory variable: a variable that potentially explains or predicts changes in the response

. . . and broadly classify study designs into one of two categories:

Observational studies: researchers observe both the explanatory and response variables without interfering in how the data arise

Experiments: researchers intervene and assign treatments (the explanatory variable) to each participant in the study

→ if this assignment involves randomization, then we call the study a *randomized experiment*

Types of Study Designs

Handwriting and SAT Scores

An article about handwriting appeared in the October 11, 2006 issue of the Washington Post. The article mentioned that among a sample of students who took the essay portion of the SAT exam in 2005-2006, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters.

Types of Study Designs

Handwriting and SAT Scores

An article about handwriting appeared in the October 11, 2006 issue of the Washington Post. The article mentioned that among a sample of students who took the essay portion of the SAT exam in 2005-2006, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters.

What are the observational units, variables, types of variables, parameter of interest, and statistic in this study?

Types of Study Designs

Handwriting and SAT Scores

An article about handwriting appeared in the October 11, 2006 issue of the Washington Post. The article mentioned that among a sample of students who took the essay portion of the SAT exam in 2005-2006, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters.

What are the observational units, variables, types of variables, parameter of interest, and statistic in this study?

What type of study is this?

Types of Study Designs

Handwriting and SAT Scores

An article about handwriting appeared in the October 11, 2006 issue of the Washington Post. The article mentioned that among a sample of students who took the essay portion of the SAT exam in 2005-2006, those who wrote in cursive style scored significantly higher on the essay, on average, than students who used printed block letters.

What are the observational units, variables, types of variables, parameter of interest, and statistic in this study?

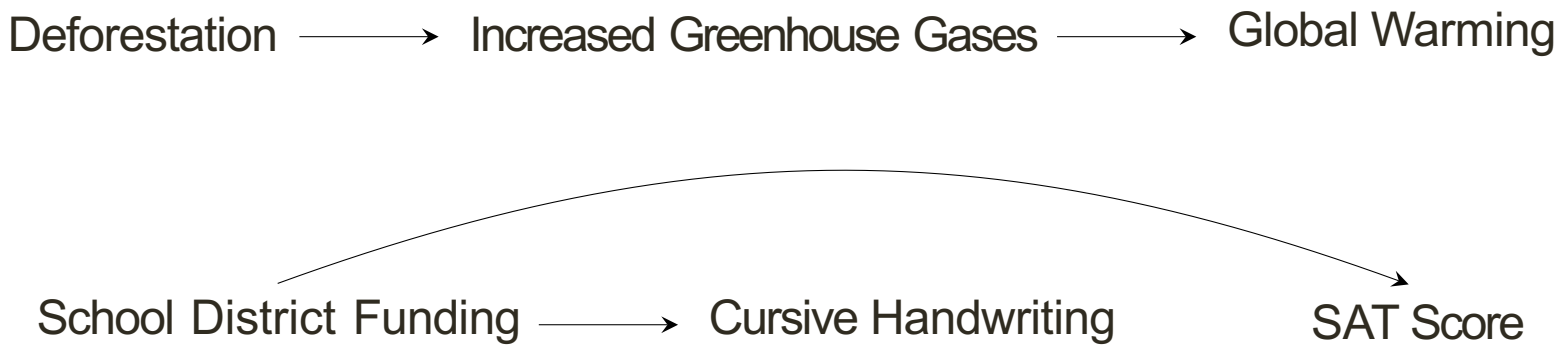
What type of study is this?

Can we conclude that writing in cursive *causes* higher scores?

Visualizing Causality

We can use Directed Acyclic Graphs (DAGs) to visualize the causal relationships that we assume exist between sets of variables:

- *Directed*: time flows from left to right via directed arrows
- *Acyclic*: no cycles or feedback loops, i.e., a variable cannot cause itself (either directly or through another variable)
- *Graphs*: intuitive way of encoding subject matter knowledge



* * * The association flows regardless of the direction of the arrows, but a path is only causal if it follows the direction of the arrows!

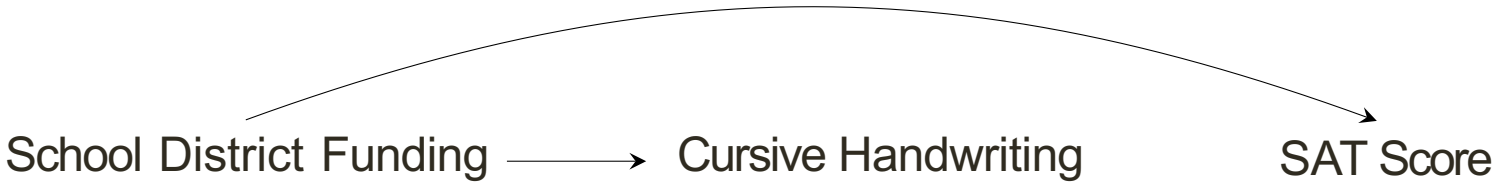
Confounding Variables

A confounder is a variable (which may or may not have been measured) that is:

1. Associated with the explanatory
2. variable Associated with the response
3. variable

   Not a downstream consequence of either the explanatory or response variable

It creates a backdoor path (i.e., path of association) between the explanatory and response variables:



Confounding Variables

A confounder is a variable (which may or may not have been measured) that is:

1. Associated with the explanatory
2. variable Associated with the response
3. variable

   Not a downstream consequence of either the explanatory or response variable

It creates a backdoor path (i.e., path of association) between the explanatory and response variables:

☣ Failing to adjust for confounders may lead to a spurious or distorted view of the explanatory/response relationship!

Randomization to the Rescue

In an observational study, the presence of confounders generally inhibits our ability to make causal claims.[1] But in randomized experiments, researchers randomly assign participants to one of two (or more) groups:

Treatment Group

Control Group

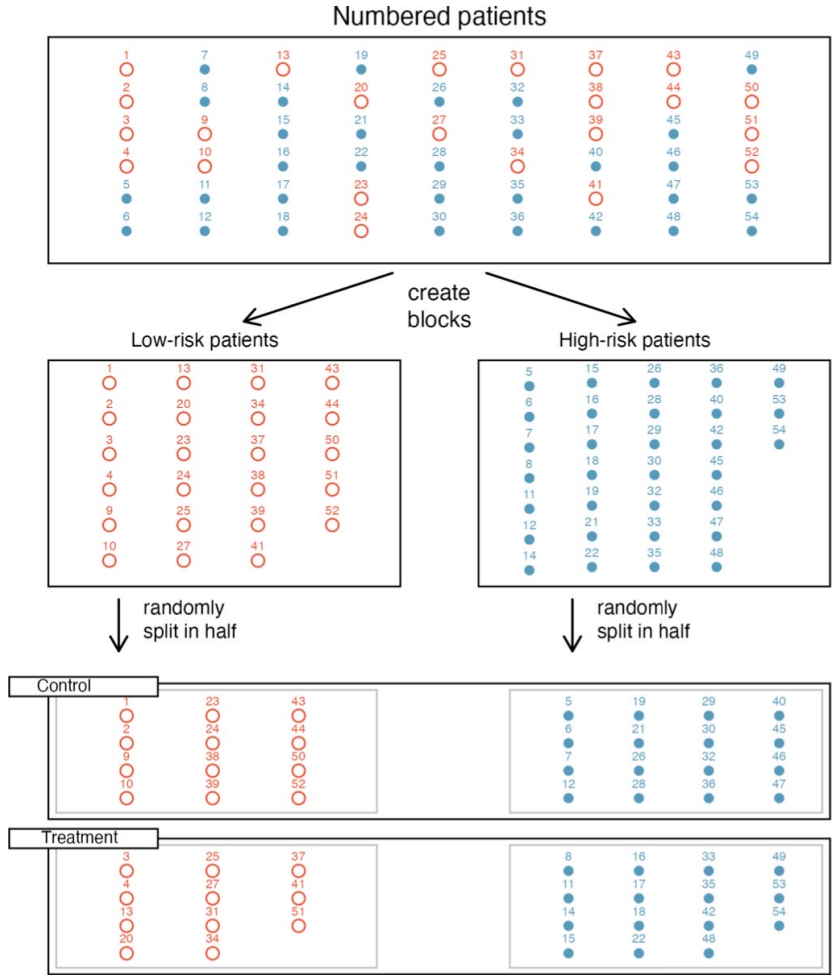
In randomly making this assignment, we ensure that nothing (beyond a coin flip) can be a cause of the explanatory variable!

⇒ randomization works to eliminate confounding by both measured and unmeasured variables!

[1] Without more advanced statistical techniques!

Blocking in Randomized Experiments

If we want to ensure that our treatment and control groups are balanced with respect to a particular variable, we may also employ blocking:



Causality and Experimental Design Practice

Form 6 groups, each will be assigned a scenario below. Work with your group to complete your scenario. *Be prepared to share your answer with the class!*

For each pair of explanatory and outcome variables below, suppose that someone conducts an observational study and arrives at the stated conclusions. Are these claims warranted? In particular, what is one possible confounder of the exposure/outcome relationship that might alternatively explain the association?

a) *We all scream for ice cream*: An investigator decides to look at the relationship between ice cream consumption and public welfare in New England. She finds that higher levels of ice cream consumption are associated with more deaths by drowning and concludes that ice cream represents a clear and present danger to the public.
b) *Nurse Ratched is ready to see you now*: A researcher observes that towns with higher numbers of doctors also report higher numbers of crimes. They conclude that doctors must commit crimes at higher rate than the general population does.
c) *Get a move on*: A physician notices that heart disease occurs more frequently in patients who are less physically active. On the basis of this observation, she starts recommending that all of her patients engage in at least thirty minutes of exercise every day.

Causality and Experimental Design Practice

Form 6 groups, each will be assigned a scenario below. Work with your group to complete your scenario. *Be prepared to share your answer with the class!*

For each pair of explanatory and outcome variables below, suppose that someone conducts an observational study and arrives at the stated conclusions. Are these claims warranted? In particular, what is one possible confounder of the exposure/outcome relationship that might alternatively explain the association?

a) *We all scream for ice cream*: An investigator decides to look at the relationship between ice cream consumption and public welfare in New England. She finds that higher levels of ice cream consumption are associated with more deaths by drowning and concludes that ice cream represents a clear and present danger to the public.

Causality and Experimental Design Practice

Form 6 groups, each will be assigned a scenario below. Work with your group to complete your scenario. *Be prepared to share your answer with the class!*

For each pair of explanatory and outcome variables below, suppose that someone conducts an observational study and arrives at the stated conclusions. Are these claims warranted? In particular, what is one possible confounder of the exposure/outcome relationship that might alternatively explain the association?

b) *Nurse Ratched is ready to see you now*: A researcher observes that towns with higher numbers of doctors also report higher numbers of crimes. They conclude that doctors must commit crimes at higher rate than the general population does.

Causality and Experimental Design Practice

Form 6 groups, each will be assigned a scenario below. Work with your group to complete your scenario. *Be prepared to share your answer with the class!*

For each pair of explanatory and outcome variables below, suppose that someone conducts an observational study and arrives at the stated conclusions. Are these claims warranted? In particular, what is one possible confounder of the exposure/outcome relationship that might alternatively explain the association?

c) *Get a move on*: A physician notices that heart disease occurs more frequently in patients who are less physically active. On the basis of this observation, she starts recommending that all of her patients engage in at least thirty minutes of exercise every day.

Causality and Experimental Design Practice

Form 6 groups. Work with your group to complete the problem below. *Be prepared to share your answer with the class!*

A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on people who wear glasses and people who don't, so they want to make sure both groups of people are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting (only desk lamps). In this experiment:

a) What is the response variable?

b) What is the explanatory variable? What are its levels?

c) What is the blocking variable? What are its levels?