Intro to NRES 710

Welcome!

Fall 2024

NRES 710 is designed to be an introductory, graduate-level statistics class. "Introductory" because you are not assumed to have much in the way of prior statistical knowledge (although an undergrad course under your belt will be helpful) or prior experience in R programming (although some prior exposure will make the learning curve less steep). "Graduate-level" because once we get rolling, it'll be full steam ahead! However, another dimension of a graduate-level course is an emphasis on collaborative knowledge development through group discussion and collective learning, and NRES 710 will provide opportunities for this.

This course provides an introduction to both statistics and computer programming. With a solid grounding in basic rules of probability and statistics, and armed with some basic computer programming abilities (and some human creativity and ingenuity), you can go extremely far in making sense of data and communicating that understanding with others!!

Install R and RStudio

CRAN website for downloading R
RStudio main site

Make sure you have the most recent versions!

Make a new RStudio "project" for this class

R projects are an **extremely useful** feature of RStudio. Just store all the code and data for this course in your project folder and it will make your life much, much easier!

Download the R code for this lecture!

To follow along with the R-based lessons and demos, right (or command) click on this link and save the script to your project directory

Why do we need (this particular) stats class?

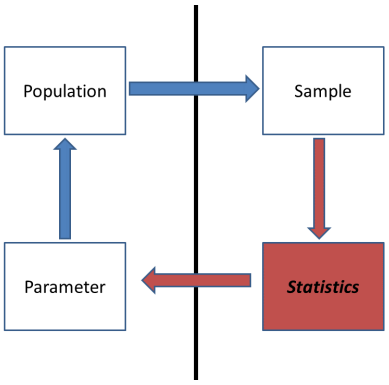
Many professors recommend that grad students take the least coursework possible and focus on gathering knowledge as needed while conducting their research. But typically, with statistics coursework as a notable exception. It can be more efficient, fun, and rewarding to learn statistics as a group rather than on your own. I hope you agree by the end of this class!

Course Overview

We will discuss the course organization and expectations. Please see course syllabus and schedule documents, provided separately.

What is statistics?

"Statistics is like grout –the word feels decidedly unpleasant in the mouth, but it describes something essential for holding a mosaic in place. Statistics is a common bond supporting all other science." -Ramsey and Schafer (2013), The Statistical Sleuth, 3rd Edition.



Statistics is the process of making inference about a property of a *population* (a parameter) from a representative *sample*. The vertical line separates the stuff we can observe and measure directly (right side) and the stuff we can't observe but want to make inference about (left side).

Commonly, statistical analysis is conducted to compare parameters from multiple populations (e.g., "are the population means different?"), or to compare parameters estimated from a sampled data set with those of an idealized model representing an underlying hypothesis of interest (or null hypothesis).

More broadly, a common dictionary definition of statistics includes "... the collection, organization, analysis, interpretation, and presentation of data" (Oxford Dictionary, Wikipedia, etc.). Given this broader definition, statistics includes key components of data collection (via study design), data analysis, data science, scientific inference, and science communication.

Different Approaches to "the Scientific Method"

Scientific inquiry is furthered through using both inductive and hypothetico-deductive approaches. Both require statistics in order to test hypotheses about how nature works. The inductive approach involves a continuous and cyclical generation of hypotheses from observations; developing predictions from hypotheses; matching new observations to predictions in order to test hypotheses; and revising hypotheses -until a hypothesis is ultimately confirmed. The emphasis is on hypothesis verification. The hypothetico-deductive approach, on the other hand, proposes multiple working hypotheses from the outset, with an emphasis on falsifying incorrect hypotheses. An "accepted truth" is one that stands up to repeated attempts at falsification. Here is an illustration of the inductive method, taken from Gotelli and Ellison (2004):

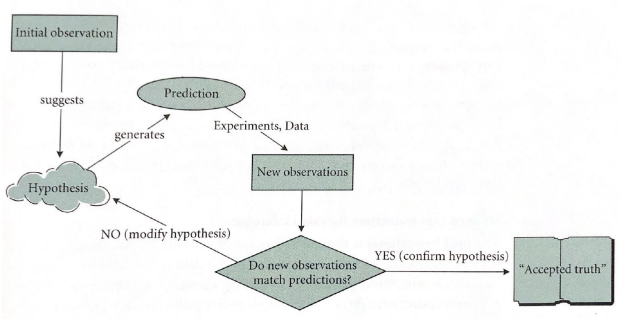


Figure 4.1 The inductive method. The cycle of hypothesis, prediction, and observation is repeatedly traversed. Hypothesis confirmation represents the theoretical endpoint of the process. Compare the inductive method to the hypothetico-deductive method (Figure 4.4), in which multiple working hypotheses are proposed and emphasis is placed on falsification rather than verification.

And here, an illustration of the hypothetico-deductive method, taken from the same source:

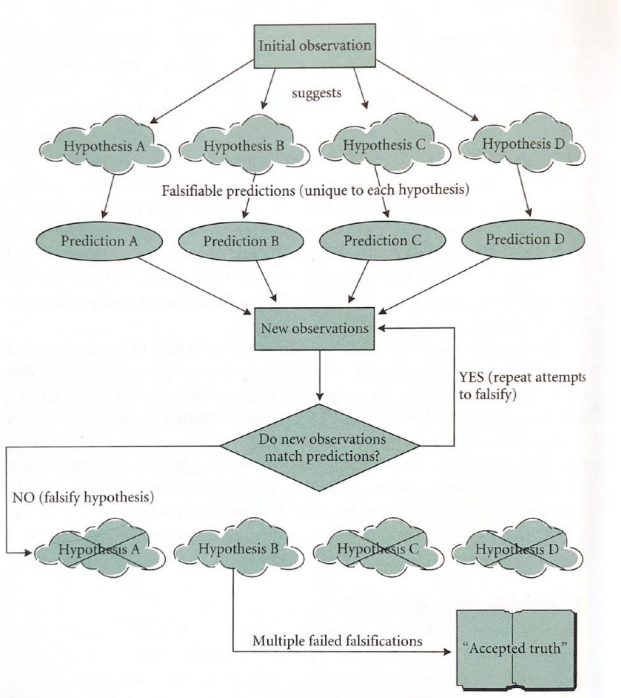


Figure 4.4 The hypothetico-deductive method. Multiple working hypotheses are proposed and their predictions tested with the goal of falsifying the incorrect hypotheses. The correct explanation is the one that stands up to repeated testing but fails to be falsified.

Data are fundamental to science!

All of science is empirical- that is, it relies on data for evaluating the validity of conclusions.

Statistics is fundamental to science because it allows us to better understand what our data can and can not tell us.

What types of data are there?

1. Categorical – represents qualitative (labeled) characteristics
    a. Nominal (> 2 categories)

    b. Dichotomous (2 categories)
2. Quantitative
    a. Discrete: can be counted (number of individuals in a population)
    b. Continuous: not countable, but can be measured (total body length, temperature)

Why do we need to know about data types? We have to analyze and visualize these data types differently!

For example:

- Nominal data can be visualized with bar plots and pie charts
- Continuous data can be visualized with histograms or scatterplots.

Working with data in R:

Let's explore some of R's built in data sets: iris, mtcars, titanic (install titanic package first)

Notes about working in R:

R is an open source project, and new packages are being added all the time.

R is incredibly powerful and feature rich. You are NOT expected to memorize syntax right away, but rather just know that the answer is always a few clicks away!

'Base R' is the default software built into R that does not including loading any additional packages. Here is a 'base R' cheat sheet; this is a great reference for most of the basic tasks you will need to perform in R.

Learn to use R scripts, and save your scripts frequently! This is the primary record of what you've done and allows you and others to reproduce your workflows.

If you have a problem, Google it! Someone has likely had the same problem as you in the past and asked for solutions online. Or, query ChatGPT for suggestions on how to solve the problem. When either Googling or prompting ChatGPT for help, be mindful that potential solutions presented to you may not be ideal for your problem.

First R demo!

NOTE: for those wishing to follow along with the R-based demo in class, click here for an R-script that contains all the code blocks in this web-based lecture.

All of you should have R and RStudio installed on your computers. See the links page for some useful references.

Starting at the most basic level, R can be used as a calculator. Try it!

```
# Getting started with R -----------------

# Use R as a calculator

2 + 2                 # use R as a calculator
```

```
## [1] 4
```

```
four <- 2 + 2         # define your first variable!
four
```

```
## [1] 4
```

```
five <- 2 + 2         # you can make mistakes and define misleading labels- R will let you!
three <- four + five
```

What about those hashtags (#) in the above code block? These are 'comments' and they are super helpful- use them early and often!

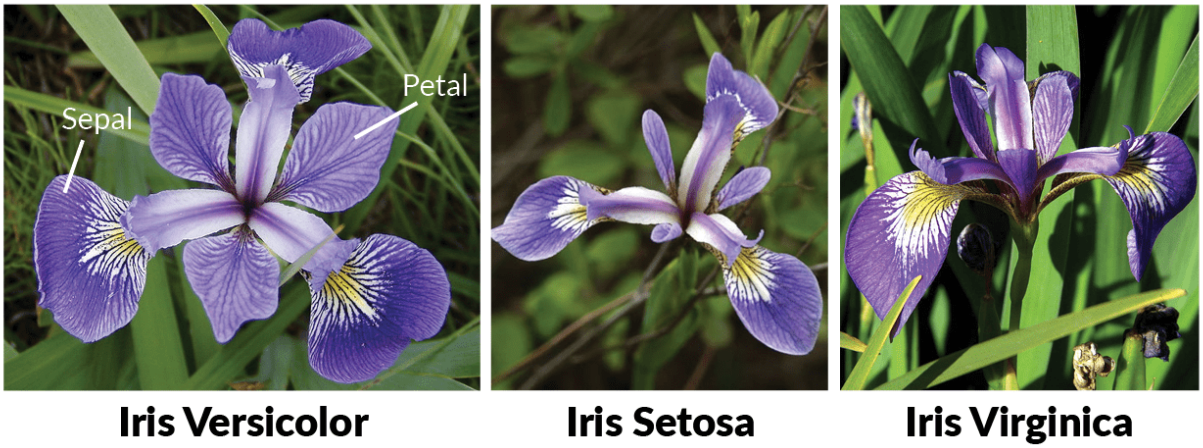
Use RStudio's autofill feature to avoid typos!

```


# R has many built in datasets

# data()    # 'uncomment' this command and run it to explore built-in datasets
            # code can be uncommented with CTRL SHIFT C (PC) or COMMAND SHIFT C (Mac)
```

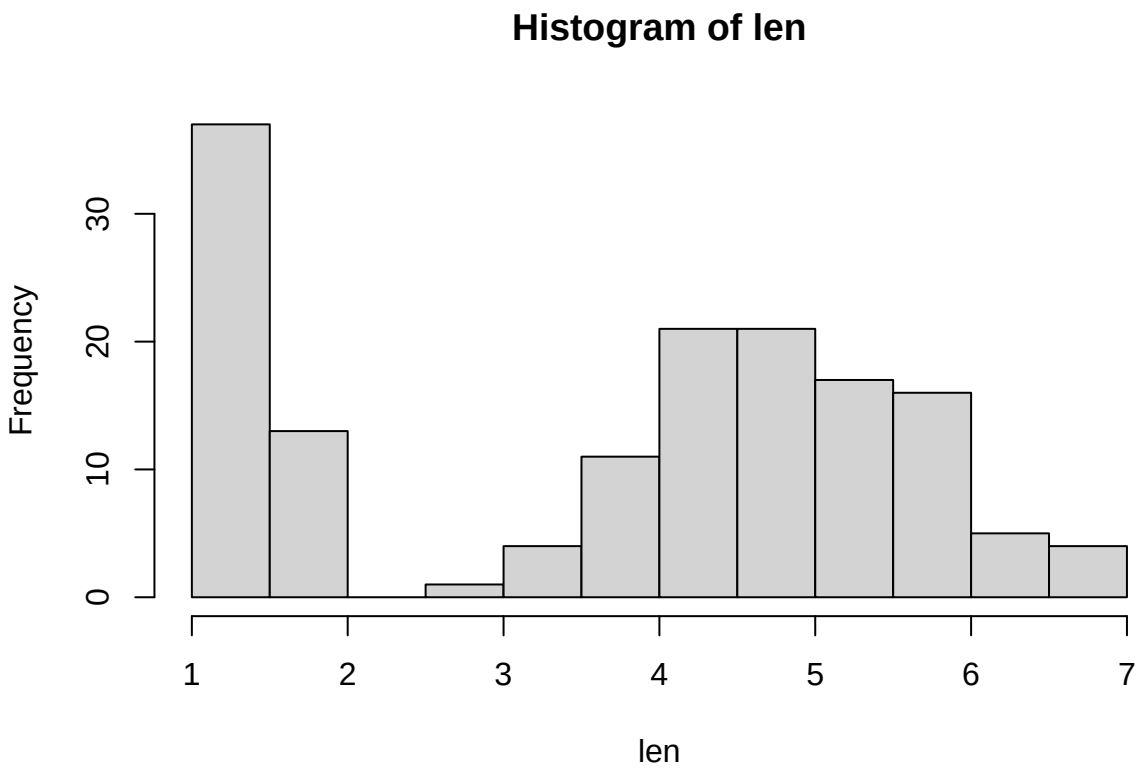
Explore R's existing datasets Let's start by working with Fisher's famous iris dataset:




```
#iris                  # this is a data frame -- the basic data storage type in R
head(iris)             # [add your own comment here!]
# tail(iris)

#  ?iris               # uncomment this to learn more about the iris dataset
# str(iris)


len <- iris$Petal.Length
hist(len)              # what does this do? How could you learn more about this 'hist' function?
```




```
# Q: what kind of data are petal lengths?
```

Now let's switch to the 'titanic' dataset. To get this dataset you need to install an R package!

```
#install.packages("titanic")     # uncomment this command to install the package- you only need to inst

library(titanic)              # this 'loads' the package and needs to be done every time you run this s

data("titanic_train")
head(titanic_train)
# ?titanic_train              # uncomment and run to learn more about the data

# Q: What kind of data are those in the "Embarked" column?
# Q: What kind of data are those in "Pclass?"
```

We can even make our own dataset!

```
# Make our own data --------------------

# lets pull 15 numbers from the standard normal distribution

a <- rnorm(15)
a <- rnorm(15, mean = 2, sd = 0.5)

# let's pull 15 numbers from the binomial distribution
```

```r
b <- rbinom(15, size = 1, prob = 0.2) # we could "weight the coin"


# we can create categories:
unit <- rep(c("Control","+N","+P","+NP"), each = 20)


# we can even create a whole dataframe
my.data <- data.frame(
  Obs.Id = 1:100,
  Treatment = rep(c("A","B","C","D","E"), each = 20),
  Block = rep(1:20, times=5),
  Germination = rpois(100, lambda = rep(c(1,5,4,7,1), each = 20)),
  AvgHeight = rnorm(100, mean = rep(c(10,30,31,25,35,7), each = 20))
)
head(my.data)
```

We can also import data from files stored on our computers (or even directly from the web)

```r
# import data from file ----------------------

# Don't forget to set your working directory (or just make sure you're using an Rstudio Project).

# setwd("~/Desktop")       # uncomment and run if you want to set the desktop as your working directory

#Read in the data. Note that the file needs to be in csv format, the name must be in quotes, and the na

# Pleach<-read.csv("PbyTime_Bio.csv", header=T) # obvs, this won't work for you because you don't have

# Use your own file to try it out. This is an example!
```

Note: I recommend always using RStudio projects whenever you are working on any statistical analysis in R (homework exercises, class projects, analysis for your thesis chapter, etc.). This reduces the hassle of setting working directories. By default, the project directory automatically becomes the working directory for your analysis!

"StatsChats"

Faculty from various departments – Paul Hurtado (math/stats dept), Ken Nussear (geography), Perry Williams (NRES), Kevin Shoemaker (NRES) and others – have hosted informal sessions for grad students in EECB, NRES, Geography etc (and faculty) to discuss data analysis questions. This can be a good opportunity to ask questions about statistics, find out more about what types of data and questions your peers are working with, and help others (thereby reinforcing concepts for yourself!). Keep an eye on that website linked above and if the group identifies a regular meeting time, I will relay that information to you all in class.

Feedback

This is my first time teaching this class. I welcome frequent and honest feedback on course topics to understand whether course topics are appropriate and/or useful. So, any feedback is welcomed - thank you!

–go to next lecture–