Vrije Universiteit Amsterdam

Faculty of Sciences
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands

Research Paper Business Analytics

Creating a benchmark for the tray optimization problem

Author:
Sven van der Kooij

Supervisor:
Dr. Kristiaan Glorie

September 28, 2015

Preface

This research paper has been written for the course Research Paper Business Analytics as part of the Business Analytics Master program at the Vrije Universiteit Amsterdam. The goal of this course is to test the individual research capabilities of the student, as well as the capability to give a clear and detailed description in both a written report and an oral presentation.
I would like to thank my supervisor Kristiaan Glorie for introducing me to this subject and for providing me with insights and support.

Sven van der Kooij
Amsterdam, September 2015

Abstract

Previous research on the Tray Optimization Problem(TOP) focused on the creation and improvement of algorithms. Researchers had to work with limited data, or they had to create data themselves. Solutions were also evaluated solely on cost. The goal of this research is to provide an environment where TOP-instances can be generated and where solutions can be properly evaluated. In order to create new instances, characteristics of provided data sets were found by means of graphical representations and statistical test. Four different methods were used to evaluate solutions on costs and failure rates, (perturbed) historical sampling and (perturbed) historical frequencies. These methods were applied to a solution that was in use while the data was gathered, the starting solution of a genetic algorithm, and the solution of the genetic algorithm. Previous research only evaluated solutions on the costs they incur. However, this research shows that it is also important to include the failure rate under changing circumstances in this evaluation, as the reduction in cost can come at a steep increase in the rate of failure. This information is essential to make a balanced decision about the practical usefulness of a solution.

Contents

Chapter 1

Introduction

Reducing costs in health care facilities is a more pressing concern now than ever before as health care costs continue to show an incredible yearly increase. In 2012 the Netherlands spent 11.8% of their Gross Domestic Product(GDP) on health care, second only to the United States of America within the organization for economic co-operation and development [OECD, 2014]. Research by van der Horst et al. [2011] has shown that these costs are expected to rise to anywhere from 19 to 31% of the GDP by 2040. These trends have shown the need for reduction of costs within health care facilities. One of the areas that are being looked into is the management of the sterile inventory within hospitals.

The sterile inventory within hospitals consists of an inventory for disposable sterile instruments and a separate inventory for reusable sterile instruments. The focus of this paper lies on the management of the inventory of reusable sterile instruments. The reusable sterile instruments are part of a return cycle within the hospital. They flow from the sterile inventory to the Operating Theatre(OT), from which they go to the Centralized Sterilization Department(CSD) and finally they return back to the sterile inventory. This cycle takes about 8-12 hours and instruments can therefore only be used once a day, but all instruments will always be available at the start of the day.
However, the instruments are not stored separately, but they are grouped in trays. Within a hospital there are different tray types, these types have a fixed composition and there can be more than one tray of each type. Each surgery requires a fixed combination of tray types. Determining the composition of tray types, the quantity to keep in stock of each tray type, and the allocation of tray types to surgeries is known as the Tray Optimization Problem(TOP). An alternative name for this problem is the Net Optimization Problem(NOP).

The first research into this problem was done by Fineman and Kapadia [1978]. The first notable research after this comes from Reymondon et al. [2008] and van de Klundert et al. [2008]. After which research into the subject picked back up with for example Florijn [2008], Glorie [2008], and more recently Kamphorst [2012], and Glorie and Dollevoet [2013]. These researches all focused on creating solution algorithms for the TOP. However, most of the researches had

problems with a lack of data. Some had to create their own data set in order to be able to test their solutions, while others did not have data to evaluate the solution given by the algorithm. Some researchers did a sensitivity analysis into the input parameters, for example costs and the maximum tray size. However, none of the researches evaluated what would happen if circumstances changed after their solution was implemented.

Therefore this research seeks to create a program where TOP-instances can be generated based on real data and where solutions can be evaluated under changing circumstances. The generated instances have to be of generic sizes, so the user will be able to choose the number of instruments and surgeries. In this paper the following four methods of evaluation will be used:

- Historical sampling
- Perturbed historical sampling
- Historical frequencies
- Perturbed historical frequencies

The structure of the paper will be as follows. First the theory behind the statistical tests used in this paper will be discussed in chapter 2. After which insight will be provided into the available data sets and characteristics will be discussed in chapter 3. New instances with most of these characteristics will be created in chapter 4. In chapter 5 a currently used net composition will be evaluated alongside two other solutions. And finally in chapter 6 some closing remarks and advice for future research will be given.

Chapter 2

Methods

Statistical tests were used in order to test for underlying distributions, or significance in differences. This chapter describes the statistical tests used throughout the paper. Section 2.1 describes the Pearson's chi-squared test statistic, used in section 3.2.2 to determine whether observed data comes from a Poisson distribution. Section 2.2 focuses on the Shapiro-Wilk test for normality, this test is used in chapter 5 to determine if the Welch's t-test, described in section 2.3, can be used. If normality is not met, the Wilcoxon–Mann–Whitney test, described in section 2.4, will be used.

2.1 Pearson's chi-squared test

The Pearson's chi-squared test was first introduced by Pearson [1900] and is a statistical test that can be used to determine whether a set of observations significantly differs from a theoretical distribution. The idea behind this test is to divide the observations into bins and to compare the number of occurrences per bin to the expected number of occurrences under the theoretical distribution. In this paper we will be using the null-hypothesis, H_0 : the data comes from a Poisson distribution. In order to test this hypothesis, the following test statistic has to be computed:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

Where:

- χ^2 = the Pearson's cumulative test statistic, which approaches a χ^2-distribution with $n - 2$ degrees of freedom under H_0
- n = the number of bins
- O_i = the number of observations in bin i

- E_i = the number of expected observations in bin i under H_0

The corresponding p-value can then be calculated by comparing the value of the test statistic with the χ^2-distribution with $n-2$ degrees of freedom. If the p-value is below the significance-level of 5%, the null-hypothesis is rejected.

2.2 Shapiro-Wilk test

The Shpairo-Wilk test was first introduced by Shapiro and Wilk [1965] and is a statistical test used to test whether a set of observations belongs to a normal distribution. There are several different tests for normality, however, research by Razali and Wah on the most popular normality tests has shown, that the Shapiro-Wilk test has the best power for a given significance. This test will be used in chapter 5 in order to test the normality assumption made by the Welch's t-test. The null-hypothesis is, that the data belongs to a normal distribution. In order to test this hypothesis the following test statistic has to be computed:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

Where:

- n = the number of data points
- a_i = the Shapiro-Wilk coefficients
- $x_{(i)}$ = the i-th smallest data point
- x_i = the i-th data point
- $\bar{x}$ = the average over all data points

The p-value is then calculated based on the value of the test statistic. If this p-value is below the significance-level of 5%, the null-hypothesis is rejected.

2.3 Welch's t-test

The Welch's t-test was first introduced by Welch [1947] and is a statistical test used to test whether two sets of observations have equal means. The advantage of Welch's t-test over the Student's t-test is, that the Welch's is more accurate when the two samples have different variances and sizes. An assumption made by this test is, that both data sets come from a normal distribution. Therefore, in order to use this test statistic, the Shapiro-Wilk test will be used to test for normality. This test will be used in chapter 5 in order to test whether the cost of one solutions is significantly lower than the other. The null-hypothesis is that the two sample means are the same. To test this hypothesis, the following test statistic has to be computed:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{2.3}$$

Where:

- $\bar{X}_i$ = the i-th sample mean
- s_i^2 = the i-th sample variance
- N_i = the i-th sample size

Under the null-hypothesis the value for this test statistic will come from a t-distribution with a number of degrees of freedom approximated by:

$$\nu \approx \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}} \tag{2.4}$$

Since we want to know if one of the solutions is significantly cheaper than the other, we use a one-tailed test to test the null-hypothesis. If the p-value is below the significance-level of 5%, or if the p-value is above 95% depending on what we choose for sample 1 and 2, the null-hypothesis is rejected and there is a significant difference between the two sample means.

2.4 Wilcoxon–Mann–Whitney test

The Wilcoxon-Mann-Whitney test was first introduced by Mann and Whitney [1947] and is a similar statistical test to the Welch's t-test. The difference is, that this test does not make the normality assumption made by the Welch's t-test. Therefore it will be applied in chapter 5 in situations where normality has been rejected. The null-hypothesis is that the two samples have the same location. In order to test this hypothesis we have to follow the following steps:

1. Combine both samples and assign a rank to each of the observations. If there is an n-way tie, all unadjusted ranks will be added up and each observation will be given a value equal to this sum divided by n.
2. Add up the ranks for both samples separately
3. Calculate the test statistic for the sample with the smallest sum of the ranks according to:

$$U = R_i - \frac{n_i(n_i + 1)}{2} \tag{2.5}$$

Where:

- R_i = the smallest sum of ranks

- n_i = the sample size belonging to the data set with the smallest sum of ranks

The p-value is then determined based on the value of this test statistic. Depending on the alternative hypothesis, we then decide whether the difference is significant and the null-hypothesis can be rejected. The alternative hypothesis that will be used are if the location of sample 1 is bigger than sample 2 and vice versa.

Chapter 3

Data

The goal of this chapter is to provide inside into the data sets provided and to show why certain decisions were made in the benchmark. Section 2.1 gives a description of the contents of the data sets and the decisions that were made because of them. And section 2.2 will focus on the analysis of the data and the factors that have to be taken into account when creating new data sets.

3.1 Data Description

In order to create realistic instances for the benchmarking tool, real hospital data was used. This data was provided by three major Dutch hospitals and therefore consists of three separate data sets. The hospitals will be referred to as hospital 1(H1), hospital 2(H2), and hospital 3(H3) for anonymity purposes.

Table 3.1 shows the contents of the data sets for each hospital. Here we can see that H2 has the most complete data set, followed by H3 and finally H1. The surgery planning contains the surgeries performed per day for a period of #Days days. The surgery frequencies are the overall number of times a specific surgery was performed during that period of time. Instrument demand contains the instruments required per surgery. The current net composition, which is only available for H2, describes the composition of the current nets and the quantities in which they are available.
In order to create instances for the static simulation, a surgery planning is required. However, this planning is not available for H1 and therefore the decision was made not to include H1 in the static simulation.
Data missing from the data sets was data on the weight and volume of the instruments, and data on the duration of the surgeries. This made it impossible to take the total weight and volume of a tray into account. It was furthermore decided that the length of a surgery day would be expressed in the number of surgeries performed and not the amount of time elapsed.
From the data sets for H2 and H3 data was missing on the sterilizations costs

of the nets and instruments. The decision was made to copy the costs that were given for H1. This sets the instrument sterilization costs to 1 and the net sterilisation and handling costs to 20.

	H1	H2	H3
Surgery planning	-	yes	yes
Surgery frequencies	yes	yes	yes
Instrument demand	yes	yes	yes
Current net composition	-	yes	-
#Surgeries	15	16	174
#Instruments	195	87	1125
#Days	365	365	337
Annual net depreciation	500	475	500
Max #instruments per net	65	60	65

Table 3.1: Characteristics of the three data sets

3.2 Data Analysis

The goal of the data analysis is to find useful characteristics in the data and to find underlying distributions that can be used to create instances of general sizes. In order to achieve this, we must first understand what is required for an instance. An instance consists of a surgery planning, an instrument demand, the maximum number of instruments per net and several cost parameters. The values for the last two are inherited from the hospital the new instance is based on. An instrument demand contains the instruments required per surgery. The data needed to create an instrument demand is further analysed in section 3.2.1. Lastly the surgery planning is a plan which describes the number of times each surgery is performed on a specific day for a predetermined amount of days. Section 3.2.2 will focus on the data analysis concerning the surgery planning.

3.2.1 Instrument demand

The difficulty in creating a new instrument demand based on another instance, is the fact that not only the number of instruments may change, but also the number of surgeries and the ratio between the two might not stay the same. The creation of an instrument demand can therefore be subdivided into the following four problems:

1. The number of instrument-types a surgery requires
2. The number of surgery-types an instrument is assigned to
3. The specific assignment of instruments to surgeries
4. The amount of an instrument-type that is required by a surgery-type

In order to solve problem one and two, an assumption has to be made. Because when the ratio between the total number of surgeries and instruments changes, either the number of surgeries per instrument, or the number of instruments per surgeries has to change. Therefore the assumption was made that the number of different instruments per surgery does not depend on the size of the instance. So in other words, increasing or decreasing the surgery to instrument ratio does not change the number of different instrument-types required by a surgery. This means that the number of surgery-types an instrument is assigned to depends on the ratio between instruments and surgeries.

In order to find a solution for problem three, we have to look at the data to see if there are any patterns that have to be preserved. First the number of different surgeries an instrument is assigned to will be analysed. Figure 3.1 shows the number of different surgery-types an instrument is assigned to for H2. We can see that there are three major peaks in the plot, one around 2 surgeries, one at 7 and one at 15. If all instruments had an equal probability of being assigned to a surgery, we would expect to see a histogram corresponding to a binomial distribution. However the data suggests that there is a distinction in instrument-types, some instruments are used very rarely, some are used more frequently and others are used for almost every single surgery. This shows that assigning instruments randomly to surgeries does not fit the data.

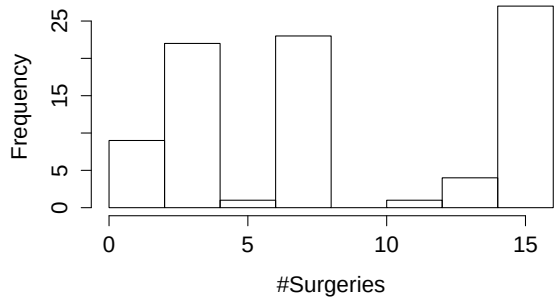


Figure 3.1: Histogram of the number of surgeries per instrument for H2

Next we will see if the division of instruments over surgeries is similar for H3. Figure 3.2 shows again the number of surgery-types an instrument is assigned to and the number of times this happens. The figure is cropped in order to be able to show the frequently assigned instruments, because the number of instruments that are assigned to a low number of surgeries is so high, that this would not have been possible otherwise. The histogram shows that there can again be made a distinction into three groups of instrument-types based on peaks seen

in the graph, however they are less pronounced than the peaks found in Figure 3.1. But the graph again shows that the spread of instruments over surgeries does not occur equally, nor with equal probability.

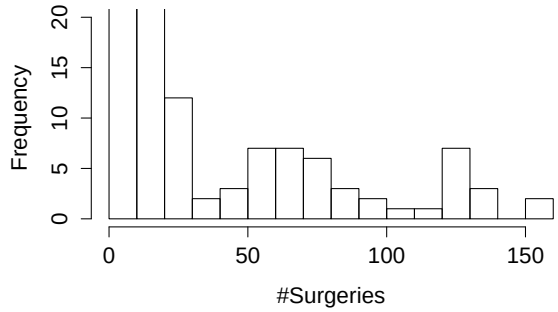


Figure 3.2: Histogram of the number of surgeries per instrument for H3

Now that we have seen that the spread of instruments over surgeries is not equal among instruments, we have to look at how the surgeries are spread over the instruments. In order to do this, we will divide the instruments in three groups based on the peaks seen in Figure 3.1 and 3.2, the division boundaries are specified in 3.2. Group 1 contains the rarely used instruments, group 2 the more frequently used instruments and group 3 contains the most frequently used instruments. What we want to test now, is whether every surgery requires an equal amount of instruments from each group. Or in other words, we want to know if the instruments from a particular group are used by all surgeries, or if they are mostly used by a small group of surgeries.

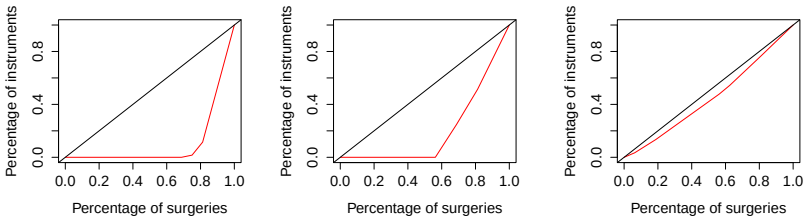
	Group 1	Group 2	Group 3
H2	0-5	6-8	8-16
H3	0-40	40-100	101-152

Table 3.2: Boundaries for the different groups of instruments

We will start by looking at H2. For each of the groups the relationship between the cumulative percentage of surgeries and the cumulative percentage of instruments is shown in Figure 3.3. These figures are similar to the Lorenz-curve of income distribution in economics. The diagonal shows the line of equal distribution, if the plot follows this line closely, the distribution is close to even. Figure 3.3a shows that only a little over 20 percent of the surgeries use the rare instruments. So this means that some surgeries use a lot of rare instruments,

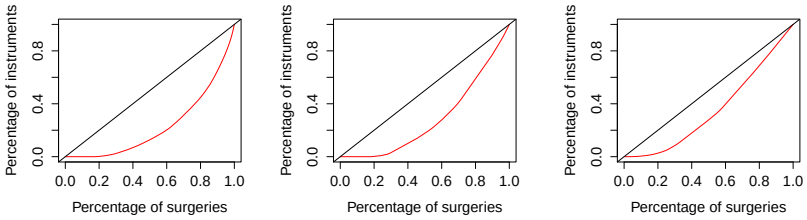
while the majority does not use them at all. If we look at Figure 3.3b, we see that the distribution is more even, but still 50 percent of the surgeries do not use these instruments. However if we look at Figure 3.3c, we can see that the distribution of frequently used instruments is close to being even. Generally it can be concluded that the distribution of instruments within groups is not even and a distinction has to be made between surgeries that use a lot of instruments from a particular group and surgeries that don't.

Now if we look at the same plots for H3 in Figure 3.4, we can see a similar pattern that is less pronounced. However it is still there and has to be taken into account when instruments are assigned to surgeries.



(a) Group 1 (b) Group 2 (c) Group 3

Figure 3.3: Equality graphs for the instrument groups for H2

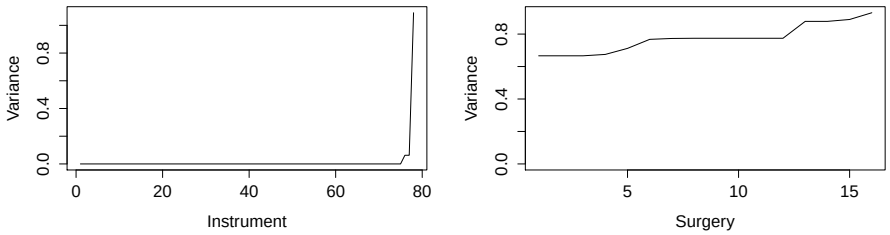


(a) Group 1 (b) Group 2 (c) Group 3

Figure 3.4: Equality graphs for the instrument groups for H3

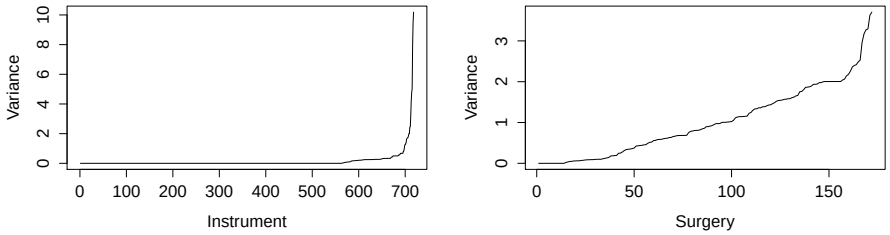
Now that the allocation of a certain instrument-type to a surgery-type has been addressed, the amount has to be determined. There are three main possibilities. One, the amount depends on the instrument type. Two, the amount depends on the surgery type. Or three, the amount follows a pattern, like a probability distribution.
Figure 3.5a shows the sorted variance of the non-zero amounts of instruments per instrument-type for instruments with more than one allocation. We can

see that the variance is zero for almost every instrument, which means that the amount heavily depends on the instrument-type. If we look into Figure 3.5b, we see the variance of non-zero amounts of instruments per surgery-type. The higher variance in this figure combined with the almost all zero variances for the instrument-types, means that the amount of instruments per allocation for H2 depends on the instrument-type and not the Surgery-type.
Looking at H3 in Figure 3.6, we can see a similar pattern for the instruments. And even though the variance for the surgery-types is different from H2, it is still mostly non-zero. This leads to the same conclusion, the amount of instruments per allocation depends on the instrument-type.



(a) (b)

Figure 3.5: The sorted instrument variance per instrument and surgery for H2



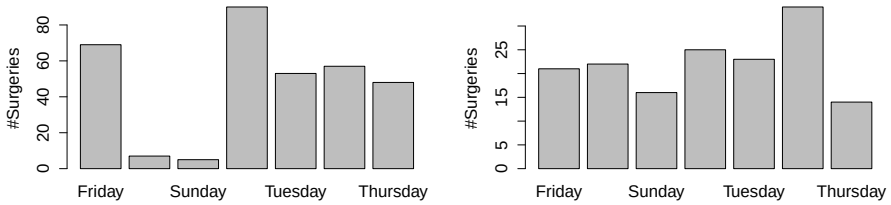
(a) (b)

Figure 3.6: The sorted instrument variance per instrument and surgery for H3

3.2.2 Surgery plan

In order to build a new surgery plan based on another instance, the data has to be studied. The first thing that was done, was looking for patterns. The first pattern that was analysed, was the pattern between different days of the

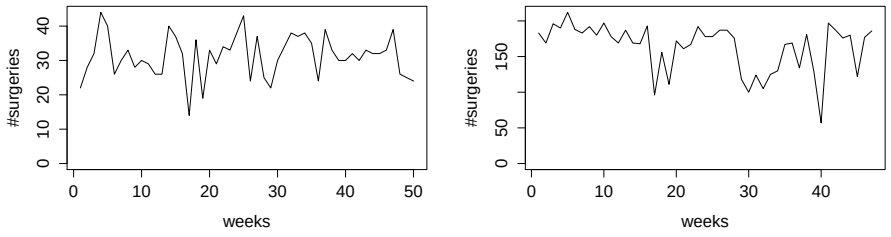
week. Figure 3.7 shows the total number of surgeries performed per day for two different types of surgery. The first surgery-type can be classified as an elective surgery, because the surgery is mostly performed on weekdays. The second surgery-type can be classified as an emergency surgery, as the number remains almost constant throughout the week. This pattern is so much different per day of the week for elective surgeries, that this has to be taken into account when making the surgery plan.



(a) Elective (b) Emergency

Figure 3.7: Examples of week-patterns for different types of surgery

The next pattern that was studied, was the year pattern. Figure 3.8a shows the year pattern for H2. The pattern appears to be random. If we look at Figure 3.8b, we see the year pattern for H3. This pattern appears to be more pronounce, with a valley just before week 20 and again around week 30 and 40. These correspond to the spring holiday, summer holiday and autumn holiday in the Netherlands. In the end it was decided that the pattern was not pronounced enough and would not be taken into consideration when creating the data set.

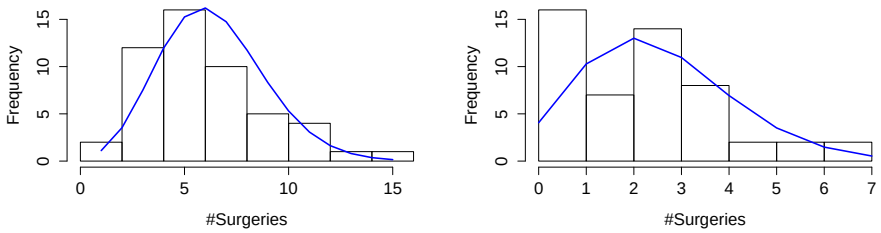


(a) H2 (b) H3

Figure 3.8: The number of surgeries per week for H2 and H3

Now that the patterns have been studied, the number of surgeries that are

performed per day of each surgery-type have to be analysed. Surgeries in an operating theater can be seen as arrivals, especially when it concerns emergency surgeries. A large number of people all have a very small probability of requiring a certain type of surgery. Therefore we expect a number of surgeries to follow a Poisson distribution, while others do not because of planning. In order to determine if a surgery-type can be modeled as a Poisson process, the week patterns have to be removed. In order to do this, we will replace the daily demand in the surgery planning with the weekly demand per surgery. This will not influence instances that can be modeled as a Poisson process as the sum of Poisson processes is again Poisson distributed with the sum of the rates.
Figure 3.9a shows the histogram of the weekly intensities of a surgery-type with corresponding Poisson distribution in the background in blue. The Poisson distribution closely matches the histogram and therefore this instance can be classified as a Poisson instance. Figure 3.9b however, shows a histogram that is much further away from the corresponding Poisson distribution with a peak at lower values that should not be there if it was a Poisson instance. This instance can therefore not be classified as a Poisson instance.



(a) Poisson (b) Non-Poisson

Figure 3.9: Histograms of the arrivals per week against the Poisson distribution

In order to classify these instances in a non subjective way, a statistical test was chosen make this decision. The test used was the Pearson's chi-squared test, described in section 2.1. The null-hypothesis is that the data is from a Poisson distribution and only when the null-hypothesis is rejected will we classify an instance as non Poisson. The significance level used to make this decision was five percent. The results are shown in Table 3.3 and show that the majority of cases can be classified as Poisson instances.

	Poisson	Non Poisson
H2	10	6
H3	105	69

Table 3.3: Classification of Poisson instances for H2 and H3

Chapter 4

Instance Generation

This chapter will show how the patterns that were found in chapter 3 were implemented when creating an instance for the TOP. An instance for the TOP consists of an instrument demand, a surgery planning, cost parameters and maximum tray sizes. A new instance can either be based on H2 or H3, as they are the data sets with all these parameters. The cost parameters and maximum tray sizes are directly inherited from the hospital the new instance is based on. In order to create new instances of arbitrary sizes, some assumptions have to be made. The first assumption is needed to create a new surgery demand. We assume that the number of different instrument-types required per surgery does not change when the ratio between instrument-types and surgery-types changes. This means that the number of surgeries an instrument is assigned to does change in these situations. The next assumption is that the number of performed surgeries changes linearly with the number of surgery-types. This assumption concerns the surgery planning.
Section 4.1 will show how a new instrument demand is created based on an existing demand and section 4.2 will do the same for the surgery planning.

4.1 Instrument Demand

The instrument demand is a matrix with the instruments as rows and surgeries as columns. It shows how many of each instrument is needed per surgery. Since this instrument demand has to be based on an already existing instance, we would like to preserve information that was in the original data set. First we will look at the instruments and after that we will talk about the surgeries.

4.1.1 Instruments

The idea behind creating new instruments is that each instrument is based on a random instrument from the original instance. The problem is that we have to

decide which characteristics have to be inherited from the original instrument in order to create the new instrument and eventually instrument demand. Figures 3.1 and 3.2 showed that instruments could be subdivided into three separate groups, with a different rate of being assigned for each group. Not only do the groups have different rates of being assigned, even within groups there are differences. To preserve these differences, the new instrument will inherit the frequency of the instrument it is based on. This frequency will determine the group it came from according to Table 3.2 and can be used to determine the rate at which an instrument should be assigned within groups.
Now that we know to which group each instrument belongs and how frequent we should assign each instrument within the groups, we would like to know how many of each instrument to assign when an assignment is made. In other words, which number should be in the cell of the matrix. This could depend on the instrument-type, the surgery-type, or on a combination of the two. However, we have already seen in Figures 3.5 and 3.6 that that number only depends on the instrument-type and not the surgery-type. So in order to create an instrument demand, the new instrument also has to inherit the amount the old instrument was assigned.
So, in order to create a new instrument demand, a new instrument will inherit the frequency of being assigned and the amount being assigned of a random instrument from the old instrument demand.

4.1.2 Surgeries

Just like the instruments, new surgeries are based on random surgeries from the original instance. The last thing left to determine before we can create the new instrument demand is how to assign instruments to surgeries. As we've seen in 3.3 and 3.4, each surgery requires a different composition of instruments. Some require a lot of specialized tools, while others only require basic tools. Therefore the number of instrument-types required per instrument group is a characteristic of the surgery-type. So when creating a new surgery a random surgery from the original instance is chosen and for this instance the number of instrument-types used from group 1, 2 and 3 are determined and passed on to the new instance.

4.1.3 Creation

Now that we have the number of instrument-types per group required by a surgery, the relative frequency an instrument should be assigned and the assignment amount, we can start to create the new instrument demand.
Let I be the group of all instrument-types. Let J be the group of all surgery-types. Let X_g be the collection of all instruments belonging to group g. Let Y_{gj} be the number of instrument-types from group g required by surgery j. Let Z be the new instrument demand, with Z_{ij} the number of instruments i needed for surgery j. Let f_i be the frequency assigned to instrument i. And let n_i be

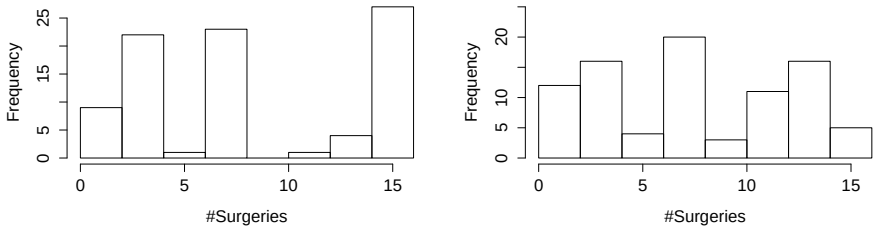
the amount instrument i should be assigned when it is matched to a surgery. The instrument demand can then be created using Algorithm 1.

set $Z \longleftarrow 0$
for *each* $j \in J$ **do**
  for *each* $g \in G$ **do**
    set $S \longleftarrow X_g$
    for $i = 1...Y_{gj}$ **do**
      compute $p_i = f_i/(\sum_{k \in S} f_k)$
      draw a random instrument l from S, where every instrument a has a probability p_a of being picked
      set $Z_{lj} \longleftarrow n_i$
      set $S \longleftarrow S \backslash \{l\}$
    end
  end
end

Algorithm 1: Creating a new instrument demand

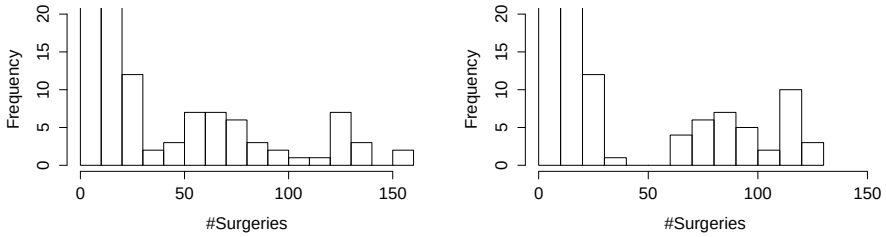
4.1.4 Results

Now that a method for creating an instrument demand has been created, we would like to know how it compares to the original instance. In order to do this, a new instrument demand was created for both H2 and H3 of the same size. Figures 4.1 and 4.2 show a comparison between the original and generated instances for H2 and H3 respectively. For H2 we see that the distinct peaks in the original instance are less pronounced in the generated instance. The cause of this is Algorithm 1 used to create a new instrument demand. It uses a probability p_i for instrument i of being assigned to a surgery. This causes randomness in the number of assignments within groups and therefore less pronounced peaks. If we now look at the same figures for H3 in Figure 4.2, we see the opposite occurring. The peaks for group 2 and 3 appear to be more pronounced in the generated instance. So in the original instance, there was more spread in the assignment of instruments to surgeries within groups than we were able to produce with Algorithm 1.
So this characteristic of the original data set is not kept in its entirety in the newly generated instance and could use improvement.



(a) Original (b) Generated

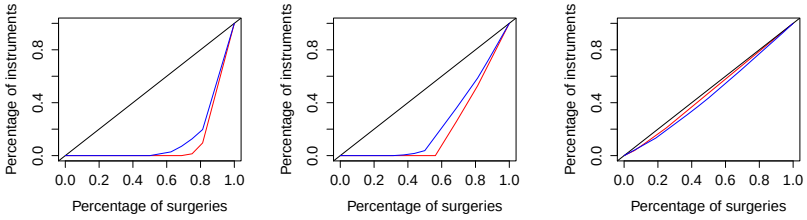
Figure 4.1: Histogram of the number of surgeries per instrument for H2 and a generated instance based on H2



(a) Original (b) Generated

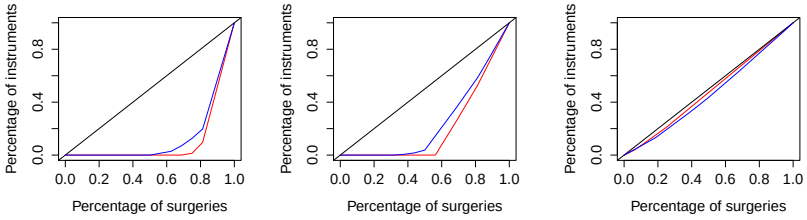
Figure 4.2: Histogram of the number of surgeries per instrument for H3 and a generated instance based on H3

The next thing we have to check, is the distribution of instrument-types within groups amongst surgery-types. Figures 4.3 and 4.4 show this distribution for respectively H2 and H3. The red line is the original instance and the blue line the generated. We can see that for both hospitals the lines are very close and the differences are caused by the randomness in the creation of new instruments and surgeries.
So in the new instance this characteristic is kept.



(a) Group 1 (b) Group 2 (c) Group 3

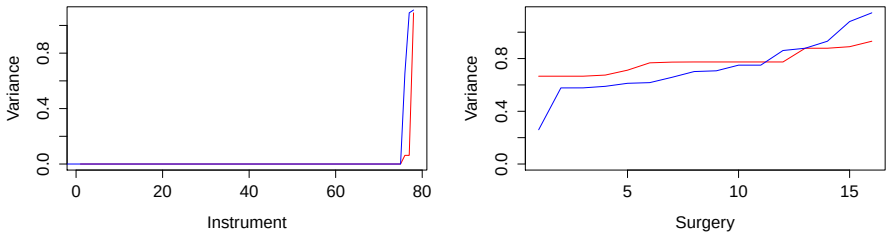
Figure 4.3: Equality graphs for the instrument groups for H2(red) and a generated instance(blue)



(a) Group 1 (b) Group 2 (c) Group 3

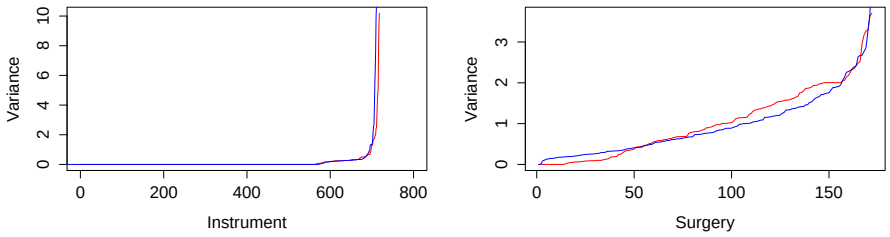
Figure 4.4: Equality graphs for the instrument groups for H3(red) and a generated instance(blue)

The last thing we have to check for the instrument demand is the variance of the amount of instruments assigned per instrument and surgery. Figures 4.5 and 4.6 show these variances for respectively H2 and H3. The original variance is shown in red and the variance belonging to the generated instance is in blue. Both graphs for both hospitals seem relatively close together with some expected variation due to the randomness of the assignments.
So we can see that this characteristic is carried on to the new instance.



(a) (b)

Figure 4.5: The sorted instrument variance per instrument and surgery for H2(red) and the generated instance(blue)



(a) (b)

Figure 4.6: The sorted instrument variance per instrument and surgery for H3(red) and the generated instance(blue)

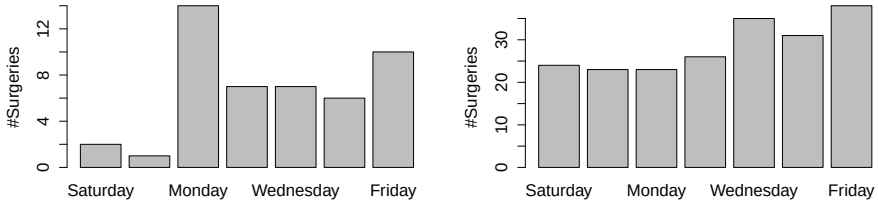
4.2 Surgery Planning

The surgery planning is a matrix with the surgeries as rows and the days as columns. The value in a cell contains the number of times that particular surgery is performed on that specific day. In order to create a new surgery planning based on an existing one, we want to preserve the characteristics of the original plan. Section 3.2.2 showed that there are two different characteristics we want to maintain. The first characteristic is the week pattern and the second is whether a surgery can be modeled as a Poisson distribution.
For instances that can be classified as Poisson instances, a parameter λ has to be determined. In order to preserve the week pattern in the new surgery plan, seven different λ's will be computed, one for every day of the week. As estimator for these parameters, the Maximum Likelihood Estimator(MLE) for the Poisson distribution will be used, which is equal to the sample mean.
For surgeries that can not be classified as Poisson instances, the historical data

is used. In order to also preserve the week pattern in this case, the value in the new surgery plan is equal to a random value in the past on the same weekday.

4.2.1 Results

In order to check whether the characteristics from the original data set are kept in the new instance, we check to see if we can find similar week patterns to the ones we saw before. Figure 4.7 shows an example for both an elective and an emergency surgery. These figures are very similar to the figures of Figure 3.7 and we can therefore conclude that the week patterns can still be found in the generated instances.



(a) Elective (b) Emergency

Figure 4.7: Examples of week-patterns for different types of surgery

Chapter 5

Solution Evaluation

A solution to the Tray Optimization Problem(TOP) consists of three parts. The first is a net composition that contains the instruments that go into each tray. The second is a tray allocation which contains the trays required to perform each surgery. The last part is the number of trays that have to be kept in stock for each tray type.
In order to evaluate such solutions of existing or created instances, four different methods will be applied. These methods will be applied to three different solutions for the original data set of H2. The first solution is the current composition of trays used by H2. The second is the solution given by a Genetic Algorithm(GA). The inner workings of this algorithm are treated like a black box in this paper. However, the algorithm requires a starting solution and for this purpose it uses a solution that is different from the original solution. Therefore, in order to test the effectiveness of the algorithm, this solution will also be examined.
Section 5.1 will discuss the results obtained using historical sampling. Section 5.2 will show the results using perturbed historical sampling. Section 5.3 will go into the results of using historical frequencies. And finally section 5.4 will talk about the results of using perturbed historical frequencies. All of the results in this chapter are obtained by taking the average over 1000 periods of five years.

5.1 Historical Sampling

Historical sampling is done by drawing a random day from the surgery plan and evaluating the solution for that day. The idea behind this method is that it is a good and fair way to compare the average cost of a solution, as none of the solutions should have a failed surgery.
Table 5.1 shows the results for historical sampling and there are indeed no failures. In order to test if the differences in costs are significant between the solutions, the Welch's t-test from section 2.3 will be used. However, before this test can be used, the normality assumption of the data has to be checked. In

order to do this the Shapiro-Wilk test from section 2.2 will be used. However, this test is not conclusive for larger data sets and has to be used in conjunction with a QQ-plot.

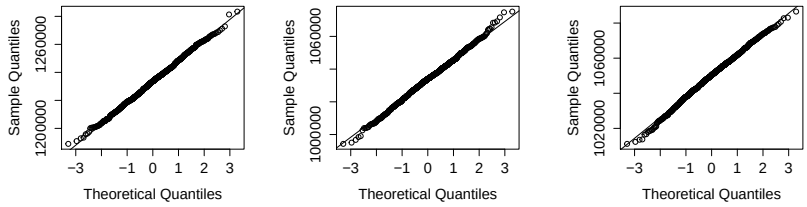
	Failures	Failure Rate(in%)	Costs
Current	0	0	246,735
Start GA	0	0	210,591
Solution GA	0	0	208,033

Table 5.1: Average yearly results for 1000 times historical sampling for 5 years

Table 5.2 shows the value for the Shapiro-Wilk test statistic and the corresponding p-value. None of the p-values are below the threshold of five percent, so the test gives no reason to withdraw the normality assumption. Since the number of data points is quite high, the Shapiro-Wilk test is inclined to classify most distributions as a normal distribution, so in addition to the test QQ-plots are required. Figure 5.1 shows the QQ-plots for the costs of the different solutions and also shows no reason to withdraw the normality assumption.

	W	p-value
Current	0.9989	0.8152
Start GA	0.99855	0.59
Solution GA	0.99792	0.2506

Table 5.2: Values for the Shapiro-Wilk test statistic for the costs



(a) Current (b) Start GA (c) Solution GA

Figure 5.1: Normal QQ-plots for the costs for historical sampling

Since both the test statistic and the QQ-plots do not reject the normality assumption we are allowed to use the Welch's t-test to see if there is a significant difference between the costs.
First we will compare the current solution to the starting solution of the GA. The null-hypothesis is that the means of both data sets are the same. The alternative hypothesis is that the costs of the starting solution of the GA are

less than the costs for the current solution. The value for the test statistic is $t = 301.47$, which, in combination with 1926.5 degrees of freedom, gives a p-value smaller than $2.2 * 10^{-16}$. Since the p-value is smaller than five percent, the null hypothesis is rejected and the costs for the starting solution of the GA are significantly smaller than those of the starting solution.
Next we will compare the starting solution of the GA to the eventual solution of the GA. The null-hypothesis is the same as before and the alternative is that the costs for the eventual solution are less than the costs for the starting solution. The value for the test statistic is $t = 29.591$ with 1997.4 degrees of freedom. This corresponds to a p-value of again $2.2 * 10^{-16}$, which means that the solution of the GA is significantly less than its starting solution.
So to conclude the historical sampling, the current solution is significantly more expensive than both other solutions and the GA produces a solution that is significantly better than its starting solution. However the differences between the current solution and the others is much greater than the difference between the starting solution and the end result.

5.2 Perturbed Historical Sampling

Perturbed historical sampling is similar to the previous method in the way that a random day from the past is drawn. However for perturbed historical sampling a certain percentage of the surgeries on a day are randomized. For this example a percentage of ten percent is chosen. The idea behind this method is to test what happens when small changes occur in the schedule. The results for this method are shown in Table 5.3. From this method onward failures will occur and therefore the costs are no longer a fair comparison. This is because failed surgeries will not incur costs, so in order to rate a solution a manager will have to look at a combination of the failure rate and the costs to determine what is best for the hospital.

	Failures	Failure Rate(in%)	Costs
Current	0.204	0.013	245,613
Start GA	1.377	0.086	209,446
Solution GA	3.115	0.195	206,605

Table 5.3: Average yearly results for 1000 times 10% perturbed historical sampling for 5 years

In order to see if the differences in failures are significant, a statistical test is again performed. First Table 5.4 shows the results for the Shapiro-Wilk test for normality. From this we can conclude that the normality assumption is not correct for this instance and therefore, the Welch's t-test can no longer be used to test for significance. Instead the Wilcoxon-Mann-Whitney(WMW) test from section 2.4 will be used.
First we will test the null-hypothesis that the number of failures in the current

situation and the starting solution are the same. Against the alternative that the current situation has more failures. The test statistic is $W = 17340$ with a p-value less than $2.2*10^{-16}$. So the starting solution for the GA has significantly more failures than the current solution.
Next we will test the null-hypothesis that the number of failures between the starting solution of the GA and then end solution are the same. With an alternative-hypothesis that the number of failures for the solution is bigger. The test statistic is $W = 38125$ with a p-value less than $2.2 * 10^{-16}$. So again the differences are significant.
So to conclude, the current solution is able to handle the small changes much better than the other solutions. And the starting solution has significantly less failures than the end result. However the maximum number of failures is still below 0.2%.

	W	p-value
Current	0.74187	$< 2.2 * 10^{-16}$
Start GA	0.99374	0.0003365
Solution GA	0.98883	$6.661 * 10^{-7}$

Table 5.4: Values for the Shapiro-Wilk test statistic for the costs

5.3 Historical Frequencies

The method historical frequencies creates days according to the historical frequencies of the surgeries. The length of a day is equal to the length of a random day in the past. The idea is that some solutions might depend on day combinations and this method tests what happens when day combinations are broken. In reality this could correspond to increasing the number of days a surgery is planned on. The results are shown in Table 5.5. When we compare these results to the perturbed historical sampling, we see that the number of failures for both GA solutions increased a lot. While the number of failures for the current solution went down.
We again want to know if the difference in failures is significant. Comparing the current solution to the starting solution for the GA gives a p-value less than $< 2.2*10^{-16}$. So the current solution has significantly less failures. The p-value corresponding to the comparison of the starting solution of the GA and the end is the same at $< 2.2 * 10^{-16}$.
Concluding we can see that both GA solutions depended heavily on day combinations and perform a lot worse than before. However, the maximum number of failures still does not exceed one percent of the total surgeries.

	Failures	Failure Rate(in%)	Costs
Current	0.157	0.010	246,668
Start GA	3.852	0.241	209,980
Solution GA	11.734	0.734	206,717

Table 5.5: Average yearly results for 1000 times historical frequencies for 5 years

5.4 Perturbed Historical Frequencies

The last method is the perturbed historical frequencies. This method is again very similar to the historical frequencies, the only difference being that the frequencies being used are perturbed by a fixed percentage. The percentage used is again ten percent. This means that if a surgery had a frequency of 100 times, the perturbed frequency will lay between 90 and 110. The idea behind it is to be able to test for both methods of variation simultaneously.
Table 5.6 shows the results and we can see that the changes between the perturbed historical frequencies and the historical frequencies are minimal. The differences are still significant, as comparing the current solution to the start yields a p-value of $< 2.2 * 10^{-16}$, which is the same p-value as we get from comparing the starting solution of the GA to the final solution.
The differences between the perturbed and non-perturbed historical frequencies were almost non-existent for this example. However, this could be different in other situations.

	Failures	Failure Rate(in%)	Costs
Current	0.166	0.010	246,446
Start GA	3.904	0.244	210,036
Solution GA	11.745	0.734	206,783

Table 5.6: Average yearly results for 1000 times 10% perturbed historical frequencies for 5 years

Chapter 6

Discussion

The goal of this research paper was to create an environment where Tray Optimization Problem(TOP)-instances could be generated based on real data and where solutions to these problems could be evaluated under changing circumstances. In addition an evaluation was to be conducted on a provided Genetic Algorithm(GA).

The creation of new instances is capable of preserving all but one of the characteristics found in the original data set. Figure 4.1 and 4.2 show that the number of surgeries an instrument is used in, is either too spread out, or too clumped up. This is caused by the random, relative to their frequencies, assignment of instruments from groups to surgeries. A possible solution for further research would be to make both the number of surgeries an instrument should be assigned to, as the number of instruments that should be assigned to a surgery deterministic. Right now only the latter is deterministic and the first is probabilistic, which appears to cause these problems.

Up until this point research has focused on algorithms that attempt to minimize the costs for a given instance. The problem is however, that the solution will not be used on that instance, but on next years instance which is not yet known. It is therefore important to determine how well a solution deals with changing conditions. This paper has shown that for a specific genetic algorithm, the reduction in costs comes at a steep increase in the number of failures. So for future research into algorithms it is advisable to also study the compromises made in order to be able to reduce the costs.

Something that could aid further research would be more detailed and bigger data sets. The current system does not work with sizes and weights of instruments and volumes of trays, nor with the length of surgeries, because this data was not available. Also if more data was available, one part of the data set could be used to create a solution, while the other could be used to evaluate the solution.

Bibliography

Stephen J. Fineman and Asha S. Kapadia. An analysis of the logistics of supplying and processing sterilized items in hospitals. *Computers & Operations Research*, 5(1):47 – 54, 1978. ISSN 0305-0548. doi: http://dx.doi.org/10.1016/0305-0548(78)90017-5. URL `http://www.sciencedirect.com/science/article/pii/0305054878900175`.

Elske Paulien Florijn. Optimisation of the distribution of surgical instruments over trays. 2008.

Kristiaan Glorie. Solving the net optimization problem. 2008.

Kristiaan Glorie and Twan Dollevoet. Return cycle inventory optimization with grouped items. 2013.

Bas Kamphorst. Optimalisatie van de samenstelling van steriele medische instrumentennetten. 2012.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18: 50–60, 1947. ISSN 0003-4851.

OECD. Health at a glance: Europe 2014. 2014.

K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900. doi: 10.1080/14786440009463897.

Nornadiah Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33. URL `http://instatmy.org.my/downloads/e-jurnal%202/3.pdf`.

Francis Reymondon, Bertrand Pellet, and Eric Marcon. Optimization of hospital sterilization costs proposing new grouping choices of medical devices into packages. *International Journal of Production Economics*, 112(1):326 – 335, 2008. ISSN 0925-5273. doi: http://dx.doi.org/10.1016/j.ijpe.2006.12.066. URL `http://www.sciencedirect.com/science/article/pii/S0925527307001491`.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. ISSN 00063444. URL http://www.jstor.org/stable/2333709.

Joris van de Klundert, Philippe Muls, and Maarten Schadd. Optimizing sterilization logistics in hospitals. *Health Care Management Science*, 11 (1):23–33, 2008. ISSN 1386-9620. doi: 10.1007/s10729-007-9037-4. URL http://dx.doi.org/10.1007/s10729-007-9037-4.

Albert van der Horst, Frank van Erp, and Jasper de Jong. Trends in de gezondheid en zorg. 2011. URL http://www.cpb.nl/sites/default/files/publicaties/download/cpb-policy-brief-2011-11-trends-gezondheid-en-zorg.pdf.

B. L. Welch. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28. URL http://biomet.oxfordjournals.org/content/34/1-2/28.short.