Detailed Analytical Report on Assignment

Introduction

We carried-out a hypothesis testing for a sample of 4857 tuples taken from Panel Study of Income Dynamics(PSID) dataset. Chose a subsample of 2000 tuples to carryout the hypothesis testing considering the dataset with 4857 tuples as the population. Hypothesis testing is done to find evidence to bring a new hypothesis by rejecting already established hypothesis.

Sampling bias

The dataset contains data about individuals of 5000 families of United States in age range of 30-50. This dataset has a bias of individuals only limited to age range 30-50 and limited to the conditions and lifestyles of United States. This cannot be generalized to the whole population of the world. Also another fact is they only considered a 18000 individuals from 5000 families which is a very small proportion from the whole population of United States which even cannot generalize for United States.

Methodology

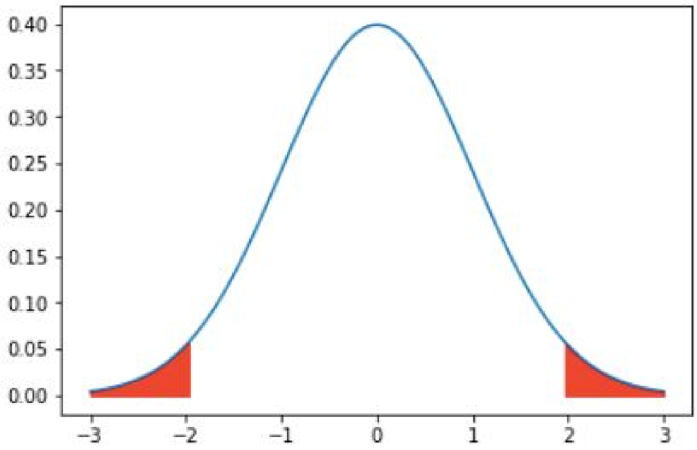
Hypothesis testing

The population we considered was dataset with 4857 tuples and the mean earnings for individuals in age range of 30-50 is 14245. So we came with the following null hypothesis(H0) and alternative hypothesis(Ha).

H0 : mean of earnings of individuals in age range of 30-50 equals to 14245
Ha : mean of earnings of individuals in age range 30-50 not equal to 14245

Since the chosen alternative hypothesis is with not equal condition we performed two tail test and a confidence level of 95% on the hypothesis made. So the level of significance (α) is 0.05. We wanted to compare the p value and the level of significance to come into a conclusion to see if we accept or reject the null hypothesis.

Here we took a sample size of 2000 samples which is greater than 30 samples allowing us to make the decision of selecting z distribution for the comparison according to the central limit theorem. Accordingly the z scores for the two tail ends are -1.96 and 1.96 respectively. The following diagram represents the z score values and level of significance in red colour.



After performing the hypothesis testing we were able to find the mean and standard deviation related to the subsample of 2000 tuples as 14375.5915 and 16144.80722600318 respectively. With these information we were able to compute the test statistic using the following equation.

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

The test statistic we obtained after the computation is 0.3667 which is in the range (0, 1.96) . The p value related to the test statistic is 0.71382. How we came into the conclusion will be discussed in Results section.

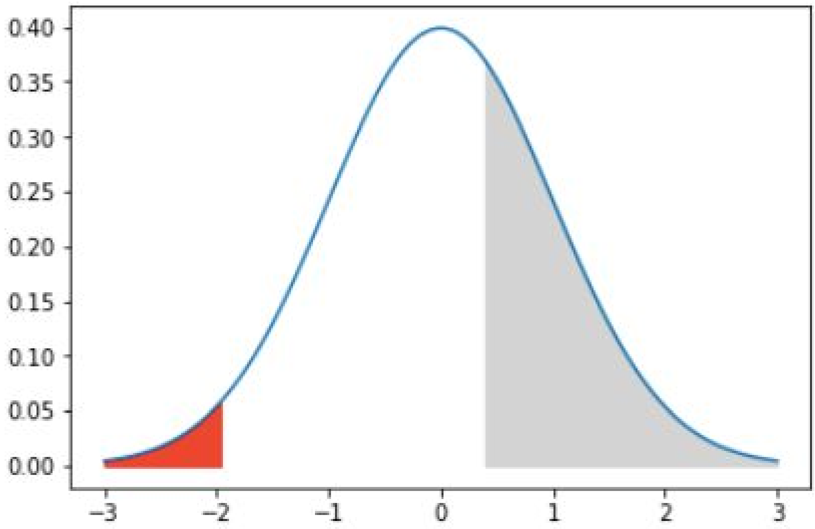
In order to perform the test we use python libraries like pandas, scipy, matplotlib and numpy.

Results

The p value obtained for the hypothesis testing is 0.71382 which is greater than the level of significance 0.05. So we failed to reject the null hypothesis. So we accept the null hypothesis which is the mean of earnings of individuals in age range of 30-50 equals to 14245.

Visualization

The computed p value is represented in the following diagram in light grey colour. We can see that the p value is greater than the level of significance value. So we came into the conclusion that we fail to reject the null hypothesis H0.



Discussion

We selected the dataset with 4857 tuples as the population but the true population is much more bigger than that. So this test doesn’t reveal the most accurate statistics and hypothesis testing results. But when we consider a sample which has features similar to the dataset we considered the results would be applicable.

The hypothesis we made was mean of earnings of individuals in age range of 30-50 equals to 14245 and by computing the test statistic and the confidence interval we chose we came into the final decision that we failed to reject the null hypothesis, hence we accepted the null hypothesis.

References

[1] http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTest-Means-Proportions/BS704_HypothesisTest-Means-Proportions3.html

[2] https://psidonline.isr.umich.edu/

[3] https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/

[4] https://en.wikipedia.org/wiki/Z-test

[5] https://en.wikipedia.org/wiki/Probability_density_function