Lab 6 - Hypothesis Testing and the T-Test

Point estimates and confidence intervals are basic inference tools that act as the foundation for another inference technique: statistical hypothesis testing. Statistical hypothesis testing is a framework for determining whether observed data deviates from what is expected. R contains an array of built in functions that make it easy to carry out hypothesis tests and analyze the results.

1. Hypothesis Testing Basics

Statistical hypothesis tests are based a statement called the null hypothesis that assumes nothing interesting is going on between whatever variables you are testing. The exact form of the null hypothesis varies from one type test to another: if you are testing whether groups differ, the null hypothesis states that the groups are the same. For instance, if you wanted to test whether the average age of voters in your home state differs from the national average, the null hypothesis would be that there is no difference between the average ages.

The purpose of a hypothesis test is to determine whether the null hypothesis is likely to be true given sample data. If there is little evidence against the null hypothesis given the data, you accept the null hypothesis. If the null hypothesis is unlikely given the data, you might reject the null in favor of the alternative hypothesis: that something interesting is going on. The exact form of the alternative hypothesis will depend on the specific test you are carrying out. Continuing with the example above, the alternative hypothesis would be that the average age of voters in your state does in fact differ from the national average.

Once you have the null and alternative hypothesis in hand, you choose a significance level (often denoted by the Greek letter α.). The significance level is a probability threshold that determines when you reject the null hypothesis. After carrying out a test, if the probability of getting a result as extreme as the one you observe due to chance is lower than the significance level, you reject the null hypothesis in favor of the alternative. This probability of seeing a result as extreme or more extreme than the one observed is known as the p-value.

The t-test is a statistical test used to determine whether a numeric data sample of differs significantly from the population or whether two samples differ from one another.

2. Testing a population parameter

Consider a simple survey. You ask 100 people (randomly chosen) and 42 say ``yes" to your question. Does this support the hypothesis that the true proportion is 50% ?

To answer this, we set up a test of hypothesis. The *null hypothesis*, denoted H_0 is that $p = .5$, the *alternative hypothesis*, denoted H_A, in this example would be $p \neq 0.5$. This is a so called ``two-sided" alternative. To test the assumptions, we use the function `prop.test` as with the confidence interval calculation. Here are the commands

```
> prop.test(42,100,p=.5)

     1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3233236 0.5228954
sample estimates:
   p
0.42
```

Note the p-value of 0.1336. The p-value reports how likely we are to see this data *or worse* assuming the null hypothesis. The notion of worse, is implied by the alternative hypothesis. In this example, the alternative is two-sided as too small a value or too large a value or the test statistic is consistent with H_A. In particular, the p-value is the probability of 42 or fewer *or* 58 or more answer ``yes" when the chance a person will answer ``yes" is fifty-fifty.

Now, the p-value is not so small as to make an observation of 42 seem unreasonable in 100 samples assuming the null hypothesis. Thus, one would ``accept" the null hypothesis.

Next, we repeat, only suppose we ask 1000 people and 420 say yes. Does this still support the null hypothesis that $p=0.5$?

```
> prop.test(420,1000,p=.5)

     1-sample proportions test with continuity correction

data:  420 out of 1000, null probability 0.5
X-squared = 25.281, df = 1, p-value = 4.956e-07
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:

 0.3892796 0.4513427

sample estimates:

   p

0.42
```

Now the *p*-value is tiny (that's 0.0000004956!) and the null hypothesis is not supported. That is, we ``reject" the null hypothesis. This illustrates the the *p* value depends not just on the ratio, but also *n*. In particular, it is because the standard error of the sample average gets smaller as *n* gets larger.

One-Sample T-Test

A one-sample t-test checks whether a sample mean differs from the population mean. Let's create some dummy age data for the population of voters in the entire country and a sample of voters in Minnesota and test the whether the average age of voters Minnesota differs from the population:

```
In [1]:
set.seed(12)
population_ages <- c(rexp(1000000,0.015)+18,    # Generate dummy age data for
population
                     rpois(500000,20)+18,
                     rpois(500000,32.5)+18,
                     rpois(500000,45)+18)

population_ages <- ifelse(population_ages<100, population_ages, population_ag
es%%100+18)


true_mean <- mean(population_ages)           # Check the population mean
true_mean

set.seed(12)
minnesota_ages <- c(rexp(100,0.015)+18,      # Generate dummy sample age data
                     rpois(50,15)+18,
                     rpois(50,25)+18,
                     rpois(50,35)+18)

minnesota_ages <- ifelse(minnesota_ages<100, minnesota_ages, minnesota_ages%%
100+18)
```

```
mean(minnesota_ages)
```

Out[1]:

51.2188371860945

Out[1]:

48.2348502354331

Notice that we used a slightly different combination of distributions to generate the sample data for Minnesota, so we know that the two means are different.

Let's conduct a t-test at a 95% confidence level and see if it correctly rejects the null hypothesis that the sample comes from the same distribution as the population. To conduct a t-test, we can use the same t.test() function we used last time to find confidence intervals:

In [2]:

```
t.test(x = minnesota_ages,      # Sample data
       mu = true_mean,          # The true population mean
       alternative = "two.sided",    # Conduct two sided test*
       conf.level = 0.95,       # Desired level of statistical significance
       )
```

Out[2]:

```
 One Sample t-test

data:  minnesota_ages
t = -2.4484, df = 249, p-value = 0.01504
alternative hypothesis: true mean is not equal to 51.21884
95 percent confidence interval:
 45.8345 50.6352
sample estimates:
mean of x 
 48.23485 
```

*Note: the alternative hypothesis can be that the sample mean is strictly less than, strictly greater than or not equal to the population parameter. For the "not equal to" hypothesis, we use a "two sided" test because an extreme test result in either direction would be evidence that the sample mean is significantly different from the population mean.

The test result shows the test statistic "t" is equal to -2.4484. This test statistic tells us how much the sample mean deviates from the null hypothesis. If the t-statistic lies outside the quantiles of the t-distribution corresponding to our confidence level and degrees of freedom, we reject the null hypothesis. We can check the quantiles with qt():

In [3]:

```
qt(p=0.025, df=249)    # Get the lower tail quantile
qt(p=0.975, df=249)    # Get the upper tail quantile
```

Out[3]:

-1.96953686764035

Out[3]:

1.96953686764035

Furthermore, we can calculate the chances of seeing a result as extreme as the one we observed (known as the p-value) by passing the t-statistic in as the quantile to the pt() function:

In [4]:

```
pt(q=-2.4484, df=249) * 2  # We multiply by 2 because we are doing a two-tail
ed test
```

Out[4]:

0.0150399648189317

Notice this value is the same as the p-value listed in the original t-test output. A p-value of 0.01504 means we'd expect to see data as extreme as our sample due to chance about 1.5% of the time if the null hypothesis was true. In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so we should reject the null hypothesis. Also notice that the 95% confidence interval in the output not does capture the true population mean of 51.2188.

Let's run the same test but change our desired confidence level to 99%:

In [5]:

```
t.test(x = minnesota_ages,
       mu = true_mean,
       alternative = "two.sided",
       conf.level = 0.99,           # Use a higher confidence level
       )
```

Out[5]:

```
 One Sample t-test

data:  minnesota_ages
t = -2.4484, df = 249, p-value = 0.01504
alternative hypothesis: true mean is not equal to 51.21884
99 percent confidence interval:
 45.07134 51.39836
sample estimates:
mean of x
 48.23485
```

Now let's calculate the upper and lower quantile bounds for a 99% confidence level:

In [6]:

```
qt(p=0.005, df=249)    # Get the lower tail quantile
qt(p=0.995, df=249)    # Get the upper tail quantile
```

Out[6]:

-2.59571775827349

Out[6]:

2.59571775827349

With a higher confidence level, we construct a wider confidence interval and increase the chances that it captures to true mean, thus making it less likely that we'll reject the null hypothesis. In this case, the test statistic -2.4484 falls within the quantile bounds for our test so the p-value of 0.015 is greater than our significance level of 0.01 and we fail to reject the null hypothesis.

Two-Sample T-Test

A two-sample t-test investigates whether the means of two independent data samples differ from one another. In a two-sample test, the null hypothesis is that the means of both groups are the same. Unlike the one sample-test where we test against a known population parameter, the two sample test only involves sample means. You can conduct a two-sample t-test by passing a second data sample into the t.test() function. Let's generate a sample of voter age data for Wisconsin and test it against the sample we made earlier:

In [7]:

```
set.seed(12)
wisconsin_ages <- c(rexp(50,0.015)+18,      # Generate more dummy sample age data
                    rpois(50,20)+18,
                    rpois(50,32.5)+18,
                    rpois(50,45)+18)

wisconsin_ages <- ifelse(wisconsin_ages<100, wisconsin_ages, wisconsin_ages%%100+18)
```

In [8]:

```
t.test(x=minnesota_ages,    # Conduct a the two sample test*
       y=wisconsin_ages,
       conf.level=0.95)
```

Out[8]:

```
 Welch Two Sample t-test

data:  minnesota_ages and wisconsin_ages
t = -2.2201, df = 442.089, p-value = 0.02692
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.2257932 -0.4397954
sample estimates:
mean of x mean of y 
 48.23485  52.06764 
```

*Note: the degrees of freedom for a two-sample test are derived from a formula based on the size and variance of each sample intended to correct for samples with unequal variance.

The test yields a p-value of 0.02692, which means there is a 2.7% chance we'd see sample data this far apart if the two groups tested are actually identical. In this case, we'd reject that hypothesis, since 2.7% is lower than our significance level of 5%.
If we changed our confidence level to 99% we would fail to reject the null hypothesis because the p-value is greater than 0.01.

Paired T-Test

The two sample t-test is designed for testing differences between independent groups. In some cases, you might be interested in testing differences between samples of the same group at different points in time. For instance, a hospital might want to test whether a weight-loss drug works by checking the weights of the same group patients before and after treatment. A paired t-test lets you check whether the means of samples from the same group differ.

We can conduct a paired t-test by passing two paired data samples to the t.test() function and including the argument paired=TRUE. Let's generate some dummy weight data and run a paired t-test on it. Note that R creates pairings based on the order of the vectors you pass in, so individuals should be in the same order in both vectors.

In [9]:

```
set.seed(80)

before_treatment_weights <- rnorm(100,250,30)    # Generate weights with mean 
250lbs

after_treatment_weights <- (before_treatment_weights + rnorm(100,-1.25,5))

weight_df <- data.frame(before=before_treatment_weights, # Pair the data in a 
data frame
                         after=after_treatment_weights,
                         change=after_treatment_weights-before_treatment_weig
hts)

summary(weight_df)              # Check a summary of the data
```

Out[9]:

```
     before           after           change       
 Min.   :160.4   Min.   :159.4   Min.   :-10.998  
 1st Qu.:231.9   1st Qu.:230.0   1st Qu.: -5.209  
 Median :249.0   Median :247.7   Median : -1.506  
 Mean   :249.7   Mean   :248.1   Mean   : -1.562  
 3rd Qu.:269.0   3rd Qu.:266.2   3rd Qu.:  1.719  
 Max.   :349.4   Max.   :348.3   Max.   : 12.045  
```

The summary shows that patients lost about 1.2 pounds on average after treatment. Let's conduct a paired t-test to see whether this difference is significant at a 95% confidence level:

In [10]:

```
t.test(before_treatment_weights,
       after_treatment_weights,
       conf.level= 0.95,
       paired=TRUE)
```

Out[10]:

```
 Paired t-test

data:  before_treatment_weights and after_treatment_weights
t = 3.1686, df = 99, p-value = 0.002038
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.583922 2.540406
sample estimates:
mean of the differences
               1.562164
```

The p-value in the test output shows that the chances of seeing this large of a difference between samples due to chance is less than 1%.

Type I and Type II Error

The result of a statistical hypothesis test and the corresponding decision of whether to reject or accept the null hypothesis is not infallible. A test provides evidence for or against the null hypothesis and then you decide whether to accept or reject it based on that evidence, but the evidence may lack the strength to arrive at the correct conclusion. Incorrect conclusions made from hypothesis tests fall in one of two categories: type I error and type II error.

Type I error describes a situation where you reject the null hypothesis when it is actually true. This type of error is also known as a "false positive" or "false hit". The type 1 error rate is equal to the significance level α, so setting a higher confidence level (and therefore lower alpha) reduces the chances of getting a false positive.

Type II error describes a situation where you fail to reject the null hypothesis when it is actually false. Type II error is also known as a "false negative" or "miss". The higher your confidence level, the more likely you are to make a type II error.

You can find the type II error rate for detecting a difference between a given distribution and a second test distribution with a known mean and standard deviation using the power.t.test() function. This function finds the power of a t-test, which is probability that the test rejects the null hypothesis when the alternative is true. The type II error rate is equal to 1-power. Let's use this function to check the type II error rate of our paired t-test:

```
In [11]:
power.t.test(n= 100,                # Size of the sample
             delta = 1.25,          # Assumed mean (avg difference) for the dist
ribution
             sd= 5,                 # Assumed standard deviation for the distrib
ution
             sig.level= 0.05,       # Significance level
             type="paired")         # t-test type
```

```
Out[11]:
     Paired t test power calculation 

              n = 100
          delta = 1.25
             sd = 5
      sig.level = 0.05
          power = 0.6969757
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

The output shows that the power of our paired t-test is 0.6969757, so the type II error is 1-0.6969757 or approximately 30% . Since the true difference between the groups is only 1.25 pounds we have a fairly high probability of failing to detect the difference.

Wrap Up

The t-test is a powerful tool for investigating the differences between sample and population means and R has nice built in functions for conducting t-tests.

Appendix: P-value approach

The *P*-value approach involves determining "likely" or "unlikely" by determining the probability — assuming the null hypothesis were true — of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed. If the *P*-value is small, say less than (or equal to) α, then it is "unlikely." And, if the *P*-value is large, say more than α, then it is "likely."

If the *P*-value is less than (or equal to) α, then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the *P*-value is greater than α, then the null hypothesis is not rejected.

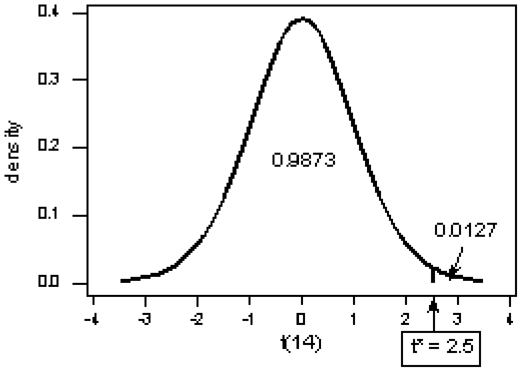
Specifically, the four steps involved in using the *P*-value approach to conducting any hypothesis test are:

1. Specify the null and alternative hypotheses.

2. Using the sample data and assuming the null hypothesis is true, calculate the value of the test statistic. Again, to conduct the hypothesis test for the population mean μ, we use the t-statistic $t^* = \dfrac{\overline{x}-\mu}{s\sqrt{n}}$ which follows a t-distribution with n - 1 degrees of freedom.
3. Using the known distribution of the test statistic, calculate the ***P*-value**: "If the null hypothesis is true, what is the probability that we'd observe a more extreme test statistic in the direction of the alternative hypothesis than we did?" (Note how this question is equivalent to the question answered in criminal trials: "If the defendant is innocent, what is the chance that we'd observe such extreme criminal evidence?")
4. Set the significance level, α, the probability of making a Type I error to be small — 0.01, 0.05, or 0.10. Compare the *P*-value to α. If the *P*-value is less than (or equal to) α, reject the null hypothesis in favor of the alternative hypothesis. If the *P*-value is greater than α, do not reject the null hypothesis.

In our example concerning the mean grade point average, suppose that our random sample of $n = 15$ students majoring in mathematics yields a test statistic t^* equaling 2.5. Since $n = 15$, our test statistic t^* has $n - 1 = 14$ degrees of freedom. Also, suppose we set our significance level α at 0.05, so that we have only a 5% chance of making a Type I error.

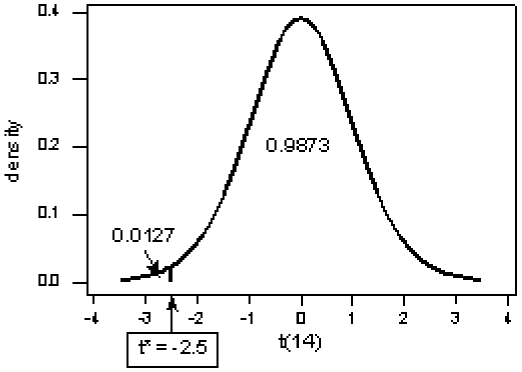
The *P*-value for conducting the **right-tailed** test $H_0 : \mu = 3$ versus $H_A : \mu > 3$ is the probability that we would observe a test statistic greater than $t^* = 2.5$ if the population mean μ really were 3. Recall that probability equals the area under the probability curve. The *P*-value is therefore the area under a $t_{n-1} = t_{14}$ curve and to the *right* of the test statistic $t^* = 2.5$. It can be shown using statistical software that the *P*-value is 0.0127:



The *P*-value, 0.0127, tells us it is "unlikely" that we would observe such an extreme test statistic t^* in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the *P*-value, 0.0127, is less than α = 0.05, we reject the null hypothesis $H_0 : \mu = 3$ in favor of the alternative hypothesis $H_A : \mu > 3$.

Note that we would not reject $H_0 : \mu = 3$ in favor of $H_A : \mu > 3$ if we lowered our willingness to make a Type I error to α = 0.01 instead, as the *P*-value, 0.0127, is then greater than α = 0.01.

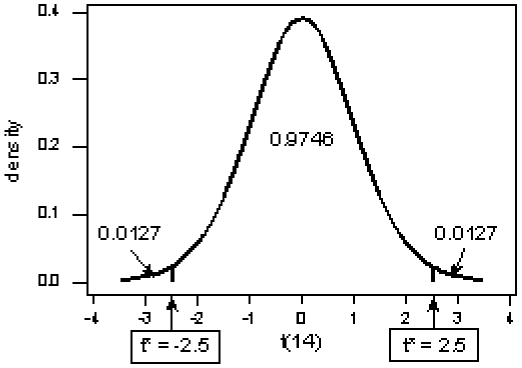
In our example concerning the mean grade point average, suppose that our random sample of $n = 15$ students majoring in mathematics yields a test statistic t^* instead equaling -2.5. The *P*-value for conducting the **left-tailed** test $H_0 : \mu = 3$ versus $H_A : \mu < 3$ is the probability that we would observe a test statistic less than $t^* = -2.5$ if the population mean μ really were 3. The *P*-value is therefore the area under a $t_{n-1} = t_{14}$ curve and to the *left* of the test statistic t* = -2.5. It can be shown using statistical software that the *P*-value is 0.0127:



The *P*-value, 0.0127, tells us it is "unlikely" that we would observe such an extreme test statistic *t** in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the *P*-value, 0.0127, is less than α = 0.05, we reject the null hypothesis $H_0 : \mu = 3$ in favor of the alternative hypothesis $H_A : \mu < 3$.

Note that we would not reject $H_0 : \mu = 3$ in favor of $H_A : \mu < 3$ if we lowered our willingness to make a Type I error to α = 0.01 instead, as the *P*-value, 0.0127, is then greater than α = 0.01.

In our example concerning the mean grade point average, suppose again that our random sample of *n* = 15 students majoring in mathematics yields a test statistic *t** instead equaling -2.5. The *P*-value for conducting the **two-tailed** test $H_0 : \mu = 3$ versus $H_A : \mu \neq 3$ is the probability that we would observe a test statistic less than -2.5 or greater than 2.5 if the population mean μ really were 3. That is, the two-tailed test requires taking into account the possibility that the test statistic could fall into either tail (and hence the name "two-tailed" test). The *P*-value is therefore the area under a $t_{n-1} = t_{14}$ curve to the *left* of -2.5 and to the *right* of the 2.5. It can be shown using statistical software that the *P*-value is 0.0127 + 0.0127, or 0.0254:



Note that the *P*-value for a two-tailed test is always two times the *P*-value for either of the one-tailed tests. The *P*-value, 0.0254, tells us it is "unlikely" that we would observe such an extreme test statistic *t** in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the *P*-value, 0.0254, is less than α = 0.05, we reject the null hypothesis $H_0 : \mu = 3$ in favor of the alternative hypothesis $H_A : \mu \neq 3$.

Note that we would not reject $H_0 : \mu = 3$ in favor of $H_A : \mu \neq 3$ if we lowered our willingness to make a Type I error to α = 0.01 instead, as the *P*-value, 0.0254, is then greater than α = 0.01.

Now that we have reviewed the critical value and *P*-value approach procedures for each of three possible hypotheses, let's look at three new examples — one of a right-tailed test, one of a left-tailed test, and one of a two-tailed test.