Lab 6 Guide

Disclaimer: This does by no means guarantee complete success in Lab 6, it is merely intended as a guide for some of you who are lost, I can not provide you with any code, but I will point to functions that should be used.

Think of the questions in a similar way you would have in STK110, but instead of working on paper with a pen and calculator, you use R as your calculator.

Important functions:

- `qnorm(probability)`
  - `qnorm()` accepts a value a probability value in the domain [0, 1], and returns the corresponding Z-score calculated with the standard normal distribution.
- `pnorm(z_score)`
  - `pnorm()` accepts a standardized (meaning it has it follows the standard normal distribution) Z value, and it returns the corresponding lower-tail probability value (unless specified otherwise as a function argument) calculated with the standard normal distribution (unless other values for the mean and standard deviation are given)
- `subset()`
- `sum() & length()`

I have made the assumption that you are familiar with `subset(), sum(), length()` by now, however, if you would like to find out more about these functions, you can inquire about them in R, by typing the function name preceded by a ? in R, e.g. `?qnorm`.

Calculating proportions and/or point estimates:

Proportions will be calculated in R by counting the number of successes and dividing it by the sample size, using R. In this lab a success is when a person is an atheist.
i.e. `data$response == "atheist"`

Counting can be done using the `sum()` function or `length()` function, along with the `subset()` function.

Below is a snapshot taken from the Lab, on how to create the subset:

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
```

Source: **https://tinyurl.com/y4qodrvb**

Full URL can be found in references section

Not all questions will be discussed/discussed in detail.

Q1)

b) For question (1b) it is important to recall the formula for the margin of error.

$MOE = Z * SE$

Z being the z score, and SE being the standard error.

The standard error can be obtained by implementing the formula for the standard error in R, i.e. $\sqrt{\frac{p(1-p)}{n}}$

p being the proportion/point estimate you calculated in 1a, and n being the sample size.

The information given below is important for, and applies to all the questions

Z: your z-score can be obtained through the `qnorm()` function that has been described above. It is important to note that since this is a two-tailed test you should divide your significance level by 2, e.g. if your confidence level is 80% (0.8), you significance level would be 20% (0.2), and therefore the value you would pass to `qnorm()` is 0.2/2 = 0.1. Or 0.9 if you pass the upper tail value, make sure that your acquired MOE is positive, you might get a negative value if you use the lower tail probability (0.1) instead of the upper tail probability (this is due to the nature of the normal distribution)

After obtaining these values, you can use it to calculate the MOE.

c) The bounds can be calculated by using the point estimate and the margin of error i.e. $bounds = point\ estimate\ \pm MOE$

Q2)

a) The MOE can be calculated in a similar fashion to the one you have calculated in Q1b, but instead of using the empirical proportion/point estimate, use the hypothesized proportion given to you.

b) The test statistic can be calculated using the usual formula for the test statistic that we were initially taught in STK110, i.e.

$$\frac{point\ estimate - hypothesised\ value}{standard\ error} \equiv \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$\hat{p}$ is the same point estimate that has been calculated in question 1.

p_0 is the given hypothesized value.

And the standard error can be calculated with the same formula as question 1, but using p_0 instead of the point estimate.

c) The p-value can be calculated using the `pnorm()` function that has been described above, and your calculated test statistic. Take note that since this is a two-tailed test, your obtained p-value should be multiplied by 2. I encourage you to consult the lecture notes/textbook/STK110 notes or online resource to understand why this is done.

Q3)

a) In this question you will set up a confidence interval about the difference between two proportions. As can be found in the lecture notes, the point estimate would now be: $point\ estimate = p1 - p2$

the lab 6 submissions instructions to see what data sets should be used to calculate these proportions, the proportions can be calculated in same manner as the proportion in q1a.

b) b)The margin of error still follows the same formula as in Q1 i.e. $MOE = Z * SE$

The Z score can be calculated in the exact same was as question one, however, a different formula is used when calculating the standard error for difference in proportions: $standard\ error = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

When these two values have been obtained, you can proceed with your calculation of the margin of error.

c) The lower and upper bounds can be calculated in the same as in question 1.

Q4)

a) The proportions can be kept the same as in Q3, however, there are some differences when conducting an hypotheses test as opposed to constructing a confidence interval, the biggest one being the calculation of the standard error.

The formula stays the same as the one in Q3, however, in this case, since we are conducting a hypotheses test to test whether or not two proportions are equal, you do not use p_1 and p_2 in your calculations, instead you used a pooled estimate.

$$p_{pooled} = \ pooled\ estimate = \frac{total\ \#\ successes\ over\ both\ samples}{total\ size\ of\ both\ samples}$$

If you do not understand why we use a pooled estimate for this hypothesis test, I encourage you to consult the lecture notes/videos for Dr van Staden's explanation.
After you have calculated your pooled estimate, you can plug in the values into the formula as follows.

$$standard\ error = \sqrt{\frac{p_{pooled}(1-p_{pooled})}{n_1} + \frac{p_{pooled}(1-p_{pooled})}{n_2}}$$

After the standard error has been calculated, you can use the formula for the test statistic value for the difference between proportions, to calculate your answer.

Formula for test statistic for difference in proportions: $\frac{(p_1-p_2)-p_0}{standard\ error}$

Where p_0 is the hypothesized difference (Keep in mind this is a test for equal proportions)

b) The p-value can be calculated in the same way as in question 3, remember to keep in mind that this is also a two-tailed test.

Q5)

For question 5 you can use the `inference()` function given in the open intro lab.

Below is a snapshot of the function for those of you without access to the website.

```
inference(us12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

Source: https://tinyurl.com/y4qodrvb

References:

Openintro website for Lab 6:
http://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs-base-R/blob/master/inf_for_categorical_data/inf_for_categorical_data.html