Bringing it all together

December 4, 2020

Final resources

- Practice final examinations
  - I will post two previous final exams. One was cumulative and we will indicate on it which questions are not applicable
  - Inference Formula sheet has been posted on the course website.
    - This is “bare bones” – you’ll have to do some work annotating this so you know when to use which formula.

R Code to know

- Code that is fair game for writing/interpretation of the code or resulting output: `qt()`, `pt()`, `qnorm()`, `pnorm()`, `pchisq()`, testing functions(`t.test()`, `binom.test()`, `prop.test()`, `chisq.test()`), `broom` functions (i.e., `tidy()`, `glance()`, and `augment()`), `lm()`, `predict()`, `confint`, `aov()`, `TukeysHSD()`, functions covered by Mi-Suk's guest lecture
- Code that is fair game for interpretation: `ggplot2`, `dplyr`, `infer`, and may have a few minor points for general R intuition (e.g. what does `<-` do?)

Bonus point!

- Screenshot and submitted to Gradescope (open now through Dec 13 at 11:59pm, absolutely no lates permitted)
- 1 bonus point added to your total grade if you complete by the deadline

Part III of the course

- Heavily focused on conducting hypothesis tests and calculating confidence intervals
- We covered many tests one by one. Your task is to be able to know what test applies when you read a question.

Parts of a hypothesis test

- What are the assumptions?
- State the null and alternative hypotheses. Are they one or two-sided?
- Calculate the test statistic
- Calculate the p-value (or write/identify the code to do so)
- Interpret the p-value in terms of how probable the result is assuming the null hypothesis is true.

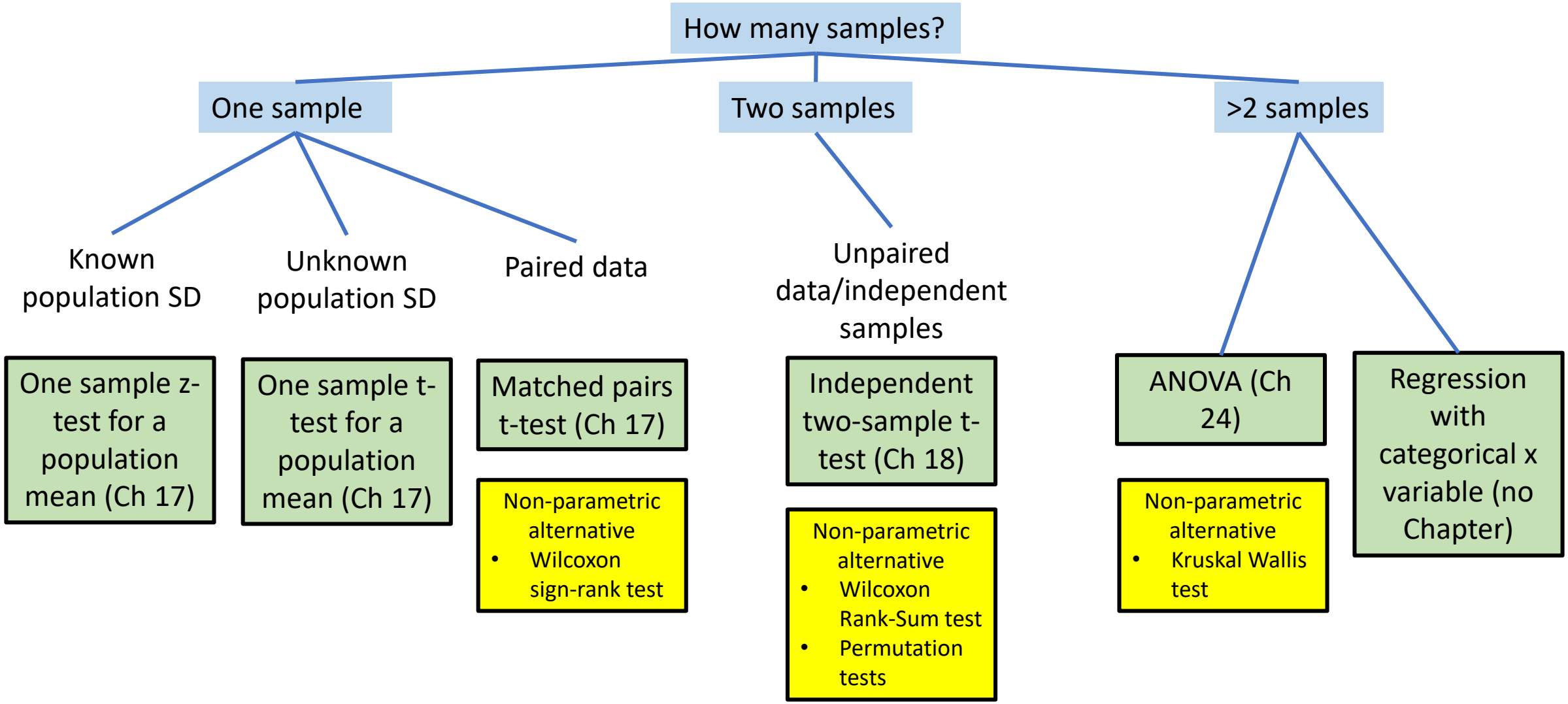
Creation of confidence interval

- Form: estimate +/- (critical value x standard error)
- Estimate is what you calculate from your data
  - The sample mean
  - The sample proportion
  - The difference in means (or proportions)
- The critical value is found using one of the R `q` functions like `qnorm()` or `qt()`. You are asking R for the value such that 95% (or 99%, say) of the area of the distribution is between +/- that value.
- The standard error is calculated using a formula, such as s/sqrt(n). The standard error decreases as the sample size n increases

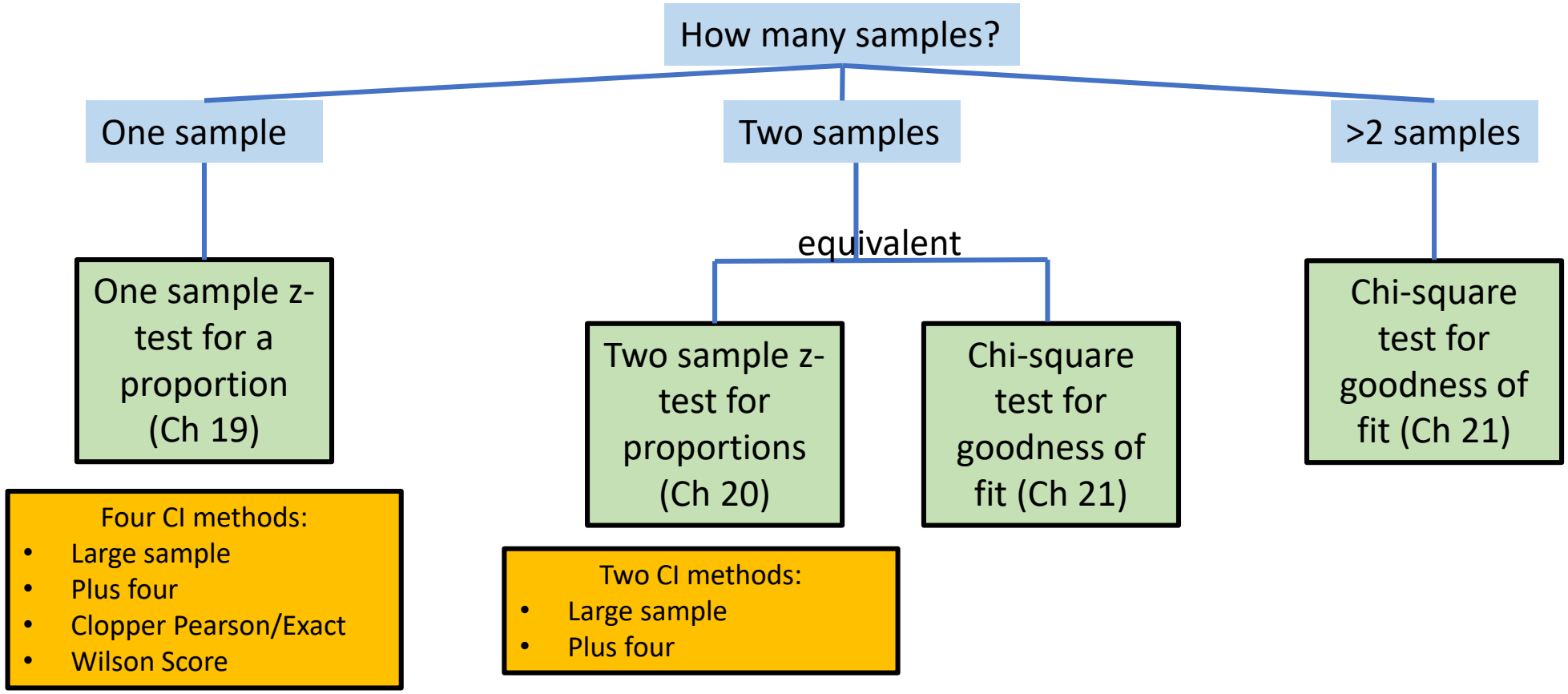
Questions to ask yourself when you read a question

- What type of data is represented?
  - Continuous/quantitative
  - Binary
  - Categorical with >2 levels
- How many samples are there?
  - One sample
  - Two samples
  - > two samples
- How many variables are there?

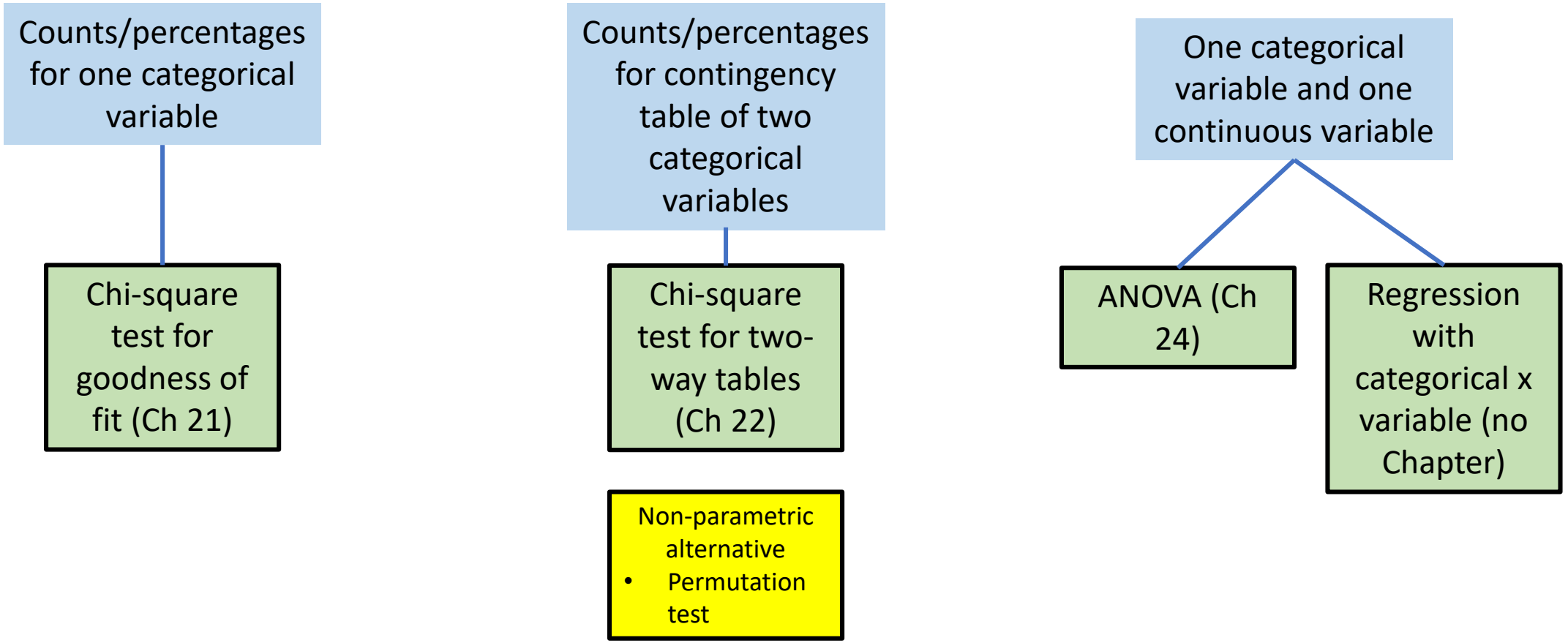
If you have continuous data



If you have binary data



If you have categorical data (>2 levels)



What about inference for regression?

- Continuous data
- One sample
- Two continuous variables: an explanatory variable x and a response variable y

t-test for the regression slope (Ch 23)

t-test for correlation (Ch 23)

We didn't cover this because it is equivalent to test for slope

Example 1: Which test to perform?

- The amygdala is a brain structure involved in the processing of memory of emotional reactions. Ten subjects were shown emotional video clips. They had their brains scanned and their memory of the clips assessed. The first three rows of the data frame looks like this:

Relative activity	Memory score
-0.417	31
-0.258	29
-0.234	29

- What type of data do you have?
- How many samples?
- How many variables?

Example 2: Which test to perform?

- A study investigated ways to prevent staph infections in surgery patients. In a first step, the researchers examined the nasal secretions of a random sample of 6771 patients admitted to various hospitals for surgery. They found that 1251 tested positive for *Staphylococcus aureus*, the bacterium responsible for most staph infections.
  - What type of data do you have?
  - How many samples?
  - How many variables?

Example 3: Which test to perform

- A study on the effects of vaping classifies people as “never vapers”, “occasional vapers”, “frequent vapers”. You interview a sample of 150 people in each group and ask a questionnaire to derive a quantitative score (between 0 and 100) on stress levels.
  - What type of data do you have?
  - How many samples?
  - How many variables?

Example 4: Which test to perform?

- Essential tremor is a neurological movement disorder characterized by involuntary rhythmic movement that typically interferes with the full use of the arms and hands. A pilot experiment examining the effectiveness of a noninvasive handheld device using active cancellation of tremor technology to stabilize tremor-induced motion in patients diagnosed with essential tremor. Tremor amplitude was measured (in centimeters) for each of 11 subjects when performing a spoon-use tasks with the ACT device turned, in random order, once on and once off.
  - What type of data do you have?
  - How many samples?
  - How many variables?

Example 5: Which test to perform?

- A random sample of 700 births from local records shows this distribution across the days of the week. Do these data give evidence that local births are not equally likely on all days of the week?

Day	Births
Monday	110
Tuesday	124
Wednesday	104
Thursday	94
Friday	112
Saturday	72
Sunday	84

  - What type of data do you have?
  - How many samples?
  - How many variables?