Practice Exam 2

Name: Sample Solutions

You may use a calculator and two 8.5" by 11" sheets of notes (front and back). Please plan on bringing your own calculator.

Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Note: This practice test will give you a general sense of the set up for exam 2. However, this practice test is not as similar to the real test as our practice quizzes have been to the real quizzes. I will ask some types of questions on the real exam that do not appear on this exam. You should also review our old labs and homework assignments 6 and 7.

I Conceptual Questions

Please answer the following in a few sentences each. (You don't need to write multiple paragraphs.)

1. Suppose I conduct a hypothesis test about the average amount of tea in a Twinings tea bag. The null hypothesis is that the population mean amount of tea is (less than or) equal to 2.5 grams, and the alternative hypothesis is that the population mean amount of tea is larger than 2.5 grams. In this context, what would a Type I error be? If I change the significance level of the test, α, from 0.05 to 0.01, how does that affect the probability of making a Type I error in this test?

A type I error would be concluding that the population mean amount of tea is larger than 2.5 grams when in fact it is (less than or) equal to 2.5 grams.

If the null hypothesis is true, the probability of making a Type I error is equal to the significance level of the test. Changing the significance level of the test from 0.05 to 0.01 reduces the probability that we will make a Type I error, if the null hypothesis is true.

2. A study found that the use of bed nets was associated with a lower prevalence of malarial infections in the Gambia, relative to the prevalence of infections with commonly used mosquito control practices there. A report of the study states that the significance is $p < 0.001$. Explain what this means in a way that could be understood by someone who has not studied statistics. You may use the word "probability", but you can't use any formulas or other statistical jargon like "statistically significant", "null hypothesis", or similar.

If in fact bed nets had no effect on the prevalence of malarial infections in the Gambia, there would be less than a $\frac{1}{10}$ % chance of seeing an effect at least as large as what we saw in this sample.

Ideas:

- p-value says what the probability of getting a sample statistic at least as extreme as the one we actually observed would be, if the null hypothesis was true.
- The answer above restates this definition of a p-value in the context of this example

3. Suppose I'm studying cedar waxwings (a type of bird common throughout North America), and I want to estimate how much weight the birds lose during migration. I capture 1000 cedar waxwing birds in Mexico during the winter months, measure their weights, and band them with a unique identifying tag. Then, at the beginning of the summer months I attempt to recapture those same birds in Canada and measure their weights again as they are finishing their migration. I plan to treat the data as paired, and I will construct a 95% confidence interval for the population mean difference in weights before and after migration using the weight differences for the sample of birds I am able to recapture. Are the following statements true or false?

   (a) If I am only able to recapture 10 of the birds again in Canada (so that $n = 10$), my 95% confidence interval will be **wider** than it would have been if I had been able to recapture all 1000 birds (so that $n = 1000$).

If the sample size is large, we get more information about the population, narrow down the possible values for the parameter.

(True) False

   (b) If I am only able to recapture 10 of the birds again in Canada (so that $n = 10$), my 95% confidence interval will be **less likely to contain the population mean weight difference** than it would have been if I had been able to recapture all 1000 birds (so that $n = 1000$).

True (False)

No matter the sample size, for 95% of samples a 95% C.I. calculated based on that sample will contain the population parameter.

4. In a t test about a population mean, what is the test statistic? Address the following points:

   (a) Write down the formula for the test statistic

   (b) Write a sentence or two describing what the statistic measures

   (c) If the test statistic has a very large value, is that strong evidence against the null hypothesis or weak evidence against the null hypothesis? Explain why, with reference to your answer to (b).

The test statistic is $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$, where

$\bar{y}$ is the sample mean, μ_0 is the value of the population mean specified in the null hypothesis, s is the sample standard deviation, and n is the sample size.

The statistic measures how far away the sample mean is from the hypothesized population mean, in units of standard errors of the sample mean. (The standard error of the sample mean is an estimate of the standard deviation of the sample mean.)

If the test statistic is very large, that means our sample mean was many standard errors away from the hypothesized population mean. Since we expect the sample mean to fall relatively close to the true population mean, a large value of the test statistic is strong evidence against the null hypothesis.

II Applied Problems

1. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

   (a) Describe the population parameter of interest and the sample statistic. What symbol is commonly used for these?

The population parameter is the proportion of the whole graduating class who found a job within a year of graduating. This is denoted by p.
The sample statistic is the number of people in the ~~graduating class~~ sample who had found jobs. We observed X=348 in this sample.

   (b) Check whether the conditions for constructing a confidence interval and conducting a hypothesis test based on these data are met. (For each condition, write a complete sentence answering whether or not that condition is met and why.)

Bias: We would need to know that our survey respondents were representative of the whole class, which is difficult to assess with the given information. For example, our results would be biased if people who had jobs were more likely to respond to the survey than people who didn't.

Categorical Variable:
For each survey respondent, we recorded a categorical variable: whether or not they had found a job.

Sample statistic is a count:
Our sample statistic is a count of the number of people in our sample who had found a job.

Independence: Since we are told this was a random sample, there is no reason to suspect that there is a connection between the different people in our sample. Independence seems OK.

Overall, let's proceed cautiously and remember to note the possibility of bias in our discussion of our results.

(c) I ran some R code and found that a 95% confidence interval for the population parameter was [0.833, 0.901]. Interpret the interval **in the context of this problem**. As part of your answer, describe what the phrase "95% confident" means.

We are 95% confident that the proportion of people in this graduating class who found jobs within a year is between 0.833 and 0.901. If we were to take many different samples from this class and calculate a different 95% confidence interval based on each of those samples about 95% of those confidence intervals would contain the true proportion of the graduating class who found jobs within a year.

(d) According to the National Center for Education Statistics, the proportion of all 20-to-24 year olds with college degrees who were employed as of 2015 is 0.89. Write down a statement of the null and alternative hypotheses for a test that the employment rate for the current graduating class is different from the national average among 20 to 24 year olds.

$H_0: p = 0.89$

The proportion of people in this class who find a job within a year is equal to the national employment rate among 20-to-24 year olds.

$H_A: p \neq 0.89$

The proportion of people in this class who found a job within a year is different from the national employment rate among 20-to-24 year olds.

(e) I used R to compute a p-value for this test, and I got a p-value of 0.2. **In the context of this problem**, what is your conclusion? Use a significance level of $\alpha = 0.05$.

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. The data from our sample do not provide enough evidence to conclude that the proportion of people in this class who find a job within a year of graduating is different from the employment rate among 20-to-24 year olds nationally.

(f) How could you have conducted this test using the confidence interval from part c?

Since the confidence interval of [0.883, 0.901] contains the value 0.89 from the null hypothesis we fail to reject the null hypothesis.
It was a 95% confidence interval, so this was a test at the significance level $\alpha = 0.05$.
The conclusion would be the same as discussed in part (e).

2. (16 points) Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 micrograms/liter and a SD of 37.74 micrograms/liter; a previous study of a large number of individuals (not all police officers) from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 micrograms/liter.

   (a) What is the population parameter of interest? Describe it in a sentence and state what symbol is commonly used for this parameter.

The population parameter is the mean concentration of lead in blood of police officers working traffic enforcement in this area. The symbol μ will be used to denote this parameter.

   (b) Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a higher concentration of lead than their neighbors in the suburbs. You may treat the value of 35 micrograms/liter from the suburban study as a fixed, known constant (i.e., we are not structuring this as a test to compare the means of two groups, but rather as a test about the mean value for the group of urban police officers).

$H_0: \mu = 35$

The mean lead concentration for police officers in this region is the same as the mean lead concentration for police officers in the nearby suburb.

$H_A: \mu > 35$

The mean lead concentration is higher for police officers in this area than it is for police officers in the nearby suburb.

(c) Explicitly state and check all conditions necessary for inference on these data. If you don't have enough information, say what you would want to know. If you would want to look at any plots, describe what plot(s) you would want to make and what you'd be looking for.

Bias: We need to know that the police officers in our sample are representative of police officers in this region more generally. It is difficult to be sure with the information we are given, but there are no specific reasons to be concerned in this case

Quantitative: for each officer in our sample, we measured the lead concentration in their blood. This is a quantitative variable.

Mean reasonable?

We would need to look at a plot of the data to check this condition. I would want to see a histogram or density plot of the blood lead concentrations for the officers in our sample, and confirm that it was unimodal, roughly symmetric, and had no outliers.

Minor asymmetry or an outlier that wasn't too severe would be OK, as long as they weren't too bad, since our sample size is 52.

Independent: We would need to be sure that there were no specific connections in the blood lead concentrations for different police officers in our sample. This seems to be OK, as there is no reason to suspect such a connection might exist.

Overall, it seems like the conditions are fairly well satisfied - but it does make me nervous when I can't look at a plot of the data!!

(d) Calculate a p-value for the test that the downtown police officers have a higher lead exposure than the group in the previous study. **For full credit, you must show all of your work!** Points are assigned to correct set-up. In doing this calculation, you may use the following facts (you just need one of these numbers - which one?):

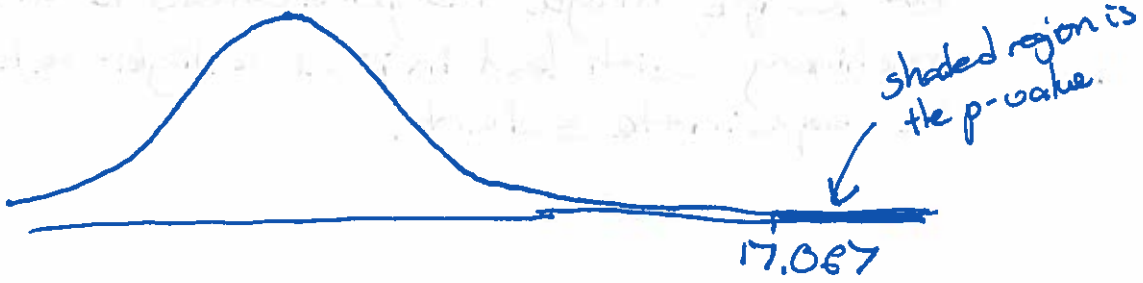
- If $T \sim t_{51}$, then $P(T > 23.754) < 10^{-16}$
- If $T \sim t_{51}$, then $P(T > 17.067) < 10^{-16}$
- If $T \sim t_{51}$, then $P(T > 3.294) = 0.001$
- If $T \sim t_{51}$, then $P(T > 2.367) = 0.011$

Our test statistic is $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{(124.32 - 35)}{37.74/\sqrt{52}}$

$= 17.067$

If the null hypothesis is correct, this statistic follows a t_{51} distribution, since our sample size is 52.
Since the alternative hypothesis was $\mu > 35$, the p-value is the probability of getting a test statistic larger than 17.067, if H_0 is true. This is $< 10^{-16}$, from the circled information above.

(e) Draw a picture of a relevant t distribution. Label it with the value of the test statistic you found in part (c) and shade in the region corresponding to the p-value.



(f) Interpret the results of your test in context. What is your conclusion?

Since the p-value is less than commonly used significance levels such as 0.05 and 0.01, we can reject the null hypothesis and conclude that police officers in this location have higher blood lead concentration than police officers in the nearby suburb.

(g) Suppose that you rejected the null hypothesis (which you may or may not have actually done based on the data above - this is a hypothetical question). Would this prove that there was a causal relationship between exposure to car exhaust and increased concentrations of lead in the blood? Explain why or why not.

No, rejecting the null hypothesis does not prove that there is a causal ~~case~~ relationship between exposure to car exhaust and increased concentrations of lead in the blood. This was an observational study, and there could be other lurking variables that are not accounted for in this analysis. For example, maybe the police officers are handling machinery with lead in it at a higher rate than the people in the suburb.

(h) Find a 99% confidence interval for the population parameter, and interpret it in context (including a description of what the phrase "99% confident" means.) You may use the following R output to help in your calculation (you only need one of these numbers - which one?)

```
> qt(0.995, df = 51)
[1] 2.676
> qt(0.995, df = 52)
[1] 2.674
> qt(0.99, df = 51)
[1] 2.402
> qt(0.99, df = 52)
[1] 2.4
```

$\bar{y} \pm t^* s/\sqrt{n}$

$124.32 \pm 2.676 \cdot \frac{37.74}{\sqrt{52}}$

[110.315, 138.325]

We are 99% confident that the mean concentration of lead in the blood of police officers working on traffic enforcement in this location is between 110.315 micrograms/liter and 138.325 micrograms/liter.

If we were to take many different samples and calculate a 99% confidence interval based on each of those samples, about 99% of those confidence intervals would contain the true mean blood lead concentration among the population of police officers working on traffic enforcement in this location.