T-testing Example

D2K Course Staff

January 17 2019

Data

Below, we show an example of how to run two sample t-tests as well as two sample proportion tests. This is done on the `iris` data set, which contains data on 3 different species of iris flowers.

```
data(iris)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

In this data set, `Species` is a categorical variable, while the sepal and petal lengths and widths are continuous variables. For this data set, we will be comparing the `setosa` and `viginica` species, so we create subsets of the data:

```
library(dplyr)

iris_setosa <- dplyr::filter(iris, Species == 'setosa')
iris_virginica <- dplyr::filter(iris, Species == 'virginica')
```

Two-Sample T-test

The two-sample t-test is used to compare the means of two groups of samples. For example, say we are interested in testing if the mean sepal length is the same for the `setosa` and `viginica` species in the `iris` data set. Formally, we are running a hypothesis test with:

H_0: The mean sepal lengths for the `setosa` and `viginica` species are the same.

H_a: The mean sepal lengths for the `setosa` and `viginica` species are different.

The assumptions for the t-test are:

- The population follows a normal distribution.
- Samples are independent.
- The samples are representative of the population.

The two-sample t-test can be run under the assumption that the population variances are equal between the two groups, or not. (The former is also known as a "pooled variance" test while the latter is known as an "unpooled variance" test.) The results of both are shown below:

```
# Equal variance
t.test(iris_setosa$Sepal.Length, iris_virginica$Sepal.Length,
       var.equal = TRUE)
```

```
    Two Sample t-test

data:  iris_setosa$Sepal.Length and iris_virginica$Sepal.Length
t = -15.386, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.786042 -1.377958
sample estimates:
mean of x mean of y
    5.006     6.588
```

```
# Unequal variance
t.test(iris_setosa$Sepal.Length, iris_virginica$Sepal.Length,
       var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  iris_setosa$Sepal.Length and iris_virginica$Sepal.Length
t = -15.386, df = 76.516, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.78676 -1.37724
sample estimates:
mean of x mean of y
    5.006     6.588
```

Notice that the confidence intervals are slightly different between the two tests. However, in both of these cases, the listed p-value is extremely small, meaning that there is a significant difference between the sepal lengths for the `setosa` and `viginica` species. (Note: in the real world, only one test should be chosen and should be done so before the test is run.)

Two-Sample Proportion Test

The two-sample proportion test is used to compare the proportions of a binary categorical variable of two groups of samples. For example, say we want to test the hypothesis:

H_0: The proportions of sepal widths greater than 3 cm for the `setosa` and `viginica` species are the same.

H_a: The proportions of sepal widths greater than 3 cm for the `setosa` and `viginica` species are different.

The assumptions for two-sample proportions test are:

- Samples are not a substantial proportion of the entire population for both groups.
- Samples are independent.
- There are a sufficient number of successes and failures in each group.

First, we have to calculate the propotions:

```
# Calculate proportions
prop.test(c(sum(iris_setosa$Sepal.Width > 3), sum(iris_virginica$Sepal.Width > 3)),
          c(nrow(iris_setosa), nrow(iris_virginica)))
```

```
    2-sample test for equality of proportions with continuity
    correction

data:  c(sum(iris_setosa$Sepal.Width > 3), sum(iris_virginica$Sepal.Width >  out of c(nrow(iris_setosa)
X-squared = 23.811, df = 1, p-value = 1.062e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.313969 0.686031
sample estimates:
prop 1 prop 2
  0.84   0.34
```

The listed p-value is extremely small, meaning that there is a significant difference between the proportions of flower sepal widths that are above 3 cm for the `setosa` and `viginica` species.