IFN501 - System Modeling and Simulation

Session 6: Introduction to Statistics

Daniel Febrian Sengkey

Department of Electrical Engineering
Faculty of Engineering
Universitas Sam Ratulangi

Outline

t-Test

- t-Test is one of several commonly used test methods.
- t-Test is based on the t distribution.
- There are several types of t-test
  - One-sample t-test
  - Two-sample t-test
    - Paired sample
    - Independent sample
- Please note that this is a very short description about t-test. You should read more to gain better understanding about this method.

t-Test

One-Sample t-Test

This case was taken from `https://www.r-bloggers.com/one-sample-students-t-test/`

- There are results of intelligence test in 10 subjects: 65, 78, 88, 55, 48, 95, 66, 57, 79, 81.
- The average result of the population which received the same test is 75.
- Let us check if the sample mean above is <u>significantly similar</u> with the population or not.
- Use 95% significance level.
- First, let us assign the scores to a variable:

```
scores <- c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
```

- Then we use the `t.test(data, mean)` function.

```
t.test(scores, mu = 75)
```

t-Test

One-Sample t-Test

```r
t.test(scores, mu = 75)
```

```
## 
##  One Sample t-test
## 
## data:  scores
## t = -0.783, df = 9, p-value = 0.454
## alternative hypothesis: true mean is not equal to 75
## 95 percent confidence interval:
##  60.2219 82.1781
## sample estimates:
## mean of x 
##      71.2
```

- There are 2 ways to interpret the result:
  - Using the value of t calculated.
  - Comparing the p-value with the significance level.
- P-value is the easier way.
- Since the significance level is 95% then the α value is 0.05.
- If the p-value is higher than the α value, the we must accept the null hypothesis: the average of the test scores is significantly similar with population average, otherwise we accept the alternate hypothesis.
- In our case, the p-value is 0.453721, which is higher than the α value, therefore we accept the null hypothesis.

- Beside checking the similarity between the average of some samples to a certain number, t-test also can be used to compare averages of 2 datasets.
- As described earlier, in this context there are 2 types of test:
  - Testing similarity between 2 matched samples. In this type of test, there are 2 datasets from repeated observations of the same subject.
  - Testing similarity between 2 independent samples. Here the samples came from different populations and each sample is not affecting each other.
- Let explore the examples from R-Tutor for matched samples[1] and independent samples[2].

[1] http://www.r-tutor.com/elementary-statistics/inference-about-two-populations/population-mean-between-two-matched-samples

[2] http://www.r-tutor.com/elementary-statistics/inference-about-two-populations/population-mean-between-two-independent-samples

- We use the built-in dataset immer, the barley yield in 1931 and 1932 of the same field.
- The data are presented in the data frame columns Y1 and Y2.

```
library(MASS)
head(immer)

##   Loc Var    Y1    Y2
## 1  UF   M  81.0  80.7
## 2  UF   S 105.4  82.3
##  [ reached getOption("max.print") -- omitted 4 rows ]
```

- **Problem:** Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields between years 1931 and 1932.
- To solve this problem in R, we use the paired test by setting the `paired` argument as `TRUE`.

```
t.test(immer$Y1, immer$Y2, paired = TRUE)
```

```
t.test(immer$Y1, immer$Y2, paired = TRUE)

##
##  Paired t-test
##
## data:  immer$Y1 and immer$Y2
## t = 3.324, df = 29, p-value = 0.00241
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.12195 25.70471
## sample estimates:
## mean of the differences
##                 15.9133
```

- The p-value in the function output is less than our α value[3].
- Therefore we have a strong evidence to reject null hypothesis and accept the alternate hypothesis: the yields of years 1931 and 1932 are significantly different.

[3] 0.05 since we used 95% confidence interval

t-Test

Two-Sample t-Test: Independent Samples

- We use the dataframe column `mpg`[4] of the built-in dataset mtcars.
- On the other hand, the `am` data column from the same data set indicates the transmission types of the automobile model (0 = automatic, 1 = manual).

```
mtcars[, c("mpg", "am")]

##                    mpg am
## Mazda RX4         21.0  1
## Mazda RX4 Wag     21.0  1
## Datsun 710        22.8  1
## Hornet 4 Drive    21.4  0
## Hornet Sportabout 18.7  0
##  [ reached getOption("max.print") -- omitted 27 rows ]
```

- Particulary, these 2 columns are independent data population.
- **Problem:** Assuming that the data in mtcars follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean gas mileage of manual and automatic transmissions.

[4] Gas mileage data of various 1974 U.S. automobiles.

- First, we must split the data into 2 set of data, one for the automatic transmission model, and one for the manual transmission model[5].

```
L = mtcars$am == 0  # select the automatic transmission model (0)
mpg.auto = mtcars[L, ]$mpg  # select automatic transmission mileage
mpg.auto

##  [1] 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3
##  [ reached getOption("max.print") -- omitted 9 entries ]
```

- The gas mileage for manual transmission can be found by using the negation of L.

```
mpg.manual = mtcars[!L, ]$mpg
mpg.manual

##  [1] 21.0 21.0 22.8 32.4 30.4 33.9 27.3 26.0 30.4 15.8
##  [ reached getOption("max.print") -- omitted 3 entries ]
```

[5] See a tutorial for data frame row slice here.

t-Test

Two-Sample t-Test: Independent Samples

- After having data for both models in separated variables, now we can use t-test to compute difference of means between the automatic and manual transmission models.

```
t.test(mpg.auto, mpg.manual)

##
##  Welch Two Sample t-test
##
## data:  mpg.auto and mpg.manual
## t = -3.767, df = 18.33, p-value = 0.00137
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28019  -3.20968
## sample estimates:
## mean of x mean of y
##   17.1474   24.3923
```

- The result shows that the p-value is lower than the predefined α, therefore we can reject H_0 and conclude that the mileages of automatic transmission and the manual transmission are significantly different.

- In addition to your knowledge, the `t.test()` function in R also support formula interface.
- Our latest case can also be solved by using

```
t.test(mpg~am, data=mtcars)

##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.00137
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28019  -3.20968
## sample estimates:
## mean in group 0 mean in group 1
##         17.1474         24.3923
```

Analysis of Variance

- t-Test is limited only to 2 samples.
- To test the difference between 3 or more samples we use the Analysis of Variance (ANOVA) method.
- The basic package in R already has this function.
- As exercise for today we use the case in `http://www.r-tutor.com/elementary-statistics/analysis-variance/completely-randomized-design`.
  - The case is about a fast food franchise that testing the market for 3 menu items.
  - The management put these new items in 18 franchisee restaurants, 6 for each restaurant.
  - The new items are randomly allocated.
  - After a week, the sales for each item is:

Item1	22, 42, 44, 52, 45, 37
Item2	52, 33, 8, 47, 43, 32
Item3	16, 24, 19, 18, 34, 39

- **Problem:** At .05 level of significance, test whether the mean sales volume for the 3 new menu items are all equal.

Analysis of Variance

The solution includes several steps of data preparation.

1. First, let's create a data frame to store the data.

```
df1 <- data.frame(
        Item1 = c(22, 42, 44, 52, 45, 37),
        Item2 = c(52, 33,  8, 47, 43, 32),
        Item3 = c(16, 24, 19, 18, 34, 39)
)
```

2. Check the content

```
df1

##   Item1 Item2 Item3
## 1    22    52    16
## 2    42    33    24
## 3    44     8    19
## 4    52    47    18
## 5    45    43    34
## 6    37    32    39
```

Analysis of Variance

3. Concatenate the data rows of `df1` into a single vector `r`.

```
r = c(t(as.matrix(df1)))  # response data
r

##  [1] 22 52 16 42 33 24 44  8 19 52
##  [ reached getOption("max.print") -- omitted 8 entries ]
```

4. Assign new variables for the treatment levels and number of observations.

```
f = c("Item1", "Item2", "Item3")  # treatment levels
k = 3  # number of treatment levels
n = 6  # observations per treatment
```

5. Create a vector of treatment factors that corresponds to each element of `r` in step 3 with the `gl()` function.

```
tm = gl(k, 1, n * k, factor(f))  # matching treatments
tm

##  [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
## [10] Item1
##  [ reached getOption("max.print") -- omitted 8 entries ]
## Levels: Item1 Item2 Item3
```

Analysis of Variance

6. Apply the function `aov` to a formula that describes the response `r` by the treatment factor tm and assign it to a new variable `av`.

```
av <- aov(r ~ tm)
```

7. Print out the ANOVA table using the `summary()` function.

```
summary(av)

##             Df Sum Sq Mean Sq F value Pr(>F)
## tm           2    745     373    2.54   0.11
## Residuals   15   2200     147
```

8. The p-value in the output is greater than the significance level ($\alpha = 0.05$). Therefore we accept the null hypothesis: The mean sales volume of the new menu items are equal.

Non-parametric Statistics

- t-Test and ANOVA are parts of statistics methods known as parametric methods.
- The usage of parametric methods is highly advised.
- However, there are several assumptions that have to be satisfied before using the parametric methods.
- Distribution normality is one of the assumptions in parametric methods.
- Unfortunately, not all data that normally distributed.
- If we can not satisfy the assumption for parametric methods, then we can use the non-parametric methods.

Non-parametric Statistics

- The examples of non-parametric methods and their parametric counterparts are:
  - 1-sample Wilcoxon test for 1-sample t-test
  - Mann-Whitney test for 2-sample t-test
  - Kruskal-Wallis for One-way ANOVA
- To check whether the data is normally distributed we can use:
  - Shapiro-Wilk test
  - Kolmogorov-Smirnov test
- R already has all these tests, you just need to browse the Internet. There a lot of tutorials there.

Non-parametric Statistics

Some Tutorials

- Normality Test
- Wilcoxon Signed Rank Test
- Mann-Whitney-Wilcoxon Test
- Kruskal-Wallis Test

After learned some statistical tests and know that these tests have assumptions that have to be fulfilled, we can write down steps for data analysis:

1. Check whether the data is normally distributed or not.
2. If the data is normally distributed, go with the parametric methods.
3. If the data is NOT normally distributed, we have to use the non-parametric methods.

Next session..

- Cellular Automaton

References I