Lab 6: More hypothesis testing

STAT218

The objective of this lab is to learn how to perform -tests for a population mean in R. We will cover:

* specifying null and alternative hypotheses
* interpreting R output

## Datasets

Throughout this lab we’ll use the datasets:

* sleep are 135 observations of average hours of sleep per night from NHANES respondents

You’ll need to run the commands below in order to load these datasets before proceeding with the activity.

library(oibiostat)  
  
data(nhanes.samp.adult)  
sleep <- nhanes.samp.adult$SleepHrsNight  
str(sleep)

int [1:135] 9 3 6 6 6 6 6 4 7 4 ...

By way of review, you’ll recall from last time that hypothesis tests are performed in R using t.test().

# do US adults sleep either more or less than 8 hours a night? [result: yes]  
t.test(sleep, mu = 8, alternative = 'two.sided')

One Sample t-test  
  
data: sleep  
t = -9.1966, df = 134, p-value = 6.328e-16  
alternative hypothesis: true mean is not equal to 8  
95 percent confidence interval:  
 6.658933 7.133659  
sample estimates:  
mean of x   
 6.896296

# do US adults sleep more than 8 hours a night? [result: no]  
t.test(sleep, mu = 8, alternative = 'greater')

One Sample t-test  
  
data: sleep  
t = -9.1966, df = 134, p-value = 1  
alternative hypothesis: true mean is greater than 8  
95 percent confidence interval:  
 6.69752 Inf  
sample estimates:  
mean of x   
 6.896296

# do US adults sleep less than 8 hours a night? [result: yes]  
t.test(sleep, mu = 8, alternative = 'less')

One Sample t-test  
  
data: sleep  
t = -9.1966, df = 134, p-value = 3.164e-16  
alternative hypothesis: true mean is less than 8  
95 percent confidence interval:  
 -Inf 7.095073  
sample estimates:  
mean of x   
 6.896296

|  |
| --- |
| Choosing your alternative wisely |
| It can be tricky to pick alternatives in directional tests. Suppose you want to test whether adults sleep less than 8 hours. Should you use an upper-sided or a lower-sided test?  Consider the difference in interpretation between the upper-sided () and lower-sided () tests above:   * [upper sided test fails to reject] the data *do not* provide evidence that U.S. adults *don’t* sleep less then 8 hours a night * [lower sided test rejects] the data provide evidence that U.S. adults don’t sleep more than 8 hours a night, and favor instead the alternative that they sleep less than 8 hours a night   Both answers are consistent with adults sleeping less than 8 hours, but they frame the question in different ways, and the results are not logically equivalent: the former is a double negative and basically inconclusive; the latter is a stronger conclusion.  If you want to answer the question whether adults sleep less than 8 hours, the better way to do it is with a lower-sided test, so that you potentially find evidence for the claim of interest (should you reject the upper-sided test) rather than potentially fail to find evidence against the claim of interest.  The best rule of thumb for choosing the null and alternative in a direction test is: *the null should be the claim you hope to refute, and the alternative should be the claim you hope to support with evidence*. |

## Reporting test results

Here you’ll practice reporting test results. You’ll work with a group on just *one* of the questions below, and your objective is to write a complete report of the test outcome:

* hypotheses tested
* test conclusion, interpreted in context
* test statistic, degrees of freedom, -value
* confidence interval and point estimate, interpreted in context

These elements should be summarized together in complete sentences; see the lecture slides for a complete example. Be sure to consider how to frame the hypotheses appropriately to answer the question (see remark above on choosing alternatives).

1. Do US adults sleep 7.5 hours per night on average?
2. Do US adults sleep less than 7.5 hours per night on average?
3. Do US adults sleep more than 7.5 hours per night on average?
4. Do US adults sleep more than 6.5 hours per night on average?

|  |
| --- |
| Your turn |
| Work with your group to answer one of the four questions above. Use significance level to make a decision.  # 1. Do US adults sleep 7.5 hours per night on average?  # 2. Do US adults sleep less than 7.5 hours per night on average?  # 3. Do US adults sleep more than 7.5 hours per night on average?  # 4. Do US adults sleep more than 6.5 hours per night on average?  Write a complete report of the test results. |

## Changing the confidence level

You will have noticed by now that t.test produces a confidence interval. The interval calculated is designed to match the test in the following sense:

* a lower-sided test produces a lower confidence bound
* an upper-sided test produces an upper confidence bound
* a two-sided test produces the usual interval

In each case, the critical value is chosen to provide a specific coverage. It will *not* be the same in each formula above, even for the same coverage level.

**The test rejects at level just in case the interval excludes the null value** . To change the confidence level, add the argument conf.level = ... to the t.test function:

# a 99% interval  
t.test(sleep, mu = 8, alternative = 'two.sided', conf.level = 0.99)

One Sample t-test  
  
data: sleep  
t = -9.1966, df = 134, p-value = 6.328e-16  
alternative hypothesis: true mean is not equal to 8  
99 percent confidence interval:  
 6.582703 7.209890  
sample estimates:  
mean of x   
 6.896296

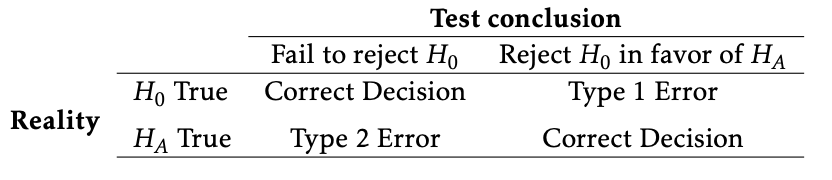
# an 80% interval  
t.test(sleep, mu = 8, alternative = 'two.sided', conf.level = 0.8)

One Sample t-test  
  
data: sleep  
t = -9.1966, df = 134, p-value = 6.328e-16  
alternative hypothesis: true mean is not equal to 8  
80 percent confidence interval:  
 6.741733 7.050860  
sample estimates:  
mean of x   
 6.896296

|  |
| --- |
| Your turn |
| Compute an upper 87% confidence bound for the mean nightly hours of sleep for U.S. adults. Interpret the interval in context.  # fill in the '...' parts to obtain an 87% upper confidence bound t.test(sleep,   mu = ...,  alternative = ...,  conf.level = ...) |

## Exploring decision errors

There are two ways to make an error in a hypothesis test.



We can explore type I error rates by testing a hypothesis known to be true; we can explore type II error rates by testing a hypothesis known to be false. To do so, let’s again pretend that the following 3,179 NHANES survey responses are a population.

load('data/nhanes.RData')  
  
# "true" population mean  
pop\_mean <- mean(nhanes$TotChol)

### Type I errors

A type I error is rejecting a true null hypothesis. To explore the rate of type I errors, let’s test the following:

5.043 is the *true population mean*, so in point of fact is true and we should *not* reject; any rejections are therefore type I errors. We’ll all generate a sample from 3,179 observations of total HDL cholesterol. Each of us will obtain a different sample. Using our respective samples, we’ll each test whether the population mean is 5.043 and see how many of us produce an erroneous conclusion. This will provide a sense of the error rate.

To start, run these commands:

# draw a sample   
samp <- sample(nhanes$TotChol, size = 20)  
  
# test a true null  
t.test(samp, mu = pop\_mean, alternative = 'two.sided')

Recall that test decisions depend on the rule: reject if . Let’s tally type I errors for different significance levels :

| Significance level | Error frequency |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

We should see that the type I error rate is approximately equal to the significance level. The significance level is, in fact, a cap on type I error.

### Type II errors

Now let’s test whether the population mean is some . In this case, will be false, and a correct decision is to reject . Any failures to reject will be considered type II errors.

Try running the following.

# draw a sample   
samp <- sample(nhanes$TotChol, size = 20)  
  
# test a false null  
t.test(samp,   
 mu = 4.2, # change this for exercise  
 alternative = 'two.sided')

The error rate will depend on how far the null hypothesis is from the truth. Presumably, if we test that , we’ll have a high type II error rate because the null value is so close to the truth; this hypothesis would be very hard to refute. If, on the other hand, we test , we whould have a low type II error rate because the null value is far from the truth. Let’s use significance level throughout and tally errors for several different null values:

| Null value | Error frequency |
| --- | --- |
| 4.2 |  |
| 4.6 |  |
| 4.9 |  |
| 5.1 |  |
| 5.4 |  |
| 5.7 |  |

Notice that the type II error is quite high for null values near the true mean; this is because the test prioritizes avoiding type I errors, and as a result has little power to detect such alternatives.