CS 251 Statistical Computing

HOP 5: R for statistical project

7/13/2020 Developed by Aya Khalil

10/21/2020 Reviewed by Shanshan Yu

School of Technology and Computing @City University of Seattle (CityU)

**Before You Start**

* If you already finished this module through any CityU Technology Institute (TI) courses,  
  just skim this module and skip it.
* Version numbers may not match with the guide. But that should be fine.  
  If given the option to choose between stable release (long-term support) or most recent, please choose the stable release.
* This guide targets Windows OS users. So, MacOS users may have different commands to input in the shell/terminal.
* We cannot explain every step. **This cookbook always needs your own creative judgement.**
* **For your working directory, use your course number.** The hands-on tutorial may use a different course number as an example.

**Learning Outcomes**

* Hypotheses Testing

**Resource**

* Hypotheses Testing: <http://www.r-tutor.com/elementary-statistics/hypothesis-testing>
* Schmuller, J. (2017). Statistical Analysis with R for Dummies. John Wiley & Sons. (ISBN: 9781119337065)
* Hypothesis Testing: <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTest-Means-Proportions/BS704_HypothesisTest-Means-Proportions3.html>
* R Programming for Statistics and Data Science 2020 online course: <https://www.udemy.com/course/r-programming-for-statistics-and-data-science/learn/lecture/10158720#overview>

**Section1: Hypotheses Testing**

A hypothesis is a guess about the way the world works. It's a tentative explanation of some process, whether that process occurs in nature or in a laboratory.

**For example**

A hypothesis is an idea that can be tested. So, you can say apples in Washington are more expensive than $1.8. In that case you can try to test your idea by gather some samples and data.

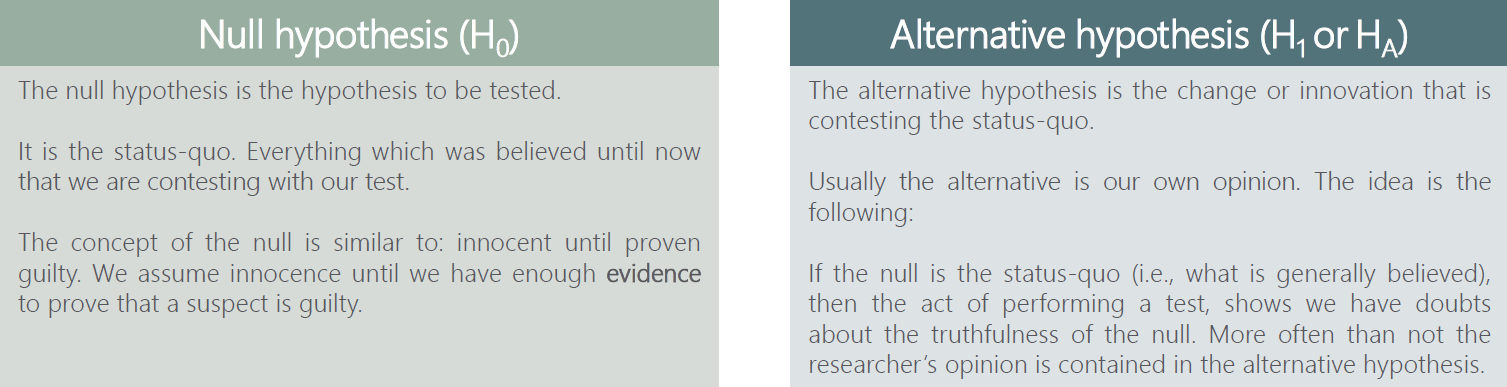
However, if you said apples in Washington are expensive, this is a statement not a hypothesis and cannot be tested.

In hypothesis testing, you

* Formulate null and alternative hypotheses
* Gather data
* Decide whether to reject or not reject the null hypothesis.

**Null & Alternative Hypotheses**

**A null hypothesis** *H0* is a hypothesis that says there is no statistical significance between the two variables. It is usually the hypothesis a researcher or experimenter will try to disprove or discredit. **An alternative hypothesis** *Ha* is one that states there is a statistically significant relationship between two variables.

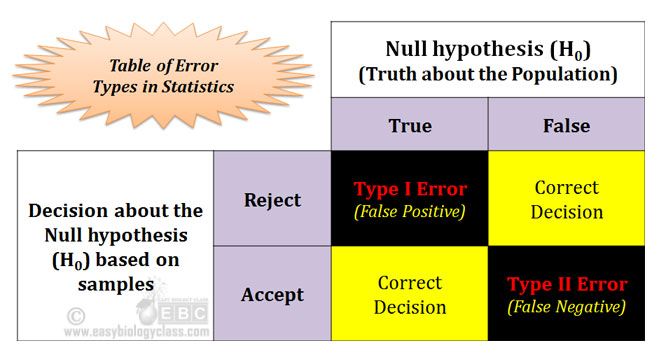


**Please watch the following video to understand the concept and the difference between *H0* & *Ha***

Hypothesis Testing - Null and Alternative Hypotheses. Retrieved from: <https://www.youtube.com/watch?v=wiYJWyfdGg4&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=49>

**Type I & Type II errors**

there are two types of errors we can make while testing Type I error (False positive) and Type II Error (False negative)



The probability of committing Type I error (False positive) is equal to the significance level α

The probability of committing Type II error (False negative) is equal to the beta β and is called ‘power of the test’

**Please watch the following video to understand the concept and the difference between TypeI and TypeII errors**

How To Identify Type I and Type II Errors In Statistics. Retrieved from: <https://www.youtube.com/watch?v=9yQm9F2_yIk&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=50>

**Level of Significance**

The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

**Common significance levels:**

0.01

0.05

0.10

**Decisions you can take**

When testing, there are two decisions that can be made to accept the null hypothesis or to reject the null hypothesis

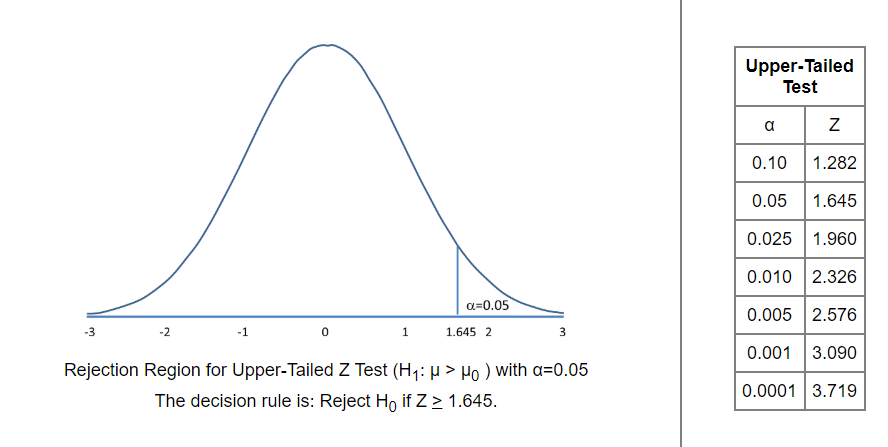
To accept the null means that there isn’t enough data to support the change or the innovation brought by the alternative

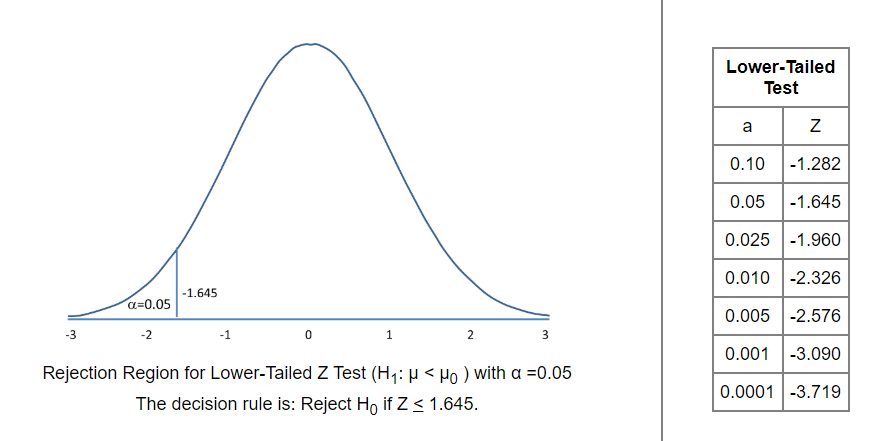
To reject the null means that there is enough statistical evidence that the status quo is not representative of the truth

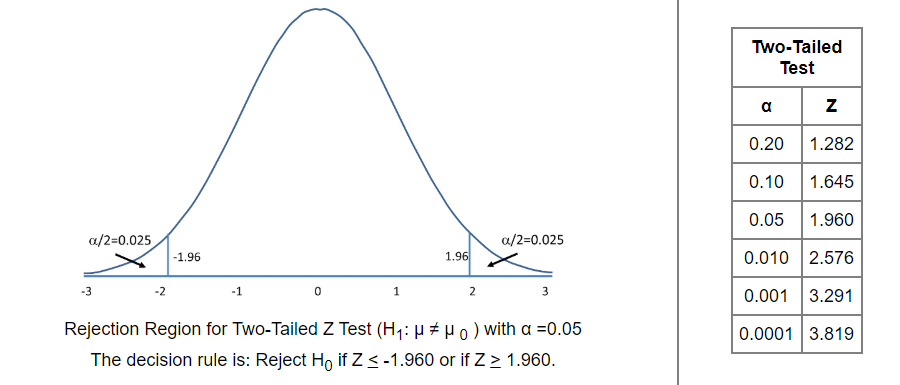
The decision rule is a statement that tells under what circumstances to reject the null hypothesis. The decision rule is based on specific values of the test statistic (e.g., reject H0 if Z > 1.645). The decision rule for a specific test depends on 3 factors: the research or alternative hypothesis, the test statistic and the level of significance. Each is discussed below.

1. The decision rule depends on whether an upper-tailed, lower-tailed, or two-tailed test is proposed. In an upper-tailed test the decision rule has investigators reject H0 if the test statistic is larger than the critical value. In a lower-tailed test the decision rule has investigators reject H0 if the test statistic is smaller than the critical value.  In a two-tailed test the decision rule has investigators reject H0 if the test statistic is extreme, either larger than an upper critical value or smaller than a lower critical value.
2. The exact form of the test statistic is also important in determining the decision rule. If the test statistic follows the standard normal distribution (Z), then the decision rule will be based on the standard normal distribution. If the test statistic follows the t distribution, then the decision rule will be based on the t distribution. The appropriate critical value will be selected from the t distribution again depending on the specific alternative hypothesis and the level of significance.
3. The third factor is the level of significance. The level of significance which is selected in Step 1 (e.g., α =0.05) dictates the critical value.   For example, in an upper tailed Z test, if α =0.05 then the critical value is Z=1.645.

The following figures illustrate the rejection regions defined by the decision rule for upper-, lower- and two-tailed Z tests with α=0.05. Notice that the rejection regions are in the upper, lower and both tails of the curves, respectively. The decision rules are written below each figure.







**Please watch the following video to understand the concept of lower, upper, both tailed tests.**

One Tailed and Two Tailed Tests, Critical Values, & Significance Level - Inferential Statistics. Retrieved from: <https://www.youtube.com/watch?v=XHPIEp-3yC0&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=51>

**Please watch the following video to understand the how you can solve hypothesis test problems**

Hypothesis Testing Problems Z Test & T Statistics One & Two Tailed Tests 2. Retrieved from: <https://www.youtube.com/watch?v=zJ8e_wAWUzE&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=53>

**Questions (***You do not have to submit the answers to these questions, if you can solve it, then go to the coding part, if not, then review the theory part one more time***)**

* What is the difference between H0 & Ha?
* What’s α?
* How can you determine if you are going to use two tailed or one test?
* What is the Z value, how can you determine, and what’s the z table?
* What is the relation between z and α?
* What is the difference between z & t testing? When can you use them?

**Section1: Z-Test**

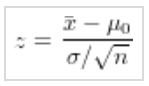
**Lower Tail Test of Population Mean with Known Variance**

The null hypothesis of the lower tail test of the population mean can be expressed as follows:



where μ0 is a hypothesized lower bound of the true population mean μ.

Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ:



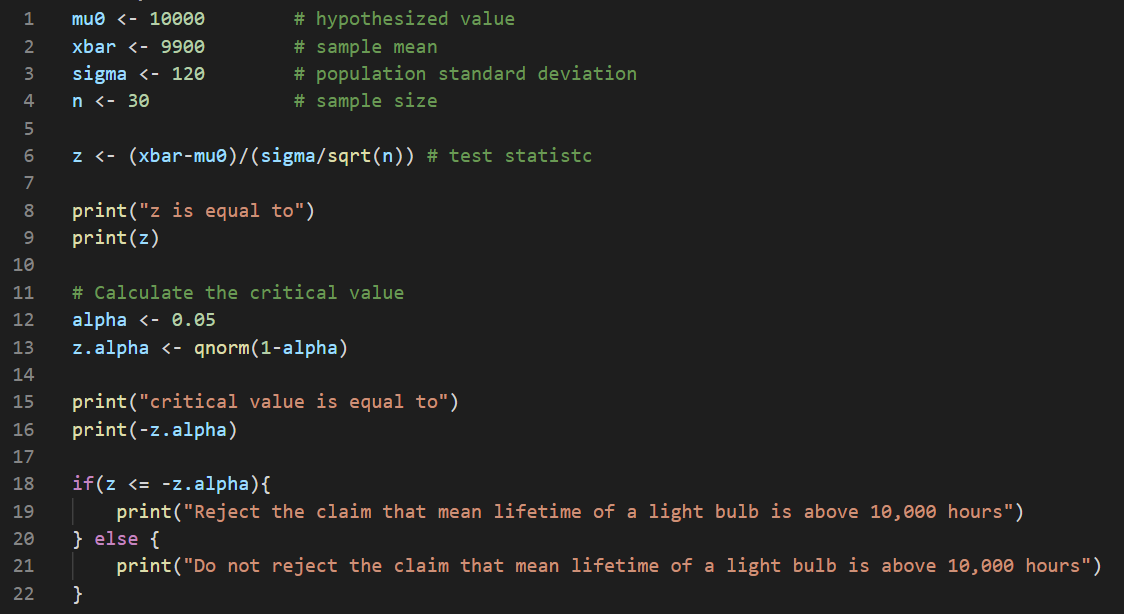
Then the null hypothesis of the lower tail test is to be rejected if z ≤−zα , where zα is the 100(1 − α) percentile of the standard normal distribution.

Problem:

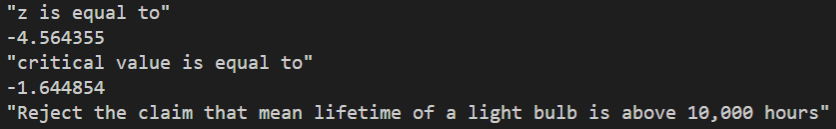
Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

Answer:

* In Module5 folder, create lwrTailMean.R file
* Type the following in lwrTailMean.R



Run your code

Output: 

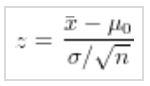
**Upper Tail Test of Population Mean with Known Variance**

The null hypothesis of the upper tail test of the population mean can be expressed as follows:



where μ0 is a hypothesized upper bound of the true population mean μ.

Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ:



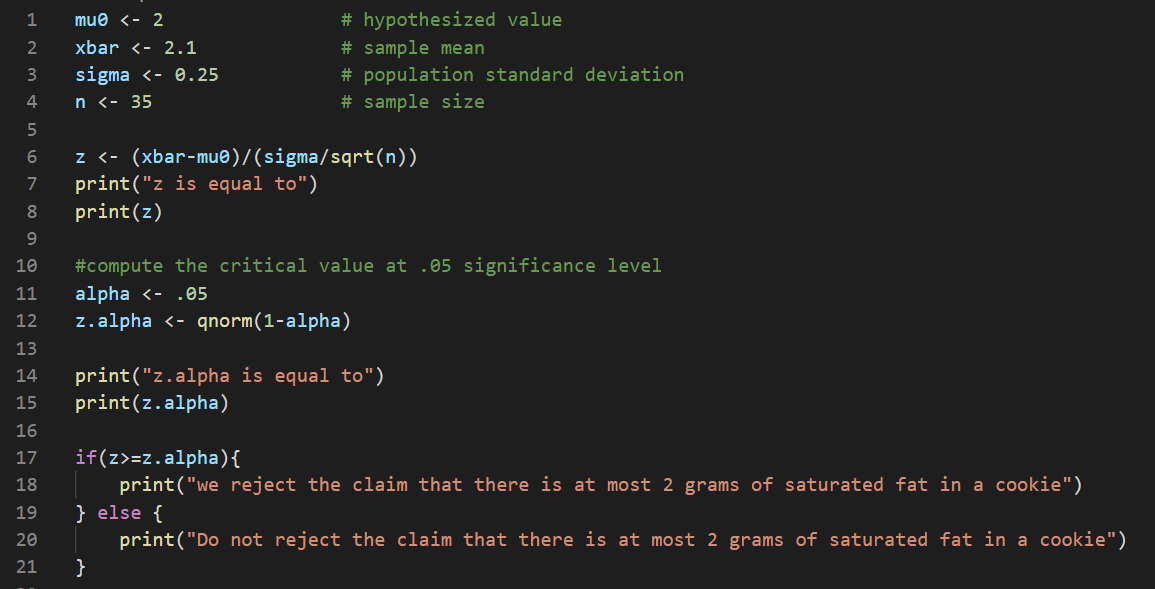
Then the null hypothesis of the upper tail test is to be rejected if z ≥ zα , where zα is the 100(1 − α) percentile of the standard normal distribution.

Problem:

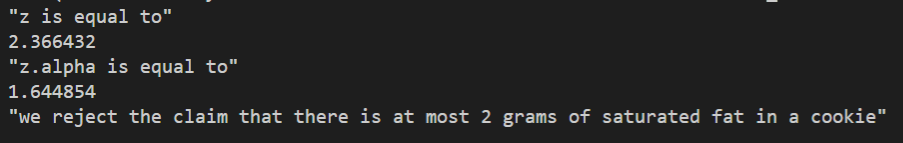
Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At .05 significance level, can we reject the claim on food label?

Solution:

* **In Module5, create Upertail.R file**
* Type the following in Uprtail.R



Run your code

Output: 

So, the answer is the test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.

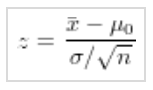
**Two-Tailed Test of Population Mean with Known Variance**

The null hypothesis of the two-tailed test of the population mean can be expressed as follows:



where μ0 is a hypothesized value of the true population mean μ.

Let us define the test statistic z in terms of the sample mean, the sample size, and the population standard deviation σ:



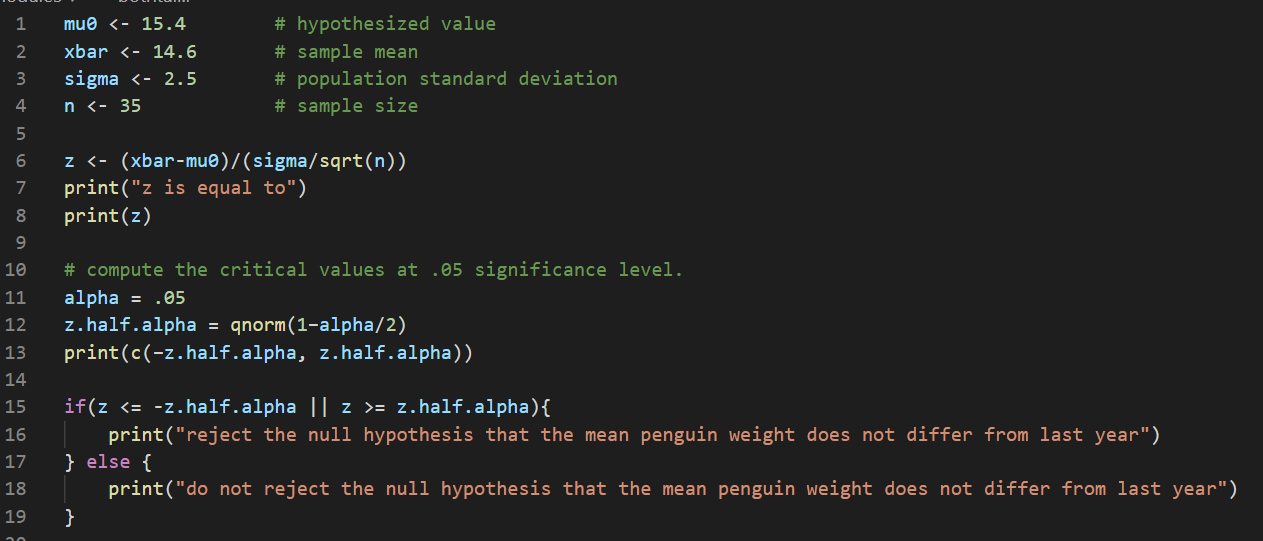
Then the null hypothesis of the two-tailed test is to be rejected if z ≤−zα∕2 or z ≥ zα∕2 , where zα∕2 is the 100(1 − α∕2) percentile of the standard normal distribution.

Problem:

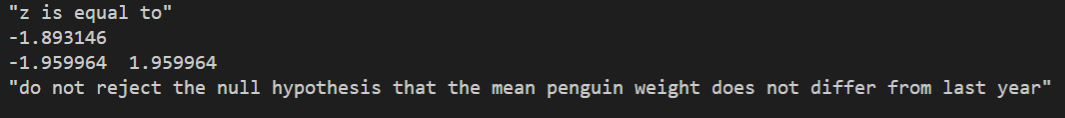
Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution:

* **In Module5, create Bothtail.R file**
* Type the following in Bothtail.R



Run your code

Output: 

The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

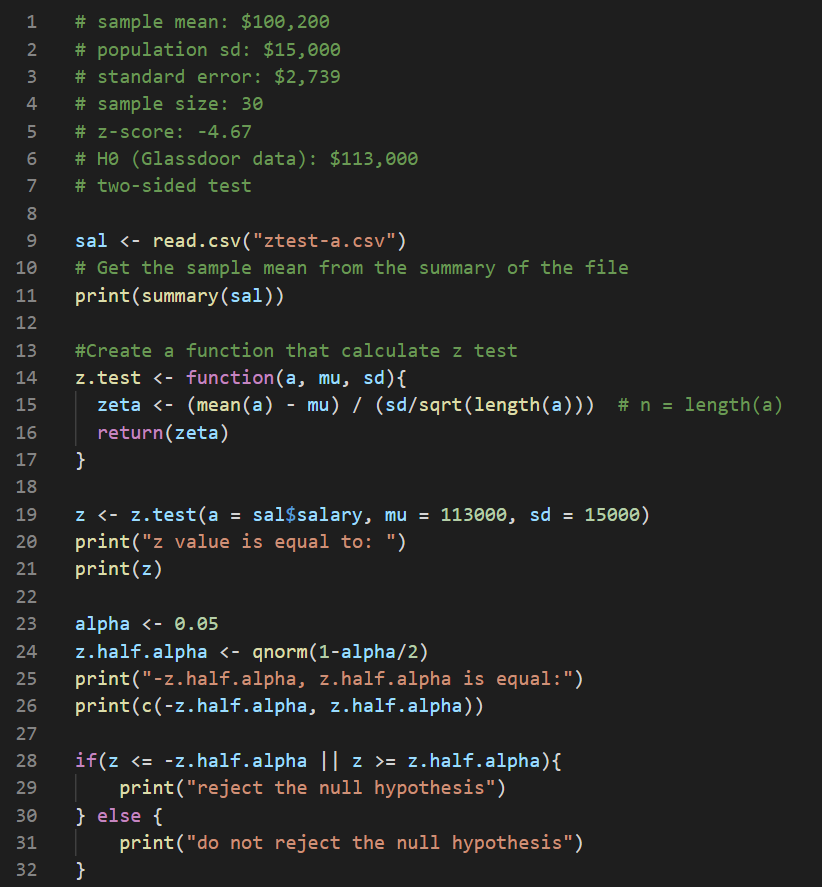
**Let’s try to do an example that read the sample data from a file**

Problem:

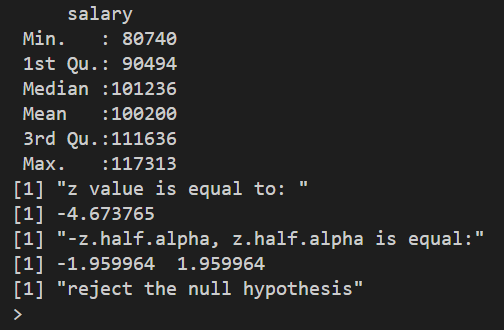
According to Glassdoor, the main data scientist salary is $113,000 per year. The **ztest-a.csv** file that we are using is a sample based on self-reported numbers on the website, so we would like to see if this cost or value is correct, in other word see if you can reject the null hypothesis at 10% significance level and the population standard deviation is 15,000.

Solution:

* **In Module5, create zVal-ReadFile.R file**
* Type the following in **zVal-ReadFile**.R



Run your code

Output: 

**Section2: t-Test**

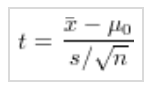
**Lower Tail Test of Population Mean with Unknown Variance**

The null hypothesis of the lower tail test of the population mean can be expressed as follows:



where μ0 is a hypothesized lower bound of the true population mean μ.

Let us define the test statistic t in terms of the sample mean, the sample size and the sample standard deviation s:



Then the null hypothesis of the lower tail test is to be rejected if t ≤−tα , where tα is the 100(1 − α) percentile of the Student t distribution with n − 1 degrees of freedom.

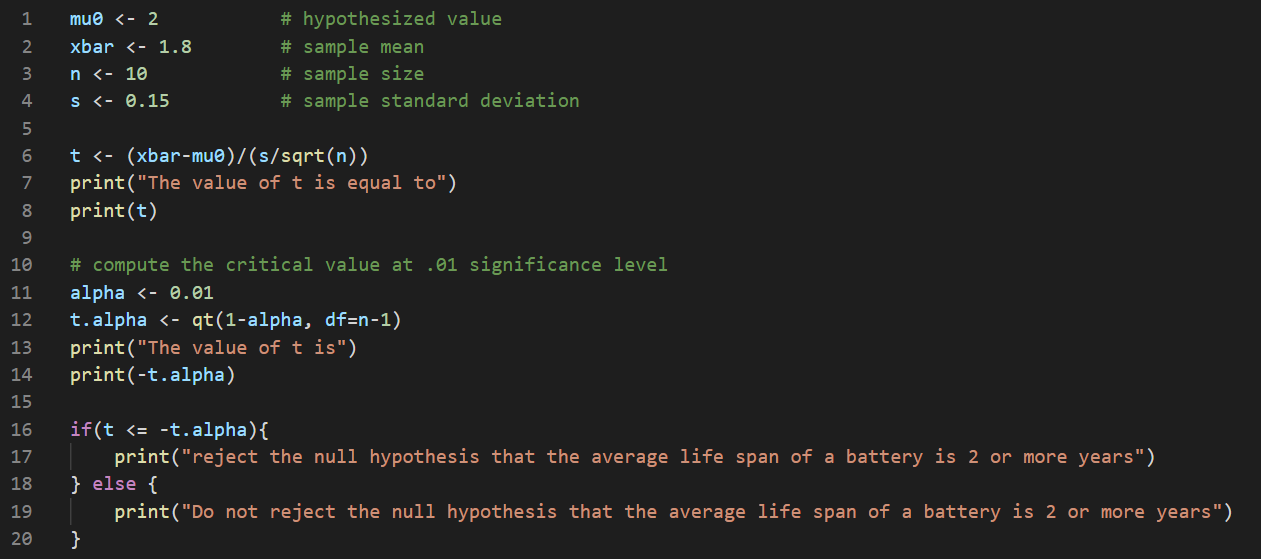
Problem:

A company manufactures car batteries with an average life span of 2 or more years. An engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 with a sample standard deviation of 0.15. At 0.01 significance level, can we reject the null hypothesis

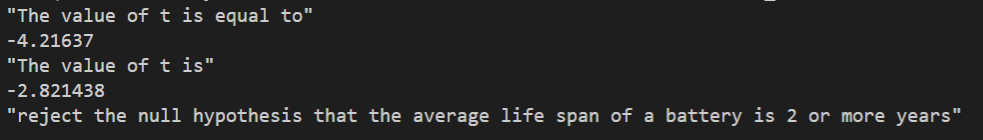
Answer:

**-In Module5, create tTestLwr.R file**

-Type the following in **tTestLwr.R file**



Run your code

Output: 

The test statistic -4.21637 is less than the critical value of -2.821438. Hence, at .01 significance level, we can reject the claim that the average life of a battery is 2 or more.

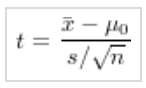
**Two-Tailed Test of Population Mean with Unknown Variance**

The null hypothesis of the two-tailed test of the population mean can be expressed as follows:



where μ0 is a hypothesized value of the true population mean μ.

Let us define the test statistic t in terms of the sample mean, the sample size and the sample standard deviation s:



Then the null hypothesis of the two-tailed test is to be rejected if t ≤−tα∕2 or t ≥ tα∕2 , where tα∕2 is the 100(1 − α) percentile of the Student t distribution with n − 1 degrees of freedom.

Problem

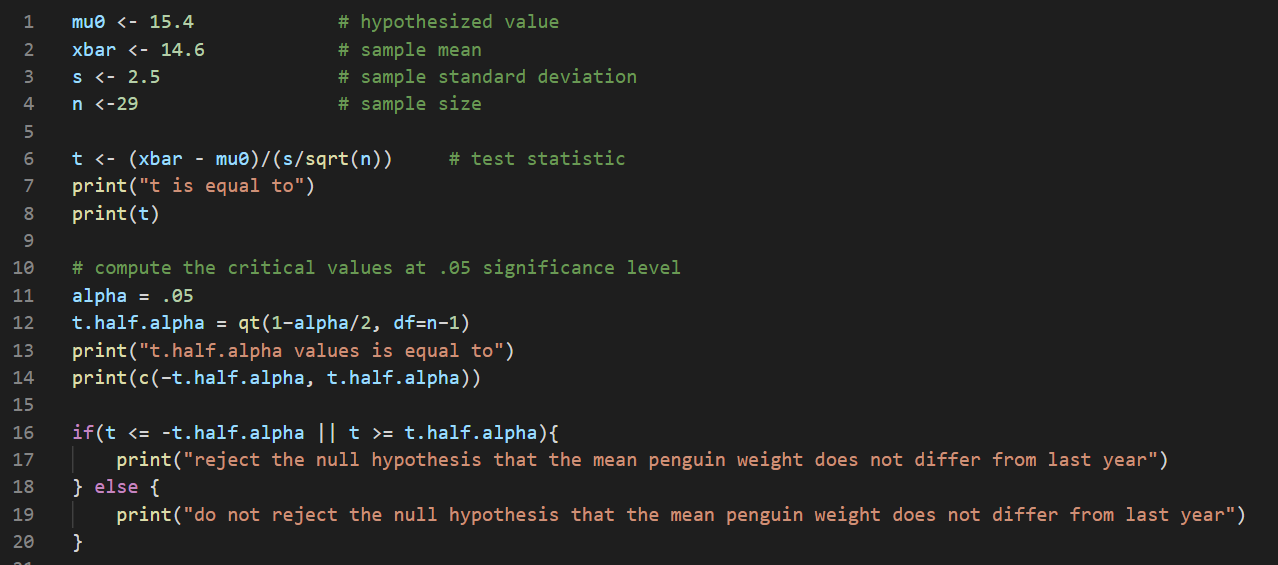
Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 29 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution

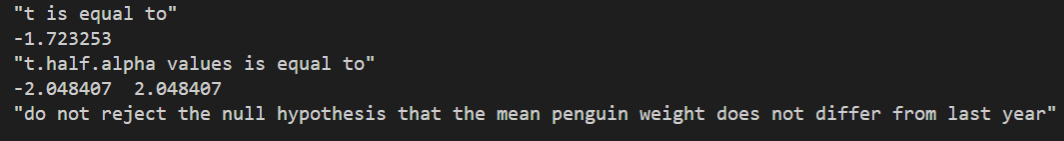
Answer:

**-In Module5, create tTestBoth.R file**

-Type the following in **tTestBoth.R file**



**Run your code**

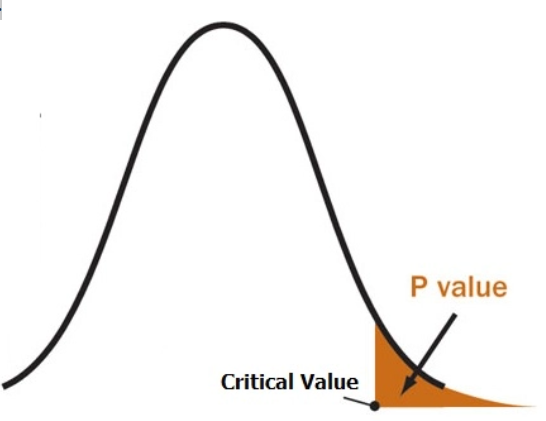
**Output** 

**Section3: P-value**

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That’s pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important (“significant”) your results.

Graphically, the p value is the area in the tail of a probability distribution. It’s calculated when you run hypothesis test and is the area to the right of the test statistic (if you’re running a two-tailed test, it’s the area to the left and to the right).



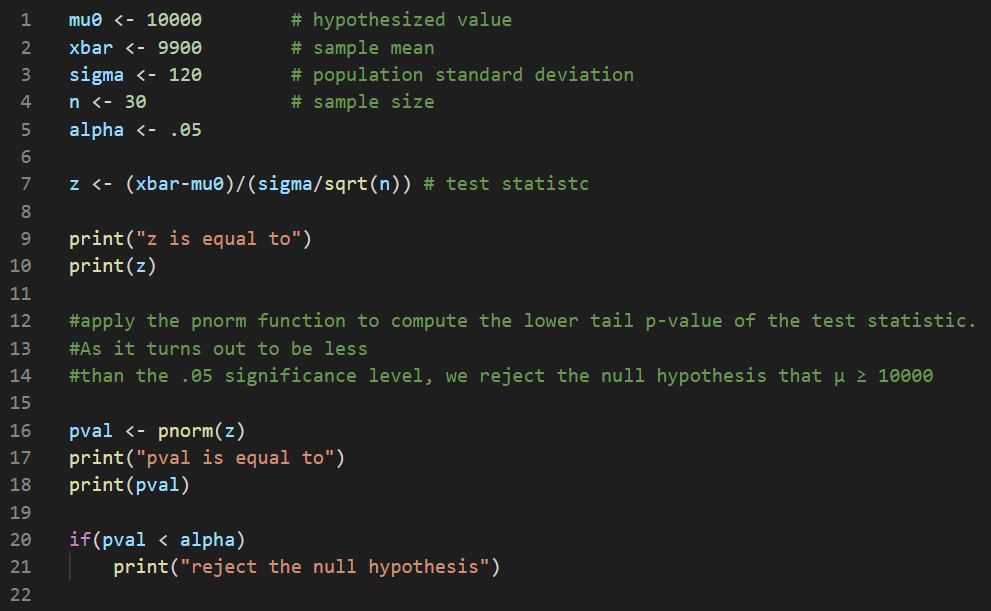
**Please watch the following video to understand the concept of P-value.**

P-Value Method For Hypothesis Testing. Retrieved from: <https://www.youtube.com/watch?v=8Aw45HN5lnA&list=PL0o_zxa4K1BVsziIRdfv4Hl4UIqDZhXWV&index=54>

Let’s try to solve the first example “Lower Tail Test of Population Mean with Known Variance” again using p-value

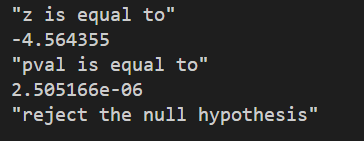
**In Module5, create PvalueEx1.R file**

Type the following in PvalueEx1.R file



Run your code

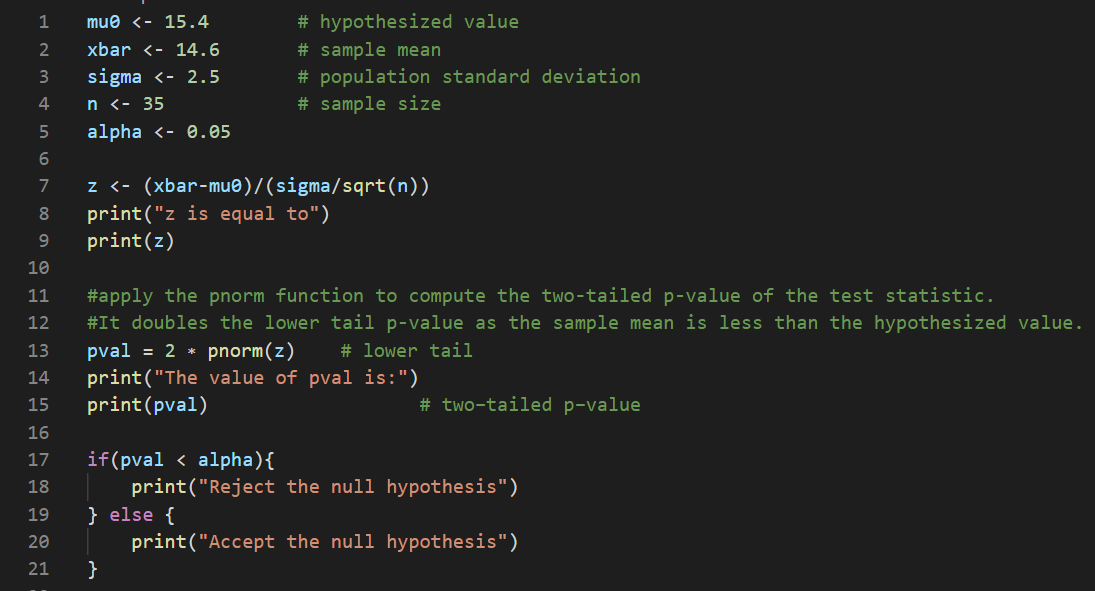
Output:



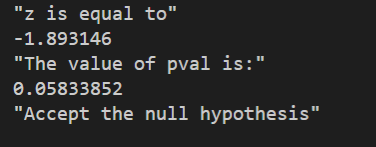
Let’s try to solve the first example “Two-Tailed Test of Population Mean with Known Variance” again using p-value

**In Module5, create PvalueEx2.R file**

Type the following in PvalueEx2.R file



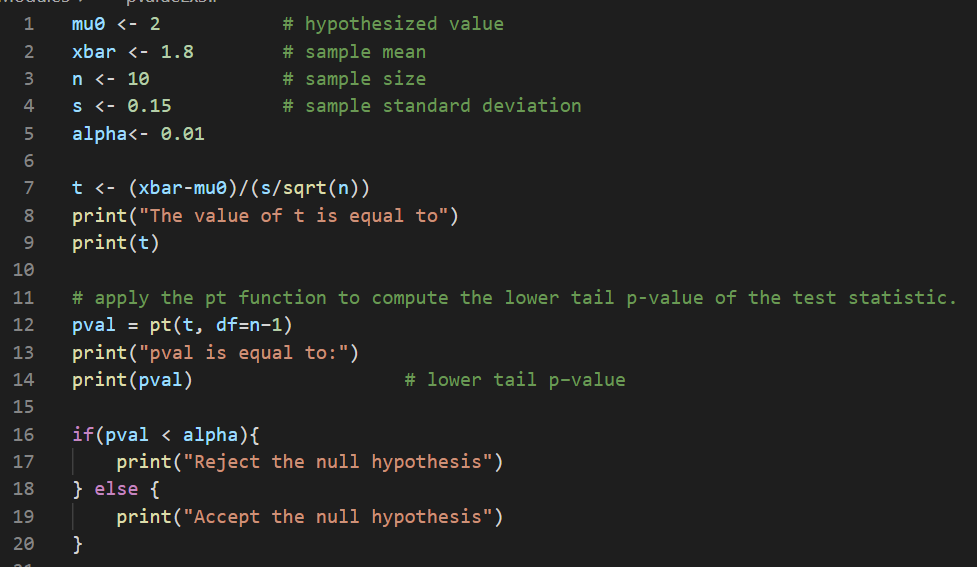
Run your code

Output: 

Let’s try to solve the first example “Lower Tail Test of Population Mean with Unknown Variance” again using p-value

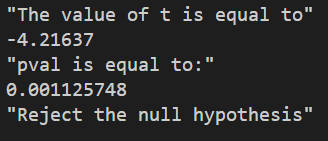
**In Module5, create PvalueEx3.R file**

Type the following in PvalueEx3.R file



Run your code

Output



**Push your work to GitHub**

**Make sure you are in**

Desktop\CS251\hop05-YourUserName

Run the following commands to push your work to the GitHub repository:

Open the terminal from the VSCode by hit the **control + ~** key and type the following command:

>>> git add .

>>> git commit -m “Submission for Module 5”

>>> git push origin master