NTNU - NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET
Faculty of Engeneering Science and Technology
Department of Civil and Transport Engineering
TBA4925 - Master Thesis

Optimizing the micro-tasking workflow and exploring it's usage potential within geospatial data

Anne Sofie Strand Erichsen
Trondheim, June 2017

DAIM page

Background

HEI

Task Description

The micro-tasking method is becoming more and more popular. Companies like Amazon develop micro-tasking web applications where people can earn money by doing micro-tasks for others. The method is used for tasks that involve both use of technology and a large number of people. By using the micro-tasking methodology, this thesis aims to study how people solves micro-tasks within geospatial data imports, which is a very complex and large process.

This study will have an emphasis on the data validation and conflict handling part of the import. These parts are complicated to do fully automatic through scripts. By varying the number of objects to solve at a time, adding rewards on some tasks, among other factors, the study will hopefully find a significant approach to prefer when using the micro-tasking method within geospatial data. What are the number of objects optimal within a task to get it completed as quickly as possible? Does the quality of the work vary between the different tasks given? Do amateurs manage to do the tasks? Do rewards have an impact on how the tasks are solved?

This thesis will also explore the micro-tasking methods usage potential within geospatial data. Can other organizations doing a process that needs humans to interfere take advantage of this method? An example is OpenStreetMap, who has taken good advantage of the method both in mapping and import projects.

Specific tasks:

- Study related literature
- Do a micro-tasking survey
- Examine how many elements are optimal when creating geospatial micro-tasks

Abstract

This paper propose a method for extracting buildings in satellite photos. The proposed network makes use of a digital surface model and multispectral satellite data. It

Sammendrag

Sammendrag på norsk

Preface

This paper is a master thesis written for the Department of Civil and Transport Engineering at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. It is a part of the study program Engineering and ICT - Geomatics, and was written in the spring of 2017.

I would like to thank my supervisor Terje Midtbø for his help and feedback, and also Atle Frenvik Sveen for his support and help every time I needed it.

Trondhiem, 2017-06-16?
Anne Sofie Strand Erichsen

Contents

Abstract . . . v

Sammendrag . . . vii

Preface . . . ix

1 Methodology and experiment . . . 1
- 1.1 Survey . . . 1
- 1.2 Web-application . . . 1
- 1.3 Pilot test . . . 1
  - 1.3.1 Execution of the pilot test . . . 2
  - 1.3.2 Results from the pilot test . . . 2
- 1.4 Sample Size . . . 4

2 Statistics . . . 7
- 2.1 Sample data from Survey . . . 7
- 2.2 Statistics theory . . . 7
  - 2.2.1 Normal testing . . . 7
  - 2.2.2 Hypothesis testing . . . 9
- 2.3 Survey results . . . 12
  - 2.3.1 Bar Charts . . . 12

Appendices . . . 13

A Tets . . . 15

1 Methodology and experiment

1.1 Survey

The survey is a part of this thesis. The survey is used to answer different hypothesis around geospatial micro tasks. To be able to answer the hypothesises three tasks containing the same two questions was developed. The questions will represent two different micro-tasks, while the three tasks will contain one, three and six elements that the participant will use to answer the questions with. The participant will always answers the two question's on six elements, but the tasks vary how many element's need's to be handled at the same time. The variation of number of elements in the task's is to hopefully find out if the number of elements in an micro-task effects how well people solves the task.

There are two variable types used in this survey, dependent- and independent variables. The dependent variables are time, number of correctly chosen elements in both questions and also difficulty. The independent variables are number of elements in the task, experienced or non-experienced participant, gender, age and if the participant know micro tasking.

1.2 Web-application

This thesis used an online web-based survey to conduct the experiment. An online survey avoid the cost and effort of printing, distributing, and collecting paper forms. Many people prefer to answer a brief survey displayed on a screen instead of filling in and returning a printed form (Ben and Plaisant, 2009).

In a self selected sample, which is some the case here, there is potentially a bias in the sample (Ben and Plaisant, 2009).

1.3 Pilot test

It is important to pilot test the survey prior to actual use (Ben and Plaisant, 2009). A pilot test provides an opportunity to validate the wording of the tasks, do the participants understand the tasks? It also helps understand the time necessary for completing the survey, which should be communicated to the participants in prior to the survey (Schade, 2015). The pilot-test will be conducted with a small sample of users. Results from the pilot test kan be used to do improvements to the actual survey, to the web-application hosting the survey and to find errors or weaknesses in the databasemodels.

A pilot test was conducted with a total of eight participants, five experienced and three non-experienced participants aged from 22 to 64 years.

After the pilot test the usability was measured. The standard ISO 9241-11 suggests that measures of usability should cover effectiveness, efficiency and satisfaction (ISO, 1998). Measuring these three classes of metric can vary widely and makes it difficult to make comparisons of usability between different systems. "[...] just because a particular design feature has proved to be very useful in making one system usable does not necessarily mean that it will do so for another system" (Brooke, 1996). Usability in this thesis will be measured with the *System Usability Scale*(SUS) because it gives an subjective measure of usability. The *System Usability Scale* questionnaire consists of ten statements where the participants rate their agreement in an five-point scale (Ben and Plaisant, 2009). Subjective measure of usability is usually obtained through the use of questionnaire and attitude scales (Brooke, 1996). SUS was developed to be quick and simple, but also reliable enough to be able to compare performance changes between versions (Brooke, 1996). It is also easy to administer the participants through the usability test and it can be used on small sample sizes an still give reliable results (Affairs, 2013).

The usability is important to measure. If the participants doesn't understand how the web-application works, they will probably not do the survey since they then have to invest time in understanding what to do. It is also important to get enough participants to do the whole survey and not quit halfway in frustration of not understanding it properly. The *System Usability Scale* can effectively differantiate between usable and unusable systems (Affairs, 2013).

1.3.1 Execution of the pilot test

The pilot test started with a brief information about this study and the survey. They where told to talk out load during the survey, no help or guidance was given to the participants. The author observed the participants while they conducted the survey. The author took notes and watched if the participants understood the questions in the survey correctly. After the survey a *System Usability Scale* questionnaire was answered by the participants. At the end the participants was asked to give general feedback on the web-application. The SUS score and the feedback was then used to determine the usability of the web-application and to determine which improvements to be done.

1.3.2 Results from the pilot test

- Did someone knew micro-tasking? Can we see something here?

The average SUS score was 84.64 out of 100.

All participants thought that the instruction movie was confusing. It was short, the instructions went too fast and it lacked voice descriptions. The movie needed major improvements.

Overall feedback on the tasks was that it was difficult to understand which building was which and also if a building layer was selected or not in question one. The lack of labels on the buildings was done on purpose to get the task as much as possible realistic. The prosess of selecting the best fitting shape needed improvements, it had to be made clearer that selection was done by clicking on the shape, not by using the layer control as some thought.

Also, some pages had too much information and long sentences. The task progress bar was removed, no one noticed it, only the survey progress bar on the top right is necessary.

The two oldest participants spent almost twice as much time on the test than the younger. Maybe it where too much cognitive load on them. Learning a new application and at the same time understanding how to do the survey and answer the questions given to them. One of them where experienced and the other non-experienced, so this is a surprising result. CHECK THE TIME ON EACH TASK FOR THE OLDER PARTICIPANTS. Figure 1.1 show the task results from all participants ordered by age. There are three entries per participant, so three and three bars are results from the same participant.

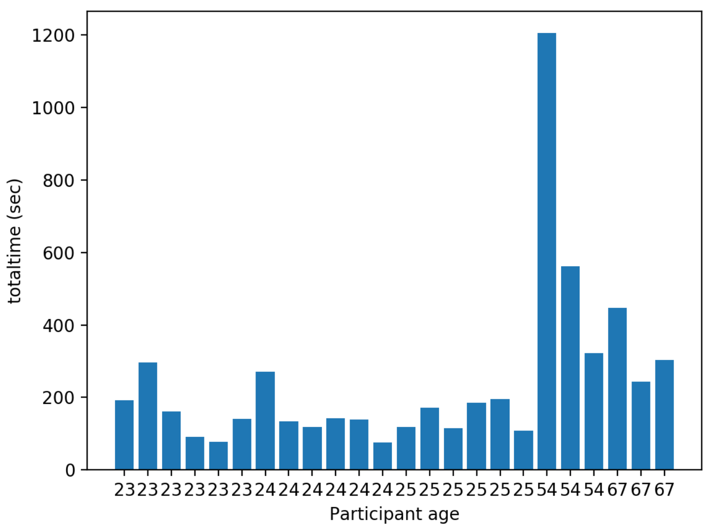

Figure 1.1: Total time - all participants ordered by age

The average time spent on the survey was 18 minutes. The two oldest participants used on average 33 minutes, while the rest of the participants spent on average 13 minutes to complete the survey.

The pilot test the data was used to test some of the hypothesis to find errors or weaknesses in the databasemodel. The data was extracted with the help of Django QuerySets and saved in csv files. There were a few errors and weaknesses found, most of them are listed under:

1. Add foreign key from TaskResult model to TaskSurvey
2. Added four other fields in TaskResult model
    - Total correct elements
    - Task order
    - Task number
    - List of correctly chosen building numbers in both questions

The additional fields will maninly help with creating plots to interpret the data better and visualize the different results.

1.3.2.1 Statistics result

First the pilot-test result need to be normality tested. As mentioned in 2.2.1 this thesis will use the Anderson-Darling test. The python library Scipy has an Anderson-Darling test function (The Scipy community, 2017), this was used to answer the normality test with the Anderson-Darling hypothesis. Testing the total time data form all participants (in total 32 entries) in Anderson-Darling gave a *p-value* of 0.717. The null hypothesis cannot be rejected, then there are two possible cases. One can either accept the null hypothesis or the sample size is not large enough to either accept or reject the null hypothesis (The Pennsylvania State University, 2017). An acceptance of the null hypothesis implies that the evidence was insufficient, the result does not necessary accept H_0, but fails to reject H_0 (Walpole et al., 2012).

Anderson-Darling test
Data: All participants - total time, 32 entries
P-value: 0.717
P-value > 0.05
H_0**:** Accepted, failed to reject

1.4 Determining the sample size

The sample size is influenced by a number of factors, including the purpose of the study, population size, the risk of selecting a "bad" sample and the allowable sampling error (Israel, 1992). In this survey there are three possible ways of determining the sample size.

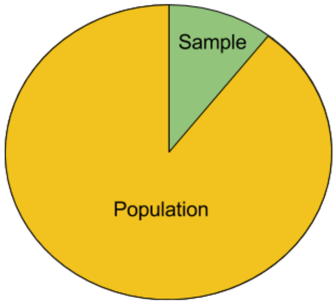

Figure 1.2: Population vs. sample

A sample is a collection of observations and is the subset of a population, illustrated in figure 1.2. The population size in this survey is not easily determined. A population is the collection of individuals of a particular type (Walpole et al., 2012). All individuals with access to a computer and internet interested in contributing to micro-tasks is basicly the population.

There are three possible ways of determining the sample size in this study. The first option is to use a sample size from a similair study. The risk is to repeat errors that were made in determining the sample for another study. The second option is to rely on published tables, depending on precision, confidence levels, and variability. According to Israel (1992) table 1, a precision of 0.05, confidence level of 95% and a size of population greater than 100'000, the neccessary sample size is 400. If the precision is changed to 0.1, the sample size neccessary increases to 100 (Israel, 1992). The numbers found in the table reflects the number of obtained responses. The last approach is to use formulas to calculate the sample size. The formulas requires the standard deviation and how much variance to expect in the response (Smith, 2013)(Israel, 1992). Israel (1992) mentions that the table gives a useful guide for determening the sample size, and that formulas are used if the study has a different combination of precision and confidence. This study will use the table result since the combinations matches this study.

It's important to mention that the quality of the sample is as important as it's size. The more variable the sampled data is, the larger the sample size is required (Israel, 1992). It's also desirable to choose a random sample, which means that the observations are made idependently and random. The main purpose of using a random sample is to obtain correct information about the unknown population parameters (Walpole et al., 2012).

2 Statistics

2.1 Sample data from Survey

Independence of observations. This is mostly a study design issue and, as such, you will need to determine whether you believe it is possible that your observations are not independent based on your study design (e.g., group work/families/etc). A lack of independence of cases has been stated as the most serious assumption to fail. Often, there is little you can do that offers a good solution to this problem.

Designed the survey so that the observations should be random and independent
- Random order on the tests
- Random color on the layers
- Random which order the layers was drawn on the map
- Random which order the metadata was written in the table

2.2 Statistics theory

This section will give an introduction to the statistics used in this thesis. The thesis will examine the data with parametric methods but also with non-parametric methods if the assumption of a normally distributed samples fails. A nonparametric method is much more efficient than the parametric procedure when the set of data used in the test deviates significantly from the normal distribution (Walpole et al., 2012). There are also some disadvantages using nonparametric methods. The methods will be less efficient, and to acchieve the same power as the corresponding parametric method a larger sample size is required. If parametric and nonparametric tests are both valid on the same set of data, the parametric test should be used (Walpole et al., 2012).

2.2.1 Normal testing

The sampling distribution of a statistic depend on the distribution of the population, the size of the samples, and the method of choosing the samples (Walpole et al., 2012). Sampling distibution describes the variability of sample averages around the population mean μ. All parametric statistics assumes normally distributed, independent observations. Parametric tests are preffered in statistics because it got more statistical power than nonparametric tests (Frost, 2015). The power of a test is the probability of correctly rejecting a false null hypothesis, which in this case is the ability to detect if the sample comes from a non-normal distribution. To determine if a sample is normally distrubuted there exists both visual methods and normality tests to assess the samples normality. A visual inspection of the sample's distribution is usually unreliable and does not guarantee that the distribution is normal (Pearson et al., 2006). Presenting the data visually gives the reader an opportunity to judge the distribution themselves. In this thesis histograms are used to visualize the data for normality.

Normality tests compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation (Ghasemi and Zahediasl, 2012). There are multiple normality tests, and deciding which test to use is not easy. This study needs a test that doesn't require every value to be unique, a test that can handle ties (identical observations). The survey used to collect the samples in this study do not guarantee unique values.

The D'Agostino-Pearson omnibus test stand out as the best choice. This test first computes the skewness, see figure 2.1, and kurtois, see figure 2.2, to quantify how far from the normal distribution the sample is from the terms of assymetry and shape. Then it calculates how far each of these values differs from the value expected with a normal distribution (Pearson et al., 2006). It works well even if all values are not unique (Motulsky, 2013). The test also works well on both short- and long-tailed distributions (Yap and Sim, 2011).

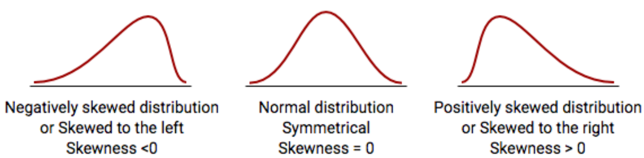

Figure 2.1: Skew (MedCalc Software bvba, 2017)

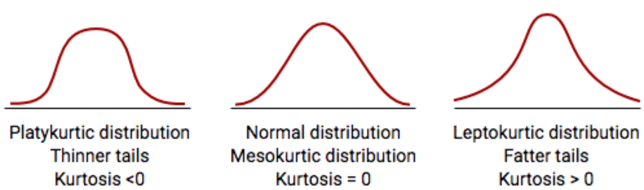

Figure 2.2: Kurtois (MedCalc Software bvba, 2017)

The D'Agostino-Pearson test uses the following hypothesis:

$$H_0\text{: The data follows the normal distribution}$$
$$H_A\text{: The data do not follow the normal distribution}$$

For small sample sizes, normality tests have little power to reject the null hypothesis, therefore small sample sizes most often pass normality tests. For large sample sizes, significant results would be derived even in the case of a small deviation from normality (Pearson et al., 2006). When the null hypothesis cannot be rejected, then there are two possible cases. First case is to accept the null hypothesis or the second case is that the sample size is not large enough to either accept or reject the null hypothesis (The Pennsylvania State University, 2017). An acceptance of the null hypothesis implies that the evidence was insufficient, the result does not necessary accept H_0, but fails to reject H_0 (Walpole et al., 2012).

2.2.2 Hypothesis testing

The null- and alternative hypothesis are statements regarding a difference or an effect that occur in the population of the study. The alternative hypothesis (Ha) usually represents the question to be answered or the theory to be tested, while the null hypothesis (H_0) nullifies or opposes Ha (Walpole et al., 2012). The sample collected in the study is used to test which statement is most likely (technically it's testing the evidence against the null hypothesis). When the hypothesis is identified, both null and alternative, the next step is to find evidence and develop a strategy for or against the null hypothesis (Lund Research Ltd, 2013a).

The first step, after identifying the hypothesis, is to determine the level of statistical significance, often expressed as the *p-value*. A statistical test will result in the probability (*the p-value*) of observing your sample results given that the null hypothesis is true. A significance level widely used in academic research is 0.05 or 0.01 (Walpole et al., 2012).

You should not report the result as "significant difference", but instead report it as "statistically significant difference". This is because your decision as to whether the result is significant or not should not be based solely on your statistical test. Therefore, to indicate to readers that this "significance" is a statistical one, include this is your sentence (Lund Research Ltd, 2013b).

2.2.2.1 Two sample t-test

When estimating the difference between two means a two-sample t-test is used (Walpole et al., 2012). A two sampled test assumes two independent, random samples from distributions with means [μ_1 , μ_2] and variances [σ_1^2, σ_2^2]. The hypothesis on two means can be written as:

$$H_0\text{: } \mu_1 \text{ - } \mu_2 = 0 \text{ or } \mu_1 = \mu_2$$
$$H_A\text{: } \mu_1 \text{ - } \mu_2 > 0 \text{ or } \mu_1 > \mu_1$$

Then the hypothesis refer to a one-tailed two sampled t-test. Before doing tests on the two means, the Levene's Test is used to test if the samples are from populations with equal variances. It tests the hypothesis:

H_0: Input samples are from populations with equal variances
H_A: Input samples are from populations that do not have equal variances

If we can assume equal variances in the two samples and the samples are normal distributed, a two-sampled t-test may be used.

Because the one-sided tests can be backed out from the two-sided tests. (With symmetric distributions one-sided p-value is just half of the two-sided pvalue). It goes on to say that scipy always gives the test statistic as signed. This means that given p and t values from a two-tailed test, you would reject the null hypothesis of a greater-than test when p/2 < alpha and t > 0, and of a less-than test when p/2 < alpha and t < 0.

Relevant hypothesis in this study that can be tested with a two-sampled t-test (if the conditions mentioned above are valid) is listed under.

Hypothesis - Two sample t-test

H_0**:** Mean task time between participants are equal
H_A**:** Experienced participants finish the tasks faster, use less time

H_0: Total number of correct elements between participants are equal
H_A: Experienced participants have a higher number of correct elements

H_0: There are no difference in total number of correct elements between the tasks
H_A: Participants have more correct elements on the one element task

H_0: There are no difference in mean time between the tasks
H_A: Participants finish the one element task faster

Before solving the hypothesis the conditions needs to be testet. More on this later.

2.2.2.2 Analysis-of-Variance

Analysis-of-Variance (*ANOVA*) is according to Walpole et al. (2012) a very common procedure used for testing population means. Where a two sample t-test are restricted to consider no more than two population parameters, *ANOVA* can test multiple population parameters. A part of the goal of *ANOVA* is to determine if the differences among the means of two or more samples are what we would expect due to random variation alone, or due to variation beyond merely random effects. *ANOVA* assumes normally distributed, independent, samples with equal variance. The equal variance assumtion will be tested with Levene's Test also mentioned in subsection 2.2.2.1.

One-way *ANOVA* tests the null hypothesis that two or more groups have the same population mean given that the mean is measured on the same factor or variable in all groups(Lund Research Ltd, 2013b). The hypothesis test can be written like this:

$$H_0\colon\ \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A\colon \text{ At least two of the means are different}$$

μ equals the group mean and k represents the number of groups. It is important to check that each group are normally distributed, not only the sample (Lund Research Ltd, 2013b). The weakness of one-way *ANOVA* is that it cannot tell which specific groups were significally different from each other if H_0 is rejected. To be able to determine which group a *post hoc test* is used.

In an one-way *ANOVA* test there should be one variable and minimum three independent groups, which is an relevant approach considering the data produced from this thesis survey. There are at least two variables in the survey data, task time and number of correctly chosen elements. The survey result can be divided into three groups, one element task, three elements task and six elements task. Each entry in the sample should only be assigned to one group. Relevant hypothesis from the study that can be used in an one-way *ANOVA* test is shown under.

Hypothesis - One-way ANOVA

H_0: Mean task time is not different between the three tasks
H_A: Mean task time is different between at least two of the tasks
Variable = time, group = tasks

H_0: Total number of correct elements between the three tasks are equal
H_A: Total number of correct elements between at least two of the tasks are not equal
Variable = Number of correct elements, group = tasks

The hypothesis written above will be testen in the section blabla.

2.2.2.3 Wilcoxon Rank-Sum test

The Wilcoxon Rank-Sum test is an appropriate alternative to the two-sample t-test (see subsection 2.2.2.1) when the normality assuptions do not hold, but the samples are still independend and have a continous distribution (Walpole et al., 2012). Since this method is nonparametric (or distribution-free) it do not require the assumption of normality.

The hypothesis for Wilcoxon Rank-Sum Test is:

$$H_0\colon\ \tilde{\mu}_1 = \tilde{\mu}_2$$
$$H_A\colon\ \tilde{\mu}_1 > \tilde{\mu}_2 \text{ or } \tilde{\mu}_1 < \tilde{\mu}_2 \text{ or } \tilde{\mu}_1 \neq \tilde{\mu}_2$$

The alternative hypothesis depends on what the test should determine. If the sample with mean $\tilde{\mu}_1$ is greater than, smaller than or unequal to the sample with mean $\tilde{\mu}_2$. First select a random sample from each population with means $\tilde{\mu}_1$ and $\tilde{\mu}_2$. If the sample sizes are different, let n_1 be the number of observations in the smallest sample and n_2 for the largest sample. Then $\tilde{\mu}_1$ will be the mean for the smallest sample. If there are ties (identical observations) in the sample a Mann-Whitey U test is preffered (The Scipy community, 2017).

2.2.2.4 Mann-Whitey U test

Skriv om nødvendig, om vi bare bruker tid er det liten sannsynlighet at det et identiske målinger..

2.2.2.5 Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric alternative to one-way *ANOVA* (see subsection 2.2.2.2) (Walpole et al., 2012). This test should be used if the assumption of normal distribution failed. As mentioned in this sections introduction, a nonparametric method does not assume normality. This test is an generalization of the rank-sum test when there are more than 2 samples.

Kruskal-Wallis is used to test equality of means in one-way *ANOVA*, so the hypothesis for the Kruskal-Wallis test is:

$$H_0\colon\ \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A\colon \text{ Minimum two of the } \mu_k\text{'s are different}$$

Here μ_k is the rank mean for the group k. As in Wilcoxon Rank-Sum test (subsection 2.2.2.3), the number of observations in the smallest sample is assigned to n_1, the second smallest to n_2 and the largest sample is assigned to n_k.

2.3 Survey results

2.3.1 Bar Charts

- All participants ordered by age
- All participants ordered by age, excluded by task 4
- All results in one task, ordered by age

Can use it to explain the data

Appendices

A | Tets

Fbox

Some text esfljsf lskj lksdjflsk slk

Some text kduhaszkdh aszkdjhs zkjdfh skdj skd

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

text

dwkjdkwjdh wkjdhw kjdh wkjhd qwkjhd kwd qw .

Bibliography

Affairs, A. S. f. P. (2013). System Usability Scale (SUS).

Ben, S. and Plaisant, C. (2009). *Designing the User Interface*. Pearson, fifth edition.

Brooke, J. (1996). *SUS-A quick and dirty usability scale. "Usability Evaluation In Industry"*. Taylor & Francis.

Frost, J. (2015). Choosing Between a Nonparametric Test and a Parametric Test. Date accessed: 2017-04-23 URL: http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test.

Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486–9.

ISO (1998). ISO 9241-11:1998(en), Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability.

Israel, G. D. (1992). Determining Sample Size 1.

Lund Research Ltd (2013a). Hypothesis Testing - Significance levels and rejecting or accepting the null hypothesis.

Lund Research Ltd (2013b). One-way ANOVA - Its preference to multiple t-tests and the assumptions needed to run this test | Laerd Statistics. Date Accessed: 2017-04-25 URL: https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-2.php.

MedCalc Software bvba (2017). Skewness and Kurtosis. Date accessed: 2017-04-23 URL: https://www.medcalc.org/manual/skewnesskurtosis.php.

Motulsky, H. (2013). Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, 3rd edition. In *Intuitive Biostatistics*, chapter 24.

Pearson, A., Sözcükler, A., düzeltmeli Kolmogorov-Smirnov, L., Pearson ve Jarqua-Bera testleri Derya ÖZTUNA Atilla Halil ELHAN Ersöz TÜCCAR, A., and Öztuna, D. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turk J Med Sci*, 36(3):171–176.

Schade, A. (2015). Pilot Testing: Getting It Right (Before) the First Time. Date accessed: 2017-04-19 URL: https://www.nngroup.com/articles/pilot-testing/.

Smith, S. (2013). Determining Sample Size: How to Ensure You Get the Correct Sample Size | Qualtrics. Date accessed: 2017-04-19 URL: https://www.qualtrics.com/blog/determining-sample-size/.

The Pennsylvania State University (2017). 7.5 - Power and Sample Size Determination for Testing a Population Mean | STAT 500.

The Scipy community (2017). scipy.stats.anderson — SciPy v0.19.0 Reference Guide. Date accessed: 2017-04-21 URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability & Statistics*. Pearson Education, ninth edition.

Yap, B. W. and Sim, C. H. (2011). Journal of Statistical Computation and Simulation Comparisons of various types of normality tests Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155.