UPPSALA UNIVERSITET

Fisher's Randomization Test versus Neyman's Average Treatment Test

By Andreas Estmark, Kajsa Georgii

Department of Statistics
Uppsala University

Supervisor: Per Johansson

2019

Abstract

The following essay describes and compares Fisher's Randomization Test and Neyman's average treatment test, with the intention of concluding an easily understood blueprint for the comprehension of the practical execution of the tests and the conditions surrounding them. Focus will also be directed towards the tests' different implications on statistical inference and how the design of a study in relation to assumptions affects the external validity of the results. The essay is structured so that firstly the tests are presented and evaluated, then their different advantages and limitations are put against each other before they are applied to a data set as a practical example. Lastly the results obtained from the data set are compared in the Discussion section.

The example used in this paper, which compares cigarette consumption after having treated one group with nicotine patches and another with fake nicotine patches, shows a decrease in cigarette consumption for both tests. The tests differ however, as the result from the Neyman test can be made valid for the population of interest. Fisher's test on the other hand only identifies the effect derived from the sample, consequently the test cannot draw conclusions about the population of heavy smokers.

In short, the findings of this paper suggests that a combined use of the two tests would be the most appropriate way to test for treatment effect. Firstly one could use the Fisher test to check if any effect at all exist in the experiment, and then one could use the Neyman test to compensate the findings of the Fisher test, by estimating an average treatment effect for example.

Keywords

Nonparametric, Parametric, Monte Carlo approximation, Inference.

Contents

List of Figures

List of Tables

Introduction

Ronald Fisher (1890–1962) and Jerzy Neyman (1894–1981) were two statisticians active during the 20th century. They had different views on how to test for the effect of a treatment on two different groups. Fisher believed that testing for an effect on the individual level was superior, whereas Neyman wanted to test for an average effect. The aim of this paper is to explore and compare Fisher's Randomization Test (FRT) and Neyman's average treatment test, the former being a non-parametric test and the latter a parametric test. The advantages and limitations of these tests will be discussed, as well as how they compare when applied to a data set.

Theoretical Framework

An Experiment

Imagine that some researchers want to conduct an experiment to investigate if a treatment (called W) has any effect. This treatment can be anything from different types of medication to a simple conversation, but the intention is the same: one wants to analyze the effect of the treatment (in this case it will be assumed that the experiment is conducted upon humans, but it could equally be performed upon animals or bacteria). This is done by comparing the fixed outcomes under different treatments for several units in a given population. The researchers choose n individuals at random from a population of N individuals to participate in the experiment. Since one cannot observe the true treatment effect on an individual level, also called the fundamental problem of inference, the individuals are divided into two groups: one treatment group ($W_i = 1$, where i = 1, 2, ..., n_1) and one non-treatment group ($W_i = 0$, where i = 1, 2, ..., n_0). The underlying problem is that one cannot record the effect of a certain individual both taking the treatment and not taking the treatment, for the same time span, to compare the results. The assignment is done through a stochastic assignment mechanism, like the flipping of a coin. Then $Y_i(w)$ becomes the notation of potential outcome of either treatment or no treatment to an individual i (Imbens and Rubin, 2015).

After having conducted the experiment there are two separate ways to test the effect, either one can test for average treatment effect or one can test if any effect exists at all. This is done separately through Neyman's test and FRT respectively.

Conditions

If the experiment described above is well conducted, the assignment to treatment will be unconfounded (also known as the exchangeability assumption). This means that there will be no external variables affecting the result of the test. Commonly confoundedness occurs if the researchers involved in the experiment are aware of the assignment of subjects and therefore treat them accordingly. Subjects assigned to the treatment group may be given more attention if their allocation is known. Double blinding of both subjects (if human) and researchers is usually necessary to avoid this problem (Kendall, 2003).

The Stable Unit Treatment Value Assumption (SUTVA), which underlies both Neyman and Fisher's tests, is made up of two parts: firstly no interference between the assignment of units and secondly no variation of the treatment within the units (Schwartz, Gatto and Campbell, 2012). If fulfilled, then $Y_i(W_i) = Y_i$, which means that an individual's outcome only depends on

the treatment given and not any interaction effects received from other units (Venmans, 2016).

Violation of SUTVA may occur in different situations. If researchers were to investigate the effect of rowing on weight loss, then the effect of the treatment on different individuals could vary depending on the power of the stream and the weather conditions during certain days. The SUTVA can also be violated if units have potential of affecting each other such as when studying infectious deceases, where the infection of one unit may be the cause of infection in another unit due to exposure (Schwartz, Gatto and Campbell, 2012).

When both SUTVA and unconfoundedness is fulfilled, then the treatment is independent from the potential outcome:

$$(y(1), y(0)) \perp\!\!\!\perp x_1 \tag{1}$$

where x_1 is the treatment given that it is constant across units (Venmans, 2016).

For the Neyman test, random sampling is vital if inference is to be made for the target population (the process of generalizing results from the sample to the population is refereed to as *external validity*). Random sampling is in many cases not correctly fulfilled, since the ostensible randomness of the selection is more than often fallacious and the fault itself is very difficult to detect since the latent bias of the selection may not be visible to the eye nor through any statistical tests. Bias occurs when a systematic deviation exists between the expected values observed by the study and the true values found in the source population. This paper will mostly focus on *selection bias*, which is caused by a sample which does not accurately represent the source population. There are several factors which may contribute to selection bias. Some of them are: under-coverage (certain groups in the source population are not represented in the sample), over-coverage (including respondents in the study who are not from the targeted population), non response bias (meaning that people who refuse to participate in the study are underrepresented) and voluntary bias (occurs when respondents are self-selected volunteers, which tends to over-represent individuals who have strong opinions about the research topic)(Glen, 2013).

Non-parametric Permutation Test

Non-parametric permutation tests are a part of Randomization tests, which rely on reorderings of the data to calculate the desired test statistic. It was initially introduced by Ronald Fisher in his book *The Design of Experiment*, published in 1935. Usually the tests concern themselves with evaluating differences in group means, medians or two variables. It can also apply to analysis

of variance or permutation tests on factorial ANOVA design (Howell, 2018).

For non-parametric permutation tests little interest is shown in the characteristics of the source population. The only assumption made about the population is that it is finite and tangible (Dransfield and Brightwell). These tests do not concern themselves with any parameters in a population model, meaning that the parameters are not estimated. If one were to test the effect of painkillers on two different groups (one placebo group and one group which actually received the pills), the following would be tested: no effect versus effect. The hypotheses could also be written as $H_0 : Y_i(0) = Y_i(1)$ and $H_1 : Y_i(0) \neq Y_i(1)$, where $Y_i(0)$ and $Y_i(1)$ are the outcomes of no treatment and of treatment respectively (i = 1, 2, ..., N).

Since non-parametric tests do not estimate parameters this creates no need for assumptions concerning the distribution of the potential outcomes (Martin and McFerran, 2017). As an example, a population from which a random sample is drawn is not required to be normally distributed, compared to Neyman's test which requires large sample normal approximation for its validity. Furthermore, the sampling of the data does not need to be random, the error distribution does not need to be known and homoscedasticity is not essential for the analysis (Imbens and Rubin, 2015).

Even though non-parametric tests may rely on very few assumptions, there are still some assumptions which are of great importance. The primary assumptions which need to be fulfilled is the concept of *exchangeability* (also refereed to as the unconfounding assumption) and the SUTVA. The former states that if the null hypothesis is true (there is no effect of treatment in our painkiller experiment), then any subject from any of the two groups will have the same outcome no matter which group they are placed in. Thus, the two groups will have the same expected values no matter how the subjects are shuffled around between the groups (Howell, 2018). This also mean that the data are free from confounding factors, meaning that there are no other variables which are affecting the outcome of the treatment (Schwartz, Gatto and Campbell, 2012).

To understand the process of an FRT, a simple example will follow:

Table 1: Example of Fisher's Randomization Test

Control	Treatment
8	7
6	4

A sample size of n = 4 is collected where the observations are divided into two groups of equal sizes: $n_c = 2$ and $n_t = 2$ (c = control group, t = treatment group), by using complete randomization (each subject has the same proba-

bility of receiving treatment). The observations are in table 1. The absolute difference between means for the two groups is calculated, which is t^{obs}_{Fisher} = 1.5 (group means are 7 and 5.5 respectively). The equation for calculating the difference is given below, where $\overline{Y}^{obs}_t$ is the mean for the treatment group and $\overline{Y}^{obs}_c$ is the mean for the control group (subscripted t and c will denote calculations or values for the treatment group and the control group respectively through out the essay).

$$t^{obs}_{Fisher} = \mid \overline{Y}^{obs}_t - \overline{Y}^{obs}_c \mid . \tag{2}$$

The second step is to calculate the number of ways in which the four observations can be divided into the two groups using complete randomization, which in this case is six different ways: $n!/(n_c!n_t!)$.

Let W_i = 1 if individual i is treated and W_i = 0 if assigned to the placebo group. The vector below illustrates the six different ways in which the four individuals can be shuffled around.

$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

For these six reorderings the mean difference should be calculated for each of them, as given by equation (2). The results are given below, where the first value is the mean difference for the first column in the above vector, and so forth. Each value is in other words t_{Fisher} for each reordering, while t_j (j = 1, 2, ..., 6) is the distribution of the estimates under H_0.

$$t_j = (\ 1.5\ ,\quad 1.5\ ,\quad 0.5\ ,\quad 0.5\ ,\quad 2.5\ ,\quad 2.5\ )$$

With the information obtained one should consider if any of these six different group options produce any combination of numbers where the group means differ equally or more than what they did in the observed sample. The matrix above is exhaustive in the sense that it contains all possible allocations for the data. This means that all possible permutations are used to obtain an exact p-value (the exactness of the p-value will be further discussed in the section: Advantages and Limitations). For this example there are four different ways in which the group means differ 1.5 or more.

For a one-tailed test, the test statistic is calculated by taking the four different permutations in where the group means are 1.5 or greater and divide it by the total number of permutations (which is six in total). This would be $\frac{2}{3}$. A two tailed test is obtained by calculating the proportion of the total number of permutations where the absolute difference is greater than or equal to 1.5, which is $\frac{2}{3}$. This method calculates the likelihood of observing the recorded

values for the two groups compared to if the groups had been completely randomized (Lane).

A significance level of 0.05 is used for this test. The null hypothesis states that the groups have the same effect while the alternative hypothesis says that there is a differential effect for at least one subject under the affect of another group (Edgington, 1986). The hypotheses can be written as the following:

$$H_0 : Y_i(0) = Y_i(1) \text{ versus } H_1 : Y_i(0) \neq Y_i(1) \ .$$

In this case one cannot reject the null hypothesis for either a one-tailed or a two-tailed test. The current sample size does not allow for rejection of H_0 for any reasonable size of α no matter the values of the groups, since the smallest p-value will never be lower than 0.1667. This is a good indicator that a bigger samples size is needed. The following equation obtains the p-value for a two-sided Fisher's Randomization Test:

$$\hat{P} = P(|t_j| \geq t^{obs}_{Fisher}|H_0) = \frac{\sum_{j=1}^{A} I(|t_j| \geq t^{obs}_{Fisher})}{A}, \tag{3}$$

where A is the number of or possible allocations. I(.) is one if the expression is true and zero if it is false.

The process of comparing values between groups using FRT can be summarized into four easy steps.

- Step one: identify the null hypothesis, i.e. decide what you want to measure.
- Step two: conduct a randomized experiment and calculate the test statistics for the sample.
- Step three: rearrange the sample in all possible reorderings and calculate the sample statistic for each new order.
- Step four: calculate the proportion of the sample statistics which have values that are equal to or more extreme than the ones calculated for the original sample.

Parametric Test

Jerzy Neyman developed a parametric test during the 1920s to analyze the difference in average treatment effect (ATE) between groups. As said previously, for each individual participating in a study, there would exist two outcomes: $Y_i(0)$ and $Y_i(1)$, which means the outcome under no treatment and treatment respectively. Given random sampling, inference can be made to the population

average treatment effect (PATE). It is defined as the difference between the averages of the not treated and treated outcomes (Imbens and Rubin, 2015):

$$PATE = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)). \tag{4}$$

While the *sample* average treatment effect (SATE) is defined as:

$$SATE = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)). \tag{5}$$

For this test the null hypothesis' statement is that the average effect is zero and the alternative hypothesis' statement is that the average effect is nonzero:

$$H_0 : PATE = 0 \quad H_1 : PATE \neq 0$$

$$H_0 : SATE = 0 \quad H_1 : SATE \neq 0 \ .$$

The Neyman test requires certain assumptions to be fulfilled. The purpose of the assumptions is to make the test as optimal as possible, and therefore robust, given its attributes. The primary assumption is the SUTVA. As discussed in the section Conditions, these restrictions may be exceedingly difficult to ensure since unintentional bias or external factors may affect the implementation of the treatment (Imbens and Rubin, 2015). But if the SUTVA is in fact fulfilled, then an unbiased estimator of the SATE can be obtained from the following estimator:

$$\hat{\tau} = \overline{Y}_t^{obs} - \overline{Y}_c^{obs}. \tag{6}$$

If random sampling can be assumed as well, then the above equation of mean difference is also an unbiased estimator of the PATE.

To test the average treatment effect of the population through a test of inference, the hypotheses would be the same: $H_0 : PATE = 0$ and $H_1 : PATE \neq 0$. To calculate the test statistic for such a test, one would firstly have to estimate the sample variance. In the equation below, s_c^2 and s_t^2 are the sample variance for the control group and the treatment group respectively. The number of individuals in each group is denoted by n_c and n_t (control and treatment).

$$\hat{\mathbb{V}}^{Neyman} = \frac{s_c^2}{n_c} + \frac{s_t^2}{n_t} \tag{7}$$

where

$$s_c^2 = \frac{1}{n_c - 1}\sum_{i \in w=0}^{n_c}(Y_i - \overline{y}_c)^2 \tag{8}$$

and

$$s_t^2 = \frac{1}{n_t - 1}\sum_{i \in w=1}^{n_t}(Y_i - \overline{y}_t)^2. \tag{9}$$

The sample variance estimator above is used even if there is not an additive treatment effect. For additive treatment effect $Y_i(1) = Y_i(0) + \delta$ follows under H_1, where δ is the effect of the treatment. The estimate is conservative (meaning that it is too big) and allows for unrestricted treatment effect. Moreover, the estimator is unbiased for a *constant* additive effect (Imbens and Rubin, 2015). The test statistic for a Neyman test is the difference between the two groups divided by the square root of the variance.

$$t_{Neyman} = \frac{\bar{Y}_t^{obs} - \bar{Y}_c^{obs}}{\sqrt{\hat{\mathbb{V}}^{Neyman}}} \sim t(n-1). \tag{10}$$

This test depends on large-sample normal approximation for its validity if inference is to be made for the PATE. If CLT and i.i.d. sampling is ensured, the mean difference estimator will rule to the following:

$$\hat{\tau} \longrightarrow N(\tau, \mathbb{V}^{Neyman}).$$

Advantages and Limitations

FRT and Neyman's test are dissimilar in many ways. Non-parametric permutation tests are advantageous since they do not depend on random sampling to be executed. Consequently non-parametric tests are fit to be applied to data which does not fulfill the classical assumptions of traditional statistical tests. Furthermore, since FRT is not limited by modular assumptions, the test becomes an exact one, meaning that there are no assumptions which may tinker with the precision of the p-value if not completely fulfilled. This means that a test with a significance level of 5% is going to have the null hypothesis rejected exactly 5% of the time if the test is redone numerous times (Metha and Petal).

The FRT can be considered less advantageous in other areas, as when considering computational mass for example. Undoubtedly the complexity of non-parametric permutation tests when one has to handle data containing huge amounts of observations was a severe problem fifty years ago, but with the exponential increase of computational power of computers this problem has become less acute. If one happens to encounter a problematic mass of possible reorderings for a data set (envision an assignment vector of millions of columns), one can use a Monte Carlo approximation to limit the amount of permutations. This is done by taking a small random sample out of all the possible number of permutations (Fay and Follmann, 2002). The exactness of the non-parametric test becomes arbitrary when using Monte Carlo, but it is still close enough to be considered an exact test (Lunneborg). Unfortunately one might loose some efficiency in the selection process, but this is dependent on the sample size of the permutations: the bigger the sample size, the smaller the loss of efficiency will be (Ernst M.D, 2004).

The FRT uses a sharp null hypothesis for testing (a hypothesis which allows the researcher to infer values for all different outcomes of the sample), while the typical Neyman test tests for the difference in ATE. The latter definition may be problematic if the preparation tested has both a negative and a positive treatment effect on the subjects, which may result in no change in ATE even though there is a considerable difference between the two groups (Imbens and Rubin, 2015).

It is possible that the Neyman test is in some ways favored due to the fact that it can produce an average effect. For example, when legislators want to examine the success of a labor market policy it may simplify the decision making process if one can present the average effect of the policy. This may be one of the reasons behind why the Neyman test is more widely used than FRT.

Method

Empirical Application

The following example aims to test the hypothesis that there is no average difference or no effect for the placebo and treatment group. To illuminate the differences between the Neyman and the Fisher test, a data set from a randomized control experiment is used (Laerd). A total of 30 heavy smokers were randomly and independently sampled and divided into two groups of equal size, where 15 of the heavy smokers received treatment for their use of cigarette consumption by wearing a nicotine patch and the remaining 15 were given a placebo treatment (a non-nicotine patch). To be categorized as a heavy smoker the individual smoked an average of 40 cigarettes or more per day. The data is numeric and represents the average number of cigarettes consumed per day, three months after the start of the treatment. The individuals were not aware of which group they were placed in.

Placebo is, in the context of clinical trials, the measurement of psychological rather the physiological effects. When testing the effect of two groups, it is normally desired for the groups to be as homogeneous as possible to be able to observe only the effect of the pertinent variable (in this case the nicotine patch) and no other unobserved factors. One should not confuse the placebo group with being equal to the test subjects in the other group without the given treatment. The placebo group still measures the psychological effect caused by the fake nicotine patch. This means that the inference calculated does not compare the effect of treatment to no treatment, but the effect of the treatment in relation to the fake treatment.

Table 2 and 3 present the raw data and descriptive statistics for the assembled data. The mean difference of the two groups are considerably different and the standard deviations are almost the same. Nothing displayed seems out of the ordinary.

Table 2: Average cigarette consumption per day three months after start of experiment

ID	Placebo	ID	Treatment
1	26	16	21
2	18	17	27
3	17	18	30
4	23	19	17
5	27	20	20
6	40	21	23
7	38	22	16
8	34	23	18
9	34	24	14
10	25	25	27
11	27	26	12
12	42	27	19
13	33	28	34
14	19	29	8
15	24	30	36

Table 3: Summary statistics of the two groups

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Placebo	15	28.467	7.981	17	23.5	34	42
Treatment	15	21.467	8.026	8	16.5	27	36

The data set will be used in evaluating Fisher and Neyman tests.

Fisher's Randomization Test

The observations in table 2 will be reallocated to the two groups 10,000 times and the distribution of the mean differences between the two groups will be used to obtain a p-value. The sampled permutations is called a Monte Carlo approximation and statisticians usually rely on everything from 1,000 to 10,000 reorderings (Influential Points). The number of permutations can be imagined as an enormous assignment vector, displaying all of the 10,000 reorderings for the 30 observations. The total number of allocations for this data is $\binom{30}{15} = 155\ 117\ 520$ which is an unfeasible amount to compute on a regular computer. A significance level of 5% will be used in this example.

The hypotheses of the test are different from traditional ones that hypothesize about a parameter, since the null hypothesis is a sharp null hypothesis that states that there is no treatment effect of the nicotine patches versus the alternative that there is a non-zero treatment effect. Setting up the two hypotheses that way enables differences to be made visible that otherwise may have been masked if centered around the mean or variance, since the negative and positive values can cancel out each other. Under H_0, with data taken from table 2, the potential outcomes can be expressed as:

$$Y^0(1) = \begin{bmatrix} 21 \\ 27 \\ 30 \\ 17 \\ \vdots \\ 36 \end{bmatrix} = Y^0(0) = \begin{bmatrix} 21 \\ 27 \\ 30 \\ 17 \\ \vdots \\ 36 \end{bmatrix}. \tag{11}$$

With the null hypothesis of no treatment effect, it is possible to assign the placebo group from the data the same values as the treatment group, as is done above. Additionally, it is important to analyze if there are extreme observations in the data. With outliers present and a test statistic of average differences, tests will have relatively low power. The low power is due to the effect of outliers on the mean estimate, but it can be addressed and mended by other statistics such as rank order statistics or the median. (Imbens and Rubin, 2015). Figure 2 shows how there are no obvious outliers in the data.

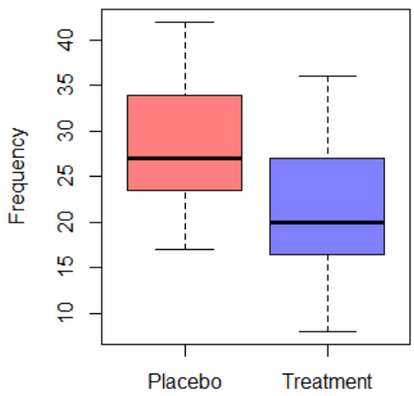

Figure 1: Boxplot of the treatment and placebo group

Following the calculation of the difference in means (-7) of the observed data, an FRT reallocates the observed data randomly into two groups. For every reallocation, the difference between the group means is calculated. In figure 2, the 10,000 permutations are visualized in a histogram.

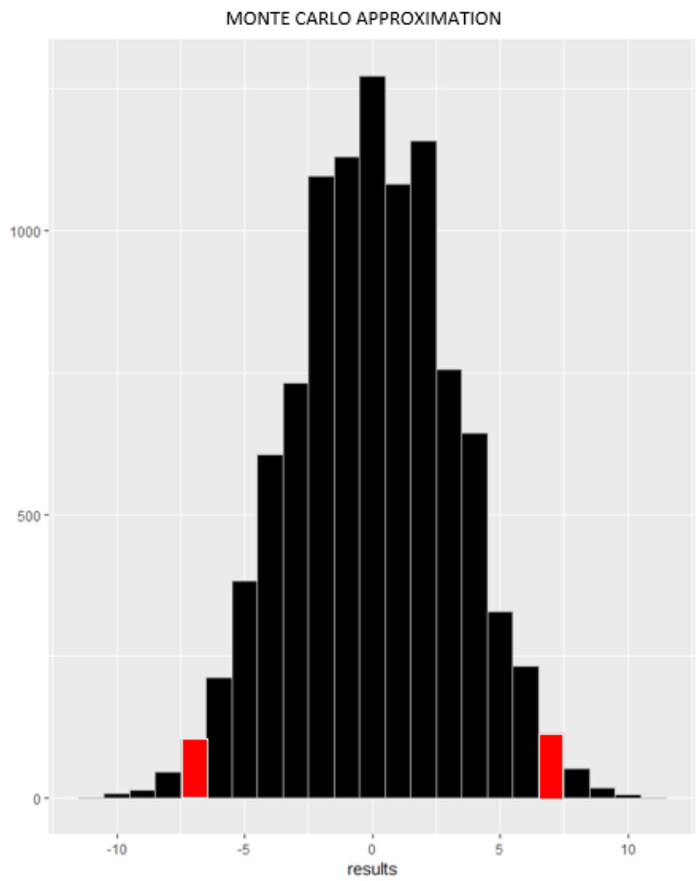

Figure 2: Randomization distribution of mean difference

The bin in red in figure 1 shows where the observed mean difference of -7, and +7 is located. By the looks of it, one can suggest that the null hypothesis of no treatment effect for any individual is rejected.

Test statistic

The absolute mean difference between the groups is calculated by subtracting the mean for one group with the mean for the other group:

$$t_{Fisher}^{obs} = \mid \overline{Y}_t^{obs} - \overline{Y}_c^{obs} \mid = 7. \tag{12}$$

.

To calculate the p-value, the sum of all mean differences in absolute value greater than or equal to seven is divided by the number of permutations. It is appropriate to use this test statistic for the reasons explained in the previous pages regarding assumptions and outliers.

Equation 13 shows how the p-value is obtained. The t_j variable is the placeholder for when the absolute mean differences exceed 7.

$$\hat{P} = P(|t_j| \geq 7|H_0) = \frac{\sum_{j=1}^{10,000} I(|t_j| \geq 7)}{10,000} = 0.0254. \tag{13}$$

Due to the Monte Carlo approximation, the precision of the estimation of the p-value can be estimated using the following equation:

$$\text{Standard Error of } \hat{p} = \sqrt{\hat{p}(1-\hat{p})/10,000} = 0.00157. \tag{14}$$

The standard error from equation 13 enables an interval of the obtained p-value. The exact p-value is obtained by going through all 155,117,520 permutations.

$$95\% \text{ interval of p-value} = [0.0285, 0.0223]. \tag{15}$$

The value of 0.0254 is smaller than the significance level of 0.05 and the interval is still within the limit of 0.05. It means that the probability of the result, given that the H_0 is true, is 0.0254. With the given significance level of α, the null hypothesis is therefore rejected.

A fiducial interval is derived for the test, sourced from Rosenbaum (2002), and is calculated through inversion of a constant effects hypothesis. The interval only covers negative values and does not touch zero, which is a good indicator that the null hypothesis was previously correctly rejected.

$$CI_{Fisher}^{0.95} = [-13.0000891, -0.9999703].$$

Fisher's Randomization test is very flexible regarding the test statistic. Contrary to Neyman's test, it is possible to device many different test statistics.

Other statistics such as the median or the rank order statistics can be used for the FRT. Figure 3 illustrates the randomization distribution of the median differences.

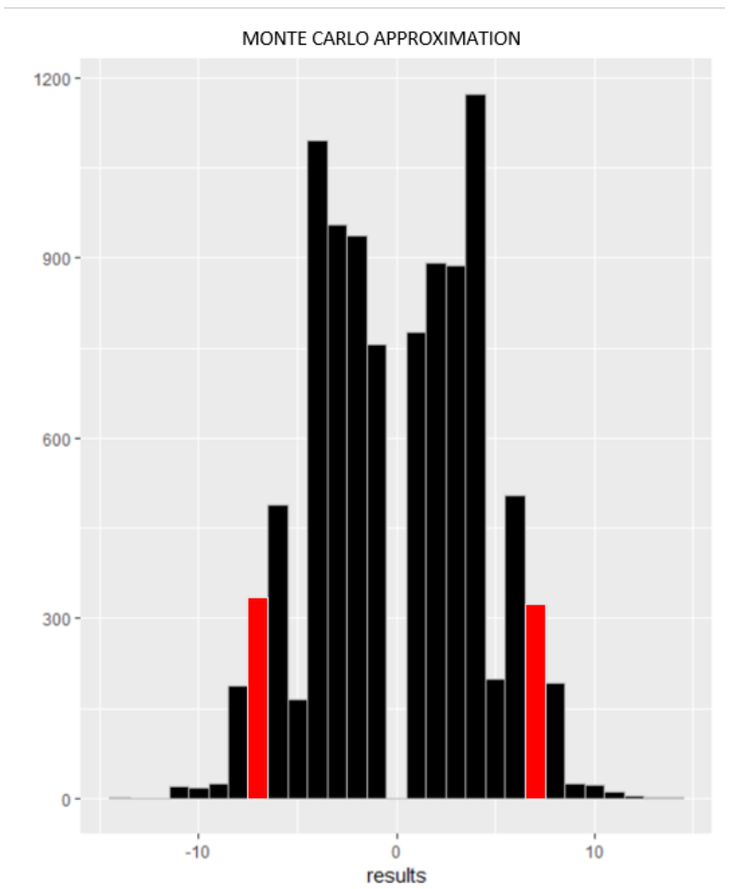

Figure 3: Randomization distribution of median difference

The observed difference in medians is 7 and the p-value is calculated in the same way as for the mean differences. The p-value of the test is 0.1143, thus it is higher than the p-value from the mean difference test. With the significance level of 0.05, the null hypothesis of no difference in medians, is not rejected. One possible explanation for the relatively high p-value is that no allocation resulted in a median difference of 0. Another possible explanation is that the variance of the sample median is larger than the variance of the sample mean in general and it is thus less efficient than the mean (Mendenhall, Scheaffer and Wackerly, 2011).

Neyman Test

As stated earlier, parametric tests use a null hypothesis formulated as a function about a parameter. In this case the parameter is the mean and this test requires assumptions that are not needed in FRT. When testing for average treatment effect, the hypotheses are different from the ones in the FRT. The null hypothesis statement is that the average effect of the nicotine patches is zero and the alternative hypothesis statement is that the average effect is non-zero. The significance level is once again set to 0.05.

$H_0^{Neyman} : PATE = 0$ vs $H_1^{Neyman} : PATE \neq 0$
or
$H_0^{Neyman} : SATE = 0$ vs $H_1^{Neyman} : SATE \neq 0$

In order for Neyman's test to be correct, large sample normal approximation of the test statistic, i.e $\bar{Y}_t^{obs} - \bar{Y}_c^{obs}$ is needed for the validity of the test. It is assumed that the sample of 30 individuals is a random sample from the population of heavy smokers, which makes it possible to infer the results for the whole population.

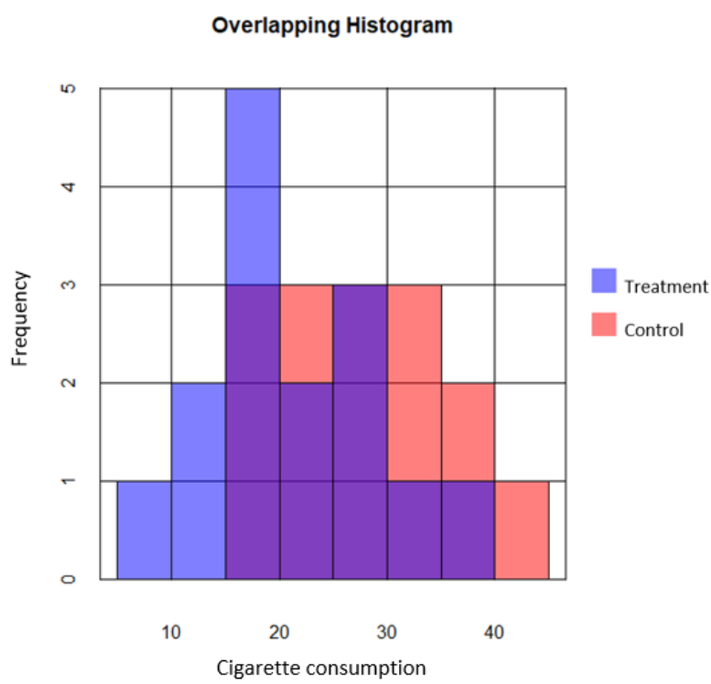

Figure 4: Overlapping histogram of the two groups

As the sample size is small, we therefore check for whether the outcomes are normal. For small n, the approximate normality of the difference in mean estimator can be questioned. However, the difference in mean estimator is normal if the outcomes themselves are normally distributed. Figure 4 displays

the outcomes for the treated and placebo groups. As the data set is made up of heavy smokers, there is a lower bound of the distributions. For example, the frequency cannot be zero or negative. A histogram is nevertheless not sufficient in determining the normality of the data. To formally test for normality, the Shapiro-Wilk normality test is used on respective group.

With the H_0 suggesting that the data is normally distributed, the p-value of the treatment test is 0.9092 and for the control test the p-value is 0.474. The H_0 is thus not rejected for the two tests. One can hence proceed by calculating a test statistic for the Neyman test, since the normality assumption of the difference in mean estimator is judged to be fulfilled.

Test statistic

In order to determine whether the use of nicotine patches has a non-zero average difference in cigarette consumption compared to the placebo group, the following test statistic is used.

$$t_{Neyman} = \frac{\bar{Y}_t^{obs} - \bar{Y}_c^{obs}}{\sqrt{\hat{\mathbb{V}}^{Neyman}}}. \tag{16}$$

The numerator of the test statistic is the observed mean difference between the treatment and placebo group, which is $\bar{Y}_t^{obs} - \bar{Y}_c^{obs} = -7$. The variance is estimated to be:

$$\hat{\mathbb{V}}^{Neyman} = \frac{s_c^2}{n_c} + \frac{s_t^2}{n_t} = 8.540317. \tag{17}$$

The $\mathbb{V}^{Neyman}$ is used to calculate the test statistic. This version of the variance is used because it is widely used even when constant treatment effect cannot be ascertained (Imbens and Rubin, 2015).

$$t_{Neyman} = \frac{\bar{Y}_t^{obs} - \bar{Y}_c^{obs}}{\sqrt{\hat{\mathbb{V}}^{Neyman}}} = \frac{-7}{2.922382} = -2.395306. \tag{18}$$

The resulting test statistic is -2.395306 and the p-value of the test is: 0.03114764 (from Student's t distribution). The null hypothesis of no average treatment effect is therefore rejected with a significance level of $0.05 = \alpha$.

The 95% confidence interval is:

$$CI_{Neyman}^{0.95} = \Big(-7 - \overbrace{1.96}^{\alpha/2} * \sqrt{\hat{\mathbb{V}}^{Neyman}}, -7 + 1.96 * \sqrt{\hat{\mathbb{V}}^{Neyman}} \Big) = \\ [-12.72787, -1.272131] \ .$$

The confidence interval does not include zero and the difference between the means is therefore statistically significant, but it is is rather wide.

Discussion

There are both theoretical and practical differences between the two tests discussed. Some differences have already been approached in the section *Advantages and Limitations* and will be further highlighted in this section, in which the aim is to discuss the differences between the tests in relation to the data set.

The p-value for the Neyman test is bigger than that of the FRT for the mean difference (observe table 4). This may be caused by numerous reasons, one being that the number of permutations will affect the p-value slightly, since the p-value of a Monte Carlo approximation only is an approximation of the true p-value. Furthermore, the probability coverage of the confidence interval, also found in table 4, may be erroneous if the normality assumption is not correctly fulfilled. The interval for the Fisher test is a fiducial interval and not a confidence interval, which means that the intervals cannot be interpreted the same way.

Table 4: Comparison between the two tests

Test	P-value	Interval
Fisher	0.0254	(-13.0000891, -0.9999703)
Neyman	0.03114764	(-12.72787,-1.272131)

When applying the average treatment effect estimator to the data, the value of -7 cigarettes is obtained. Neyman's test regards the difference in means whereas this is not the case for the Fisher test. A pharmaceutical company may decide to start the production of nicotine patches if they have evidence that they decreased cigarette consumption by 7 cigarettes, inferred from the experiment. At the same time, testing whether nicotine patches have any effect at all is also valuable. It can illustrate effects that are hidden when performing a statistical test on a parameter. Furthermore one still has the -7 estimate for the Fisher test, even though it may not be an estimator.

Another difference between the tests regards the estimated variance. The variance can be calculated in different ways when employing the Neyman test, which can result in an erroneous variance estimate. It is not the case in a permutation test as there is no need to include a variance estimate in any of the calculations.

The last subject to be approached in the comparison of the tests is the question regarding inference. Can the results derived for the two tests be generalized to the target population? Can one trust the accuracy of the assumptions enough to be able to assess any external validity?

Before being able to discuss the differences between Fisher's and Neyman's tests' assumptions and their implication on inference, it is proper to

have a few words be said about causality. Causation means that one event is the result of another event. Two variables that are correlated with each other do not imply that one is the result of the other. One example illustrates this difference: warm weather *causes* an increase in crime rates (there are several reasons for this, for example: heat makes people more irritable), this makes crime rates and ice cream consumption correlated but neither is the cause of the other (Marchand, 2017).

For both Neyman's and Fisher's test one assumes SUTVA and unconfoundedness. If the assumptions are fulfilled the results are more reliable, however if this is not the case and the SUTVA is violated the estimate of casual effect become unstable and less credible, since the conditions underlying the effect of the treatment will not be homogeneous and the uniqueness of each potential outcome for the units cannot be ensured. The exposure needs to have the same effect for all the individuals in the treatment group if the construct validity is not to be questioned (the extent to which a test measures what it claims to be measuring), while no interference between units needs to be absolute if external validity is to apply to the test. On the other hand if the assumption of unconfoundedness is violated, the internal validity of the test is jeopardized, meaning that the association estimated not automatically equals to causation. If exchangeability cannot be ensured, even the local effect of nicotine patches can be questioned and consequently all of the study results would peril (Schwartz, Gatto and Campbell, 2012).

What is out of the researchers control however is that the treatment effect may not have an exclusive impact on the participants' consumption of cigarettes. Some units may be affected by other factors which can contribute to an increase or decrease in effect, such as teaming up with friends to try to stop smoking (this could have a negative impact on cigarette consumption). These are latent factors which are almost impossible to detect and very difficult to compensate for. Researchers remain incognizant of their impact.

The assumption of no confounding is also debatable concept. It states that there should be no difference between the treatment and placebo group. We argue however that the units within the groups are not applicable to represent the effect for a hypothetical person who supposedly is using the nicotine patches versus not using the nicotine patches under the same conditions and during the same time period. This is due to the use of a placebo group. The placebo group cannot represent a person not participating in the study, since the person still experiences the effect of the placebo both contributed from the fake nicotine patches, but also the effect generated from the commitment of participation and the effects which may arise from the attention of researchers and reflection of behavior. This makes the assumption of unconfoundedness questionable, if it is expected that the same effect as in the experiment is to apply to people who are not using the nicotine patches versus the same people

actually using them. This causes questioning of the external validity of the test.

Random assignment and random sampling are assumptions which underlies the PATE. It is important to understand why these assumptions plays a crucial role in statistical inference. In a quixotic world where both conditions are met one is able to draw accurate conclusions about the causal relationships within the source population. Since random sampling rarely is a reality and only occasionally do one succeed in random assignment, this ends up having implications on the inference of the test. When neither assumption is fulfilled the results only describe the relationships within the sample. No other statistical conclusions can be drawn. In cases when random sampling is not fulfilled, but random assignment is, determination of causal relationships will be possible, but only within the sample, as for SATE and the FRT. On the contrary, if random sampling is fulfilled but random assignment is not, relationships within the population will be uncovered, but no causality can be determined (Khan Academy).

In the context of the given data it is difficult to assess if it is a randomized controlled trial. The statistician has to trust that the assignment of subjects to the two test groups has been done at random by researchers. The randomness of the selection process requires more investigation. The source population in this experiment are heavy smokers and therefore the target population has to be heavy smokers as well. We cannot hope to include "light" smokers or likewise in the generalization of the study results. There might have been some fall out in the selection process, for example: people who feel ashamed of their smoking habits would be less inclined to participate in the study. Other aspects of the selection process which should be put through questioning is how the subjects were obtained. Was the participation of the experiment decided by an application submission initiated by the subjects or were the subjects chosen and contacted directly by the researchers. These are questions to which we lack answers with the given data set and ultimately we have to approximate the validity of the randomization of the selection process if we wish for the results to apply to the target population.

In conclusion, the uncertainty underlying all assumptions spoken of and their implications on rudimentary inference leaves many to believe that less is more. In the case of the nicotine patch experiment, the random sampling assumption may or may not be correctly fulfilled, whereas the other assumptions' fulfillment is more believable. Since FRT only relies on SUTVA and unconfoundedness to conduct "truthful" inference for the sample, this seems like a more favourable choice of test since this would minimize the chances of encountering falsely accepted assumptions. This however entirely depends on the purpose of the test, since the ATE still is useful if one wishes to derive an estimate, or if one has the intention of having the research results apply to a

wider population. Perhaps the most optimal solution would be not to choose one or the other, but to combine the information obtained from both tests and from there on infer appropriate conclusions. For example one could start the experiment by using FRT to estimate if any effect at all exists and then, if an effect is detected, go on to compensate the analysis with results taken from the Neyman test. In regards to future research, analyzing how the two tests are utilized today and how their usage differs between fields of study, would be of interest.

Conclusion

To conclude: two different tests calculating if a treatment effect exists have been evaluated and compared to each other, with the purpose of mapping out their different structures, advantages and limitations. The essay focused on investigating the different tests' assumptions and their implications on inference.

By using the data set example for this paper, both tests showed that the treatment decreased average cigarette consumption by rejecting the null hypotheses. The tests also differed as the result of inference from the Neyman test can be made valid for the target population, while Fisher's test only identifies the effect and does not draw conclusions about the population of heavy smokers. To do this one would need a qualitative discussion to make appropriate inference to the overall population.

The findings of this paper suggest that the best utilization of these methods is to combine the use of them. For example, one could firstly use the Fisher test to calculate if any effect exists and then use Neyman's test to calculate an estimated average for this effect if one is detected.

References

Dictionary - *random*, https://www.dictionary.com/browse/random

Baron-Cohen.S and Brockman.J. (2013). *Thinking*. Edge Foundation. p.158.

Basu, D. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Rejoinder. Journal of the American Statistical Association, 75(371). p.576.

Clen.S. (2013). *Bias in Statistics: Definition, Selection Bias Survivorship Bias*. Statistics How To. https://www.statisticshowto.datasciencecentral.com/what-is-bias/ [03-04-2019].

Edgington.E.S. (1986). *Randomization Tests for Censored Survival Distributions*. Biometrical Journal. p.531.

Ernst M.D. (2004). *Permutation Methods: A Basis for Exact Inference*. Statistical Science. ://projecteuclid.org/download/pdfview$_1$/*euclid.ss*/1113832732.19 : 4.

Esthermsmth. (2017). *Chaos Theory*. Learning Theories. https://www.learning-theories.com/chaos-theory.html [01-04-2019].

Dransfield B. and Brightwell B. Influential Points. *Permutation tests*. http://influentialpoints.com/Training/permutation_tests.htm [25-04-2019].

Duignan, J. (2016). In: A Dictionary of Business Research Methods. Oxford University Press.

Fay M.P. Follmann D.A. (2002). *Designing Monte Carlo Implementations of Permutation or Bootstrap Hypothesis Tests*. 56:63.

Hills R.K. (2013). *Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis*. National Center of Biotechnology Information.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706448/ [02-04-2019].

Hooijmans.C.R. (2014). *SYRCLE's risk of bias tool for animal studies*. BMC.
https://www.academia.edu/34251950/SYRCLE_s_risk_of_bias_tool_for_animal_studies.

Howell.D.C. (2018). *Overview of Randomization Tests*. University of Vermont. https://www.uvm.edu/~dhowell/StatPages/Randomization%20Tests/RandomizationTestsOverview.html [03-04-2019].

Howell D.C. (2018). *The null hypothesis*. University of Vermont. https://www.uvm.edu/~dhowell/StatPages/Randomization%20Tests/null_hypotheses.html [05-04-2019].

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge

University Press. (Chapter 5, 6 and 12).

Influential Points. http://influentialpoints.com/Training/permutation_tests.htm [17-04-2019].

Kendall J.M. (2003). *Designing a research project: randomised controlled trials and their principles*. Emergency Medicine Journal. 20:2.

Kestenbaum.D. (2004). *The not so random coin toss*. NPR. https://www.npr.org/templates/story/story.php?storyId=1697475&t=1553678587466 [01-04-2019].

Khan Academy. *Random sampling vs. random assignment (scope of inference)*. https://www.khanacademy.org/math/ap-statistics/gathering-data-ap/statistics-experiments/a/scope-of-inference-random-sampling-assignment. [06-04-2019].

Kitchen, C. (2009) Nonparametric vs Parametric Tests of Location in Biomedical Research American Journal of Ophthalmology , Volume 147 , Issue 4 , p.571 - 572.

Laerd, *Independent t-test using Stata*, electronic dataset, https://statistics.laerd.com/stata-tutorials/independent-t-test-using-stata.php [20-03-2019]

Lane D. *Randomization tests*. HyperStat. http://davidmlane.com/hyperstat/B163479.html [02-04-2019].

Lang J.B. (2015). *A Closer Look at Testing the "No-Treatment-Effect" Hypothesis in a Comparative Experiment*. Statistical Science 30:3. https://arxiv.org/pdf/1509.03108.pdf.

Lunneborg C.E. *Random Assignment of Available Cases: Let the Inferences Fit the Design*. University of Washington. https://www.uvm.edu/~dhowell/StatPages/Randomization%20Tests/referencePapers/LunneborgPapers%20Folder/randomiz.pdf.

Marchand A. (2017). *Why Ice Cream Isn't Deadly: Correlation vs. Causation*. Plebian Science. https://plebeianscience.wordpress.com/2017/06/18/why-ice-cream-isnt-deadly-correlation-vs-causation/ [06-04-2019].

Martin, E. and McFerran, T. (2017). In: A Dictionary of Nursing, 7th ed. Oxford University Press, p.346.

McGregor.A. (2015) *Why medicine often has dangerous side effects for women*. Ted. https://www.ted.com/talks/alyson_mcgregor_why_medicine_often_has_dangerous_side_effects_for_women?language=en.

Mendenhall W., Scheaffer R.L and Wackerly D.D. (2011). *Mathematical statistics with applications 7th edition*. Cengage Learning. p.445

Metha C.R. Patel N.R. *Exact tests*. University of Sussex. http://www.sussex.ac.uk/its/pdfs/PASW_Exact_Tests.pdf.

Mewhort.D.J.K, Johns.B.T and Kelly.M. (2010) *Applying the permutation test to factorial designs*. The SCIP Conference. https://link.springer.

com/content/pdf/10.3758%2FBRM.42.2.366.pdf.

Michelangeli.M. Wong B.B.M. Chapple D.G. (2015) *It's a trap: sampling bias due to animal personality is not always inevitable*. Behavioural Ecology. https://academic.oup.com/beheco/article/27/1/62/1742406. 27:62-67.

Pearl J. (2010). *An Introduction to Causal Inference*. Int J Biostat (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836213/.

Rosenbaum P.R. (2002). *Observational Studies*. New York: Springer, 2nd Edition.

Schwartz S., Gatto N.M. and Campbell U.B. (2012). *Extending the sufficient component cause model to describe the Stable Unit Treatment alue Assumption (SUTVA)*. Epidemiol Perspect Innov. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3351730/#B12. 9:3.

Venmans F. (2016). Potential outcome and randomized experiments. [Powerpoint presentation]. University of Mons. http://homepages.ulb.ac.be/~frycx/Slides%20Venmans%203.pdf [2019-05-27].

Appendices

Abbreviations

- **SUTVA** - Stable Unit Treatment Value Assumption
- **FRT** - Fisher's Randomization Test
- **ATE** - Average Treatment Effect
- **SATE** - Sample Average Treatment Effect
- **PATE** - Population Average Treatment Effect

Additional subjects

- **Randomness**

  The dictionary definition of the word *random* is in a general sense described as something "*...occurring without definite aim, reason or pattern*". This statement seems reasonable as a broader denotation of the word, but the concept of randomness is not completely unambiguous. The British dictionary refers to randomness as something which "*...cannot be determined but only described probabilistically*". This suggest that randomization is a process which outcome is not dependent on a deterministic pattern, which implies that if a generation of numbers cannot be predicted or influenced in any way then it is random. This leaves the question whether anything is truly random put in the light of philosophy suggesting that the world is entirely built up by determinism.

  The assumed randomness of a coin flip for example can with the right measurements be mathematically proven to be biased in some sense (Kestenbaum, 2004). Though one can still argue that the current calculations lack ground to stand on, it is still most likely inevitable that in some distant future we will be able to calculate it properly which strips the randomness of its unbiasedness and unpredictability. The human bias could be compromised if one had the mathematical foundation to build a robot which would toss the coin in a manner which would guarantee absolute randomness.

  Even though many findings indicate that coin flips and other traditional stochastic processes (dice throwing is another classical textbook example) are governed by deterministic outcomes, other theories such as the Chaos Theory suggests that this does not strip the process of its randomness since the sequence will still be unpredictable regardless of the process' deterministic nature (Esthermsmth, 2017). Consider the endless number of crossroads a deterministic process evolves from and all

the alternative paths which could have been chosen during the procedure before the end result unfolded (i.e. the butterfly effect). This suggests that the final outcome is one of an infinite number of possibilities, which makes the result unpredictable since the number of possible outcomes are too great to be intelligible. In the future this theory could be more susceptible to fallacy if computers strong and fast enough to compute all possible outcomes of a condition would be invented and the concept of infinity would have to be reconstructed.

This paper will not make any assumptions regarding the entropy of randomness. Randomness will only be dealt with in relation to the unconfoundedness assumption, which states that randomized selection or assignment of subjects is free from the dependence of potential outcomes (Imbens and Rubin, 2015).

- **Faults in the selection process**

  As highlighted throughout this essay, the supposed randomness of the selection process is in many instances fallacious. To understand this properly a few examples will follow to illustrate this occurrence through empirical context.

  Firstly, in research involving laboratory animals the choice of animals are rarely drawn at random from the main population. This is partly due to incentives urging researchers to pick subjects which are easily accessible, such as local livestock, or subject which pose little economical constraint. In addition, unwanted trapping bias may arise when one collects wild animals. This could be the bias which follows from trapping a certain species with some sort of luring technique (like baited traps), a technique which would attract a certain behavioural type, such as especially bold and risk taking animals (Michelangeli, Wong and Chapple, 2015).

  Even when randomization of animals has been ensured in the selection process, other factors emerging during the process may affect the results. Small things such as placing of cages may affect some pharmacological agents that are to be tested, since temperatures may vary depending on if the cages are placed on a top shelf or the floor (Hooijmans, 2014). This could mean that a subject which once accurately represented a certain type of the population may now obtain new characteristics, as an effect of the warmer environment, which would skew the distribution in different directions. In addition, the strength of the treatment would not be constant across units, which would violate the SUTVA.

  Faults in the selection process can also be found in human studies

where scientist let patient volunteers constitute the whole sample size, which obviously gives rise to some considerable bias. This is problematic if one wishes to have the findings apply to a broader population outside of patient volunteers. This means that the inference calculated using these subjects become restricted only to the population to which it applies, meaning that the results generated cannot be generalized (which most likely defeats the initial purpose of the experiment)(Hills, 2013).

Generalizing research results to wider populations than what is considered statistically appropriate is relatively common in the medical industry. Many studies entirely rely on male participants when testing the effect of new drugs on humans. This can be problematic since males and females may react differently to different preparations. Further implications may be that dosage recommendations are then set by male standards for females (McGregor, 2015).