Hypothesis-testing-in-R

Gian Carlo Sanfuego

2024-03-11

#Hypothesis Testing in R

What is Hypothesis Testing? Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed depends on the nature of the data used and the reason for the analysis. For example, you have an assumption, you think something should be like that - so you test the assumption. You want to be sure that the assumption is not just based on intuition, and that’s what hypothesis testing is really about. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

In other words, a statistical hypothesis is a hypothesis that is testable based on observed data; that is modeled as the realized values taken by a collection of random variables. Its like you have an idea and you think something should be correct. Hypothesis testing helps you to clarify you doubts and validate your intuition

# Task One: Getting Started

In this task, you will learn to set and check your current working directory. Also, you will learn two important rules for hypothesis testing

### Set and get the working directory

# setwd("C:/Users/gianc/OneDrive/Desktop/DATA ANALYSIS/PROJECTS/R/HYPOTHESIS TESTING IN R")

### Get the working directory

getwd()

## [1] "C:/Users/gianc/OneDrive/Desktop/DATA ANALYSIS/PROJECTS/R/HYPOTHESIS TESTING IN R"

Rule One: When p-value < level of significance or when test statistic (calculated value) > tabulated value, we have evidence against the null hypothesis, hence, do not accept the null hypothesis

The p-value in statistics testing is just a probability of obtaining test result that is at least as extreme as a result we actually observed under the assumption that null hypothesis is correct. Usually in statistics, the level of significance is 0.05.

Rule Two: On the other hand if p-value > level of significance or when test statistic (calculated value) < tabulated value, we have weak evidence against the null hypothesis, so we fail to reject the null hypothesis

# Task Two: Test for Proportions

In this task, we will perform hypothesis test for proportions using a one sample test

### 2.1: Consider a survey asking 100 randomly selected people if they had

### breakfast on Saturday morning.

### Suppose 42 people say yes. 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 H0 is that the proportion p is 0.5, the alternative hypothesis, denoted HA, in this example would be that the proportion p is not 0.5. This is a so called two-sided alternative test.

### To test the assumptions, we use the function prop.test as

### with the confidence interval calculation as follows:

prop.test(42, 100, p = 0.5,  
 alternative = "two.sided",  
 conf.level = 0.95,  
 correct = TRUE)

##   
## 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

The function performs a Z-test, that is, it compares this proportion to the hypothesized value; the input of our test is 42, the number of total sample 100 people, the hypothesized value of the proportion under the null hypothesis is p equals to 0.5 or 50%. Setting correct = TRUE applies Yates’ continuity correction, which adjusts the observed counts in each cell by 0.5 before calculating the test statistic. This correction is often used in the context of 2x2 contingency tables when performing hypothesis tests on proportions. If correct = TRUE, the function applies Yates’ continuity correction. If correct = FALSE (the default), no continuity correction is applied.

The p value is 0.1336. Usually our level of significance is 0.05. Do not confuse this with the p equals to 0.5. The p here is the proportion. In this the level of significance or alpha is 0.05, that is like a 95% interval.

So looking at this, obviously, the p value is greater than the 0.05. We can see in this case that since the p value is greater than the 0.05 alpha level, we then accept the null hypothesis. That also means that we don’t have enough evidence to reject the null hypothesis.

### 2.2: What if we ask 1000 people and 420 say yes.

### Does this still support the

### null hypothesis that p = 0.5?

prop.test(420, 1000, p = 0.5,  
 alternative = "two.sided",  
 conf.level = 0.95,  
 correct = TRUE)

##   
## 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

We see in this case that the p-value is significantly small, very minimal about times ten raised to power of -7. This means that we do not accept the null hypothesis. This simply illustrates that the p-value does not just depend on the ratio it also depends on the n, the value , the number of samples. In particular, this is because the standard error of the sample average gets smaller as n gets larger. So we can see now how to make decisions.

# Task Three: Test for Means

In this task, we will perform hypothesis testing for means for one sample.

### 3.1: Suppose a flash light manufacturer claims a customer gets

### 25 hours of light on average on a full charge.

### A consumer group asks 10 owners of this flashlight to calculate

### their hours of usage on a full charge and the mean value was 22

### with a standard deviation of 1.5.

### Is the manufacturer’s claim supported?

### In this case H0: mu = 25 against the one-sided

### alternative hypothesis that mu < 25.

### To test using R we simply need to tell R about the type of test.

### (As well, we need to convince ourselves that the t-test is appropriate

### for the underlying parent population.)

For this example, the built-in R function t.test isn’t going to work as the data is already summarized so we are on our own. We need to calculate the test statistic and then find the p-value.

### Compute the t statistic. Note we assume mu = 25 under H0

xbar = 22  
s = 1.5  
n = 10

What we want try to calculate manually is the test statistics.

### Calculate the t statistic

t <- (xbar - 25)/(s/sqrt(n))

The reason we want to calculate in this case is because our sample size is small. Remember the very important condition in t-test is when sample size is less than 30.

### Now use pt to get the distribution function of t

### To get the p-value

pt(t, df=n-1)

## [1] 6.846828e-05

The resulting value is 6.84 times 10 raise to -5. What does that mean to us? That this is a very small p-value which is less than 0.05, and the decision rule states that when p-value is less than 0.05, we do not accept the null hypothesis. This means that the manufacturers claim is very suspicious.

Now, lets see how to use the t.test function

### 3.2: The following is the results of the measurements on each of

### 6 randomly selected members of a population whose distribution is

### normal with unknown mean and unknown variance: 11,19,16,21,24,27

### Test the hypothesis mu = 14.0 against the alternative mu > 14.0

### at 5% level of significance.

result <- c(11,19,16,21,24,27)  
t.test(result, mu = 14, alternative = "greater")

##   
## One Sample t-test  
##   
## data: result  
## t = 2.4286, df = 5, p-value = 0.02974  
## alternative hypothesis: true mean is greater than 14  
## 95 percent confidence interval:  
## 14.96489 Inf  
## sample estimates:  
## mean of x   
## 19.66667

In this case, we can see that the p-value is 0.02, approximately 0.03 and that is still less than 0.05. Thus rejecting the null hypothesis that the mean is 14 although, we do not have a very strong evidence unlike when we had 0.00 raised to power minus five. Comparing the mean of the sample data to a hypothesized mean of 14 suggests that the population mean is significantly greater than 14 at a 5% significance level.

### Task Four: Two sample test for proportions

In this task, we will learn how perform hypothesis test for proportions for two samples

### 4.1: A survey is taken two times over the course of two weeks.

### The pollsters wish to see if there is a difference in the results

### as there has been a new advertising campaign run.

### Here is the data:

Week1 <- c(45,35)  
Week2 <- c(56,47)   
advert <- data.frame(Week1, Week2)  
rownames(advert) <- c("Favorable", "Unfavorable")   
advert

## Week1 Week2  
## Favorable 45 56  
## Unfavorable 35 47

### The standard hypothesis test is H0 : P1 = P2 against the alternative

### (two-sided) HA: P1 neq P2.

### The function prop.test is used to being called

### as prop.test(x,n) where x is the number favorable and n is the total.

### Here it is no different, but since there are two x’s it looks slightly

### different as follows:

prop.test(c(45, 56), c(45+35,56+47))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(45, 56) out of c(45 + 35, 56 + 47)  
## X-squared = 0.010813, df = 1, p-value = 0.9172  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.1374478 0.1750692  
## sample estimates:  
## prop 1 prop 2   
## 0.5625000 0.5436893

We see that in this case, the p-value is 0.9172 and that is greater than the level of significance 0.05, this suggests that we will accept the null hypothesis. This means that our proportion one is equal to the proportion two. Meaning there is no difference for week 1 and week 2 of favorable and unfavorable responses.

# Task Five: Two sample test for means

In this task, we will learn how to perform hypothesis test for means for two samples with equal and unequal variances

### 5.1: Suppose the recovery time for patients taking a new drug is

### measured (in days). A placebo group is also used

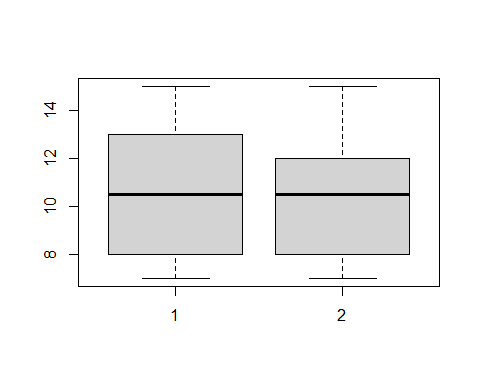
### to avoid the placebo effect. The data is as follows:

withdrug <- c(15,12,13,7,9,8)  
placebo <- c(15,11,8,12,10,7)

First thing, we want to confirm if the assumptions are met. Is the data normally distributed? Is the data equal in variance? Are the data continuous? How about independence? We can test assumptions visually and formal statistics test. But this time we will do it visually:

### Plot a boxplot for the two variables

boxplot(withdrug, placebo)



The boxplot supports the assumptions of equal variances and normality.

Lets confirm it using var function to check the variance.

### Calculate the variances

var(withdrug)

## [1] 9.866667

var(placebo)

## [1] 8.3

### We now test the null hypothesis of equal means against the one-sided

### alternative that the drug group has a smaller mean.

t.test(withdrug, placebo, alternative = "less", var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: withdrug and placebo  
## t = 0.095783, df = 10, p-value = 0.5372  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 3.320442  
## sample estimates:  
## mean of x mean of y   
## 10.66667 10.50000

Here, the p-value is 0.5372, and that is way greater than 0.05, which means we cannot reject the null hypothesis, thus accepting the null hypothesis that drug group has a equal mean than placebo group.

What if we have a case where variance is not equal? The counterpart is by using var.equal = FALSE.

### 5.2: Two-sampled t-test with unequal variances

### Let’s use the same example as above, but assuming unequal variances

t.test(withdrug, placebo, alternative = "less", var.equal = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: withdrug and placebo  
## t = 0.095783, df = 9.9262, p-value = 0.5372  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 3.322823  
## sample estimates:  
## mean of x mean of y   
## 10.66667 10.50000

Notice that gives us a same result. Although in some cases, the result are slightly different. Thus, the conclusion are the same. The difference between the two can be seen in degrees of freedom.

# Task Six: Matched Samples

In this task, we will learn how to test hypothesis for matched samples

Matched samples are also called matched pairs or paired samples, or sometimes called dependent samples and these are paired up so that the participants share every characteristics except for the one under investigation.

### 6.1: In order to promote fairness in grading an entrance examination, each

### candidate was graded twice by different graders. Based on the grades,

### can we see if there is a difference between the two graders?

grader1 <- c(3, 0, 5, 2, 5, 5, 5, 4, 4, 5)   
grader2 <- c(2, 1, 4, 1, 4, 3, 3, 2, 3, 5)

### Clearly there are differences. Are they described by random fluctuations

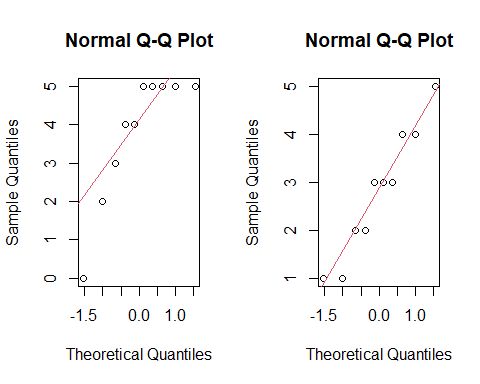
### (mean mu is 0), or is there a bias of one grader over another?

### (mean mu neq 0)

Testing this should have conditions set and met. First is that, the sampling method for each sample is simple random, meaning that each person has an equal chance of being selected. Second is that they are dependent. Third is that data is normally distributed. We will use qqplot to check for normality.

### We should check the assumption of normality with normal plots

par(mfrow=c(1,2))  
qqnorm(grader1)  
qqline(grader1, col = 2)  
qqnorm(grader2)  
qqline(grader2, col = 2)



This looks to be normal

### A matched sample test will give us some insight.

### We apply the t-test as follows:

t.test(grader1, grader2, paired = TRUE, alternative = "two.sided")

##   
## Paired t-test  
##   
## data: grader1 and grader2  
## t = 3.3541, df = 9, p-value = 0.008468  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## 0.325555 1.674445  
## sample estimates:  
## mean difference   
## 1

We can see that the result is 0.008468 leading us to not accept the null hypothesis because the p-value is less than the level of significance of 0.05, This agrees with our earlier intuition because we saw a clear difference in the values of grader1 and grader2. Clearly we saw that there was difference but we wanted to be sure that assumptions is correct, and the result shows that there is a clear difference. Thus, accepting the alternative hypothesis. This will help us prompt an action why are these two graders giving different results even though it is the same paper given at different point in time. This testing helps us validate decision making.