Lab 6: 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 three datasets:

* ddt are the 15 measurements from lecture of DDT level (ppm) in kale
* body.temps are 130 observations of body temperature from a sample of US adults
* sleep are 135 observations of average hours of sleep per night from NHANES respondents

library(oibiostat)  
  
ddt <- MASS::DDT  
str(ddt)

num [1:15] 2.79 2.93 3.22 3.78 3.22 3.38 3.18 3.33 3.34 3.06 ...

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

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

data(thermometry)  
body.temps <- thermometry$body.temp  
str(body.temps)

num [1:130] 96.3 96.7 96.9 97 97.1 97.1 97.1 97.2 97.3 97.4 ...

## Performing a test in R

We’ll illustrate the use of t.test with the ddt data. The parameter of interest here is:

### Upper-sided test

Consider first testing the hypotheses:

The null parameter value is the cutoff point . This is an *upper-sided* test because the alternative specifies that the parameter *exceeds* the null value. We implement this test by providing as arguments to t.test:

* data vector ddt
* null parameter value mu = 3
* upper sided alternative alternative = 'greater'

# upper sided test H0: mean ddt <= 3 vs HA: mean ddt > 3  
t.test(ddt, mu = 3, alternative = 'greater')

One Sample t-test  
  
data: ddt  
t = 2.9059, df = 14, p-value = 0.005753  
alternative hypothesis: true mean is greater than 3  
95 percent confidence interval:  
 3.129197 Inf  
sample estimates:  
mean of x   
 3.328

Take a moment to locate each of the following outputs:

* test statistic value
* degrees of freedom for the model
* -value
* confidence interval for the mean
* point estiamte for the mean

Since , we reject at significance level . Interpret the result:

The data provide sufficiently strong evidence to reject the null hypothesis that mean DDT in kale is at most 3 ppm in favor of the alternative hypothesis that mean DDT in kale exceeds 3ppm (p = 0.00575).

### Lower-sided test

Now suppose you wish to test whether the mean DDT in kale is at least 3.5ppm.

The null parameter value is the cutoff point . This is a *lower-sided* test because the alternative specifies that the parameter *is smaller than* the null value. We implement this test by providing as arguments to t.test:

* null parameter value mu = 3.5
* upper sided alternative alternative = 'less'

# lower-sided test, H0: mean ddt >= 3.5 vs HA: mean ddt < 3.5  
t.test(ddt, mu = 3.5, alternative = 'less')

One Sample t-test  
  
data: ddt  
t = -1.5238, df = 14, p-value = 0.07491  
alternative hypothesis: true mean is less than 3.5  
95 percent confidence interval:  
 -Inf 3.526803  
sample estimates:  
mean of x   
 3.328

Since , we fail to reject at significance level . Interpret the result:

The data **do not** provide sufficiently strong evidence to reject the null hypothesis that mean DDT in kale is at least 3.5 ppm in favor of the alternative hypothesis that mean DDT in kale exceeds 3.5ppm (p = 0.075).

### Two-sided test

Lastly, suppose we wish to test whether mean DDT is 3 or not.

The null parameter value is the point . This is a *two-sided* test because the alternative specifies that the parameter *is either greater or smaller than* the null value. We implement this test by providing as arguments to t.test:

* null parameter value mu = 3
* upper sided alternative alternative = 'two.sided'

# two-sided test, H0: mean ddt == 3 vs HA: mean ddt =!= 3  
t.test(ddt, mu = 3, alternative = 'two.sided')

One Sample t-test  
  
data: ddt  
t = 2.9059, df = 14, p-value = 0.01151  
alternative hypothesis: true mean is not equal to 3  
95 percent confidence interval:  
 3.085913 3.570087  
sample estimates:  
mean of x   
 3.328

Since , we reject at significance level . Interpret the result:

The data provide sufficiently strong evidence to reject the null hypothesis that mean DDT in kale is 3 ppm in favor of the alternative hypothesis that mean DDT in kale is not 3ppm (p = 0.012).

Notice that, the evidence against the null is slightly weaker with respect to the two-sided alternative than with respect to the upper-sided alternative. This makes sense, because the alternative comprises a larger range of values, some of which are not very consistent with the data.

## Answering questions with tests

### Body temperatures

Answer the following questions with hypothesis tests using the body.temps data. Be sure to consider how to frame the hypotheses appropriately to answer the question.

1. Is the mean body temperature different from 98.6 °F?
2. Is the mean body temperature higher than 98 °F?
3. Is the mean body temperature higher than 98.2 °F?
4. Is the mean body temperature lower than 98.2 °F?
5. Is the mean body temperature actually 98.2 °F?
6. Is the mean body temperature actually 98.3 °F?

|  |
| --- |
| Your turn |
| Write commands to perform tests to answer questions 1-6 above.  # 1. Is the mean body temperature different from 98.6 °F?  # 2. Is the mean body temperature higher than 98 °F?  # 3. Is the mean body temperature higher than 98.2 °F?  # 4. Is the mean body temperature lower than 98.2 °F?  # 5. Is the mean body temperature actually 98.2 °F?  # 6. Is the mean body temperature actually 98.3 °F?  Interpret the result of each test in context. Use significance level to make a decision. |

|  |
| --- |
| Remark |
| The hypothesis tests suggest that in fact, mean body temperature is lower than commonly thought. However, a visual inspection of the data reveals an interesting twist: the body temperature data are actually bimodal!    This doesn’t invalidate any of our inferences, but suggests that we might be asking the wrong question by focusing on the population mean if, in fact, there is no *one* mean body temperature. We should perhaps ask instead, what feature of the population explains the bimodal distribution? |

### Hours of sleep

Answer the following questions with hypothesis tests using the sleep data. Be sure to consider how to frame the hypotheses appropriately to answer the question.

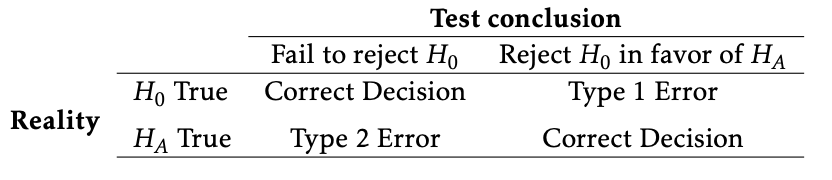
1. Do US adults sleep 7.5 hours per night on average?
2. Do US adults sleep less than 7.5 hours per night on average?
3. Do US adults sleep more than 7.5 hours per night on average?
4. Do US adults sleep more than 6.5 hours per night on average?

|  |
| --- |
| Your turn |
| Write commands to perform tests to answer questions 1-4 above.  # 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?  Interpret the result of each test in context. Use significance level to make a decision. |

## Exploring decision errors

If there is time in class, we’ll explore decision errors a little. If not, you can simulate this activity by repeatedly running the commands and making tallies on your own.

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



### Type I errors

First we’ll all generate a sample from 3,179 observations of total HDL cholesterol, pretending that the full set of observations constitutes a population. 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.

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.

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

One Sample t-test  
  
data: samp  
t = -0.4551, df = 19, p-value = 0.6542  
alternative hypothesis: true mean is not equal to 5.042938  
95 percent confidence interval:  
 4.586964 5.336036  
sample estimates:  
mean of x   
 4.9615

Recall that the significance level (denoted ) determines the decision rule: we reject if . Let’s tally errors as follows:

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

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

# 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')

One Sample t-test  
  
data: samp  
t = 4.745, df = 19, p-value = 0.0001409  
alternative hypothesis: true mean is not equal to 4.2  
95 percent confidence interval:  
 4.654665 5.372335  
sample estimates:  
mean of x   
 5.0135

Let’s use significance level throughout and tally errors as follows:

| 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.