Statistical Hypothesis Testing with R

This tutorial is an attempt to demonstrate how to perform statistical hypothesis testing with R.

t-Tests

Download the cardio dataset from eLearn@USM. Load the dataset into R and name the data frame as cardio. The dataset contains medical data of 70000 patients such as blood pressure, glucose reading and cholesterol.

```
> cardio <- read.table('cardio.csv', sep=';', header=T)
```

Let's test if the population mean weight of patients with well above glucose reading and patients with normal glucose reading is equal. Select the patients' weights with well above glucose reading and normal glucose reading.

```
> weight_gluc_1 <- cardio$weight[cardio$gluc==1]
> weight_gluc_3 <- cardio$weight[cardio$gluc==3]
```

We randomly select 250 patients

```
> index <- round(runif(250,1,2000))
> weight_gluc_1 <- weight_gluc_1[index]
> weight_gluc_3 <- weight_gluc_3[index]
```

Assuming the variance of the population is similar, perform the independent t-test as follows.

```
> t.test(weight_gluc_1, weight_gluc_3, var.equal=T)
```

Based on the results, the p-value is 0.0003 which is less than the significance level (0.05). Hence, we can reject the null hypothesis. Note that, you might get a different result because the samples are randomly selected.

```
        Two Sample t-test

data:  weight_gluc_1 and weight_gluc_3
t = -3.6368, df = 498, p-value = 0.0003048
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.574875 -2.261125
sample estimates:
mean of x mean of y
   73.650    78.568
```

Now, let's perform another test. The claim is that the difference in population mean weight between patients with well above glucose reading and patients with normal glucose reading is equal and greater than 4 kg. To perform the independent t-test, we specify the argument mu and alternative to 4 and 'less' respectively. The argument alternative indicate the alternative hypothesis of the hypothesis test.

```
> t.test(weight_gluc_3, weight_gluc_1, mu=4, var.equal=T,
  alternative='less')
```

```
        Two Sample t-test

data:  weight_gluc_3 and weight_gluc_1
t = 0.67885, df = 498, p-value = 0.7512
alternative hypothesis: true difference in means is less than 4
95 percent confidence interval:
     -Inf 7.146446
sample estimates:
mean of x mean of y
   78.568    73.650
```

The p-value is 0.7512 which is greater than the significance level. Thus, we cannot reject the null hypothesis.

Download the spider-anxiety ("spider.long.csv") dataset from eLearn@USM. Load the dataset into R. Name the data frame as spider. The dataset contains measures of anxiety of 12 subjects when they were shown real tarantula spider and picture of the same tarantula spider.

It is hypothesized that the sample means are the same. Let's perform independent t-test first on the dataset. Since the data is stored in a dataframe, we perform the t-test as follows. Note, that paired argument is FALSE by default.

```
> t.test(Anxiety ~ Group, spider, var.equal=T)
```

We will get an output as follows.

```
        Two Sample t-test

data:  Anxiety by Group
t = -1.6813, df = 22, p-value = 0.1068
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.634222   1.634222
sample estimates:
    mean in group Picture mean in group Real Spider
                       40                       47
```

As we can see, the p-value is 0.1068 which is more than 0.05. Thus, we cannot reject the null hypothesis.

Now, let's perform a paired t-test on the dataset.

```
> t.test(Anxiety ~ Group, spider.long, paired=TRUE)

        Paired t-test

data:  spider.wide$real and spider.wide$picture
t = 2.4725, df = 11, p-value = 0.03098
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.7687815 13.2312185
sample estimates:
mean of the differences
                      7
```

As we can see, the p-value is 0.03098 which is less than 0.05. Thus, we can reject the null hypothesis. Paired t-test is more appropriate for the dataset since the experiment involves the same subjects.

ANOVA

Download medicine dataset from eLearn@USM and load the dataset into R.

Explore the dataset using summary() function and visualization such as plotting the scatter plot of the data.

We use aov() to perform ANOVA test. The general form of aov() is as follows.

aov(outcome ~ predictor(s), data = dataFrame, na.action = an action)

where

- outcome is the variable that you're trying to predict, also known as the dependent variable.
- predictor(s) lists the variable or variables from which you're trying to predict the outcome variable, also known as the independent variable(s). To add more predictor, simply write "+ variable_name" e.g. aov(outcome ~ predictor1 + predictor2, dataFrame)
- dataFrame is the name of the dataframe from which your outcome and predictor variables come.
- na.action is an optional command – if you have complete data you can ignore it, but if you have missing values then it can be used to exclude NAs (na.action = na.exclude)

To perform ANOVA test:

```
> medModel <- aov(effect ~ dose, medicine)
```

The statement generate the model that contains information about how well dose predicts effect. To add more To see the summary statistics, type the following statement.

```
> summary(medModel)
            Df Sum Sq Mean Sq F value  Pr(>F)
dose         1  19.60  19.600   10.56 0.00634 **
Residuals   13  24.13   1.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first row (dose) is the parameters of variation between groups. The second row is the parameters of variation within groups. The F value is the F-ratio and the last column, Pr(>F) is the p-value of the ANOVA test. As we can see, the p-value is less than 0.05. Thus, we can reject the null hypothesis.

Chi-square Test

Download student dataset from eLearn@USM and load the dataset into R.

The last column (G3) contains the final scores of the students. Let's convert the data to categorical data as follows.

```
> breaks <- c(0, 10, 20)
> student$G3.groups <- cut(student$G3, breaks=breaks, labels=c("low",
  "high"), include.lowest=T)
```

Let's determine the relationship between family educational support with final score. First, we create a contingency table from the attributes.

```
> cols <- c('famsup', 'G3.groups')
> subset_data <- student[cols]
> c_table <- table(subset_data)
> print(c_table)
```

To perform the chi-square test, type the following statement.

```
> chisq.test(c_table, correct=F)

        Pearson's Chi-squared test

data:  c_table
X-squared = 1.0375, df = 1, p-value = 0.3084
```

As we can see, the p-value is 0.3084 way above the significance level. Thus, we cannot reject the null hypothesis and conclude that the variables are independent.

Exercise

Load the cardio dataset into R.

1. Extract the ap_hi rows where the cholestrol is 1 (normal) and name the list as hi_chol1
2. Extract the ap_hi rows where the cholestrol is 3 (well above normal) and name the list as hi_chol3
3. Reduce the number of observations in hi_chol1 to 250 (use random sampling).
4. Reduce the number of observations in hi_chol3 to 250 (use random sampling).
5. Perform t-test to test if the difference in means between hi_chol3 and hi_chol1 equals 0.
6. Perform t-test to test if the difference in means between hi_chol3 and hi_chol1 is equal or greater than 10.

Download the customer dataset and load it into R.

1. Change negative and zero values of column income to a value that is close to zero e.g. 0.001. You may create a new column income_fix.
2. Extract the income rows where the health.ins is equals to TRUE and name the list as income_t.
3. Extract the income rows where the health.ins is equals to FALSE and name the list as income_f.
4. Calculate the mean and standard deviation of income_t and income_f. Based on the descriptive statistics, what do you think of the two distributions?
5. Perform ANOVA test on the health.ins and income attributes. Specify health.ins and income as outcome and predictor respectively. State the conclusion based on the outputs of the test.