AS.280.347 CLASS 1.1

Course Introduction

- Course information
- Recall of previous topics
- NMES data
- Using R markdown and GitHub Classroom

Health Data Analysis Practicum (AS.280.347)

- **Objective:** to enable each student to enhance his/her quantitative, scientific reasoning *and to achieve a functional standard in statistical data analysis using the R statistical language*

- **Modular Organization:**
  - **Module 1:** Risk of smoking-caused disease (LC, CVD, etc), the contribution of smoking, and possible effect modification by sex and SES
  - **Module 2:** Particulate air pollution and mortality in U.S. cities
  - **Module 3:** Individual projects!

Health Data Analysis Practicum (AS.280.347)

- **Teams of 3-5 students for the first two projects**
  - You are strongly encouraged to work together in groups prior to meetings to develop your teamwork skills, in particular listening and teaching

- **Student evaluation based on:**
  - **knowledge and skills in data analysis:** quality of the project
  - **contribution to group:** quality of group presentations; critiques by team members

- **Presentations:**
  - Group presentations for each of the first two projects
  - Individual mini-presentations for final project

Health Data Analysis Practicum (AS.280.347)

- **Computation:** Statistical software R
  - Bring your laptop (with R installed) to each course meeting
  - You will create all of your assignments using R markdown
  - You are encouraged to complete online modules on the DataCamp platform to learn and improve your R skills

- **Version control/collaboration:** GitHub
  - GitHub is an online compendium of file repositories where people can share their work, work collaboratively with others, and easily use a version control system to track development of software and projects.
  - We will share course materials through GitHub; you will collaborate in your teams using Guthub; you will turn in your work through GitHub

- **Class structure:**
  - We will usually start class by sharing YOUR work that has been done in the previous week
  - We will ask you to turn in your work to us (over email) by Sunday night at midnight so we can prepare for Monday's class
  - Everyone should be prepared to share their work and provide constructive feedback to their classmates each week

Communicating with instructors

- If you need to email us about a course-related matter:

phbiostats@jhu.edu

- This account is accessed by both Dr. Jager and Dr. Taub.
- **Emails to our individual accounts about a course-related matter will NOT receive a reply.**
- If asking a question about code or other work for an assignment, please also copy Ruthe (rhuang16@jhu.edu) on your email as well.

Discussion questions: Recall…

- What are the goals and steps in data analysis?
- What do we mean by “cause”?
- What is confounding?
- What is effect modification?

Counterfactual definition of “causal effect” of “treatment”

Our definition of “cause”:

- The difference between a population characteristic (e.g. mean) having given the treatment to everyone and the same population characteristic absent the treatment
- **Potential** for intervention – to have either, if not both worlds

Counterfactual data table

Person	Treatment(0/1)	Y(0)	Y(1)	Y(1)-Y(0)
1		16	22	6
2		17	18	1
3		15	20	5
4		18	20	2
5		16	18	2
6		14	22	8
Average		**16**	**20**	**4**

Actual data table

Person	Treatment(0/1)	Y(0)	Y(1)	Y(1)-Y(0)
1	0	16	?	?
2	0	17	?	?
3	0	15	?	?
4	1	?	20	?
5	1	?	18	?
6	1	?	22	?
Average		**16**	**20**	**4**

Module 1: Smoking and risk of disease

- Question 1.1 (Q1.1): How does the risk of disease compare for smokers and otherwise similar non-smokers?

- Question 1.2 (Q1.2): Does the contribution of smoking to the risk of disease vary by sex or SES?

- To address each question, we want:
  - a data display (graph or table!)
  - a statistical analysis

- We will answer these questions using data from the National Medical Expenditures Survey (NMES)

NMES data

```
> head(nmes.data)

     id totalexp lc5 chd5 eversmk current former packyears
1 20449 25951.58   1    0       0      NA      0         0
2 15534   378.33   0    0       1       1      0         3
3  9503    51.18   0    0       1       0      1        40
4 15024  1899.20   0    0       0      NA      0         0
5 17817   153.50   0    0       1       1      0        86
6 31716   270.00   0    0       0      NA      0         0


  yearsince      bmi beltuse educate marital poor age female
1         0 23.96408       2       4       1    1  78      1
2         0 26.68133       3       1       5    0  30      1
3         9 22.32027       3       4       1    0  72      1
4         0 25.06986       3       4       2    0  64      1
5         0 20.23634       3       1       1    0  59      1
6         0 22.19736       2       1       5    0  25      0
```

NMES data

- age: age in years
- female: 1=female, 0=male
- eversmk: 1=has ever been a smoker, 0=has never been a smoker
- current: 1=current smoker, 0=not current smoker
- former: 1=former smoker, 0=not former smoker, NA if eversmk=0
- packyears: reported packs per year of smoking (0 if eversmk = No
- yearsince: years since quitting smoking (0 if eversmk = No)
- totalexp: self-reported total medical expenditures for 1987
- lc5: 1=Lung Cancer, Laryngeal Cancer or COPD, 0=none of these
- chd5: 1=CHD, Stroke, and other cancers (oral, esophageal, stomach, kidney and bladder), 0=none of these
- beltuse: 1=Rare, 2=Some, 3=Always/Almost always
- educate: 1=College grad, 2=Some college, 3=HS grad, 4=other
- marital: 1=Married, 2=widowed, 3=divorced, 4=separated, 5=never married
- poor: 1=Poor, 0=Not poor

Discussion questions: Recall…

- What do we mean by “cause”?

- What is confounding?

- What is effect modification?

Counterfactual definition of "causal effect" of "treatment"

Our definition of "cause":

- The difference between a population characteristic (e.g. mean) having given the treatment to everyone and the same population characteristic absent the treatment
- **Potential** for intervention – to have either, if not both worlds

Counterfactual definition of “causal effect” of “treatment”

Our definition of “causal effect”:

- The difference (or other comparison) between a population characteristic (e.g. mean, risk) having given the treatment to everyone and the same population characteristic absent the treatment
- **Potential** for intervention – to have either, if not both worlds

In this case:

- Treatment = **smoking**
- Population characteristic = **risk of disease**
- We want to compare the risk of disease between two worlds where (1) everyone smokes and (2) no one smokes

Counterfactual data table

		no one smokes	everyone smokes
Person	**Treatment** 0 = doesn’t smoke 1 = smokes	**Outcome (0)** 0 = no disease 1 = disease	**Outcome (1)** 0 = no disease 1 = disease
1		0	1
2		1	1
3		0	0
4		0	1
5		0	0
6		0	1
Risk		**1/6 = .17**	**4/6 = .67**

Difference in risk = Risk (1) – Risk (0) = .67 - .17 = .5

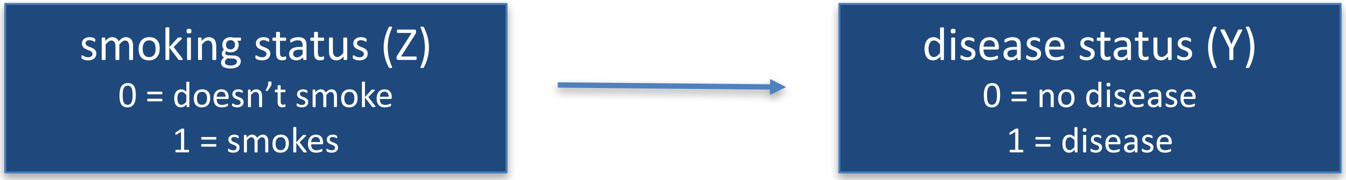
Actual data table

		non-smokers	smokers
Person	**Treatment** 0 = doesn’t smoke 1 = smokes	**Outcome (0)** 0 = no disease 1 = disease	**Outcome (1)** 0 = no disease 1 = disease
1	0	0	?
2	0	1	?
3	0	0	?
4	1	?	1
5	1	?	0
6	1	?	1
Risk		**1/3 = .33**	**2/3 = .67**

Difference in risk = Risk (1) – Risk (0) = .67 - .33 = .34

Confounding

When do we have **confounding** when studying the effect of an treatment Z (e.g., smoking) on an outcome variable Y (e.g., disease status)?



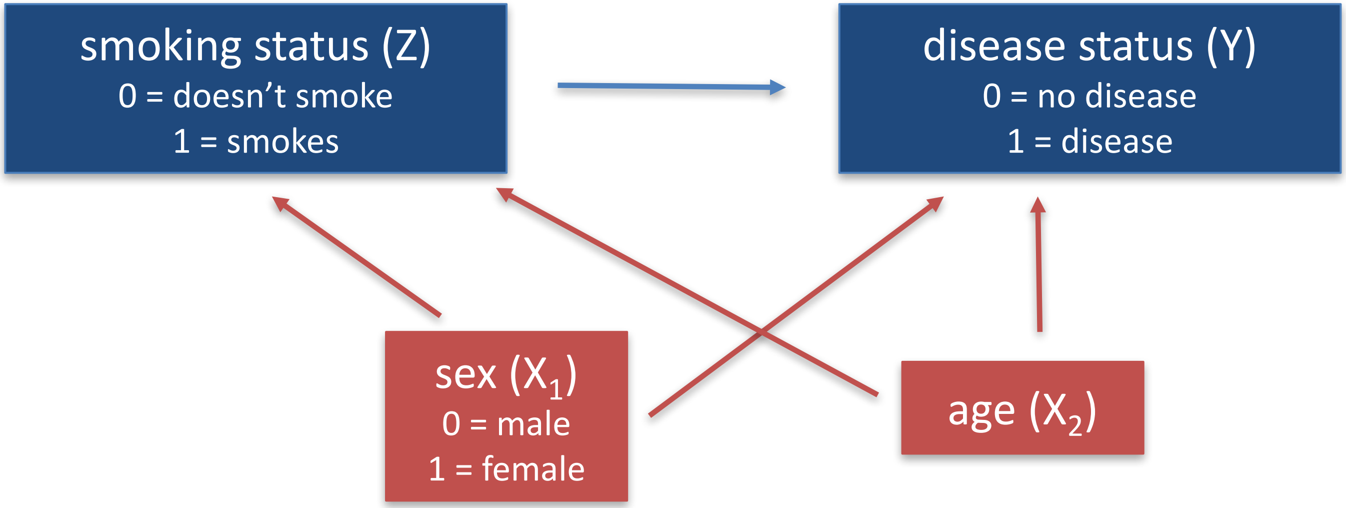
smoking status (Z)
0 = doesn't smoke
1 = smokes

disease status (Y)
0 = no disease
1 = disease

Confounding

When do we have **confounding** when studying the effect of an treatment Z (e.g., smoking) on an outcome variable Y (e.g., disease status)?

When we fail to compare **otherwise similar** units and, as a result, attribute to Z what is **actually caused by factors X** that differ between the Z=0 and Z=1 observations.



Assignment 1.1

- Create a data display with the NMES data to answer Q1.1:

  How does the risk of disease compare for smokers and otherwise similar non-smokers?

  – Work together in groups!

  – Submit your display in R markdown through GitHub by Sunday @ midnight

    - If you have trouble using GitHub, we will have a submission link available on Blackboard as well. By Assignment 1.2, we will REQUIRE all homework submissions to be through GitHub.

- Next week in class we will start with discussion/critiques of your displays.

  – Class brainstorming on ideas to improve these displays.