MVE051 Lab 1

Erik Johnsson, Vidar Magnusson

May 2019

1 Thesis

Our hypothesis is that in general most employees at Microsoft work less on open source projects (public repositories) during the summer compared to the rest of the year. The summer was decided to include the days between the 157th day of the year to the 244th day of the year. This generally being (with exception to leap years) the 6th of June to the 1st of September.

2 Process of gathering data

To draw conclusions to support our thesis we first needed to gather a sufficient amount of data for it to be reliable. To do this we decided to use the Github api to gather information about members of the Microsoft Organisation on Github. This made available to us a large amount of information, entire list can be seen here, we decided to use the contributions data. It should be noted that only contributions to public Github repositories could be gathered due to the rest being protected for privacy reasons. The data we received was then reformatted into a file where for each user, for each day of the year we had the total number of contributions done by that person on that day of the year for all years since they created their Github accounts. This data file will be an annex to this document. Finally we calculated the average number of contributions done for each person during the summer days (defined above) and then the average number of contributions during the entire year. If then the average number of contributions throughout the year were higher than the average during the summer, the person was assigned a 1 otherwise a 0. It is from this data we did the rest of our analysis.

3 Data

Summary of the data gathered, for more explanation see "Process of gathering data".

Total users surveyed: 210

Number of users who worked less during the summer compared to the rest of the year (Success): 140

Number of users who didn't (Fail): 70

If success = 1, fail = 0 $\rightarrow \bar{X} = \frac{2}{3}$.

4 Statistical analysis

So our hypothesis was that most employees observed would work less during the summer than the rest of the year therefor our alternative hypothesis is the same as saying that our population mean should be larger than 0.5, i.e. $H_1 = P > 0.5$. From this we can see that our null hypothesis must be less or equal to 0.5 (meaning that most people worked more or as much during the summer) i.e. $H_0 = P \leq 0.5$. We also decided that we wanted a certainty of at least 99.9% to make sure that we could be relatively certain that we drew the right conclusion.

We begin by assuming that our null hypothesis is correct, we then calculate the variance of the null hypothesis. To make sure that our null hypothesis has the best possible chance of succeeding, we assume a population proportion of 50% or 0.5 is the value of p (the critical value for H_0). As our data is Bernoulli distributed, our assumed standard deviation becomes the root of the variance for the Bernoulli distribution $\sigma_{H_0} = \sqrt{p(1-p)} = \sqrt{0.5 \cdot 0.5} = 0.5$. From this we know that the sample means are normally distributed with the parameters $\mu_{\bar{X}} = p$ and $\sigma_{\bar{X}} = \frac{\sigma_{H_0}}{\sqrt{n}} = \frac{0.5}{\sqrt{210}}$
The sample mean $\bar{X} = \frac{2}{3}$. To calculate the chance of getting a sample mean which is $\leq \bar{X}$ we calculate the z statistic which is how many standard deviations away is from the mean $\mu_{\bar{X}}$. $z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\frac{2}{3} - \frac{1}{2}}{\frac{0.5}{\sqrt{210}}} = 4.83$
From this we can look in a z table to find the 99.9% standard deviation interval, we found this to be 3.08. Since $4.83 > 3.08$ we reject the null hypothesis due to it being outside the accepted interval, this means that we accept our alternative hypothesis with a certainty of at least 99.9%.

5 Conclusions

Since our alternative hypothesis was accepted with a certainty of at least 99.9% we conclude that our hypothesis can be assumed to be true. This means that we conclude that a majority of the members of the Microsoft Github organization work less on open source projects during the summer than the rest of the year.