In this lab, we’ll look at data from an observational study\(^1\) to examine the relationship between smoking and getting coronary heart disease.

**Cases:** 609 males from Evans County, GA followed for 7 years

**Variables:**
- CHD (Response): Presence (coded 1) or absence (coded 0) of coronary heart disease.
- SMK (Explanatory): Coded 1 if subject ever smoked, 0 if never smoked.
- ID: Subject identifier
- AGE: Age
- CAT: High (coded 1) or normal (coded 0) catecholamine level.
- CHL: Cholesterol
- ECG: Presence (coded 1) or absence (coded 0) of electrocardiogram abnormality.
- DBP: Diastolic blood pressure.
- SBP: Systolic blood pressure.

1) **The Problem with Observational Studies**

This portion of lab is designed to demonstrate the problem with observational studies when looking for causal effects, and to illustrate confounding variables.

a) Get the data into Minitab. Download the data [here](#), do File -> Open Worksheet in Minitab, and choose the file. (Make sure to search for all types of files).

b) Explore the relationship between smoking and presence of coronary heart disease.

i) You can create a table of the data with Stat -> Tables -> Descriptive Statistics. You may want to add percent getting CHD for each category of smoking by clicking Display summaries for Categorical Variables.

ii) You can visualize the relationship with Graph -> Barchart -> Cluster or Stack or Graph -> Piechart -> Multiple graphs (by Variables tab to do piecharts of CHD by smoking status). Which visualization do you find most informative?

iii) Does there appear to be an **association** between smoking and getting coronary heart disease?

iv) Can you conclude, from these data alone, that smoking causes coronary heart disease? Why or why not?

---

c) There may be several confounding variables here, but we’ll explore age as a confounding variable. A confounding variable is a third variable that is associated with both the explanatory and the response variables. Explore the association between age and smoking status, and the association between age and CHD. For each, you may want to visualize the association with Graph -> Boxplot -> With Groups or Graph -> Dotplot -> With Groups. (Which visualization do you prefer?)

i) How is age associated with smoking?

ii) How is age associated with CHD?

iii) Does age appear to be a confounding variable?

d) What does this tell you about the relationship between smoking and CHD – do you think the association is stronger or weaker than you observed in 1b?

2) The power of randomization.

a) Suppose we could randomly assign people to smoke or not (we can’t actually do this, but pretend hypothetically we could). We can create a new variable that is randomly assigned, by randomizing the values in SMK. Like last lab, use Calc -> Random Data -> Sample From Columns.

How many rows do you want to randomly sample? Which column do you want to sample values from? Which column do you want to store the random smoking assignments in?

Name this new variable RandomSMK or something informative.

b) By the power of randomization, this new randomly assigned smoking variable should not be associated with any other potential confounding variables! Examine the relationship between RandomSMK and AGE (using the same methods as 1c). Are they associated?

c) Visually examine the relationship between RandomSMK and any other variable in the dataset. It shouldn’t be associated with anything! Randomization eliminates confounding variables!!!

d) If we had done this randomization at the beginning of the study, and if we still observed an association between smoking and CHD, we could then conclude causality because we could be sure the association wouldn’t be due to any confounding variables. (Note: this was an observational study so didn’t do this, but other studies have since established a true causal link between smoking and coronary heart disease.)
3) **Evaluate methods of data collection.** Use google scholar to investigate a topic you are interested in, preferably something related to your major. Find a study that involves data, and answer the following questions:

a) What are the cases? What are the variables?

b) If you were to represent the data as a dataset with cases as rows and variables as columns, how many rows and how many columns would it have?

c) How was the sample selected? How does this influence the types of conclusions that can be made?

d) Was it a randomized experiment or an observational study? How does this influence the types of conclusions that can be made?

e) If it was a randomized experiment, did they use a placebo? Was it double-blind? How does this influence how much you trust the results?