Bringing it all together

December 4, 2020

1

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.

2

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?)

3

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

4

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.

5

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.

6

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

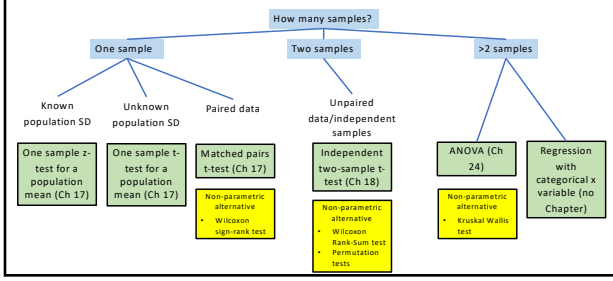
7

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?

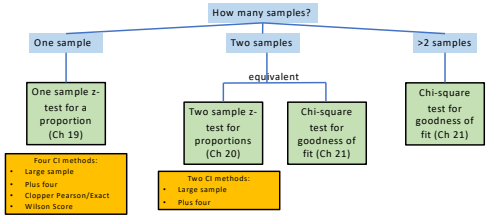
8

If you have continuous data



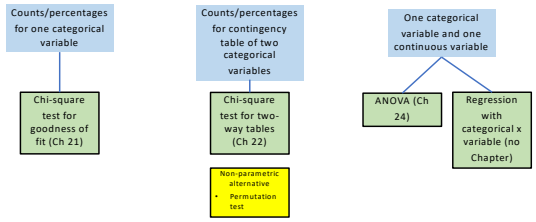
9

If you have binary data



10

If you have categorical data (>2 levels)



11

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

12

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?

13

Example 1: Which test to perform? (From Ch 23)

- Continuous data
- One sample (on ten participants)
- Two variables

→ t-test for the regression slope (Ch 23)

While it isn't yet apparent (because a question wasn't asked), this should signal linear regression because that is what we try when we have two continuous variables on one sample

Questions that could be asked (if you had all the data):

- Make a scatter plot. Does it appear linear?
- Interpret the provided r or r^2 values.
- Write code to test the slope coefficient. Write the null and alternative hypotheses for the corresponding test. Interpret the output from the test (including the p-value).
- Assess the provided diagnostic plots. Does the data meet the assumptions?

14

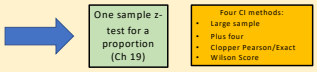
Example 2: Which test to perform?

- 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?

15

Example 2: Which test to perform? (Ch 19)

- Binary data (staph or not)
- One sample
- One variable



Questions that could be asked:

- Perform the test (write your hypotheses, evaluate the assumptions, calculate the test statistic, write code to find the p-value, interpret the p-value
- Calculate the confidence interval by hand using large sample or plus four method (when to use plus four method?). Write code for other two methods
- Evaluate the assumptions for using the method

16

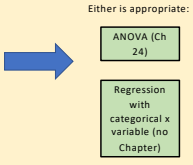
Example 3: Which test to perform

- 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?

17

Example 3: Which test to perform (Ch 24)

- You have one categorical variable (level of vaping), and one continuous variable (stress level)
- The categorical variable corresponds to three samples – one for each level of the category
- Two variables: one categorical and one continuous



18

Example 3: Which test to perform (Ch 24)

Questions that could be asked:

- State the null and alternative hypotheses for an ANOVA test
- What are the numerator and denominator degrees of freedom for the ANOVA test?
- With more information, you could perform the ANOVA calculate, or write code to perform the calculation, or interpret ANOVA R output
- You could be asked to write a model to perform regression with x as a categorical variable (if you knew the name of the dataset and variables).
- You could be asked to interpret the output from a regression model performed on these data

19

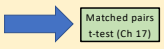
Example 4: Which test to perform?

- 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?

20

Example 4: Which test to perform? (Ch 17)

- Tremor amplitude is a continuous variable
- Two samples, but paired data (before and after)
- If helpful, can think of "on" and "off" as a second, categorical variable.



Questions that could be asked:

- (If you had the data or some statistics): Perform the appropriate test. Write code for the p-value.
- What is the key feature of these data that determines which test is appropriate?
- How could you plot the data before preforming the test to visualize the statistic of interest?
- Define wash out period, and carry over effects.

21

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?

22

Example 5: Which test to perform? (Ch 21)

- Categorical variable with > 2 levels: Day of the week
- One sample
- Note that you have a table of data where the numeric information is only one column and you're asked if data is evenly distributed across the days
  - This signals that the null is an "even distribution" across the days and should remind you to calculate observed counts (provided) and expected counts (under the null)
- Questions that could be asked:
  - Calculate the expected counts, calculate the test statistic, what are the degrees of freedom, write code to calculate the p-value. What day of the week contributes data that is furthest from the null hypothesis?

23