Reflection Questions for Math 58 & 58B

Johanna Hardin - Pomona College

Spring 2020

The questions below come from an introductory statistics course which follows closely **Introductory Statistics with Randomization and Simulation** by Diez, Barr, & Çentinkaya-Rundel. `https://www.openintro.org/` The indicated chapters and sections are from that text.

Chapter 2, Sections 1-4 – hypothesis foundations

1. What is the difference between a statistic and a parameter?
2. In a typical study, do you have one statistic or more than one statistic? And do you know the value of the statistic?
3. In a typical study, do you have one parameter or more than one parameter? And do you know the value of the parameter?
4. Explain what it means for a statistic to have a distribution.
5. What is a p-value?
6. What is the difference between a one- and two-sided hypothesis?
7. What is the difference between a null hypothesis and an alternative hypothesis?

Chapter 2, Sections 5-7 – normal model

1. What does it mean for something to have a normal distribution?
2. How can you use the normal curve to calculate percentages or probabilities?
3. What does it mean for $\hat{p}$ to have a distribution? Can you explain in words?
4. What does the central limit theorem tell us about the distribution of $\hat{p}$?
5. What technical conditions are important in order for the central limit theorem to apply?
6. What does a Z score measure?

Chapter 2, Section 8 – confidence intervals

1. What is a confidence interval?
2. Part of the CI interpretation includes a phrase "95% confident." Explain what 95% means.
3. How can you find the appropriate Z* value?
4. What is the difference between a Z score and Z*?
5. When computing a confidence interval (i.e., when we don't have a preconceived idea for p), how is the standard deviation of $\hat{p}$ estimated?
6. When using the normal distribution to create a confidence interval for p, how is the critical value for, say, a 94.7% interval calculated?

Chapter 1, Sections 3-4 – sampling

1. Why is it good to take random samples?
2. What is a simple random sample?
3. Why don't researchers always take random samples?
4. What benefit(s) does a large sample provide to the study?
5. What is the difference between practical significance and statistical significance?

Chapter 2, Section 3 – errors & power

1. Why is it never okay to accept H_0?
2. What is the difference between a Type I and Type II error?
3. Which is worse: a Type I error or a Type II error?
4. What is power? How is power calculated? What does power depend on?

(no ISRS) Binomial probabilities (Math 58 only)

1. How can the binomial distribution be used to calculate probabilities?
2. What are the technical conditions of the binomial distribution?
3. How is the normal distribution different from the binomial distribution? (one answer is that the normal describes a continuous variable and the binomial describes a discrete variable. what does that mean? what is another distinction?)

4. What are the technical conditions allowing the normal distribution to approximate the binomial distribution?

5. What is one reason to choose to use the normal distribution?

6. What is one reason to choose to use the binomial distribution?

(no ISRS) Relative Risk & Odds Ratios (Math 58B only)

1. What is the differences between cross-classification, cohort, and case-control studies?

2. When is it not appropriate to calculate differences or ratios of proportions? Why isn't it appropriate?

3. How are odds calculated? How is OR calculated?

4. What do we do when we we can't calculate statistics based on proportions? Why does this "fix" work?

5. What is the statistic of interest? What is the parameter of interest?

6. Why do we look at the natural log of the RR and the natural log of the OR when finding confidence intervals for the respective parameters?

7. How do you calculate the SE for the $\ln(\hat{RR})$ and $\ln(\hat{OR})$?

8. Once you have the CI for $\ln(RR)$ or for $\ln(OR)$, what do you do? Why does that process work?

Chapter 3, Section 2 – 2 binary variables

1. What is the statistic of interest? What is the parameter of interest?

2. How does the inference *change* now that there is binary (response) data taken from two populations?

3. How does the inference *stay the same* now that there is binary (response) data taken from two populations?

4. What does the Central Limit Theorem say about *two* sample proportions?

5. When is it appropriate to apply a hypothesis test to the data? And when is it appropriate to apply a confidence interval to the data?

6. How do we calculate $\text{SE}(\hat{p}_1 - \hat{p}_2)$?

7. What technical conditions must hold for the Central Limit Theorem to apply?

Chapter 1, Sections 4-5 – types of studies

1. What is the difference between an observational study and an experiment?
2. Why aren't all studies done as experiments?
3. What is a confounding variable?
4. Have you looked at page 199 of ISCAM? (also copied into the class notes) Do you understand that page? [Random sampling vs. Random allocation]
5. How is the statistical meaning of the word *cause* different from the usage in the sentence: *The ball that hit me in the head caused me to get a headache.*
6. What are the meanings of the words: randomized, double-blind (single-blind), control, placebo, significant, and comparative. Why are these ideas important to interpreting study results?

Chapter 3, Section 3 – 2 categorical variables

1. How would you describe the data seen in $r \times c$ tables?
2. Describe the simulation mechanism that creates a sampling distribution under the assumption that the null hypothesis is true (like the cards in the first week of class using the gender discrimination example).
3. What is the test statistic (for both the infer simulation and the chi-square test with the mathematical model!!)? Why do we need a complicated test statistic here and we didn't need one with 2×2 tables?
4. How do you compute the expected count? What is the intuition behind the computation?
5. What is one benefit that the two sample z-test of proportions has? That is, what is one thing we can do if we have a 2×2 table instead of an $r \times c$ table?
6. Describe the directionality of the test statistic. That is, what values of X^2 make you reject H_0?
7. What are the technical assumptions for the chi-square test? Why do you need the technical assumptions?
8. What are the null and alternative hypotheses?

Chapter 4, Section 1 – one quantitative variables

1. What changed about the studies (data structure) from Chapters 2 & 3?
2. What is the statistic of interest now? What is the parameter of interest?
3. What is the difference between the distribution of the data and the distribution of the statistic? There is a theoretical difference as well as a computational difference.
4. What is the limiting sampling distribution of the statistic? (Note, the answer here is for big samples, that is the Central *Limit* Theorem works only where there is a limit... i.e., the sample size is big.)
5. If interest is in a statistics other than the sample mean, what is a tool we can use for finding the alternative statistic's sampling distribution?
6. Explain the intuition behind bootstrapping.
7. Explain how the SE for the statistic is calculated using bootstrapping.
8. What is the difference between a normal distribution and a t distribution?
9. When do we use a z and when do we use a t?
10. When would you use a confidence interval and when would you use a hypothesis test?
11. What different information does a boxplot give versus a histogram?

Chapter 4, Section 3 – 2 means

1. What changed about the studies (data structure) from section 4.1?
2. What is the statistic of interest now? What is the parameter of interest?
3. What is the sampling distribution for the statistic of interest?
4. How is the t-distribution become relevant?
5. What are degrees of freedom in general? What are the actual degrees of freedom for the test in section 4.3?
6. How is the null mechanism different across the three analysis methods in section 3.2: randomization test, two-sample t-test, random sampling test (n.b. this is also called the parametric bootstrap)?
7. How do you create a CI? How do you interpret the CI?
8. What if your data are NOT normal? What strategies can you try out?

Chapter 4, Section 2 – paired samples

1. What changed about the studies (data structure) in section 4.2 as compared with 4.1 or 4.3?
2. What is the statistic of interest now? What is the parameter of interest?
3. What is the sampling distribution for the statistic of interest?
4. What benefit does pairing have on the analysis?
5. What happens if a paired study is analyzed as if it were an independent two sample study? (What happens to the p-value? What happens to the CI?)
6. What is the easiest way to think of / analyze paired data?

Chapter 4, Section 4 – ANOVA

1. Why are these tests called ANalysis Of VAriance (ANOVA)?
2. Describe the variability in the numerator and the variability in the denominator. What does each measure?
3. What are the null and alternative hypotheses for ANOVA?
4. What features of the data affect the power of the test? What does power mean here?
5. What are the technical conditions? Why do we need equal variances here?

Chapter 5, Sections 1-4 – correlation & simple linear regression

1. Describe the linear model with multiple variables.
2. Describe the error / residual term and how it is calculated with multiple variables.
3. What are the (three-ish) statistics of interest in this chapter? What are the parameters of interest?
4. What does correlation measure?
5. How do we find the values of b_0 and b_1 for estimating the least squares line?
6. Why is it dangerous to extrapolate?
7. How do we interpret R^2? Why is that?
8. What does it mean to say that b_1 has a sampling distribution? Why is it that we would never talk about the sampling distribution of β_1?

9. Why do we need the LINE technical conditions for the inference parts of the analysis but not for the estimation parts of the analysis?

10. Is linear regression always appropriate when comparing two continuous variables?

11. What are the LINE technical conditions? How are the conditions assessed?

12. What are the three factors that influence the $SE(b_1)$? (Note: when something influences $SE(b_1)$, that means the inference is also effected. If you have a huge $SE(b_1)$, it will be hard to tell if the slope is significant because the t value will be small.)

13. What does it mean to do a randomization test for the slope? That is, explain the process of doing a randomization test here. (See shuffle options in the Analyzing Two Quantitative Variables applet.)

14. Why would someone transform either of the variables?

15. What is the difference between a confidence interval and a prediction interval? Which is bigger? Why does that make sense? How do the centers of the intervals differ? (They don't. Why not?)

Chapter 6, Sections 1-3 – multiple regression

1. Describe the linear model with multiple variables.

2. Describe the error / residual term and how it is calculated with multiple variables.

3. How does the model change when multiple variables are included?

4. How are p-values interpreted now that there are multiple variables?

5. How is R^2 interpreted? What is the difference between R^2 and R^2_{adj}?

6. How are variables chosen for the final model?

7. How are the model conditions assessed?

Chapter 6, Section 4 – logistic regression

1. The main difference between logistic regression and linear regression is the response variables. Explain that difference.

2. Why isn't it appropriate to use linear regression when the response variable is binary?

3. How is the b_1 coefficient interpreted? (Hint: it has something to do with odds ratios.)

4. What does it mean for b_1 to have a sampling distribution?

5. How is the significance of the coefficient measured?

6. What are the statistics of interest in this section? What are the parameters of interest?