DESIGNING STUDIES

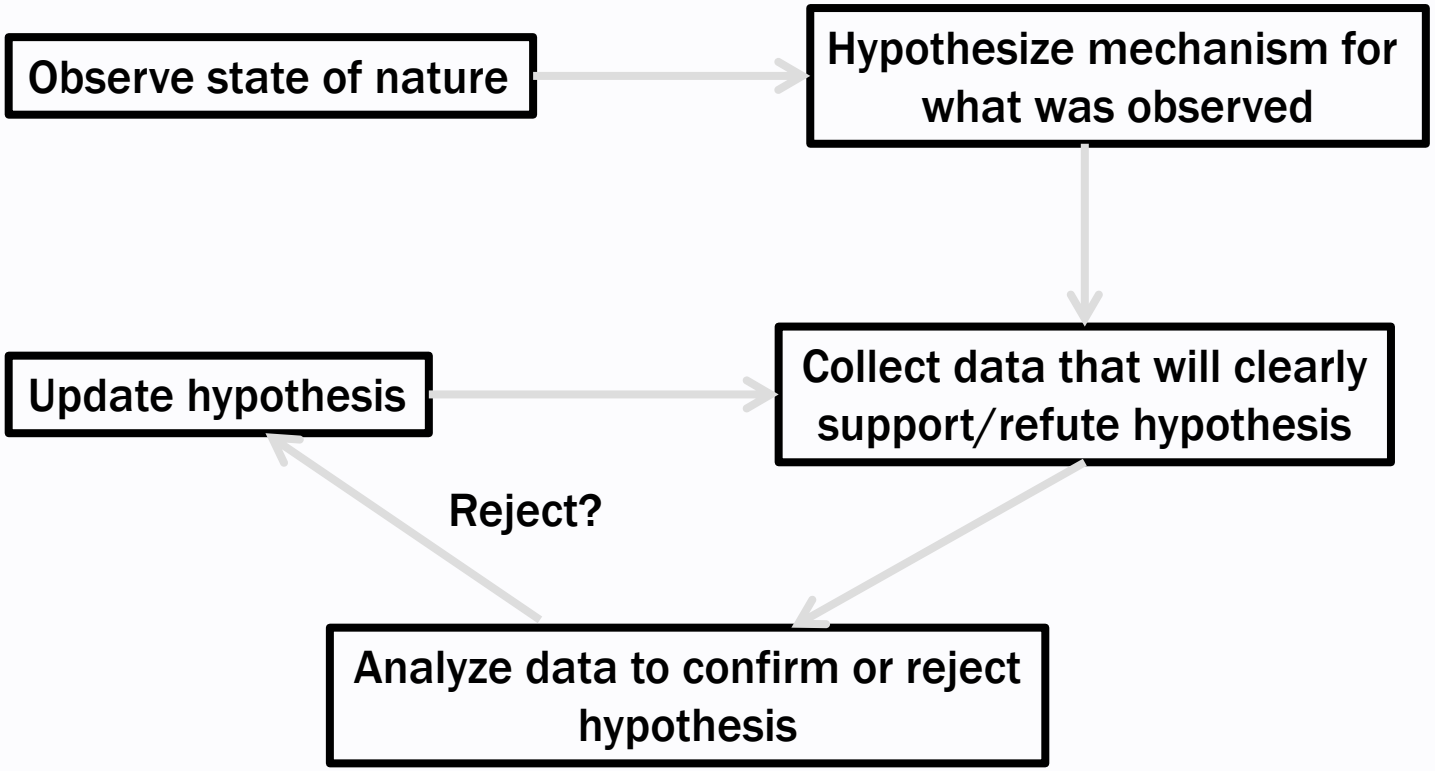
Chapters 1, 2

LEARNING OBJECTIVES

- Identify if a given data collection procedure is an observational study or randomized, comparative experiment
- Define experimental unit, observational unit, treatment factor, etc.
- Explain what pseudo-replication is
- Explain purpose of randomization

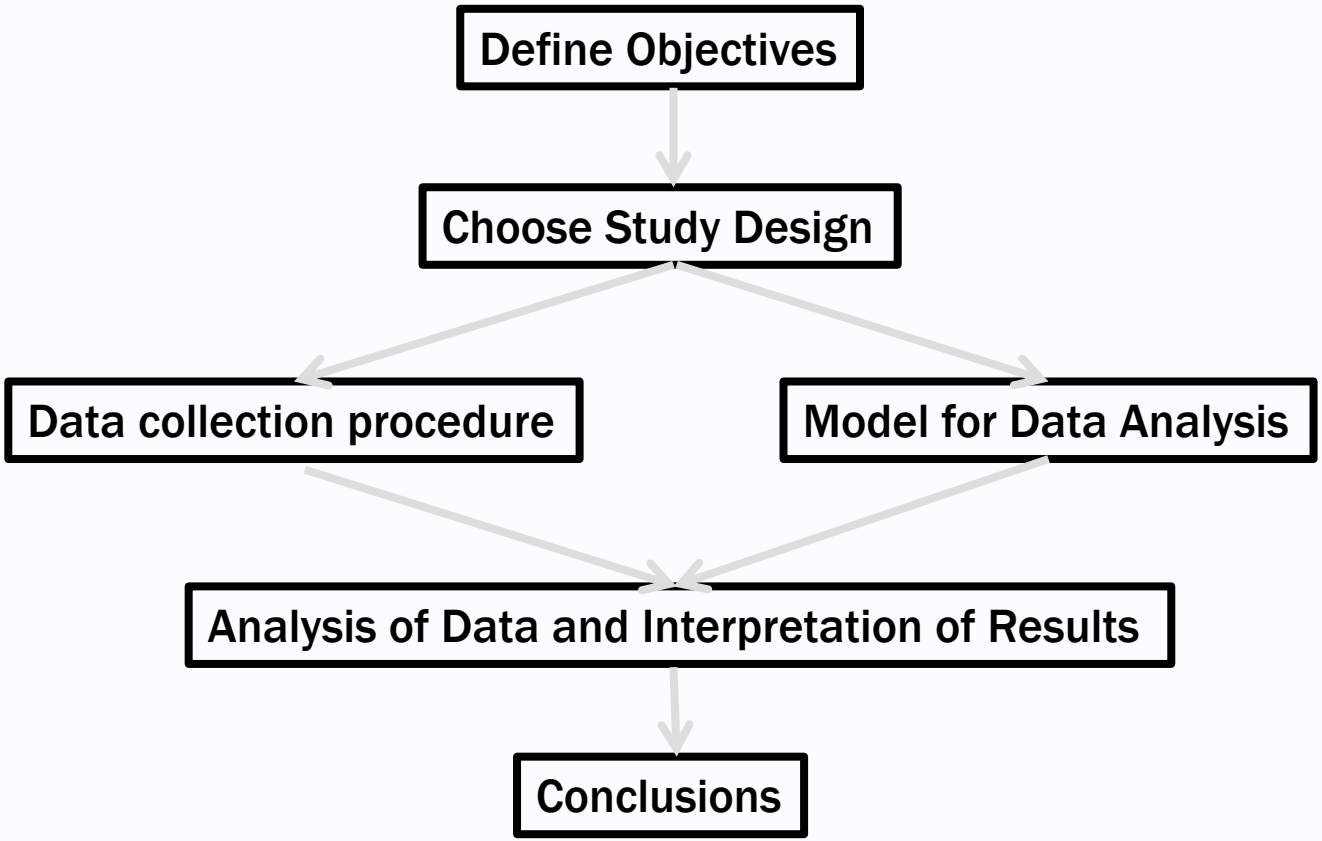
SCIENTIFIC METHOD

- Organized approach to avoid false starts and incomplete answers to research questions



EXPERIMENTATION FLOW DIAGRAM

- Hypothesis → data collection phase: clearly state the objectives of experiment



DEFINING OBJECTIVES

- "What scientific questions hoping to answer?"
- Data and analysis methodology to answer question?
  - Argue how the pairing can answer broad questions
- What is the statistical objective?
  - Understand distribution of a single response?
  - Determine relationships between multiple variables?
  - Build a predictive model?
  - Determine causes of variation of a response?
  - Find conditions that optimize response?

DOES SMOKING CAUSE LUNG CANCER? STUDY DESIGN 1

- Does frequent smoking of cigarettes cause lung cancer?
- Study design idea: randomly sample many smokers and nonsmokers and compare the relative proportions of those with lung cancer
- Analysis goal: determine whether smoking causes an increase in the probability of lung cancer
- Is there an analysis that could give valid **causal** inferences?

SMOKING STUDY DESIGN 1: TWO-SAMPLE PROPORTION TEST

- Two-sample proportion test and find a statistically-significant difference between the proportions
- Does this tell us smoking causes lung cancer?

- If there is a causal effect, we expect difference

- What else could explain the observed difference?
  - Random chance associated with Type I error
  - Any other thing that is related to or causes lung cancer such as genetics, work/home environment, age, etc.

SMOKING STUDY DESIGN 1: REGRESSION ADJUSTMENT

- Partition subjects based on extra variables and compare proportions within a group
  - Subjects in group influenced similarly by extra variables
  - Within-group differences more likely due to smoking alone
- Issues:
  - Group sizes can vary and may be small for some cases
  - Accounted for every possible extra variable?
  - Works in theory but under a lot of assumptions
- Upshot: need techniques to “adjust” for researcher’s lack of control over how subject becomes a smoker

DOES SMOKING CAUSE LUNG CANCER? STUDY DESIGN 2

- What if we could assign subject to be smoker?
  - Yeah…you can't do that
  - Let's just pretend we can for now
- Best way to assign subjects?
  - Limit possibility proportion differences occur due to known or unknown external variables
- How to guard against variables we don't know about?

RANDOM ASSIGMENT

SMOKING STUDY DESIGN 2: RANDOM ASSIGNMENT

- Randomization reduces probability external variables are what cause differences
  - Guaranteed to remove effects for studies with many subjects
  - Still the gold standard for establishing causality
- Different types of random assignments:
  - Completely randomized designs for unknown external variables
  - Block designs for some known external variables
  - Split plot designs for complicated assignments of treatments
- "Block what you can, randomize what you can't"

OBSERVATIONAL STUDIES AND RANDOMIZED, CONTROLLED EXPTS

- Study 1 is an observational study
  - Observe and record without controlled intervention
  - Only choose what and how you observe
- Study 2 is randomized, comparative experiment
  - Controlled intervention where external variables held constant and others are purposefully changed
- Observational studies aren't bad, you just have to be careful about scientific conclusions made from data

SOURCES OF VARIATION
INDEPENDENT AND TREATMENT FACTORS

- **Independent variable (or factor):** variable thought to influence the dependent variable
  - Not in every observational study but is in every experiment
  - Values of a factor are called **levels**
- **Treatment factor:** factor that experimenter intentionally varies
  - Unique to designed experiments
- **Dependent variable (or response):** measured variable we think may be influenced by changes in independent variables

TYPES OF EXTERNAL VARIABLES AND RELATIONSHIPS BETWEEN VARIABLES

- **Nuisance factor:** known variable that could influence response but not of particular interest
  - Sometimes fixed to specific level
  - Otherwise should be adjusted in the analysis
- **Lurking variables:** like nuisance factor except it is unknown or unobservable
- Randomization reduces chance treatment factor(s) are correlated with these types of variables
  - Two factors are **confounded** if perfectly correlated
  - E.g. every time assign subject to not smoke they must also exercise three times a week, smokers not allowed to exercise

SMOKING STUDY DESIGN 2: APPLYING DEFINITIONS

- **Independent variables (sources of variation):**
  - Smoker/non-smoker
  - Genetic information
  - Age
  - Environment information
- **Dependent variable:**
  - 0/1 indicator for whether subject has lung cancer
- **Activity:** identify the treatment factor and at least one nuisance and lurking factor

TREATMENT APPLICATION PROCESS AND EXPERIMENTAL UNITS

- **Treatment application process:**
  - How is a treatment level applied?
  - How much control do you have over application?
  - What is it applied to?
  - Guarantee applications of same treatment are independent?

- **Experimental unit (EU):** "subject" or "material" receiving an independent application of a treatment
  - How do you expect the treatment to affect the EU?

- **Observational unit (OU):** part of the EU that measurements are taken on

RUNS, REPLICATES, AND TREATMENT APPLICATION ERROR

- **Run:** the application of a treatment level to an EU and dependent variable measurement(s) from EU
  - # of EUs = # of runs
  - Yields one or more response measurements

- **Treatment replicate:** Independent application of a treatment to a new EU

- EU is potentially influenced by **ALL steps involved in the treatment application process**
- Expect variation between replicates of same treatment (**treatment application error**)

SMOKING STUDY DESIGN 2: APPLYING DEFINITIONS

- Activity: propose a treatment application process for a completely randomized, smoking study
  - Take multiple measurements from each person throughout the study period
  - 500 people recruited in the study

- Based on your description, what are EUs and OUs?
- How many replicates would you recommend per treatment level?
- Give hypothetical cause of treatment application error

SMOKING STUDY DESIGN 3

- Randomly partition 50 mice into 10 chambers with 5 mice each

- Randomly assign 5 of 10 chambers to receive smoke
- Mice put in same chamber each day
- Detect presence/absence of cancer in each mouse
- What are the EUs and OUs in this experiment?

PSEUDO-REPLICATION: CONFUSING OUS WITH EUS

- Every EU receives independent treatment application
- OUs from EU receive same treatment application
  - Different from “receiving the same treatment”
- Thinking of OUs as EUs is called pseudo-replication of a treatment
  - Assumes you have more information about treatments than you really do
  - Leads to larger Type I errors

LEARNING OBJECTIVES REVIEW

- Identify if a given data collection procedure is an observational study or randomized, comparative experiment
- Define experimental unit, observational unit, treatment factor, etc.
- Explain what pseudo-replication is
- Explain purpose of randomization