ST517 Note Outline 12: Wrap up and Big Ideas

We have covered many topics in this course, focusing on each topic independently. This section is intended to wrap the course up by connecting the various topics, while summarizing the important aspects and big ideas to take from the course.

Big Ideas (Part 1)

- Statistics is the science of using data to make informed decisions
- No single study "proves" or "disproves" an idea, but rather there are "levels of evidence" a statistical study provides in support of an idea
    - Some studies provide good evidence that can help you in making a decision
    - Others provide poor evidence and should not be considered when making a decision
    - Many studies are in-between these extremes
    - Need to figure out the quality of a study for yourself
- Studies that make good use of some basic statistical principles provide stronger evidence of an idea than studies that don't; for example:
    - A randomly-selected sample provides a better representation of the population than does a non-random sample
    - A well-designed, randomized experiment provides stronger evidence of a cause-and-effect relationship than does an observational study
- It is important to carefully consider the strengths and weaknesses of a study's design, data collection, and data analysis when evaluating the study's conclusions and the quality of evidence that study provides
- It is also important to carefully consider the results of several studies when evaluating the cumulative "body of evidence" of an idea

Goals of (most) Statistical Studies

1. To learn about a population—a large group of people or things that you would like to learn about but typically cannot measure in full
2. To compare two (or more) populations
3. To summarize the relationship between two (or more) variables

Accomplishing these Goals

- Main tool broadly referred to as **statistical inference**
  i.e. the process of learning about the population(s) by taking a sample(s)
    - Main tools of statistical inference: confidence intervals and hypothesis tests
        - Can be used to help accomplish all 3 goals
    - Another tool useful specifically for goal 3 is regression analysis
- The data collection process affects your ability to conduct statistical inference
    - If the sample is well-chosen and the study is well-designed: Statistical inference is relatively straightforward; we can say that what we learn about sample also holds true for the entire population
    - If the sample is not well-chosen or the study is not well-designed (e.g. there are lurking variables, response bias, or other forms of bias): Inference is not as straightforward
        - Need to think more carefully about the study
        - Ex: what population can the results extend to? (E.g. If sample only contains students from Statistics, results only generalize to students from Statistics, not all students at NCSU)
- Probability distributions (specifically the **sampling distribution**) provide the foundation for conducting statistical inference
    - Statistical inference is possible because the sampling distribution follows a predictable pattern
    - Sampling distributions allows us to quantify the **sampling variability** in sample statistics, including how they differ from the parameter (this distance is known as the **margin of error**) and what type of variability would not be expected to happen by random chance (by using a p-value to determine **statistical significance**)
    - By understanding the sampling distribution we can understand how statistics act and answer questions about them
    - Allows us to address research questions using what we know about how statistics are expected to behave

Overview: Confidence Intervals (CI)

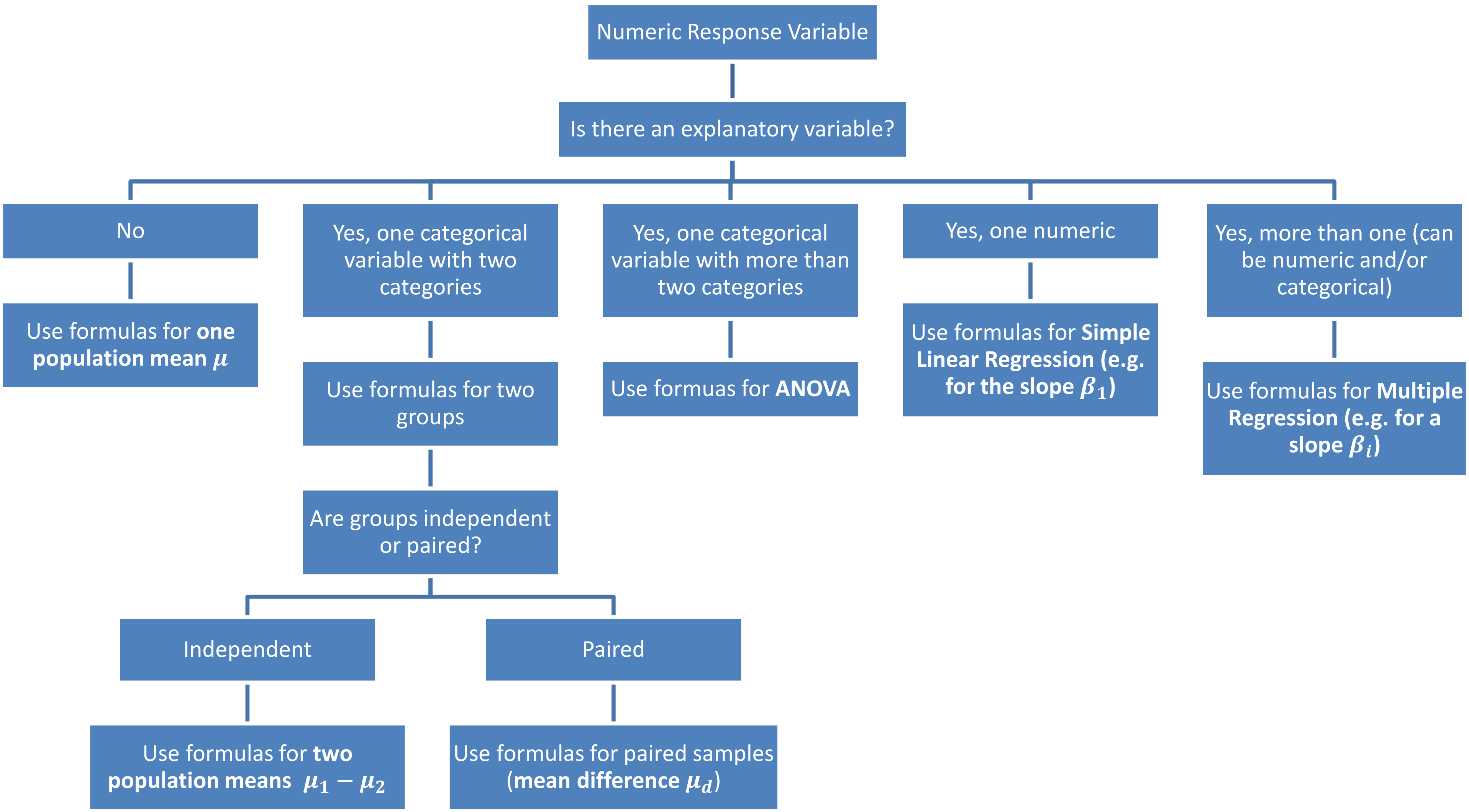
- General purpose: To estimate what is true in the population
- Specific types of confidence intervals we have seen this semester:
  - For the mean of a single population (μ), including paired differences (μ_d)
  - For the difference in means of two independent populations ($\mu_1 - \mu_2$), pooled and un-pooled
  - For a proportion from a single population (p)
  - For the difference in proportions of two independent populations ($p_1 - p_2$)
  - For a the slope of a regression model (β_1)
  - For the average value of y at a particular value of x
  - For an individual value of y at a particular value of x

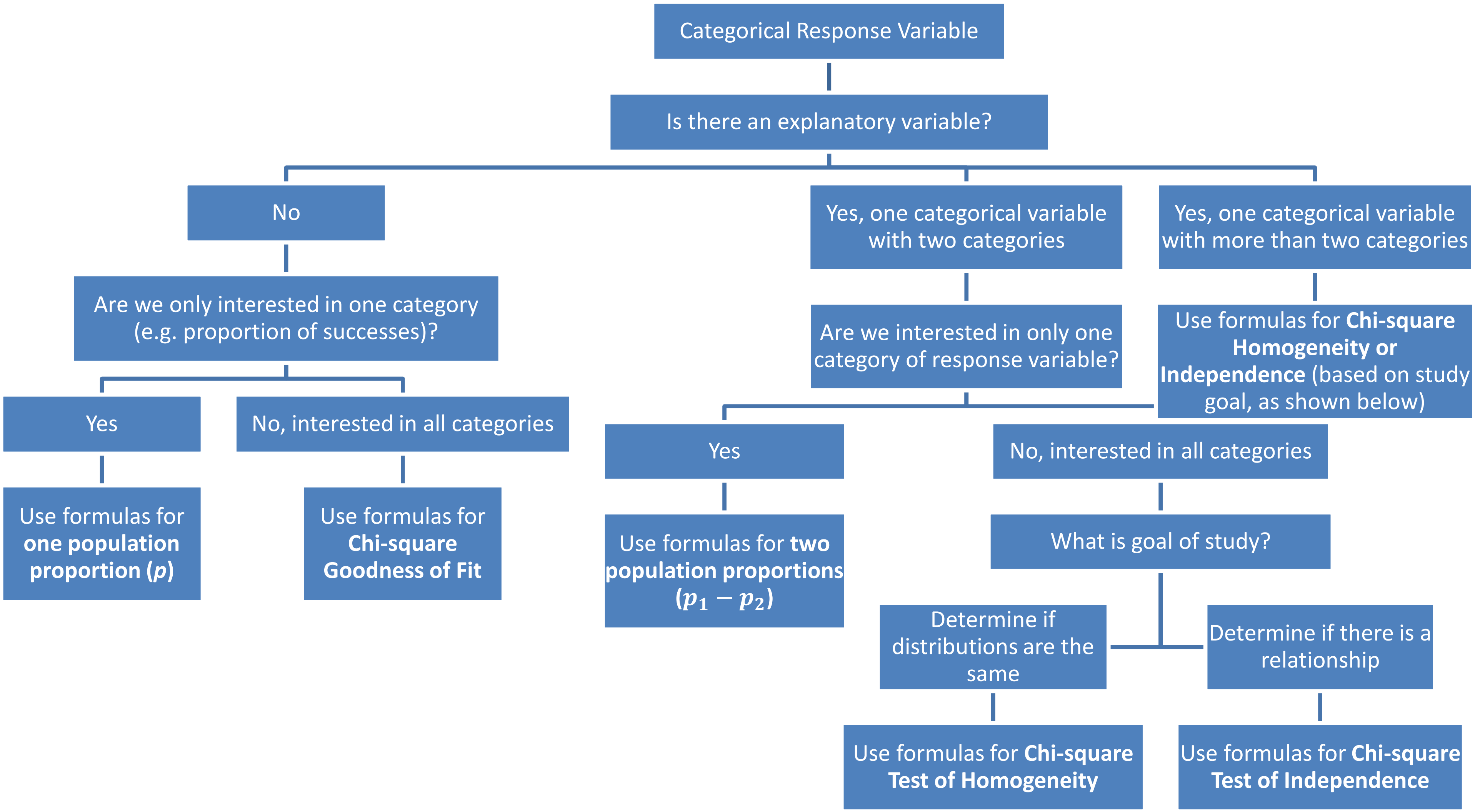
Overview: Hypothesis Tests (HT)

- General purpose: To determine if a particular hypothesis about a population is supported by evidence or not
- Specific types of hypothesis tests we have seen this semester:
  - Same scenarios as those for confidence intervals (except the last two)
  - For the mean of three or more independent populations (ANVOA)
  - For categorical variables with more than two categories (Chi-square)

Determining Which Formula to Use

- In practice, confidence intervals and hypothesis tests should be used together to give a more holistic understanding of the conclusions
  - Determining formulas = identifying parameter
- For this course, I will often assess one at a time
  - Determining formula = identifying tool, then identifying parameter
- Identifying the tool to use:
  - I will specify this, so look for keywords, for example:
    - CI: interval, margin of error, multiplier, standard error
    - HT: hypothesis, test, test statistic, null distribution, p-value, statistical significance, significance level
    - Regression: relationship, model, equation, predict
- Identifying the Parameter:
  - Ask yourself a series of questions, starting with: Is the response variable quantitative or categorical?
  - For the rest of the questions, see the flow charts on the next two pages


Numeric Response Variable
Is there an explanatory variable?
No
Use formulas for **one population mean** μ
Yes, one categorical variable with two categories
Use formulas for two groups
Are groups independent or paired?
Independent
Use formulas for **two population means** $\mu_1 - \mu_2$
Paired
Use formulas for paired samples (**mean difference** μ_d)
Yes, one categorical variable with more than two categories
Use formuas for **ANOVA**
Yes, one numeric
Use formulas for **Simple Linear Regression (e.g. for the slope** β_1)
Yes, more than one (can be numeric and/or categorical)
Use formulas for **Multiple Regression (e.g. for a slope** β_i)


Categorical Response Variable
Is there an explanatory variable?
No
Are we only interested in one category (e.g. proportion of successes)?
Yes
Use formulas for **one population proportion (p)**
No, interested in all categories
Use formulas for **Chi-square Goodness of Fit**
Yes, one categorical variable with two categories
Are we interested in only one category of response variable?
Yes
Use formulas for **two population proportions** $(p_1 - p_2)$
No, interested in all categories
What is goal of study?
Determine if distributions are the same
Use formulas for **Chi-square Test of Homogeneity**
Determine if there is a relationship
Use formulas for **Chi-square Test of Independence**
Yes, one categorical variable with more than two categories
Use formulas for **Chi-square Homogeneity or Independence** (based on study goal, as shown below)

Big Ideas (Part 2)

- There are many other types of confidence intervals and hypothesis tests that we did not have time to cover this semester
- However, they all have things in common with the examples we did cover
  - See Note Outline 5 for important points about confidence in general
  - See Note Outlines 6 for important points about testing in general
- All forms of statistical inference (confidence intervals, hypothesis tests, regression analysis) have their own conditions that must hold in order for them to be valid
  - Ensure procedure is correct for the situation
  - Ensure formulas used for margin of errors and test statistics are valid
  - Specific conditions depend on the type of inference (e.g. type of parameter)
- Basic steps for conducting the hypothesis test are the same for any test:
  - Conducted within overall process of data analysis (i.e. the scientific method)

Overview: The Scientific Method (The Outline for this Course)

1. Identify the research Question(s)
   - Discussed throughout the course via examples
2. Conduct background Research
   - Not discussed this semester
3. Form a Hypothesis or make a Prediction
   - Discussed primarily on Note Outlines 6 to 10 via examples
4. Experiment / Collect Data
   - Discussed in Note Outline 1
5. Explore, summarize, and Analyze data
   - Tools for EDA presented in Note Outline 2
   - Tools for formal analysis = tools of statistical inference
     - Foundations of inference (probability and sampling distributions) presented in Note Outlines 3 and 4
     - Confidence intervals presented in Note Outline 5, 7 to 10
     - Hypothesis tests presented in Note Outlines 6 to 10
     - Regression analysis presented in Note Outlines 9 to 11
6. Make Conclusions

Final Thought: People are Necessary

- Software calculates the results
- People can think, software doesn't
- The method for collecting the data is important
- The interpretation is important
- Thinking about the real world situation is important