

# STAT 439

## Introduction to Categorical Data Analysis

Prof. Stacey Hancock

Welcome!


Course webpage:

<https://staceyhancock.github.io/stat439/>

2

## Motivating Case Study: Challenger Explosion



 The Space Shuttle Challenger lifts off (left) and explodes shortly after over the Kennedy Space Center, Fla., Tuesday, Jan. 28, 1986. All seven crew members died in the explosion, which was blamed on faulty o-rings in the shuttle's booster rockets. The Challenger's crew was honored with burials at Arlington National Cemetery. (Bruce Weaver/AP Photo)

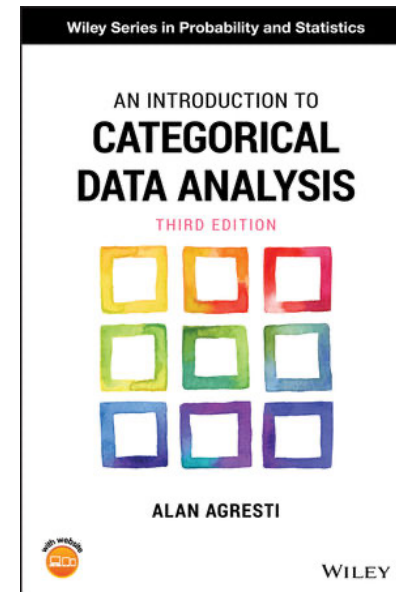
3

## Syllabus Highlights

<https://staceyhancock.github.io/stat439/>

- Office hours
- D2L Discussion forums
- Course description and goals
- Prerequisites
- Textbook
- R and RStudio
- Assessment
- Policies

4



## Data Analysis

What steps do we go through when conducting a complete inferential data analysis?

1. Aims of a data analysis
  - Description, exploration, confirmation, prediction
2. Establish the context of the analysis
  - Statistics produces inference about a population based upon a sample
  - Need to understand the population sampled
  - Understand the data collection procedure (true random sample?)
  - Understand the background science
  - The scientific goals of the analysis/experiment

## Data Analysis

3. Develop a statistical model
  - Clearly defined (measurable) outcome is essential
  - Predictor(s) of interest
  - Covariates or confounders we need to consider
  - If we cannot decide which parameters would be appropriate when measurements are available on the entire population, then there is *no chance* that statistics can be of help!
4. Evaluation of the properties of the design, model, and estimation procedure
  - Essential that these aspects be addressed as completely as possible *prior* to data analysis
  - Clear specification of outcomes and predictors
  - Use of robust statistical methods

## Data Analysis

5. Look at the data – visualize!
6. Computation
  - Turn the handle...
7. Interpretation of results
  - Present results clearly and precisely at a level appropriate to the audience
  - If applicable, present scientific justification for why results agree with the hypothesis (should have been done at the design stage)
  - The most elegant experiment/data analysis is meaningless unless it can be easily explained to the scientific community it was designed to impact!

# CHAPTER 1

---

Introduction to Categorical Response Data

Sections 1.1-1.2

## Types of Variables

- A “**variable**” is a characteristic of an observation.
  - Its value “varies” across observational units.
- **Categorical** (qualitative) **variables** are variables whose values consist of a label or category.
  - Examples?
- **Quantitative variables** are variables whose values consist of a number that has meaning.
  - Examples?

## Types of Categorical Variables

- Categorical variables for which the categories form a natural ordering are called **ordinal** categorical variables.
  - Examples?
  - Statistical analyses should take advantage of the ordering for scientific and efficiency reasons.
- Categorical variables whose possible outcomes have no natural ordering are called **nominal** categorical variables.
  - Examples?
  - Statistical methods for ordinal data should *not* be used on nominal data.
- Categorical variables with only two possible outcomes are called **binary** variables (often coded as 0 or 1).
  - Examples?

## Response vs. Explanatory Variables

- We would like to explain or predict the variation in the response (dependent) variable Y using the variation in the explanatory (independent) variable X.
  - Not necessarily a *causal* relationship!
- Examples?
  - X = aspirin or placebo; Y = heart attack (yes/no)
  - X = annual income; Y = job satisfaction
  - X = sex; Y = hourly wage

## Summary of Statistical Methods

		Response Variable Y	
		Categorical	Quantitative
Explanatory Variable(s) X	Categorical	Contingency tables, e.g., chi-squared test, Fisher's Exact test (Ch. 2) GLMs (Ch. 3-7)	2-sample t-test ANOVA
	Quantitative	GLMs (Ch. 3-7)	Simple and multiple linear regression

STAT 439

STAT 216-217  
(Quantitative response assuming normal distribution)