Homework 10

David Östling
dostl@kth.se
DA2210

December 1, 2021

1 Confidence intervals and statistical significance

a) To begin with there will always be a sampling error at play here, 100 samples are way too few in order to be able to draw any spot on conclusion on what is being experimented on. In order to estimate how accurate our proposed mean value is expected to be we can turn to a confidence interval. Using this interval we can find a set of values where the exact mean lies somewhere in between. If we know take the provided samples and calculate the sample mean we get the following for research group 1: 9,86890326. The width of the confidence interval depends on the sample variation and the amount of samples. Now, we just need the standard deviation, which after calculation in excel came out to be: 1,426644728. We can now find the margin of error as well as the range for the confidence interval using the following formula [1]:

$$\overline{X} \pm t\frac{s}{\sqrt{n}}$$

Figure 1: The formula for finding the confidence interval where X = sample mean, t = standard error, s = standard deviation, n = sample size

For a research paper I would like to be pretty confident that my mean is rather close to exact, in my eyes 95% accuracy (confidence level) should be enough [1, 2]. Now, as we have everything for the formula except t we will need to find it by using a *t-distribution table*, this gives us t = 1,984 [2]. If we place each value into the formula we can calculate the confidence interval and we get:

$$9,86890326 + 1,984 * \frac{1,426644728}{\sqrt{100}} \approx 10.152$$

$$9,86890326 - 1,984 * \frac{1,426644728}{\sqrt{100}} \approx 9.586$$

Figure 2: Returns the confidence interval where the upper bound is 10,152 and the lower bound is 9,586

From the above calculation we know that the confidence interval is between 9,586 and 10,152 and that with 95% accuracy the exact mean value can be found somewhere in between these values.

b) Now we simply do the same thing as we did in a and try to spot any potential differences, we calculate all the values for the formula in excel and find the following:

$$9,63796179 + 1,984 * \frac{1,949211388}{\sqrt{100}} \approx 10.02$$

$$9,63796179 - 1,984 * \frac{1,949211388}{\sqrt{100}} \approx 9.251$$

Figure 3: Returns the confidence interval where the upper bound is 10,02 and the lower bound is 9,251

Judging by the calculation the confidence interval here is between 9,251 and 10,02 and that with 95% accuracy the exact mean value can be found somewhere in between these values. This is a slightly different interval though when compared to what we found in a). This can be partly explained by the sample size being so small (100 is not sufficient for any exact results) but also that the variety in the sample values were rather different as seen in the graph below *(see figure 3)*.

c) After putting the values into excel yet again we got the following mean: 12,465067, median: 7,580873671. They are different simply because the mean is the average for all sample values whereas the median is the value in the middle of all of them. If the sample values had been exceedingly similar then the median could be the same as the mean but this is not the case here. The variety of the sample values here is way too large for that to even be a

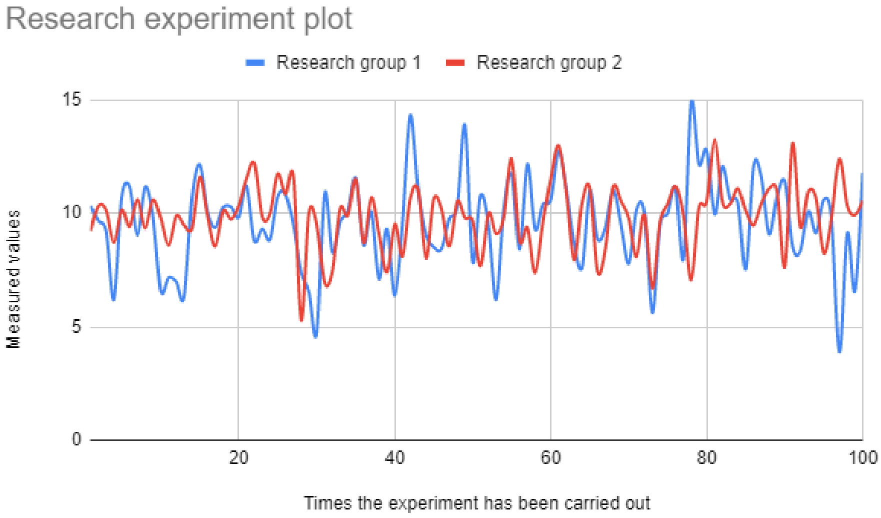

Figure 4: A graph of the research group tests showing the variety in samples

possibility. The histogram below can be used to see the exact distribution between the sample values, judging from the picture below the most common values would be between 0 to 10:

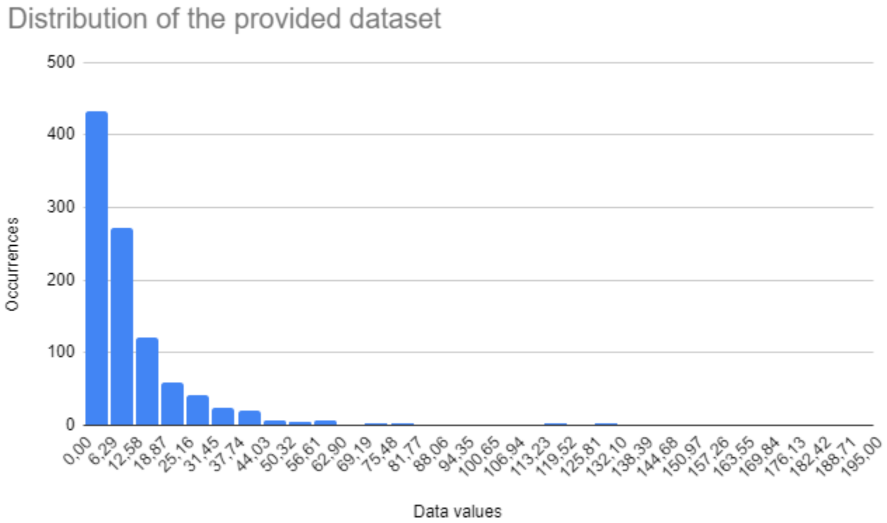

Figure 5: A histogram showing the distribution of samples

Now as for calculating the confidence interval using the median we can use the formulas below [3]:

$$nq - z\sqrt{nq(1-q)}$$

$$nq + z\sqrt{nq(1-q)}$$

Figure 6: A formula for the confidence interval for the median whereas n = amount of samples, z = critical value (very much like t in the formula for the confidence interval for a mean, q = quantile interest in this case being 0,5 as we have a median

As we have a median we know that q = 0.5 (50% below the median and 50% above the median) and that we have n = 1000 samples. Now, if we want to get a 95% accuracy (confidence level) as last time we need to use z = 1.96 [3]. If we put it all together we get the following:

$$1000*0.5+1.96*\sqrt{(1000*0.5)*(1-0.5)} \approx 531$$

$$1000*0.5-1.96*\sqrt{\left(1000*0.5\right)*(1-0.5)} = 500$$

Figure 7: A calculation which returns the index for the element of the upper bound: 530 and the lower bound: 500 for the confidence interval for the median

Now we simply need to sort the data set using e.g. Python and check what elements reside in each of the indexes found by the above calculation. After sorting we can clearly see that they are: 8.2156178955294 (index = 531, upper bound) and 7.56690999842661 (index = 500, lower bound). This means that the confidence interval for the median is between 7.56690999842661 and 8.2156178955294 with 95% accuracy. The median: 7,580873671 we found is within this range but is very close to the lower bound.

2 Reproducibility in research

Chosen article:) Comparing anomaly-detection algorithms for keystroke dynamics [4]

I believe they could have been more precise about what data sets they used for the study. They do still describe how they achieved what they achieved in the study but they do not (at least from what I can see) show what exact data sets were used. I believe this could pose a rather large obstacle when trying to replicate the study even if they did describe how the sets were trained and implemented. Another factor is the running subjects, they recruited 51 people from one university, 30 males and 21 females where 8 of them were left-handed and 43 of them were right-handed [4]. While this is a pretty accurate description of which people were tested it could be pretty difficult to replicate. If all the people were from the same university then it could mean that they have more things in common when compared to other potential subjects thus it becomes harder to replicate the test. In my eyes they should have been more generous with their data sets and perhaps even shared a link to an excel file or similar so that it would have been easier for people to replicate their study. Although I do still think that it is good enough to get an approximate replication of what was achieved here.

References

[1] Hazra, Avijit. ”Using the confidence interval confidently.” Journal of thoracic disease 9.10 (2017): 4125.

[2] Nakagawa, Shinichi, and Innes C. Cuthill. ”Effect size, confidence interval and statistical significance: a practical guide for biologists.” Biological reviews 82.4 (2007): 591-605.

[3] Zach, et al.“How to Find a Confidence Interval for a Median” Statology, May. 2021, [Available:] https://www.statology.org/confidence-interval-for-median/

[4] K. S. Killourhy and R. A. Maxion, ”Comparing anomaly-detection algorithms for keystroke dynamics,” 2009 IEEE/IFIP International Conference on Dependable Systems Networks, 2009, pp. 125-134, doi: 10.1109/DSN.2009.5270346.