Analysrapport 2017:1

ARBETSMILJÖ VERKET

Non-response analysis of NU2015

Based on two Bachelor Theses

Förord

Den här rapporten är en bortfallsanalys av Arbetsmiljöverkets Nulägesundersökning 2015, även kallad NU2015, om arbetsorganisation och arbetsmiljö inom svenskt arbetsliv. Underskningen är en uppföljning av Arbetsmiljöverkets Nulägesundersökning från 2012 (NU2012). För mer information om NU2015, se Teknisk rapport (Statistics Sweden, 2016).

Rapporten sammanfattar och analyserar de huvudsakliga resultaten från två kandidatuppsatser från Institutionen för statistik på Stockholms universitet. De båda uppsatserna bifogas i sin helhet i slutet av rapporten.

Den sammanfattande rapporten är skriven av Niklas Toorell, junior analytiker/statistiker på Arbetsmiljöverket under sommaren 2016, tillsammans med Annette Nylund, senior analytiker på Arbetsmiljöverket, som också fungerar som projektledare för NU2015. Hans-Olof Hagén, konsult år 2017, tidigare senior adviser på Statistiska centralbyrån, har bidragit med kunskaper i metodologi. Karin Sjödahl, konsult på Perido, har språkgranskat rapporten.

Rapporten publiceras i Arbetsmiljöverkets serie för analysrapporter. Annette Nylund svarar på frågor om innehållet i rapporten.

Arbetsmiljöverkets Analysrapport 2017:1

För Arbetsmiljöverket
Februari 2017

Ann Ponton Klevestedt

Chef för enheten statistik och analys

Table of Contents

Sammanfattning **1**
Abstract **2**
Introduction **3**
The Aim of Non-response Analysis 3
Two Bachelor Theses 3
The NU2015 Survey **4**
Objective 4
Random Sample 4
Response Rate 5
Results in the nonresponse analysis **6**
Differences in Background 6
Mean Comparison *6*
Regression Analysis *7*
Differences in Organisation 8
Partly Overlapping Sample in NU2012 and NU2015 ..... *9*
Statistical Tests of Composite Indicators *9*
The Synthesised Conclusions **10**
References **11**
Appendix A: Industry Groups **12**
Appendix B: Independent Variables **13**
The Two Theses **14**

Diagrams and Tables

Diagrams and Figures

Diagram 1. Graph of mean comparison ratios, ranked by the response/sample ratios across size and industry groups 6
Figure 2. Illustration of sub-sample used in the thesis: Differences in Organisation 9

Tables

Table 1. Response rates for industry and size in strata in the NU2015 sample, percent 5
Table 2. Differences in background and probability of responding in NU2015 8
Table 3. Tests of equality between groups in Differences in Organisation 10

Tables in Appendix

Table A. Descriptions of the industry group codes in the NU2015 sample 12
Table B. Labels and descriptions of independent variables in both theses 13

Sammanfattning

Den här rapporten har sin grund i *Arbetsmiljöverkets Nulägesundersökning 2015*, kallad NU2015, som har till syfte att undersöka organisation och arbetsmiljö inom svenskt arbetsliv. Den är en urvalsundersökning i huvudsak baserad på intervjuer, genomförd tillsammans med Statistiska Centralbyrån (SCB) under sista kvartalet 2015 och första kvartalet 2016. Undersökningen baseras på ett representativt slumpmässigt urval om 3035 organisationer, som dras ur det svenska arbetslivet. Organisationerna inkluderar både privata företag och arbetsplatser i offentlig regi. Av urvalets 3035 organisationer har 1619 svarat, vilket motsvarar en svarsandel på 53,3 procent. Det är en lägre svarsfrekvens än i den föregående undersökningen som genomfördes tre år tidigare, kallad NU2012.

Låg svarsfrekvens i urvalsundersökningar kan medföra vissa risker. För att ta ett exempel: låt oss säga att det görs en studie som vill undersöka proportionen höger- och vänsterhänta i Sverige. Om en relativt stor del av gruppen inte svarar, och den delen bara består av vänsterhänta, blir resultatet av undersökningen uppenbarligen missvisande. Därför är det nödvändigt att i samband med urvalsundersökningar också göra en analys av dem som inte svarat, en så kallad bortfallsanalys. Syftet med sådana analyser är att bekräfta att den svarande delen på ett tillförlitligt sätt representerar målgruppen som vi vill studera, i det här fallet det svenska arbetslivet.

För NU2015 har två bortfallsanalyser genomförts och de sammanfattas och analyseras i denna rapport. Rapporten baseras på två kandidatuppsatser, författade under vårterminen 2016, vid Statistiska Institutionen på Stockholms Universitet. Tillsammans belyser de två uppsatserna eventuella skillnader i bakgrund och organisation, mellan gruppen svarande och icke svarande, i målgruppen svenskt arbetsliv. Den första uppsatsen, *Two Methods of Nonresponse Analysis – A case study on The Swedish Work Environment Authority's Baseline Survey 2015*, kommer i rapporten att kallas *Differences in Background*. Den andra uppsatsen, *Bortfallsanalys av systematiska fel i Arbetsmiljöverkets Nulägesundersökning 2015*, kommer att kallas *Differences in Organisation*. Båda uppsatserna bifogas i sin helhet.

De två uppsatserna kompletterar varandra i tillvägagångssätt och har samma syfte: att undersöka om NU2015 kan användas för studier av svenskt arbetsliv. Tillsammans svarar de två bortfallsanalyserna på frågan om de svarande organisationerna i undersökningen NU2015 på ett tillräckligt bra sätt representerar det svenska arbetslivet, och om undersökningen har tillräcklig kvalité för att användas i studier av olika organisationsformer och i studier av hur arbetsmiljöarbete bedrivs. Slutligen hjälper de två studierna till att belysa om undersökningen NU2015 är användbar i studier av samband mellan arbetets organisation och arbetsmiljöarbete respektive verksamhetens resultat och de anställdas utveckling, samt balans mellan arbete och familj, och genusperspektiv.

I den första uppsatsen, *Differences in Background*, undersöks om de svarande och icke-svarande organisationerna skiljer sig åt när det gäller köns-, ålders- och utbildningsstruktur, samt näringsgren och storleksgrupp. Detta görs med hjälp av registerdata från Statistiska centralbyråns Lisadatabas (Statistiska centralbyrån, 2001). Dessutom inkluderas information om skillnader i innovationer och IT-användning mellan de två grupperna av organisationer (Statistiska centralbyrån, 2014, 2015).

Den andra uppsatsen, *Differences in Organisation*, fokuserar på fem sammansatta mått som beskriver olika sätt att organisera arbetet och arbetsmiljöarbetet. Analysen tar hjälp av de organisationer i NU2015 som också har svarat på den tidigare genomförda undersökningen, NU2012. Information om organisation och arbetsmiljö hämtas från den tidigare

undersökningen år 2012, då sådan information inte var tillgänglig i NU2015 när detta uppsatsarbete startade. I denna bortfallsanalys genomförs flera olika typer av test för att se om det förekommer skillnader mellan de svarande och icke-svarande organisationerna, med hänsyn till de fem sammansatta måtten om organisation och arbetsmiljö.

Resultaten från de båda uppsatserna tillsammans visar att risken för att dra felaktiga slutsatser på grund av bortfall av svarande är mycket låg. Undersökningen NU2015 kan på ett tillförlitligt sätt användas för att beskriva organisation av arbetet och arbetsmiljöarbete i det svenska arbetslivet, och sätta in dessa frågor i relevanta sammanhang.

Abstract

This report is a nonresponse analysis of the *Swedish Work Environment Authority's Baseline Survey 2015 (NU2015)*. The conclusion of the report is that NU2015 is unbiased, and therefore represents the Swedish working life adequately, the probability of nonresponse bias is low.

The report presents and analyses the main findings of two bachelor theses from the Department of Statistics at Stockholm University. The two theses are appended in full. The first thesis title is: *Two Methods of Nonresponse Analysis – A case study on the Swedish Work Environment Authority's Baseline Survey 2015*, abbreviated in this report Differences in Background. The second thesis titel is: *Bortfallsanalys av systematiskt fel i Arbetsmiljöverkets Nulägesundersökning 2015*, abbreviated in this report *Differences in Organisation*. Both were written during the spring term of 2016 at the Department of Statistics at Stockholm University.

NU2015 is a survey about management practices and work environment management of the Swedish working life, and a follow-up of the NU2012 Survey, which was conducted in 2012. For detailed information about the collection of the survey, see the technical report (Statistics Sweden, 2016).

The report is written by Niklas Toorell, junior analyst/statistician at the Swedish Work Environment Authority during the summer of 2016, together with Annette Nylund, senior analyst at the Swedish Work Environment Authority, who also serves as the project leader of NU2015. Hans-Olof Hagén, consultant in the year 2017, previously a senior adviser at Statistics Sweden, contributed in methodology. Proof-reading was done by Karin Sjödahl, consultant at Perido.

Questions about the report can be put forward to Annette Nylund, at the Swedish Work Environment Authority.

Introduction

The Aim of Non-response Analysis

The potential aim of nonresponse analysis can be illustrated with an example. Suppose that a survey is constructed with the aim of measuring the proportions of right- and left-handers in Sweden. This simple question can be described as the study variable of interest. If a relatively large portion of the sample does not answer, for one reason or another, and this set only consists of left-handers, then the results of the study would obviously be wrong. Put in statistical terms, we would say that the estimates of the survey are biased.

Nonresponse and its implications is a study field that has increased in awareness among academics in recent years, due to the greater general understanding of sample selection and also the tendency of falling response rates in Sweden and the rest of the western world. Research has shown that it's not the response rate itself that is causing the bias. Rather, the biasedness occurs when there are substantial and characteristic differences between the respondents and the non-respondents in the sample. If these differences get too large, then the estimates of the study variables risk becoming skewed (Groves, 2006).

A properly executed survey should always be accompanied by a nonresponse analysis, given the circumstances explained above. The purpose of such an analysis is to determine whether or not the data set can be used according to the purpose of the survey, and that the results based on the survey are reliable and not exposed to nonresponse bias.

Two Bachelor Theses

The results of the nonresponse analysis are based on two bachelor theses performed by two teams of students from the Department of Statistics of Stockholm University in the spring term of 2016. These two theses are respectively labelled with abbreviations throughout this report, and they are appended in full. The first thesis is called *Two Methods of Nonresponse Analysis – A case study on The Swedish Work Environment Authority's Baseline Survey 2015,* and is abbreviated *Differences in Background.* The second thesis is called; *Bortfallsanalys av systematiska fel i Arbetsmiljöverkets Nulägesundersökning 2015,* and is abbreviated *Differences in Organisation.*

Although the objective is the same for both, the approaches are quite different. The objective of both theses is to find out if there is nonresponse bias in the NU2015 survey. The method of the first thesis involves testing whether the two groups – the responding and nonresponding part of the sample – differ from each other, based on background factors such as gender, age and education. The second thesis utilises the sample units that is overlapping with the previous survey, NU2012. With this latter method, it is possible to test the differences between the two groups concerning management practices and work environment management practices in Sweden, before access to further NU2015 data. These variables are of main interest.

Differences in Background utilises register data provided by Statistics Sweden, such as the gender, education and age structure of the staff of the organisations included in the sample. Since a large proportion of the NU2015 sample has also participated in two earlier mandatory surveys regarding the level of innovation and ICT usage, this type of information is also considered. The idea is to identify variables that, hypothetically, could affect the study variables of interest. If they are also found to influence the willingness to participate in the survey significantly, it could serve as an indication of nonresponse bias.

Differences in Organisation, on the other hand, focuses on a sub-sample consisting of organisations in the NU2015 sample that have also participated (and responded) in a previous version of the same survey. This study was carried out in 2012 and is abbreviated NU2012. Among

other things, in this earlier survey, five composite indicators were used to measure various management practices. These are, in essence, equivalent to the study variables of interest in NU2015, but on an aggregated level. The general idea behind *Differences in Organisation* is to test whether there are any distinct differences in the composite indicators extracted from NU2012, between those who have responded in NU2015 and those who have not. If the differences are significant for any of the five variables, there is a reason to believe that the estimates of NU2015 are biased.

The two theses show complementing results. All things considered, the joint conclusion is that the probability of nonresponse bias is low. This confirms that the response set of NU2015 adequately represents the target population and that the study is well suited for the purpose of describing the Swedish working life in 2015.

The NU2015 Survey

Objective

The overall objective of NU2015 is to study different organisation of management practices, with focus on learning practices at work in the year 2015, alongside work environment management. The Swedish working life is the target population of the survey. The NU2015 survey is a study conducted in collaboration with Statistics Sweden.

Random Sample

A random sample of organisations in NU2015 is drawn to achieve the objective of the survey, and is based on 3035 organisations. The term organisation is used here to emphasise that both firms and public workplaces are included. The sample selection is organised in four parallel sub-samples. The main sample is firms in the private sector that are included in the Swedish Innovation Community Survey (CIS) or in the survey called IT-use in firms (including the survey about micro firms). The rest is drawn directly according to industry and size in the same way as in NU2012. All the sub-samples are randomised according to industry and size.

The sample was drawn in the year 2015 and data was collected throughout the last quarter of 2015 and the first quarter of 2016.

As mentioned above, the aim of NU2015 is to study the management practices and work environment of the Swedish working life. Examples of questions included are "Who is in charge of the everyday quality control?" and "How does your organisation go about examining the work environment?". Most of the responses were collected through telephone interviews, and the ambition is to let the CEO, or a delegated person, of each organisation, answer the questions. Again, the term organisation includes both firms and public workplaces (Statistics Sweden, 2016).

After subtracting the over coverage of 66 organisations, 3035 organisations constitutes the sample of the survey. They have been randomly selected, through stratified sampling, to adequately represent the stated target population. With stratification, the population of a survey gets divided into several homogeneous subgroups, strata, before the actual sampling is done – per stratum. Rather than one big sample, there are several small ones that together represent the target population. With NU2015, the strata are defined by five different size classes, based on the number of employees, and 21 industry groups. The codes of the industry groups follow the SNI2007/Nace Rev 2 classification standard and are listed in Appendix A. The total number of strata is 5×21=105.

The sample in the NU2015 survey is partly drawn within sample frame of the Swedish Community Innovation Survey (CIS) or the survey called IT-use in firms (including the IT-survey about micro firms). This method is described as that NU2015 is "piggy-backing" two other surveys. For industry groups and size classes that are not included in these two other surveys the sample is directly drawn from five size classes, based on the number of employees, and the industries, organised in the sample as 21 groups. In both cases the observations are randomized. The earlier survey NU2012 is directly drawn from size classes, based on the number of employees, and 21 industry groups. Out of the 3035 sampled organisations merely 1619 answered, yielding a total response rate of 53.3 percent. A detailed presentation on the response distribution among the size groups, industry classes and strata is seen in Table 1. Since the main focus was on the firms with at least ten employees, the efforts to get answers was on these.

Table 1. Response rates for industry and size in strata in the NU2015 sample, percent

Industry Group	**Size Group** Number of Employees					
Code	**5-9**	**10-19**	**20-49**	**50-199**	**200+**	
A	*30.0*	*65.5*	*36.7*	*56.7*	*83.3*	48.8
C1	*51.7*	*63.3*	*33.3*	*53.3*	*56.7*	51.7
C2	*36.7*	*69.0*	*46.7*	*65.5*	*60.0*	55.4
C3+B	*53.3*	*50.0*	*46.7*	*50.0*	*46.4*	49.3
D+E	*36.7*	*43.3*	*60.0*	*76.7*	*75.9*	58.4
F	*26.7*	*53.3*	*41.4*	*51.7*	*63.3*	47.3
G	*23.3*	*40.0*	*57.1*	*53.3*	*43.3*	43.2
H	*32.1*	*43.3*	*63.3*	*43.3*	*64.3*	49.3
I	*36.7*	*23.3*	*34.5*	*46.7*	*51.7*	38.5
J	*31.0*	*33.3*	*36.7*	*36.7*	*55.2*	38.5
K	*46.7*	*50.0*	*63.3*	*44.8*	*71.4*	55.1
L	*40.0*	*53.6*	*65.5*	*46.7*	*62.5*	53.2
M	*46.7*	*43.3*	*60.0*	*50.0*	*60.7*	52.1
N	*48.3*	*39.3*	*48.3*	*43.3*	*44.8*	44.8
O+U	*68.4*	*73.9*	*81.5*	*86.2*	*72.4*	77.2
Poff	*46.4*	*63.3*	*70.0*	*76.7*	*75.0*	66.4
Ppriv	*56.7*	*72.4*	*56.7*	*75.9*	*65.5*	65.3
Qoff	*59.3*	*53.6*	*63.3*	*53.3*	*63.0*	58.5
Qpriv	*55.2*	*62.1*	*70.0*	*65.5*	*36.7*	57.8
R	*40.0*	*55.2*	*58.6*	*46.7*	*55.2*	51.0
S+T	*53.3*	*50.0*	*70.0*	*66.7*	*68.0*	61.4
	43.3	52.1	55.3	56.6	59.7	**53.3**

Source: Statistics Sweden, 2016.

Response Rate

Of the full sample, 1619 organisations responded, which corresponds to a rate of 53.3 percent. See further in the technical report (Statistics Sweden, 2016). The sampling process and response rate of NU2015 is further described in the next section of this report. The response rate is lower than the previous survey, NU2012. The sample of NU2012 is described in Stelacon (2013) and Arbetsmiljöverket (2014).

Results in the nonresponse analysis

The results from the two nonresponse theses are presented respectively.

Differences in Background

Introduction to the Thesis

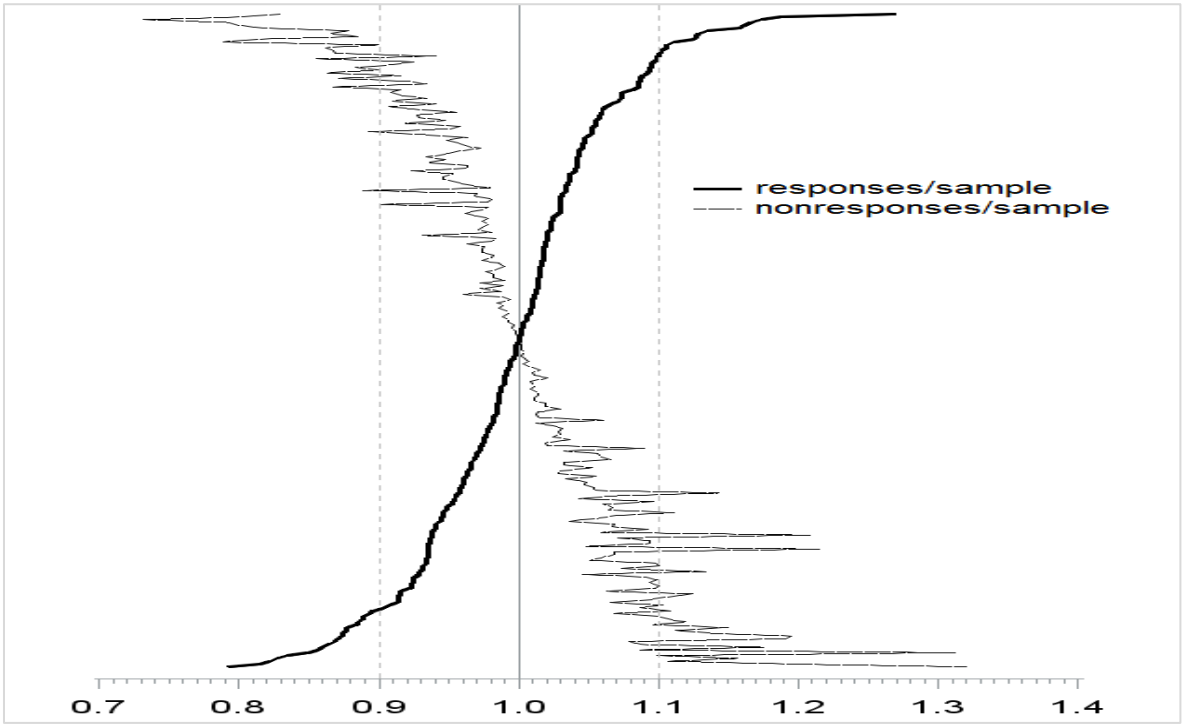
The primary source of information in *Differences in Background* is register data provided by Statistics Sweden. In particular, the aim is to determine if the responding and nonresponding sample differ with regard to the following register variables: gender, foreign background, age, education level and income (the definitions of these variables are explained in detail in Appendix B). These variables are explicitly selected because they could influence the dependant variables. Two analytical methods are used in order to achieve this objective: mean comparison and regression analysis.

Mean Comparison

The mean within the response subset of the sample is calculated for each of the variables described above. The same is done for the full sample. Dividing the response mean by the sample mean yields a ratio. Naturally, when the ratio is close to one, it is an indication that there are small or no differences between the two groups. This, in turn, would suggest that there is only a small probability of nonresponse bias in the NU2015 study, concerning that specific variable.

When applying this technique to the full sample (this is the same method that was used in the response analysis concerning NU2012 (Arbetsmiljöverket, 2014)), the differences are negligible. Ratios are also produced on a detailed level for each of the size and industry classes, yielding a total of 234 ratios. This approach is presented in a graph, shown below in Figure 1.

Diagram 1. Graph of mean comparison ratios, ranked by the response/sample ratios across size and industry groups



Source: Löwing, Martin and Toorell, Niklas (2016). Two Methods of Nonresponse Analysis – A case study on the Swedish Work Environment Authority's Baseline Survey 2015. Statistiska Institutionen, Stockholms Universitet. No 2016:26.

The bold line shows the ratios of the mean for the responding firms divided by the mean for all firms in the sample, ranked from the smallest to the largest, while the thin line shows the mean for the non-responding firms divided by the mean for the whole sample for the same variable.

As seen in the graph, the vast majority of the ratios are placed within 10 percent deviation from one (1.0), which is quite good. In fact, 84.6 percent of the ratios are within these boundaries. Overall, the results from using this method imply that there is no reason to believe that there is nonresponse bias present in NU2015.

Regression Analysis

The second method in the analysis of *Differences in Background* is a regression analysis. One of the big advantages of regression analysis is that it is based on a model that estimates effects of all included information simultaneously.

Regression analysis generally includes several analyses, that are often defined according to some presumptions and specified in models. According the results there are mainly two figures of interest.

The first figure worth noting is the *p-value* that specifies the probability of obtaining a significant estimate. The p-value of 0.05 means that there is a five percent chance of risk that the estimate is a random result; that 1 in twenty times you incorrectly assume a relationship where there is no relation. In social science 0.05 is normal and a very good significance level is 0.01. When the p-value is less than 0,05, the estimated effect is statistically significant, marked in the table with stars or a hash. The estimates shall be interpreted in relation to their respective reference group.

The second figure is the *estimate* that measures the size and the direction of the independent variables' effect on the dependent variable. In this case the independent variables are the background information. The analysis is a test of the probability that the background affect the belonging to the group that responded in NU2015. All effects of the independent variables are estimated simultaneously but can be interpreted separately, per group of information. In other words, the effect can be studied for one variable at a time, given all other information in the analysis. A positive estimate means that the specific information provide a higher probability for the organisation to belong to the group that responded. A negative estimate means that the probability to be a part of the responding group decreases. The best result in this case, is that there is no relation - a zero effect - between the background and respondents in NU2015.

In this case, the objective is to measure the effects of the variables described above (see section

Introduction to the Thesis*)*, on the probability of an organisation responding in NU2015 – in so called randomisation tests. The model also includes the stratifying variables, i.e. the size and industry class of the organisations. Here, most of the variables are not numerical, they are categorical, which requires a reference variable in each group of variables[1].

As seen in Table 2, the majority of the estimates are non-significant (marked with hash), indicating that there appear to be small or no differences between the responding and non-responding organisations in the survey. It should be stressed that only one of the 21 industry groups, which are excluded in the table to save space, is significant. Further, upwards of 30 additional variables, regarding the level of innovation and ICT usage, are tested in alternative models, not presented here, where all effects are found to be either non-significant or irrelevant. However, when testing a very large set of variables, one is almost guaranteed to find at

[1] Logistic models are preferable when the dependent variable is dichotomous, meaning that it only has two possible outcomes. Here, it indicates whether or not the sample unit has responded in NU2015.

least some significant effect, because they increase the probability of obtaining so-called "false positives".

First of all, the significant effects (p-value) of the variables size and industry groups are addressed. Even though it is reasonable to control for the stratifying variables in the regression model, potential bias issues regarding these aspects are already taken care of in the sampling process to some extent. For more information about the stratification, see the previous section.

Table 2. Differences in background and probability of responding in NU2015

Variable	Short description[2]	Estimate	p-value	Significance
Wom	Employees that are women	0.2273	0.2533	#
ForBack	Employees with foreign background	-0.6548	0.0043	***
Age0_34	Employees of ages 34 and younger	-0.5561	0.0355	**
Age35_54	Employees between 35 and 54 years of age	*Reference group*		
Age55	Employees of ages 55 and older	0.5863	0.0802	#
EducElm	Employees with elementary school education	-0.1578	0.6960	#
EducUpSec	Employees with upper secondary education	*Reference group*		
EducHigh	Employees with higher education	0.0970	0.6568	#
AvgInc_log	Average income	-0.1170	0.1611	#
sgrupp1	First size group, 5-9 employees	-0.3707	0.0023	***
sgrupp2	Second size group, 10-19 employees	*Reference group*		
sgrupp3	Third size group, 20-49 employees	0.1208	0.3094	#
sgrupp4	Fourth size group, 50-199 employees	0.1890	0.1122	#
sgrupp5	Fifth size group, more than 200 employees	0.3068	0.0128	**

*Note: Significance: # Non-significant, *** Significant at a 0.01 level, ** Significant at a 0.05 level. Definitions of the variables se appendix B.*
Source: Löwing and Toorell (2016). Two Methods of Nonresponse Analysis – A case study on the Swedish Work Environment Authority's Baseline Survey 2015. Statistiska Institutionen, Stockholms Universitet. No 2016:26.

The variables measuring the proportion of employees with foreign background, labelled *ForBack,* and age, specifically the age-group *Age0_34*, measuring the proportion of employees of ages 34 and lower, both have significant negative effects that could be worth noting. It could be possible that organisations with higher proportions of these categories in their workforce indicate different practices and that they, because of this, are less inclined to participate in a survey with questions about management practices and work environment. Therefore, these parts of the Swedish working life could be under-represented in NU2015. These findings are further discussed below.

Differences in Organisation

Introduction to the Thesis

As mentioned in the introduction, roughly one-half of the NU2015 sample constitute the nonresponse set; they did not answer the questions in the survey. A nonresponse analysis, conducted in ideal theoretical conditions, would have access to what these organisations would have answered. But this is not possible since they do not participate in the survey. The general idea behind *Differences in Organisation* is to utilise the actual responses of an earlier version of

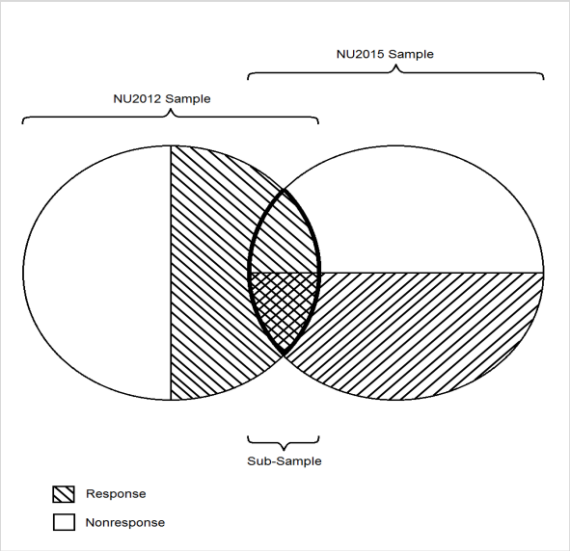
[2] For a detailed variable description list, see Appendix B.

the survey, NU2012, on the sample of NU2015. These responses have been compiled into five composite indicators ranging from 0 to 1, further described below. Statistical tests are deployed to determine whether or not the observed values of the composite indicators are significantly different between the responding and nonresponding part of the NU2015 sample.

Partly Overlapping Sample in NU2012 and NU2015

The official nonresponse analysis of the earlier version of the survey, NU2012, concludes that the results of that study are unbiased (Arbetsmiljöverket, 2014). This premise, along with the fact that some of the responding organisations in NU2012 are also included in the NU2015 sample, is important for the choice of method in *Differences in Organisation*. The focus is on the overlapping sample of organisations in the two surveys; the sample in NU2015 that is overlapping with the responding part of the NU2012 sample, here called the sub-sample. With this technique, variables of interest are available for both the responding and the nonresponding part of the NU2015 sub-sample. This sample of organisations consists of 325 units and can be described as an intersection between the two samples, as illustrated in Figure 2.

Figure 2. Illustration of sub-sample used in the thesis: Differences in Organisation



Note: The displayed sets have non-actual proportions

Statistical Tests of Composite Indicators

The study variables used are five composite indicators labelled k_ind, k_strukt, k_dec, k_num and sam-index. These are designed to measure the management practices of the organisations and are described in greater detail in Appendix B. They are all based on several questions, and their answers have been aggregated in composite indicators. The observed values of the composite indicators are calculated for the response and nonresponse subset of the sub-sample separately. With the help of established statistical techniques, described below, the aim is to test whether or not these figures differ from each other. If there are large enough, significant differences for any of the five composite indicators, it would suggest that there is nonresponse bias present in the NU2015 study.

Student's t-test is generally conceived to be the more accurate of the two but does, on the other hand, assume that the tested variables are normally distributed. The Wilcoxon-Mann-Whitney test does not impose such a constraint. Because of this, whether or not the composite indicators are normally distributed is tested separately, both graphically and numerically, with mixed

results. Some of the variables, k_ind and k_strukt in particular, do not seem to be normally distributed. However, as seen in Table 3, all p-values are exceeding 0.05 (and even 0.10), indicating that the differences between the responding and nonresponding part of the sub-sample are non-significant[3]. It should be noted that the outcomes are the same regardless of the choice of test. All in all, the results of *Differences in Organisation* indicate that there is no reason to believe that there is a nonresponse bias present in the NU2015 study.

Mainly two of the statistical tests presented in the thesis are considered here in the summary for the comparison of the groups: Student's t-test and the Wilcoxon-Mann-Whitney test. The results of both tests are seen in Table 3 below.

Table 3. Tests of equality between groups in Differences in Organisation

	Student's t-test			Wilcoxon-Mann-Whitney test		
Variable	**t-statistic**	**p-value**	**Significance**	**Z-statistic**	**p-value**	**Significance**
k_ind	-1.21	0.23	#	-1.57	0.12	#
k_strukt	-0.29	0.77	#	-0.04	0.97	#
k_dec	-1.51	0.13	#	-1.56	0.12	#
k_num	0.46	0.65	#	0.48	0.63	#
samindex	0.41	0.68	#	-0.25	0.80	#

 # Non-significant The Welch's t-test is used for the sam-index due to the variance for this variable
Source: Axelsson, Merrisha and Åberg, Edvard (2016). Bortfallsanalys av systematiskt fel i Arbetsmiljöverkets Nulägesundersökning 2015. Statistiska Institutionen, Stockholms Universitet. No 2016:28.

The Synthesised Conclusions

Given some of the results of the first thesis, *Differences in Background*, one could perhaps stress some legitimate concerns over possible bias that the organisations with larger proportions of young people and people of foreign background are under-represented in NU2015. It is, however, important to put these findings into context. First of all, when testing a very large set of variables as in this case, it's almost unavoidable not to find significant effects among some of the independent variables. Very large sets increases the possibility of obtaining so-called "false positives" that seem to be distinct but are just due to chance.

The analyses in the second thesis, *Differences in Organisation*, has tested if there is a difference between the organisations that responded to NU2015 compared to those that did not, with the help of the five composite indicators about work organisation and work environment. No significant difference between those who responded and those who did not was found. Thus, one could safely conclude that the results of NU2015 are indeed unbiased.

While a low response rate for a survey might seem problematic at first, it's important to remember that it doesn't necessarily entail estimation complications. The reliability of NU2015 has been thoroughly reviewed, since two separate nonresponse analyses have been conducted. All things considered, the joint conclusion of the two papers confirms that the probability of bias is low and that the responding part of the survey does represent the target population with high credibility. Therefore, the estimates of NU2015 are well suited for researching the management practices and work environment of the Swedish working life.

[3] Interpretation of p-value and the concept of statistical significance, see section *Regression Analysis.*

References

Arbetsmiljöverket (2014). *Bortfallsanalys - Representerar de svarande organisationerna i Arbetsmiljöverkets Nulägesundersökning 2012 svenskt arbetsliv?* Arbetsmiljöverkets analysrapport 2014:1.

Groves, R. M. (2006). *Nonresponse Bias in Household Surveys*. The Public Opinion Quarterly, 70(5), 646-675. www.tc.umn.edu/~alonso/Groves_POQ_2006.pdf.

Statistiska centralbyrån (2001). *Longitudinell Integrationsdatabas för Arbetsmarknads och Sjukförsäkringsstudier (LISA)*, SCB. http://www.scb.se/sv_/Vara-tjanster/Bestalla-mikrodata/Vilka-mikrodata-finns/Longitudinell-integrationsdatabas-for-sjukforsakrings--och-arbetsmarknadsstudier-LISA/.

Statistiska centralbyrån (2014). *Innovationsverksamhet i Sverige (CIS), 2012-2014*. http://www.scb.se/Statistik/UF/UF0315/_dokument/UF0315_DO_2012-2014_AS_160302.pdf.

Statistiska centralbyrån (2015). *Företagens användning av IT 2015*. http://www.scb.se/Statistik/_Publikationer/NV0116_2015A01_BR_00_IT02BR1501.pdf

Statistics Sweden (2016). *Teknisk Rapport - En beskrivning av genomförande och metoder. Nulägesundersökningen 2015*.

Stelacon (2013). Teknisk beskrivning – Arbetsmiljöverkets Nulägesundersökning SAM 2012. https://www.av.se/globalassets/filer/statistik/arbetsmiljostatistik-teknisk-beskrivning-nulagesundersokning-sam-2012-analysrapport-2012.pdf https://www.av.se/sok/?qry=bortfallsanalys.

Appendix A: Industry Groups

Table A. Descriptions of the industry group codes in the NU2015 sample

Industry Group Code	Description
A	Agriculture, forestry and fishing
C1	Manufacturing, labour intense sectors
C2	Manufacturing, science intense sectors
C3+B	C) Manufacturing, capital intense sectors and B) Mining and quarrying
D+E	D) Electricity, gas, steam and air conditioning supply E) Water supply, sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O+U	O) Public administration and defence, compulsory social security
Poff	Education, public sector
Ppriv	Education, private sector
Qoff	Human health and social work activities, public sector
Qpriv	Human health and social work activities, private sector
R	Arts, entertainment and recreation
S+T	S) Other service activities T) Activities of households as employers, undifferentiated goods- and services-producing activities of households for own use

Source: Teknisk rapport. En beskrivning av genomförande och metoder. Nulägesundersökning 2015. Statistics Sweden, 2016

Appendix B: Independent Variables

Table B. Labels and descriptions of independent variables in both theses

	Independent variables	Description
Differences in Background	Wom	Proportion of women in the organisation.
	ForBack	Proportion of employees with foreign background in organisation. Foreign background is here being defined as someone who is born outside Sweden or who has two parents born outside Sweden.
	Age0_34	Proportion of employees between 0 and 34 years old.
	Age35_54	Proportion of employees between 35 and 54 years old.
	Age55	Proportion of employees older than 54 years.
	EducElm	Proportion of employees with elementary school as maximum education level. *Primary education* in Swedish.
	EducUpSec	Proportion of employees with upper secondary school as maximum education level. *Secondary education* in Swedish.
	EducHigh	Proportion of employees with higher education as maximum education level. *Higher education* in Swedish.
	AvgInc	Average yearly income of employees in organisation.
Differences in Organisation	k_ind	Level of *individual learning* within organisation, in the sense that the employees are permitted to learn and develop as individuals.
	k_strukt	Level of *structural learning*, in the sense that the organisation, as a whole, learns and develops from its surroundings.
	k_dec	Level of *decentralisation*, in the sense that the employees "on the floor" get more power.
	k_num	Level of *numerical flexibility*, in the meaning of the organisation's just in time practices and ability to reduce labour costs in short time.
	Samindex	Level of compliance with the Swedish regulation regarding the systematic work environment management.

The Two Theses

Differences in Background

Löwing, Martin and Toorell, Niklas (2016). *Two Methods of Nonresponse Analysis – A case study on the Swedish Work Environment Authority's Baseline Survey 2015.* Statistiska Institutionen, Stockholms Universitet. No 2016:26. I sammanfattningen används ett kortnamn på engelska; In the summary the used abbreviation is: *Differences in Background.*

Differences in Organisation

Axelsson, Merrisha and Åberg, Edvard (2016). *Bortfallsanalys av systematiskt fel i Arbetsmiljöverkets Nulägesundersökning 2015.* Statistiska Institutionen, Stockholms Universitet. No 2016:28. I sammanfattningen används ett kortnamn på engelska; In the summary the used abbreviation is: *Differences in Organisation.*

Stockholms universitet

Bachelor thesis

Department of Statistics

Kandidatuppsats, Statistiska institutionen

No. 2016:26

Two Methods of Nonresponse Analysis

- A case study on The Swedish Work Environment Authority's Baseline Survey 2015

Två metoder för bortfallsanalys

– En fallstudie på Arbetsmiljöverkets Nulägesundersökning 2015

Martin Löwing and Niklas Toorell

Independent writing 15 ECTS credits in Statistics III, spring-term 2016

Supervisor: Dan Hedlin

Abstract

The Swedish Work Environment Authority has, in co-operation with Statistics Sweden, conducted their Baseline Survey for the year 2015 (NU2015). Our main objective with this thesis has been to investigate how well the responding part of the survey sample represent the target population. Due to the co-ordination of NU2015 with other surveys and databases we have had access to a rich set of possible auxiliary variables. We have approached the problem with two different techniques, namely *mean comparisons* among some of these variables and *logistic regression*. These methods were deployed as means to finding out which of these aspects, if any, had a significant and substantial effect on the *propensity to answer*. In order to cause bias, they also had to *affect the study variables of interest* in the survey. Here, prior research, theory and intuition had to guide us. A *Stepwise Backward Elimination* procedure was conducted in order to narrow down the number of auxiliary variables under consideration. For the vast majority of analyzed factors we could not find any reasons to suspect that the respondents and nonrespondents differed. We are, however, providing evidence that the working life experiences of *younger people* and *people with foreign background* are underrepresented in the survey. Organizations with relatively high shares of these categories respond to the survey to a lesser extent. It is also known that these subgroups of the population have different work environment experiences in relation to the average employee. We propose that inference based on NU2015 are to take these findings in to consideration, perhaps through *calibration* or *post-stratification*. We also conclude that logistic regression is to be preferred in favor of the mean comparison approach with regard to nonresponse analyses.

Keywords: Survey, Bias, Nonresponse, Logistic Regression, Backward Elimination

Sammanfattning

Arbetsmiljöverket har, i samarbete med Statistiska Centralbyrån (SCB), utfört sin Nulägesundersökning för året 2015 (NU2015). Vårt övergripande mål med den här uppsatsen har varit att undersöka hur väl den del av undersökningsurvalet som svarade kan antas representera målpopulationen. Tack vare hur NU2015 är koordinerad med andra undersökningar och databaser, så har vi haft ett stort antal möjliga bakgrundsvariabler till vårt förfogande. Vi har närmat oss ämnet med två olika tekniker, nämligen *skattningsjämförelser* mellan vissa av variablerna samt *logistisk regression*. Dessa metoder har företagits för att mäta i vilken utsträckning dessa aspekter har en signifikant och betydande effekt på *benägenheten att svara*. För att det ska föreligga risk för snedvridning måste de också påverka *variablerna som skattas i undersökningen*. Här var vi tvungna att förlita oss på tidigare forskning, teori och intuition. En stegvis baklänges eliminationsprocess företogs med syftet att sålla bland de möjliga variablerna. För de allra flesta kunde vi inte finna något som tyder att de skulle vara källor till olikheter mellan de svarande och icke-svarande. Vi presenterar dock bevis för att arbetslivserfarenheterna hos *unga* och *personer av utländsk härkomst* är underrepresenterade i undersökningen. Organisationer med relativt höga andelar av anställda med dessa attribut har svarat i mindre omfattning. Det är också känt att dessa grupper i genomsnitt har erfarenheter av arbetsmiljöer som skiljer sig från populationen i stort. Vi föreslår därför att slutsatser som baseras på NU2015 ska ta dessa fynd i beaktande; möjliga metoder för detta är *kalibrering* och *post-stratifiering*. Vi sluter oss också till att logistisk regression har otvivelaktiga fördelar i förhållande till skattningsjämförelsen vad gäller dess användning i bortfallsanalyser.

Nyckelord: Survey, Bias, Systematiskt skattningsfel, Bortfall, Logistisk regression

Preface

We want to extend our warmest gratitude towards Hans-Olof Hagén and the staff at ES/IFI at Statistics Sweden, Stockholm. For the opportunity to write this thesis, for their invaluable expertise and for their hospitable reception.

Martin Löwing Niklas Toorell

Table of contents

1 Introduction 1
1.1 Abbreviations of surveys and datasets 1
1.2 Thesis objective 1
1.3 Theory behind nonresponse bias and earlier research 2
2 Methodology 4
2.1 Mean comparison approach 4
2.2 Logistic regression approach 5
3 Data 8
3.1 The NU2015 Survey 8
3.1.1 Response rate of NU2015 10
3.2 The LISA Database 12
3.3 The FDB Database 13
3.4 The CIS2014 Survey 13
3.5 The IT2015 Survey 13
3.6 Loss of information when combining the datasets 14
4 Results and Analysis 15
4.1 Mean comparisons of main structural factors 15
4.2 Logistic regression models 18
4.2.1 General Model 18
4.2.2 Stepwise Backward Elimination procedure 21
4.2.3 Extended Model 22
4.2.4 Common Effects 25
4.2.5 Assessing the Models 27
5 Conclusions 28

References 32
Appendix A: Variables removed with SBE 34
Appendix B: SAS code 35

1 Introduction

1.1 Abbreviations of surveys and datasets

NU2015 – The Swedish Work Environment Authority's Baseline Survey, 2015 *(Arbetsmiljöverkets Nulägesundersökning 2015)*

CIS2014 – Community Innovation Survey 2014 *(Innovationsverksamhet i Sverige 2014)*

IT2015 – ICT Usage and e-Commerce in Enterprises 2015 *(Företagens användning av IT 2015)*

LISA - The Longitudinal Integration Database for Health Insurance and Labour Market Studies *(Longitudinell Integrationsdatabas Sjukförsäkrings- och Arbetsmarknadsstudier)*

FDB - Statistical Business Registry *(Företagsdatabasen)*

1.2 Thesis objective

In each survey, one has to consider the quality of the collected data. There are several types of errors which can occur so that the data do not represent what one really wanted to measure. Common sources are *measurement errors* and *bias due to missing data*. In this thesis we are only concerned with the second type of error, specifically the bias that arises from *nonresponse.*

At hand is the NU2015 survey produced by the Swedish Work Environment Authority in co-operation with Statistics Sweden. For us, the overall question is: *Do the organizations that did respond to the survey represent the target population in a statistical sense?* And if that is not the case: *What aspects of the population are worth taking into consideration with regard to nonresponse bias?*

In order to answer these questions we will use two different methods: comparing the mean values of a certain set of variables and modeling the response propensity[1] through logistic regression. *Do the results of these two methods correspond? Is one of the methods to prefer over the other?* These are the questions we hope to answer with this thesis.

[1] Propensity could be understood as probability or likelihood. In the context of nonresponse bias, propensity is the appropriate term.

1.3 Theory behind nonresponse bias and earlier research

Ever since the breakthrough of survey sampling in the 1930s there have been an awareness among statisticians of the theoretical implications of nonresponse. As the average response rates started dropping quite dramatically during the last decades, the topic has gained traction as a research subject in statistical literature. Many valuable insights on the subject are, therefore, quite recent. (Särndal and Lundström, 2005)

In this thesis we lean heavily on the findings of Groves (2006). His main argument is that there really is no evidence to support the theory that higher response rates always lead to more precise results. It may actually be the other way around. He uses Bethlehem's (2002) definition of approximate bias:

$$Bias(\bar{y}_r) \approx \frac{\sigma_{yp}}{\bar{p}} \tag{1.1}$$

where $Bias(\bar{y}_r)$ is the nonresponse bias of the (unadjusted) respondent mean $\bar{y}_r$, $\bar{p}$ is the mean propensity to answer the survey and σ_{yp} is the population covariance between the survey variables and the propensity. It states that the bias of the estimator $\bar{y}_r$, say the mean value of a study variable investigated in the survey, is dependent on the propensity to answer and the relationship between the variables themselves and the propensity.

There are two important features with this expression. First, it assumes that there are no measurement errors and that all bias stem from nonresponse. Secondly, it does not contain the actual response rate, at least not explicitly. $\bar{p}$ is, of course, correlated with the response rate – if $\bar{\rho}$ is high, the response rate tends to be high as well and vice versa.

This is crucial, as the key argument supplied by Groves is that it is σ_{yp} that is the governing parameter when it comes to nonresponse bias. The results from a survey with a low response rate could be well suited be used to estimate, say the mean value of **Y** (a set of study variables in the survey) in an unbiased way. If only the propensity to answer is not affected by a) the study variables **Y** themselves, or b) some variables **Z** that affect both **Y** and p, the result will be on point.

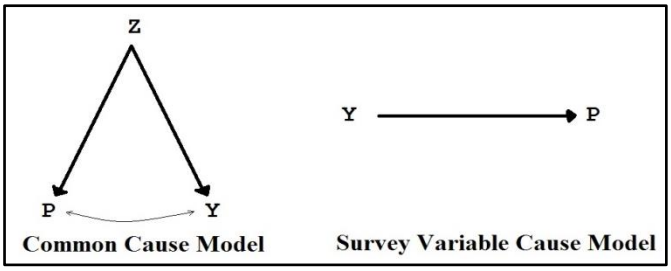


Figure 1.1: Idealized causal models of response propensity. Source: Groves (2006)

These are two alternatives, illustrated in Figure 1.1, out of five which Groves describes, and we will mostly focus on the Common Cause Model - an underlying vector **Z** which has an impact on both the response propensity, p, and **Y.**

The conclusion Groves makes is that there is not, as has been implicitly implied in earlier research, a "hard-core nonrespondent group" in each survey. These hard-core individuals would always have had an answering propensity of zero, no matter what types of questions the survey contained. This is, in Groves mind, not realistic and so he lands in the expression of Equation (1.1) which allows for a continuum of response propensities.

Särndal *et al.* (1992) underlines the view that the hard-core perspective is much too simplistic. They argue that a *response homogeneity model* can be fitted and the resulting propensities to answer the survey in each (homogeneous) group used to weight the estimates. Tivesten *et al.* (2012) apply this method. They weigh their car accident data with the inverse of the propensity to answer which is estimated with logistic regression.

Valuable insights has also been gained from Lineback *et al.* (2010). They point out that both weighting and imputation as methods depend on the condition that the data is Missing At Random (MAR). If this is the case, the respondents can be seen as a random subsample and inference based on this group can be generalized to the sample and then to the population. If, however, the data is not MAR, imputation and weighting might worsen the bias.

Rather than weighting the data, we will attempt to find out which variables, if any, might influence the propensity to respond. That is, in the response homogeneity context, which variables might be used to form groups within which all observations have (at least roughly) the same propensity to respond?

When assessing whether or not the variables in question belong to **Z** we consult three main topics, namely; *theory, effect* and *significance*. We can use statistical procedures to test whether or not the potential variables have a substantial and significant effect on $\boldsymbol{p}$. Of course, we cannot directly observe if the **Z** candidates affect the elements of **Y**, since we do not have **Y** data on the nonrespondents. Therefore, theory and intuition will have to play a role in deciding which variables might be of interest.

2 Methodology

As stated above, the main objective is to investigate if the responding group of NU2015 is representative of the whole target population – the Swedish working life. In order to achieve this we're applying two different methods which are thoroughly described below.

There are, of course, other methods than those we use for investigating nonresponse bias. One such method is *extrapolation,* where subjects who respond late in the survey are assumed to be more similar to the nonrespondents. You then compare their answers with the average. You could also try to contact a subsample of the nonrespondents and interview them regarding a small, well chosen, subset of survey questions. If these values correspond well with the values among the respondents, nonresponse bias might be less of a problem. Both of these methods are described by Armstrong and Overton (1977).

We are assuming that the sampling frame from which the survey is sampled adequately represent the target population and that there are no measurement errors. For all computational procedures we have used SAS® 9.4. Unless otherwise stated, we're applying the standard significance level of 5 % for all statistical tests.

2.1 Mean comparison approach

The analytical approach of the official nonresponse report on NU2012 (Arbetsmiljöverket, 2014a) focuses mainly on mean comparisons among a certain set of auxiliary variables. One of our aims have been to replicate this method in order to ease comparisons between the two papers. The variables used, which we refer to as the *main structural factors*, are presented in Table 4.1 of the results section.

For each variable, x_i, a ratio is calculated as

$$\frac{\bar{x}_r}{\bar{x}} \tag{2.1}$$

where $\bar{x}_r$ is the mean of the response subset and $\bar{x}$ is the mean of the whole sample. This is done for both weighted and unweighted means.

In order to calculate the weighted mean one first needs to determine what weight to use. In a survey setting this is typically the design weight, which is a quantity that takes the inclusion probability of each population unit into account. The value of a design weight tells us how many population units a sample observation represents. If the inclusion probability is the same across the full sample there will be no difference between a weighted and unweighted mean. This implies that it will only make sense to calculate the weights when stratified sampling is applied, which is the case with NU2015. With stratification, the population of a survey gets divided into several homogeneous subgroups, *strata*, before the actual sampling is done –per *stratum*.

Formally, the design weight of a stratum j will be calculated as

$$w_j = \frac{N_j}{n_j} \tag{2.2}$$

where n_j is the sample size of the stratum and N_j is the population size of the stratum.[2] Then, the weighted mean can be defined as

$$\bar{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i}. \tag{2.3}$$

Depending on the magnitude, a ratio that deviates from unity could serve as an indication that there are differences between the response subset and the population with regard to that specific factor. The ratios will be presented and analyzed for both the whole sample as well as on a detailed level for each of the size and industry classes.

2.2 Logistic regression approach

One of the main advantages with regression modelling in general is the ability to estimate certain effects and test their significance, while controlling for the effects of other variables. In our case the dependent variable (whether or not the organization in question has responded) is binary. The logistic models superiority over Ordinary Least Squared (OLS) when it comes to estimating dichotomous dependent variables is well known. (Pampel, 2000) Consequently, fitted logistic models will be used to assess whether or not different auxiliary variables have a substantial and significant effect on the propensity to answer the NU2015 survey. For theoretical background and application of logistic regression, we have leaned heavily on Allison (2012), Tivesten *et al.* (2012) and Särndal *et al.* (1992).

The model reads

$$log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2.4}$$

where $\log\left[\frac{p}{1-p}\right]$ is known as the *log-odds*, p is the propensity of an organization responding, α is the intercept, which also serves as the baseline, and β_k is the coefficient of the k_{th} variable x_k. Expressing (2.4) in terms of the propensity p gives

$$p = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}. \tag{2.5}$$

At first, a model utilizing only the main structural factors, called the *General Model*, will be estimated and analyzed. Since the sample of NU2015 is highly coordinated with those of CIS2014 and IT2015, it provides us with a rich set (upwards 400) of other possible auxiliary variables. In order to fit an alternative model with any of these, one first has to decide which

[2] Notice that this weight calculation is not the same as the one outlined in the technical specification of NU2015. (SCB, 2016) One of the reasons for this is that we, naturally, can't make the assumption that there is no nonresponse bias in the study. The weight definition we're using is taken from Svenska Statistiskersamfundet (2005).

variables to include. Whenever possible, aggregated levels are used and as a first rough cut, most of the variables are tested separately with the Wald Chi-Square test. Then, a *Stepwise Backward Elimination* (SBE) procedure is conducted to conclude which of these to consider.

With SBE, SAS will fit a logistic model with all potential variables and then exclude the one with the highest p-value, given that it is higher than a specified threshold. Then, a new model is estimated without that specific variable. This procedure is repeated until no more variables can be omitted. With our SBE specification the cut-off point is set to 0.2. Also, the variables of the General Model have been specified to remain in all steps, regardless of their respective p-values. Given the results of the SBE procedure, a new model, called the *Extended Model,* is estimated with additional auxiliary variables. Due to the population differences of NU2015 and the other two surveys, the number of observations are much less in the Extended Model than in the General model. See Section 3.6 for further details.

The stratifying variables, `sgrupp` and `bgrupp`, will enter both models. Although this type of specification might be less efficient when it comes to estimating the standard errors than if we fitted (2.4) within strata (using "strata bgrupp sgrupp" in the model statement in SAS), it is more valuable for us to receive estimations on the effects of the different covariate levels. If, say, the beta estimates of certain size groups are influential, it is a sign that it wise to stratify using size of an organization.

For all tests of significance, both in the SBE procedure and otherwise, Wald Chi Square tests are used. The test statistic is defined as

$$\frac{\beta_i^2}{SE_{\beta_i}^2} \tag{2.6}$$

where β_i is the i_{th} parameter coefficient, and the denominator is its variance. The Wald statistic is known to be overly conservative when the estimate of the parameter is large, since this also tends to inflate the estimated standard error. This, in turn, increases the risk of a type-II error. One alternative is to conduct likelihood ratio tests, which are known to be more exact. However, trying to fit the best possible model is not the stated objective of this thesis.

To display the effects of belonging to, for example, different size groups, *odds ratios* are calculated. The interpretation of an odds ratio is closely related to that of a propensity or probability, and can be descried mathematically as

$$OR_{X=1,X=0} = \exp(\beta_X) \tag{2.7}$$

where OR is the odds ratio, and β_X is the estimated log odds ratio. In this scenario, X=0 is the baseline group. $\exp(\beta_X)$ measures the deviation stemming from belonging to X=1 in relationship to this baseline group. If the odds ratio is, say, equal to 2, the straight forward interpretation is that the odds of success (in our case, responding to the NU2015) is two times as likely if the observation belong to X=1 compared to the baseline.

All reported coefficients are estimated with *Maximum Likelihood Estimation* (MLE), calculated with Fisher's Scoring iteration technique. MLE is known to be a (or have approximately

similar properties to) *uniformly minimum-variance unbiased estimator*, especially in large samples. The MLE maximizes the following expression:

$$log\ \mathcal{L}(\theta; x_1 \dots x_n) = \sum_{i=1}^{n} log\, f(x_i|\theta) \tag{2.8}$$

where $\mathcal{L}$ is the *likelihood function,* θ is the parameter which is being optimized in order to maximize the expression and the *x*'s are the values of the auxiliary variables. This is done, through iteration, by assigning values to the different parameters being estimated**.**

Although the aim isn't to fit the best possible model, but rather finding out which variables that are affecting the response propensity, we will provide measures of predictive power for both models. The measurement we're using, called *tau-a*, varies between 0 and 1 and is utilizing all possible pairs of the sample observations (without pairing an observation with itself). It is calculated as

$$Tau - a = \frac{C - D}{N} \tag{2.9}$$

where C is the number of concordant pairs, D is discordant pairs and N is total number of pairs. SAS matches the observation into all possible pairs, and disregards the ones where both have the same value of the response dummy (i.e. both *did* or *did not* answer). Then it compares the estimated propensities given by the model, and if the model predicts a higher value for the observation that did answer than for the one that did not, then that pair is *concordant*. If the model predicts a higher value for the one that did not answer, it is a *discordant*. (Allison, 2012)

3 Data

All the data used in this thesis is produced and recorded at Statistics Sweden. Most of the data provided for us comes in the form pre-existing datasets, such as the LISA, the FDB, the CIS and the IT datasets, whereas the final version of the NU2015 dataset was provided about halfway into the period of the thesis course. All datasets are described in greater detail below.

Data produced by Statistics Sweden is generally believed to have high reliability and is the base for calculating the CPI the GDP and other statistics regarding the Swedish economy, to name a few. It also serves as basis for decisions at both local and national political arenas.

3.1 The NU2015 Survey

The Swedish Work Environment Authority's Baseline Survey for 2015 (NU2015) is a survey regarding the systematic work environment management in the Swedish working life. The survey is second in line, following NU2012, and is carried out as a co-operation between the Swedish Work Environment Authority and Statistics Sweden. Both studies focus on questions regarding the health and hazards in the workplace. Not as much their occurrences but their preventions and the steps taken to avoid them. The questions in NU2015 range from "Who is in charge of the everyday quality control?" and "How does your organization go about examining the work environment?" to "Do you know of the existence of the regulation known as the Systematic Work Environment Management?" (SCB, 2016)

One of the specific aims with the study, among others, is to compute a Systematic Work Environment index, with which the systematic work environment management in organizations[3] can be assessed. The index generates a value between 0 and 1, where values closer to 1 indicate that the work environment management is carried out in accordance with regulations. (Arbetsmiljöverket, 2013) The target population of the study is, as stated above, the Swedish working life. This means that both private and public organizations, ranging from municipal kindergartens to multinational car manufacturers, are included.

NU2015 is a non-mandatory survey, a fact that quite possibly could have a negative impact on the response rate. Most of the data was collected through telephone interviews. According to Swedish regulation it is the leadership of an organization that is ultimately responsible for the work environment. Because of this, the goal have been to let the CEO or the workplace manager answer the questions.

It's important to stress the fact that the NU2015 sample consists of organizations, for which we have various auxiliary data, rather than the individuals at the workplaces who are answering (or supposed to answer) the survey. In other words, we might, for instance, be able to determine the gender structure of a specific organization but not the gender of the person who is answering the survey.

The sampling frame, from which the NU2015 sample was drawn, was created in November 2014 in order to allow for coordination with CIS2014 and IT2015. It consists of 103 189 organizations and is assumed to suitably represent the target population. The sample of 3101

[3] The term organization is used in order to emphasize that both private enterprises and public establishments, such as municipal workplaces, constitute the population.

organizations was drawn with stratified sampling. The strata are defined by 5 size groups and 21 industry classes, creating a total of 105 strata. The number of employees defining each of the size groups are: 1) 5 to 9, 2) 10 to 19, 3) 20 to 49, 4) 50 to 199 and 5) more than 200. Table 3.1 presents the industry groups involved.

Table 3.1: Descriptions of the industry group codes in the NU2015 sample. Source: SCB (2016) and Eurostat (2008)

Industry Group Code	**Description**
A	Agriculture, forestry and fishing
C1	Manufacturing, labour intense sectors
C2	Manufacturing, science intense sectors
C3+B	C) Manufacturing, capital intense sectors and B) Mining and quarrying
D+E	D) Electricity, gas, steam and air conditioning supply E) Water supply, sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O+U	O) Public administration and defense, compulsory social security U) Activities of extraterritorial organizations and bodies
Poff	Education, public sector
Ppriv	Education, private sector
Qoff	Human health and social work activities, public sector
Qpriv	Human health and social work activities, private sector
R	Arts, entertainment and recreation
S+T	S) Other service activities T) Activities of households as employers, undifferentiated goods- and services-producing activities of households for own use

It should be noted that some of the strata were "closed" in the middle of the data collecting period. That includes all strata belonging to the smallest size class. One of the reasons for doing this is that it is relatively costly to peruse smaller organizations. (SCB, 2016)

For each stratum, the ambition have been to sample 30 observations applying the following order of priority: 1) choose an organization that have participated in CIS2014, 2) choose an organization that have participated in IT2015 and 3) simple random sampling. This procedure explains how the three surveys are coordinated.

3.1.1 Response rate of NU2015

There are several different suggestions on how to calculate and present the response rate of a survey. For this thesis we are mostly following the recommendations from the Swedish Statistical Society in their publication on standardized nonresponse calculations. (Svenska statistikersamfundet, 2005) They suggest that the survey sample gets divided into four main categories: responses, nonresponses, unknown status and over coverage. These categories should preferably be presented in sub-levels, such as "partial response", "refused participation" and so on. Sample units that are found not to belong to the target population are placed in the over coverage category. The units for which one is unable to determine whether they belong to the population or not, should be placed in the unknown status category.

Table 3.2: Distribution of responses in the NU2015 sample.

Category	**Frequency**	**Percent**
Responses	1619	52.2 %
Out of which:		
Telephone interviews	*1327*	*42.8 %*
Online surveys	*243*	*7.8 %*
Partial responses	*49*	*1.6 %*
Nonresponses	1416	45.7 %
Out of which:		
Refusals	*716*	*23.1 %*
Non-contacts	*651*	*21.0 %*
Others	*49*	*1.6 %*
Over coverage	66	2.1 %
Sum:	**3101**	**100 %**

The category distribution of the NU2015 sample is displayed in Table 3.2. Worth noting are the two different methods for participating in the survey. Most of the data collecting was done through telephone interviews. The online survey was sent as an alternative to those for

whom contact was not achieved but who didn't actively refuse to participate. Although possible, we will not differentiate between the two methods in the analysis of this thesis. Similarly, no weight will be put on the partial responses which only constitute a negligent proportion of the sample. However, it should be noted that the theoretical issues with non-answered questions are very similar and closely related to those of unit nonresponse. No certain unknown status category was necessary for this particular sample.

For NU2015, the over coverage mostly consist of companies that have ceased to exist since the sampling frame was created. Since the over coverage is determined not to belong to the target population of the study, this quantity should be omitted when calculating the response rate. Thus, the unweighted response rate of NU2015 could be expressed as

$$RR_u = \frac{n_r}{n_r + n_q} \tag{3.1}$$

where n_r is number of respondents and n_q is the number of non-respondents in the sample. With the definition of design weights as in equation (2.2), the weighted response rate could be expressed as

$$RR_w = \frac{\sum_r w_k}{\sum_r w_k + \sum_q w_k} \tag{3.2}$$

where $\sum_r w_k$ is the sum of the weights for those who have answered the survey and $\sum_q w_k$ is the sum of the weights for those who have not. Using equations (3.1) and (3.2), the response rates for NU2015 are found to be $RR_u = 0.533 = 53.3\ \%$ and $RR_w = 46.5\ \%$. The nonresponse rates are consequently $NR_u = 1 - RR_u = 46.7\ \%$ and $NR_w = 53.5\ \%$ respectively. Note that the weighted response rate will put more weight on small organizations, where the response rate is lower, because these are dominant in the population numerically. However, since the main focus of NU2015 is on larger organizations it could be appropriate to provide the figures when the smallest size group (5 to 9 employees) is excluded from the sample. In that case the response rates become $RR_u = 55.9\ \%$ and $RR_w = 53.1\ \%$.

The unweighted response rates for each stratum, as well as the size and industry groups, are presented in Table 3.3. As one might suspect, the response rate seems to steadily increase as an organization gets larger. The dispersion among the different industry groups is quite substantial, on the other hand. Industry group *O+U* comes out on top while *I*, *J* and *N* have rates quite much lower than the total average of 53.3 %. This could serve as an indication of possible nonresponse bias but one should be cautious to draw any conclusions using this information alone.

Table 3.3: Unweighted response rates for industry groups, size groups and strata in the NU2015 sample.

Industry Group	**Size Group**					
Code	5-9	10-19	20-49	50-199	200+	
A	*30.0 %*	*65.5 %*	*36.7 %*	*56.7 %*	*83.3 %*	48.8 %
C1	*51.7 %*	*63.3 %*	*33.3 %*	*53.3 %*	*56.7 %*	51.7 %
C2	*36.7 %*	*69.0 %*	*46.7 %*	*65.5 %*	*60.0 %*	55.4 %
C3+B	*53.3 %*	*50.0 %*	*46.7 %*	*50.0 %*	*46.4 %*	49.3 %
D+E	*36.7 %*	*43.3 %*	*60.0 %*	*76.7 %*	*75.9 %*	58.4 %
F	*26.7 %*	*53.3 %*	*41.4 %*	*51.7 %*	*63.3 %*	47.3 %
G	*23.3 %*	*40.0 %*	*57.1 %*	*53.3 %*	*43.3 %*	43.2 %
H	*32.1 %*	*43.3 %*	*63.3 %*	*43.3 %*	*64.3 %*	49.3 %
I	*36.7 %*	*23.3 %*	*34.5 %*	*46.7 %*	*51.7 %*	38.5 %
J	*31.0 %*	*33.3 %*	*36.7 %*	*36.7 %*	*55.2 %*	38.5 %
K	*46.7 %*	*50.0 %*	*63.3 %*	*44.8 %*	*71.4 %*	55.1 %
L	*40.0 %*	*53.6 %*	*65.5 %*	*46.7 %*	*62.5 %*	53.2 %
M	*46.7 %*	*43.3 %*	*60.0 %*	*50.0 %*	*60.7 %*	52.1 %
N	*48.3 %*	*39.3 %*	*48.3 %*	*43.3 %*	*44.8 %*	44.8 %
O+U	*68.4 %*	*73.9 %*	*81.5 %*	*86.2 %*	*72.4 %*	77.2 %
Poff	*46.4 %*	*63.3 %*	*70.0 %*	*76.7 %*	*75.0 %*	66.4 %
Ppriv	*56.7 %*	*72.4 %*	*56.7 %*	*75.9 %*	*65.5 %*	65.3 %
Qoff	*59.3 %*	*53.6 %*	*63.3 %*	*53.3 %*	*63.0 %*	58.5 %
Qpriv	*55.2 %*	*62.1 %*	*70.0 %*	*65.5 %*	*36.7 %*	57.8 %
R	*40.0 %*	*55.2 %*	*58.6 %*	*46.7 %*	*55.2 %*	51.0 %
S+T	*53.3 %*	*50.0 %*	*70.0 %*	*66.7 %*	*68.0 %*	61.4 %
	43.3 %	52.1 %	55.3 %	56.6 %	59.7 %	**53.3 %**

3.2 The LISA Database

The database of Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA) contains data on all people in Sweden older than 16. The focus of the data is on health, labour and educational variables, such as student aid, unemployment benefits, income, education level, sick leave benefits, and the like. It provides a stepping stone towards creating sampling frames for surveys regarding individuals, as well contributing valuable background data for such studies. For an individual who has an employment, the

organizational number (or work place number if the employer is a publically owned enterprise) of their employer is available and the individual can thusly be linked to an organization. By creating combinations of LISA and NU2015, age structures, gender shares and the like can be extracted on an organizational level.

We had some initial concerns with the fact that the LISA set at our disposal was from 2013. Values for specific organizations might have changed to some extent, as people do switch jobs or quit working. However, on an aggregated level, these differences should not have an impact on our overall analysis.

3.3 The FDB Database

The Statistical Business Registry (FDB) at Statistics Sweden keeps updated information on all Swedish enterprises, government offices and organizations, as well as their work places. It has a key role for constructing sampling frames for various surveys such as NU2015 and CIS2014. The database contains data on financial ratios, staffing, legal form, geographical information, industry class among others.

3.4 The CIS2014 Survey

With CIS2014 Statistics Sweden tries to measure the level and form of innovation in the Swedish business sector. It is a joint venture with Eurostat, and the survey is carried out across several European countries. The survey consists of 34 main questions, many with several sub questions, generating roughly 170 variables. Most of them are coded as dummies.

The questions range from "During the years 2012 to 2014, did your enterprise introduce new or significantly improved products?" and "During the years 2012 to 2014, did your enterprise receive any public financial support for innovation activities from The European Union (EU)" to enquiries regarding pending patents. In all, the survey attempt to capture the elusive subject of innovation. With this data, we can control for whether different variables regarding innovation affect the propensity to answer the NU2015 survey.

With the exception of some industry and size groups that were included altogether, CIS2014 applied stratified sampling with aim of getting 7 observations within each stratum. The stratifying variables were: size group, industry group and geographical region. In total 9348 enterprises were included out of which 8159 responded, yielding a response rate of 87.3 %. (SCB, 2014)

3.5 The IT2015 Survey

The IT2015 survey is designed much like the CIS2014, in so that it also focuses on the Swedish business sector and is carried out in co-operation with Eurostat. It consists of 39 main questions, with a fair deal of sub questions generating somewhere around 100 variables. The question ranges from "Does your enterprise employ ICT specialists?" and "Does your enterprise have a Website?" to "How does your enterprise share supply chain management information electronically?" The aim is to capture the matter from many angles to give as good picture as possible of the ICT usage in enterprises.

The sample was drawn much in the same way as CIS2014 was. With some exceptions, a stratified random sample was drawn, based on size, industry and geocode. At least 5 enterprises was drawn in each stratum, wherever possible. If the number of enterprises in a strata was less than five, all were sampled. The number of units in the sample was 4595 and 3855 responded to the survey, giving a response rate of 83.9 %. (SCB, 2015)

3.6 Loss of information when combining the datasets

First of all, for the organizations in NU2015 that have been created after 2013 we won't be able to extract any information from the LISA database. This quantity forms an under coverage of sorts and consists of 74 organizations. Subtracting these and the over coverage of 66 units, specified in section 3.1.1, gives a total of $3101 - 66 - 74 = 2961$ observations.

Further, since the target populations of CIS2014 and IT2015 constitute a subset of the NU2015 population (Swedish business sector as opposed to the Swedish working life), the size and industry groups of the surveys will be somewhat different. Because of this, the usage of study variables from CIS2014 and IT2015 on the NU2015 sample will result in loss of eligible observations. Table 3.4 specifies how many of the units in NU2015 that remain when utilizing information from the other two surveys, separately and in combination. As seen in the table, when utilizing information from both, roughly a third of the total NU2015 sample units remain. Note that the "No"-frequencies of the table should *not* be interpreted as the sample nonrespondents of those surveys.

Table 3.4: Frequency table of CIS2014 and IT2015 respondents in NU2015 sample.

		Have answered CIS2014		
		Yes	No	**Total**
Have answered IT2015	Yes	993 (33.5 %)	171 (5.8 %)	1164 (39.3 %)
	No	255 (8.6 %)	1542 (52.1 %)	1797 (60.7 %)
	Total	1248 (42.1 %)	1713 (57.9 %)	**2961** (**100 %**)

4 Results and Analysis

4.1 Mean comparisons of main structural factors

The set of variables, as well as the general procedure of this subchapter, is chosen with the intention of facilitating a follow-up using the official nonresponse analysis conducted on NU2012 as a reference. (Arbetsmiljöverket, 2014a) The names and explanations of these 9 variables, which are referred to as the *main structural factors* throughout this thesis, are listed in Table 4.1. All of them have been constructed with the LISA database of individuals, which is described in section 3.2.

With one exception, the factors used here are essentially the same as in the earlier analysis. The reasoning from the authors as to why these specific variables should be analyzed was primarily based on theory, which we agree upon. Following the recommendations of that report, the number of intervals for both the age and education variables have been reduced from six to three in our study. The purpose of doing this is to avoid getting excessively small sub-groups when analyzing the variables on industry or size group levels.

Table 4.1: Names and descriptions of main structural factors.

Variable	Description
Wom	Proportion of women in organization
ForBack	Proportion of employees with foreign background in organization
Age0_34	Proportion of employees between 0 and 34 years old
Age35_54	Proportion of employees between 35 and 54 years old
Age55	Proportion of employees older than 55 years
EducElm	Proportion of employees with elementary school as maximum education level
EducUpSec	Proportion of employees with upper secondary school as maximum education level
EducHigh	Proportion of employees with higher education as maximum education level
AvgInc	Average yearly income per employee in organization

We have also added a variable which wasn't included in the earlier analysis. `ForBack` specifies the proportion of people with foreign background employed by each organization

in the sample.[4] We suspect that it might help to explain both the propensity to respond to the survey as well as the study variables of interest.

For each of the main structural factors the means, both weighted and unweighted as defined in equations (2.2) and (2.3), are calculated for the response subset as well as the sample as a whole. In practice, the comparisons of these two quantities yield a ratio, calculated as in equation (2.1), which is rounded to its first decimal place. Naturally, when this ratio equals unity it is an indication that there are no or small differences between the two groups. This, in turn, would suggest that there is a small probability of nonresponse bias in the study estimates if that specific auxiliary variable is used in estimation.

Table 4.2: Comparisons of unweighted and weighted means among the main structural factors using the full NU2015 sample.

	Unweighted means			**Weighted means**		
Variable	Sample	Responses	Ratio	Sample	Responses	Ratio
Wom	0.435	0.456	1	0.442	0.486	1.1
ForBack	0.169	0.155	0.9	0.170	0.156	0.9
Age0_34	0.326	0.300	0.9	0.345	0.317	0.9
Age35_54	0.468	0.476	1	0.453	0.459	1
Age55	0.206	0.224	1.1	0.201	0.224	1.1
EducElm	0.115	0.108	0.9	0.122	0.109	0.9
EducUpSec	0.499	0.485	1	0.541	0.527	1
EducHigh	0.386	0.407	1.1	0.337	0.364	1.1
AvgInc	339 182	341 123	1	305 009	300 095	1

The results of this approach on the whole sample can be seen in Table 4.2. They suggest that organizations with a higher degree of employees with foreign background as well as the lowest age and education level are slightly underrepresented in the survey. Similarly, the organizations with an older and more educated work force are slightly overrepresented. Although most the ratios depart from unity these figures aren't particularly worrisome. While a deviation of 10 % isn't ideal it must be deemed acceptable in this context. This assessment is in accordance with the earlier nonresponse analysis conducted on NU2012. (Arbetsmiljöverket, 2014a) Also worth noting is that almost all of the unweighted and weighted comparison ratios, with the exception of `Wom`, correspond. This could be an indication that smaller organizations with a higher proportion of women respond to a larger extent.

[4] A person with foreign background is being defined as someone who is born outside Sweden or who has two parents born outside Sweden.

Figure 4.1 illustrates the ratios when applying the same technique with only unweighted means for each level of the stratifying variables `sgrupp` and `bgrupp`. The number of resulting ratios becomes 9 factors $\times$ 5 size groups $+$ 9 factors $\times$ 21 industry classes$=$ 234. From the graph it is clearly visible that the vast majority of ratios are quite close to 1. In fact, 84.6 % of the response mean ratios are within 10 % deviation from unity. Even the values at the end of the tails aren't particularly extreme. This shows that the responses of NU2015 are quite suitable for descriptions on a detailed, size or industry group, level.

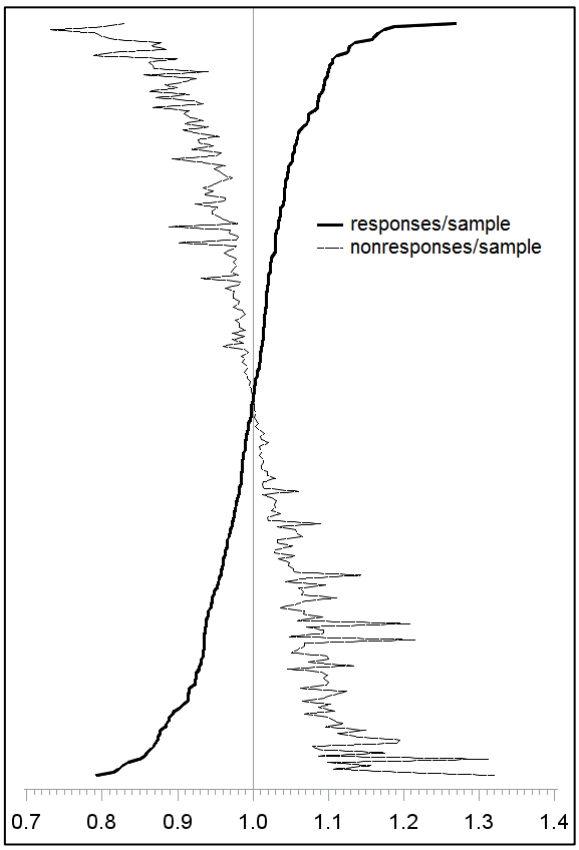


Figure 4.1: Distribution of 234 unweighted mean comparison ratios.

All in all, the results of this approach show that the response subset of the sample represent the target population – the Swedish working life – relatively well.

4.2 Logistic regression models

For the logistic regression approach two principal models are fitted. The results of the General Model, which only utilizes the main structural factors (presented in Table 4.1) along with the stratifying variables, are presented in section 4.2.1. In section 4.2.2, additional auxiliary variables from the CIS2014 and IT2015 surveys are selected through the Backward Elimination procedure. In section 4.2.3, the results of the Extended Model are presented.

The tables that are presented in each model section display the estimated betas, their standard errors and the Wald Chi Square p-values. The categorical stratifying variables `sgrupp` and `bgrupp` follow a standard dummy design where the levels of each variable become separate dummies. This is achieved by instructing SAS to use *reference* (as opposed to *effect*) *parametrization*. The choice of reference groups are discussed below.

Since the age and education intervals are defined as proportions they, respectively, sum to 1 for each observation. Because of this one of the groups has to be omitted in order to avoid multicollinearity. We have chosen to omit the second group in both cases, as to display the effects of having a more or less educated and older or younger working force relative to the middle group. Further, because of the unit measurement of the average income variable, we have taken the natural logarithm of `AvgInc` as to ease the interpretation of the estimated betas.

Specific variables of interest are discussed throughout the chapter. Those that have significant effects in both models are analyzed in section 4.2.3. Lastly, there will be some comments on the predictive power of both models in section 4.2.4.

4.2.1 General Model

For the General Model we are able to use the whole NU2015 sample consisting of 2961 observations, out of which 1580 responded and 1381 did not. The second size group (10-19 employees) and *L* (Real estate activities) are set as the baseline levels for `sgroup` and `bgroup` respectively. These were chosen because they are the ones closest to the total average in terms of response rate. Thus, the number of included explanatory variables are 31 out of which 7 are main structural factors, 4 are size groups and 20 are industry classes. The model that is being estimated can be expressed as

$$\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 Wom + \beta_2 ForBack + \beta_3 Age0\_34 + \beta_4 Age55 + \beta_5 EducElm + \beta_6 EducHigh + \beta_7 \log(AvgInc) + \beta_7 sgrupp_1 + \ldots + \beta_{11} sgrupp_5 + \beta_{12} bgrupp_A + \ldots + \beta_{31} bgrupp_{S+T}$$

where p is the propensity to respond in the survey. The results of the estimation are presented in Table 4.3.

Table 4.3: Results of the General Model. Significant estimates are bolded.

Parameter	Group	Estimate	S.E.	p-value
Intercept		1.5227	2.1150	0.4715
Wom		0.2273	0.1990	0.2533
ForBack		-0.6548	0.2295	**0.0043**
Age0_34		-0.5561	0.2645	**0.0355**
Age55		0.5863	0.3351	0.0802
EducElm		-0.1578	0.4037	0.6960
EducHigh		0.0970	0.2184	0.6568
AvgInc_log		-0.1170	-0.1170	0.1611
sgrupp	**1**	-0.3707	0.1215	**0.0023**
sgrupp	3	0.1208	0.1188	0.3094
sgrupp	4	0.1890	0.1190	0.1122
sgrupp	**5**	0.3068	0.1232	**0.0128**
bgrupp	A	0.0244	0.2589	0.9249
bgrupp	C1	0.0745	0.2454	0.7615
bgrupp	C2	0.2192	0.2443	0.3696
bgrupp	C3+B	-0.0785	0.2471	0.7508
bgrupp	D+E	0.3190	0.2460	0.1948
bgrupp	F	-0.0051	0.2534	0.9841
bgrupp	G	-0.2311	0.2451	0.3457
bgrupp	H	-0.0397	0.2494	0.8736
bgrupp	I	-0.3069	0.2632	0.2436
bgrupp	J	-0.4273	0.2519	0.0898
bgrupp	K	0.1668	0.2488	0.5026
bgrupp	M	0.1093	0.2498	0.6618
bgrupp	N	-0.0960	0.2526	0.7040
bgrupp	**O+U**	0.9581	0.2781	**0.0006**
bgrupp	Poff	0.4219	0.2732	0.1225
bgrupp	Ppriv	0.4874	0.2622	0.0631
bgrupp	Qoff	0.1178	0.2661	0.6580
bgrupp	Qpriv	0.1845	0.2555	0.4704
bgrupp	R	0.0214	0.2486	0.9314
bgrupp	S+T	0.3340	0.2502	0.1819

Size Group

From the results in Table 4.3 it seems like the size groups do play an important role for the propensity to respond to NU2015. One would suspect that bigger organizations, with bigger resources, are more likely to respond to surveys in general. The signs and effects of the estimated coefficients for the levels of `sgrupp` (although they are not all significant) show a clear pattern, consistent with this theory. These results shouldn't come as a surprise given the response rates reported in Table 3.3 (the same goes for the industry groups, discussed further below). However, calculating the odds ratios clarifies this picture even more.

Table 4.4: Odds ratios of size classes vs. the reference group in the General Model.

			95 % Wald C.I.	
Parameter	**Group**	**Point estimate**	**Lower**	**Upper**
sgrupp	1 vs. 2	0.690	0.544	0.876
sgrupp	3 vs. 2	1.128	0.894	1.424
sgrupp	4 vs. 2	1.208	0.957	1.525
sgrupp	5 vs. 2	1.359	1.067	1.730

With the estimates of Table 4.4 we can compare the organizations in the smallest size group with those in the second smallest. The odds of responding is 31 % lower in the smaller group. Also, note that the 95 % Wald C.I. does not contain the value 1 indicating that the effect indeed is significant. For the largest organizations, the odds are 36 % higher compared with the second largest.

Since `sgrupp` is one of the variables on which the stratification process is based and the estimates are weighted by, it was *a priori* a top candidate for **Z**. The results are in accordance with this prior belief. They are also in accordance with the response rates reported in Table 3.3. On an aggregated level, the propensity to answer rises steadily from the smallest size group to the largest. Inside the various industry groups, this picture is less clear, but still quite persistent.

As mentioned briefly in Section 3.1, some of the strata, including the smallest size group, were "closed" in the middle of the data collection period. Had these strata not been closed, the response rate might have been slightly higher. In any case, the closing of these strata might have *exaggerated*, but not *caused*, the low response rate.

Industry Group[5]

As seen in Table 4.3, `bgrupp` seem to affect the extent to which organizations answer the NU2015 survey. The most striking result is the estimated, highly significant, effect of belonging to the industry group *O+U*. Per definition, these organizations are governmental authorities, such as Statistics Sweden. The odds ratio between *O+U* and the baseline group *L* is 2.607, with 95 % Confidence Limits 1.511 and 4.496, which is the largest deviation

[5] See Table 3.1 for detailed descriptions of the industry codes.

among all groups. After industry group *O+U,* the second and third largest positive effects are those of *Ppriv* and *Poff;* private and public education.

On the other hand, the largest *negative* (and second largest in absolute terms) coefficient is that of the *J* group. These organizations deal with information and communication, producing and broadcasting television shows, publishing and the like. Calculating the estimated odds ratio between *J* and the baseline *L* gives that the odds ratio 0.652. The boundaries of the Confidence Interval are 0.398 and 1.069.

Overall, the industry groups concerned with official matters, such as health and education have positive coefficients. One possible explanation is that these organizations feel more obliged to respond to surveys, since they (at least through public funding in the case of private education and health care) are operating in the public sphere. The response rates presented in Table 3.3 are somewhat in accordance with these results.

4.2.2 Stepwise Backward Elimination procedure

The number of variables entering the *Stepwise Backward Elimination* (SBE) procedure are 49, out of which 7 are the main structural factors, 3 are size groups, 10 are industry classes and where 29 stem from the CIS2014 and IT2015 surveys. Only these 29 survey variables are considered for elimination in the SBE procedure. By design, SAS is instructed to keep the other ones in order to control for their effect on the propensity. The dataset used consists only of 993 observations. The reason for this, and the reduced number of size and industry classes, are the loss of information that occurs when we combine the datasets. This is further explained in section 3.6.

The cutoff point at which variables are omitted in each step was set to 0.2. The study variables of CIS2014 and IT2015 that made the cut are displayed in Table 4.5.[6] Each of these are discussed and considered below.

The indicator `Mareur14` is set to be 1 if the organization sold goods or services on the European market in the years 2012 - 2014 and 0 otherwise. `newprocessD14` is a dummy indicator which is set to be 1 if the organization has answered yes to at least one of the following questions:

> During the three years 2012 to 2014, did your enterprise introduce
>
> a) new or significantly improved methods of manufacturing or producing goods or services
> b) new or significantly improved logistics, delivery or distribution methods for your inputs, goods or services
> c) new or significantly improved supporting activities for your processes, such as maintenance systems or operations for purchasing, accounting, or computing?

and 0 otherwise. (SCB, 2014)

`itspt2` is also a dummy indicator which is 1 whenever an organization has offered their ICT specialist staff any ICT related education during 2012 through 2014. Given the results of Table 4.5 and the lack of theories for why the variable should be relevant, we have chosen

[6] The eliminated variables and their associated p-values can be seen in Appendix A.

to exclude it from the Extended Model. First of all, the effect on the propensity to respond is non-significant and relatively small. Secondly, and perhaps more importantly, we cannot find any theoretical reason for how this variable should affect the way organizations answer the NU2015 questions. The question becomes if different values of `itspt2` implies different answers on the NU2015 questions, and we believe that they should not, to any greater extent. One consequence of disregarding `itspt2` from further analysis is that we do not need to account for the IT2015 data in the Extended Model. This grants us 255 additional observations, otherwise omitted.

Table 4.5: Variables kept in Stepwise Backward Elimination procedure with cutoff point set to 0.2.

Parameter	Estimate	S.E.	p-value
Mareur14	-0.5355	0.1595	0.0008
newprocessD14	0.4137	0.1576	0.0086
itspt2	0.2634	0.1891	0.1637

4.2.3 Extended Model

The dataset used for estimating the Extended Model contains 1248 eligible observations, out of which 668 responded to NU2015 and 580 did not. The number of explanatory variables are 23 where 7 are the main structural factors, 2 stem from CIS2014, 3 are size groups and 11 are industry classes. Again, the second size group is set to be the reference group. For `bgrupp` *M* is chosen with the same logic as described in section 4.2.1. The Extended Model reads:

$$\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 Wom + \beta_2 ForBack + \beta_3 Age0\_34 + \beta_4 Age55 + \beta_5 EducElm + \beta_6 EducHigh + \beta_7 \log(AvgInc) + \beta_8 newprocessD14 + \beta_9 mareur14 + \beta_{10} sgrupp_3 + \ldots + \beta_{12} sgrupp_5 + \beta_{13} bgrupp_{C1} + \ldots + \beta_{23} bgrupp_N \, .$$

Table 4.6: Results for the Extended Model. Significant estimates are bolded.

Parameter	Group	Estimate	S.E.	p-value
Intercept		4.5347	3.5824	0.2056
Wom		0.2465	0.3421	0.4712
ForBack		-1.2149	0.4226	**0.0040**
Age0_34		-1.0960	0.4750	**0.0210**
Age55		-0.0932	0.6833	0.8915
EducElm		-0.5564	0.7989	0.4861
EducHigh		0.2922	0.4205	0.4872
AvgInc_log		-0.3057	0.2743	0.2649
newprocessD14		0.3464	0.1335	**0.0095**
Mareur14		-0.4709	0.1373	**0.0006**
sgrupp	3	0.1386	0.1684	0.4104
sgrupp	4	0.1430	0.1683	0.3956
sgrupp	5	0.3223	0.1728	0.0621
bgrupp	C1	0.1036	0.3412	0.7615
bgrupp	C2	0.3822	0.3164	0.2271
bgrupp	C3+B	-0.1398	0.3428	0.6835
bgrupp	D+E	0.1886	0.3160	0.5505
bgrupp	F	-0.0090	0.3450	0.9791
bgrupp	G	-0.1993	0.3107	0.5213
bgrupp	H	0.0977	0.3355	0.7708
bgrupp	I	-0.1792	0.3587	0.6174
bgrupp	**J**	-0.5736	0.2853	**0.0444**
bgrupp	K	-0.0263	0.3043	0.9312
bgrupp	N	-0.1942	0.3273	0.5530

New Processes

As seen in Table 4.6, `newprocessD14` seems to affect the propensity to respond positively, although with a moderate magnitude. The point estimate of the odds ratio is 1.414, and the 95 % CLs are 1.088 and 1.837, making the odds for organizations with newly implemented processes 41 % larger.

For an instance, let's remember how the `newprocessD14` was constructed, as described in section 4.2.2. The indicator can itself be broken down to three other dummies, namely `newprocess_prod14`, `newprocess_logis14` and `newprocess_supp14`. These, in turn, contain information on whether the organizations have implemented new production, logistic and support processes, respectively. If the `newprocessD14` is supplemented with

these three in the Extended Model, the individual effect of each can be measured. As it turns out, the effects of all three are non-significant.

There is also another variable that was not included in the SBE step, because its information was included through other variables. It is called `noiprodprocinnovD14`, and is an indicator which is 1 whenever both `newprocessD14` and `newproductD14` are 0. Thus, it is an indicator of non-innovative companies. This aggregated indicator has a significant negative effect which implies that more innovative organizations have a higher propensity to respond. When dividing the aggregated `noiprodprocinnovD14` into the two sub-indicators, it is clear that this effect comes from the process part – the product indicator does not pass the SBE step. This is illustrated in Figure 4.2.

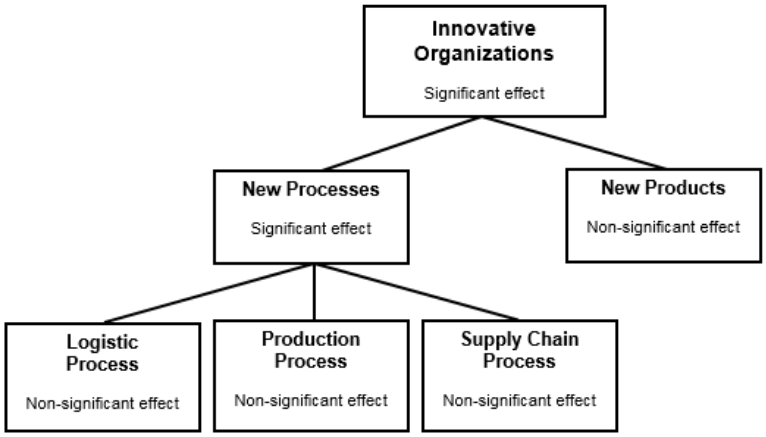


Figure 4.2: Breakdown of the aggregated innovation variable.

We have not found any research on the subject of why this might be, but one hypothesis is that organizations with newly implemented processes might be eager to "show them off". If, as there might be reasons to suspect, new processes in some way affect the answers in NU2015, this could be a source of bias.

However, there are also reasons to believe that this is an artifact of the data. There is another dummy indicator called `neworgD14`, which measures organizational innovation. It seems to us that this kind of activities should have a stronger bound to the questions in NU2015 and its effect was removed in the SBE step. The same is true for the new product indicator. The CIS surveys are conducted in such a way that information regarding new product innovation are sought after in more detail than new processes. The reason for this is that it is believed that product innovation simply is more important than the process contra part, although this is a question of some debate. If, however, the "show off" hypothesis were to be true, this should show in the `newproductD14` as well. Out of those who *did not* implement new processes, 50 % responded to the NU2015 survey. Out of those who *did,* 60 % responded.

Selling goods or services on the European market

The other effect that is specific for the Extended Model is that of `Mareur14`. It is set to be 1 if the organization sold goods or services on the European market. The odds ratio of those who were active on this market compared to those who were not is 0.624 with the 95 %

Confidence Limits being 0.477 and 0.817. Thus, companies who act on the European market had 37.6 % lower odds than those who did not, according to our model.

We could not find any prior research on the subject. One hypothesis might be that organizations that act on the European market are less concerned with domestic matters. Companies use official statistics to make decisions and if they only act on local or domestic arenas, the incentives to "pull their weight" might be larger.

There is also evidence against this. A world market indicator, `Worldmark14`, entered in the SBE stage, and the effect was removed. For the offered hypothesis to hold true, it is reasonable to believe that it would be so also at the world market level. The raw number tells us that out of those who *did* sell goods in Europe, half did respond to the NU2015. Out of those who *did not* sell goods at the European market, 56 % responded.

In addition, it is hard to come up with any plausible hypothesis regarding this variable's effect on work environment. So, even if the effect on the propensity to answer is real, the results of having relatively few responses from these companies should not introduce bias in the results.

4.2.4 Common Effects

Age Group
It seems like having a younger working force affects the propensity to respond negatively. It is known that younger people in general respond to surveys to a lesser extent, so the result is not all that surprising. The Swedish Work Environment Authority conclude in a report from 2013 that younger people to a higher degree work in environments that are associated with less secure forms of employment (such as part time and fixed-term contracts) and relatively high degrees of work related hazards. (Arbetsmiljöverket, 2013)

Table 4.7: Odds ratios of the age intervals, both models.

	General Model			Extended Model		
Parameter	**Point estimate**	**95 % C.L. Lower**	**95 % C.L. Upper**	**Point estimate**	**95 % C.L. Lower**	**95 % C.L. Upper**
Age0_34	0.573	0.341	0.963	0.334	0.132	0.848
Age55	1.797	0.932	3.467	0.911	0.239	3.477

According to our models, the odds ratios read as in Table 4.7. Note that given how these variables are defined, a unit increase is equivalent to when the proportion of an age group goes from 0 to 100 %. This fact underlines the importance of a very cautious interpretation. However, regardless of which specification is used, the effect of having a younger staff is clear. Even though `Age55` only is significant at a 10 % level in the General Model specification (and not at all in the extended one), it follows the predicted pattern. It is shown that age matters when it comes to work environment and there is a risk that organizations that employ predominantly youths opt out of responding to surveys such as NU2015 due to this. These results are in line with those of the mean comparisons in Table 4.2.

Share of employees with foreign background
From the results of both models it is clear that organizations with a larger share of employees with foreign background seem to have responded to a lesser extent. Once again, compared to the findings in Table 4.2 these results are in accordance. This might have the effect that important aspects of the working environment is lost. It is known, for instance, that individuals with foreign background are less likely to be unionized. (LO, 2002) This subset of the population, like with the young, is also more likely to work in environments that are associated with less secure forms of employment. (SCB, 2009)

One way to illustrate this tendency is the graph of Figure 4.3. The odds ratio function is calculated as $\exp(\hat{\beta}_2 ForBack)$ for both models when the foreign background variable goes from 0 to 1 (since it's measured as a proportion). Again, given how the variable is defined and the unrealistic nature of the respective intercepts, it's important to be careful with the conclusions about this graph. What we want to show is that the estimated odds of responding to the survey decrease when the share of employees with foreign background increases.

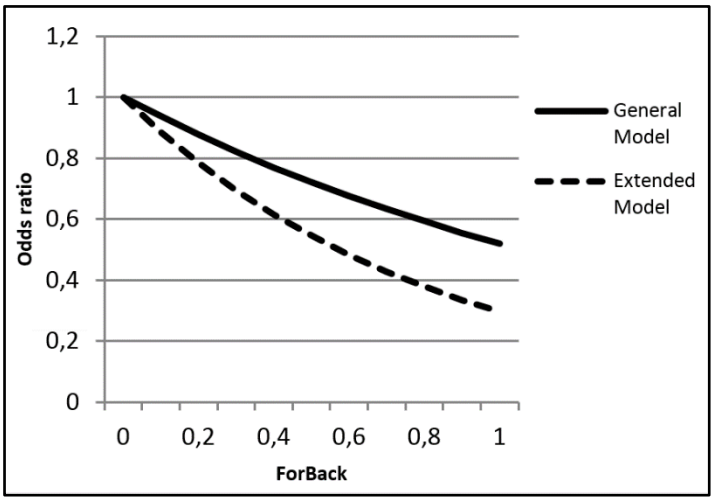


Figure 4.3: Odds ratios of the foreign background variable for both models, relative to their respective intercepts.

For both models, we have tried to add several different interaction combinations between `ForBack` and other variables, such as `sgrupp` and `bgrupp`. None of these attempts resulted in significant effects, which is why these estimates aren't reported.

Share of female employees
The share of female employees was under special consideration in the NU2012 nonresponse analysis, the verdict then being that it did not seem to matter to any greater extent. (Abetsmiljöverket, 2014a) In our case, as per Table 4.2, the unweighted mean share was 0.435 in the sample and 0.456 in the response group. In fact, according to the same table, women seem to be slightly overrepresented according to the weighted means. The logistic models estimate that the effect of having different shares of female employees does not matter. In contradiction with our prior belief, even when fitting a model with just `Wom` as an explanatory variable, the effect turns out to be small and nonsignificant. Our prior belief was that `bgrupp` sufficiently should eliminate the effect of having more women employed. Also, we

were surprised to see that the interaction between the share of women and employees with foreign background was nonsignificant.

Average Income and Education
It might be reasonable to assume that both the average income and education level affect the work environment. However, according to our results, they do not seem to affect the response propensity of NU2015.

4.2.5 Assessing the Models

The *tau-a* measurements of predicted power, defined in equation (2.9), are reported in SAS to be 0.135 and 0.132 for the General and Extended Model respectively. Both are undoubtedly weak figures. Since the Extended Model has a lower figure than the General Model, one could definitely argue that the added auxiliary variables of CIS2014 do not seem to be of much importance. However, despite its name, the Extended Model has less explanatory variables than the General Model. The reason for this is that the target population of CIS2014 constitute a subset of the NU2015 population. Again, this is further explained in section 3.6. It's quite possible that the difference between the two models, with regard to the *tau-a* figures, stems from the discrepancy in number of variables, rather than the variables themselves.

The stated objective when fitting these models was never to predict the propensity to respond, but to find out which variables that might affect that propensity. As such, we do not believe that the reported measures of *tau-a* are particularly worrisome.

5 Conclusions

Let us now return to Figure 1.1. Assuming that the effects of the variables discussed above are real, do they belong to **Z** or are they indicating something about the relationship between the response propensity, p, and the study variables, **Y**? Please note that the conclusions drawn in this segment are relevant for the NU2015 survey only.

In this thesis we have found several possible sources of nonresponse bias. The auxiliary variables in question are discussed separately below. Assuming that the effects of these variables are real, with the theoretical framework of Groves (2006) they can be explained by two different scenarios:

1) the variables affect the response propensity and the survey variables separately (Common Cause Model), or
2) the organizations have different characteristics, some of which measured in NU2015, that affect the propensity to respond to the survey (Survey Variable Cause Model).

First, let us consider the estimated effects of the CIS2014 variables, measured in the Extended Model. For the case of selling goods on the European market, it is hard to come up with a hypothesis regarding the effect on the work environment. In addition, it seems unlikely that activities on the European market should matter, when the world market contra part does not. *We conclude that* `Mareur14` *does not belong to* **Z**, *regardless of whether the effect on* p *is real or a false positive.*

According to our model, implementing new processes in an organization seem to affect p positively. Although we lack prior research on the topic, one hypothesis might be that organizations that are implementing new processes want to "show them off". It is possible that these new processes in some way affect the work environment. However, the evidence is weak and as stated above, it is plausible that the significance is a false positive. Implementing new organizational innovations or products should, if our hypothesis were to hold, also invoke this willingness to show off. *It is possible, however unlikely, that* `newprocessD14` *belongs to* **Z**.

Now, let us turn our attention towards the effects of the stratifying variables `sgrupp` and `bgrupp`. The most interesting finding regarding the size of the organizations is that the pattern matches the preexisting theory well. Smaller organizations respond less often, bigger respond more. It is also highly plausible that the number of employees affect the work environment, at least in the context of the NU2015 survey. Having a smaller organization decreases the likelihood that knowledge regarding various important aspects of the work environment exist within that organization. *Thus, our conclusion is that* `sgrupp` *belongs to* **Z** *and that the use of it as a stratifying variable is wise.*

What about the industry classes? There is only one significant deviation from the baseline, measured in the General Model, and that is if the organizations belong to *O+U*. These organizations are, as previously mentioned, mostly governmental. It is not all that surprising that these organizations respond to a higher degree than the rest, even though NU2015 is an explicitly non-mandatory survey. One could argue that authorities are obliged to respond anyway due to the regulations of Sweden. A rundown on The Public Access to Information and Secrecy Act states that the "authorities must on request also provide each other with

such information at their disposal that is not subject to secrecy." (Regeringskansliet, 2009, p. 20) It is also known that employees in different industries have different experiences regarding work environment issues. According to recent research, industry group is the third most important factor when it comes to explaining individual work-related hazards. (Arbetsmiljöverket, 2014b) Since these deviations are very important, although the results from the logistic regression approach are relatively weak, *we conclude that* `bgrupp` *belongs to* **Z**.

When it comes to the main structural factors – such as education, average income, age, gender and foreign background – a more detailed discussion might be in place. With the share of female employees, we start by noting that there is prior research that indicates that males are more over confident than females. This is one plausible explanation as to why males tend to answer surveys to a larger extent than women. (Bengtsson *et al.*, 2004) However, it is not the company as a whole that answers the survey, it is one person – and so should the share of females really matter? It seems like it does not in this case. Even fitting a model with just `Wom` as the explanatory variable turned out to be nonsignificant. Male and female employees have, on average, different experiences regarding the work environment and there exist important differences that the statistics need to reflect in order to be relevant. (LO, 2014) It seems like that is the case with NU2015 since the share of female employees does not seem to affect $\boldsymbol{p}$. *Therefore, we conclude that share of female employees does not belong to* **Z**.

It seems plausible that the average income and education level among the employees of an organization do affect the work environment through a positive correlation. However, according to our analysis these variables do not affect the propensity to answer, and therefore *neither average income nor level of education do belong to* $\boldsymbol{Z}$.

According to our analysis, organizations with a higher proportion of young employees respond to the NU2015 survey to a lesser extent. It is known that younger people, in general, are poor respondents in relation to older contra parts. (Groves *et al.*, 2000) However, it is not the organization as a whole that respond to a survey, and so should the age structure affect the propensity to answer? According to our models, it does. We suspect that the age structure is a proxy variable for other aspects of the organizations. As discussed in the analysis section, younger people do, in general, have different work environment experiences than the average employee. For instance, they are more likely to have insecure forms of employment.

The same is true for people with foreign background. We have found strong evidence to suggest that organizations with a relatively larger share of employees with this attribute respond to a lesser extent. Just as with the youth, this subgroup is also overrepresented in having relatively insecure forms of employment.

Do this qualify age structure and `ForBack` as candidates for **Z**? *We believe that they do not.* To us, it does not seem likely that these variables should have a direct effect on the propensity to answer. Rather, they should be viewed as indicators, or proxies, for a variable for which we cannot control – the actual work environment itself. To us, it seems plausible that organizations with less satisfying work environment have lower propensity to respond to a survey regarding this specific subject. This would correspond to what Groves (2006) calls the Survey Variable Cause Model, mentioned above. Regardless of if these two factors belong to **Z** or not, we believe that this conclusion could be used to produce more precise estimations.

We suggest that it would be reasonable to adjust the inference made from the collected data of NU2015, as well as similar studies in the future, with respect to the age structure and share of employees with foreign background in each organization. *Calibration* and *post-stratification* are two methods often used to achieve this objective. The application of these techniques are beyond the scope of this thesis, but we would recommend the works of Särndal and Lundström (2005) on the subject. Although our logistic models indeed produce estimates that could be used for these purposes, we wouldn't necessarily recommend it without further investigation.

As stated in the Thesis Objective section, we will also draw conclusions about the two methods used in this study. First, although the outcome of the mean comparison approach is not as clear, most of the results of the two methods correspond to some degree. One noticeable difference is that with the logistic regression approach there are no significant estimates regarding the education level of the work force.

As a method, comparing the weighted and unweighted means is a straight forward and easy way of detecting deviations in the main structural factors. However, in accordance with the earlier nonresponse analysis (Arbetsmiljöverket, 2014a) the comparison on a detailed level is presented as a graph in Figure 4.1. The interpretation is not really straight forward, and there is no way of telling which of these deviations that are caused by which structural factor. As such, the method loses in analytical power in our opinion. Also, the method of rounding to one decimal place and stating that a 10 % deviation is acceptable does, indeed, seem quite arbitrary.

The main strength of fitting logistic models is that they allow us to test and measure the effects of several variables at once while controlling for their effect on one another. Although statistical tests can, and do, over- and underestimate the significance of effects, they do provide clear-cut values. With these we can discriminate between the variables. We have tried not to lean too heavily on the p-values, but rather examine the trends and general directions of the data. Logistic regression serves this purpose sufficiently, in our opinion. For the end-user, however, the method might be harder to comprehend. Odds and odds ratios are not as settled in the common users mind as proportions and probabilities. As briefly mentioned above, the estimates produced with this method could actually be used to reduce the nonresponse bias. For the objective of this thesis, we conclude that the logistic regression approach is far superior.

Although no additional investigation was conducted on *mode effects* (i.e. the possible differences between those who responded via telephone interview compared to the internet questioner), there might be reasons to do this in the future. If there are differences between those who responded to the survey and those who did not with regard to the study variables **Y**, the bias might actually increase.

Another suggestion is for Statistics Sweden to start measure the work effort of those who are collecting the data from the sample. This could be done by simply counting the number

of attempts to contact each sample unit. Such work effort variable would be of great value when conducting a nonresponse analysis, such as this one, in the future.

References

Allison, P. D. (2012). *Logistic Regression Using SAS®: Theory and Application.* Cary, NC: SAS Institute Inc.

Arbetsmiljöverket (2013). *SAM-index. Ett sätt att belysa systematiskt arbetsmiljöarbete i svenskt arbetsliv – baserad på Arbetsmiljöverkets Nulägesundersökning SAM 2012.* Arbetsmiljöverkets analysrapport 2013:2.

Arbetsmiljöverket (2014a). *Bortfallsanalys - Representerar de svarande organisationerna i Arbetsmiljöverkets Nulägesundersökning 2012 svenskt arbetsliv?* Arbetsmiljöverkets analysrapport 2014:1.

Arbetsmiljöverket (2014b). *Riskfaktorer för arbetsolycka - bakomliggande faktorers inverkan på individens olycksrisk.* Arbetsmiljöverkets analysrapport 2014:2.

Armstrong, J. S. and Overton, T. S. (1977). Estimating Nonresponse in Mail Surveys. *Journal of Marketing Research,* (14), 396-402.

Bengtsson, C., Persson, M., and Willenhag, P. (2004). Gender and overconfidence. *Economics Letters,* (86), 199-203.

Bethlehem, J. (2001). Weighting Nonresponse Adjustments Based on Auxilary Information. In R.M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse.* New York: Wiley.

Eurostat (2008). *Statistical Classification of Economic Activities in the European Community, Rev. 2.* http://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.pdf

Groves, R. M. (2006). Nonresponse Bias in Household Surveys*, The Public Opinion Quarterly, 70*(5), 646-675.

Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-Saliency Theory of Survey Participation: Description and an Illustration, *The Public Opinion Quarterly, 64*(3), 299-308.

Lineback, J. F., and Thompson, K. J. (2010). Conducting Nonresponse Bias Analysis for Business Surveys. Section on Government Statistics – JSM 2010.

LO (2002). *Röster om facket och jobbet.* http://www.lo.se/home/lo/res.nsf/vres/lo_fakta_1366027492914_rosteromfacket1_text_pdf/$file/rosteromfacket1_text.pdf

LO (2014). *Arbetsmiljö 2014 - klass & kön.* http://www.lo.se/home/lo/res.nsf/vRes/lo_fakta_1366027478784_arbetsmiljo_2014_klass_kon_pdf/$File/Arbetsmiljo_2014_klass_kon.pdf

Pampel, F. C. (2000). *Logistic Regression: A Primer.* Thousand Oaks, CA: SAGE Publications.

Regeringskansliet (2009). *Public Access to Information and Secrecy Act*. http://www.regeringen.se/contentassets/2c767a1ae4e8469fbfd0fc044998ab78/public-access-to-information-and-secrecy-act

SCB (2009). *Integration – utrikes födda på arbetsmarknaden*. http://www.scb.se/statistik/_publikationer/LE0105_2009A01_BR_BE57BR0901.pdf

SCB (2014). *Innovationsverksamhet i Sverige (CIS), 2012-2014*. http://www.scb.se/Statistik/UF/UF0315/_dokument/UF0315_DO_2012-2014_AS_160302.pdf

SCB (2015). *Företagens användning av IT 2015*. http://www.scb.se/Statistik/_Publikationer/NV0116_2015A01_BR_00_IT02BR1501.pdf

SCB (2016). *Teknisk Rapport - En beskrivning av genomförande och metoder. Nulägesundersökningen 2015*.

Svenska Statistikersamfundet (2005). *Standard för Bortfallsberäkning*. http://statistikframjandet.se/survey/wp-content/uploads/2011/05/bortfallsrapport.pdf

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Tivesten, E., Jonsson, S., Jakobsson, L., and Norin, H. (2012). Nonresponse analysis and adjustment in a mail survey on car accidents. *Accident Analysis and Prevention 48,* 401-415.

Appendix A: Variables removed with SBE

Note: No (additional) effects met the 0.2 significance level for removal from the model.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	anylegprotectionD	1	37	0.0020	0.9647	
2	worldfirstD14	1	36	0.0126	0.9105	
3	web	1	35	0.0134	0.9078	
4	supportD14	1	34	0.0257	0.8726	
5	Foradlingsvarde2014	1	33	0.0826	0.7738	
6	Maroth14	1	32	0.0773	0.7809	
7	Marloc14	1	31	0.1300	0.7184	
8	itust2	1	30	0.1449	0.7034	
9	iterp1	1	29	0.1560	0.6929	
10	turn	1	28	0.1744	0.6762	
11	itsp2	1	27	0.2160	0.6421	
12	crman	1	26	0.2227	0.6370	
13	crmstr	1	25	0.3209	0.5711	
14	mobbb	1	24	0.3483	0.5551	
15	Marnat14	1	23	0.4200	0.5169	
16	newproductD14	1	22	0.4452	0.5046	
17	newmarket_D14	1	21	0.4095	0.5222	
18	europeanfirstD14	1	20	0.5625	0.4533	
19	nationfirstD14	1	19	0.4268	0.5136	
20	Localmark14	1	18	0.8887	0.3458	
21	Nationmark14	1	17	0.2543	0.6141	
22	Worldmark14	1	16	0.3667	0.5448	
23	Euromark14	1	15	0.8185	0.3656	
24	utvp	1	14	0.9588	0.3275	
25	neworg14	1	13	1.0425	0.3072	
26	secpol1	1	12	1.3668	0.2424	

Figure A.1: Screenshot of SBE output in SAS.

Appendix B: SAS code

Defining the response variable

```
data indata.mn_nu2015_B;
set indata.mn_nu2015;
if resultatkod =11 or resultatkod=19 or resultatkod =10 then mn_nu2015_resp=1;
else mn_nu2015_resp=0;
if resultatkod => 90 then delete;
run;
```

Defining the main structural factors

```
data mn_nu2015bset1_ind2013_mrgclass;
set individ.mn_nu2015bset1_ind2013_merge;
if 0=< Alder =<34 then AldInt1=1;
else AldInt1=0;
if 35=< alder =<54 then AldInt2=1;
else AldInt2=0;
if alder =>55 then AldInt3=1;
else AldInt3=0;
if UtlSvBakg = 11 or UtlSvBakg = 12 then UtlBakg = 1;
else UtlBakg = 0;
if Kon = 2 then Kon2 = 1;
else Kon2 = 0;
if 0=< SUN2000Niva_Old =<2 then UtbInt1=1;
else UtbInt1=0;
if 3=< SUN2000Niva_Old =<4 then UtbInt2=1;
else UtbInt2=0;
if SUN2000Niva_Old =>5 then UtbInt3=1;
else UtbInt3=0;
inkomst = forvink ;
intTot=1;
run;

data mn_nu2015bset1_ind2013_mrgclass2;
set mn_nu2015bset1_ind2013_mrgclass;
by peorgnr;
if first.peorgnr then
            do;
            AldInt1b=0;
            AldInt2b=0;
            AldInt3b=0;
            UtlBakg2=0;
            Kon3=0;
            UtbInt1b=0;
            UtbInt2b=0;
            UtbInt3b=0;
            inkomst2=0;
            TotInOrg=0;
            end;
AldInt1b + AldInt1;
AldInt2b + AldInt2;
AldInt3b + AldInt3;
UtlBakg2 + UtlBakg;
Kon3 + Kon2;
UtbInt1b + UtbInt1;
UtbInt2b + UtbInt2;
UtbInt3b + UtbInt3;
inkomst2 + inkomst;
TotInOrg + IntTot;
```

```
if last.peorgnr;
run;

data utdata.mn_nu2015b_ind2013_var (drop= AldInt1b AldInt2b AldInt3b UtlBakg2
Kon3 UtbInt1b UtbInt2b UtbInt3b inkomst2);
set mn_nu2015bset1_ind2013_mrgclass2;
Age0_34 = AldInt1b/TotInOrg;
Age35_54 = AldInt2b/TotInOrg;
Age55 = AldInt3b/TotInOrg;
ForBack = UtlBakg2/TotInOrg;
Wom = Kon3/TotInOrg;
EducElm = UtbInt1b/(UtbInt1b+UtbInt2b+UtbInt3b);
EducUpSec = UtbInt2b/(UtbInt1b+UtbInt2b+UtbInt3b);
EducHigh = UtbInt3b/(UtbInt1b+UtbInt2b+UtbInt3b);
AvgInc = (inkomst2/TotInOrg)*100;
run;
```

Calculating the means of Table 4.2

```
* UNWEIGHTED MEANS;
proc means data=utdata.mn_nu2015b_ind2013_var;
*class mn_nu2015_resp;
var Age0_34 Age35_54 Age55 ForBack Wom EducElm EducUpSec EducHigh AvgInc;
*id mn_nu2015_resp ;
run;

proc means data=utdata.mn_nu2015b_ind2013_var;
class mn_nu2015_resp;
var Age0_34 Age35_54 Age55 ForBack Wom EducElm EducUpSec EducHigh AvgInc;
id mn_nu2015_resp ;
run;

* WEIGHTED MEANS;
proc means data=utdata.mn_nu2015b_ind2013_var;
*class mn_nu2015_resp;
var Age0_34 Age35_54 Age55 ForBack Wom EducElm EducUpSec EducHigh AvgInc;
*id mn_nu2015_resp ;
weight designvikt;
run;

proc means data=utdata.mn_nu2015b_ind2013_var;
class mn_nu2015_resp;
var Age0_34 Age35_54 Age55 ForBack Wom EducElm EducUpSec EducHigh AvgInc;
id mn_nu2015_resp ;
weight designvikt;
run;
```

Extracting the mean ratios to produce Figure 4.1

```
* ###### MEAN RATIOS #########;
* ### SIZE GROUPS (sgrupp) ###;
* ###   Variable: Age0_34  ###;

proc means data=utdata.mn_nu2015b_ind2013_var noprint;
class sgrupp mn_nu2015_resp;
var Age0_34;
id sgrupp mn_nu2015_resp ;
output out=sgrupp_means;
run;

data sgrupp_means;
set sgrupp_means;
```

```
if _TYPE_= 0 or _TYPE_=1 then delete;
if _STAT_ ne 'MEAN' then delete;
run;

data sgrupp_tempTot (rename=(Age0_34=var_tot));
set sgrupp_means;
if _TYPE_=3 then delete;
run;

data sgrupp_temp1  (rename=(Age0_34=var_r));
set sgrupp_means;
if _TYPE_=2 then delete;
if mn_nu2015_resp=0 then delete;
run;

data sgrupp_temp0  (rename=(Age0_34=var_nr));
set sgrupp_means;
if _TYPE_=2 then delete;
if mn_nu2015_resp=1 then delete;
run;

data means.rat_sg_AldersInt1;
merge sgrupp_tempTot sgrupp_temp1 sgrupp_temp0;
by sgrupp;
meanratio_r = var_r/var_tot;
meanratio_nr = var_nr/var_tot;
auxvar = 'AldersInt1';
run;

data utdata.totalmeans;
set means.rat_sg_AldersInt1 (keep=meanratio_r meanratio_nr auxvar);
run;

[...]

* ###### MEAN RATIOS #########;
* # INDUSTRY GROUP (bgrupp)  #;
* ###   Variable: AvgInc   ###;

proc means data=utdata.mn_nu2015b_ind2013_var noprint;
class bgrupp mn_nu2015_resp;
var AvgInc;
id bgrupp mn_nu2015_resp ;
output out=bgrupp_means;
run;

data bgrupp_means;
set bgrupp_means;
if _TYPE_= 0 or _TYPE_=1 then delete;
if _STAT_ ne 'MEAN' then delete;
run;

data bgrupp_tempTot (rename=(AvgInc=var_tot));
set bgrupp_means;
if _TYPE_=3 then delete;
run;

data bgrupp_temp1  (rename=(AvgInc=var_r));
set bgrupp_means;
if _TYPE_=2 then delete;
if mn_nu2015_resp=0 then delete;
run;

data bgrupp_temp0  (rename=(AvgInc=var_nr));
set bgrupp_means;
```

```
if _TYPE_=2 then delete;
if mn_nu2015_resp=1 then delete;
run;

data means.rat_bg_AvgInc;
merge bgrupp_tempTot bgrupp_temp1 bgrupp_temp0;
by bgrupp;
meanratio_r = var_r/var_tot;
meanratio_nr = var_nr/var_tot;
auxvar = 'AvgInc';
run;

data utdata.totalmeans;
set utdata.totalmeans means.rat_bg_AvgInc (keep=meanratio_r meanratio_nr
auxvar);
run;
```

Logistic regression procedures of section 4.2

```
* # 4.2.1 General Model # *;

title 'General Model';
proc logistic data=indata.mn_nu2015b_ind2013_var out=genmod;
 class mn_nu2015_resp sgrupp (ref='2') bgrupp (ref='L0000') /param=ref;
 model mn_nu2015_resp(event='1') =
 Wom ForBack
 Age0_34 /*Age35_54*/ Age55
 EducElm /*EducUpSec*/ EducUpHigh
 AvgInc_log
 sgrupp bgrupp ;
run;

* # 4.2.2 Stepwise Backward Elimination #;

title 'Stepwise Backward Elimination';
proc logistic data=steg.nu_set1_indcisit ;
class       mn_nu2015_resp sgrupp(ref='2') bgrupp(ref='M0000')
 Localmark14 (ref='0') Nationmark14 (ref='0')
 Euromark14 (ref='0') Worldmark14 (ref='0')
 Marloc14 (ref='0') Marnat14 (ref='0')
 Mareur14 (ref='0') Maroth14 (ref='0')
 nationfirstD14 (ref='0') europeanfirstD14 (ref='0')
 worldfirstD14 (ref='0') newproductD14 (ref='0')
 newmarket_D14 (ref='0') newprocessD14 (ref='0')
 anylegprotectionD (ref='0') supportD14 (ref='0')
 neworg14 (ref='0') itsp2 (ref='0') itspt2 (ref='0')
 itust2 (ref='0') mobbb (ref='0') web (ref='0')
 iterp1 (ref='0') crmstr (ref='0') crman (ref='0')
 secpol1 (ref='0') utvp (ref='0') /param=ref;

model mn_nu2015_resp(event = '1')=
 Wom ForBack
 Age0_34 /*Age35_54*/ Age55
 EducElm /*EducUpSec*/ EducUpHigh
 AvgInc_log
 sgrupp bgrupp

 /* CIS2014 variables */
 Foradlingsvarde2014
 Localmark14 Nationmark14
 Euromark14 Worldmark14
 Marloc14 Marnat14 Mareur14
 Maroth14 nationfirstD14
 europeanfirstD14 worldfirstD14
 newproductD14 newmarket_D14
```

```
 newprocessD14 anylegprotectionD
 supportD14 neworg14

 /* IT2015 variables */
 itsp2 itspt2 itust2 mobbb web
 iterp1 crmstr crman secpol1
 utvpv turn

 /* SBE specifications */
 /selection=backward include=9 slstay=0.20 maxiter= 25;

run;

* # 4.2.3 Extended Model #;
title 'Extended Model';
proc logistic data=steg.nu_set1_indcis out=extmod;
class mn_nu2015_resp sgrupp (ref='2') bgrupp (ref='M0000')
 newprocessD14 (ref='0') Mareur14 (ref='0') /param=ref;

model mn_nu2015_resp(event='1') =
 Wom ForBack
 Age0_34 /*Age35_54*/ Age55
 EducElm /*EducUpSec*/ EducUpHigh
 AvgInc_log newprocessD14 Mareur14
 sgrupp bgrupp ;
run;

* # Extended Model, Alterations # ;
title 'Extended Model, noiprodprocinnovD14 instead of newprocessD14';
proc logistic data=steg.nu_set1_indcis; *out=extmod;
class mn_nu2015_resp sgrupp (ref='2') bgrupp (ref='M0000')
noiprodprocinnovD14 (ref='0') Mareur14 (ref='0') /param=ref;
model mn_nu2015_resp(event='1') = Wom ForBack
 Age0_34 /*Age35_54*/ Age55
 EducElm /*EducUpSec*/ EducUpHigh
 AvgInc_log noiprodprocinnovD14 Mareur14
 sgrupp bgrupp ;
run;

title 'Extended Model, detailed indicators instead of newprocessD14';
proc logistic data=steg.nu_set1_indcis out=extmod_npdetailed;
class mn_nu2015_resp sgrupp (ref='2') bgrupp (ref='M0000')
 newprocess_prod14 (ref='0') newprocess_logis14 (ref='0')
 newprocess_supp14 (ref='0') Mareur14 (ref='0') /param=ref;
model mn_nu2015_resp(event='1') = Wom ForBack
 Age0_34 /*Age35_54*/ Age55
 EducElm /*EducUpSec*/ EducUpHigh
 AvgInc_log
 newprocess_prod14 newprocess_logis14 newprocess_supp14
 Mareur14
 sgrupp bgrupp ;
run;
```

Stockholms universitet

Kandidatuppsats

Statistiska institutionen

Bachelor thesis, Department of Statistics

Nr 2016:28

Bortfallsanalys av systematiskt fel i Arbetsmiljöverkets Nulägesundersökning 2015

Non-response analysis of non-response bias in Swedish Work Environment Investigation 2015

Merrisha Axelsson och Edvard Åberg

Självständigt arbete 15 högskolepoäng inom Statistik III, VT2016

Handledare: Frank Miller

Sammanfattning

En bortfallsanalys har genomförts utav Arbetsmiljöverkets Nulägesundersökning 2015 (NU2015) med ett relativt stort urval av organisationer som svarat (J-NU15), samt inte svarat (N-NU15) på NU2015. Där båda svarat på Arbetsmiljöverket Nulägesundersökning 2012 (NU2012). I denna analys har det testats hur grupperna skiljer sig åt i variabler av intresse för att undersöka om svarande i NU2015 påverkas av bortfallet i NU2015 gällande kompositvariablerna. Skiljaktigheterna av grupperna testas med *parametriska* och *icke-parametriska* tester som utförs på *kompositvariabler*. *Kompositvariablerna* har utformats för att förklara de mer väsentliga delarna av arbets- och näringslivet. Följaktligen har *histogram, Q-Q plots* samt *Anderson-Darling* test använts för att grafiskt, samt numeriskt, analysera normalitet av kompositvariablerna, då parametriska test är baserade på antagandet av normalitet. *Wilcoxon-Mann-Whitney* (WMW) test har utförts som ett univariat *icke-parametriskt* test samt *Student's*/*Welch's* t-test har nyttjats som univariata *parametriska* test. De kombinerade resultaten av både univariata *parametriska* och univariata *icke-parametriska* test har sedan applicerats till *Zimmerman's* regel om vilket test som är mest lämpat att utgå från, då normalitetsantagandet är svagt. Korrelationen har även mäts mellan *kompositvariablerna* för de två grupperna, vilket varit intressant för det slutliga multivariata testet *Hotelling's Two-sample Squared T-test*. Resultatet har visat på svaga förhållanden till normalitet av vissa *kompositvariabler*. Dock så har samtliga *parametriska* och *icke-parametriska* test visat på att det inte finns någon signifikant skillnad mellan *J-NU15* och *N-NU15* av deras värden för *kompositvariablerna*. Zimmerman's regel har utgjort valet av vilket resultat som man främst bör utgå från. Resultatet har även visat på svaga korrelationer mellan samtliga kompositvariabler. Utförandet av *Hotelling's Two-sample Squared T-test* har även visat på att det inte förekommer någon skillnad mellan grupperna *J-NU15* och *N-NU15*. Samtliga test syftar på att det inte finns någon skillnad mellan grupperna. Det indikerar på att bortfallet i NU2015 inte påverkar resultatet för *kompositvariablerna*. Och att resultatet ifrån NU2015 kan vara användbart för att förklara hur svenskt arbets- och näringsliv ser ut 2015.

Nyckelord: *parametriska test, Wilcoxon-Mann-Whitney, Zimmerman's regel, Hotelling's Two-sample Squared T-test, kompositvariabler*

Abstract

This nonresponse analysis has been made of the Swedish Work Environments Investigation 2015 (NU2015) with a relative good sample for those organizations that have answered (J-NU15) and not answered (N-NU15) on NU2015. Both groups have answered the Swedish Work Environments Investigation 2012 (NU2012). Testing has been made for how the groups compare in variables of interest. Comparisons between the groups have been made with *parametric* and *nonparametric* tests that have been performed on *composite indicator* variables for the two group's answers from NU2012. The *composite indicator* variables have been composited to explain the more essential parts of the working environment. *Histograms, Q-Q plots* and the *Anderson-Darling* test have been applied to perform visual and numerical tests of normality. This since normality is a base requirement for some of the tests that have been used. *Wilcoxon-Mann-Whitney* (WMW) test has been used as a univariate *nonparametric* test and *Student's/Welch's* t-tests have been applied as *parametric* tests. The combined results have been applied to *Zimmerman*'s rule of which test that is most valid to use when the assumptions of normality is weak. The correlation has been tested between the *composite indicator* variables, the correlation is of interest when performing the multivariate test *Hoteling's Two-sample Squared T-test*, a test which searches for comparisons between all five *composite indicator* variables at the same time with one test. The results have proved small hints of normality on some of the *composite indicator* variables, however both the *nonparametric* test and the *parametric* tests have proved that there is no significant difference between J-NU15 and N-NU15 when it comes to the values of the *composite indicator* variables. This indicates that the non-response from NU2015 does not effect on the result from the *composite indicator* variables and that the result from NU2015 can be useful to explain Swedish Work Environment for 2015.

Keywords: *parametric tests, Wilcoxon-Mann-Whitney, Zimmerman's rule, Hotelling's Two-sample Squared T-test, composite indicator*

Förord

Denna uppsats utförs i uppdrag av Arbetsmiljöverket i samarbete med SCB. Vi vill tacka Hans-Olof Hagén (statistiker) på SCB och Annette Nylund (senior analytiker) på Arbetsmiljöverket som tilldelat oss detta uppdrag. Vi vill även tacka vår handledare Frank Miller (universitetslektor) på statistiska institutionen vid Stockholms Universitet som väglett oss vid de statistiska metoderna.

Innehållsförteckning

1 Inledning....1
1.1 Tidigare forskning....1
2 Data....3
2.1 Bransch och storleksklasser....5
3 Metod....6
3.1 Parametriska tester....6
3.1.1 Grafisk analys av normalitet....7
3.1.2 Anderson-Darling goodness of fit test....8
3.1.3 Student's/Welch's T-test....8
3.2 Icke-parametriska test....9
3.2.1 Wilcoxon-Mann-Whitney test....10
3.3 Zimmerman's regel....11
3.4 Multivariata metoder....12
3.4.1 Two-sample Hotelling's T-squared test....12
4 Resultat....14
5 Diskussion....21
Litteraturförteckning....24
Appendix A: Variabler och Branscher....26
A.1 Branscher....27
Appendix B: Test av normalitet....28
B.1 Normalitet av J-NU15 och N-NU15 för kompositvariablerna....28

1 Inledning

Detta är en undersökning av *kvalitén* i Arbetsmiljöverkets Nulägesundersökning 2015 (NU2015), om arbetsorganisation och arbetsmiljö i svenskt arbetsliv. Undersökningen är en bortfallsanalys av NU2015. Analysen av NU2015 genomförs genom att använda sig av uppgifter från den tidigare undersökningen Arbetsmiljöverkets Nulägesundersökning 2012 (NU2012). Med tillgång till uppgifter om svarade i NU2012 och vilka av dessa som ingår i urvalet av NU2015, och om de har svarat eller icke-svarat.

Fem av dem viktigaste kompositvariablerna som är framtagna med hjälp av uppgifter från undersökningen i NU2012 jämförs, för svarande och icke-svarande i NU2015. Medelvärdet av svarande och icke-svarande i NU2015 jämförs med hjälp av uppgifter från NU2012. Dessutom görs en genomgång av vilka test som är lämpliga att göra, för att undersöka eventuella skillnader.

Svarsandelen för NU2012 är 65 procent. När denna bortfallsanalys startade hade knappt hälften svarat på undersökningen NU2015. Insamlingen var således ännu inte fullbordad. Svarsfrekvenser under 80 procent kan resultera i förkastliga undersökningar (Karin Dahmström, 2011), men genom att ha data från samma observationer under olika tidsperioder så kan man statistiskt försöka estimera svar. För den här analysen finns bara tillgång till uppgifter från NU2012. Dessa används i analysen av svarande och icke-svarande i NU2015.

Syftet med bortfallsanalysen är undersöka om svarande i NU2015 påverkas av bortfallet i NU2015 gällande kompositvariablerna.

1.1 Tidigare forskning

År 2014 utfördes en bortfallsanalys av NU2012, där svarande var 1993 organisationer av de 3054 som utgjort den fullkomliga urvalspopulationen. Bortfallsanalysen har utförts genom att testa om det finns en skillnad mellan svarande NU2012 och hela urvalspopulationen. Likhetsgranskningarna har utförts för att analysera dess likheter gällande variabler som beskriver olika strukturfaktorer som bransch, ålder, andel kvinnor, män etc., Jämförelsen av variablerna har utförts med både viktade och oviktade beräkningar av medelvärdet för de olika strukturfaktorerna. Resultatet av bortfallsanalysrapporten konstaterade att det inte förkom någon signifikant skillnad mellan grupperna och därav går det att uttala sig om att NU2012 går att generalisera för hela urvalpopulationen. I denna studie är följande resultat från Arbetsmiljöverket en huvudgrund till antaganden av att följande data är representativt för den svenska arbetsmiljön år 2012. I NU2015 är de mest huvudsakliga delarna uppbyggda på samma sätt som i NU2012. Främst i urval och enkätfrågor (Arbetsmiljöverket, 2014).

1.2 Systematiskt fel

Ett väsentligt problemen med låg svarsfrekvens är risken för systematiskt fel. Det kan uppstå när man har ett för stort urval av svarande i en undersökning som tillhör en viss grupp. Det skapar en skevhet då vissa grupper blir under- eller överrepresenterade. Systematiskt fel har ofta kopplats till hur stort bortfallet i undersökningen är, litet bortfall har ansetts vara samma som liten risk för systematiskt fel om man inte saknar en specifik grupp. Karin Dahmström (2011, s. 355-357) menar att bortfall över 20-30 procent resulterar i förkastliga undersökningar, där hon menar att felet är direkt proportionell mot bortfallets storlek och mot den faktiska skillnaden mellan medelvärdet i gruppen som svarat respektive inte svarat.

Groves (2006) visade däremot i studier att undersökningar utförda med simulerad data gav resultatet att systematiskt fel ökat då svarsfrekvensen ökat. Vilket ger ett annat perspektiv på hur man ser på relationen mellan systematiska fel och bortfall. Merkel & Edelman (Wright, 2015) fann i sin observationsstudie att det inte fanns någon relation mellan systematiskt fel och bortfall. Deras jämförelser skedde mellan estimerade valresultat från partisympatiundersökningar, gentemot hur de riktiga valresultaten såg ut.

Undersökningar med låg svarsfrekvens kan skapa problem, men med tillräcklig data kan man göra jämförelser av hur grupper skiljer sig åt med bakgrundsvariabler om det objekt man försöker analysera. Betlehem & Kersten (Wright, 2015) ser systematiskt fel som determinerad av den relativa storleken av gruppen som inte svarat och kontrasten mellan medelvärden för svarande och icke-svarande på en variabel av intresse. De kompositvariabler som testas i denna bortfallsanalys är sådana variabler som gör det möjligt att undersöka olika delar av de väsentliga delarna av svenskt arbets- och näringsliv.

1.2.1 Bortfallshantering

Hanteringen av bortfallet för att göra undersökningar mer representativa utförs mest effektivt beroende på vad för bakgrundsvariabler som man har tillgängligt. Antingen hanteras bortfallet med att reducera eller komplettera data med tillämpning av olika metoder.

Det finns många olika sätt att komplettera data. Viktningar för att göra mindre grupper mer representativa är ett sätt som fungerar för reducerad data. Man kan även göra regressionsmodeller för att estimera svar av bortfallen och skapa estimerade svar genom bakgrundsvariabler. Beroende på data kan hanteringen variera på vad som minskar risken för systematiskt fel mest. Att reducera data genom att använda bara vissa observationer ur det egentliga populationsurvalet är också en metod. Denna bör användas då man anser att de observationer som man reducerar data till är tillräckliga för att visa på extern validitet för hela urvalet (Longford 2005, s. 39-52).

2 Data

Variablerna som analyseras är kompositer som indikerar på hur väl en organisation fungerar inom de mer väsentliga delarna av arbets- och näringslivet; de fyra första variablerna representerar kompetensen hos organisationerna för grenarna *decentralisering, individuellinlärning, strukturellinlärning* och *numerisk flexibilitet*.

Med *Individuell inlärning* menas hur väl en organisation är bra på att låta sina anställda utvecklas som individer. Såväl som att låta dem vara i en kontext som utvecklar, lär ut och tillfredsställer ett antal mänskliga behov. Om individer lär sig mer så kan de ändra mer, därav kan de anpassas till ändringar av miljön.

Strukturell inlärning baseras på teorier om att anpassning är viktig för organisationen. Den delen beskriver hur väl organisationen lär sig från sin miljö. Organisationen bör kunna lyssna på nya kunders behov, anpassa sig efter förändringar hos konkurrenter samt andra förändringar utanför organisationen.

Numerisk flexibilitet menar hur företaget anpassar sig till att kunna reducera intern och extern arbetskraft under korta perioder(bemanningspersonal, konsulter etc.)

En annan del är *decentralisering* som menar att omfördelning av makt är viktigt inom organisationer, om individer får mer makt och känner sig mer involverade i beslut som sker, så ökar det deras välmående (Statistiska centralbyrån, 2011 sid. 21-49).

En annan variabel representerar *SAM-index*, ett index som väger samma olika delar i arbetsmiljöarbetet, som mäts med hjälp av fyra olika delindex; *Vad*, *Vem*, *Riskundersökning* och *Uppföljning*. Sammantaget handlar dessa om arbetsgivarens sätt att undersöka, genomföra och följa upp verksamheten så att ohälsa och olycksfall i arbetet förebyggs, så att en tillfredställande arbetsmiljö uppnås (Arbetsmiljöverket, 2013).

Frågorna har standardiserats och fått värden från 0 till 1, beroende på svaret av enkätfrågan. Höga värden ger stark indikation på att företaget med stor förmåga är noggranna inom de delar av arbets- och näringslivet som kompositvariablerna utgör. Kompositvariablerna är aritmetiska medelvärden komponerade av 1-8 standardiserade enkätfrågor.[1] Exempelvis har en organisation med värdet 1 i variabeln för *strukturellinlärning* en bra förmåga inom organisationen för de delar som *strukturellinlärning* utgör.

Variablerna som analyseras och representerar de fem delarna av arbetsmiljön är följande:

[1] $M(x) = \frac{x_1+x_2\cdots+x_n}{n}$ Där x_i är de standardiserade frågorna och n är antalet frågor.

k_ind*(individuellinlärning*), k_strukt*(strukturellinlärning)*, k_dec(*decentralisering*), k_num(*numerisk flexibilitet*) och samindex(*SAM-index*) (se Bilaga A.1).

För att förenkla benämns grupperna med förkortat namn. Gruppen svarande som varit med i båda undersökningarna NU2012 och NU2015 går under namnet J-NU15 (svarat i både NU2012 och NU2015). Gruppen icke-svarande som varit med i båda undersökningarna NU2012 och NU2015 går under namnet *N-NU15* (inte svarat i NU2015 men i NU2012).

Av de 1993 som svarat på NU2012 har vi haft tillgång till 317 observationer som varit med urvalet i NU2015, där 187 svarat och 138 inte svarat vilket resulterar i 325 organisationer. Av dessa finns 3 organisationer som svarat och 5 organisationer som inte svarat där värden av kompositvariablerna saknas, detta kan bero på felkodning och vi väljer att titta vidare på de observationer vi har som utgör 317 organisationer.

2.1 Bransch och storleksklasser

I NU2012 är fördelningen av storleksklasser och bransch, är ungefärligt likfördelad. Där svarande från NU2012 i varje grupp av storleksklass är ungefär jämnfördelade vilket kan ses i Tabell 2.1. Däremot så ses det i Tabell 2.2 att det förekommer skillnader mellan storleksklasser för svarande och icke-svarande i NU2015.

Av de 317 organisationer så överrepresenteras storleksklassen 250+ om man jämför med hur fördelningen ser ut i NU2012. Där storleksklassen 250+ utgör 48 % i urvalen från J-NU2015 och N-NU15. Det kan bero på att det förekommer färre stora företag av det slaget i Sverige. Det ökar sannolikheten för att samma företag är med i två undersökningar av samma slag under två olika tidsperioder. Skillnaderna mellan J-NU15 och N-NU15 är dock inte lika stora (Arbetsmiljöverket, 2014).

Fördelningen av *branscher* är även ojämn i urvalsfördelningen, den är överrepresenterad av bransch 16(Offentlig förvaltning, försvar (O) respektive internationella organisationer i Sverige (U)). Fördelningen av bransch 16 för *vårt* urval NU2015 är ca 18 %, jämfört med det övriga branscher som ligger ca mellan 1-9%.

För fördelningen av *bransch* mellan *J-NU15* och *N-NU15* är skillnaden dock inte lika stor. Där fördelningen mellan grupperna för branscherna är ungefär likfördelade (se Bilaga A.1.1) (Stelacon, 2012).

Tabell 2.1 Fördelningen av organisationers storleksklasser för svarande i NU2012

Storleksklass	5-9	10-19	25-49	50-249	250+	Totalt
Svarat NU2012	396	393	399	412	393	1993
Fördelning (%)	19,87	19,72	20,02	20,67	19,72	100

Källa: Arbetsmiljöverket Nulägesundersökning SAM 2012

Tabell 2.2 Fördelningen av storleksklass för de organisationer som deltagit i det urval för NU2015 som svarat i NU2012

Storleksklass	5-9	10-19	25-49	50-249	250+	Totalt
Urval NU2015	21	33	36	74	153	317
Fördelning (%)	6,62	10,41	11,36	23,34	48,26	100
J-NU15	19	12	24	46	83	184

Fördelning(%)	10,33	6,52	13,04	25,00	45,11	100
N-NU15	9	14	12	28	70	133
Fördelning(%)	6,77	10,53	9,02	21,05	52,63	100

Källa: Arbetsmiljöverkets Nulägesundersökning 2015

3 Metod

Undersökningen som utförs är jämförandet av skillnader mellan grupp *J-NU15* och *N-NU15* för kompositvariablerna. Antagandet som görs är att svaren från organisationerna inte ändrats signifikant mellan NU2012 och NU2015, då det är för kort tid för organisationer att ändra arbetsmiljö. Samt att resultaten från NU2012 är representativt för svenskt arbets- och näringsliv år 2012.

Det genomförs signifikanstest för nollhypotesen: att det inte finns skillnad mellan *J-NU15* och *N-NU15* inom kompositvariablerna. I detta fall har vi 5 stycken variabler som skall testas.

För att kunna välja de mest lämpligaste testerna för att jämföra skillnaden mellan grupperna, behöver vi välja mellan att använda icke-parametriska eller parametriska test. Valet av test kommer att grunda sig i huruvida data följer antaganden som metoderna baseras på.

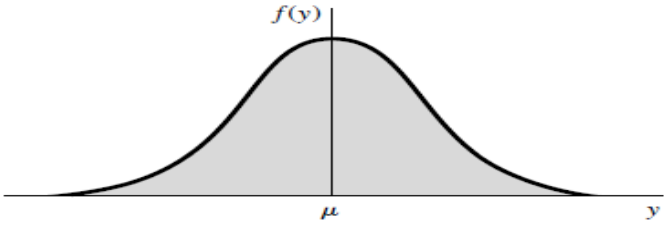
3.1 Parametriska tester

Det lämpligaste parametriska testet för jämförelse mellan gruppernas medelvärde av kompositvariablerna är *Student´s* t-test. *Student's* t-test baseras på följande antaganden;

1. Observationerna är oberoende samt identisk fördelade.
2. Observationerna är normalfördelade.
3. Grupperna har samma varians. (Wackerly *et al.*, 2008).

3.1.1 Grafisk analys av normalitet

För att kunna använda sig utav ett *Student's* t-test undersöks om data är normalfördelad eller har någorlunda likheter med fördelningen. I en normalfördelning är observationerna symmetrisk fördelade kring medelvärdet μ.



Figur 3.1 Täthetsfunktion av en normalfördelning (Wackerly *et al.*, 2008).

För stickprov med normalfördelade stokastiska variabler är stickprovsmedelvärdet normalfördelad med:

$$N \sim (\mu, \frac{\sigma^2}{n}),$$

där μ är medelvärdet, σ^2 är variansen och n antalet observationer.

Analys av normalitet görs både grafiskt och genom statistiska test. Grafiskt görs det med hjälp av histogram och Q-Q plottar för att visuellt se om det förekommer ett mönster som liknar en normalfördelning, men även se hur observationerna är fördelade. Det har visats i studier att grafiska metoder kan ge bättre resultat än numeriska metoder som exempelvis *Anderson-Darling* test av normalitet (Hazelton, 2003).

Då endast en variabel analyseras åt gången är grafiska metoder, som att visuellt titta på fördelningen av variabeln med hjälp av ett histogram lämpligt. Med ett histogram kan man se fördelningens symmetri, spridning och eventuella skevhet (Thode, 2002).

Vid mindre urval kan det vara problematiskt att använda sig utav ett histogram. En annan grafisk metod som är lämpad är kvantil-kvantil plot, känd som Q-Q plot. Q-Q plot är en sannolikhetsplot som används vid test av normalitet. Genom att plotta observationerna mot förväntade värden av en standard normalfördelning och undersöka hur bra plottarna från urvalet följer den teoretiska sannolikhetslinjen. Avviker punkterna ifrån sannolikhetslinjen indikerar detta på att data inte följer en normalfördelning (Thode, 2002).

3.1.2 Anderson-Darling goodness of fit test

En annan metod för att testa normalitet är *Anderson-Darling(AD)* test, ett *goodness-of-fit* test som utförs för att testa hur bra observationerna följer en specifik fördelning. Vid test av normalitet används *AD* statistiskan:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(x_{(i)}) + \ln(1-(x_{(n+1-i)}))), \quad (1)$$

där n är antalet observationer och $x_{(1)}$ $x_{(2)}$ ... $x_{(n)}$ är observationerna sorteras efter storlek, från det minsta till det största, $F(x)$är den underliggande fördelningsfunktionen för normalfördelning som jämförs med observationerna. Följande nollhypotes H_0 testas mot mothypotesen $H_{1:}$ för att få svar gällande observationernas fördelning (Engmann & Cousineau, 2011);

$H_{0:}$ $uravlet\ kommer\ ifrån\ en\ normalfördelning$
$H_{1:}$ $urvalet\ kommer\ inte\ ifrån\ en\ normalfördelning$

Valet av signifikansnivå i denna studie är $\alpha = 0{,}05$. Om *AD* testet påvisar ett resultat där p-värdet, P≥α indikerar detta att vi inte kan förkasta H_0 att urvalet följer en normalfördelning. Visar resultatet ett p-värde som understiger signifikansnivån på 5 %, P<α antyder detta att H_0 förkastas och att urvalet inte kommer från en normalfördelning (Engman and Cousineau, 2011).

3.1.3 Student's/Welch's t-test

Ett av kriterierna för att kunna genomföra *Student's* t-test krävs att variansen följer antagande för homogenitet mellan gruppernas varianser, för att inte påverka validiteten av resultatet. Beroende på gruppernas storlek kan detta medföra instabilitet hos gruppernas varianser. Större urval tenderar att få mindre varians och mindre urval större varians. Detta måste tas i beaktning där *Student's* t-test använder den poolade variansen. Förekommer skiljaktigheter av gruppernas varians används den icke-poolade variansen vid utförandet av *Welch's* t-test (Aishah and Yahya, 2014).

För att kunna jämföra grupperna med hjälp av *Student's/Welch's* t-test skattas medelvärde och varians i grupperna. Populationens medelvärde och varians är okänd, därav används stickprovsvarians och det skattade medelvärdet (Liao, 2002). Vid icke-poolade varianser används den skattade variansen utifrån vardera gruppen.

Stickprovsvariansen har följande ekvation:

$$s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}, \quad (2)$$

där x_i= stokastiska variabler i observationer $\bar{x}$= urvalsmedelvärde n= antal observationer.

Med hjälp av *Student's*/*Welch's* t-test kommer följande nollhypotes och mothypotes besvaras. Nollhypotesen: $H_0: \mu_1 = \mu_2$ gruppernas medelvärde är lika, Mothypotes: $H_1: \mu_1 \neq \mu_2$ gruppernas medelvärde är olika (Liao, 2002).[2]

Under nollhypotesen är följande formel för *Student's* t-test statistiska med den poolade variansen (Wackerly *et al.*, 2008):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{(p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \tag{3}$$

där $\bar{x}_1, \bar{x}_2$= medelvärde för gruppen *i*= 1, 2, $s_{(p)}$= poolade variansen för grupperna, n_1, n_2 =storleksurval i grupperna *i*= 1, 2.

Under nollhypotesen är följande formel för *Welch's* t-test med den icke-poolade variansen (Aishah & Yahya, 2014):

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_1^2}{n_2}}} \tag{4}$$

Vid Student's/Welch's t-test görs ett dubbelsidigt test då vi är intresserade av skiljaktigheten mellan grupperna. P-värdet för dubbelsidigt test definieras som; $p = 2P(t > t_o \mid t \sim t_{\alpha,df})$, t_o= uträknade t-värdet och $t_{\alpha,df}$= kritiska t-värdet, med α=0,05 och df (n_1+n_2-2) (Aishah and Yahya, 2014).

Nollhypotesen förkastas gällande att populationen har samma medelvärde om Student's/Welch's t-test visar att p-värdet < α, som i detta fall är α=0,05. Och förkastar inte nollhypotesen om p-värdet överstiger α=0,05.

Skulle det visa att variablerna inte följer en normalfördelning och vi inte kan specificera fördelningen så blir icke-parametriska tester ett intressant alternativ.

3.2 Icke-parametriska test

Statistiska metoder baseras ofta på olika antaganden. I många fall brukar inte data följa antaganden, som exempelvis för *Student's* t-test. Oftast faller detta in gällande antagande om att båda grupperna ska komma ifrån en normalfördelning och ha samma varians. Detta kan i sin tur påverka valet av metod och resultatets validitet. Vid parametriska tester är fördelningen för populationen känd men med okända medelvärden och varianser (Wackerly *et al.*, 2008). I detta fall när man inte har kunskapen om populationens

[2] μ_1= medelvärde i J-NU15 μ_2= medelvärde i N-NU15

fördelning kan icke-parametriska tester vara lämpliga, då testet kan utföras utan att baseras på några antaganden om fördelningen (Liao, 2002).

3.2.1 Wilcoxon-Mann-Whitney test

Mann-Whitney U test, som även kallas *Wilcoxon rank sum* test är ett test som används vid jämförelse av två grupper (Liao, 2002)[3]. Testet är baserat på att ranka observationernas värde i varje grupp från det lägsta (lägsta siffran får värde 1) till det högsta värdet. Testet tar även hänsyn till så kallade *Ties* som innebär att grupperna emellan har observationer med samma värde. Detta handskas med att ta medelvärde av ranknumret, som blir det nya rank-värdet för de relevanta observationerna. Där efter summeras rang i grupperna för att jämför gruppernas fördelningar (Wackerly *et al.*, 2008).

Normalfördelningsantagandet krävs inte för detta test.

Ekvation för *Wilcoxon-Mann-Whitney* (WMW) test statistika*:*

$$U = n_1 n_2 + \frac{n_1(n_2+1)}{2} - W \ , \tag{5}$$

där *W=rang-summan av observationerna*
n_1*= antal observationer i grupp 1* n_2*= antal observationer i grupp 2.*

Nollhypotesen är att fördelningen mellan grupperna är likartade (förekommer ingen skillnad i medelvärdet mellan grupperna). Och mothypotesen är att fördelningen mellan grupperna skiljer sig i läge (medelvärdet mellan grupperna är olika, tvåsidigt test).

Vid ett stort urval där observationerna är $n_1 > 10$ och $n_2 > 10$ används Z-statistikan:

$$Z = \frac{U - \left(\frac{n_1 n_2}{2}\right)}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}, \tag{6}$$

där U-värdet beräknas utifrån *ekvation (5).*

Med en signifikansnivå på α=0,05 kan vi inte förkasta H_0 att det inte förekommer någon skillnad mellan grupperna när p-värdet >α=0,05. Vi förkastar H_0 om att det inte förekommer en skillnad mellan grupperna, om p-värdet understiger signifikansnivån α=0,05 (Wackerly *et al.*, 2008).

[3] Vi utgår från att i denna undersökning kalla det för *Wilcoxon-Mann-Whitney* (WMW) test.

3.3 Zimmerman's regel

I en artikel publicerad i *British Journal of Matematical and Statistical Psychology* skriver *Zimmerman* (2011) om problematiken gällande antagande för normalitet vid t-test och dess brister gällande validiteten av resultatet, om data inte fullföljer kriterierna. Icke-parametriska tester som *WMW* används då observationerna inte följer antaganden. Dock kan det oftast uppstå komplikationer gällande val av parametriska och icke-parametriska tester om man inte har någon form av kunskap gällande populationen. Då det inte finns någon given regel vid val av att välja det bästa testet, kan valet av metod vara problematiskt.

Zimmerman menar att olika typer av studier har påvisat att när data följer antagandena för parametriska tester förekommer det inte stora skillnader i resultatet som de icke-parametriska testerna ger (om man utför båda testerna på samma data). Följer data inte antagandena förekommer det oftast skiljaktigheter i resultaten mellan de parametriska och icke-parametriska testerna.

Han resonerar att om både *Student´s* t-test och *WMW* används på ett urval där antagande för normalitet är oklart, borde testerna ge liknande resultat om urvalet kommer ifrån en normalfördelning. *Zimmerman* förklarar i nämnd artikel att vid stora skillnader mellan parametriska och icke-parametriska metoder skall icke-parametriska metoder vara mer giltiga. Han menar dock att det kan vara svårt att skapa en algoritm eller någon form av regel som kan besluta val av metod, speciellt då det oftast inte går att kombinera parametriska och icke-parametriska metoder. Zimmerman menar när det kommer till *Student's* t-test och *WMW* så kan följande metod någorlunda appliceras, eftersom att båda metoder baseras på rank transformationer.

Metoden går ut på att först utföra ett t-test på observationerna. Sedan gör ytterligare test där man beräknar t_r som är ett t-test baserad på rangern. Därav görs en jämförelse av t-värden från *Student's* t-test och rang-baserade t-test t_r.

Ekvationen för test-statistikan är:

$$t_r = \frac{\theta}{\sqrt{(n-1-\theta^2)/(n-2)}}, \tag{7}$$

där θ är desamma som z-värdet ifrån *WMW* testet och $n = n_1 + n_2$.

Absoluta värdet 0,4 använd som gränsvärde[4]. Om skillnaden mellan t-statistiken ifrån *Student's/Welch's* t-test och t-statistiken för rangerna t_r överstiger gränsvärde på 0.4 så rekommenderas *WMW* testet annars *Student's/Welch's* t-test. Denna metod är komplex men skulle möjligtvis kunna vägleda vid komplikationer gällande val av parametriska och icke-parametriska tester (Zimmerman, 2011 s. 388- 390).

3.4 Multivariata metoder

Vid univariata metoder som *Student´*s t-test och *WMW* test observeras endast en variabel åt gången för att hitta skillnader mellan grupperna. I många fall kan det förekomma korrelationer mellan variabler vilket inte tas hänsyn till vid univariata metoder. Av det är multivariata metoder användbart om man förmodar att ett sådant samband finns. Multivariata metoder undersöker alla variabler samtidigt för att kunna hitta ett underliggande mönster, som kan finna om det förekommer någon form av skiljaktigheter mellan grupperna (Rencher and Christensen, 2012).

Då variablerna är kvantitativa och kontinuerliga så används den mest välkända och vanligaste formen av korrelationskoefficient som är *Pearsons produktmomentkorrelationskoefficient.*[5]

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} \quad (8)$$

Denna metod kan vara lämplig då vi vill se hur sambandet mellan variablerna ser ut, som vidare kan ge oss en indikation på kommande metoder. Korrelationen har värden inom intervallet -1 och 1, där högt värde av absolut värdet 1 syftar på att det förekommer ett linjärt samband (Haning, 1991).

3.4.1 Two-sample Hotelling's T-squared test

En följande metod som kan appliceras är *Two-sample Hotelling's T-squared* test. Testet kan utföras med hjälp av en multivariat teknik känd som MANOVA *(multivariate analysis of variance).* MANOVA kan användas för att hitta signifikanser givet flertal variabler. Med hjälp av ett *Two-sample Hotelling's T-squared* test kan test utföras för att se om det förekommer en skillnad i medelvärdena för kompositvariablerna mellan *J-NU15* och *N-NU15.*

Antaganden för multivariata metoder är likartad med antagande vid univariata fall. Antagande för följande metod för MANOVA bör vara att;

1. Data ska komma ifrån en multivariat normalfördelning.

[4] *Cut-off värde* kan översättas som brytningspunkt, där värdet 0,4 har bestämts efter simuleringsstudie.

[5] $\bar{x}, \bar{y}$ *och* s_x, s_y = medelvärden och standardavvikelser för grupperna.

2. Kovariansmatris av gruppernas kovarians ska vara lika.

3. Observationerna ska vara oberoende.

Att undersöka om data är multivariat normalfördelad är merendels invecklat. Statistiska undersökningar har påvisat att det är sällsynt att variabler som är normalfördelade inte kommer ifrån en multivariat normalfördelning. Och av den anledningen kan univariata tester (att titta på variablerna för sig) vara en form för att undersöka detta antagande (Sharma, 1996).

Detta kan göras genom att använda sig av grafiska metoder som QQ-plots eller t.ex. Anderson-Darling test.

Det andra antagandet är att gruppens kovariansmatriser ska vara lika. En metod för att testa detta är Box's M, som testar om det förekommer någon signifikant skillnad mellan gruppernas kovarians matriser.

Följande hypotes ställs; $H_0: \Sigma_1 = \Sigma_2$ grupp 1 och 2 kovarianser är lika. $H_1: \Sigma_1 \neq \Sigma_2$ där grupp 1 och 2 kovarianser är skilda. Påvisar Box´s M testet ett p-värde som understiger signifikansnivån α=0.05, förkastar vi H_0om att gruppernas kovarians matriser är lika. Effekten av att genomföra MANOVA trots olika kovarians matriser kan ge invalid svar (Sharma, 1996).

Two-Sample Hotelling's T-squared test-statistiska är:

$$T^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \left\{ \boldsymbol{S}_p \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2), \tag{9}$$

där $\bar{\boldsymbol{x}}_1$och $\bar{\boldsymbol{x}}_2$ står för vektorerna av medelvärden för alla variabler *i*, för grupperna *j*= 1, 2.[6] $\boldsymbol{S}_p$= Poolad varians-kovarians matris i grupp 1 och 2.

Vid större urval approximeras och transformeras T^2 statistiska till en F- statistiska.

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 \sim F_{p, n_1 + n_2 - p - 1}, \tag{10}$$

p= antal frihetsgrader (STAT505, 2016)

[6] $\bar{\mathbf{x}}_{\mathrm{j}} = \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{X}_{\mathrm{ji}}$

Med *Two-Sample Hotelling's T-squared* testas följande:

$$H_o: \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{i1} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{i2} \end{pmatrix} H_1: \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{i1} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{i2} \end{pmatrix},$$

där μ_{ij}= följande medelvärde i anslutning till variabeln *i*, i grupp *j*= 1, 2. Nollhypotesen H_o, är att det inte förekommer skillnad i medelvärdena hos grupperna. Mothypotes H_1, är att det förekommer en skillnad i medelvärdena hos grupperna (Sharma, 1996). Följaktligen förkastas H_0 när p-värdet av F-statistikan påvisar ett p-värde < $\alpha(0.05)$ (STAT505, 2016).

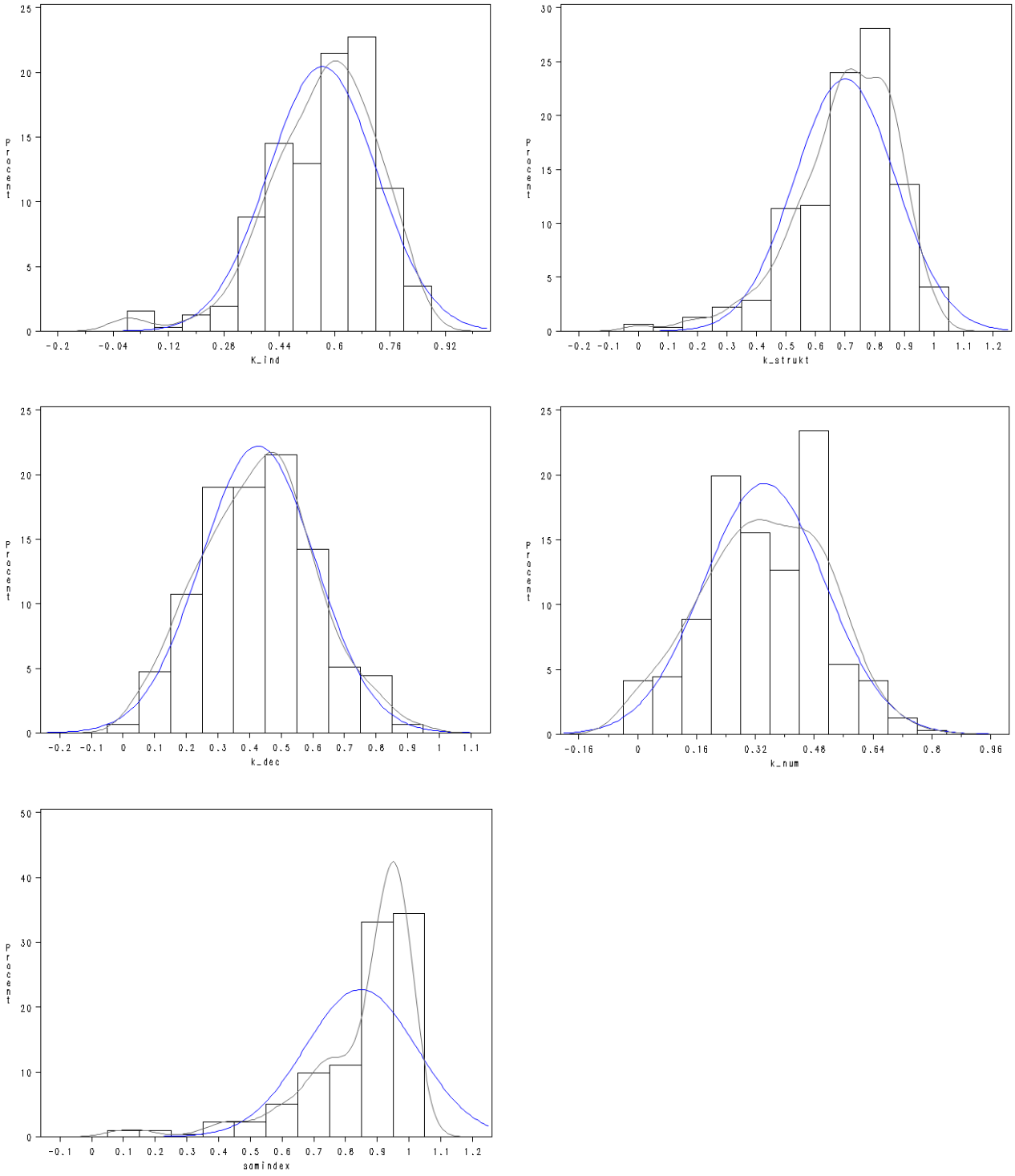
4 Resultat

För att kunna utföra metoderna har vi analyserat antaganden för samtliga test.

För *Student´s/ Welch's* t-test har vi testat antagande för: oberoende, normalfördelning och variansen.

För samtliga test så antas oberoende av svarande då de deras svar inte påverkats av varandra.

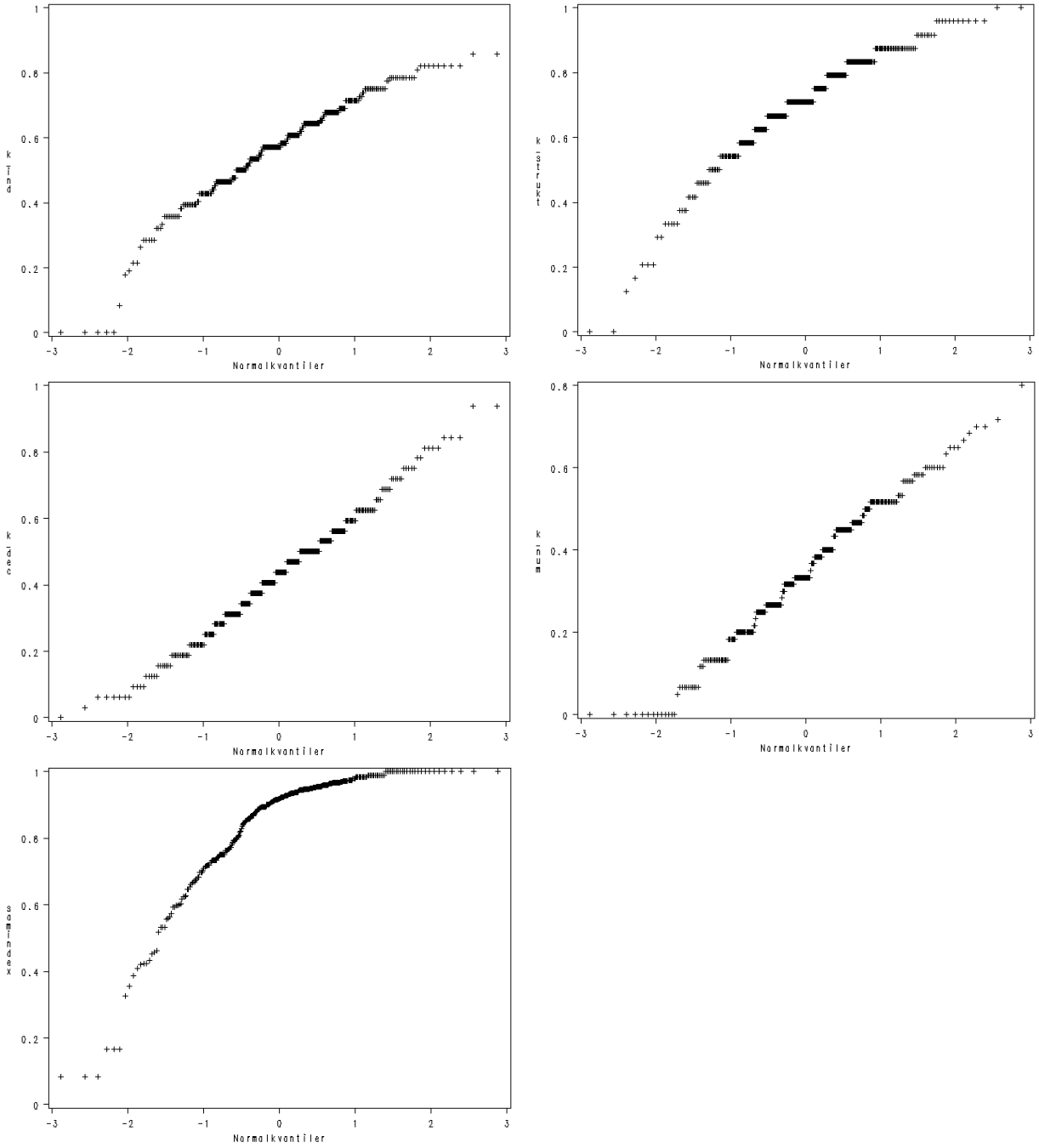
Det andra steget som gjorts i analys av data är att observera fördelningen av kompositvariablerna. De grafiska metoderna för test av normalitet har utförts genom att analysera kompositvariablernas fördelning med hjälp av histogram samt *Q-Q* plots.



Figur 4.1 Histogram för kompositvariablerna; *k_ind*, *k_strukt*, *k_dec*, *k_num* samt *samindex*, i nämnd ordning. Visar fördelningen av kompositvariablernas värden för de 317 organisationer, samt en teoretisk normalfördelningslinje och en slätad linje av kompositvariablernas fördelning. Y-axeln visar procentuella fördelningen, X-axeln visar värdet av kompositvariabeln.

Utifrån histogrammen i *Figur 4.1* kan vi hitta tydliga mönster som påminner oss om normalfördelningen då vi har analyserat fördelningen av kompositvariablerna. Dessa mönster förekommer starkast hos variabeln *k_ind* och *k_dec*, där observationerna anses vara symmetriska kring medelvärdet. För variabeln *k_strukt* syns en fördelning som påminner om normalfördelningen, men som inte är lika stark och tydlig som för tidigare nämnda variabler. Samma gäller för variabeln *k_num* men som även den inte är lika klar. För variabeln *samindex* syns däremot en tydlig högerfördelning av observationerna som tyder på att den inte är normalfördelad.

Grafisk analys av histogram har även gjorts för fördelningen mellan grupper *J-NU15* och *N-NU15* där resultaten har varit liknande som för hela urvalet av NU2015 (se Bilaga B.1).



Figur 4.2 Resultat av Q-Q plottar för kompositvariablernas *k_ind*, *k_strukt*, *k_dec*, *k_num* samt *samindex*, i nämnd ordning. Y-axeln visar kompositvariabelns värde och X-axeln visar normalkvantiler**.**

Från Q-Q plottarna i *figur 4.2* får vi likartade svar som observationen i histogrammen; *k_ind* och *k_dec* är de variabler som följer sannolikhetslinjen närmast. Vi ser dock otydligheter i QQ-ploten av *k_strukt* och *k_num*, *k_num* som skiljer sig från sannolikhetslinjen i nedre delen, då det förekommer flertal observationer där som inte följer linjen. Samma slutsats dras om *samindex* där observationerna i förhållande till sannolikhetslinjen avviker, linjen blir kurvig i den övre delen och följer inte sannolikhetslinjen på ett sätt som förespråkar för normalitet.

En Q-Q plot har även analyserats mellan grupperna *J-NU15* och *N-NU15*, där fördelningen är liknande som för hela urvalet (se Bilaga B.1).

För att stärka resultatet av normalitetstest så har även *AD* test utförts. Resultatet visar att fördelningen avviker statistisk signifikant från normalfördelningen för de flesta kompositvariabler, då p-värdet understiger signifikansnivån $\alpha = 0{,}05$. Dock är kompositvariabeln *k_dec* inte signifikant med ett p-värde på 0,081.

Detta betyder att H_0 att urvalet kommer från en normalfördelning förkastas för kompositvariablerna; *k_ind*, *k_strukt*, *k_num* samt *samindex*. H_0 behålls dock för kompositvariabeln *k_dec* som indikerar att urvalet är ifrån en normalfördelning.

Som i de grafiska metoderna har även *AD* test utförts för kompositvariablerna mellan grupperna. Där resultatet varit liknande; att H_0 förkastas för samtliga variabler förutom för *k_dec*.

Utifrån resultatet är endast *k_dec* normalfördelad, vilket indikerar på att resterande kompositvariabler *k_ind*, *k_strukt*, *k_num* samt *samindex* inte följer antagande för normalitet. Därav är *Student's*/*Welch's* t-test endast lämplig för kompositvariabeln *k_dec*.

Likväl skall man vara kritisk till numeriska metoder för att testa normalitet då det i många fall ger missvisande resultat och därav är det bättre att använda sig av grafiska metoder (Hazelton, 2003). Utifrån resultatet så är *k_dec* och *k_num* de variabler som mest följer en normalfördelning. Detta måste även också beaktas eftersom grafiska metoder är subjektiva, då det förekommer våra egna tolkningar av data (Sharma, 1996).

Student´s t-test och *WMW* test utförts på samtliga variabler.

För *Student´s* t-test så baseras antaganden att grupperna har samma varians. Vid test av jämlik varians så visade sig endast kompositvariabeln *samindex* vara signifikant med ett p-värde på $0{,}0147$, resterande hade ett p-värde som var långt över den kritiska gränsen på $\alpha = 0{,}05$. Vilket betyder att för kompositvariabel *samindex* förkastar vi H_0 att grupperna har samma varias. Vilket betyder att *Welch's* t-test använts för *samindex* och för resterande kompositvariabler har *Student's* t-test använts.

Tabell 4.1 T-värde och två sidigt p-värde av *Student's/Welch's* t-test.

Variabel	T-värde	Two-sided pr. >|T|
k_ind	–1,21	0,23
k_strukt	–0,29	0,77
k_dec	–1,51	0,13
k_num	0,46	0,65
samindex	0,41	0,68

Från tabell 4.1[7] kan det ses att *Student's/Welch's* t-test påvisas vara icke-signifikanta för alla kompositvariabler, då p-värdet för samtliga ligger över den kritiska gränsen $\alpha = 0.05$. Vilket menar på att vi behåller nollhypotesen om att det inte finns någon skillnad mellan grupperna *J-NU15* och *N-NU2015* utifrån kompositvariablernas värden för respektive grupper. Det kan tolkas som att grupperna inte skiljer sig signifikant åt mellan resultaten av kompositvariablerna.

WMA test utförs utan antaganden om urvalets fördelning.

Tabell 4.2 Z-värde och två sidigt p-värde av *Wilcoxon-Mann-Whitney* test

Variabel	Z-värde	Two-sided pr >|Z|
k_ind	−1,57	0,12
k_strukt	−0,04	0,97
k_dec	−1,56	0,12
k_num	0,48	0,63
samindex	−0,25	0,80

I Tabell 4.2 visas det att samtliga kompositvariabler är icke-signifikanta, då p-värdena överstiger den kritiska gränsen. Detta medför samma slutsats som i *Student's/Welch's* t-test; att det inte förekommer någon skillnad mellan grupperna *J-NU15* och *N-NU2015*. Nollhypotesen behålls på signifikansnivån $\alpha = 0{,}05$.

För att välja vilket av *Student's* t-test och *WMW* vars resultat är lämpad att utgå ifrån, görs genomförandet av *Zimmerman's* regel med hjälp av ekvationen (7) för beräkningen av t_r.

Av den orsaken att *Welch's* t-test används för variabel *samindex* har inte *Zimmerman's* regel applicerats, då regeln omfattar t-värden ifrån *Student´s* t-test. Utifrån tidigare resultat kunde vi konstatera att *samindex* inte följer antagandet för normalitet, där av används *WMW* resultatet för denna kompositvariabel.

[7] I denna upplaga har det korrigerats tryckfel som skett i den originella uppsatsen. Z-värde och Two-sided pr >|Z| för k_ind och k_strukt är korrekt i denna upplaga. Revidering har utförts 2016-08-27.

För variabeln *k_num* finner vi att i (7) $n_1 + n_2 = 316$, med $\theta = 0{,}4821$ (z-värdet ifrån *WMW*) och $t_r = 0{,}415$. Skillnaden mellan t-värdet av t-testet och t_r i *k_num* beräknas därefter: $0{,}46 - 0{,}4815 = -0{,}0215$. Skillnaden mellan dessa t-statistikor överstiger alltså inte det absoluta gränsvärdet på 0,4, då skillnaden är $-0{,}0215$. Detta indikerar på att resultatet ifrån *Student's* t-testet i denna variabel skall användas.

Med samma formel fås följande resultat av differenser mellan *Student's* t-test och t_r för resterande variabler: *k_ind*=1,2792, *k_strukt*=1,1722 och *k_dec*=0,05. Skiljaktigheterna har påpekat att variabeln som understiger gränsvärdet är *k_dec*. Detta indikerar på att resultatet ifrån *Student's* t-testet i denna variabler skall användas. Gällande *k_ind* och *k_strukt* konstateras att det överstiger gränsvärdet och därmed används resultatet ifrån *WMW* testet.

Med *Zimmerman's* regel har vi fått en indikation på val av test för det flesta variabler, men oavsett regeln kan vi ändå konstatera att det inte finns någon signifikant skillnad mellan grupperna inom varje variabel. Oavsett vilket test vi använder, så har både *Student's/ Welch's* t-test och *WMW* påvisat att det inte förekommer någon signifikant skillnad mellan *J-NU15* och *N-NU15.*

Det slutgiltiga för att testa skillnader mellan grupperna görs en multivariatanalys för att testa alla kompositvariabler samtidigt. För att fortsätta med den multivariata metoden analyseras korrelationen mellan variablerna.

Tabell 4.3 Korrelationsmatrisen visar korrelationssambandet mellan variablerna, p-värde och antalet organisationer för urvalet från NU2015. Tabellen visar i följande ordning: 1. Korrelationskoefficienten 2. P-värdet och 3. Antal observationer

	k_ind	k_strukt	k_dec	k_num	samindex
k_ind	1				
k_strukt	0,40 <,0001 317	1			
k_dec	0,14 0,0100 316	0,22 <,0001 316	1		
k_num	0,14 0,0114 316	0,16 0,0035 316	-0,12 0,0291 315	1	
samindex	0,20 0,0004 317	0,26 <,0001 317	0,03 0,5478 316	0,11 0,0464 316	1

I Tabell 4.3 ses hur samtliga variabler är signifikanta, då samtliga variabler har ett p-värde som understiger den kritiska gränsen. Detta antyder att det förekommer ett linjärt samband mellan variablerna. Förutom mellan kompositvariabeln *samindex* och *k_dec*

som har ett p-värde på ca $0{,}55$. Dock måste man beakta hur starkt det linjära sambandet är. Det förekommer en korrelation mellan variablerna, men korrelationen mellan samtliga variabler är väldigt svagt, då korrelationskoefficienterna endast antar värden mellan $-0{,}12$ och $0{,}4$.

Med låg korrelation är slutgiltigt steget att analysera samtliga variabler samtidigt, som utförs genom *Two-sample Hotelling's Squared t-test*. Utgångspunkten för normalitet vid detta test är de grafiska analyserna. Där vi kan se att alla variabler med undantag av *samindex* följer en approximativ normalfördelning. Dock så inkluderar vi *samindex* för enkelhetens skull vid utförandet av multivariata testet.

Kommande antagande för att testa om det förekommer någon signifikant skillnad mellan grupperna kovarians matriser har *Box´s M* genomförts. Resultatet har påvisat att nollhypotesen behålls vilket betyder att det inte förekommer någon skillnad mellan grupperna kovarians matriser då det visar ett p-värde på $0{,}505$, vilket överstiger den utvalda kritiska gräns på $\alpha = 0{,}05$.

Two-sample Hotelling´s Square t-test ger följande resultat visar ett F-värde på $0{,}80$ och att p-värde>F är $0{,}5525$; nollhypotesen förkastas inte och testet pekar även här på att det inte förekommer någon skillnad mellan *N-NU15* och *J-NU15* för kompositvariablerna *(se bilaga).*

Slutsatsen av resultatet har påvisat att i alla kompositvariabler så förekommer det inte någon signifikant skillnad mellan grupperna svarande och icke-svarade i NU2015.

5 Diskussion

Utifrån resultatet utav de statistiska metoder som utförts så kan man dra slutsatsen att det inte finns någon signifikant skillnad mellan svarande och icke-svarande NU2015, vad gäller *kompositvariablernas* värden från samma organisationer i NU2012.

Det indikerar på att bortfallet i NU2015 inte påverkar resultatet för kompositvariablerna. Och att resultatet ifrån NU2015, kan vara användbart för att förklara hur svenskt arbets- och näringsliv ser ut 2015. Skulle det förekomma någon skillnad för kompositvariablerna mellan grupperna så skulle det indikera på att icke-svarande i NU2015 har en påverkan på resultatet av kompositvariablerna för *vårt* urval i NU2015. Och därför försämra kvalitén i NU2015.

Bedömningen bör tas kritiskt, då fördelningen utav urvalet vi tittar på inom branscher och storleksklasser inte varit likgiltig till fördelningen utav hela populationsurvalet. Det kan skapa skillnader i svar då urvalet vi undersöker är överrepresenterat av stora organisationer och av bransch 16 (offentliga organisationer etc.) Det har dock inte förekommit någon stor skillnad mellan bransch och storleksklasser mellan grupperna svarande och icke-svarande i urvalet från NU2015. Följaktligen finns det ingen skillnad inom vårt urval, men skillnad mellan vårt urval och urvalspopulationen.

Vidare bör bedömningen om kvalitén av NU2015 i detta fall tas med försiktighet, då vi uttalar oss om NU2015 med endast data ifrån NU2012. Och enbart information om de som varit med i båda urvalen. Att resultatet visat att bortfallet inte påverkar värdena på de 5 *kompositvariablerna*, stärker undersökningen för NU2015. Med antagandet att organisationer skulle svarat likartat i båda undersökningarna. Dock är detta bara en estimering, men kan möjligtvis stärka användbarheten av NU2015.

Problematiken kring normalitet har skapat problem i våra test då vi har använt metoder som baseras på normalfördelad data, och antagande inte alltid fullföljts när vi analyserat för normalfördelning. Det som förespråkar för de test som vi gjort är däremot att test av normalitet kan vara komplext, då data i sin helhet oftast inte är normalfördelad. Med *Zimmerman's* regel kan vi trots problematiken kring normalitet kunnat få indikationer på vilket av *Student's* t- test eller *WMW* som varit mest lämpad. Då samtliga test påvisat samma resultat – att nollhypotesen inte kunde förkastas, har denna regel inte påverkat resultatet av jämförandet. *Zimmerman's* regel skulle ha haft mer betydelse om testerna påvisat olika resultat gällande skillnader av grupperna. Regeln skulle fylla funktionen om att tala om vilket resultat som är bäst lämpad.

Gällande resultatet ifrån multivariata metoden, då vi analyserar alla variabler samtidigt, resulterat i samma svar. Vi är dock medvetna att riktiga tester kring multivariata normalfördelning inte skett då vi endast har teoretisk gjort detta antagande. Detta kan i sin tur påverka resultatet, men eftersom att det inte avviker ifrån resterande metoder. Där av kan vi utgå från samtliga resultat att det inte finns någon skillnad mellan grupperna. Skulle fallet varit så att metoderna påvisat olika resultat, skulle detta behöva tas i beaktning för att göra djupare analyser.

Det finns statistiska metoder som kan användas i framtida studier som möjligtvis kan stärka kvalitén av NU2015. Analysen som gjorts har använts genom att endast arbeta med data som varit tillgänglig. Det har inte gjorts några försök att komplettera data för bortfallet. Men med informationen som finns anser vi att möjligheten finns för framtida bortfallsanalyser.

Arbetsmiljöverket – I samarbete med SCB – arbetar för att få fram kompositvariablerna för företagen som svarat i NU2015. Då enkäterna är uppbyggda på samma sätt i de mer väsentliga delarna så kan man testa för skillnader mellan hur kompositvariablernas värden ändrats mellan NU2012 och NU2015. Detta kan analyseras i individuella organisationer som varit med i båda undersökningarna men även i större utsträckningar. Som mellan storleken på organisationerna, bransch och sektor tillhörighet. Skulle det visa sig att det finns små skillnader mellan organisationerna, eller inga skillnader alls, så skulle bakgrundsvariabler kunna användas för att estimera organisationers svar och där av hantera bortfallet genom att komplettera data.

Systematiska fel är ett stort problem för statistiker, och som det ser ut för kommer problemet att kvarstå i framtiden. Självfallet är det önskvärda att minska på bortfallen, men det problemet kan vara mer inriktat som ett arbete för politiker än för statistiker. Att öka informationen om undersökningsdeltagare från olika perioder ökar dock sannolikheterna att minska på dessa fel genom bakgrundsvariabler. Med tillräcklig

bakgrundsinformation så kan bearbetningen av data och estimering av vad undersökningsdeltagare skulle ha svarat bli mer precist.

Litteraturförteckning

Arbetsmiljöverket 2013. *SAM-index: Ett sätt att belysa systematiskt arbetsmiljöarbete I svenskt arbetsliv - baserad på Arbetsmiljöverkets Nulägesundersökning SAM 2012.* Arbetsmiljöverkets analysrapport 2013:2.

Arbetsmiljöverket 2014. *Bortfallsanalys – Representerar de svarande i organisationerna i Arbetsmiljöverkets Nulägesundersökning 2012 svenskt arbetsliv?* Arbetsmiljöverkets analysrapport 2014:1.

Betlehem, J., & Bakker, B. (2014*)* The impact of nonresponse on survey quality. *Statistical Journal of the IAOS* 30*:*243-248.

Dahmström K (2011) *Från datainsamling till rapport – att göra en statistisk undersökning*, 5:e upplaga. Lund: Studentlitteratur.

Engmann, S. & Cousineau, D. (2011) Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods* 6(3): 1-17.

Groves, R. M. (2006) *Nonrespone Rates and Nonresponse Bias in Household Surveys*, Public opinion Quarterly 70(4): 646-75

Haning, R. (1991) Bivariate correlation with Spatial data: *Geographical Analysis.* 23: 210-227.

Hazelton, L. M. (2013) A Graphical Tool for Assessing Normality, *The American Statistican*, 57:4, 285-288.

Liao T. F (2002) *Statistical Group Comparison.* New York: John Wiley & Sons.

Longford, T. N. (2005) *Missing Data and Small-Area Estimation - Modern Analytical Equipment for the Survey Statistician.* New York: Springer Science+Business Media, Inc.

Aishah, A. & Yahya, S.S. (2014*) Sensitivity Analysis of Welch's t-Test.* AIP Conference Proceedings 1605: sid. 888-893.

Rencher, A. C. & Christensen, W. F. (2002) *Methods of Multivariate Analysis.* New York: John Wiley & Sons.

Olson, K. (2006) Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, Vol. 70, No. 5, Special Issue 2006, 737-738.

Sharma, S (1996) *Applied Multivariate Techniques*. New York: John Wiley & Sons.

Statistiska centralbyrån (2011) Quality of Data in the Swedish Meadow Survey, Learning organisations matters, Flex-3: 21-49

Statistics solutions, (2016) *Mann-Whitney U test.* Tillgänglig: http://www.statisticssolutions.com/mann-whitney-u-test/ [datum, 20 mars 2016].

Stat505, Applied Multivariate Statistical Analysis (2016). *The Two-Sample Hotelling's T-square Test Statistic.* Tillgänglig: https://onlinecourses.science.psu.edu/stat505/node/124 [datum 2 maj 2016]

Stelacon (2012) *Teknisk beskrivning Arbetsmiljöverkets Nulägesundersökning SAM 2012*. Stockholm: AB Stelacon

Thode, H. C (2002) *Testing for Normality*. USA: CRC Press.

Wackerly, D. D., Mendenhall III, .W. & Scheaffer, R. L. (2008*) Mathematical Statistics with Applications, 7rd edn.* USA: Brooks/Cole.

Wagner, J. (2012) Research Synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, Vol. 76, No. 3, Fall 2012, 555-575.

Wright, G. (2015) An empirical examination of the relationship between nonresponse rate and nonresponse bias. *Statistical Journal of the IAOS 31* 305-315 DOI 10.3233/SJI-140844.

Zimmerman, D. W. (2011) A simple and effective decision rule for choosing a significance test to protect against non-normality. *British Journal of Mathematical & Statistical Psychology* 64(3): 388-409.

Appendix A: Variabler och Branscher

Tabell A.1 Visar vilka standardiserade frågor som konstruerat 4 av 5 kompositvariabler som används i denna bortfallsanalys	
Kompositvariablers konstruktion	
k_ind	(F66NY+F63NY+F65dailylearn+ F71NY+F70NY+F69NY+F64NY) / 7
k_strukt	(F59NY+F62NY+F60NY+F55NY+F56NY+F61NY) / 6
k_num	(F57NY+F47NY+F46NY+F48NY+F53NY+F51B_NY+F50NY+F49NY) / 8
K_dec	(F67rotate+F36NY+F35NY+F37NY+F38NY) / 5

Källa: Kodbok 2012, Arbetsmiljöverket 2014:1

Tabell A.2 Visar vilka standardiserade frågor och delvariabler som konstruerat samindex (Totalindex i kodbok 2012)	
Konstruktion av samindex	
VAD	(VAD9+VAD10+VAD11+VAD13+VAD14) / 5
VEM	(HUR19+HUR20+HUR21+HUR22) / 4
REF	F18
RISKUND	VAD7
samindex	(VAD+VEM+REF+RISKUND)

Källa: Kodbok 2012, Arbetsmiljöverket 2014:1

A.1 Branscher

Tabell A.1.1 Fördelningen av branscher för de olika grupperna Svarat i NU2012, N-NU15, J-NU15 samt urvalet från NU2015

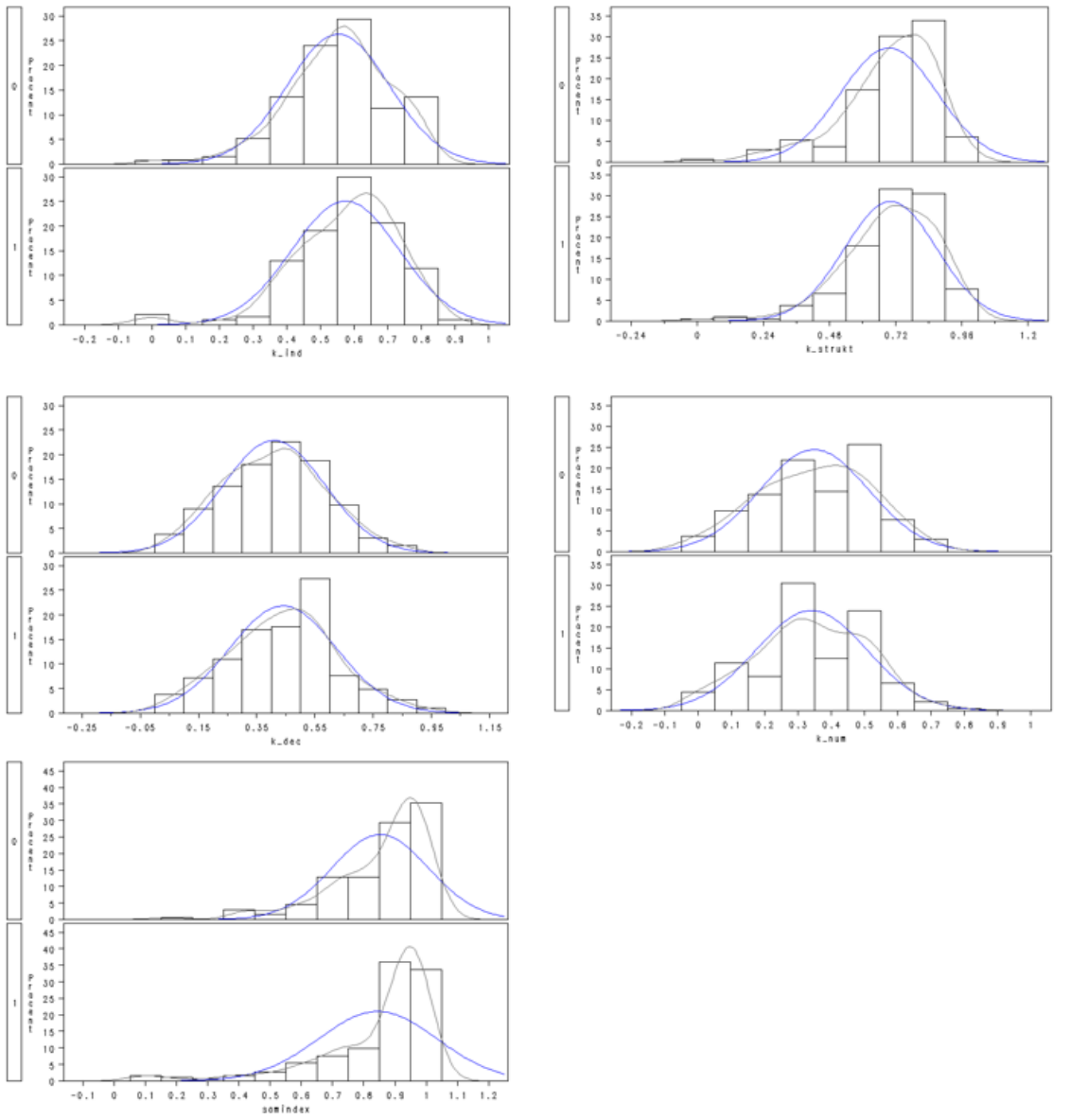
Näringsgren(SNI 2007)			Svarat i NU2012 Bransch Förd. (%) n=1993	N-NU15 Bransch Förd. (%) n=133	J-NU15 Bransch Förd. (%) n=184	Urval NU2015 Bransch Förd. (%) n=317
Bransch-grp (Ng1)		Beskrivning				
1	A	Jordbruk, skogsbruk, fiske	4,06	6,77	9,78	8,52
2	C1	Tillverkning arbetskraftsintensiv industri	4,77	1,50	4,35	3,15
3	C2	Tillverkning kunskapsintensiv industri	5,02	3,01	3,26	3,15
4	C3+B	Tillverkning kapitalintensiv industri	5,07	7,52	1,63	4,10
5	D+E	El, gas, värme, kyla (D), Vattenförsörjning, avlopps-rening, avfall, sanering (E)	4,82	6,02	9,78	8,20
6	F	Byggverksamhet	4,97	4,51	3,26	3,79
7	G	Handel	4,38	2,26	0,54	1,26
8	H	Transport, magasinering	4,33	1,50	4,35	3,15
9	I	Hotell, restaurang	4,33	4,51	3,26	3,79
10	J	Information, kommunikation	4,62	1,50	4,89	3,47
11	K	Finans, försäkring	4,08	9,77	7,07	8,20
12	L	Fastighetsverksamhet	4,67	6,02	7,07	6,62
13	M	Ekonomi, juridik, vetenskap, teknik	4,62	4,51	2,17	3,15
14	N1	Uthyrning, fastighetsservice, resetjänster, andra stödtjänster exkl. personaluthyrning N78.2	4,52	1,50	1,09	1,26
15	N78.2	Personaluthyrning	4,33	3,01	1,09	1,89
16	O+U	Offentlig förvaltning, försvar (O) respektive internationella organisationer i Sverige (U)	5,32	18,05	17,93	17,98
17	P priv	Utbildning i privat sektor, inkluderar näringslivets och hushållens organisationer	5,23	1,50	2,17	1,89
18	P off	Utbildning i offentlig sektor, inkluderar näringslivet och hushållens organisationer	5,48	5,26	7,07	6,31
19	Q priv	Vård och omsorg; social tjänster i privat sektor, inkluderar näringslivets och hushållens organisationer	4,82	1,50	1,63	1,58

20	Q off	Vård och omsorg; sociala tjänster i offentlig sektor, inkluderar arbetsställen i stat, kommun och landsting	5,27	3,76	2,72	3,15
21	R	Kultur, nöje och fritid	4,67	6,02	4,89	5,36

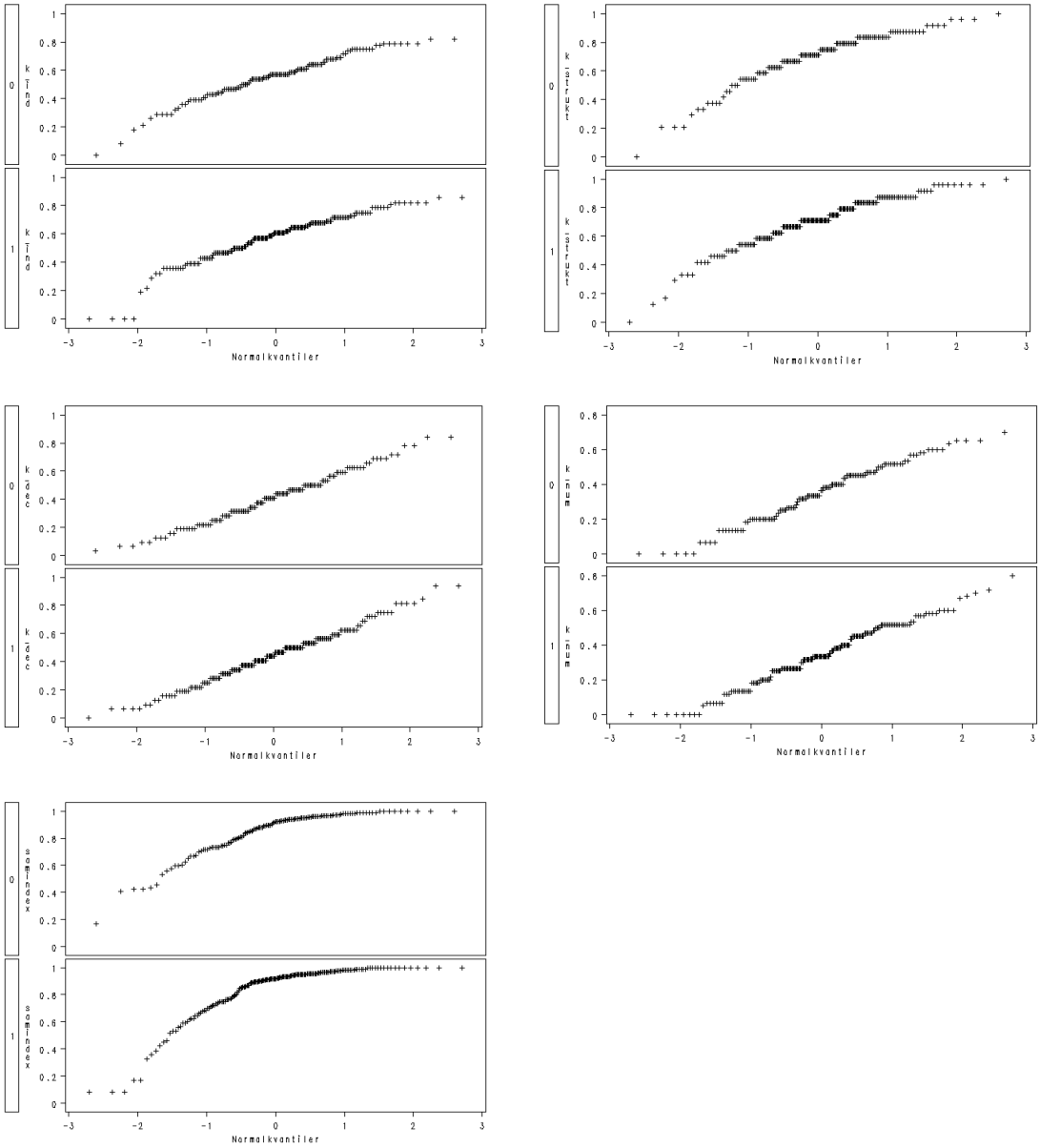
Källa: Arbetsmiljöverket Nulägesundersökning SAM 2012 samt Arbetsmiljöverkets Nulägesundersökning 2015

Appendix B: Test av normalitet

B.1 Normalitet av J-NU15 och N-NU15 för kompositvariablerna



Figur B.1.1 Histogram av k_ind, k_strukt, k_dec, k_num samt samindex för grupper *J-NU15* och *N-NU15*



Figur B.1.2 Q-Q plottar av k_ind, k_strukt, k_dec, k_num samt samindex för grupper *J-NU15* och *N-NU15*

www.av.se

Vår vision: Alla vill och kan skapa en bra arbetsmiljö

ARBETSMILJÖ
VERKET

Arbetsmiljöverket | 112 79 Stockholm | Telefon 010-730 90 00 | av.se