VRIJE
UNIVERSITEIT
BRUSSEL

METHODS FOR SCIENTIFIC RESEARCH

ASSIGNMENT-2

Submitted by: Deepika Chandrababu
Student Id: 0572230

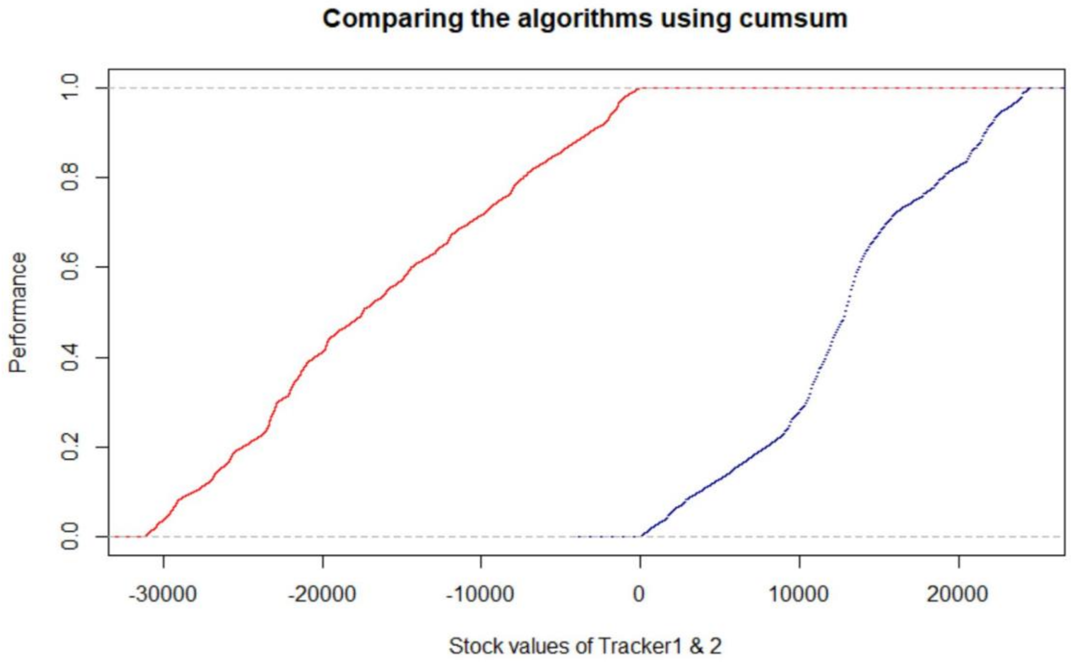
19 December 2021

Question 1_Dataset 3

In this experiment, there are two independent interval variables(Tracker1 , Tracker2) that holds the stock values of two trackers.

1. Cumulative sum Plot

The distribution of both the trackers was heavy tailed. To make the spread of the data points more readable, the cumulative sum is calculated, which holds the total sum of data as it grows. For example, if the first value is a, then the second value will be a+b, and the third will be a+b+c and so on. Since the stock values are random, each data point can be summed up to achieve a smooth curve. It is also used to monitor the total contribution of the given variable.

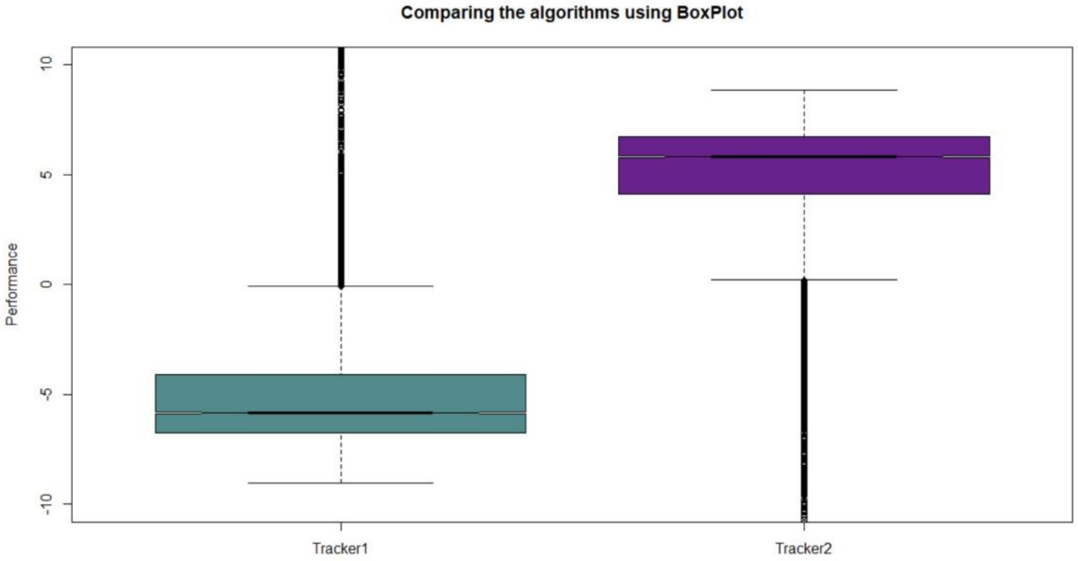


It is clearly seen that both the trackers data points are different from their spread (range) of the stock values.

Tracker 1(red line) leads to losses as it falls under the negative values of the stocks. Whereas the Tracker 2 (blue line) grows from 0 to 20000 meaning it has extreme profits eventhough there are breaks in between which causes losses. These are rare events which could have happened because of the outliers in large sample. Tracker 2 has most profits and sometimes loses which are seen in breaks (points with space) in the blue curve. Tracker 2 ends up with a positive stock value in the end while the other one ends up with a negative stock value.

2. Box Plot

Boxplots are useful to show the spread and centers of the dataset as it gives the lower whisker, lower quartile, median, upper quartile and upper whisker values. Here, boxplot is used to compare the median lines of both the trackers to spot if there are any difference between them.



It is clearly seen that, The median lines of both the box plots are not equal. Hence they are different.

- For tracker 1, The mimimum and maximum values are seen in the whiskers, the lower whisker is found to be (-9) and the higher wishker(0). The lower quartile lies in -7 and the median line between the box indicates the average value which is -6.5 for first tracker. The upper quartile is found to be -4.
- For tracker2, the lower whisker is found to be (0) and the higher wishker(9.5), the lower quartile is 3.5 and upper quartile is 6. The median line lies in 5.5
- Most of the stock values are numerically distant from the rest of the data which are the outliers.

The stock values of tracker 1 are on the negative side while the stock values of tracker 2 on the positive.This indicates that tracker 2 has most profits and miminum losses.

Wilcoxon Rank sum test

Wilcoxon rank sum test is used to compare the indepedent groups of samples. It can be done without the assumption that the data is normally distributed. It is a non-parametric test with the null hypothesis that the probability of x>y is equal to the probability of y>x, where x and y are randomly collecedsamples from the distribution.

The test is performed using the r-command,

```
wilcox.test(data$Tracker1, data$Tracker2, paired=FALSE,
alternative="two.sided", conf.level=0.95)
```

The results obtained were,

W = 9079097, p-value < 0.05

p-value is less than 0.05, hence there is significant difference between the two trackers.

Effect size: Cliff's delta

Cliff's delta is calculated by the formula,

2W/(n.m)

using the r-command ,

```
cliff.delta(data$Tracker1, data$Tracker2)
```

The value of delta from the test is as follows,

delta estimate: -0.8184181 (large)

There is a Large effect which means the difference is important.

Difference in terms of statistics:

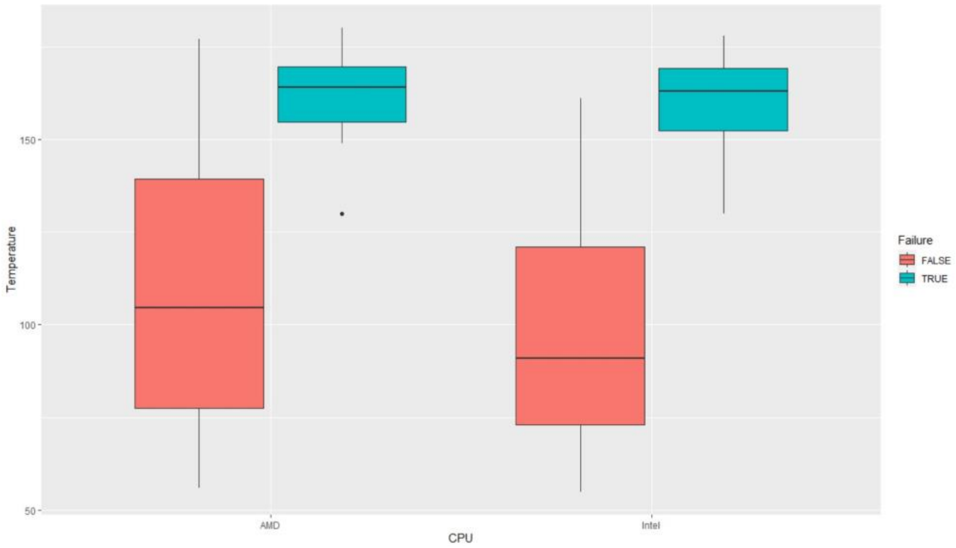
The wilcoxon test was performed under the assumption of the null hypothesis that the stock values of tracker 1 and tracker 2 are equal.

The test statistic of wilcoxon rank sum test states that "alternative hypothesis: true location shift is not equal to 0" which implies the distribution is either skewd to the left or to the right from the other. Hence, the medians for both the distribution are different. Moreover, the exact difference were seen in the box plot where the median of tracker 2 was on the positive range and the median of tracker 1 was on the negative range. Also, the p-value is less than the significance level alpha = 0.05, which rejects the null hypothesis and concludes that the median of tracker 1 is statitically different from the median of tracker2.

To conclude, the best tracker is tracker 2 which holds positive stock values and that can be used to train the model to maximize return.

Question 2_Dataset 4

In this experiment, there is one nominal independent variable(CPU), one interval indepedent variable(temperature) and one dependent variable (Failure).



Making a boxplot to better understand the outcome of the binary variable. Both the CPU brands have the failure rate as false below the temperature of 150 degrees centigrade. Which means they do not fail below a certain temperature and they seem to fail above the temperature of 150 degrees centigrade.

To test whether the CPU brands influence the failure rates, Generalized linear model is choosen because the output variable has only two different values true or false.

Generalized linear model is applied using the r command,

```
Model <- glm( formula=Failure ~ Temperature+CPU, data= ques2, family = binomial)
Summary(model)
```

where,

the formula depicts that the temperature and cpu predicts the Failure, data is the dataset and family is the output binary variable.

The results obtained were,

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-17.87053	3.11543	-5.736	9.69 x 10^{-9}
Temperature	0.11192	0.01983	5.644	1.66 x 10^{-8}
CPUIntel	1.47443	0.59905	2.461	0.0138

Fitted Model:

Failure = -17.9 + 0.1*Temperature + 1.5*CPUIntel

The p-value of Temperature is less than 0.05 which shows that temperature is statistically significant and influences the failure rates. While the p-value of CPUIntel is 0.01. Thus, the brands of CPU does not influence failure rate.

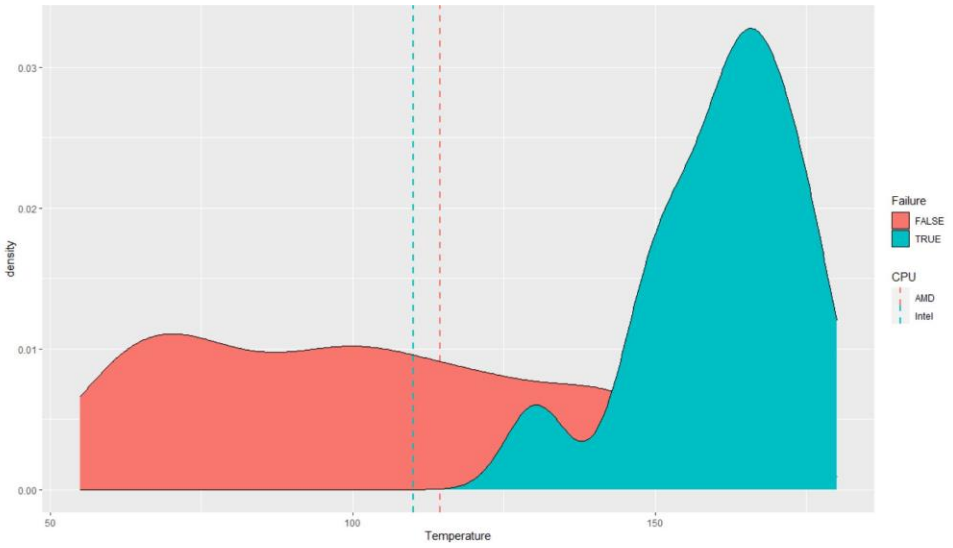
Effect size: odds ratio

Odds ratio is the probability of one outcome divided by the probability of the alternative outcome, which is calculated by the formula,

Odds = p/(1-p)

- For every independent variable there is an associated effect size (odds ratio) from the fitted logistic model.
- The beta value of the slope associated with temperature is 0.1, so the **odds ratio** will be **$e^{0.1}$**
- The beta value for CPUIntel is 1.5, so the **odds ratio** will be **$e^{1.5}$**
- The probability of 0.1 means that there is 1 in 10 chance of event occurring.
- On the other hand, the odds ratio greater than 1, in our test (1.5) mean that the event is more likely to occur in the first group.
- Hence, temperature influences the failure rate.

To check the reliability of the CPU brands, a density plot was made to visualise the distribution with the median lines for both the brands of the CPU



It is seen from the median lines that when the temperature is 125 degrees centigrade, the failure rate for both the brands are found to be false. Hence, these brands of CPU fulfil the US military specifications by having lesser failure rates at this temperature.

Question 3_Dataset 1

In this dataset, there are independent ≥2 nominal variables (type1,type2,type3) and dependent ordinal variables which are ordered from A being best to F being worst.

Since each customer is asked only once, the samples are independent between type1,type2 and type3. To test if the user experience of different varaints are different, kruskal walli test is conducted as this test is generally used in experiments of three or more groups to decide whether the variables are identical. Assuming the null hypothesis is that the user experience for the different variants is similar.

The r command to compute kruskal wallis chi squared test,

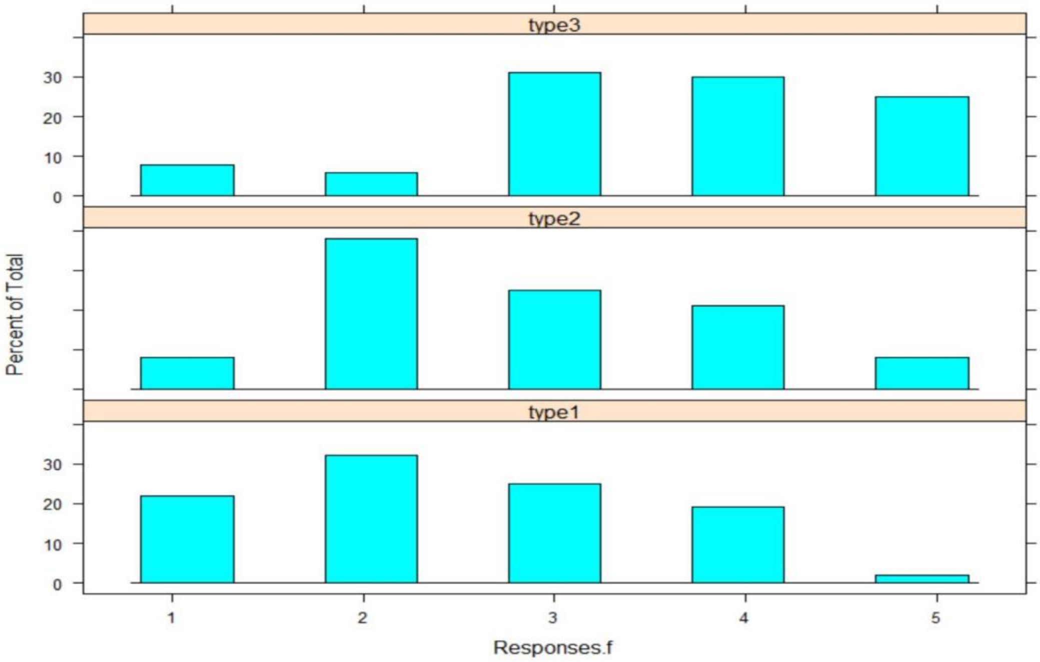
kruskal.test(formula=Responses~Interfaces, data=data3)

The results obtained were

Kruskal-Wallis, χ^2 (2)=43, $p<0.05$

From the results of kruskal test, $p<0.05$ which rejects the null hypothesis and concludes that the medians of different groups are different and user experiences of the different variants are different.

A barplot is created to visualise the spread of the dependent variable in different groups such as A=1,B=2,C=3,D=4,F=5, between the three variants type1, type2, type3.



The bar plot also confirms that the expereinces for different variants are different. Each type has different lengths of the bar for each grading 1 to 5 except for type 1 and type 2 in score 3 which are equal. It is also seen that the distribution of type 3 is skewed towards the right (worst scores), type 2 is normally distrbuted and type 1 is skewed towards left (best scores).

Effect size:

Effect size for Kruskal-wallis test is calculated using the formula

$$\boldsymbol{\eta^2 = (H - k + 1)/(n - k)}$$

Where,
k is the number of groups, here k=3 since we have 3 variants.
n is the number of observations, N=300 since our sample has 300 obs.
H is the value obtained in the Kruskal-Wallis test, here H= 43

The effect size obtained is

$$\boldsymbol{\eta^2 = 0.14}$$

The interpretation values ≥ 0.14 are condisered to have large effect from this test. A large effect size has been detected and so the difference is important.

Contingency table

Contingency table was created to interpret the mimimum and maximum responses for each group under each type. In this way, we can easily find the differences under each category in numbers.
The table is formed using the r command;
xtabs(~ Interfaces + Responses, data = question3)

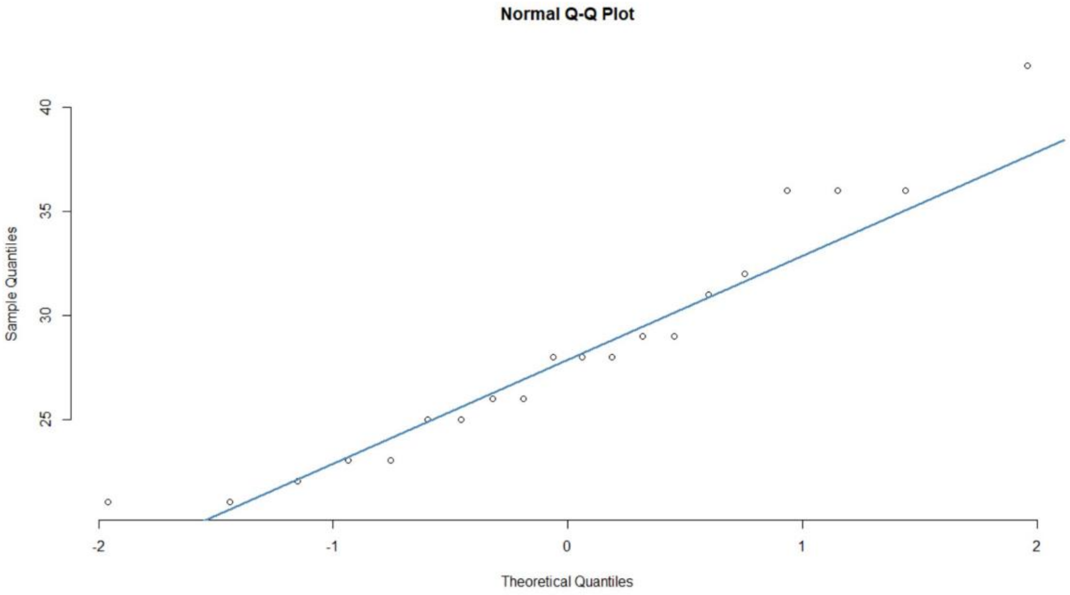
	Responses				
Interfaces	A	B	C	D	F
Type1	22	32	25	19	2
Type2	8	38	25	21	8
Type3	8	6	31	30	25

From the contingency table, it is clear that type 1 is the best variant which has many responses falling under the best category (A=22, B=32) and very less (F=2) worst experiences, since the grading goes from A as the best score and type 3 is the worst variant as it has most scores on the worst experiences category (F=25) than on the best scores (A=8).

Question 4_Dataset 1

The dataset had the performance of ten programmers completing the assignments with and without static analysis. It had 2 independent nominal variable(yes/no) and a dependent interval variable(Times). So the appropriate test should be t-test that can check the signifance.

To perform the t-test, the data should be normal distributed. Lets make qqplot to see the distribution of the data points.



Most of the data points are close to the straight line. However, they could be outliers.

Hence, the normality was also checked using **Shapiro-Wilk** normality test and the results were

W = 0.93123, p-value = 0.1631

p > 0.05 which concludes the data is normally distributed and a t-test is appropriate for this type of data. Since the data is collected from the experience of same people, a paired t-test is conducted using the r command,

```
yes<-question4$Times[question4$Static_Analysis=="yes"]
no<-question4$Times[question4$Static_Analysis=="no"]
```

t.test(yes,no,var.equal = TRUE)

The results obtained were,

$t(18) = -2.7699$, $p<0.05$

The result conforms there is statistically significant difference with and without static analysis. The p-values is less than 0.05 which rejects the null hypothesis and accepts the alternate hypothesis. Also, the smaller the t-value, the more similarity exists between yes and no.

Effect size: Cohen's d

The formula for Cohen's d (for equally sized groups) is the difference between the means of two groups divided by standard deviation.

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\overline{s}}$$

Where,

$\overline{x}_1$ is the mean of group 1. In this test, the subset of sample belonging to yes.

$\overline{x}_2$ is the mean of group 2. In this test, the subset of sample belonging to no.

$\overline{s}$ is the pooled standard deviation of the two groups calculated by the formula.

$$\overline{s} = \sqrt{\frac{(n_1 - 1)\overline{s}_1^2 + (n_2 - 1)\overline{s}_2^2}{n_1 + n_2}}$$

The results obtained;

Cohen's d= 1.8

The effect size is larger than 1.20 which is very large effect for cohen's d and almost huge which implies that the mean values of the two groups differ by 1.8 standard deviations.

The estimated sample size to achieve power of 80% with p-value 0.05 is n = 6

The r command used was,

```
pwr.t.test(d = 1.8, sig.level = 0.05, power = 0.8)
```

The original data has a sample size of 20 but to achieve a power of 0.8 with alpha 0.05 we need a sample size of 6.

Estimating sample size using Leave one out procedure:

Estimating the sample size leaving one data point in each subset, The samples are subsetted using the following r-command,

```
subsampleyes1 <- yes[-1]; subsampleno1 <- no[-1]
subsampleyes2 <- yes[-2]; subsampleno2 <- no[-2]
subsampleyes3 <- yes[-3]; subsampleno3 <- no[-3]
subsampleyes4 <- yes[-4]; subsampleno4 <- no[-4]
subsampleyes5 <- yes[-5]; subsampleno5 <- no[-5]
subsampleyes6 <- yes[-6]; subsampleno6 <- no[-6]
subsampleyes7 <- yes[-7]; subsampleno7 <- no[-7]
subsampleyes8 <- yes[-8]; subsampleno8 <- no[-8]
subsampleyes9 <- yes[-9]; subsampleno9 <- no[-9]
subsampleyes10 <- yes[-10]; subsampleno10 <- no[-10]
```

For each small subset, power analysis was done as mentioned the above and the table concluding the p-values, effect size, sample size are listed below.

Sample	t-value	p-value	Effect size (d)	Estimated Sample Size (n), when p-value=0.05, Power = 0.8
Sub sample1	-2.7196	0.01	1.8	6
Sub sample2	-2.5831	0.02	1.7	7
Sub sample3	-2.1892	0.04	1.5	8
Sub sample4	-2.3676	0.03	1.6	7
Sub sample5	-2.4846	0.02	1.7	7
Sub sample6	-3.281	0.004	2.2	4
Sub sample7	-3.0837	0.007	2	5
Sub sample8	-2.5831	0.02	1.7	7
Sub sample9	-2.8235	0.01	1.9	6
Sub sample10	-2.2103	0.04	1.5	8

The significant differences (p-values) are gradually decreasing between each small subgroup leaving one data point.

The relationship between effect size and sample size are the contrary, i.e A smaller cohen's d indicates the necessity of larger sample sizes and a larger cohen's d indicates the neccesity of smaller sample sizes.

From the estimated sample sizes of each small subgroup, it it clear that the sample size estimates for this pilot experiment completely relied on the effect size of the subgroup since even a small variability in the effect size influenced the sample size to increase or decrease.

To conclude, there is no stability in the estimated sample sizes from the small group. Thus, from a very small data of only ten programmers, the sample sizes measured are not reliable.