Hypothesis Testing

Hamid Abdulsalam

Introduction

In this section, we will look at hypothesis testing with practical examples. There are two main branches of Statistics, Inferential and descriptive statistics.Descriptive statistics derives a summary from the data by using central tendencies like mean and median, dispersion like variance and standard deviation, and skewness. In a nutshell, in descriptive statistics, you are basically describing the data.

In inferential Statistics, Inferences are made about the population and the sampled data through hypothesis testing and estimation of parameters.

Hypothesis Testing

In testing hypothesis, you try to answer a research question. A research question can be, Is there a significant difference between the salary of female and male workers of company B? The hypothesis can be, **There is a significant difference between the salary of female and male workers of company B.**

The research question begins with **Is there** and the hypothesis begins with **There is**.Based on the research question,the null and alternative hypothesis will be obtained.

Hypothesis Testing

The null hypothesis and alternative hypothesis can be expressed as:

- Ho: u1 = u2
- Ha: u1 != u2

Where *u1* represent the mean salary of female workers and *u2* represent the mean salary of male workers of Company B. We can use inference tests to obtain the p-value and make conclusion. In testing hypothesis, there are five major steps to follow.

Hypothesis Testing

The five major steps are :

- Step 1: Declare the null hypothesis H0 and the alternative hypothesis Ha
- Step 2: Specify an acceptable level of significance, alpha, normally 0.05 or 0.01.
- Step 3: Select a test statistic.
- Step 4: Compute the value of the statistic from the data and obtain your P-value
- Step 5: Compare your P-value with the level of significance, alpha, in step 2

Hypothesis Testing

If the p-value is less than or equal to alpha, which is usually 0.05, you reject the null hypothesis and say that the alternative hypothesis is true. If the p-value is more than 0.05, you fail to reject the null hypothesis.

When we reject a true null hypothesis, we have committed **type I** error. Similarly, a rejection of a true alternative hypothesis means that we have committed **type 2** error.

T-Test

T-test is one of the most important tests in data science.A one Sample T-test is used to determine whether the sample mean is statistically different from a known or hypothesized population mean

Assumptions of One Sample T-test

- The dependent variable must be continuous .
- The samples are randomly sampled from their population
- The dependent variable should be approximately normally distributed.

Case Study 1

A manufacturer of motorcycle exhaust systems wants to test its new exhaust system to determine whether it meets state air pollution standards. The mean emission of all motor-cycle exhaust systems of this type must be less than twenty parts per million of carbon. Fourteen exhaust systems are manufactured for testing purposes. The data below represents the emission level of the fourteen exhaust systems tested. Do the data supply sufficient evidence to allow the manufacturer to conclude that the new exhaust system for motorcycles meets the pollution standard? Assume that the manufacturer has set the level of significance to 1%.

- 17 16 20 18 16 19 22 13 15 17 21 14 18 12

Case Study 1

- Ho: $u = 20$
- H1: $u < 20$
- Level of Significance : 0.01
- Test

```
data <- c(17,16, 20, 18, 16, 19, 22, 13 ,
          15 ,17 ,21, 14, 18, 12)
result<-t.test(data, mu=20, alternative = "less")
```

Output

```
result
```

```
##
##  One Sample t-test
##
## data:  data
## t = -3.8243, df = 13, p-value = 0.001054
## alternative hypothesis: true mean is less than 20
## 95 percent confidence interval:
##      -Inf 18.38923
## sample estimates:
## mean of x
##        17
```

Conclusion

Since the P-value is less than 0.01, we reject the null hypothesis and conclude that the mean emission of all motor-cycle exhaust systems is less than twenty parts per million of carbon

Two Sample T-Test

The two-sample unpaired t-test is when you compare two means of two independent samples. Given two samples or group, you want to see if the means are the same or significantly different from each other.

Case Study 2: Do students taking Introduction to Statistics in early morning classes perform better than their peers in late afternoon classes? Data on their final exams were obtained. Using a 5% level of significance, does the early morning class seem to perform better on the final exam than the late afternoon class?

Case Study 2

- Ho: u1 = u2
- H1: u1 > u2

where u1 represent the mean final exam of morning students and u2 represent the mean final exam of afternoon students.

- Level of Significance : 0.05

```
data2<- read.csv("data/class.csv", header = T)
str(data2)
```

```
## 'data.frame':    17 obs. of  2 variables:
##  $ MorningClass: int  80 82 71 90 69 75 78 87 95 98 ...
##  $ EveningClass: int  79 83 73 77 84 90 91 95 80 94 ...
```

Output

```
result2<-t.test(data2$MorningClass, data2$EveningClass,
                alternative = "greater")
result2
```

```
## 
##  Welch Two Sample t-test
## 
## data:  data2$MorningClass and data2$EveningClass
## t = 0.16577, df = 30.283, p-value = 0.4347
## alternative hypothesis: true difference in means is grea
## 95 percent confidence interval:
##  -4.176315       Inf
## sample estimates:
## mean of x mean of y 
##  82.68750  82.23529
```

Conclusion

Since we have a P-value which is greater than 0.05, we fail to reject the null hypothesis and conclude that students taking Introduction to Statistics in early morning classes do not perform better than their peers in late afternoon classes

Contigency Test

If you have two categorical variables and you want to compare whether there is a relationship between two variables, you can use the contingency test. The null hypothesis means that the two categorical variables have no relationship. The alternate hypothesis means that the two categorical variables have a relationship.

Case Study 3

Let’s examine the relationship between Number of cylinders(cyl) and Number of carburetors (carb) which are variables are in ‘mtcars’ dataset. We want to know if Number of cylinders and Number of carburetors are dependent or not.

Ho : Number of cylinders(cyl) and Number of carburetors (carb) are Independent

H1: Number of cylinders(cyl) and Number of carburetors (carb) are dependent

```r
data("mtcars")
chi_t<-chisq.test(mtcars$carb, mtcars$cyl)
```

Output

```
chi_t
```

```
##
##  Pearson's Chi-squared test
##
## data:  mtcars$carb and mtcars$cyl
## X-squared = 24.389, df = 10, p-value = 0.006632
```

Conclusion

We have a high chi-squared value and a p-value that is less than 0.05 significance level. So we reject the null hypothesis and conclude that there is a significant relationship between Number of cylinders(cyl) and Number of carburetors (carb).

The End