Graphical user interface, text

Description automatically generated with medium confidence

Logo

Description automatically generated

**Abhilash Dikshit**

Course: ALY 6010

Instructor: Mohsen Soltanifar

Week 4: Final Milestone 2

Date: 2022/12/04

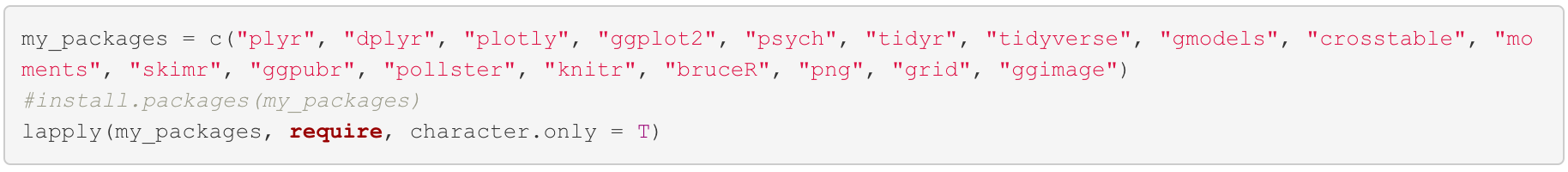
**Introduction**

Hypothesis testing is the art of testing if variation between two sample distributions can just be explained through random chance or not. If we have to conclude that two distributions vary in a meaningful way, we must take enough precaution to see that the differences are not just through random chance. A statistical hypothesis is an assumption made by the researcher about the data of the population collected for any experiment. It is not mandatory for this assumption to be true every time.Hypothesis testing, in a way, is a formal process of validating the hypothesis made by the researcher.

In this report, we will be conducting one sample and two sample testing on the data set used in Assignment 1. For one-sample t-test, we will identify the Life expectancy of each continent is greater than 60 or not and for two-sample t-test, we will identify if Asia has higher GDP per capita as compared to Europe. Later, we evaluate if Africa has higher population as compared to Americas using two sample t-test.

**Installation of R packages:**

Initially we installed the required package for the T-test and for visualisation.



Then we re-used the same data set from assignment 1 which we downloaded from the GitHub.



Top and Bottom 3 Of Cleaned Dataset

Graphical user interface

Description automatically generated with medium confidence

In order to clean up the data, it was sorted by descending *year*; column headers were converted to capital letters, and the regular expression was removed using gsub; column class type was changed from character to integer and the final stage was to add the life expectancy percentage column.

**Summary of Dataset:**

Table

Description automatically generated with low confidence

The mean population value across every country from the year 1952 till 2007 is 30 million and the median is 70 million. The minimum Life expectancy is 23 and maximum is 83, and the average life expectancy is at 59 while the median is 60.

### Part 1: One-sample t-test:

### Life expectancy of each continent greater than 60 or not?

Our null and alternate hypothesis are:

H0 = The Life expectancy of each continent is equal or lower than 60

H1 = The Life expectancy of each continent is more than 60

|  |  |
| --- | --- |
|  | The P-value of 0.001012 is lesser than significance level. Hence, we reject the null hypothesis H0 to identify if the Life expectancy of each continent is greater than 60. |
|  |  |

Boxplot representation of life expectancy of each continent:

|  |  |  |
| --- | --- | --- |
| Chart, box and whisker chart  Description automatically generated | | For the life expectancy report of each continent, first we identified the average life expectancy using mean functionand the output was 58.9689 which is denoted as Red Dash Line in boxplot.  Based on the data visualization from year 1952-2007, all the continents median values are above the average life expectancy except African continents due to scarcity of food and water that plays a major factor for the illness in that region. Moreover, both maximum and minimum life expectancy can be seen in Asia. |
| Here the density plot of Life expectancy for each continent has been shown because histograms will block each other. Due to the transparency, we can clearly identify Oceania and Asia were having higher expectancy as compared to other continents. We will be discussing more later using other plots by comparing the populations. |  | |

After that, we did statistical measures for this question.

We calculated the mean and median of Life expectancy: Text

Description automatically generated

Next, we identified the attributes that stored in the t-test of Life expectancy: Text

Description automatically generated with medium confidence

|  |  |
| --- | --- |
|  | Later, we extracted the conf.int, null.value and parameter as follows which shows the mean, the confident level and the parameter from the test |

**Part 2: Two-sample t-test**

### Question 2: Asia has higher GDP Per Capita as compared to Europe. Is it true?

Our null and alternate hypothesis are:

H0= Asia has higher GDP per capita as compared to Europe

H1= Asia does not have higher GDP per capita as compared to Europe

Firstly, we separated the vectors from the continent column as below:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

The P-value of 0.00000000000007902 is lesser than significance level. Hence, we reject the null hypothesis H0 to identify if Asia has higher GDP Per Capita as compared to Europe.

Chart, bar chart

Description automatically generated

After that, we did statistical measures for this question.

We calculated the mean and median of Life expectancy: Text

Description automatically generated with medium confidence

Next, we identified the attributes that stored in the t-test of GDP per capita For Asia and Europe:

|  |  |
| --- | --- |
|  | Later, we extracted the conf.int, null.value and parameter as follows which shows the mean, the confident level and the parameter from the test |

### Question 3: Africa has higher Population as compared to Americas. Is it true?

Our null and alternate hypothesis are:

H0= Africa has higher population as compared to Americas

H1= Africa does not have higher population as compared to Americas

Firstly, we separated the vectors from the continent column as below:

Chart

Description automatically generated with medium confidence

Text, letter

Description automatically generated

The P-value of 0.000001915 is lesser than significance level. Hence, we reject the null hypothesis H0 to identify if Africa has higher population as compared to Americas.

Chart, bar chart

Description automatically generated

After that, we did statistical measures for this question.

We calculated the mean and median of Life expectancy: A picture containing text

Description automatically generated

Next, we identified the attributes that stored in the t-test of Population for Africa and Americas:

|  |  |
| --- | --- |
|  | Later, we extracted the conf.int, null.value and parameter as follows which shows the mean, the confident level and the parameter from the test |

**Reference**:

1. Bevans ([2022, November 11](http://127.0.0.1:55106/rmd_output/3/#ref-R-Career));Datanovia ([2019, December 26](http://127.0.0.1:55106/rmd_output/3/#ref-R-Action));Brian Ripley ([October 12, 2022](http://127.0.0.1:55106/rmd_output/3/#ref-R-Cran));*RPubs - How Do i Get p-Values and Critical Values from r?* ([2017, March 1](http://127.0.0.1:55106/rmd_output/3/#ref-R-Material1));Investopedia ([2022, August 31](http://127.0.0.1:55106/rmd_output/3/#ref-R-Material2));*S.3.2 Hypothesis Testing (p-Value Approach) | STAT ONLINE. (N.d.). PennState: Statistics Online Courses.* ([n.d.](http://127.0.0.1:55106/rmd_output/3/#ref-R-Material3));*8. Calculating p Values — r Tutorial.* ([n.d.](http://127.0.0.1:55106/rmd_output/3/#ref-R-Material4))
2. *8. Calculating p Values — r Tutorial.* n.d. <https://www.cyclismo.org/tutorial/R/pValues.html>.
3. Bevans, R. 2022, November 11. *Hypothesis Testing | a Step-by-Step Guide with Easy Examples*. <https://www.scribbr.com/statistics/hypothesis-testing/>.
4. Brian Ripley, Douglas M. Bates, Bill Venables. October 12, 2022. *Support Functions and Datasets for Venables and Ripley’s MASS*. <https://cran.r-project.org/web/packages/MASS/MASS.pdf>.
5. Datanovia. 2019, December 26. *How to Do a t-Test in r: Calculation and Reporting*. <https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/>.
6. Investopedia. 2022, August 31. *What Is a Confidence Interval and How Do You Calculate It?*<https://www.investopedia.com/terms/c/confidenceinterval.asp>.
7. *RPubs - How Do i Get p-Values and Critical Values from r?* 2017, March 1. <https://rpubs.com/mdlama/spring2017-lab6supp1>.
8. *S.3.2 Hypothesis Testing (p-Value Approach) | STAT ONLINE. (N.d.). PennState: Statistics Online Courses.* n.d. <https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/hypothesis-testing/p-value-approach>.