MSc Program
Engineering Management

Revenue Forecasting Based on Business Opportunities

A Master's Thesis submitted for the degree of
"Master of Science"

supervised by
Univ.Prof. Dr.techn. Dr.h.c.mult. Peter Kopacek

Fabian Hauser

0853898

April 2016, Vienna

TU WIEN CONTINUING EDUCATION CENTER

Affidavit

I, **FABIAN HAUSER**, hereby declare

1. that I am the sole author of the present Master's Thesis, "REVENUE FORECASTING BASED ON BUSINESS OPPORTUNITIES", 73 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted this Master's Thesis as an examination paper in any form in Austria or abroad.

Vienna, 18.04.2016

Signature

Abstract

CRM Opportunity data from a corporation are analysed for patterns. These data is created from salesmen and provide business opportunity information with win probabilities and volumes.

The first step of the analysis is to create crosstables. In the table, observed probabilities and subjective probabilities vary a lot. We also check for probability independencies of opportunities at MDF compared to EFY volumes. As a result, we do not find dependences.

The sum of all open weighed MFY opportunities from FY 2012 – 2015 calculated as a forecast for the EFY total volume is € 168.712.641. The realized volume is however € 53.178.551, that is only 33.8% of all weighted opportunities.

Even if we assume that we have only the prior knowledge, the 30.6% of the FY are successful, the subjective probabilities are actually pointing into the wrong direction and we obtain a much better EFY estimation than using the subjective probabilities.

There is no relationship between the subjective probability judgment and the won opportunities. Finally, a way is found to smooth the crosstable with a logarithm, where again similar probability categories are aggregated to make relationships clearer.

In the end, it is clear to see that small volume opportunities tend to have higher success rates than large volume opportunities. Further a model with a deviation of less than 1% is found and a 10% VaR is calculated. It turns out that the model with three sized categories is the favourable, because 10 size categories seem too many and likely create an overfitting effect within the data.

The main conclusion is that the company has an issue of very overoptimistic salesmen. We propose not to rely on the “experience” of the salesmen but consider the size of an OI as relevant indicator. Small and medium OIs have higher success probabilities than the average, while large OIs have much smaller ones.

Big opportunities are won rather rarely (18%). Maybe the corporation should invest more in their acquirement efforts for big projects.

However, a goal needs to be for this company to train their sales staff to make them more sensitive for their business estimations to get better data input and concluding to more data that are reliable to process.

Index

1 Introduction 1

2 Fundamentals 3

2.1 Probability and Statistics 3

2.1.1 Probability 3

2.1.2 Normal Distribution 4

2.1.3 Covariance..6

2.1.4 Categorical Data: Contingency Table 6

2.1.5 Chi squared distribution 9

2.1.6 Chi squared Test of Independence..10

2.1.7 Kolmogorov–Smirnov Test...12

2.1.8 Multiple Linear Regression 15

2.1.9 Simple Linear Correlation Coefficient 19

2.1.10 Coefficient of Determination...20

2.1.11 Quantile 21

2.1.12 Box Plot 22

2.1.13 Value at Risk (VaR) 22

2.2 CRM – Customer Relationship management..23

2.2.1 Knowledge of Customer Needs 23

2.3 Forecasting 26

3 Data construction 29

4 Data analysis...35

4.1 Revenue Probabilities at MFY..35

4.2 OI: Mid Fiscal year and End Fiscal year 41

4.2.1 Expected FY Volume 41

4.2.2 Distribution of Volume at MFY and EFY 41

4.2.3 Distribution of MFY and EFY Volume by Probability...........................43

4.3 VaR of won volumes at EFY based on subjective probabilities at MFY.....45

4.3.1 Usage of the subjective probabilities 45

4.3.2 Usage of the observed probabilities...47

4.3.3 Usage of the estimated probabilities..48

4.4 The role of Mid Fiscal Year Volume 49

4.4.1 Relation of volume and Won/Lost 49

4.4.2 The relationship between MFY and successful EFY Volume 52

4.4.3 Relation of Volume and the assignment of OIs to a subjective probability category 54

4.4.4 VaR model of expected total Volumes based on MFY Volumes 57

5 Reasons for the deviation of subjective and objective probabilities 60

5.1 Lottery effect 60

5.2 Motivational Biases 62

5.3 Estimating and Forecasting Biases 63

5.3.1 Overconfidence...65

5.3.2 Overoptimism..66

6 Summary 68

7 Outlook 71

8 References 73

1 Introduction

The goal of this thesis is to find a business opportuntiy forecasting model for prediction of future business based on data from a coporate worldwide CRM system. Only this part of the system is considered, which refers to opportunities which are boosting sales revenues. This means customers which are buying every year approximately the same amount, like wholesalers or regular customers, are not pictured.

For the so called run rate business a different forecasting model is used. This thesis focuses on findings obtained by observed probabilities, aggregation of probabilities, Chi squared tests, Kolmogoroff-Smirnoff test, regressions and correlations between opportunity volume and probability, and looking on the distributions to better forecast opportuntity volumes which are building up new customers and new projects.

The most useful forecasting methode would be the tobi II models. Combining regular and potential new customers forecasts would result into the actual revenue forecast. The data used are about one worldwide Business Sector restricted to a small region. A fiscal year (FY) starts always on the 1st of October and ends at the next year's 30th of September. Available data is from the fiscal years (FY) 2010 until 2015, whereupon 2010 and 2011 were the test/introduction phase of the system. Therefore they are excluded due to a lack of data quality and amount of data. For the FY 2012 the CRM system became mandatory. So we decide to take the data beginning with FY 2012 until the end of FY 2015.

The expected method to predict future revenues should result to a modified version of value at risk that is based on Monte Carlo simulations. All simulations are programmed and performed in R.

The thesis beginns with an introduction where the topic is going to be explained followed by the chapter 2 Fundamentals what a CRM system is doing and what forecasting methods are implemented there and how they do work, but it does not come into work as an important variable. The subjective win probabilities of the salesmen, contains serious measurement problems. Further in chapter 3 Data construction it is described how the data, is build up, how they look like and how they were created. Chapter 4 Data analysis uses the previous mentioned methods, observed probabilities, aggregation of probabilities, Chi squared tests …, to analyse the data for a better understanding of the sales people. The next chapter 5 Reasons for the deviation of subjective and objective probabilities uses psychological literature combined with business cases to explain some of observed behaviour. All together is summarized in a final summary chapter. Finally the last chapter Outlook proposes a tobi II model and argues why it makes no sense in using it with the actual data set.

2 Fundamentals

This section gives a summary of the statistical methods used in this thesis. Reference is Doge (2008).

2.1 Probability and Statistics

2.1.1 Probability

We can define the probability of an event either by using the relative frequencies or through an axiomatic approach. In the first approach, we suppose that a random experiment is repeated many times in the same conditions. For each event A defined in the sample space Ω, we define $\mathrm{n_A}$ as the number of times that event A occurred during the first n repetitions of the experiment. In this case, the probability of event A, denoted by $\mathrm{P(A)}$, is defined by:

$$\mathrm{P(A)} = \lim_{\mathrm{n}\to\infty} \frac{\mathrm{n_A}}{\mathrm{n}},$$

which means that $\mathrm{P(A)}$ is defined as the limit relative to the number of times event A occurred relative to the total number of repetitions. In the second approach, for each event A, we accept that there exists a probability of A, $\mathrm{P(A)},$ satisfying the following three axioms:

1. $0 \leq \mathrm{P(A)} \leq 1,$

2. $\mathrm{P}(\Omega) = 1,$

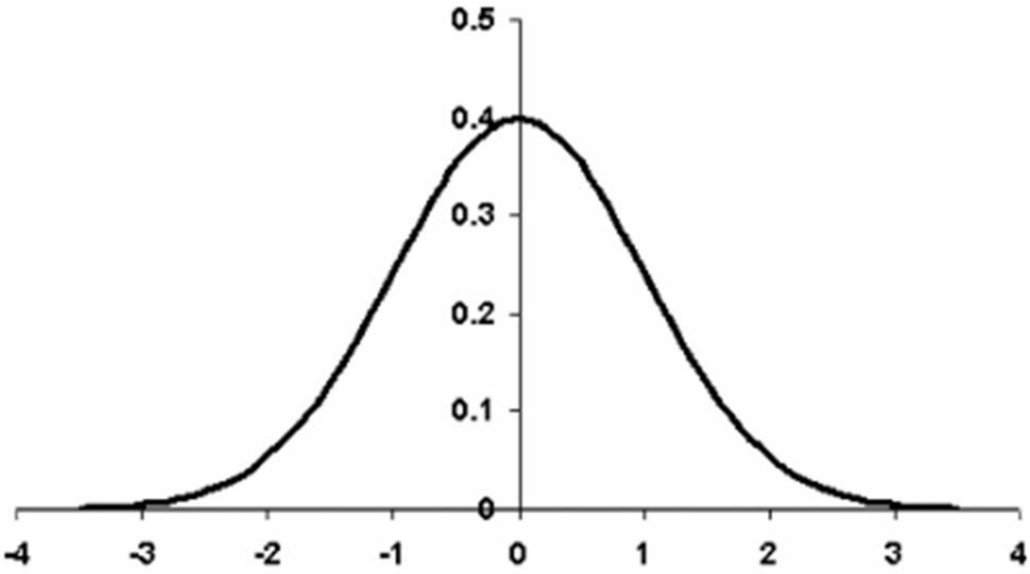
3. For each sequence of mutually exclusive events $\mathrm{A_1, A_2, \ldots}$ (that is of events $\mathrm{A_i} \cap \mathrm{A_j} = \phi \text{ if } \mathrm{i} \neq \mathrm{j}$):
   $\mathrm{P}[\cup_{\mathrm{i}=1}^{\infty} \mathrm{A_i}] = \sum_{\mathrm{i}=1}^{\infty} \mathrm{P(A_i)}.$

2.1.2 Normal Distribution

Random variable X is distributed according to a normal distribution if it has a density function of the form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right),$$

$$(\sigma > 0).$$



Normal distribution, $\mu = 0, \sigma = 1$

We will say that X follows a normal distribution of mean μ and of variance σ^2. The normal distribution is a continuous probability distribution.

The expected value of the normal distribution is given by:

$E[X] = \mu.$

The variance is equal to:

$\mathrm{Var}(X) = \sigma^2.$

Variance is a measure of dispersion of a distribution of a random variable. Empirically, the variance of a quantitative variable X is defined as the sum of squared deviations of each observation relative to the arithmetic mean divided by the number of observations.

Variance is generally denoted by S^2 when it is relative to a sample and by σ^2 when it is relative to a population N. We also denote the variance by $\mathrm{Var}(X)$ when we speak about the variance of a random variable.

By definition, the variance of a population is calculated as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i\text{-}\mu)^2}{N},$$

where N is the sum of the population and μ it’s mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}.$$

If the mean μ is equal to 0, and the variance σ^2 is equal to 1, then we obtain the standard normal distribution whose density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp(\frac{\text{-}x^2}{2}).$$

If a random variable X follows a normal distribution of mean μ and variance σ^2, then the random variable

$$Z = \frac{X\text{-}\mu}{\sigma}$$

Follows a standard normal distribution.

The normal distribution plays a central role in the theory of probability and its statistical applications. Many measurements such as the size or weight of individuals, the diameter of a piece of machinery, the results of an IQ test, etc. approximately follow a normal distribution. The normal distribution is frequently used as an approximation, either when the normality is attributed to a distribution in the construction of a model or when a known distribution is replaced by a normal distribution with the same expected value or variance. It is used for the approximation of the chi-square distribution, the Student distribution with large degrees of freedom and discrete probability distributions such as the binomial distribution and the Poisson distribution for large N. The normal distribution is also a fundamental element of the theory of sampling, where its role is important in the study of correlation, regression analysis, variance analysis, and covariance analysis.

2.1.3 Covariance

The covariance between two random variables X and Y is a measure of how much two random variables vary together.

If X and Y are independent random variables, the covariance of X and Y i is zero. The converse, however, is not true.

Consider X and Y, two random variables defined in the same sample space. The covariance of X and Y, denoted by $\mathrm{Cov}(\mathrm{X},\mathrm{Y})$, is defined by

$$\mathrm{Cov}(\mathrm{X},\mathrm{Y}) = \mathrm{E}[(\mathrm{X}\text{-}\mathrm{E}[\mathrm{X}])(\mathrm{Y}\text{-}\mathrm{E}[\mathrm{Y}])],$$

where $\mathrm{E}[.]$ is the expected value.

2.1.4 Categorical Data: Contingency Table

A category represents a set of people or objects that have a common characteristic. If we want to study the people in a population, we can sort them into “natural” categories, by gender (men and women) for example, or into categories defined by other criteria, such as vocation (managers, secretaries, farmers . . .).

Categorical data consists of counts of observations falling into specified classes. We can distinguish between various types of categorical data:

- Binary, characterizing the presence or absence of a property;

- Unordered multi categorical (also called “nominal”);

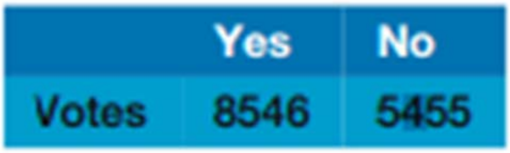
- Ordered multi categorical (also called “ordinal”);

- Whole numbers.

We represent the categorical data in the form of a contingency table.

Variables that are essentially continuous can also be presented as categorical variables. One example is “age”, which is a continuous variable, but ages can still be grouped into classes so it can still be presented as categorical data.

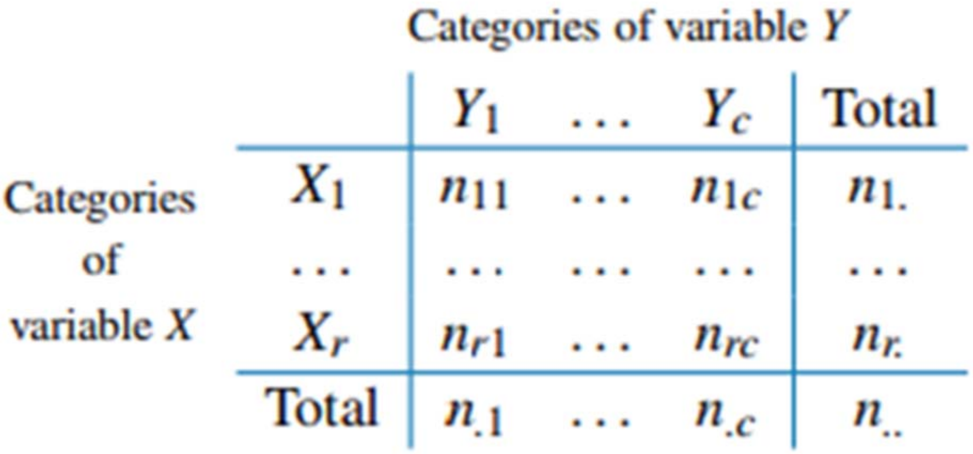
In a public opinion survey for approving or disapproving a new law, the votes cast can be either “yes” or “no”. We can represent the results in the form of a contingency table:



	Yes	No
Votes	8546	5455

If we divide up the employees of a business into professions (and at least three professions are presented), the data we obtain is unordered multi categorical data (there is no natural ordering of the professions). In contrast, if we are interested in the number of people that have achieved various levels of education, there will probably be a natural ordering of the categories: “primary, secondary” and then university. Such data would therefore be an example of ordered multi categorical data. Finally, if we group employees into categories based on the size of each employee’s family (that is, the number of family members), we obtain categorical data where the categories are whole numbers.

Consider a contingency table relating to two categorical qualitative variables X and Y that have, respectively, r and c categories:



		Categories of variable Y			
		Y_1	...	Y_c	Total
Categories of variable X	X_1	n_{11}	...	n_{1c}	$n_{1.}$
	...	...	...	...	...
	X_r	n_{r1}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	...	$n_{.c}$	$n_{..}$

where

n_{ij} j represents the frequency that category i of variable X and category j of variable Y is observed,

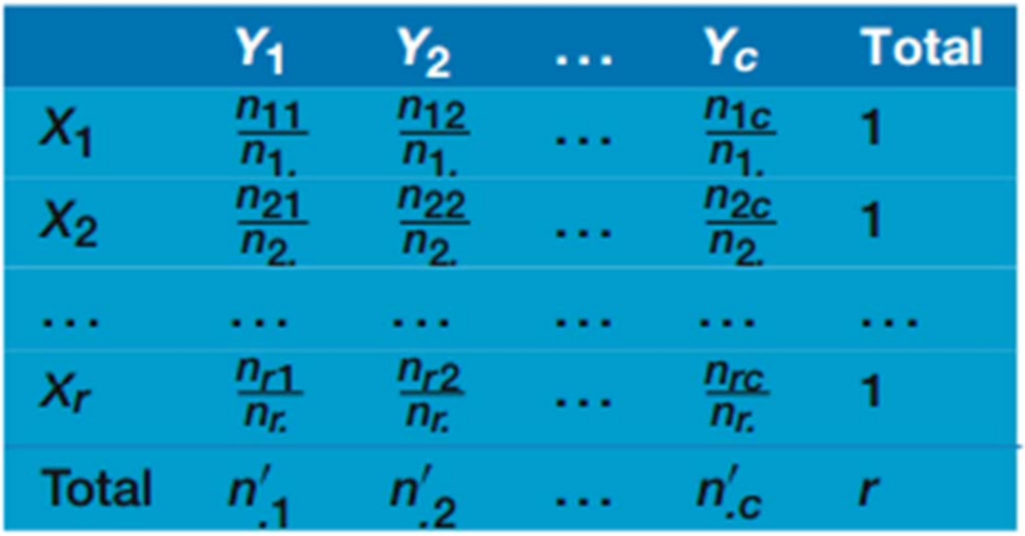
n_i represents the sum of the observed frequencies for category i of variable X,

$n_{.j}$ represents the sum of the observed frequencies for category j of variable Y,

$n_{..}$ represents the total number of observations.

Tables of row profiles X_I and column profiles X_J are constructed.

For a fixed line (column), the line (column) profile is the line (column) obtained by dividing each element in this row (column) by the sum of the elements in the line (column). The line profile of row i is obtained by dividing each term of row i by $n_{i.}$, which is the sum of the observed frequencies in the row. The table of row profiles is constructed by replacing each row of the contingency table with its profile:



	Y_1	Y_2	...	Y_c	Total
X_1	$\frac{n_{11}}{n_{1.}}$	$\frac{n_{12}}{n_{1.}}$	...	$\frac{n_{1c}}{n_{1.}}$	1
X_2	$\frac{n_{21}}{n_{2.}}$	$\frac{n_{22}}{n_{2.}}$	...	$\frac{n_{2c}}{n_{2.}}$	1
...	...	...	...	...	...
X_r	$\frac{n_{r1}}{n_{r.}}$	$\frac{n_{r2}}{n_{r.}}$	...	$\frac{n_{rc}}{n_{r.}}$	1
Total	$n'_{.1}$	$n'_{.2}$	...	$n'_{.c}$	r

The column profile matrix is constructed in a similar way, but this time each column of the contingency table is replaced with its profile: the column profile of column j is obtained by dividing each term of column j by $n_{.j}$, which is the sum of frequencies observed for the category corresponding to this column. The tables of row profiles and column profiles correspond to a transformation of the contingency table that is used to make the rows and columns comparable.

2.1.5 Chi squared distribution

Consider a frequency table with r rows and p columns, it is possible to calculate row profiles and column profiles. We can define the distances between these profiles. The Euclidean distance between the components of the profiles, on which a weighting is defined (each term has a weight that is the inverse of its frequency), is called the chi square distance. The name of the distance between rows i and i' is derived from the fact that the mathematical expression defining the distance is identical to that encountered in the elaboration of the chi square goodness of fit test.

$$d(i,i') = \sqrt{\sum_{j=1}^{c} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2 * \frac{1}{f_{.j}}},$$

Where

$f_{i.}$ Is the sum of the components of the ith row;

$f_{.j}$ Is the sum of the components of the jth column;

$\left[\frac{f_{ij}}{f_{i.}}\right]$ Is the ith row profile for $j = 1,2,3, \dots, c$.

Likewise, the distance between two columns j and $j^{'}$ is given by:

$$d(j, j^{'}) = \sqrt{\sum_{j=1}^{r} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2 * \frac{1}{f_{i.}}},$$

Where $\left[\frac{f_{ij}}{f_{.j}}\right]$ is the jth column profile for $j = 1, \dots, r$.

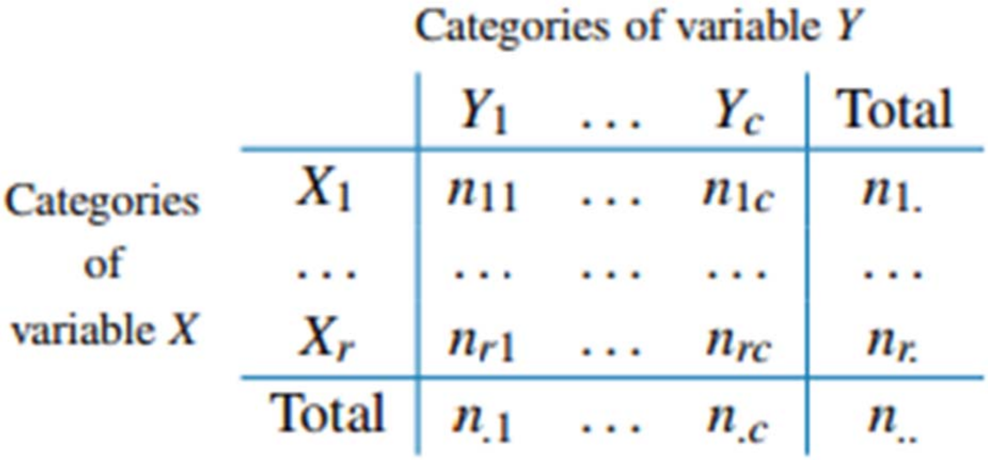
The chi-square distance incorporates a weight that is inversely proportional to the total of each row (or column), which increases the importance of small deviations in the rows (or columns) which have a small sum with respect to those with more important sum package. The chi-square distance has the property of distributional equivalence, meaning that it ensures that the distances between rows and columns are invariant when two columns (or two rows) with identical profiles are aggregated.

2.1.6 Chi squared Test of Independence

The chi-square test of independence aims to determine whether two variables associated with a sample are independent or not. The variables studied are categorical qualitative variables. The chi-square independence test is performed using a contingency table.

Consider two qualitative categorical variables X and Y. We have a sample containing n observations of these variables.

These observations are summarized in a contingency table. We denote the observed frequency of the category i of the variable X and the category j of the variable Y as n_{ij}.



		Categories of variable Y			
		Y_1	$\dots$	Y_c	Total
Categories of variable X	X_1	n_{11}	$\dots$	n_{1c}	$n_{1.}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	X_r	n_{r1}	$\dots$	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	$\dots$	$n_{.c}$	$n_{..}$

The hypotheses to be tested are:

Null hyp. H_0: The two variables are independent,

Alternative hyp. H_1: The two variables are not independent.

Steps of the test:

1. Compute the expected frequencies, denoted by e_{ij}, for each case in the contingency table under the independence hypothesis:

$$\mathrm{e}_{\mathrm{ij}} = \frac{\mathrm{n}_{\mathrm{i}}{*}\mathrm{n}_{\mathrm{j}}}{\mathrm{n}_{..}},$$

$$\mathrm{n}_{\mathrm{i}} = \sum_{\mathrm{k}=1}^{\mathrm{c}} \mathrm{n}_{\mathrm{ik}} \text{ and } \mathrm{n}_{.\mathrm{j}} = \sum_{\mathrm{k}=1}^{\mathrm{r}} \mathrm{n}_{\mathrm{kj}},$$

2. Calculate the value of the χ^2 (chi-square) statistic, which is really a measure of the deviation of the observed frequencies n_{ij} from the expected frequencies e_{ij}:

$$\chi^2 = \sum_{\mathrm{i}=1}^{\mathrm{c}} \sum_{\mathrm{j}=1}^{\mathrm{r}} \frac{(\mathrm{n}_{\mathrm{ij}}\text{-}\mathrm{e}_{\mathrm{ij}})^2}{\mathrm{e}_{\mathrm{ij}}}.$$

3. Choose the significance level α to be used in the test and compare the calculated value of χ^2 with the value obtained from the chi-square table, $\chi^2_{v,\alpha}$. The number of degrees of freedom correspond to the number of cases in the table that can take arbitrary values; the values taken by the other cases are imposed on them by the row and column totals. So, the number of degrees of freedom is given by:

$v = (r\text{-}1)(c\text{-}1).$

4. If the calculated χ^2 is smaller then the $\chi^2_{v,\alpha}$ from the table, we do not reject the null hypothesis. The two variables can be considered to be independent. However, if the calculated χ^2 is greater then the $\chi^2_{v,\alpha}$ from the table, we reject the null hypothesis for the alternative hypothesis. We can then conclude that the two variables are not independent.

Certain conditions must be fulfilled in order to be able to apply the chi-square test of independence:

1. The sample, which contains n observations, must be a random sample;

2. Each individual observation can only appear in one category for each variable. In other words, each individual observation can only appear in one line and one column of the contingency table.

Note that the chi-square test of independence is not very reliable for small samples, especially when the estimated frequencies are small, that means < 5. To avoid this issue we can group categories together, but only when these groups obtained are sensible.

2.1.7 Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov test is a nonparametric goodness-of-fit test and is used to determine whether two distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution. It is used when we have two samples coming from two populations that can be different. Unlike the Mann–Whitney test and the Wilcoxon test where the goal is to detect the difference between two means or medians, the Kolmogorov–Smirnov test has the advantage of considering the distribution functions collectively. The Kolmogorov– Smirnov test can also be used as a goodness of-fit test. In this case, we have only one random sample obtained from a population where the distribution function is specific and known.

Consider two independent random samples:

$(X_{1,}X_2, \dots, X_n)$, a sample of size n coming from a population 1, and $(Y_{1,}Y_2, \dots, Y_m)$, a sample of dimension m coming from a population 2. We denote by, respectively, $F(x)$ and $G(x)$ their unknown distribution functions.

The hypotheses to test are as follows:

A: Two-sided case:

$H_0 : F(x) = G(x) \text{for each } x$

$H_1 : F(x) \neq G(x) \text{or at least one value of } x$

B: One-sided case:

$H_0 : F(x) \leq G(x) \text{for each } x$

$H_1 : F(x) > G(x) \text{or at least one value of } x$

C: One-sided case:

$H_0 : F(x) \geq G(x) \text{for each } x$

$H_1 : F(x) < G(x) \text{or at least one value of } x$

In case A, we make the hypothesis that there is no difference between the distribution functions of these two populations. Both populations can then be seen as one population. In case B, we make the hypothesis that the distribution function of population 1 is smaller than those of population 2. We sometimes say that, generally, X tends to be smaller than Y. In case C, we make the hypothesis that X is greater than Y. We denote by $H_1(x)$ the empirical distribution function of the sample

$(X_{1,}X_2, \dots, X_n)$ and by $H_2(x)$) the empirical distribution function of the sample $(Y_{1,}Y_2, \dots, Y_m)$. The statistical test are defined as follows:

A: Two-tail case

The statistical test T_1 is defined as the greatest vertical distance between two empirical distribution functions:

$$T_1 = {}^{\text{sup}}_{\text{x}}|H_1(x)\text{-}H_2(x)|.$$

B: One-tail case The statistical test T_2 is defined as the greatest vertical distance when $H_1(x)$ is greater than $H_2(x)$:

$$T_2 = {}^{\text{sup}}_{\text{x}}|H_1(x)\text{-}H_2(x)|.$$

C: One-tail case The statistical test T_3 is defined as the greatest vertical distance when $H_2(x)$ is greater than $H_1(x)$:

$$T_3 = {}^{\text{sup}}_{\text{x}}|H_2(x)\text{-}H_3(x)|.$$

We reject H_0 at the significance level α if the appropriate statistical test $(T_1, T_2 \text{ or } T_3)$ is greater than the value of the Smirnov table having for parameters $\text{n}, \text{m}, \text{and } 1\text{-}\alpha$, which we denote by $t_{n,m,1\text{-}\alpha}$ that is, if

$$T_1(\text{or } T_2 \text{ or } T_3) > t_{n,m,1\text{-}\alpha}.$$

If we want to test the goodness of fit of an unknown distribution function $F(x)$ of a random sample from a population with a specific and known distribution function $F_0(x)$, then the hypotheses will be the same as those for testing two samples, except that $F(x)$ and $G(x)$ are replaced by $F(x)$ and $F_0(x)$.

If $H(x)$ is the empirical distribution function of a random sample, then the statistical tests T is defined as follows:

$$T = {}^{\text{sup}}_{\text{x}}|F_0(x)\text{-}H(x)|.$$

The decision rule is as follows: reject H_0 at the significance level α if T is greater than the value of the Kolmogorov table having for parameters n and $1\text{-}\alpha$, which we denote by $t_{n,1\text{-}\alpha}$ that is, if

$$T \ > t_{n,1\text{-}\alpha}.$$

To perform the Kolmogorov–Smirnov test, the following must be observed:

1. Both samples must be taken randomly from their respective populations.

2. There must be mutual independence between two samples.

3. The measure scale must be at least ordinal.

4. To perform an exact test, the random variables must be continuous, otherwise the test is less precise.

2.1.8 Multiple Linear Regression

A regression analysis where variable Y may linearly depend on many independent variables $X_1, X_2, \dots, X_k$ is called multiple linear regression.

A multiple linear regression equation is of the form:

$$Y = f(X_1, X_2, \dots, X_k),$$

where $f(X_1, X_2, \dots, X_k)$ is a linear function of $X_1, X_2, \dots, X_k$.

A general model of multiple linear regression containing $k = (p\text{-}1)$ independent variables (and n observations) is written as:

$$Y_i = \beta_0 + \sum_{j=1}^{p\text{-}1} X_{ji}\beta_j + \epsilon_i, \quad i = 1, \dots, n \ ,$$

Y_i is the dependent variable $X_{ji}, j = 1, \dots, p\text{-}1$, are the independent variables, $\beta_j, j = 0, \dots, p\text{-}1$, are the parameters to be estimated, and ε_i is the term of random non observable error.

In the matrix form, this model is written as:

$Y = X\beta + \varepsilon$**,**

where Y is the vector $(n \times 1)$ of observations related to the dependent variable (n observations), β is the vector $(p \times 1)$ of parameters to be estimated, ε is the vector $(n \times 1)$ of errors,

and $X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1(p\text{-}1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{n(p\text{-}1)} \end{pmatrix}$ is the $(n \times p)$ matrix of the independent variables.

Starting from the model

$Y = X\beta + \varepsilon$**,**

with minimizing the sum of squared errors (least-squares method)

$$\min_{\beta}[\epsilon'\epsilon] = \min_{\beta}[(Y\text{-}X\beta)'(Y\text{-}X\beta)] = \min_{\beta}[Y'Y\text{-}\beta'X'Y\text{-}Y'X\beta + \beta'X'X\beta]$$

$$= \min_{\beta}[Y'Y\text{-}2Y'X\beta + \beta'X'X\beta]$$

$$\vartheta[Y'Y\text{-}2Y'X\beta + \beta'X'X\beta]/\vartheta\beta = \text{-}2X'Y + 2X'X\beta = 0$$

we obtain the normal equations and further the estimate $\hat{\beta}$ of the vector β:

$\hat{\beta} = (X'X)^{-1}X'Y,$

and we calculate an estimated value $\hat{Y}$ for Y**:**

$\hat{Y} = X\hat{\beta}$**.**

At this step, we can calculate the residuals, denoted by vector e, that we find in the following manner:

$e = Y\text{-}\hat{Y}$.

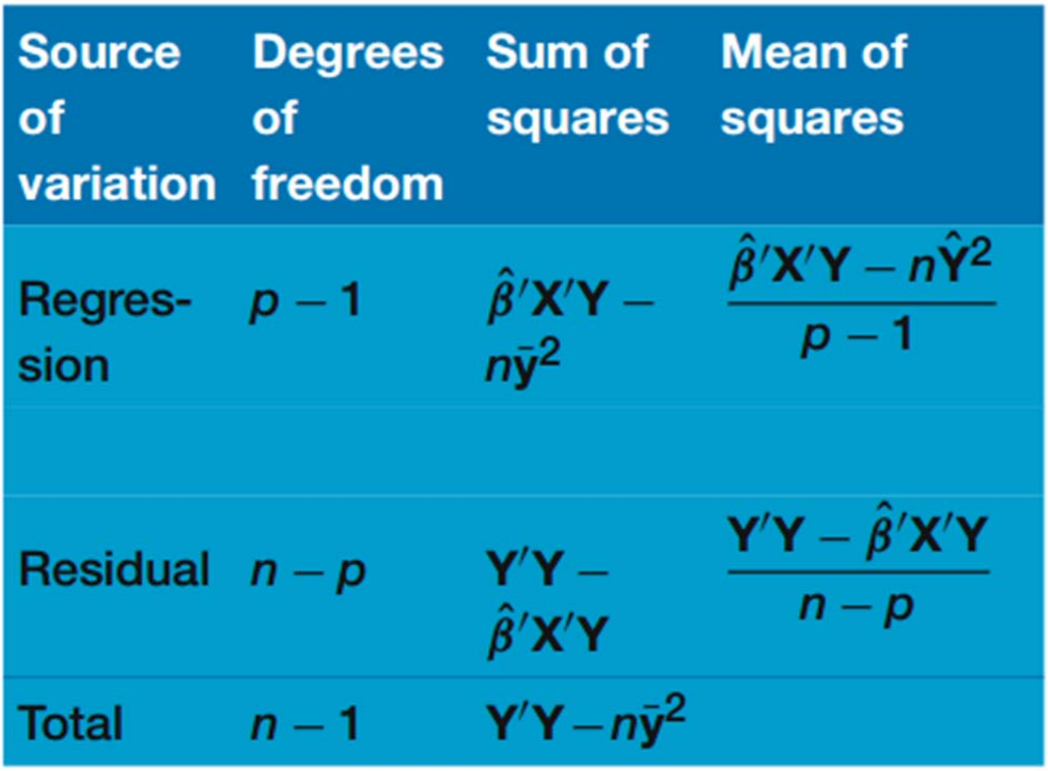
To know which measure to trust for the chosen linear model, it is useful to conduct an analysis of variance and to test the hypotheses on vector β of the regression model. To conduct these tests, we must make the following assumptions:

- For each value of $X_{ji}$and for all $i = 1, \ldots, n$ and $j = 1, \ldots, p\text{-}1$, there is a random variable Y distributed according to the normal distribution.

- The variance of Y is the same for all X_{ji}; it equals σ^2 (unknown).

- The different observations on Y are independent of one another but conditioned by the values of X_{ji}.

In the matrix form, the table of analysis of variance for the regression is as follows:



Source of variation	Degrees of freedom	Sum of squares	Mean of squares
Regression	$p-1$	$\hat{\beta}'X'Y - n\bar{y}^2$	$\frac{\hat{\beta}'X'Y - n\hat{Y}^2}{p-1}$
Residual	$n-p$	$Y'Y - \hat{\beta}'X'Y$	$\frac{Y'Y - \hat{\beta}'X'Y}{n-p}$
Total	$n-1$	$Y'Y - n\bar{y}^2$	

If the model is correct, then S^2 the variance

$$S^2 = \frac{Y^{'}Y\text{-}\hat{\beta}'X'Y}{n\text{-}p}$$

is an unbiased estimator of σ^2.

The analysis of variance allows us to test the null hypothesis:

$H_0\colon \beta_j = 0 \text{ for } j = 1, \ldots, p\text{-}1$

against the alternative hypothesis:

$H_1\colon \text{at least one of the parameters } \beta_j, j \neq 0, \text{ is different from zero}$

calculating the statistic:

$F = \frac{EMSE}{RMSE} = \frac{EMSE}{S^2},$

where EMSE is the mean of squares of the regression, RMSE is the mean of squares of residuals, and TMSE is the total mean of squares.

This statistic F must be compared with the value $F_{\alpha,p\text{-}1,n\text{-}p}$ of the Fisher table, where α is the significance of the test.

⇒

$\text{If } F \leq F_{\alpha,p\text{-}1,n\text{-}p} \text{ , then we accept } H_0$

$\text{If } F > F_{\alpha,p\text{-}1,n\text{-}p} \text{ , then we reject } H_0 \text{for } H_1.$

The coefficient of determination R^2 is calculated in the following manner:

$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}'X^{'}Y\text{-}n\hat{y}^2}{Y^{'}Y\text{-}n\hat{y}^2},$

Where ESS is the sum of squares of the regression and TSS is the total sum of squares.

2.1.9 Simple Linear Correlation Coefficient

The simple correlation coefficient is a measure of the strength of the linear relation between two random variables. The correlation coefficient can take values that occur in the interval [−1; 1]. The two extreme values of this interval represent a perfectly linear relation between the variables, “positive” in the first case and “negative” in the other. The value zero implies the absence of a linear relation. The correlation coefficient presented here is also called the Bravais–Pearson correlation coefficient

Simple linear correlation is the term used to describe a linear dependence between two quantitative variables X and Y (see simple linear regression). If X and Y are random variables that follow an unknown joint distribution, then the simple linear correlation coefficient is equal to the covariance between X and Y divided by the product of their standard deviations:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

Here $Cov(X, Y)$ is the measured covariance between X and Y; σ_X and σ_Y are the respective standard deviations of X and Y.

Given a sample of size n, (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) from the joint distribution, the quantity

$$r = \frac{\sum_{i=1}^{n}(X_i\text{-}\overline{X})(Y_i\text{-}\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i\text{-}\overline{X})^2 \sum_{i=1}^{n}(Y_i\text{-}\overline{Y})^2}}$$

is an estimation of ρ. It is the sampling correlation.

If we denote $(X_i\text{-}\overline{X})$ by x_i and $(Y_i\text{-}\overline{Y})$ by y_i, we can write this equation as:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)}}$$

To test the null hypothesis:

$$H_0: \rho = 0$$

against the alternative hypothesis

$H_0: \rho \neq 0,$

we calculate the statistic t:

$t = \frac{r}{S_r},$

where S_r is the standard deviation of the estimator r:

$S_r = \sqrt{\frac{1-r^2}{n-2}}.$

Under H_0, the statistic t follows a Student distribution with $n-2$ degrees of freedom. For a given significance level α, H_0 is rejected if $|t| \geq t_{\frac{\propto}{2},n-2}$; the value of $t_{\frac{\propto}{2},n-2}$ is the critical value of the test given in the Student table.

2.1.10 Coefficient of Determination

The coefficient of determination denoted by R^2 determines whether the hyperplane estimated from a multiple linear regression is correctly adjusted to the data points.

The value of the multiple determination coefficient R^2 is equal to:

$$R^2 = \frac{\text{Explained varation}}{\text{Total variation}}$$

$$= \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

It corresponds to the square of the multiple correlation coefficient. Notice that

$0 \leq R^2 \leq 1.$

In the case of simple linear regression, the following relation can be derived:

$r = \text{sign}(\widehat{\beta}_1)\sqrt{R^2},$

where $\widehat{\beta}_1$ is the estimator of the regression coefficient β_1, and it is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i\text{-}\bar{X})(Y_i\text{-}\bar{Y})}{\sum_{i=1}^{n}(X_i\text{-}\bar{X})^2}.$$

If there is an exact linear relation between two variables, the correlation coefficient is equal to 1 or −1. A positive relation (+) means that the two variables vary in the same direction. If the individuals obtain high scores in the first variable (for example the independent variable), they will have a tendency to obtain high scores in the second variable (the dependant variable). The opposite is also true. A negative relation (−) means that the individuals that obtain high scores in the first variable will have a tendency to obtain low scores in the second one, and vice versa. Note that if the variables are independent the correlation coefficient is equal to zero. The reciprocal conclusion is not necessarily true. The fact that two or more variables are related in a statistical way is not sufficient to conclude that a cause and effect relation exists. The existence of a statistical correlation is not a proof of causality.

Statistics provides numerous correlation coefficients. The choice of which to use for a particular set of data depends on different factors, such as:

- The type of scale used to express the variable;
- The nature of the underlying distribution (continuous or discrete);
- The characteristics of the distribution of the scores (linear or nonlinear).

2.1.11 Quantile

Quantiles measure position (or the central tendency) and do not necessarily try to determine the center of a distribution of observations, but to describe a particular position. This notion is an extension of the concept of the median (which divides a distribution of observation into two equal parts).

The most frequently used quantiles are:

- Quartiles, which separate a collection of observations into four parts,

- Deciles, which separate a collection of observations into ten parts,

- Centiles, which separate a collection of observations into a hundred parts.

The calculation of quantiles makes sense only for a quantitative variable that can take values on a determined interval. Note that the greater the number of observations, the more sophisticated the separation of the distribution can be. Quantiles can generally be used for any distribution. The calculation of deciles and, a fortiori, centiles requires a relatively large number of observations to obtain a valid interpretation.

2.1.12 Box Plot

The box plot is a way to represent the following five quantities for a set of data: the median; the first quartile and the third quartile; the maximum and minimum values. The box plot is a diagram (a box) that illustrates:

- The measure of central tendency (in principal the median);

- The variability, and;

- The symmetry.

It is often used to compare several sets of observations.

2.1.13 Value at Risk (VaR)

The VaR of a random variable revenue is defined as in the following. (Tsay, 2005)

$$\mathrm{Prob}(\mathrm{Revenue} < \mathrm{VaR}_{\mathrm{Revenue}}) = 0{,}10$$

2.2 CRM – Customer Relationship management

Customer relationship management (CRM) is the customer-focused business strategy which is not a new concept. Although CRM is more about the customer, it cannot be successful by this definition alone. CRM should be performed in organizations as the combination of three main concepts: people, processes, and technology. CRM is a combination of people, processes, and technology that seeks to provide understanding of customer needs, to support a business strategy, and to build long-term relationships with customers. To increase relationships with all customers the integration of these three is essential. Applying CRM's system in one organization means a change in different areas of the business and seeks a proper balance of people, processes, and technology. One of the main reasons of CRM failures is considering technology as the main part of the system. CRM project success will happen if the CRM users investigate people, process, and technology either one by one or together.

The goals of CRM are:

- Building long-term and profitable relationships with chosen customers,
- Getting closer to those customers at every point of contact with them

The whole idea of studying, analysing and creating new customers while trying to keep the current customers happy and satisfied is known as CRM. The very core of CRM is nothing more than collecting customer data and analysing it to make decisions that bring in new customers apart from satisfying the existing ones. (Arockia Raj, 2012)

2.2.1 Knowledge of Customer Needs

CRM allows an organization to develop a knowledge base that all employees can accessible easily. This allows the company to analyse available data and provides employees with accurate information about customers. It also empowers the organization to arrive at correct and well-informed decisions. In addition, it helps the company be as close to the client base as possible so it can effectively anticipate their changing needs and cater to such needs. With the knowledge base, employees can easily share and update any piece of information to any department with ease.

Gummesson (2004) also addresses the value of CRM in B2B contexts, emphasizing the shortage of empirical research in the area. He remarks that it is not easy to operationalize return on relationships in a B2B setting and proposes indicators that focus on the customer, employees, and the information technology interface between the customer and the company. Furthermore, (Boulding, Staelin, Ehret, & Johnston, 2005)

Relationship marketing and customer relationship management (CRM) in general have become central business issues. With more intense competition in many mature markets companies have realized that development of relationship with more profitable customer is a critical factor to staying in the market. Thus, CRM techniques have been developed that afford new opportunities for businesses to act well in a relationship market. The focus of CRM is on the customer and the potential for increasing revenue, and in doing so it enhances the ability of a firm to compete and to retain key customers.

The relationship between a business and customers can be described as follows. A customer purchases products and services, while business is to market, sell, provide and service customers. Generally, there are three ways for business to increase the value of customers:

- increase their usage (or purchases) on the products or service that customers already have;
- sell customers more or higher-profitable products;
- keep customers for a longer time.

A valuable customer is usually not static and the relationship evolves and changes over time. Thus, understanding this relationship is a crucial part of CRM. This can be achieved by analysing the customer life-cycle, or customer lifetime, which refers to various stages of the relationship between customer and business. A typical customer life-cycle is shown in Figure 1. (Olafsson S. , 2008)

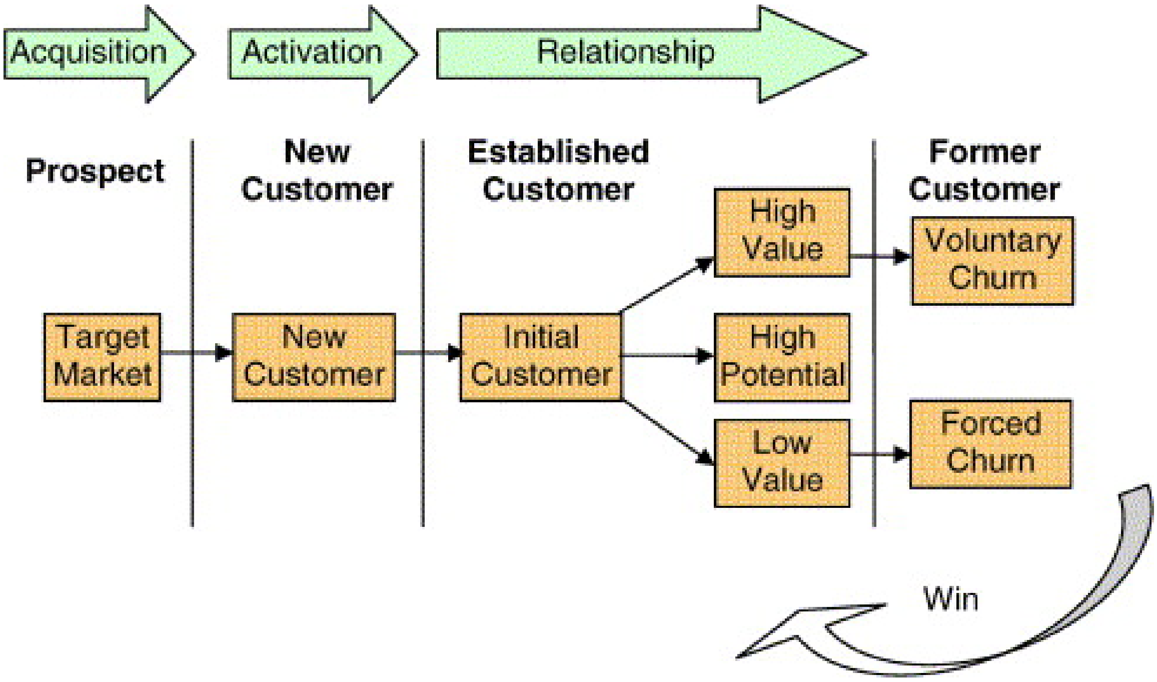


Figure 1: Illustration of a customer life-cycle (Olafsson S. , 2008)

According to Figure 1 the first step acquisition are marketing campaigns that are directed to the target market and seek to interest prospects in a company’s product or service. If prospects respond to company’s inquiry then they will become respondents. Responders become established customers when the relationship between them and the companies has been established. For example, they have made the initial purchase or their application for a certain credit card has been approved. At this point, companies will gain revenue from customer usage. Furthermore, customers’ value will be increased not only by cross-selling that encourages customers to buy more products or services but also by up-selling that encourage customers to upgrading existing products and services. On the other hand, at some point established customers stop being customers (churn). There are two different types of churns. The first is voluntary churn, which means that established customers choose to stop being customers. The other type is forced churn, which refers to those established customers who no longer are good customers and the company cancels the relationship. The main purpose of CRM is to maximize customers' values throughout the life-cycle. (Olafsson S. , 2008)

2.3 Forecasting

Forecasts are the foundation of many business decisions, wether expectations for the future are based upon quatative information, qualittative information, intuition, or other means. As the quantity and quality of information contious to expand, we have an opportunity to expand our perspectives regarding the factors influencing and driving demand, as well as the process by which purchase decisions are made. Predivtive analytics moves beyond the patterns of demand and facilitates a view into more behavior-based understanding of our customers, their interests and needs, and their consumption of our prdoucts and services. This is an emerging and developmental approach to getting a more complete view of the future for use in business decision processes.

Companies invest in planning and forecasting processes, technology systems, methods and metrics, inventories and business analytics in an effort to improve their ability to satisfy demand. Also opportunity forecasting based on CRM opportunity data from the past is a common strategy to support business decisions and provide an overview about possible future business figures. (Fildes, Nikolopoulos, Crone, & Syntetos, 2008)

But all attempts to predict and forecast are uncertain and riddled with inaccuracy. Nevetheless, anything that improves accucarcy and reduces uncertainty has the potential to materially improve a company's performance and position in the markets that is serves. Thus, the quest for improved approaches to forecasting, new methologies and more robust data continous.

The traditional forecasting approach done by businesses has been based upon mathematically extrapolating past demand into the future on the assumption that the future demand will follow the same mathematical patterns as the past. Often, time series methods such as averages, trend models, seasonal model, and decomposition model, are used to extrapolate the data. One of the widely used time series methodes in business is exponential smoothing. A blending of these methodological approaches predicting future demand. Of course, this assumes that future states and conditions will be the same as or at least similar to the past. This is an important and strong underlying assumption. And it is one that doesn't facilitate an understanding of what kinds of factors affect demand, and what kinds of influences affect purchase behaviours. The low cost of producing time series forecasts, the limited information required, as well, as the ease of calculation has contributed materially to its widespread use. Quite often the time series projection are qualitativley modified to lelive the ceteris paribus assumption of time series methods, in an effort to incoparte business intelligence consideration of other factors influencing the estimate. Therefore, the predictions are a result of a collage of systematic and non systematic methods that may improve or may damage the reliability of the estimate, and may introduce additional sources of judgemental error and bias. Certainly it would be better if predictions could be made with a more robust approach that takes into account the various conditions, factors, and influences for the varaible to improve forecasting.

The new approach to forecasting is called predictive analytics. This is the practice of extracting information from existing data sets in order to dertermine patterns and predict future outcomes and trends. Predicitve nalytics does not tell you what will happen in the future. It forecasts what might happen in the future with an acceptable level of reliabilty, and includes what-if scenarios and risk assesment. Predicte models and anlaysis are typically used to analyze current data and historical facts in order to better understand customers, products, and partners as well as identifying potential risks and opportunities for a company.

There are various models that are used in predicitve analytics system. Linear and logistic regression models are widely used, as are neural networks. These are classic methodologies included with the “Cause and Effect” grouping of models.

Predicitive analytics is general purpose approach that can be applied to a variety of questions, problems, and business needs. It has been applied to predict electiricy consumption, stock markets prices, product demand, couppon redemption rates, personnel trunover, liablity rates, loan delinquency rates and much more. (Lawless, 2015)

In the following are the most important used methods which are used in this thesis to find a fitting forecasting method for this usecase:

- Contingency table to structure and aggregate the data
- Chi squared test of independence
- Kolmogorov-Smirnov test to compare distributions
- Linear regression
- Test for correlations
- Box plots to visualize the data.

3 Data construction

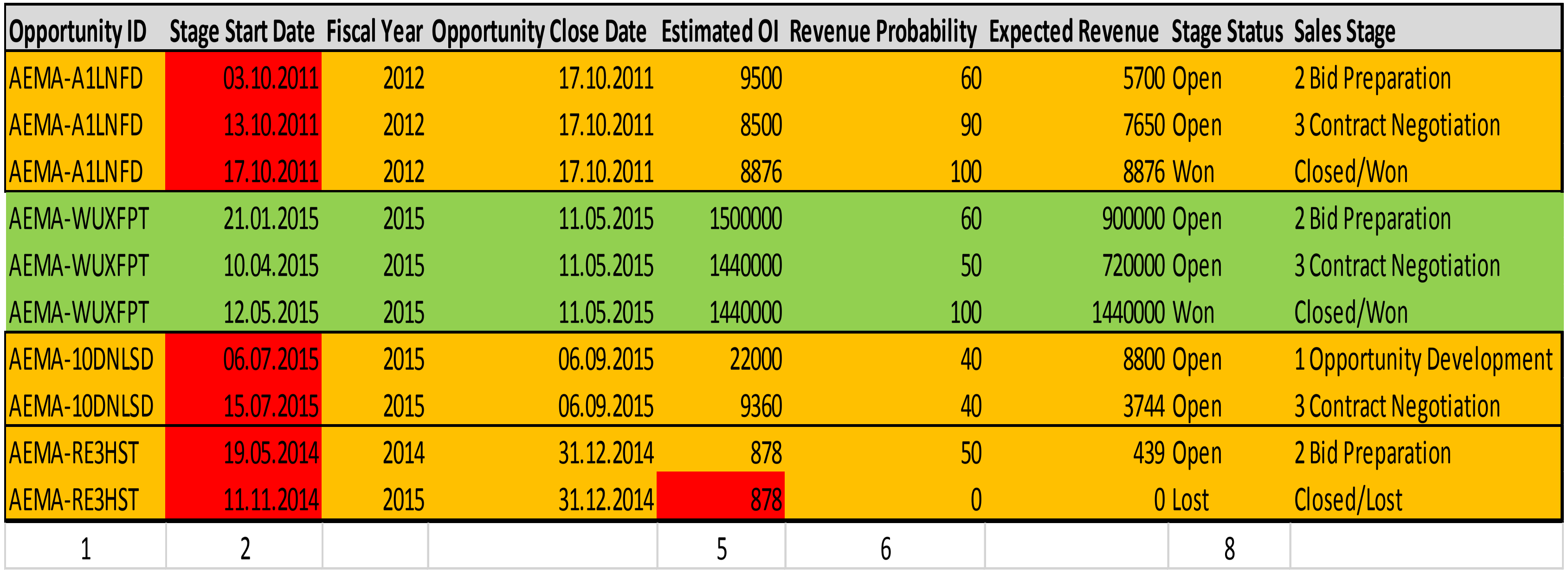
The data which is going to be analyzed in this thesis is called “Opportunity” (OPP). In this corporation it is used for documentation purposes of possible business opportunties with external customers.

The salesmen have to create these OPP within the coporate CRM System and to fill out the online form according to coporate CRM process standards. Only for this thesis relevant processes are going to be discussed.

Table 1 gives an example of the raw historic OPP data downloaded from the corporate CRM system, historic because each change of a OPP is documented with a time stamp, see column two called “Stage Start Date”. Fitting to this column, the third column called “Fiscal Year”, FY, is the actual FY of the “Stage Start Date”.

A coporate fiscal year (FY) starts always on the 1st of October and ends at the next year’s 30th of September. The first column shows the OPP ID. Every OPP has an “Unique Opportunity ID” within the corporate CRM system. The fifth column, “Estimated OI”, the volume of each business opportunity which is estimated by salesmen. “Estimated OI” stands for estimated order intake. Each timestamp of an OPP can have one of three statuses “Open”, “Won” or “Lost”. These statuses can be seen in the column “Stage Status”. In case the sales man wins a business opportunity a new record is introduced with a new timestamp and the status “Won”. The same for “Lost”. “Open” means the salesman is still in negotiation with the customer. Column six “Revenue Propability” tells us the propability in percentage to win this OPP. This value is an estimation coming from the salesmen experience. He can choose out of several pre given probability categories: 0%, 1%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%.

Based on previous information the system calculates the next column “Expected Revenue” which is “Estimated OI” multiplied with “Revenue Propability”. In case the “Stage Status” changes from “Open” to “Won” the “Revenue Propability” is going to be overwritten with 100 and so the “Expected Revenue” is becoming the “Estimated OI”. The same is if the status changes to “Lost” the “Revenue Propability” changes automatically to zero and so does also the “Expected Revenue”. In the last column called “Sales Stage” the salesman has the possibility to enter the opportunity status in detail. “Stage Status” and “Sales Stage” are linked together. If “Sales Stage” changes to a “Lost” or “Won” status so automatically does “Stage Status”. There is also one “Sales Stage” status called “Closed/Canceled”. In this stage the OPP is “Lost”, but has been canceled by the customer. In the column “Opportunity Close Date” the Salesman enters the estimated closure date of an OPP. The descption for all variables is available in Table 2.



Opportunity ID	Stage Start Date	Fiscal Year	Opportunity Close Date	Estimated OI	Revenue Probability	Expected Revenue	Stage Status	Sales Stage
AEMA-A1LNFD	03.10.2011	2012	17.10.2011	9500	60	5700	Open	2 Bid Preparation
AEMA-A1LNFD	13.10.2011	2012	17.10.2011	8500	90	7650	Open	3 Contract Negotiation
AEMA-A1LNFD	17.10.2011	2012	17.10.2011	8876	100	8876	Won	Closed/Won
AEMA-WUXFPT	21.01.2015	2015	11.05.2015	1500000	60	900000	Open	2 Bid Preparation
AEMA-WUXFPT	10.04.2015	2015	11.05.2015	1440000	50	720000	Open	3 Contract Negotiation
AEMA-WUXFPT	12.05.2015	2015	11.05.2015	1440000	100	1440000	Won	Closed/Won
AEMA-10DNLSD	06.07.2015	2015	06.09.2015	22000	40	8800	Open	1 Opportunity Development
AEMA-10DNLSD	15.07.2015	2015	06.09.2015	9360	40	3744	Open	3 Contract Negotiation
AEMA-RE3HST	19.05.2014	2014	31.12.2014	878	50	439	Open	2 Bid Preparation
AEMA-RE3HST	11.11.2014	2015	31.12.2014	878	0	0	Lost	Closed/Lost
1	2			5	6		8	

Table 1: Raw opportunity data

OPP	Opportunity
FY (Fiscal Year)	Fiscal year
MFY	Mid fiscal year (31.03. of every year)
EFY	End fiscal year (30.09. of every year)
OI	Order intake of an Opportunity
Opportuntiy ID (1)	Unique OPP database identification number
Stage Start Date (2)	OPP change date of a modification
Fiscal Year	OPP change year converted into actual FY
Sales Status	Represent the sales status in which the OPP is
Stage Status (8)	Can be Open/Won/Lost. Depends on the Sales Status
An OPP is “Closed”	When the “Stage Status” is either “Won” or “Lost”
Opportunity Close Date	OPP close date, estimated by the salesmen
Estimated OI (5)	OPP possible OI, estimated by the salesmen (Unweighted OI)
Revenue Probability (6)	Chance of success to win an OPP
Expected Revenue	= “Revenue Probability” * “Estimated OI” = Weighted OI

Table 2: List of variables

“Estimated OI”, “Revenue Propability”, “Opportunity Close Date” and “Sales Stage” is updated manually by the salesmen and he can choose all those values based on his experience with the customer. Each OPP modification will create a new row. For forecasting we use only columns 1, 2, 5, 6 and 8.

Column 1: Describes unique ID within the database table of this opportunity.

Column 2: Is the date when the opportunity was created, modified or closed.

Column 3: Set the FY according to the “Stage Start Date”.

Column 4: Is the closing date of the opportunity estimated by the salesmen

Column 5: Is the value of this opportunity

Column 6: Is the chance to win of this opportunity estimated by the salesmen

Column 7: Is the weighted volume of the opportunity, which is calculated (chance to win x opportunity volume)

Column 8: Tells the status of the opportunity, if it is open, won or closed

Column 9: Goes more in detail and explains the exact phase of the opportunity status, in which sales phase the opportunity is.

The other variables are listed because of completeness and give extra information about the status of ongoing business.

Assumption for the data construction:

- Data samples are going to be taken on the mid of the FY (MFY) is the 31th march of each year, based on the OPPs “Stage Start Date” and “Stage Status” “Open”. Then they are compared with the sample at EFY.

- Data from FY 2010 and 2011 are not used if they get “Closed” in these FYs because of the introduction period of the CRM system. Only “Closed” OPPs from FY 2012 – 2015 are used.

- In the following we will not distinguish between OPPs which are canceled by the customer and OPPs which are lost because of a competitor. So OPPs which are in the “Sales Status” “Closed”/”Canceled” count as a “Closed”/”Lost”.

- Only OPPs which have an “Estimated OI” > €1.000 are going to be considered. Due to coporate guidlines only OPPs with a value over €1.000 have to be entered into to the system.

- Only OPPs which are “Closed” between MFY and EFY of a fiscal year are considered in the data analysis, e.g. OPP ID AEMA-PLTIRE.

The OPP data is transformed to e.g.:

AEMA-WUXFPT	1500000	1440000	1	2	2015

Table 3: Transformed OPP

1,500,000 is the volume of the OPP AEMA-WUXFPT at MFY 2015. 1,440,000 is the volume of the OPP at EFY 2015. 1 means 100% “Revenue Propability” and 2 stand for status “Won”. “Lost” would be number 1. 2015 is the FY in which the analysis was performed.

With these restrictions the number of all considered OPPs from FY 2012 until 2015 is 415.

Description of the examples given in Table 1:

- OPP with the ID AEMA-A1LNFD is in the final status “Won”. It has an expected OI over €1,000, but has been created and closed in the first half of FY 2012. So it is not relevant for forecast, as the OI is already known at MFY. That’s why this OPP is not in our analysis included.

- OPP with the ID AEMA-WUXFPT is in the final status “Won”, has an expected OI over €1,000. It has been created in the first half of FY 2015 and was “Closed” in the seconde half of FY 2015, so this OPP is included in our analysis included.

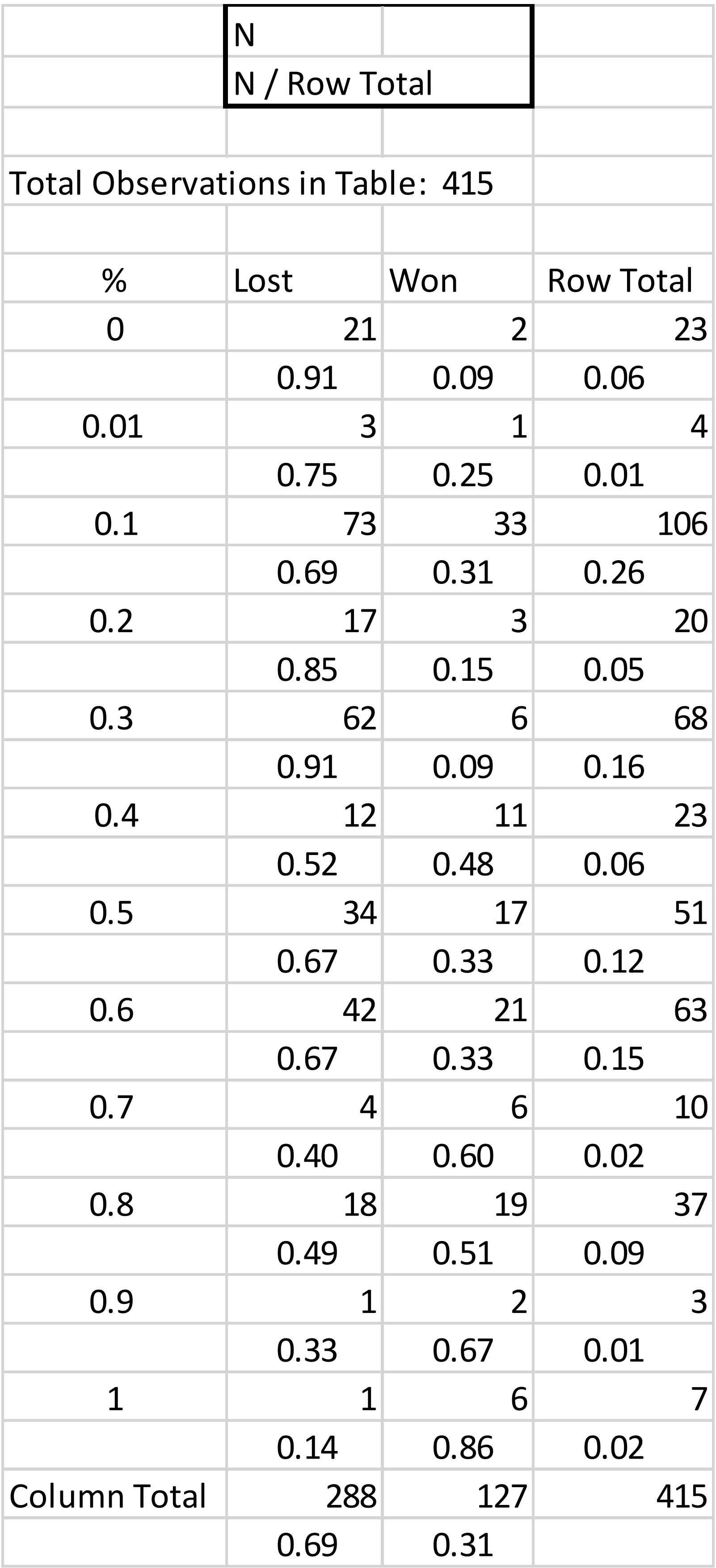
- OPP with the ID AEMA-10DNLSD is in the final status “Open”, has an expected OI over €1,000, but has been created in the seconde half of FY 2015. So it didn’t come over MFY 2015. That’s why this OPP is not included in our analysis.

- OPP with the ID AEMA-RE3HST is in the final status “Lost” but has an expected OI under €1,000. So this OPP is not included in our analysis.

4 Data analysis

4.1 Revenue Probabilities at MFY

In a first step, we are going to check whether the subjective probabilities assigned by the salesmen to the OPPs according to their experience correspond to the observed share of successful OPPs.



N
N / Row Total

Total Observations in Table: 415

%	Lost	Won	Row Total
0	21	2	23
	0.91	0.09	0.06
0.01	3	1	4
	0.75	0.25	0.01
0.1	73	33	106
	0.69	0.31	0.26
0.2	17	3	20
	0.85	0.15	0.05
0.3	62	6	68
	0.91	0.09	0.16
0.4	12	11	23
	0.52	0.48	0.06
0.5	34	17	51
	0.67	0.33	0.12
0.6	42	21	63
	0.67	0.33	0.15
0.7	4	6	10
	0.40	0.60	0.02
0.8	18	19	37
	0.49	0.51	0.09
0.9	1	2	3
	0.33	0.67	0.01
1	1	6	7
	0.14	0.86	0.02
Column Total	288	127	415
	0.69	0.31	

Table 4: Contingency table of revenue probability times “Lost”/”Won”

Table 4 shows the contingency table revenue probabilities at mid year times Won/Lost OPPs within the second half of the fiscal year. The OPPs are considered only for the fiscal year, which is assigned to them.

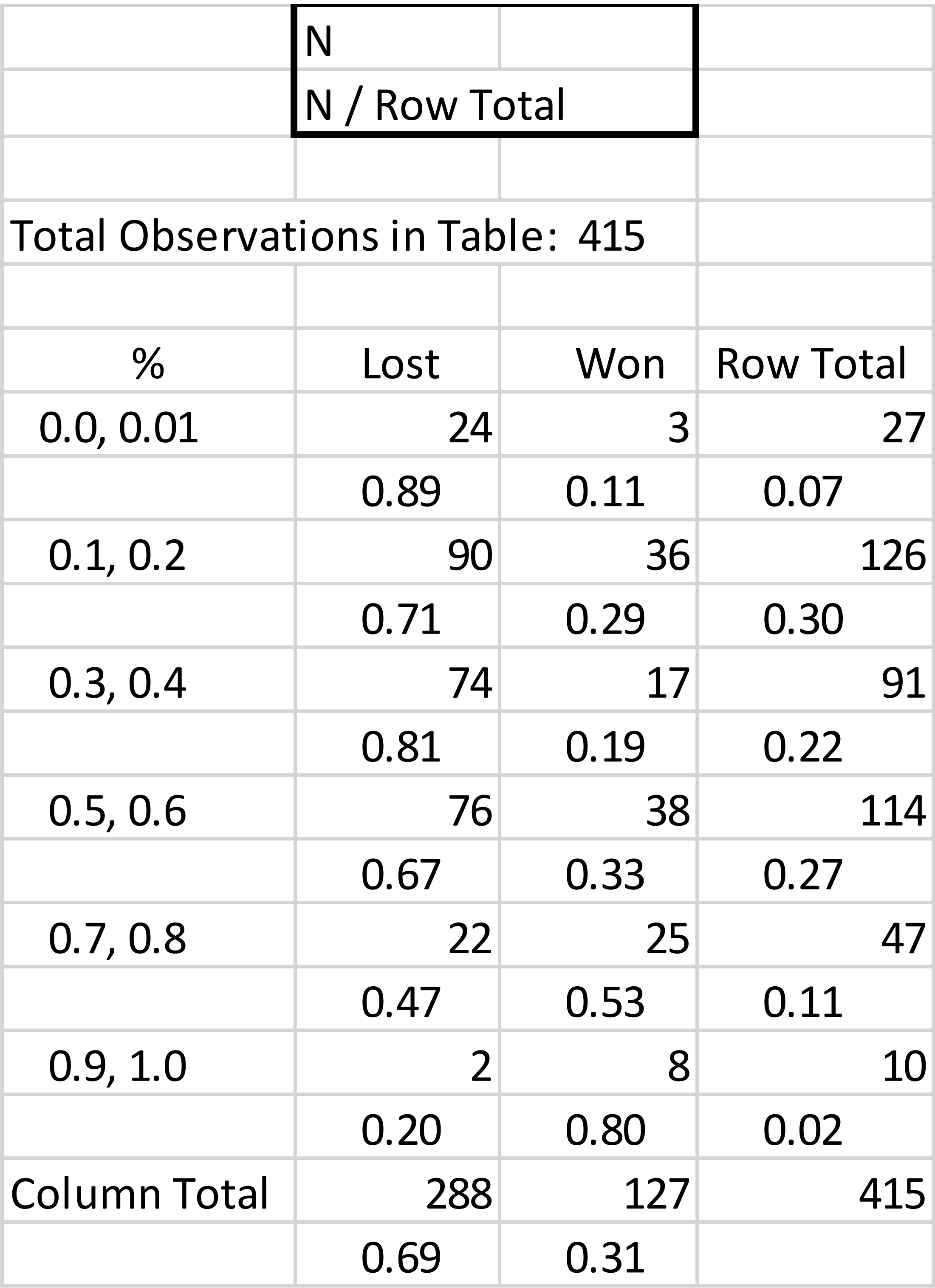
Actually, this table is the aggregation of the tables calculated only for OPPs ending within the fiscal year 2012 and the table of those ending 2013, etc. In total there are 415 OPPs within the period 2012 to 2015. 69% of them are lost and 31% are won.

The rows are labeled with the subjective probability categories provided by the sales men. The table gives for each category how many OPP were lost and how many were won, together with the row percentages. E.g. for OPPs in the 50% and 60% categories, the observed shares of won OPPs are both 33% of 51 OPPs and 63 OPPs. This indicates that the revenue probabilities are much larger in these cases, so that they overestimate the observed shares. For the lowest categories, the observed shares are underestimated. For the categories above 0.7 we find again a clear overestimation.

We are interested to find out whether the revenue probabilities are helpful for predicting the observed share of won OPPs. At first we test for the independence between both variables using the χ^2 contingency table test. As this test requires a minimum of elements in each cell, we aggregate the revenue probabilities to 6 categories and obtain Table 5.

However we can not manage to get at least four observations for every cell. E.g. (0%, 1%) won with a sample size of 30% and (90%, 100%) lost with a sample size of two. Table 5 looks from the perspective of distribution of the observed probabilities, better and is easier to read than Table 4.

Still we can monitor some strange behavior in the observed probabilities at the subjective probability category (30%, 40%). With an average of 35% we have an observed value of 19%. Compared to the subjective category (10%, 20%) with an average of 15% the observed value with 29% is much higher. In conclusion, it looks as if the sales men overestimate the actual probabilities for the first two subjective probability categories (0, 0.01) and (0.1,0.2), while they underestimate them for 30% onwards, see Figure 7. In the next step we check for independence between subjective probabilities and the “Lost”/”Won” status, to see whether there is any relationship between both variables.



	N		
	N / Row Total		
Total Observations in Table: 415			
%	Lost	Won	Row Total
0.0, 0.01	24	3	27
	0.89	0.11	0.07
0.1, 0.2	90	36	126
	0.71	0.29	0.30
0.3, 0.4	74	17	91
	0.81	0.19	0.22
0.5, 0.6	76	38	114
	0.67	0.33	0.27
0.7, 0.8	22	25	47
	0.47	0.53	0.11
0.9, 1.0	2	8	10
	0.20	0.80	0.02
Column Total	288	127	415
	0.69	0.31	

Table 5: Contingency table of Revenue Probability times (Lost/Won)

Test: Chi-square test for independence

H_0: The revenue probability at MFY and “Won”/”Lost” of OPP at EFY are independent

H_1: Both are not independent.

Result: Equation 1 based on Table 5 shows that hypothesis H_0 of independence is rejected. The p-value is very low (<<1%), so there is no relationship between both variables.

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  34.34683     d.f. =  5     p =  2.030881e-06
```

Equation 1: Chi-squared test for Table 3

In the following, the question is to find out to what extent do the sales men approximate the share of won OPPs?

Thereby we try to explain the observed probabilities by subjective ones of the sales men. So we regress the observed probabilities on the probability categories using the data from Table 4. Equation 2 gives the result. The number of observation is 12.

```
Obs_prob(i) = 0.807 * prob(i) + u-hat(i)   R2 = 0.914
             (0.075)
```

Equation 2: Regression, N=12, Standard error in brackets

The multiple coefficient of determination, R^2, is 0.914, which is quite high. The assigned probabilities correlate highly with the observed probabilities. The coefficient of the revenue probabilities is 0.807, which is smaller than 1. The value is highly significantly different from zero as its t-value is 0.807/0.075 = 10.76. The p-value with 11 df (degree of freedom) is very small. Its corresponding t-value is (0.807 – 1.0)/0.075 = -2.573 with a p-value of 0.023. Besides with a significant level of 5% it is significantly smaller than 1.

The coefficient of 0.807 indicates that the probabilities of the sales men overestimate the observed probabilities. The model unfortunately does not distinguish between the low probabilities, which are underestimated, and the larger ones, which are overestimated.

4.2 OI: Mid Fiscal year and End Fiscal year

4.2.1 Expected FY Volume

The expected volume at EFY is based on the expected OI data at MFY and is calculated as a weighted mean. Equation 3 tells us the sum of all OPPs at MFY weighted with the revenue probabilities for each FY of 2012-2015. This corporate rule is used for forecasting the EFY volume.

$$\text{Expected FY Volume} = \sum \text{Revenue prob}(\text{OPP})*\text{Estimated OI}(\text{OPP})$$

$$= 168.712.641$$

Equation 3: Expected EFY volume

Equation 3 results For the FYs 2012 – 2015 into €168,712.641. We compare this value with the sum of all won OPPs at EFY 2012-2015 and find a value of €53,178.551. In conclusion, the share of final realized volume is only 31.5% of the forecasted.

4.2.2 Distribution of Volume at MFY and EFY

Figure 2 compares the histograms of OIs of all OPPs observed at MFY with the volumes of the won OPPs at EFY. “Vol mid yr” stands for all OPP volumes status at MFY. “Vol fisc yr” stands for all OPP volumes at EFY. Both distributions look somewhat similar, although the first is skewed to the right, the second skewed to the left. Especially at the right end, the area of the high volume OPPs is much smaller in the second histogram. It seems that approximately more than 30% of the high volume OPPs are lost.

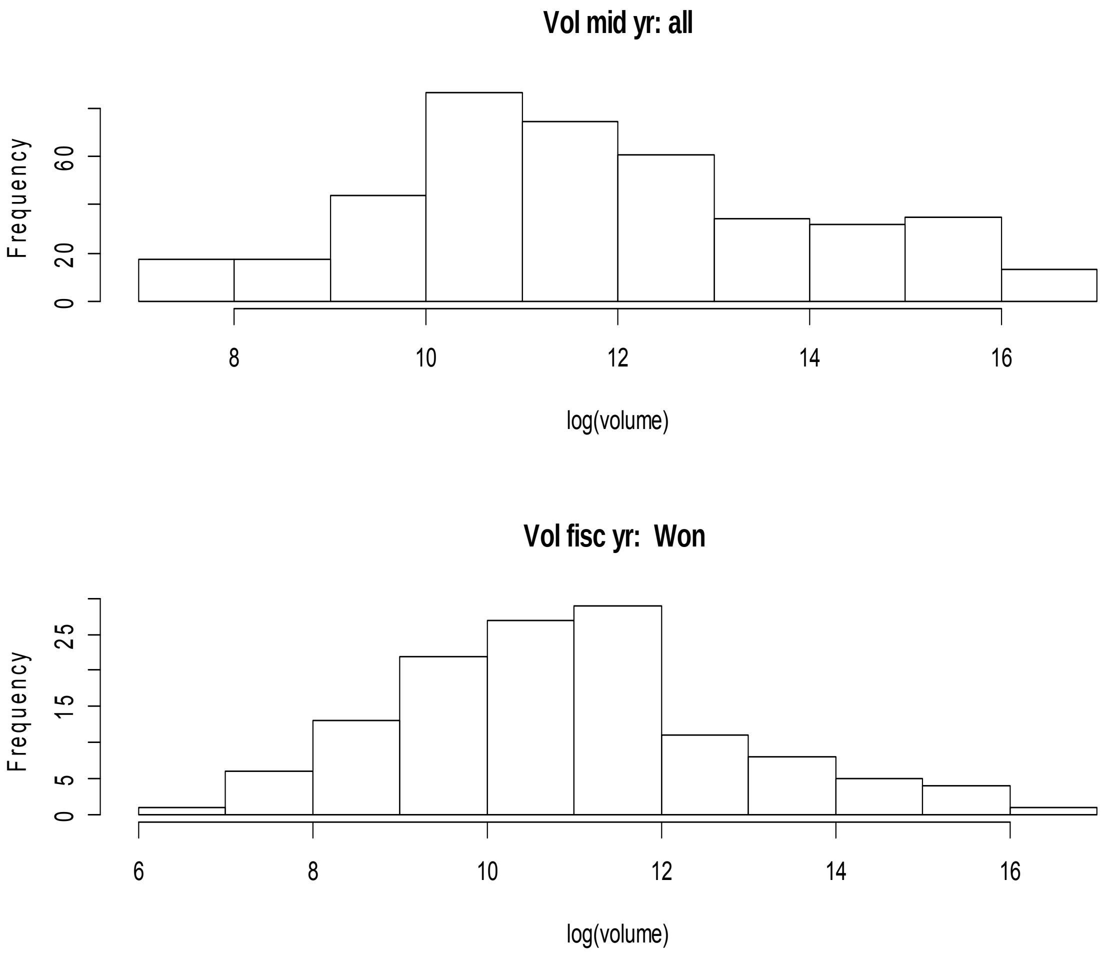


Figure 2: Histograms of OPP log(volume) MFY and OPP log(volume) won at EFY

In order to get an interpretable visualization the log of the volume has to be taken. We also perform a statistical test to confirm our visual impression.

The hypotheses of the two sample Kolmogorov-Smirnov tests are

H_0: Both distributions are identical.

H_1: The distributions differ.

```
Test Results:
  STATISTIC:
      D | Two Sided: 0.1921
    D^- | Less: 0.1921
    D^+ | Greater: 7e-04
P VALUE:
Alternative Two-Sided: 0.001276
Alternative Exact Two-Sided: 0.001276
Alternative Less: 0.0006378
Alternative Greater: 0.9999
```

Table 6: MFY versus EFY Kolmogorov-Smirnov two sample test

Result: The test rejects the hypothesis that both distributions are identical with a p=0.0013, see Table 6, the previous visual impression is confirmed.

Below we try to explain this discrepancy by the uncertainty which of the OPPs belonging to one category are realized, and by the difference of the subjective and observed probabilities.

4.2.3 Distribution of MFY and EFY Volume by Probability

Further, we look at the distributions of the prospective volumes for each probability class at MFY and the won volumes at the EFY, see Figure 3. Also here because of a the behaviour of the OPP volumes the log(volumes) are used to smooth the graphs and make the graphs comparable with each other.

We find that the distributions of different probability categories are not so much different from each other.

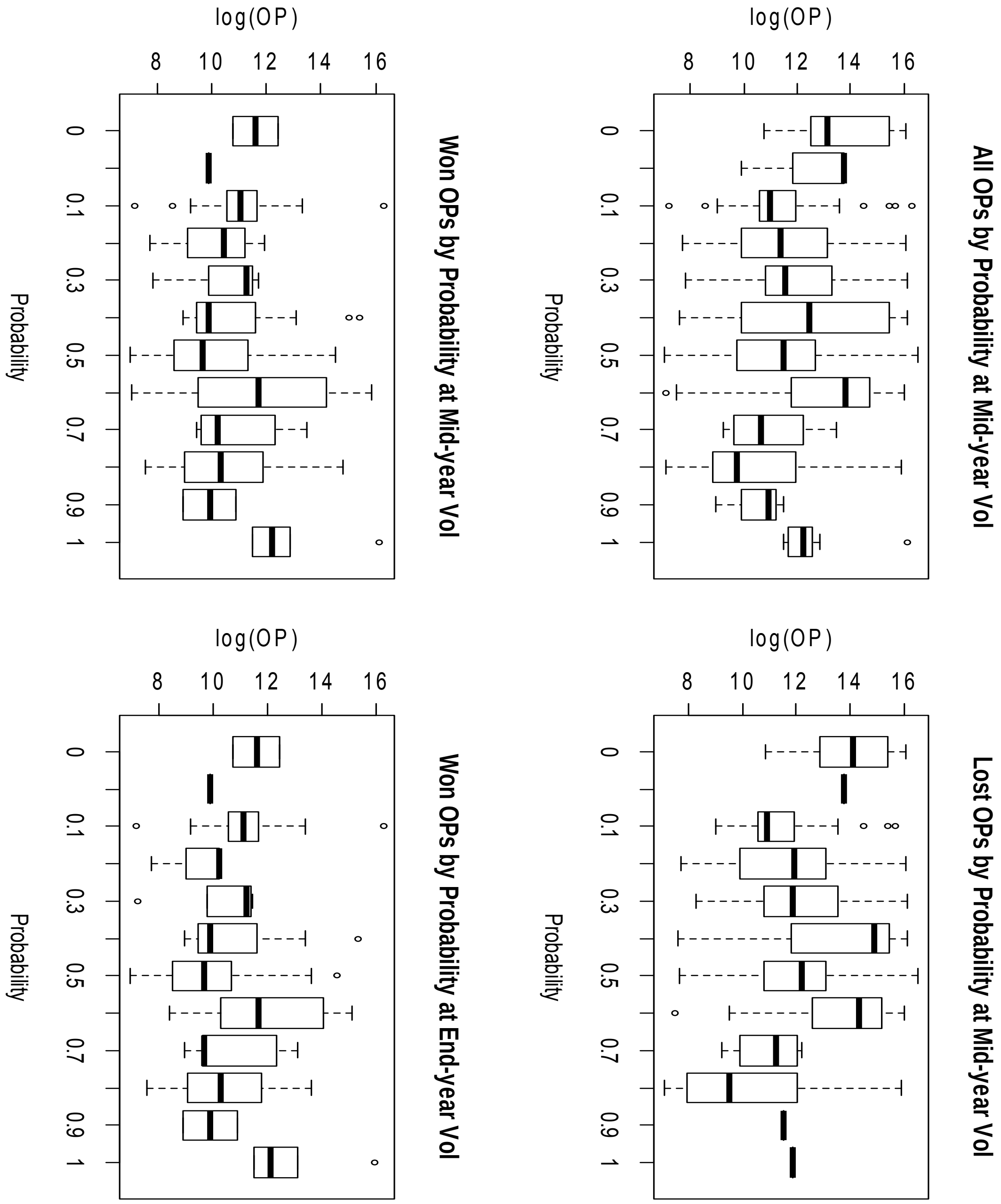


Figure 3: Boxplots for log(OPP Volume) for Revenue Probability categories

4.3 VaR of won volumes at EFY based on subjective probabilities at MFY

4.3.1 Usage of the subjective probabilities

We are interested in the effect of uncertainty in the expected total volume at EFY, as it is not known in advance which of the OPPs in each category will be won. Only the proportion of the number of realized OPPs is assumed to be known, i.e. the probability that is assigned by the sales men to each OPP.

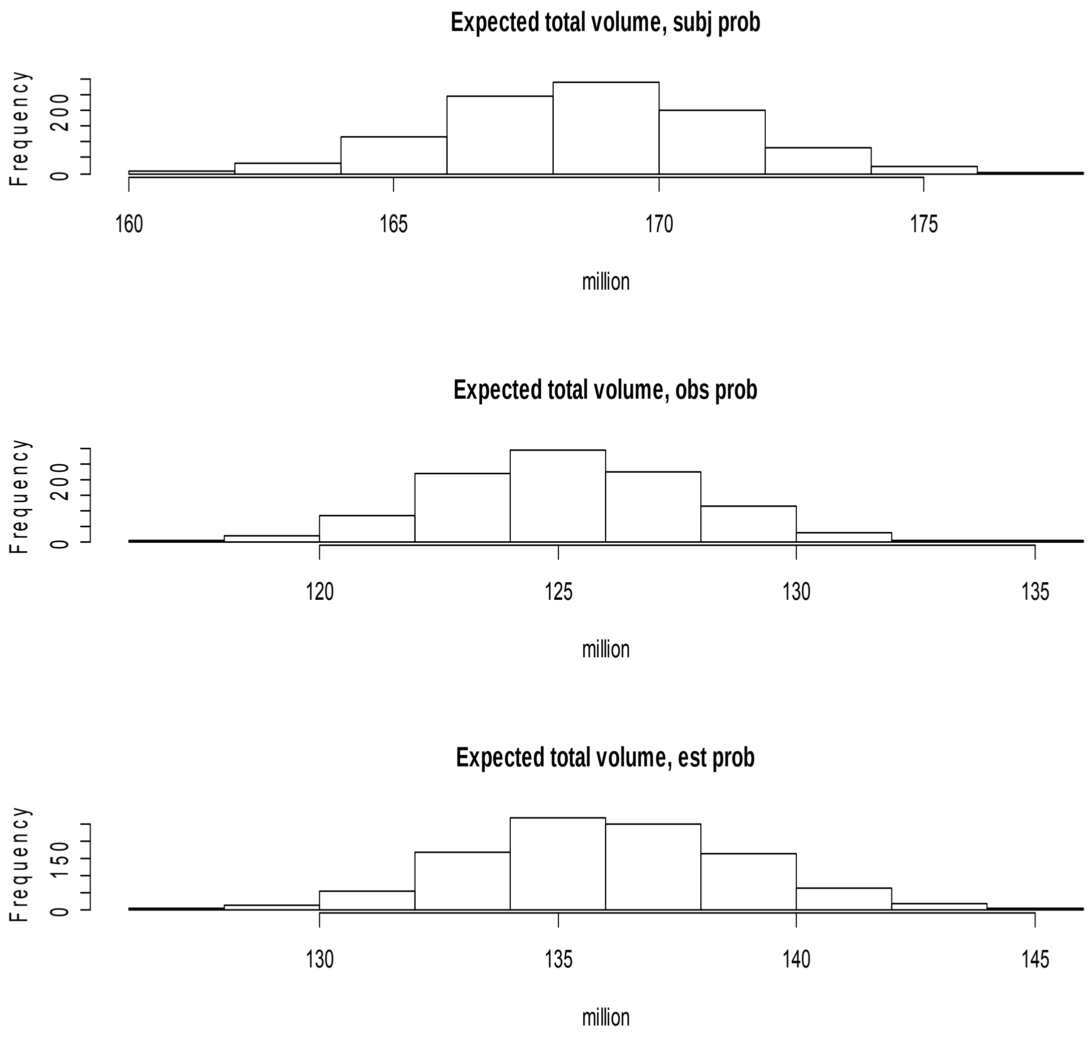
This is accomplished by resampling out of observed distribution of volumes at MYR for each category, choose the OPPs according to the assigned probability and sum those volumes. The assumptions are:

1. The assigned probability is correct.
2. The number of OPPs in each category is known.
3. The OPP volumes in each category are known.
4. Unknown is which of the OPPs within each probability category will be realized (only the proportion of OPPs within each category is known)

Technically, we resample from a population where the observed volume values are repeated 50 times. This guarantees that the number of OPPs to be taken is integer. E.g. for the category 1% we have four different OPPs, so we have to choose 2 out of each resampled data.

The results are pictured by a histogram, given in the upper part of Figure 4.

“subj prob” stands for subjective probability and takes the probability provided by the sales men. “obs prob” are the observed probabilities which are given in Table 4. The third the “est prob” = estimated probability uses the probabilities given by the regression model, Equation 2, i.e. the subjective probabilities * 0.0807. We can see in Figure 4 that the shape of the histograms look almost the same. But the mean of the estimated and observed probabilities shift a lot to left compared to the subjective probabilities, further the variance shrinks too.



Expected total volume, subj prob

Expected total volume, obs prob

Expected total volume, est prob

Figure 4: Histograms of estimated OIs at EFY, 3 methods

For further interpretation, we compare the differences between the means. With the usage of subjective probabilities, we obtain an average total volume of €168,680.171, see Table 7. This is very close to €168,712.641 from Equation 3 above. The 10% quantile, the 10% VaR, value-at-risk, is €165,334.398, which is only 2% below the mean. This indicates that the uncertainty in the different volumes cannot explain the low observed EFY value.

```
Basic statistics:
Nobs=           1000
+NAs=              0
Min=            158794363.38860
Max=            175740415.22500
Mean=           168680171.25024
Median=         168758955.58590
StDev=          2668909.56415
Trim05 Mean=    168689428.15170
Skewness=     -0.09750
Kurtosis=      3.09146
Jarque-Bera=   1.93305
p-value=       0.38040
quantile(sum_rs,0.10) = 165334398 at 10%
quantile(sum_rs,0.90) = 171979674 at 90%
```

Table 7: Weighted sums obtained by resampling: subjective probability

4.3.2 Usage of the observed probabilities

As the subjective probabilities diverge from the observed shares considerably, we calculate the distribution of the total volumes at EFY as if we knew the realized shares in each category.

The histogram of the resampling results is the second in Figure 4. Compared to the first the distribution shifts to the left, by approximately €43,476.744, see Table 8, to a mean of €125,203.427 but is still far away from the actual realized volume of €53,178.551.

```
Basic statistics:
Nobs=             1000
+NAs=                0
Min=            116740905.05440
Max=            134644821.70720
Mean=           125203427.33394
Median=         125206599.95350
StDev=          2646691.17313
Trim05 Mean=    125192722.22119
Skewness=      0.04271
Kurtosis=      3.01909
Jarque-Bera=   0.31916
p-value=       0.85250
quantile(esum_rs,0.10)= 121752338 at 10%
quantile(esum_rs,0.90)= 128606004 at 90%
```

Table 8: Weighted sums obtained by resampling: observed probability

4.3.3 Usage of the estimated probabilities

Finally, we use the estimated relation between subjective and observed shares within each category. I.e. we multiply the subjective probabilities by 0.807 and calculate the distribution of the expected OIs at EFY. In this case, the mean is €136,069.331 but also still far away from the actual €53,178.551.

The third histogram in Figure 4 again shifts a bit back to the right compared to the second, as expected.

```
Basic statistics:
Nobs=             1000
+NAs=                0
Min=            129018625.09380
Max=            145075953.25400
Mean=           136056993.56594
Median=         136069331.68370
StDev=          2620854.03262
Trim05 Mean=    136048742.70945
Skewness=      0.08176
Kurtosis=      2.95576
Jarque-Bera=   1.19581
p-value=       0.54996
quantile(esum_rs,0.10)= 132722895 at 10%
quantile(esum_rs,0.90)= 139457371 at 90%
```

Table 9: Weighted sums obtained by resampling: estimated probability

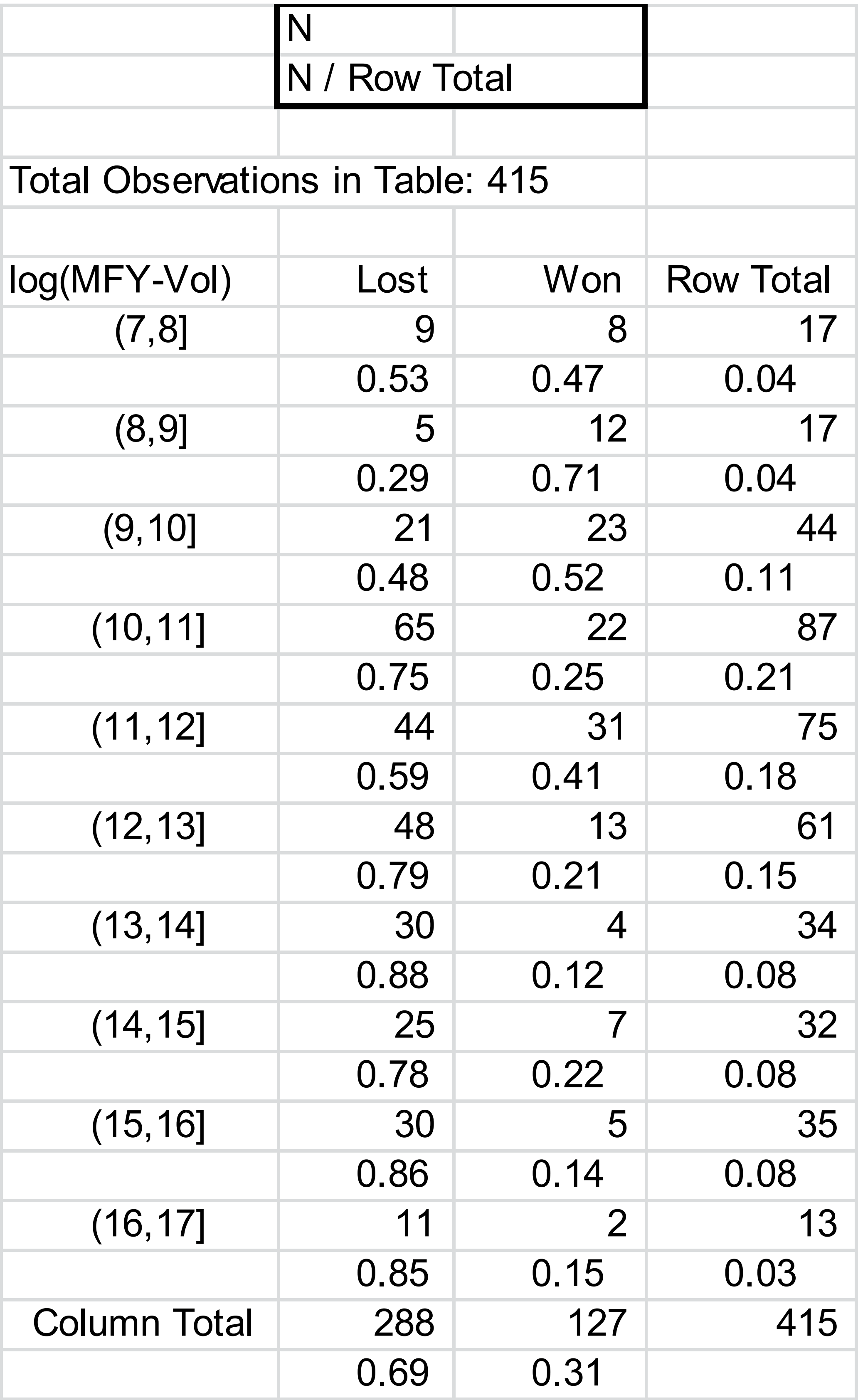
4.4 The role of Mid Fiscal Year Volume

4.4.1 Relation of volume and Won/Lost

In conclusion, there is still a tremendous gap between expected and realized total volume, so we add the volume to our analysis to find maybe dependences.

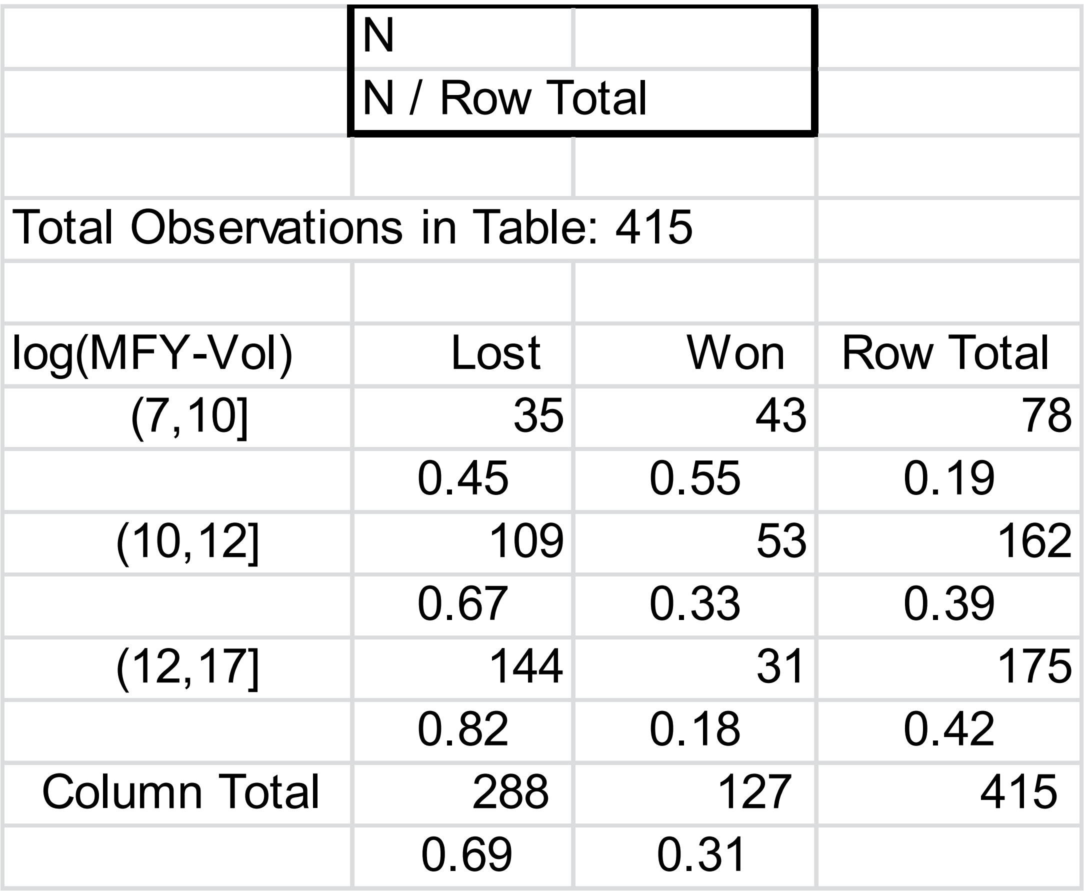
The contingency table in Table 10, log(MFY volume) x Lost/Won, shows some dependence between log(volume) measured at MFY and whether the OI is successful or not.

Especially small OIs tend to have a considerably higher success probability, than the largest ones. To make the relationship clearer we aggregate similar categories. See Table 11.



	N		
	N / Row Total		
Total Observations in Table: 415			
log(MFY-Vol)	Lost	Won	Row Total
(7,8]	9	8	17
	0.53	0.47	0.04
(8,9]	5	12	17
	0.29	0.71	0.04
(9,10]	21	23	44
	0.48	0.52	0.11
(10,11]	65	22	87
	0.75	0.25	0.21
(11,12]	44	31	75
	0.59	0.41	0.18
(12,13]	48	13	61
	0.79	0.21	0.15
(13,14]	30	4	34
	0.88	0.12	0.08
(14,15]	25	7	32
	0.78	0.22	0.08
(15,16]	30	5	35
	0.86	0.14	0.08
(16,17]	11	2	13
	0.85	0.15	0.03
Column Total	288	127	415
	0.69	0.31	

Table 10: Crosstabs with 10 size categories



	N		
	N / Row Total		
Total Observations in Table: 415			
log(MFY-Vol)	Lost	Won	Row Total
(7,10]	35	43	78
	0.45	0.55	0.19
(10,12]	109	53	162
	0.67	0.33	0.39
(12,17]	144	31	175
	0.82	0.18	0.42
Column Total	288	127	415
	0.69	0.31	

Table 11: Crosstabs with 3 size categories

Table 11 shows that the lowest 19% of the OIs have a success probability of 55% instead of 31%. In the middle, the probability of 33% is approximately correct, for the highest 42% OIs only 18% succeeded.

The expected sum of successful OIs used in combination with the success probabilities given in Table 10 yields €71,159.852, see Equation 4. This is only 33,8% too large compared to the actual realized value of €53,178.551.

```
md-yr-vols  weighted with obs probs for log(size OI)
categories
sum(tab2a[,3]*tab2a[,8])
71,159.852 (+33.8%, *1.33)
```

Equation 4: Total Volume with observed probabilities of 10 size categories without share

4.4.2 The relationship between MFY and successful EFY Volume

Now we look at the relationship between MFY OPP and realized EFY OPP. First results are summarized in Table 12. It shows no systematic deviation of the ratio of EFY to MFY volumes for won OIs. The resulting mean share is hardly different from 1 and the t-test for the ratio equal to 1 can not reject the null hypothesis.

```
mean(share) is 1
share <- dtab12[,4] / dtab12[,3]
basic_stats(share)
Basic statistics:
Nobs=              127
+NAs=                0
Min=           0.07532
Max=           3.71212
Mean=          0.98746
Median=        1.00000
StDev=         0.47059
Trim05 Mean=   0.94816
Skewness=      3.06173
Kurtosis=     19.14409
Jarque-Bera= 1577.59728
p-value=       0.00000
t_mean1 <- (0.98746 - 1)/sqrt(0.47059^2/126)
t = -0.299
```

Table 12: Ratios of MFY and EFY Volume

There is a single outlier with a share of 3.7. By correcting the one outlier the result does not differ.

Alternatively, a log-log relationship between the volumes is considered. A more refined model for the relationship between both volumes is in logarithm. The estimated model is given in Equation 5. It turns out that the elasticity is significantly smaller than 1, even 0.931.

```
log(EFY-Vol) = 0.625 + 0.931*log(MFY-Vol)+ v-hat   R2=0.937
              (-0.243)(0.022)
```

Equation 5: Relationship between log(EFY-vol) & log(MFY-Vol)

The corresponding confidence interval is a lot smaller than

$[0.931 \pm 1.96]$*$0.022 = [$-$0.022638, 0.06302]$.

The log-log model yields a better result, as the assumption that the proposed volume is equal to the realized one at EFY, see Table 20.

```
The approximation for the share (EFY-Vol / MFY-Vol) is
 share = 1.867405*MFY-Vol^(-0.069)*exp(0.4876^2/2)
```

Equation 6: Calculation of the share

The model expressed in the variable share is given in Equation 6. There the term exp(0,4876^2/2) corrects the expected value as the assumed normal distributed regression errors are transformed by exp(). The correction factor is exp(sigma^2_residuals/2). This comes from the formula of the expected value of the log normal distribution. Note that in Equation 6, if the realized MFY volume increases the realized share decreases.

If we consider that, the expected realized volume per OI is very close to the true value, using the 10 MFY size categories, see Table 13. This indicates that the subjective success probabilities should depend in an essential way on the volume size.

```
MFY-Vol * mod_share weighted with obs probs for log(size OI)
10 categories
sum(tab2a[,3]*mod_share*tab2a[,8])
53,654.639 (+0.9%)
```

Table 13: Total Volume with observed probabilities of 10 size categories with share

If we consider only 3 size categories the expected value is 13.6 % too large, see Table 14.

```
MFY-Vol * mod_share weighted with obs probs for log(size OI)
3! categories
sum(tab2a[,3]*mod_share*tab2a[,11])
60,418.125 (+13.6%)
```

Table 14: Total Volume with observed probabilities of 3 size categories with share

Table 14 however, is the more preferable model, because it has less volume categories and so will likely provide a smoother forecast output. 10 categories might be criticized for “overfitting”.

4.4.3 Relation of Volume and the assignment of OIs to a subjective probability category

To find a relationship between the volume and the subjective probability we first look at the correlation between the assigned probability of OIs and the MFY-logarithm volume to find out whether the assignment of an OI to a subjective probability category depends on its size. However, the correlation between logarithm volume and subjective probabilities is -0.052, see Table 15. In the following, the correlation coefficient is essentially zero. The correlation coefficient is an indicator for a monotonic relationship only. Table 15 give the results for the test of the correlation coefficient of zero. The hypothesis cannot be rejected with a p-value of 0.294.

```
cor(tab2a[,7],log(tab2a[,3]))
p= 0.294
ρ= -0.05159005
```

Table 15: Correlation between subjective probability and log(volume)

If independency between both variables is tested, independence however, is clearly rejected, because the p-value is <<1%, see Table 16. In conclusion, there is a rejection of independency although there is no correlation between volume and subjective probability.

```
Chi^2 =  114.1203     d.f. =  10     p =  7.855855e-20
```

Table 16: Chi squared test

		N		
		N / Column Total		
		N / Table Total		
subj. prob	log(MFY-Vol) (7,10]	log(MFY-Vol) (10,12]	log(MFY-Vol) (12,17]	Row Total
(0.0, 0.01)	1,00	4,00	22,00	27,00
				0.07
	0.01	0.02	0.13	
(0.1, 0.2)	11,00	85,00	30,00	126,00
				0.30
	0.14	0.52	0.17	
(0.3, 0.4)	16,00	34,00	41,00	91,00
				0.22
	0.21	0.21	0.23	
(0.5, 0.6)	24,00	24,00	66,00	114,00
				0.27
	0.31	0.15	0.38	
(0.7, 0.8)	25,00	10,00	12,00	47,00
				0.11
	0.32	0.06	0.07	
(0.9, 1.0)	1,00	5,00	4,00	10,00
				0.02
	0.01	0.03	0.02	
Column Total	78,00	162,00	175,00	415,00
	0.19	0.39	0.42	
Avg. subj. prob	0,51	0,31	0,39	

Table 17: Crosstab MFY subjective probabilities x MFY volumes

Table 17 gives in the last line the average subjective probability for 3 OI size classes. The first two values reflect very closely the observed probabilities for the small and medium size OIs. However, for large OIs, that is above €162,000 $(\sim e^{12})$ the average subjective probability is 0.39 while only 0.18 are observed. This is a large overestimation by a factor of 2.

According to this most of the opportunities in Table 17 should be within the subjective probability category (0.1, 0.2). However, only 17% of this volume category lays in this probability category and the figures tend to point to an overweighting of opportunities with a big volume. This is also proven by the average subjective probability of only 39%, which lays a lot over 18%.

On the other hand, for opportunities with logarithm volume from 7 to 10 the probabilities seem to be correct distributed. In Table 11 you can find a win rate of 55% for this volume category. In Table 17 31% of the small opportunities lays in the probability category (0.5, 0.6).

Nevertheless, all in one the average subjective probability is according to Table 17 51%, which is smaller than the 55% and indicate a little under weighting of opportunities. However, the middle category-logarithm-volumes from 10 to 12 have according to Table 11 a win rate of 33% resulting in a correct distribution of 22% in probability category (0.3, 0.4). Here an average subjective probability is calculated of 31%, which is also tending into the direction of under weighting.

4.4.4 VaR model of expected total Volumes based on MFY Volumes

Under the assumption, the win probability of each size category is known, at MFY we obtain the VaR by resampling in each size category. Two histograms are created. Figure 5 depicts the volume distributions for the cases of 3 and 10 size categories. The corresponding 10% VaR-values are in Table 18 and 19.

The 10 size category histogram has nearly no variance because the number of possible variation in each size category is very small. 10 size categories seem to be too many and generate an overfitting effect, see Table 19. The standard deviation is very small compared to Table 18.

```
Basic statistics:
Nobs=             1000
+NAs=                0
Min=           56662549.27320
Max=           64432611.97435
Mean=          60424133.38112
Median=        60392439.34704
StDev=         1086109.00499
Trim05 Mean=   60414094.20186
Skewness=      0.13502
Kurtosis=      3.28157
Jarque-Bera=   6.34204
p-value=       0.04196
quantile(v3_sum_rs,0.10)= 59003364 at 10%
quantile(v3_sum_rs,0.90)= 61741653 at 90%
```

Table 18: Weighted sums obtained by resampling with estimated probabilities categorized in 3 size categories

```
Basic statistics:
Nobs=             1000
+NAs=                0
Min=           52988016.04247
Max=           54510264.23834
Mean=          53647697.25981
Median=        53652935.86941
StDev=         227377.01985
Trim05 Mean=   53645517.55617
Skewness=      0.13096
Kurtosis=      3.20413
Jarque-Bera=   4.59458
p-value=       0.10053
quantile(v10_sum_rs,0.10)= 53366789 at 10%
quantile(v10_sum_rs,0.90)= 53916344 at 90%
```

Table 19: Weighted sums obtained by resampling with estimated probabilities categorized in 10 size categories

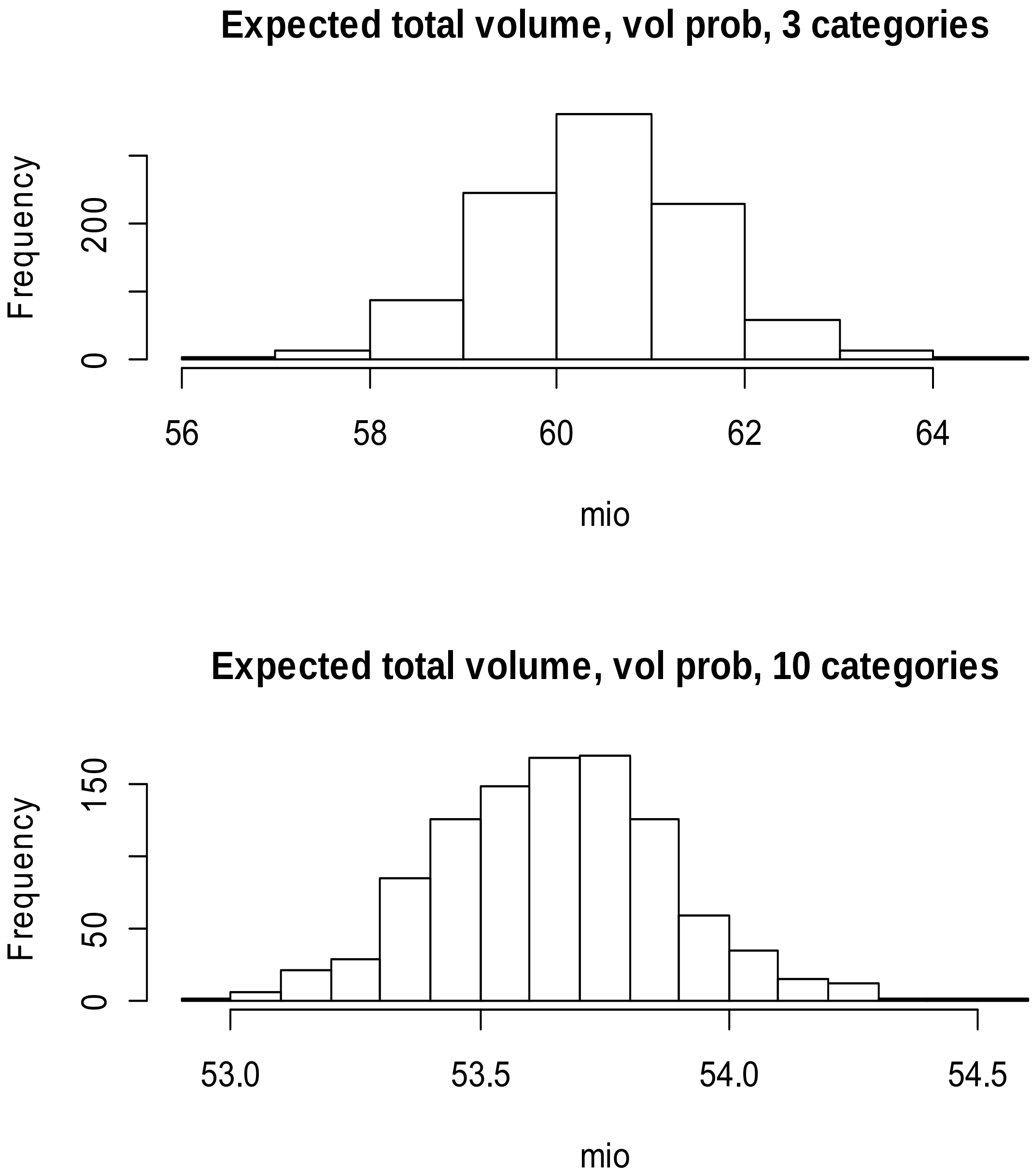


Figure 5: VaR for 10 & 3 size categories

The distribution for 3 volume categories, does not cover the observed total EFY volume, but is not far away, see Table 18. This is into sharp contrast to the usage of the subjective probabilities which we had expected to work reasonably well before the project was started. Some more refinements are necessary to obtain a reasonable forecasting model.

5 Reasons for the deviation of subjective and objective probabilities

5.1 Lottery effect

In settings where risky decisions have to be made many people favour riskier options which offer a small probability of large gains, that is, where the distribution of payoffs has positive skew. Some examples are betting on long-shot horses where the odds are very low to win but on the other hand the skew much higher is compared to the favourites with the greatest expected return. (Golec & Tamarkin, 1998)

Similar habits also appear when people are buying lottery tickets, (Garrett & Sobel, 1999) and (Forrest, Simmons, & Chesters, 2002) outcome is, that people are more focused on the size of the top prize than the chance of success and finally the estimated payback of the lottery. Positive skew also has an impact on economic choices besides gambling.

Hamilton (2000) shows that three quarter of all people who enter self-employment face higher variance and skew but on the other hand lower expected return than in employment. Also 97% of inventors will not break even on their investments, but face a very skew distribution of return conditional on succeeding (Åstebro & Rotman, 2003).

However, one of the main factors, which brings in the gambling factor are the sales man contracts. In most sales man contracts the sales volume plays a big role in his incentive and further in his overall payment. In fact, one of the most popular key performance indicator (KPI) of a sales man is the overall volume he sells.

As a result, it is in his interest is to win, as many opportunities as possible and the best thing would be they all should have a high volume. However, working time is limited. So he needs to weight the time he is spending on an opportunity to win it. Mostly big projects consume less time than small projects, which have together the same volume as the big project. As an example, a salesman has very high volume opportunity, which has a low chance of success.

The best example is Lotto. In this case, very small winning chances going together with very big volumes are going to be overestimated. Åstebro investigated this skew seeking behaviour and found that most individuals are prudent and make skew seeking choices. “We also find evidence on skew seeking choices, as subjects in our experiment make riskier choices when lotteries display greater positive skew”, (Åstebro, Mata, & Santos-Pinto, 2015).

Nevertheless, in our case, the lottery top prize is the high volume opportunity and the money we would spend on the lottery ticket is the time we spend for the opportunity. As a result, our sales man acts the same way like people in the study and overestimate the opportunity to win this big project.

However, how can one explain these choices favouring options with positive skew, high risk and low expected return (Åstebro, Mata, & Santos-Pinto, 2015)? Maybe this incentive-based contracts combined with the nature of a sales man to sell as many as possible are reasons for this misestimating of probabilities habit.

Additionally, to the lottery example in the test candidates were not driven by love for risk, but rather by optimism and likelihood insensitivity. By being optimistic, the test candidates overweight the probability of getting larger prizes and underweight the probability of getting lower prizes regardless of the probabilities of the prizes. So in conclusion small volumes are going to be underestimated and very big volumes are going to be overestimated because also of a skewness seeking behaviour. (Åstebro, Mata, & Santos-Pinto, 2015)

In chapter 4.4.3 this behaviour has been shown, low volume opportunities are slightly underweighted and high volume opportunities are overweighed.

However, in the end a we find a pattern of the volumes in combination with the subjective probabilities from the salesmen. Volumes and subjective probability categories are not dependent from each other. However, with a correlation of (Volume, subjective probabilities) = -0.052, see Table 15, there is hardly any linear relationship between volume of an opportunity and the subjective probability provided by the salesmen.

5.2 Motivational Biases

Motivational biases can affect estimates and forecasts whenever estimators believe that the quantities expressed may affect them personally. For example, managers may have an incentive to overstate productivity forecasts to reduce the risk that the capital dollars allocated to their business units will be reduced.

More subtle biases also affect estimates provided from managers, and the effect can depend on the individual. For example, project managers who are anxious to be perceived as successful may pad cost and schedule estimates to reduce the likelihood that they fail to achieve expectations.

On the other hand, project managers who want (consciously or unconsciously) to be regarded as high-performers may underestimate the required work and set unrealistic goals. Most managers are overly optimistic. When companies collect data on the financial returns from projects, they usually find that actual returns are well-below forecasted returns. Motivational biases can also cause people to minimize the uncertainty associated with the estimates that they provide.

For example, sometimes managers become defensive when asked to estimate the potential risks associated with a proposed project, even in environments where it is well known that projects can fail. They feel that admitting to downside potential would suggest deficient risk management practices or the fallibility of their project management skills.

Experts likewise face disincentives to fully acknowledging uncertainty. They may think that someone in their position is expected to know, with high certainty, what is likely to happen within their domains of expertise. We do, in fact, appear to value the opinions of highly confident individuals more highly. Studies show that consultants and others who sell advice are able to charge more when they express great confidence in their opinions, even when their forecasts are more often proven wrong (Radzevick & Moore, 2011).

As we previously salesmen have mostly incentive based contracts and are also a kind of project manager of their business opportunity, because they are full responsibility about their business. Maybe we can put a sales man on the same environment like a manager or project manager, so according to Radzevick (2011) they likely won’t admit, that business opportunities are looking bad and also don’t change their subjective probabilities within the system. This behaviour may lead to incorrect values in system.

Poorly structured incentives, obviously, can distort decisions as well as estimates. For example, any company that rewards good outcomes rather than good decisions motivates a project manager to escalate commitments to failing projects, since the slim chance of turning the project around is better from the manager’s perspective than the certainty of project failure. (Widemann, 2004)

5.3 Estimating and Forecasting Biases

It is already common known that people’s intuitive decisions are often strongly and systematically biased. The conclusion reached by Tversky is that people use unconscious shortcuts, termed heuristics, to help them make decisions. "In general, these heuristics are useful, but sometimes they lead to severe and systematic errors" (Tversky & Kahneman, 1987).

However, people are notoriously poor at estimating and forecasting as we as far the analysis showed. They ignore or do not correctly use probabilities when making choices and make overly optimistic forecasts that cannot be justified.

Studies show that people make systematic errors when estimating how likely uncertain events are. As shown in Figure 9, likely outcomes (above 40%) are typically estimated to be less probable than they really are. In addition, outcomes that are quite unlikely are typically estimated to be more probable than they are.

Furthermore, people often behave as if extremely unlikely, but still possible outcomes have no chance whatsoever of occurring. Figure 6 of Widemann (2004) however refers to the probabilities of an undesired event, like the sinking of the Titanic. In our core, the events are winning of an opportunity, so the s-shaped curve is mirrored around the diagonal. Two possible reason were found to additionally explain the systematic wrong probability estimations of the salesmen.

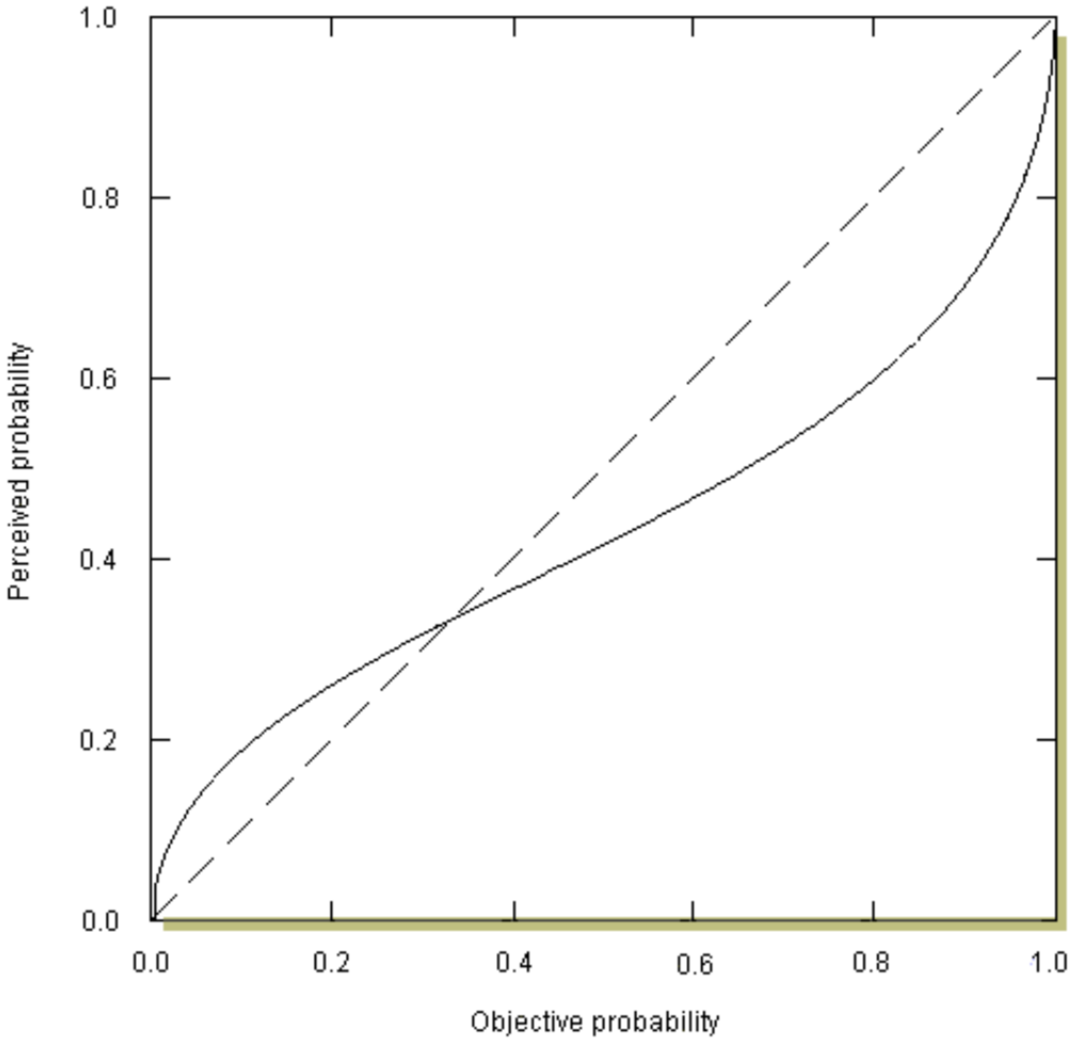


Figure 6: People systematically over- or under-estimate probabilities (Widemann, 2004)

5.3.1 Overconfidence

Overconfidence has been called, "perhaps the most robust finding in the psychology of judgment" (DeBondt & Thaler, 1995). In fact, we believe we are better at making estimation than we actually do. For illustration purposes Widemann introduced the so called "2/50 rule".

Test candidates are asked to provide 98% confidence intervals different uncertain quantities lie. Example questions were "What's the elevation of the highest mountain in Texas?" "Give me low and high values within which you are 98% sure that the actual values fall.” After the true values were revealed, up to 50% of the answers were outside of the specified confidence intervals. In conclusion, if people were not overconfident, values outside 98% confidence intervals would occur just 2% of the time. Popular phrases like the British mathematician Lord Kelvin said, “Heavier-than-air flying machines are impossible.” or Thomas Watson, founding chairperson of IBM, reportedly said, "I think there is a world market for about five computers." are underlining the previous statement (Widemann, 2004).

Overconfidence of the salesmen might be a reason explaining why the probabilities are biased.

5.3.2 Overoptimism

Overoptimism describes the human tendency to believe things are turning out more likely for the good than for the bad. However, optimism has been blamed for a variety of problems in decision-making. This includes also over estimating the likelihood of positive events and under-estimating the likelihood of negative events, as seen in Figure 6.

Economists believe the bias contributes to the creation of economic bubbles; during periods of rising prices, investors are overoptimistic about their investments. It has been suggested that in many cases of corporate disclosure fraud, the offending officers and directors were not consciously lying but instead were expressing honestly held but irrationally optimistic views of their firms condition and prospects. A related bias is wishful thinking, a tendency to focus on outcomes that are pleasing to imagine rather than what is objectively most likely. (Widemann, 2004)

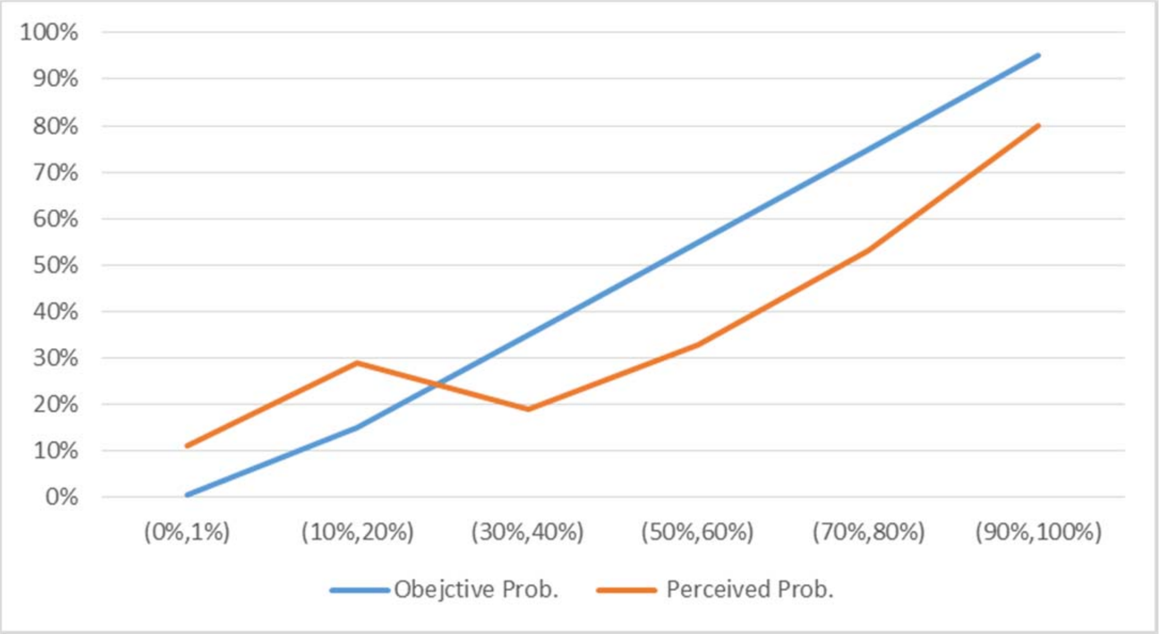


Figure 7: Subjective probability of opportunities to be won (subjective x actual prob.)

Figure 7 was made based on the data from Table 5. The curves in Figure 7 have obviously the same behaviour like in the literature in Figure 6. In conclusion, our sales obviously over and under estimating their opportunities.

6 Summary

CRM opportunity data from a corporation are analysed for patterns. These data is created from salesmen and provide business opportunity information with win probabilities and volumes. These data is created from salesmen and provide win probabilities and volumes.

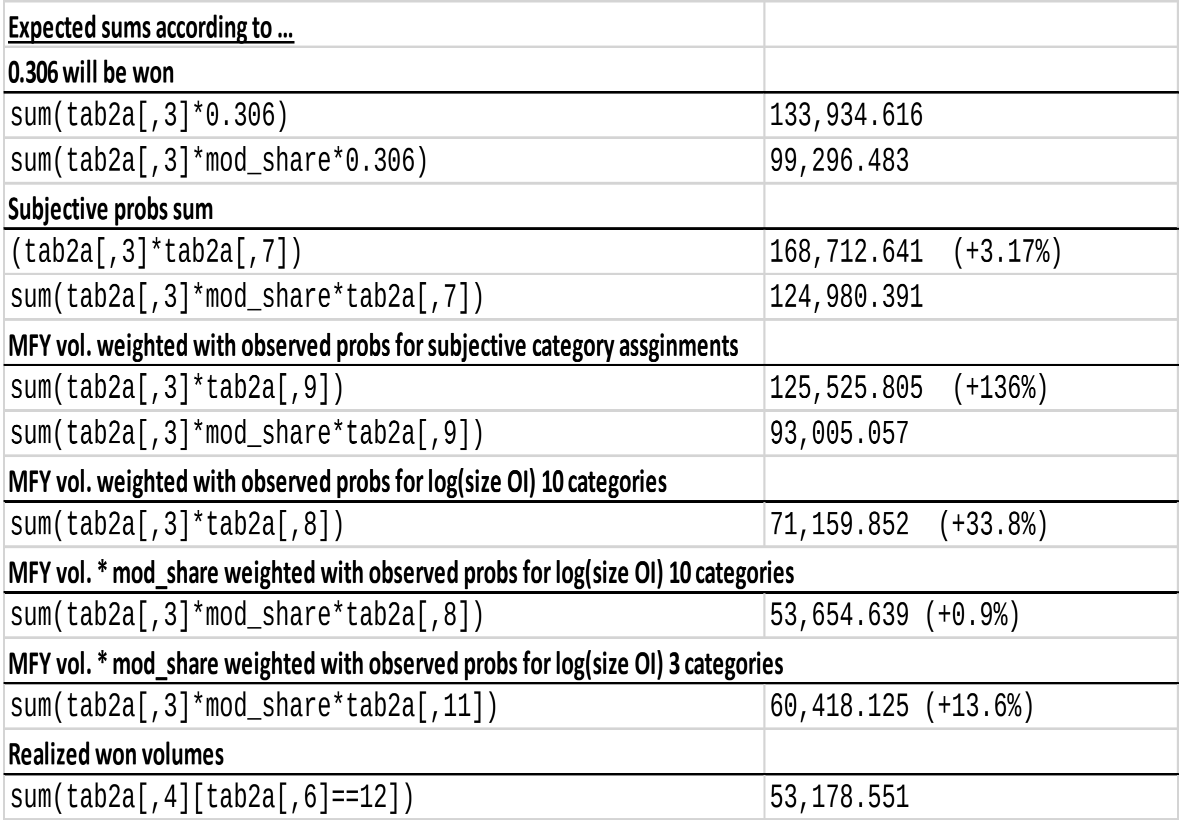
The first step of the analysis is to build up cross tables with the data and to aggregate similar probability categories to make relationships clearer. In the table, observed probabilities and subjective probabilities vary a lot. We also check for probability independencies of opportunities at MDF compared to EFY volumes. We do not find dependences. The sum of all open weighed MFY opportunities from FY 2012 – 2015 calculated as a forecast for the EFY total volume is € 168.712.641. The realized volume is however € 53.178.551, that is only 33.8% of all weighted opportunities.

Even if we assume that we have only the prior knowledge, the 30.6% of the FY are successful, the subjective probabilities are actually pointing into the wrong direction we obtain a much better EFY estimation than using the subjective probabilities. There is no relationship between the subjective probability judgment and the won opportunities.

Finally, a way was found to smooth the crossable with a logarithm were again similar probability categories were aggregated to make relationships clearer. In the end, it is clear to see that small volume opportunities tend to have higher success rates than large volume opportunities. A model with a deviation of less than 1% is found and a 10% VaR calculated. It turns out that the model with three sized categories is the favourable, because 10 size categories seem too many and likely create an overfitting effect within the data.

According to Figure 7, the company has an issue of very overoptimistic salesmen. As a result, the subjective probabilities are false and should not be used for forecasting. In Table 20 you can see the impact of those probabilities.

If we simply assume that, each opportunity has the same probability to be won say the observed 0.306, the expected EFY volume would be € 133,934.616. The expected outcome using the subjective probabilities only however gives € 168,712.641. Considering the uncertainty in the EFY estimates as pointed out by the VaR analysis, this is a clear overstatement by 26% with respect to the (almost) no knowledge situation. Our conclusions from this evidence are that the salesmen might not only selling the company products to potential customers, but also might "sell" their activities to the company where they are employed. The subjective probabilities may also be seen as proposed successes of the salesmen in e.g. a wage bargaining process.



Expected sums according to ...	
0.306 will be won	
sum(tab2a[,3]*0.306)	133,934.616
sum(tab2a[,3]*mod_share*0.306)	99,296.483
Subjective probs sum	
(tab2a[,3]*tab2a[,7])	168,712.641 (+3.17%)
sum(tab2a[,3]*mod_share*tab2a[,7])	124,980.391
MFY vol. weighted with observed probs for subjective category assginments	
sum(tab2a[,3]*tab2a[,9])	125,525.805 (+136%)
sum(tab2a[,3]*mod_share*tab2a[,9])	93,005.057
MFY vol. weighted with observed probs for log(size OI) 10 categories	
sum(tab2a[,3]*tab2a[,8])	71,159.852 (+33.8%)
MFY vol. * mod_share weighted with observed probs for log(size OI) 10 categories	
sum(tab2a[,3]*mod_share*tab2a[,8])	53,654.639 (+0.9%)
MFY vol. * mod_share weighted with observed probs for log(size OI) 3 categories	
sum(tab2a[,3]*mod_share*tab2a[,11])	60,418.125 (+13.6%)
Realized won volumes	
sum(tab2a[,4][tab2a[,6]==12])	53,178.551

Table 20: Comparison between expected OPP volumes

The main conclusion is that the uncertainty in forecasting the EFY realized volume does not depend in an essential way on the randomness of the incoming projects, but on the process to find appropriate subjective success probabilities. We propose not to rely on the “experience” of the salesmen but consider the size of an OI as relevant indicator. Small and medium OI have higher than average success probabilities, while large OI much smaller ones. Big opportunities are won rather rarely (18%). Maybe the corporation should invest more in their acquirement efforts for big projects.

However, a goal needs to be for this company to train their sales staff to make them more sensitive for their business estimations to get better data input and concluding to more data that are reliable to process.

7 Outlook

As we use in the previous models only the information of the volume, the next step would be to include also MFY opportunity probabilities. In our case a reasonable measured probability of the salesmen could be the decision variable.

According to Verbeek (2012, p.239ff) the traditional way to describe the Tobi II is a sample selection problems. In his example, we have a linear wage equation. The variable w_i denotes the wage of employee i. $x_{1i}^{'}$ and $x_{2i}^{'}$ denote vector describing variables which can describe certain aspects like age, gender, education… With the second equation h_i, it is possible to describe whether the person is working or not, which is a binary outcome.

Further there are also some logical rules which apply, w_i only has only an output when h_i^* > 0. The so called unobserved errors $(\epsilon_{1i}, \epsilon_{2i})$ are usually assumed to obey a bivariate normal distribution with expectations zero, variances σ_1^2 , σ_2^2 and covariance σ_{12}. β_1 and β_2 denote the regression coefficients.

$$w_i^* = x_{1i}^{'}\beta_1 + \epsilon_{1i}$$
$$h_i^* = x_{2i}^{'}\beta_2 + \epsilon_{2i}$$
$$w_i = w_i^*, \qquad h_i = 1 \text{ if } h_i^* > 0$$
$$w_i \text{ not observed}, \qquad h_i = 0 \text{ if } h_i^* > 0$$

A next step would be to build up a Tobit II Model to also include reasonable measured subjective probabilities of the salesmen together with opportunity volume. In order to do so we would need to have 12 dummy variables for each objective probability category. (0%, 1%, 10%, …) So for each category we name a variable (x_1, x_2, x_3, …) according to table 4. The variable w denotes the won volume per OPP. To include our information about the phases at MFY won or lost we set a variable y for each OPP (Won = 1, Lost = 0). So h_i indicates a binary variable, count the volume in or not.

$$x_{1i}^{'} = (\text{subjective probabilities}_i, \log(\text{EFY Volume})_i)$$

$$x_{1i}^{'} = x_{2i}^{'}$$

Since the subjective probabilities of the salesmen are estimated badly, we do not see any advantage to proceed along this line. After an improvement of the measurement of the MFY probabilities, the Tobit II model could be a possible next step in future. The 3 category model for volumes together with a model of realized share of the projected volumes is going to produce estimates which are approximately only 13% too high. Maybe these estimates could be improved.

8 References

Arockia Raj, G. (2012). CUSTOMER RELATIONSHIP MANAGEMENT: AN CONCEPTUAL OVERVIEW. *International Journal of Retailing & Rural Business Perspectives (1)*, 171-175.

Åstebro, T., & Rotman, J. L. (2003). The return to independent invention: Evidence of risk seeking, extreme optimism or skewness-loving? *The Economic Journal*(113), 226-239.

Åstebro, T., Mata, J., & Santos-Pinto, L. (2015). Skewness seeking: risk loving, optimism or overweighting of small probabilities? *Theory and Decision*(78), 189-208.

Boulding, W., Staelin, R., Ehret, M., & Johnston, W. (2005). A customer relationship management roadmap: what is known, potential pitfalls, and where to go. *Journal of Marketing (69)*, 155-166.

DeBondt, W. F., & Thaler, R. H. (1995). Financial Decision Making in Markets and Firms: A Behavior Perspective. In A. Jarraow, V. Maksimovic, & W. T. Ziemba, *Financial Handbooks in Operations Research and Management Science, Elsevier* (p. 386). North Holland: Elsevier.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics.* New york: Springer.

Fildes, R., Nikolopoulos, K., Crone, S., & Syntetos, A. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society (59)*, 1150-1172.

Forrest, D., Simmons, R., & Chesters, N. (2002). Buying a dream: Alternative models of the demand for. *Economic Inquiry*(40), 485-496.

Garrett, T., & Sobel, R. (1999). Gamblers favor skewness not risk: Further evidence from United States. *Economics Letters*(63(1)), 85-90.

Golec, J., & Tamarkin, M. (1998). Bettors love skewness, not risk, at the horse track. *Journal of Political Economy*(106), 205-225.

Gummesson, E. (2004). Return on relationships (ROR): the value of relationship marketing and CRM in business-to-business contexts. *Journal of Business & Industrial Marketing (19)*, 136-148.

Hamilton, B. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy*(108), 604-631.

Lawless, M. (2015). Predictive Analytics: An Opportunity for Better Demand Planning and Forecasting. *Journald of Business Forecasting*, 44-46.

Olafsson, S. (2008). Operations research and data mining. *European Journal of Operational Research (187)*, 1429–1448.

Radzevick, ,. J., & Moore, D. A. (2011). Competing to Be Certain (But Wrong): Market Dynamics and Excessive Confidence in Judgment,. *Management Science (57)*, 93-106.

Tsay, R. S. (2005). *Analysis of Financial Time Series.* New Jersey: Wiley.

Tversky, A., & Kahneman, D. (1987). Judgment Under Uncertainty: Heuristics and Biases. *Cambridge Press (185)*, 1124-1131.

Verbeek, M. (2012). *A Guide to Modern Econometrics.* Rotterdam: Wiley.

Widemann, M. R. (2004). *A Management Framework for Project, Program and Portfolio Integration.* Victoria: Trafford Publishing.