ANOVA – The Effect of Outliers

By Markus Halldestam

Bachelor’s thesis
Department of Statistics
Uppsala University

Supervisor: Inger Persson

2016
Abstract
This bachelor’s thesis focuses on the effect of outliers on the one-way analysis of variance and examines whether the estimate in ANOVA is robust and whether the actual test itself is robust from influence of extreme outliers. The robustness of the estimates is examined using the breakdown point while the robustness of the test is examined by simulating the hypothesis test under some extreme situations. This study finds evidence that the estimates in ANOVA are sensitive to outliers, i.e. that the procedure is not robust. Samples with a larger portion of extreme outliers have a higher type-I error probability than the expected level.

Keywords: Analysis of variance, outlier, outlying observation, type-I error, robust
# Table of Contents

1. Introduction .................................................................................................................................................. 1

2. Methodology and Definition ......................................................................................................................... 2

   2.1. One-Way Analysis of Variance ............................................................................................................... 2

   2.1.1. Assumptions and Sample Size Criteria ............................................................................................. 4

   2.1.3. Estimates and Their Distributions ..................................................................................................... 5

   2.2. Robust Parameter Estimate and Test ..................................................................................................... 6

   2.2.1. Breakdown Point ................................................................................................................................. 6

   2.2.2. Type-I and Type-II Error Probabilities of the Test ............................................................................ 7

   2.3. Outlying Observations ........................................................................................................................... 8

3. Simulation ....................................................................................................................................................... 10

   3.1. Setup ......................................................................................................................................................... 10

   3.2. Simulated situations ................................................................................................................................. 11

   3.2.1. No Outliers .......................................................................................................................................... 11

   3.2.2. One Substantial Positive Outlier ........................................................................................................ 11

   3.2.3. One Positive and One Negative Outlier in Different Groups ............................................................... 11

   3.2.4. Two Positive and Two Negative Outliers in Different Groups ............................................................ 11

   3.2.5. Three Positive and Three Negative Outliers in Different Groups ....................................................... 11

4. Results ............................................................................................................................................................. 12

   4.1. Small Sample Size .................................................................................................................................. 12

   4.2. Medium Sample Size ............................................................................................................................... 13

   4.3. Large Sample Size ................................................................................................................................ 14

5. Conclusion ..................................................................................................................................................... 15

References ......................................................................................................................................................... 16

Appendix ............................................................................................................................................................ 18
1. Introduction

Scientific and work related data are generally collected for the purpose of interpretation. Each observation or individual that the data consists of carries a fraction of the information of the population. If a method or a test is properly used for the data, information is obtained suitable for analysis. The information, and the analysis of this information, has a significant role for science, society and commerce. Even though the information extracted from collected data is vital, the analyst’s application of a method or test has a nearly as important part in the procedure as the data itself. However, not only don’t all analyses include proper method or test suitable for the purpose of the study, most tests are restricted by assumptions and sample size criteria which limits the analyst. These criteria and assumptions are more often than not overlooked or simply ignored, which obviously may cause great misinterpretations.

One commonly applied test for inference is the One-way ANOVA (Analysis of variance). The one-way ANOVA is used for inferences of between group means, where the groups are defined by one categorical variable. ANOVA brings many advantages as it is easy to apply, perform and interpret, compared to other similar methods and tests (Cobb, 1984). Because of its many advantages, the ANOVA is a vital method in explanatory and confirmatory data analysis (Gelman, 2005), not least in the field of social science (Iversen & Norpoth, 1987). After decades of use, the method of ANOVA stands as one of the basis for statistical journals (Giloni, Seshadri, & S., 2005). Although its advantages, many of the available guides of how to proceed with an ANOVA claims that the test is sensitive to outliers. However few explain to which extent and which effect outliers actually have on the test. As the test is so commonly used it is therefore of importance to examine the results of outliers, and other types of noise, while proceeding with an ANOVA.

The most reasonable concern is the effect on the parameter estimate of the groups, on which the inferences is based. Would outlying observations have an effect on the mean estimate, the result could be that the test would wrongly reject the null hypothesis that all group means are equal while in fact they are equal, or vice versa. Prior research has found that violations of the independent assumptions influence the type-I and type-II error probability (Scariano & Davenport, 1987). Analysis of means has very poor resistance to outliers (Hoaglin, Mosteller, & Tukey, 1983). There is therefore reason to examine, not only, the robustness of the parameter estimate in an ANOVA, but also, how a potential deviation, caused by (an) outlying observation(s), may affect the outcome of the actual inference.

In this paper an examination of whether the parameter estimate used in the one-way analysis of variance could be considered robust or not is performed. The intention of this study also includes what effect a potential non-robust estimator would have for the outcome and reliability of the inference of the one-way ANOVA. The following two research questions summarize what this paper intends to solve:

- Are the parameter estimates of one-way ANOVA robust against outliers?
- Will outliers limit the hypothesis test in ANOVA?
2.0. Methodology and Definition

2.1.1. One-Way Analysis of Variance

One-way ANOVA with three groups\(^1\), or more, has the fundamental purpose to test whether the means of the groups differ from each other. To initially understand this technique, one must understand what types of categorical variables that can diverge an underlying population into three or more groups. Typical categorical variables for divergence in medicine are: satisfaction with care, degree of a symptom, amount of felt pain (Jekel, 2007), in social science; political affiliation, income-classes or religious affiliation, etc. The variable should be able to categorically classify an observation/individual into \(k\) groups\(^2\). (Iversen & Norpoth, 1987)

A simple way of examining if there is any difference between the groups is to compute the mean of each separate group and compare the groups’ mean. However, by simply comparing the between group means one only examines the sample of the underlying population and no generalization can be made for the potential difference of the three groups in the actual population. A potential difference between the groups in the sample can be a coincidence made by random variation from one sample. For one to be able to make any statement of the population from the sample, one must test whether the three group means differs significantly. This is where the one-way ANOVA comes into the picture.

The one-way ANOVA with \(k\) groups contains one dependent variable that is numerical and one explanatory variable that is categorical, with \(k\) categories. The explanatory variable is the one that defines the groups, and for \(k\) groups the explanatory variable can only have \(k\) values or categories. The dependent variable is the variable of attention for the groups’ means.

In a population there is, in many cases, variation of some sort of degree. The total sum of squares, SST, could be seen as a measure of the total variation of the population. SST is calculated using each individual observation in the sample and the estimated mean of the entire population, called the grand mean\(^3\), denoted \(\bar{y}_{GM}\) (Hocking, 2013). Total sum of square is formulated\(^4\) as follows:

\[
SST = \sum_{i=1}^{n} (y_i - \bar{y}_{GM})^2
\]

\(^1\) If the ANOVA only consists of two groups the F=t\(^2\), where t is the Student’s ‘t-test.
\(^2\) Where \(k\) represents the number of different categories, three or more.
\(^3\) The grand mean is the same as; the mean of all observations or the mean of all groups’ means, the mean of means.
\(^4\) \(y_i\) is the value of the observation and where \(i\) is observation 1 to \(n\) (number of total observations), and \(\bar{y}_{GM}\) is the grand mean.
The total variation of the population could be seen as a produce out of two types of variations, (i) the variation caused by group differences, the between group variation, and (ii) the variation caused by the variance within each group, the within group variation. One of the more fundamental intents with the ANOVA is to identify and measure the degree of these produces that causes the total variation. The between group variation sum of square, SSB, is the sum of each group’s mean in relationship with the grand mean, squared. (Hocking, 2013) (Krishnaiah, 1980). Formulated as:

\[ SSB = \sum_{j=1}^{m} (\bar{y}_j - \bar{y}_{GM})^2 \]

The second part of the total variation, the within group variation, is similarly calculated using sum of squares, SSW. The sum of squares of the within group variation is calculated using each individual value in relationship with its group’s mean (Krishnaiah, 1980). SSW is formulated as:

\[ SSW = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \]

As the total variation of the population is a produce of the within and between group variation the following relationship can be stated (Hocking, 2013):

\[ SST = SSB + SSW \]

This relationship also includes the degree of freedom for each statistics, summarized in table 1.

<table>
<thead>
<tr>
<th></th>
<th>SSB</th>
<th>SSW</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df</td>
<td>(m - 1)</td>
<td>(m(n - 1))</td>
<td>(m \times n - 1)</td>
</tr>
</tbody>
</table>

As analysts often are more interested in generalizing conclusions rather than making up statements of the sample, the one-way ANOVA allows the analyst to examine a potential difference between the groups. This is possible while testing the null hypothesis, that the means in all groups are equal, against the alternative that at least one mean differs from the others.

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_m \]

\[ H_a: \text{at least one group mean diverges from the other group means} \]

---

5 \( \bar{y}_j \) is the mean of group \( j \), where \( j \) is group 1 to \( m \) (number of total groups), and \( \bar{y}_{GM} \) is the grand mean.

6 \( y_{ij} \) is the value of observation \( i \) in group \( j \). \( \bar{y}_j \) is the mean of group \( j \). \( i \) ranges from 1 to \( n_j \) (number of total observations in group \( j \)). \( j \) ranges from 1 to \( m \) (number of total groups).

7 \( m \) represents the number of groups while \( n \) represents the numbers of members in each group.
The null hypothesis is tested, while assuming the null hypothesis, using the F-statistic. The test statistic is defined as the between group variability divided with within-group variability. What the test actually does is formulating a ratio of the two variations of the population.

\[
F = \frac{\text{Between} - \text{group variability}}{\text{Within} - \text{group variability}}
\]

Consider the relationship in the F-statistic. A higher value of the between group variability, compared to the within group variability, could be interpreted as a larger portion of the variation in the underlying population being caused by differences between groups, and thereby a larger F-value. If the circumstances were the opposite, and a smaller F-value, a larger part of the portion of the population’s variation is caused by variation within the groups. The variability between and within the groups are usually represented by the mean sum of squares, MSB and MSW, respectively. The mean sum of squares between and within the groups are simply SSB and SSW divided with their respective degrees of freedom (Krishnaiah, 1980). Thus, the test statistic for the ANOVA may be reformulated as:

\[
F = \frac{\text{MSB}}{\text{MSW}} = \frac{\frac{SSB}{m-1}}{\frac{SSW}{m(n-1)}}
\]

Assuming the null hypothesis, the F-statistic allows the analyst to test this hypothesis. A critical F-value is calculated using the conditions under the null hypothesis and a suitable significance level, chosen prior the test. The critical value defines the threshold for the test, and the null hypothesis is to be rejected if the calculated F-statistic, on the basis of the information in the sample, extends the critical value.

2.1.2. Assumptions and Sample Size Criteria

Analysis of variance has four crucial assumptions that must hold. The inference of the test becomes futile if one assumption is violated. For the ANOVA to be valid the data must consist of an (i) independent random sample. The sample must contain observations independently chosen from each other. Sampling one observation must not influence the sampling of another observation. (ii) The numerical variable must be normally distributed in each group. (iii) The residuals must be normally distributed. (iv) All groups must not suffer from heterogeneity. The variance must be equal in all groups included in the large sample, aka homogeneity. (Scariano & Davenport, 1987) (Hair, 2006)

There are three general minimum sample size recommendations for each group when applying the one-way ANOVA. (i) Each group size must extend the number of dependent variables. Hence, in a one-way ANOVA with e.g. three groups, only one dependent variable exists. (ii) The number of members in each group must be larger than, or equal to, 20. (iii) Each group, within the sample, shall have approximately an equal amount of observations. (Hair, 2006)
2.1.3. Estimates and Their Distributions

The one-way ANOVA mainly uses two estimators for the test. (i) The sample mean, which is used to estimate the mean of the underlying population. The sample mean is used both for the entire sample to estimate the grand mean, and to estimate each group mean. The estimate for the grand mean and the group means are similar except with regards to the observations included. The grand mean estimate takes all observations within the sample into account while the group mean only accounts for the observations in its particular sample/group. The sample mean is formulated as:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The estimator for the grand and the group mean follows the normal distribution:

$$\hat{\mu} = \bar{y} \sim N(\mu, \frac{\sigma^2}{n})$$

The second estimator, which is most associated with ANOVA, is (ii) the variance of the sample mean. The variance of the sample mean is an estimator for the variance of the underlying population. It is also used to estimate the variation of each group. The variance of the sample mean is formulated as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The variance of the sample mean follows a \(\chi^2\)-distribution with \((n - 1)\) degrees of freedom:

$$\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n - 1)}{n - 1}$$

(Hocking, 2013)
2.2.1. Robust Parameter Estimate and Test

Many different papers across the years have different definitions of what a robust estimator is, and how to measure the robustness of it (Stewart, 1999). Robust may be defined as something that is sturdy and string in its form, or as Huber (1981) put it: “robustness signifies insensitivity to small deviations from the assumptions”, where the assumptions are of the underlying distribution of the population. Hence, many assumptions are at most approximations of the reality which potentially, with the present of noise and/or extreme observations, would lead to gross errors (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). As many assumptions are approximations the behavior of an estimate is not always as expected. The assumed underlying distribution may in fact be the distribution closest to reality but the sample, upon which the estimate is calculated, can consist of noise or extreme observations. Such influences from the sample may cause deviations from the parameter’s true value. If the estimator instead was robust the influence of extreme observations would be limited (Hoaglin, Mosteller, & Tukey, 1983), and the effect of the disturbance in the sample would be minor, if any at all. Robustness of a statistic test is the product of its used parameter estimates. Krishnaiah (1980) defined a test that is robust as: “Specifically, a test is called robust when its significance level (Type-I error probability) and power (one minus Type-II error probability) are insensitive to departures from the assumptions on which it is derived”. What this definition actual means is that a statistic test is not robust if any deviation from the test’s assumptions affects its significance level and power. What the definition does not include is what impact deviations from sample size criteria would have on the test’s significance level and power.

2.2.2. Breakdown Point

One commonly used measurement for the robustness of the estimator is the breakdown point (Rousseeuw & Leroy, 1987) (Davies & Gather, 2007). The breakdown point is the minimum fraction of outlying observations that may cause an estimator to differ randomly far from the true value of the parameter (Hoaglin, Mosteller, & Tukey, 1983). This basically means that the breakdown point is a measure of to which limit an estimator can handle substantially large observations (Huber, 1981). The breakdown point of an estimate is expressed in two parts: worst-case breakdown point and well-placed breakdown point. The worst-case breakdown point quantifies the resistance of deviated observations in a data set. The worst-case breakdown point is the fraction of \( k/n \), where \( k \) is the number of observations in a sample that can be replaced by an arbitrary value while still leaving the estimate “unharmed”, and \( n \) represents the sample size. The well-placed breakdown point quantifies the fraction of possible “bad” observations in a sample without seriously changing the estimate (Hoaglin, Mosteller, & Tukey, 1983). One could see the worst-case breakdown point as a measure of how many “bad” observations that are needed to force the estimate to become randomly off, while the well-placed breakdown point is the measure of how many “bad” observations that may be in a sample without the estimate taking an arbitrary value.

One commonly applied estimate is the sample mean. Say that a population consists of independent random variables \( \{X_1, ..., X_n\} \) and the realization\( \{x_1, ..., x_n\} \). What happens if we hold \( \{x_1, ..., x_{n-1}\} \) fixed while \( x_n \) goes towards infinity?

\[
\bar{x} = \frac{x_1 +, ..., +x_n}{n}
\]

It is obvious that as \( x_n \) goes toward infinity, given that \( \{x_1, ..., x_{n-1}\} \) are held fixed, the sample mean will go towards infinity. The impact of observation \( x_n \) drastically changes the value of the sample mean. Each observation influences the sample mean by the value \( 1/n \), and because of that the sample mean does not have any protection from (a) deviating observation(s). Only the presence of one single arbitrarily large observation may cause the sample mean to become arbitrarily large. While letting \( n \to \infty \) the worst-case breakdown point becomes more tangible, and the breakdown point is approximately zero. For the sample mean, this is also the case for the well-placed breakdown point as, under the best of situations, the sample mean cannot even handle one observation going towards infinity. The worst-
case breakdown point, and the well-placed breakdown point, of the sample mean are zero and the estimator is not robust against outliers (Hoaglin, Mosteller, & Tukey, 1983).

The median is considered to be more tolerant to deviations than the mean. Almost half of the observations in a sample may be “bad”, without damaging the estimate. We consider a sample size of \( n \) observations and we want to find the breakdown point of the median. As the sample median is the value of the observation that separates the sample’s higher half from the lower half; \( n = 2k + 1 \) for odd \( n \) and \( n = 2k + 2 \) for even \( n \). Thus the worst-case breakdown point for an odd sample size is \( k/(2k + 1) \), while \( k/(2k + 2) \) for an even sample size. This expression may be simplified using the oddness function:

\[
d(n) = \begin{cases} 
0 & \text{for even } n \\
1 & \text{for odd } n
\end{cases}
\]

And the worst-case breakdown point of the sample median may be formulated as:

\[
\frac{\frac{1}{2}n + \frac{1}{2}d(n) - 1}{n} = \frac{1}{2} - \frac{2 - d(n)}{2n}
\]

The worst-case breakdown point for the sample median is asymptotically \( 1/2 \), which is more tangible if we let \( n \to \infty \). An estimate can have a worst-case breakdown point from 0 to 0.5, where a high breakdown point is equivalent with robustness. The limit can however not extend 0.5 as from that point it is impossible to differentiate between the underlying distribution and the distribution that disturbs the estimator (Hoaglin, Mosteller, & Tukey, 1983).

The well-placed breakdown point tells how many of the observations in the sample may be “bad” without damaging the estimate. A large portion of the sample may be substantially large or small while still not influencing the sample median. The sample median has a well-placed breakdown point of \( 2k/n \), or:

\[
1 - \frac{2 - d(n)}{n}
\]

The asymptotic well-placed breakdown point is just shy of 1, while letting \( n \to \infty \). This doesn’t mean that in the sample all observations may be extreme observations. It means that all other observations than the middle observation may be outliers.

The sample median is, based on the worst-case- and well-placed breakdown point, a robust estimator. The sample mean, as earlier illustrated, is however not a robust estimator on the basis of the breakdown point. On the basis of the lack of robustness of the sample mean, the estimate used in the one-way ANOVA is sensitive to outlying observations. This also effects the second estimate of ANOVA, the variance of the sample mean. Since the sample mean is not robust, an estimator that uses the sample mean to estimate a second parameter cannot be robust. As the variance used in ANOVA is the variance of the sample mean, it cannot be robust. One sizeable outlier may cause the sample mean to deviate from the true mean, and causes a substantial variation. Thus, the variance of the sample mean must not be considered robust.

2.2.3. Type-I and Type-II Error Probabilities of the Test
As argued in section 2.2.2. the estimators, sample mean and variance, used in one-way ANOVA are not robust. But perhaps even more important than clarifying that the estimators used in ANOVA are robust or not, is to measure the effect of outlying observations on the test result. As mentioned in the introduction, it is reasonable to believe that outliers may cause the test to falsely reject the null hypothesis, if there actually are no differences between the groups, and vice versa. For the one-way
ANOVA to be considered robust, it must be insensitive towards the influence of outliers. The type-I and –II error probabilities of the test must not change due to the effect of substantial outliers.

To be able to see the effect of outliers it is most convenient to observe a population with three underlying groups with no differences in means, which is something that can be done by a simulation study, constructing the groups equally distributed \(N(0,1)\) without any outliers or noise in one overall sample while also constructing a second overall sample, with the same distribution as in the “perfect” overall sample, but with outliers. By constructing two overall samples, one with outliers and one without, it allows to simulate the testing of the null hypothesis. The simulation will examine how often the null hypothesis is wrongfully rejected. If the overall sample with outliers has a substantial different ratio of type-I error, it would be clear that outliers have an impact on the test result and the robustness of ANOVA would be questioned.

One could consider applying this technique on an actual population. However the likelihood of finding a population with such a perfect underlying distribution while just few observations are extreme is very low. Actual data would include noise which would include more factors than simply substantial outliers. Even though inferences are more often than not applied on real data, and testing ANOVA on the basis of data from the reality would contribute to a higher understanding in the application of ANOVA, a simulation allows for replications far easier. A simulation study also makes it possible to handle a large quantity of trials, which is superior to the case of handling actual data.

Even though this study does not include actual data it is important to have the fundamental idea that one-way ANOVA is commonly applied and that this study may contribute to a better understanding to potential flaws. The most commonly chosen significance level for the test, as for most tests in inference, is five percent. A significance level of five percent would result in only five percent of tests of the null hypothesis will be wrong under the null hypothesis (Sheskin, 2000). In the case of the simulation, this will be expected for the uncontaminated overall sample. How the actual type-I error probability will act in the contaminated overall sample is more unsure, and is to be investigated.

2.3. Outlying Observations

For someone who has dealt with inference, outlying observations would probably be familiar. Outlying observations, or simply “outliers”, are observations/individuals that deviate markedly from other members of the sample in which it occurs (Grubbs, 1969). Although, outlying observations are highly documented in scientific journals, there is no general threshold regarded if an observation should be considered an outlying observation. There are however a number of methods available to detect outliers (Rousseeuw & Leroy, 1987) (Hodges & Austin, 2004) (Adikaram, Hussein, Effenberger, & Becker, 2014).

In a theoretical sense an outlying observation could be seen as one observation which is going towards infinity, while all uncorrupted observations remain fixed. Given a sample \(\{x_1, \ldots, x_n\}\), where \(\{x_1, \ldots, x_{n-1}\}\) are fixed and act as the assumed distribution, \(x_n\) is an outlier with an infinitely large value. Even though this definition may be suitable from a theoretical aspect it may be an inept definition in a practical sense. In empirical inference it is more suitable for a more strict definition on the basis of when an extreme value may be problematic for the significance level and the power of the test or method. This is simply due to the fact that infinite values are seldom encountered while handling untransformed data.

One very common and applied method of detecting and defining the threshold of an outlying observation is by the use of boxplots\(^8\). A boxplot is mainly used to graphically display the data. The boxplot uses the median and the lower and upper quartiles to illustrate the behavior of the middle section of the data as well as the ends of the distribution. The lower quartile (or the 25\(^{th}\) percentile) is denoted \(Q_1\) while the

---

\(^8\) Boxplots and its components were first introduced and defined in 1969 by Tukey, see (McGill, Tukey, & Larsen, 1978).
upper quartile (or the 75th percentile) is denoted $Q_3$. Between the lower and the upper quartiles is the inner-quartile range, denoted $IQ$. To identify extreme observations, the boxplot creates two types of “fences”, the inner and the outer fences. Both types of fences have an upper and a lower bound relative to the first and third quartile and the inner-quartile range (Härdle, 2003):

1. Lower inner fence: $Q_1 - 1.5 \times IQ$
2. Upper inner fence: $Q_3 + 1.5 \times IQ$
3. Lower outer fence: $Q_1 - 3 \times IQ$
4. Upper outer fence: $Q_3 + 3 \times IQ$

An observation positioned beyond the inner fence, but within the outer fence, is considered a mild outlier. An observation beyond the outer fence is considered to be an extreme outlier. This definition is the most suitable for this study as it separates the centralized mass of a set of data from an outlier and has a clear general threshold for when an observation is considered an outlier. Another advantage with this definition is that it separates mild from extreme observations. Even though some observations may be relatively large or small compared to the majority of the sample, the observations may actually behave according to the assumed distribution. If a sample is very large it is likely to find an observation relatively far from the median, even in a simulation. This isn’t a problem when using this definition and its separation of mild and extreme observations as those “normal” outliers would be classified as mild observations. Another advantage of using boxplots to define an outlier includes the fact that the definition defines an outlier relative to the median, not the mean. As argued in section 2.2.2. the median is robust while the sample mean is not. A definition on the basis of the mean would most likely be influenced by the outlier itself, which could stain this study. Therefore a definition on the basis of e.g. the three sigma rule, see (Pukelsheim, 1994), would be insufficient.

In this study a worst case scenario is simulated, i.e. examining the effect on the type-I error probability caused by an extreme outlier. By taking this approach it is suitable to have the standpoint of treating mild outliers as something normal, while studying the effects of extreme (substantial) outliers. In that sense the standpoint of this study is that mild outliers fall under the category of statistical noise. Statistical noise, or simply noise, is in statistics often a term for unexplained variation in a sample. Unexplained variation is considered to be normal in a sample, and while this study treats mild outliers as noise, a mild outlier is therefore, in this study, considered a normal phenomenon. This study will on the basis of the definition of an extreme outlier used for boxplots examine the effects of substantial outliers.

Outliers are caused by two reasons. The deviation of one/or a few, observation(s) may be the result of random variability inherent in the data (Grubbs, 1969). For example: if one were to investigate the living area per individual in a neighborhood consisting mainly by tenements, and one villa, there is reason to be vigilant as that one household could have a much large value relative the others. In this case the outlier is merely an extreme value relative to the others in the sample, but it is a realistic and an actual value and should therefore be processed in the same manner as the others. Removing “true” outliers is not recommend as it yields improper estimates of the standard errors (Wilcox, 1997). However, outliers may also be the result of mishandling the data. These deviations could be caused by errors in calculations or recordings of values. In contrast to the other type of outlier, this is due to human error (Grubbs, 1969). As outliers caused by human error are not actual values of observations they should not be included in the inference. In the best of worlds they should not even been made in the first place. The problem with outliers, initially, is to identify what caused them and from that information decide if those observations represent the underlying population or not. In this study no separation of different causes for outliers is made. It would be irrelevant for the purpose of this study to try to define outliers on the basis of what caused them. This is due to the fact that ANOVA works exactly the same regardless of the cause of an extreme outlier.
3.0. Simulation
The main objective of this study is to test the one-way ANOVA’s sensitivity to outliers. Specifically, the sensitivity of the ANOVA’s type-I error probability when outliers are present. The type-I error probability shall not be effected by extreme outliers if the ANOVA is robust.

The base material for this study is a generated population with two variables. The software for generating, processing and simulating the data is R Studio. One of the two variables is a numerical variable, while the second is a categorical variable. The numerical variable follows the standardized normal distribution\(^9\), \(N(0, 1)\), and is denoted “the x-variable”. The generated population consists, in total, of one million observations. The reason why the population consists of one million observations is to simulate a large underlying population. Obviously, the numbers of observations could be larger or smaller but the one million mark gives the advantage of having an underlying population that is large as well easy to cope with. The observations, in the population, are distributed between three groups that are defined by the categorical variable, denoted “Group”. The categorical variable has three levels named, by convenience, “A”, “B” and “C”. The x-variable is equally distributed for each group, \(N(0, 1)\). The number of observation in each group is also the same in all three groups, i.e. 1/3 of the total number of observations.

3.1. Setup
To emulate the fact that one-way ANOVA is applied on a large variation of sample sizes, this study will showcase three levels of sample sizes: (i) a small sample size, (ii) a medium sized sample size and (iii) a large sample size. These levels of sample sizes consist of: (i) 100, (ii) 300 and (iii) 3000 observations in total. Each group has approximately one third of the total number of the sample size.

In excess of the three levels of sample sizes, this study also tests a number of different situations. These are: (i) no outliers, (ii) one substantial positive outlier, (iii) one substantial positive and one substantial negative outlier in different groups, (iv) two substantial positive outliers in group C and two substantial negative outliers in group B and (v) three substantial positive outliers in group C and three substantial negative outliers in group B.

For each sample size level and situation the null hypothesis, that all means of the x-variable are the same in all groups, will be tested. This process is repeated 10 000 times while counting the number of rejections and acceptations of the null hypothesis. Each test consists of a sample where each observation is randomly selected\(^10\), and where an observation cannot be selected twice in the same sample. This results in each test of the null hypothesis being based on a unique composition of observations. The reason why the test is repeated 10 000 times is because this will allow the computing of a type-I error ratio for each sample size and situation. While repeating the number of test more than 10 000 times would allow for a more general perception of the true type-I error probability, the time at hand limits the number of times a test may be repeated. Due to the time limit, the 10 000 mark is sufficient for an idea of what the actual type-I error probability would be.

The significance level for the test is five\(^11\) percent. The reason why the five percent level is the most suitable is mostly due to the fact that it is the most applied level of significance for the ANOVA, or any statistical test. As the fundamental purpose of this study is to question the suitability of the one-way ANOVA for inference, this significance level reflects the practical application best. The significance level of five percent shall theoretically result in a type-I error ratio of five percent (Sheskin, 2000). This is the expected ratio for at least the situation when no outliers are present. As stated before, if ANOVA

---

\(^9\) Standardized normal distribution with zero mean and one variance.

\(^10\) Note: except the outlier or outliers

\(^11\) Note: in this study the five percent mark is a cold hard threshold, which means that any p-value that extends the five percent mark is considered insignificant and the null is accepted. This means that if a p-value just extends the threshold it is statistical insignificant, i.e. 0.050000001 is insignificant while 0.049999999 is significant.
is robust the type-I error probability must not be affected by the presence of outliers. Thus the type-I error ratio, of the situation when one or more outliers are included, must be at least relatively close to when no outliers are in the sample.

The outliers in the simulation will have the x-value of ten or minus ten. A positive outlier is referred to the outlier with the value ten and the negative outlier refers to the outlier with the value minus ten. These outliers are extreme outliers and deviates therefore extremely far from the mean or median of each sample. The outliers are constructed to be drastically extreme to make it easier to examine the impact of an outlier that is very far from the central mass of a sample. Why the value is either ten or minus ten, is due to the limit of the outer fence. In the standard normal distribution with mean zero and standard deviation one, with one million observations, the outer fence begins around an x-value of 4.725\textsuperscript{12} units from the mean. By definition, all observations beyond that point are extreme outliers. To simulate an extreme case, the outliers in this study will have a value that is about twice the outer fence, which roughly is ten or minus ten.

3.2. Simulated situations

3.2.1. No Outliers
The one situation that will work as the reference group of this study is the simulation of when there are no outlying observations present at all. Here are all observations simply randomly selected from the standardized normally distributed population.

3.2.2. One Substantial Positive Outlier
In the situation with one substantial positive outlier all observations in the sample except one are randomly selected from the underlying population. The last observation is however a substantially large outlier with the x-variable value ten, i.e. a value ten times the standard deviation of the underlying population. This observation will always be present in group C, due to convenience.

3.2.3. One Positive and One Negative Outlier in Different Groups
This situation is an extension from the previous situation by one substantial positive outlier. While still including the positive outlier in group C, this situation also includes a substantially small observation. This newly added observation replaces one observation from the normally distributed population and has an x-value of minus ten. The extremely small observation is always positioned in group B.

3.2.4 Two Positive and Two Negative Outliers in Different Groups
As in all other situations, group A will not be infested by any outliers. However group B will have two substantial negative outliers. They have the x-value of minus ten. Group C also has two extreme outliers, both with the x-value of ten.

3.2.5 Three Positive and Three Negative Outliers in Different Groups
This situation is similar to the previous situation where group A is free from outliers. The minus ten outliers in group B are now three and the positive ten outliers in group C are also three.

\textsuperscript{12} In the generated population the IQ is 1.35 while $Q_1 \sim -0.675$ and $Q_3 \sim 0.675$. This results in the lower outer fence at $\sim -4.725$ and a upper outer fence at $\sim 4.725$. 
4.0. Results

4.1. Small Sample Size

In the small sample size simulation each sample contains a total observation count of 100. Each group is of about the same size, ~33 observations. Each situation is simulated 10 000 times and the null hypothesis is tested each time. The results are described in Table 2 below.

Table 2 – Ratio of rejected and accepted null hypothesis – Small sample size

<table>
<thead>
<tr>
<th></th>
<th>No outliers</th>
<th>With one outlier</th>
<th>With two outliers</th>
<th>With four Outliers</th>
<th>With six outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>null rejected</strong></td>
<td>5.19</td>
<td>3.45</td>
<td>5.38</td>
<td>31.98</td>
<td>83.36</td>
</tr>
<tr>
<td><strong>null accepted</strong></td>
<td>94.81</td>
<td>96.55</td>
<td>94.62</td>
<td>68.02</td>
<td>16.64</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*The results are presented in percent of the 10 000 trials for each situation.

The small sample performs according to the expectations when no outliers are included in the sample, the ratio of type-I error is approximately five percent for this situation (table 2). When the sample includes an extreme positive outlier in group C, the results differ expressively from the expected five percent. The failure rate of the hypothesis testing is just short of 3.45 percent. This is an unpredicted and odd improvement from the sample without any outliers, and what caused this could be discussed. It is possible that this is a special case, where the ANOVA performs better than the significance level, if the sample is small and there is one extreme outlier. From that point of view, it is reasonable to argue that the presence of an outlier effects the estimates of ANOVA and thereby effects the type-I error probability, in this special case affecting it by decreasing the error ratio. However this claim needs further evidence to back it up properly.

When the sample not only consists of one extremely large outlier in group C, but also an extreme negative outlier in group B, the type-I error ratio is similar to the no outlier situation. The error ratio cannot be argued to be different from the significance level as the ratio is 5.38 percent. Therefore the outliers cannot be proven to have a large effect on the type-I error probability in the small sample. However, during the two last situations, the ones with four and six outliers, the type-I error ratio rapidly increases. In the situation with two large outliers in group C and two negative outliers in group B, the failure rate is about 32 percent. This indicates that the outliers have an effect on the type-I error probability by largely increasing the rate. In the most extreme case, where three large outliers and three negative outliers are included in the sample, the type-I error ratio is nearly 83 percent. In this extreme case the ratio of type-I error is larger than the correct decisions. The six outliers have an extremely large effect on the type-I error probability for the ANOVA.

In the case of the small sample size, it is obvious that an increase of outliers in a small sample rises the type-I error ratio. Only considering the small sample size, it is fair to conclude that the one-way ANOVA is sensitive to substantial outliers in that sense that outliers increase the risk of committing a type-I error. From the perception of what ratio of error that is acceptable, from the significance level of five percent, the sample with four and six outliers fail to not extend that threshold. In the situation when only one large outlier is included in the sample, the type-I error ratio is smaller than the acceptable threshold and could be a special case. However it is clear that outliers do affect the type-I error probability of ANOVA.
4.2. Medium Sample Size

The medium sample size includes 300 observations in each sample. Group sizes are about 100 in each. The five different situations are simulated 10 000 times each to test the null hypothesis. Results are presented in Table 3 below.

Table 3 – Ratio of rejected and accepted null hypothesis – Medium sample size*

<table>
<thead>
<tr>
<th></th>
<th>No outliers</th>
<th>With one outlier</th>
<th>With two outliers</th>
<th>With four Outliers</th>
<th>With six outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>null reject</strong></td>
<td>5.34</td>
<td>4.68</td>
<td>6.61</td>
<td>23.57</td>
<td>56.34</td>
</tr>
<tr>
<td><strong>null accept</strong></td>
<td>94.66</td>
<td>95.32</td>
<td>93.39</td>
<td>76.43</td>
<td>43.66</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*The results are presented in percent of the 10 000 trials for each situation.

Just as when examining the small sample size, the percent of rejected null hypotheses in the sample without any outliers is roughly five percent (Table 3). The sample with one outlier in group C, has a ratio of type-I error lower than five percent indicating either no effect, or a smaller improvement in terms of failure. As the difference from the five percent mark is so little, it is a very weak statement to claim that the medium sample is effected by one substantial outlier. The sample with two outliers is, however, different from the five percent threshold. With the 6.61 percent mark it differs by more than 1.6 percentages from the significance level and indicates a small effect on the type-I error probability. The small sample size could not prove an effect of two outliers, which the medium sized sample size does. If four outliers are included in the sample, the ratio of type-I error is about 23.6 percent, which is a strong evidence that the outliers in the sample reduce the probability of a correct conclusion. The sample with six outliers is an even stronger proof of the outliers’ effect on increasing the type-I error ratio, with its 56.34 percent of type-I errors. However, worth noting is the fact that even though the ratio of errors has increased by including the numbers of outliers, the medium sample performs better than the small sample when the numbers of outliers are high. The 23.6 and 56.3 percent type-I errors can be seen in contrast to the small sample’s 32 and 83.4 percent for corresponding situations (Table 2).

The medium sized sample performs better than the small sample in the sense that it handles the situation with a higher number of outliers better. In the situation when only one outlier is included in the sample no effect can be proven, however when the sample has two outliers, there is an increase of type-I errors which indicates that the medium sample is sensitive to outliers even when only two outliers are included. The medium sample with four or six outlier has a high ratio of failures which indicates a sensitivity to a few extreme outliers.
4.3. Large Sample Size

The large sample includes 3000 observations in each sample. Each group consists of about 1000 observations each and the different situations are simulated 10,000 times. The results are presented in table 4 below.

*The results are presented in percent of the 10,000 trials for each situation.

Table 4 – Ratio of rejected and accepted null hypothesis – Large sample size

<table>
<thead>
<tr>
<th></th>
<th>No outliers</th>
<th>With one outliers</th>
<th>With two outliers</th>
<th>With four Outliers</th>
<th>With six outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>null reject</strong></td>
<td>4.98</td>
<td>4.88</td>
<td>6.17</td>
<td>9.91</td>
<td>16.1</td>
</tr>
<tr>
<td><strong>null accept</strong></td>
<td>95.02</td>
<td>95.12</td>
<td>93.83</td>
<td>90.09</td>
<td>83.9</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

In the case of the large sample with no outliers, the type-I error ratio is very close to the expected error ratio of five percent (Table 4). This is also the case when the sample has one outlier within it. The 4.98 and 4.88 percent for the no outlier and with one outlier samples cannot be proven to differ from five percent. It is easy to conclude that the effect of one outlier is nonexistent or at least small enough not to be a factor. The situation when a negative outlier in group B also is included in the sample, the ratio of type-I error extends the five percent threshold by 1.17 percentage points. This indicates that two outliers are enough to disturb the hypothesis testing, increasing the ratio of failures. For the situation when there are four outliers in the sample, the type-I error ratio is just shy of 10 percent. Even though this is a proof of sensitivity, it is also a proof of the large sample being superior in handling outliers compared to the medium or small sample (compare Tables 2 and 3). In the medium sample one is more than two times more likely to make a type-I error than in the large sample, and in the small sample one is more than three times more likely to make a type-I error, when there are four outliers included. This is also the case when the sample has six outliers within it. The large sample has a type-I error ratio of 16.1 percent while the medium sample has an error ratio of 56.34 percent and the small sample a ratio of 83.36 percent. It is therefore fair to claim that a larger sample is less sensitive to a few extreme outliers compared to a smaller sample.

The large sample cannot prove that one single outlier has an effect on the type-I error probability. However when two, four or six extreme outliers are included in the sample, all situations fail to be below the threshold of five percent. Therefore it is obvious that the one-way ANOVA is sensitive to outliers, especially if there are more than one extremely large or negative outlier. The ANOVA performs better though, if the sample is large relative to a smaller sample, as the type-I error probability is less in the large sample.
5.0. Conclusion
This paper can conclude that the parameter estimates used in the one-way analysis of variance are not robust against outliers. This is due to the fact that one single observation may cause the estimate to deviate exceptionally far from the true value. One-way ANOVA cannot be considered to be robust as outliers affect the type-I error probability. This is evidence in the line of the studies made by; (Scariano & Davenport, 1987) (Hoaglin, Mosteller, & Tukey, 1983) (Huber, 1981) (Krishnaiah, 1980). From the results of the simulations in this study, it is proven that one outlier in a small, medium or large sample has little or no effect on the type-I error probability. However from the point when two, or more, substantial outliers are included in the sample, the type-I error probability increases. A smaller sample performs in general worse, in the sense of type-I errors, than a large sample, even though any sample size is effected by outliers.

This study may also have found a special case where the one-way ANOVA performs better than the significance level. It is the case when the sample consists of 100 observations, with equal group sizes and one extremely large outlier in one of the groups. One reasonable explanation would be that the outlier in fact affects the estimates but in that sense that it decreases the probability of making a type-I error. Needless to say, the special case needs to be examined even more to find what’s causing it.

Different sample sizes have an effect on how likely one is to make a type-I error. Smaller samples have a higher risk of making type-I errors compared to larger samples. This is due to the fact that one single observation has a smaller influence in a larger sample relative to a small sample. However a larger sample doesn’t offset the fact that an outlier may cause an impact on the hypothesis testing.

From the results of this study it is reasonable to ask if the one-way ANOVA is suitable for inference based on its sensitivity towards extreme outliers. It is therefore worth mentioning that this study simulates extreme cases. Small to large sample sizes may perform in an acceptable manner when only one extreme outlier is in the sample, but are far more likely to make a type-I error if the numbers of outliers are increased. The one-way ANOVA is suitable as it is easy to apply and interpret but the analyst needs to understand that the conclusion from the hypothesis testing has an increasing probability of being wrong if a few substantial outliers are included in the sample.

This study proposes forthcoming research to, not only furthermore investigating what have caused the special case, but also issue depth and width for the understanding of outliers’ impact on the one-way ANOVA. This could be done using different levels of extreme outliers. A study regarding the outliers’ effect on the type-II error probability may also lead to further understanding of the test.
References


Appendix

Appendix include the R-code for the simulation:

```r
# Förr skall den underliggande population bildas med N=1miljon och två variabler. En numerisk och en kategorisk

#/* den kategoriska variablen har tre nivåer A, B och C*//
y <- sample( LETTERS[1:3], 1000000, replace=TRUE, prob=c(1/3, 1/3, 1/3) )
#/* Y är ett matrix av variablen y*//
y <- matrix(y)
#/* x är en standard normalfördelad variabel*//
x <- rnorm(1000000 , mean=0, sd=1)
#/* x är ett matrix av variablen x*//
x <- matrix(x)
#// en standard pop med tre grupper och variabeln x som är norm *//
xy <- cbind(x,y)

# x och y variablen kombineras till en grupp av data. Den gruppen döps om till pop för att mer enkelt ha koll på den

pop <- xy

### för att kunna dra stickprov från populationen måste den konverteras till ett data set samt att definera numerisk och kategorisk variabel

# kod för att kolla dimesionen

dim(xy)

#converterar xy till en data.frame

tab <- data.frame(xy)

summary(tab)

#försäkrar mig med att X1 är numerisk och X2 är kategorisk

tab$X1 <- as.numeric(x)
tab$X2 <- as.factor(y)

# ANOVA

### för att genomföra en ANOVA dras tre storlekar av stickprov.
# H0 alla mu är lika  
# sedan simuleras totalt 10 000 stickprov för varje stickprovsstorlek och situation  
# set.seed(1863969764) är seedkoden för simmuleringen. Denna är till för att enklare replikera testen.

# litet stickprov utan outlier:

set.seed(1863969764)#seed för de första tusen. en lägstill för nästa till tio är gjorda
N <- 10000# Antal replikat
pval1 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) { # Själva loopen
  S1 <- pop[sample(1:nrow(pop), 100, replace=FALSE),]
  ANOVA1 <- aov(S1X1 ~ S1X2)
  summary(ANOVA1)
  pval1[i] <- summary(ANOVA1)[[1]]["Pr(>F)"][1] # Stoppa in p-värden i element i i vektorn pval
}

mean(pval1 < 0.05) # Proportionen förkastningar
```
## medelstorst stickprov utan outliers

```r
set.seed(1863969764)
N <- 10000 # Antal replikat
pval2 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
S2 <- pop[sample(1:nrow(pop), 300, replace=FALSE),]
ANOVA2 <- aov(S2$X1 ~ S2$X2)
summary(ANOVA2)
pval2[i] <- summary(ANOVA2)[[1]]["Pr(>F)"][[1]] # Stoppa in p-värdet i element i i vektorn pval
}
mean(pval2 < 0.05) # Proportionen förkastningar
```

## stort stickprov utan outliers

```r
set.seed(1863969764)
N <- 10000 # Antal replikat
pval3 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
S3 <- pop[sample(1:nrow(pop), 3000, replace=FALSE),
ANOVA3 <- aov(S3$X1 ~ S3$X2)
summary(ANOVA3)
pval3[i] <- summary(ANOVA3)[[1]]["Pr(>F)"][[1]] # Stoppa in p-värdet i element i i vektorn pval
}
mean(pval3 < 0.05) # Proportionen förkastningar
```

### Outliers

```r
## En outlier skapas för att ingå i stickprovet. Outlier har x-värdet 10 och är i grupp C.
## Detta skapas som en till pop och kombineras med stickprovet

```r
cbind(x, y)
```

```r
# OutTen är data.frame for en obs som är en outlier där X1=10 och X2="C"
OutTen <- data.frame(x)
```

```r
SN <- pop[sample(1:nrow(pop), 99, replace=FALSE),]
```
#SN utan outliers slås ihop med outliern
So1<-rbind(data=OutTen, data=SN)
summary(So1)

###########################################################################################################
## med en outlier i grupp C
# H0 alla mu är lika
###########################################################################################################
## Litet stickprov med en outlier i grupp C
set.seed(1863969764)
N <- 10000 # Antal replikat
pval4 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SN <- pop[sample(1:nrow(pop), 99, replace=FALSE),]
  So1<-rbind(data=OutTen, data=SN)
  ANOVAo1 <- aov(So1$X1 ~ So1$X2)
pval4[i] <- summary(ANOVAo1)[[1]]["Pr(>F)"][[1]] # Stoppa in p-värden i element i i vektorn pval
}
mean(pval4 < 0.05) # Proportionen förkastningar

## medelstort stickprov med en outlier i grupp C
set.seed(1863969764)
N <- 10000 # Antal replikat
pval5 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SN2 <- pop[sample(1:nrow(pop), 299, replace=FALSE),]
  So2<-rbind(data=OutTen, data=SN2)
  ANOVAo2 <- aov(So2$X1 ~ So2$X2)
pval5[i] <- summary(ANOVAo2)[[1]]["Pr(>F)"][[1]] # Stoppa in p-värden i element i i vektorn pval
}
mean(pval5 < 0.05) # Proportionen förkastningar

## stort stickprov med en outlier i grupp C
set.seed(1863969764)
N <- 10000 # Antal replikat
pval6 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SN3 <- pop[sample(1:nrow(pop), 2999, replace=FALSE),]
  So3<-rbind(data=OutTen, data=SN3)
  ANOVAo3 <- aov(So3$X1 ~ So3$X2)
pval6[i] <- summary(ANOVAo3)[[1]]["Pr(>F)"][[1]] # Stoppa in p-värden i element i i vektorn pval
}
mean(pval6 < 0.05) # Proportionen förkastningar

###########################################################################################################
## nästa situation skall även en negativ outlier inkluderas i stickprovet.
## Därför skapas en population med x-värden minus 10 i grupp B.

y <- sample( LETTERS[1:3], 1, replace=TRUE, prob=c(0, 1, 0) )
#*/Y är ett matrix av variablen y*/
y <- matrix(y)
x<-rnorm(1 , mean=-10, sd=0)
#//x är ett matrix av variabeln x*//
x <- matrix(x)

xy <- cbind(x, y)
#OutmTen är data.frame for en obs som är en outlier där X1=-10 och X2="B"
OutmTen <-data.frame(xy)

OutmTen$X1<as.numeric(x)
OutmTen$X2<as.factor(y)

# Ett stickprov består av n-2 "vaniga" observationer och kombineras med de två outliers. EX:
SM <- pop[sample(1:nrow(pop), 98,
             replace=FALSE),]

#SM utan outliers slås ihop med outliers
Som1<-rbind(data=OutTen, data=SM, data=OutmTen)

summary(Som1)

# Ett stickprov består av n-2 "vaniga" observationer och kombineras med de två outliers. EX:
N <- 10000 # Antal replikat
pval7 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SM <- pop[sample(1:nrow(pop), 98,
                  replace=FALSE),]
  Som1<-rbind(data=OutTen, data=SM, data=OutmTen)
  ANOVAom1 <- aov(Som1$X1 ~ Som1$X2)
  pval7[i] <- summary(ANOVAom1)$[1]["Pr(>F)"$][1] # Stoppa in p-värdet i element i i vektorn pval
}

mean(pval7 < 0.05) # Proportionen förkastningar

# Ett stickprov består av n-2 "vaniga" observationer och kombineras med de två outliers. EX:
N <- 10000 # Antal replikat
pval8 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SM2 <- pop[sample(1:nrow(pop), 298,
                  replace=FALSE),]
  Som2<-rbind(data=OutTen, data=SM2, data=OutmTen)
  ANOVAom2 <- aov(Som2$X1 ~ Som2$X2)
  pval8[i] <- summary(ANOVAom2)$[1]["Pr(>F)"$][1] # Stoppa in p-värdet i element i i vektorn pval
}

mean(pval8 < 0.05) # Proportionen förkastningar

# Ett stickprov består av n-2 "vaniga" observationer och kombineras med de två outliers. EX:
N <- 10000 # Antal replikat
pval9 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  SM3 <- pop[sample(1:nrow(pop), 2998,
                  replace=FALSE),]
  Som3<-rbind(data=OutTen, data=SM3, data=OutmTen)
  ANOVAom3 <- aov(Som3$X1 ~ Som3$X2)
  pval9[i] <- summary(ANOVAom3)$[1]["Pr(>F)"$][1] # Stoppa in p-värdet i element i i vektorn pval
}


mean(pval9 < 0.05) # Proportionen förkastningar

### Nu slall stickprovet innehålla totalt 4 outlier. Två negativa i grupp B och två positiva i grupp C. 
### Detta skapas med:

```r
y <- sample( LETTERS[1:3], 2, replace=TRUE, prob=c(0, 0, 1) )
```  
//Y är ett matrix av variablen y*/

```r
x <- rnorm(2 , mean=10, sd=0)
```  
//x är ett matrix av variablen x*/

```r
xy <- cbind(x, y)
```  
#Out2Ten är data.frame for två obs som är outlier där X1=10 och X2="C"

```r
Out2Ten <- data.frame(xy)
```  
```r
Out2Ten$X1< as.numeric(x)
```  
```r
Out2Ten$X2< as.factor(y)
```  
```r
y <- sample( LETTERS[1:3], 2, replace=TRUE, prob=c(0, 1, 0) )
```  
//Y är ett matrix av variablen y*/

```r
x <- rnorm(2 , mean=-10, sd=0)
```  
//x är ett matrix av variablen x*/

```r
xy <- cbind(x, y)
```  
#Out2mTen är data.frame for två obs som är outlier där X1=-10 och X2="B"

```r
Out2mTen <- data.frame(xy)
```  
```r
Out2mTen$X1< as.numeric(x)
```  
```r
Out2mTen$X2< as.factor(y)
```  
### ett stickprov består av n-4 "vanliga" observation och kompineras med outliers. EX:

```r
S2M <- pop[sample(1:nrow(pop), 96, replace=FALSE),]
```  
`#SM utan outliers slås ihop med outliers

```r
S2om1 <- rbind(data=Out2Ten, data=S2M, data=Out2mTen)
```  
```r
summary(S2om1)
```  
### litet stickprov med två negativ Out. i B och två positiv Out. i C.

```r
set.seed(1863969764)
```  
```r
N <- 10000 # Antal replikat
```  
```r
pval10 <- rep(0, N) # vektor som kommer innehålla alla p-värden for (i in 1:N) {
```  
```r
S2M <- pop[sample(1:nrow(pop), 96, replace=FALSE),]
```  
`#SM utan outliers slås ihop med outliers

```r
S2om1 <- rbind(data=Out2Ten, data=S2M, data=Out2mTen)
```  
```r
ANOVA2om1 <- aov(S2om1 ~ S2om1$X2)
```  
```r
pval10[i] <- summary(ANOVA2om1)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
```  
```r
mean(pval10 < 0.05) # Proportionen förkastningar
```
## medelstort stickprov med två negativ Out. i B och två positiv Out. i C.
set.seed(1863969764)
N <- 10000 # Antal replikat
tpval11 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S2M2 <- pop[sample(1:nrow(pop), 296, replace=FALSE),]
  S2om2 <- rbind(data=Out2Ten, data=S2M2, data=Out2mTen)
  ANOVA2om2 <- aov(S2om2$X1 ~ S2om2$X2)
  tpval11[i] <- summary(ANOVA2om2)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
}
mean(tpval11 < 0.05) # Proportionen förkastningar

## Stort stickprov med två negativ Out. i B och två positiv Out. i C.
## Detta blir dock för mycket för datorn att arbeta med så simuleringen görs 2*5000.
set.seed(1863969765)
N <- 5000 # Antal replikat
tpval12a <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S2M3 <- pop[sample(1:nrow(pop), 2996, replace=FALSE),]
  S2om3 <- rbind(data=Out2Ten, data=S2M3, data=Out2mTen)
  ANOVA2om3 <- aov(S2om3$X1 ~ S2om3$X2)
  tpval12a[i] <- summary(ANOVA2om3)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
}
mean(tpval12a < 0.05) # Proportionen förkastningar

set.seed(1863969766)
N <- 5000 # Antal replikat
tpval12b <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S2M3 <- pop[sample(1:nrow(pop), 2996, replace=FALSE),]
  S2om3 <- rbind(data=Out2Ten, data=S2M3, data=Out2mTen)
  ANOVA2om3 <- aov(S2om3$X1 ~ S2om3$X2)
  tpval12b[i] <- summary(ANOVA2om3)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
}
mean(tpval12b < 0.05) # Proportionen förkastningar

#####################################################################################################
########
#####################################################################################################
## sist skall ett stickprov innehålla 6 outliers. dessa är; 3 negativa i grupp B och 3 positiva i grupp C
## först skapas outliers:
y <- sample( LETTERS[1:3], 3, replace=TRUE, prob=c(0, 0, 1) )
##/*Y är ett matrix av variablen y*/
y <- matrix(y)
x<-rnorm(3 , mean=10, sd=0)
##/*x är ett matrix av variabeln x*/
x <- matrix(x)
xy <- cbind(x, y)
#Out3Ten är data.frame for 3 obs som är outlier där X1=10 och X2="C"
Out3Ten <- data.frame(xy)
Out3Ten$X1<-as.numeric(x)
Out3Ten$X2<-as.factor(y)

y <- sample( LETTERS[1:3], 3, replace=TRUE, prob=c(0, 1, 0) )
# "Y är ett matrix av variablen y"

y <- matrix(y)
x<-rnorm(3 , mean=-10, sd=0)
# "x är ett matrix av variablen x"
x <- matrix(x)

xy <- cbind(x, y)
#Out3mTen är data.frame för 3 obs som är outlier där X1=-10 och X2="B"
Out3mTen <- data.frame(xy)

Out3mTen$X1<-as.numeric(x)
Out3mTen$X2<-as.factor(y)

## ett stickprov består av n-6 "vanliga" observationer och kombineras med 6 outliers. EX:
S3M <- pop[sample(1:nrow(pop), 94, replace=FALSE),]

#SM utan outliers slås ihop med outliers
S3om1<-rbind(data=Out3Ten, data=S3M, data=Out3mTen)

summary(S3om1)

--------------------------------------------------------------------------------

#### ANOVA

#### litet stickprov med tre negativ Out. i B och tre positiv Out. i C.
set.seed(1863969764)
N <- 10000 # Antal replikat
pval13 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S3M <- pop[sample(1:nrow(pop), 94, replace=FALSE),]
  S3om1<-rbind(data=Out3Ten, data=S3M, data=Out3mTen)
  ANOVA3om1 <- aov(S3om1$X1 ~ S3om1$X2)
  pval13[i] <- summary(ANOVA3om1)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
}

mean(pval13 < 0.05) # Proportionen förkastningar

#### Medelstort stickprov med tre negativ Out. i B och tre positiv Out. i C.
set.seed(1863969764)
N <- 10000 # Antal replikat
pval14 <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S3M2 <- pop[sample(1:nrow(pop), 294, replace=FALSE),]
  S3om2<-rbind(data=Out3Ten, data=S3M2, data=Out3mTen)
  ANOVA3om2 <- aov(S3om2$X1 ~ S3om2$X2)
  pval14[i] <- summary(ANOVA3om2)[[1]]["Pr(>F)"][1] # Stoppa in p-värdet i element i i vektorn pval
}

mean(pval14 < 0.05) # Proportionen förkastningar

#### Stort stickprov med tre negativ Out. i B och tre positiv Out. i C.
#### Detta blir dock för mycket för datorn att arbeta med så simuleringen görs 2*5000.

set.seed(1863969764)
N <- 5000 # Antal replikat
pval15a <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S3M3 <- pop[sample(1:nrow(pop), 2994, 
    replace=FALSE),]
  S3om3 <- rbind(data=Out3Ten, data=S3M3, data=Out3mTen)
  ANOVA3om3 <- aov(S3om3$X1 ~ S3om3$X2)
  pval15a[i] <- summary(ANOVA3om3)[[1]]["Pr(>F)" ][1] # Stoppa in p-värdet i element i i vektorn pval
}
mean(pval15a < 0.05) # Proportionen förkastningar

set.seed(1863969765)
N <- 5000 # Antal replikat
pval15b <- rep(0, N) # vektor som kommer innehålla alla p-värden
for (i in 1:N) {
  S3M3 <- pop[sample(1:nrow(pop), 2994, 
    replace=FALSE),]
  S3om3 <- rbind(data=Out3Ten, data=S3M3, data=Out3mTen)
  ANOVA3om3 <- aov(S3om3$X1 ~ S3om3$X2)
  pval15b[i] <- summary(ANOVA3om3)[[1]]["Pr(>F)" ][1] # Stoppa in p-värdet i element i i vektorn pval
}
mean(pval15b < 0.05) # Proportionen förkastningar