Practicum 3: Hypothesis testing and confidence intervals

Harvard Catalyst Certificate in Applied Biostatistics

Goal: This practicum will go over how to complete one sample and two sample t-tests in R.

Worked example

**One sample t-test**

Please load the dataset from Practicum 1 into R by following the steps from Practicum 1.

Based on your prior knowledge, you know that the general population has a mean score of 50. We would like to understand if the mean score in RRMS patients is the same as the general population. Although we anticipate that the mean will only be lower, we will use a two-sided hypothesis test since this is by far the most common in the medical literature. The steps for a hypothesis test are:

1. State null hypothesis
2. State type of data
3. Determine appropriate statistical test
4. State summary statistics if possible
5. Calculate p-value (stat package)
6. Decide whether to reject or not reject the null hypothesis
7. Write conclusion

Before we look at the data, we can complete the first three steps of our hypothesis test:

1. State null and alternative hypothesis: H0: =50, HA: 
2. State type of data: Continuous outcome, single sample
3. Determine appropriate statistical test: one sample t-test

Remember that we set up the null hypothesis so that we can reject it. In this case, we would like to know if RRMS patients are different than the general population so we set it up assuming that they are the same.

R will allow us to calculate our summary statistics, confidence interval and p-value using a single command, *t.test*. For the one sample t-test, we need to specify the variable of interest (blue arrow) as well as the value for the null hypothesis, 50 (red arrow):

*t.test(data\_prac1$mcs,mu=50)*

> t.test(data\_prac1$mcs,mu=50)

One Sample t-test

data: data\_prac1$mcs

t = -1.228, df = 250, **p-value = 0.2206**

**alternative hypothesis: true mean is not equal to 50**

95 percent confidence interval:

**48.14726 50.42966**

sample estimates:

mean of x

**49.28846**

The output presented here provides the sample mean (blue arrow: 49.3). The output also presents the 95% confidence interval for the mean (red arrow: 48.1, 50.4). In addition to this information, R provides the two-sided alternative hypothesis (green arrow), which corresponds to the p-value (*p=0.22,* orange arrow).

With this information, we can complete our hypothesis test.

1. State summary statistics if possible: 
2. Calculate p-value (stat package): *p=0.22*
3. Decide whether to reject or not reject the null hypothesis: Fail to reject H0
4. Write conclusion: Based on our sample, there is no statistically significant difference between the mean in RRMS patients and the mean in general population.

**Two sample t-test**

Another question of interest based on this initial dataset is whether the mental quality of life as measured by MCS is different in male vs. female RRMS patients. We assume that the male and female RRMS patients in our study represent a random sample of the male and female RRMS patients.

As always, we can complete the first three steps of the hypothesis test prior to looking closely at the data.

1. State null hypothesis:
   1. , 
   2. , 
2. State the type of data: Continuous outcome, dichotomous predictor
3. Determine appropriate statistical test: Two sample t-test

Remember that we set up the null hypothesis so that we can reject it. In this case, we would like to know if males are different compared to females so we set it up assuming that they are the same. As before, R will allow us to calculate our summary statistics and p-value using a single command, *t.test*. For the two sample t-test, we need to specify the outcome variable (blue arrow) as well as the group variable (red arrow). Note that the default in R is to fit an unequal variance t-test. Since we would like to run this type of t-test, no additional options are required.

*t.test(data\_prac1$mcs~data\_prac1$male)*

> t.test(data\_prac1$mcs~data\_prac1$male)

Welch Two Sample t-test

data: data\_prac1$mcs by data\_prac1$male

t = 1.5371, df = 93.824, p-value = 0.1276

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.6150422 4.8314837

sample estimates:

mean in group 0 mean in group 1

49.77562 47.66740

The statistic of interest here is the difference in the sample means because we can write our null hypothesis as the difference between the groups is equal to 0 (option b. above). The output presented here provides the means in the two groups, from which you can calculate the difference (blue arrows: 49.8-47.7 = 2.1). The output also presents the 95% confidence interval for the difference in the mean (red arrow: -0.62, 4.83). In addition to this information, R provides the two-sided alternative hypothesis (green arrow), which corresponds to the p-value (*p=0.13,* orange arrow).   
  
If we had completed the analysis assuming equal variances, we would have arrived at the same conclusion. Please try this on your own.

*t.test(data\_prac1$mcs~data\_prac1$male,var.equal=T)*

To learn about the additional options for any R command, please type ? prior to the command name. For example, to learn about the additional options for *t.test* type:

*?t.test*

Try on your own

Using the class dataset, formulate a hypothesis that could be tested using a two sample t-test and perform the hypothesis using R.