Statistical Modeling and Data Analysis: Factors affecting the income of individuals

Emilie Krutnes Engen

Statistics for data analysis
Carlos III University of Madrid

Universidad
Carlos III de Madrid

Contents

List of Tables

List of Figures

1 Introduction

The distribution of earnings is a topic of continuing interest and thus heavily studied in previous literature. Earnings inequality plays a crucial role in modern macroeconomic, and heterogeneity between different social groups is a continuous debated topic. Early human capital theory and Mincer's equation suggest that there is a relation between earnings and education. In this study we will investigate the distribution of earnings in detail and look into possible earnings differences between different groups in the population.

The objective of this paper is thus to investigate factors affecting earnings, using data from the Programme for the International Assessment for Adult competencies (PIAAC), conducted in Denmark in 2011. Developed by OECD, PIAAC provides comparable data about the adult population's competencies. The survey was performed in 25 countries, and is the most comprehensive survey conducted in this area. PIAAC was implemented to measure and compare cognitive and workplace skills of individuals between countries. The PIAAC survey also collects information about the background of the respondents, including age, gender, working sector, citizenship, earnings, education and experience. This data form the basis for the analysis of earnings, conducted in this study.

The interviews were performed on a random sample of the adult population ranging from 16 to 65 years. This segment of the population consist 3.629.000 persons. The survey was conducted by 7328 respondents undertaken in their homes. The data was collected through interview and an Internet- or paper based test, translated into each individuals native language. The PIAAC data set used in this report is limited to the population from 30 to 60 years. This is done to look at primary working population. This gives a total sample size of 3830 respondents. For the purpose of this study the target population is all adult between 30-60 years, currently living in Denmark.

The remainder of this paper is organized in seven sections. Section 1 introduces the variables used in this study and explains how the data was collected. Section 2 presents the univariate description of the variables. Section 3 presents a bivariate analysis to study the hourly earnings. In Section 4 confidence intervals are constructed and based on these some hypothesis tests are performed. In Section 5 stratified sampling is conducted to estimate the earnings for different groups of the population. In Section 6 the fit of different theoretical density functions are tested for the distribution of hourly earnings. Finally, Section 7 summarizes and concludes the analysis performed.

2 Description of Variables and Recoding

The target population includes adults between the age of 30 and 60 who reside in Denmark during the time of the data collection. This constitutes all adults regardless of citizenship, nationality and language. The target population does however not include people living institutions, such as prisons, hospitals and nursing homes.

The PIAAC survey employed a multistage sample design using the population register as a sample frame. According to Mohadjer et al. (2013) undocumented immigrants are not completely covered in the sample. The exclusion rate is less than 0.1 %. The sample design used in the PIAAC survey is a self-weighting design, where each individual in the target population have equal probability of selection. The sample design used is a one-stage stratified sample design where the only sample unit used is persons. The sample size was allocated to strata according to the formula in Equation 1:

$$N_h = P_{hl} \times N_h \tag{1}$$

where n_h is the number of persons sampled in stratum h, P_hl is the probability of selecting person l in strata h and N_h is the number of eligible persons in stratum h (Mohadjer et al., 2013). The size of the target population is, according to Statistics (2016) 2.324.749 person. The sample size used for the purpose of this report is 3830.

The PIAAC survey investigates cognitive and professional competencies, implied by three index variables: literacy, numeracy and problem solving. In addition to assessing the skill level, the participants also provided information regarding their background. The variables studied in this resport are age, hourly earnings, years schooling, experience, gender, working sector and immigrant status. All of the variables below were conducted through a personal interview and the answered were logged by the interviewer.

2.1 Qualitative Variables

In the data set the qualitative variables are represented by mutually exclusive dummy variables. All the qualitative variables in this data set are nominal, as none of them have any particular ordering.

Gender The qualitative variable gender includes two mutually exclusive categories: male and female. Both categories are included in the data set as dummy variables. The dummy variable *male dummy* is 1 if the respondent is male and 0 if it is female. Similarly *female dummy* defines if the respondent is female.

Immigrant status This variable is represented by two dummy variables, stating whether the respondent is native or immigrant. The *native dummy* is 1 if the individual is a Danish resident and 0 for other residences. The *immigrant dummy* similarly defines whether the individual is an immigrant. The two dummy variables are disjoint.

Working sector The working sector is represented by the two dummy variables, *public sector dummy* and *private sector dummy.* The binary variable *public sector dummy* is 1 if the respondent is currently working in a public sector company and 0 otherwise. Similarly *private sector dummy* is 1 if the respondent is currently working in a company in the private sector and 0 otherwise.

2.2 Quantitative Variables

The data set includes several quantitative variables. For the purpose of this study, the quantitative variables considered are age, hourly earnings, years of schooling and work experience. A description of the variables is presented below.

Age The age of the respondents is a discrete variable, measured in years. The data set is limited to respondents between 30 and 60 years.

Hourly Earnings Hourly earnings is a continuous variable measuring the respondents income per hour in Danish kroners (DKK). In the survey the respondents reported their weekly or monthly earnings and their working hours to the interviewer. Based on this information hourly earnings were calculated and in the data set the calculated earnings per hour is presented, measured in DKK.

Years of Schooling Years of schooling is a discrete variable that is referring to years of education undertaken by the individuals, including primary school, high school and higher education or university degrees.

Work Experience Experience is a discrete variable stating the number of years of work related experience obtained by the respondents.

3 Univariate description

In this section all variables are investigated in isolation, by calculating the distributions of the different variables from the data set. Location measures and dispersion measures are further calculated for all quantitative variables. Table 1 provides a summary of the location- and dispersion measures, including mean, median, minimum and maximum values, standard deviation (St. Dev.), coefficient of variance (CV), skewness and the interquartile range (IQR).

Table 1: Location measures and dispersion measures for the quantitative variables

Variable	Mean	Median	Min	Max	Range	St. dev.	CV	Skewness	IQR
Age	46.49	47	30	60	30	8.82	18.97	-0.1746	16
Earnings	209.70	189.50	0.00047	4479.77	4479.77	121.07	57.73	15.5102	73.76
Education	13.53	15	6	20	14	2.65	19.60	-0.0762	3
Experience	25.53	26	0	47	47	10.59	41.48	-0.1179	17

3.1 Gender

The sample show a close to equal gender distribution, having 51 % male respondents and 49 % female. According to (Statistics, 2016) this distribution corresponds to the gender distribution for the given age range at the time of the data collection. Therefore the distribution seems like a reasonable representation for the target population. By using the relative frequencies from Table 2 a barplot illustrating the distribution is presented in Figure 1. Notice that the frequencies in Table 2 are measured for each qualitative variable in isolation. By summing the absolute frequencies for each variable you get the total sample size of 3830.

Table 2: Frequency table for the different qualitative variables in isolation

	Gender		Immigrant status		Sector	
Frequencies	Male	Female	Native	Immigrant	Private	Public
Abs. freq. (n_j)	1939	1891	3098	732	2361	1469
Rel. freq. (f_j)	0.51	0.49	0.81	0.19	0.62	0.38

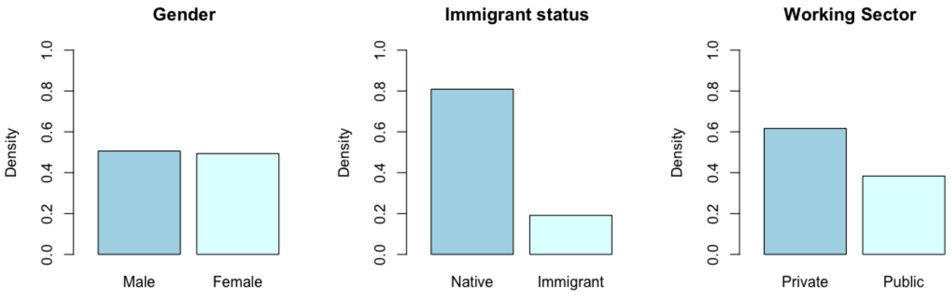

Figure 1: Barplots for distributions of gender, immigrant status and working sector

3.2 Immigrant Status

According to the frequencies in Table 2, 81 % of the sample are natives and 19 % have other backgrounds. According to Rosdahl and Jørgensen (2016) the number of immigrants were oversampled to distinguish between western and non-western immigrants. This information is not included in this study. As a result the oversampling in this context represent a misrepresentation of the target population. The immigrants may show different characteristics compared to natives. According to Statistics (2016) immigrants represent 10.20 % of the danish population. The distribution of native and immigrant respondents are illustrated using a barplot in Figure 1.

3.3 Working Sector

The distribution between the working sectors presented in Table 2, show that 62 % of the sample is working for a company in the private sector and 38 % is working in the public sector. The barplot of the distribution between the two sectors is presented in Figure 1.

3.4 Hourly Earnings

The distribution of hourly earnings is presented by grouped frequencies in the histogram in Figure 2. The recorded frequencies show that more than half of the sample recorded earnings between 100 and 200 DKK per hour and a third between 200 and 300 DKK. The calculated mean earnings is 209.70, with a standard deviation of 121.07 DKK, as presented in Table 1.

The boxplot in Figure 2 illustrate the location of the outliers, where some are far from the interquartile range (IQR). In this case the box plot is useful for identifying the outliers, but because of the large range of recorded hourly earnings the box plot fails to visually represent the data within the IQR.

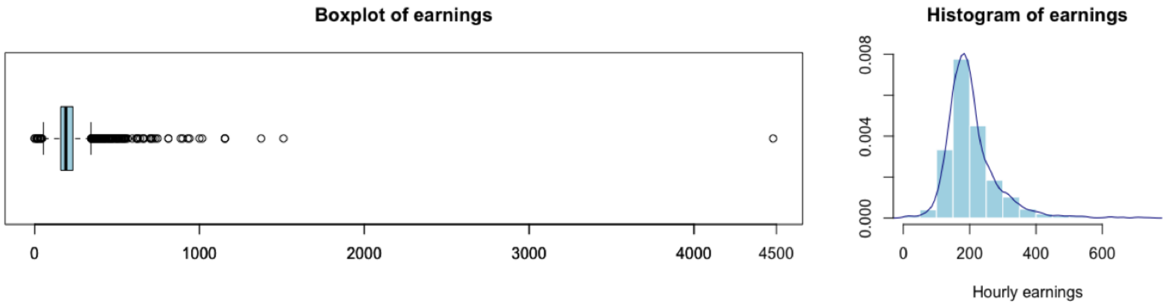

Figure 2: Boxplot and histogram for the distribution of hourly earnings

The histogram in Figure 2 show the distribution between the different earnings intervals, limited to 1000 DKK. The data points outside this interval would not be visible in a histogram. The histogram suggests that the distribution of earnings is positively skewed, also referred to as right skewed. The sample skewness is 15.51, presented in Table 1. This implies that the data is very likely to have a positive skew. Notice that the standard deviation is very high, and the coefficient of variance (CV) in Table 1 suggest that average deviation from the mean, compared to the mean itself is about 58 %. It is also worth noticing that there are 565 missing data points for this variable. The minimum value in Table 1 is slightly above zero, which imply that the large number of missing values may be due to unemployment. However this is not known with certainty.

3.5 Age

The distribution of age is presented by dividing the respondents into six different age groups. The histogram for the distribution of age is presented in Figure 3. Similarly to immigrant respondents persons aged between 55 and 60 is oversampled, due to the desire to investigate the older population's expectations about retirement (Rosdahl and Jørgensen, 2016). This explains the higher relative frequency in the age interval 55-60 years in Figure 3. Apart from the age interval 55-60 the age groups show an approximately uniform distribution and compared to data from (Statistics, 2016) at the time of the data collection, this sample is representative in terms of the age distribution.

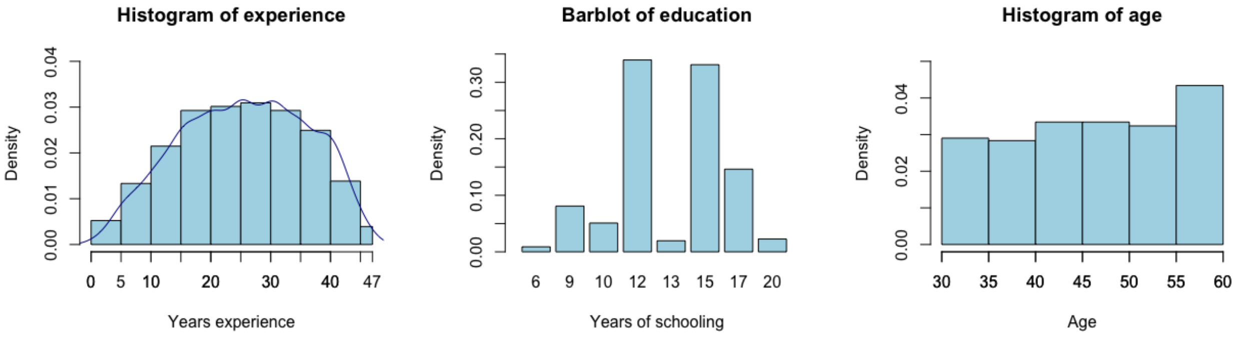

Figure 3: Histogram for age distribution, barplot for educational distribution and histogram for the distribution of work experience

3.6 Years of Schooling

The distribution of years of schooling represented by the relative frequencies is presented in a barplot in Figure 3. The barplot indicate that a large part of the population undertake 12 or 15 years of schooling. In Denmark, primary and secondary school is generally conducted in 9 or 10 years. The distribution between 12 and 15 years are close to equal.

3.7 Work Experience

The distribution of working experience in the sample is presented in the histogram in Figure 3, ranging from 0 to 47 years of experience. The interquartile range presented in Table 1 suggests that a large part of the population have work experience between 17 and 34 years. According to the histogram in Figure 3 the experience is approximately normally distributed.

The skewness presented in Table 1 is slightly negative, which may be related to the oversampling of adults aged between 55 and 60. However, with a CV about 20 % and the relatively large tails in the histogram may indicate that experience may follow a t-distribution.

4 Bivariate description

4.1 Qualitative Variables

When looking into different conditional distributions, there were particularly one that yield interesting results, the distribution of working sectors for males and females. The grouped barplot in Figure 4 suggest that almost 80 % of the male respondents work in privat sector companies, while about 55 % of the female respondents work in public sector. We recall that the distributions between the private- and public sector is 60 % and 40 %, respectively. This imply a tendency of more men working in private sector and more women working in public sector.

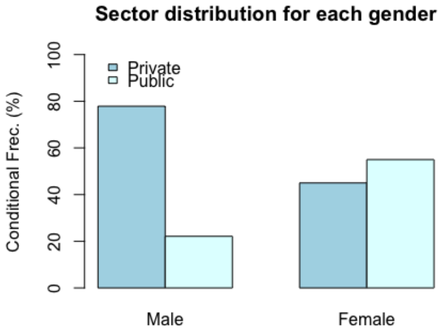

Figure 4: Grouped barplot of sector distribution for a given gender

Other conditional distribution, such as the distribution of immigrants and natives for given genders and working sectors were also investigated. However,these did not yield any interesting differences.

4.2 Hourly Earnings and Qualitative Variables

In this section, the hourly earnings is investigated in more detail, by looking at the differences within the different qualitative variables. From Table 3 we have that both the mean and the median are higher for males than for females. There is also a difference in the median, however this is not as large as the difference in the mean. Figure 5 suggest that female earning are slightly more positively skewed, than male earnings. The difference in CV suggest that female earning show a greater variation than male.

Table 3: Comparing measures for hourly earnings

	Hourly earnings			
	Mean	Median	St. Dev	CV
Male	224.1412	199.9688	111.6216	49.79967
Female	196.0755	183.0686	127.9431	65.25198
	Mean	Median	St. Dev	CV
Native	215.4205	193.7197	122.9298	57.06501
Immigrant	185.1348	164.0271	109.4320	59.10938
	Mean	Median	St. Dev	CV
Private	219.2449	196.2241	109.9562	50.15221
Public	196.9216	184.8419	133.5829	67.83558

When comparing hourly earnings for natives and immigrant, Table 3 suggest that mean hourly earnings are higher for natives than for immigrants. The median is also higher for natives. The CV is quit similar for the two groups, but they both have show large deviations from the mean. The histogram in Figure 5 suggest that the distribution of hourly earnings for immigrants have a higher positive skew, than natives. However they both show a positive skew.

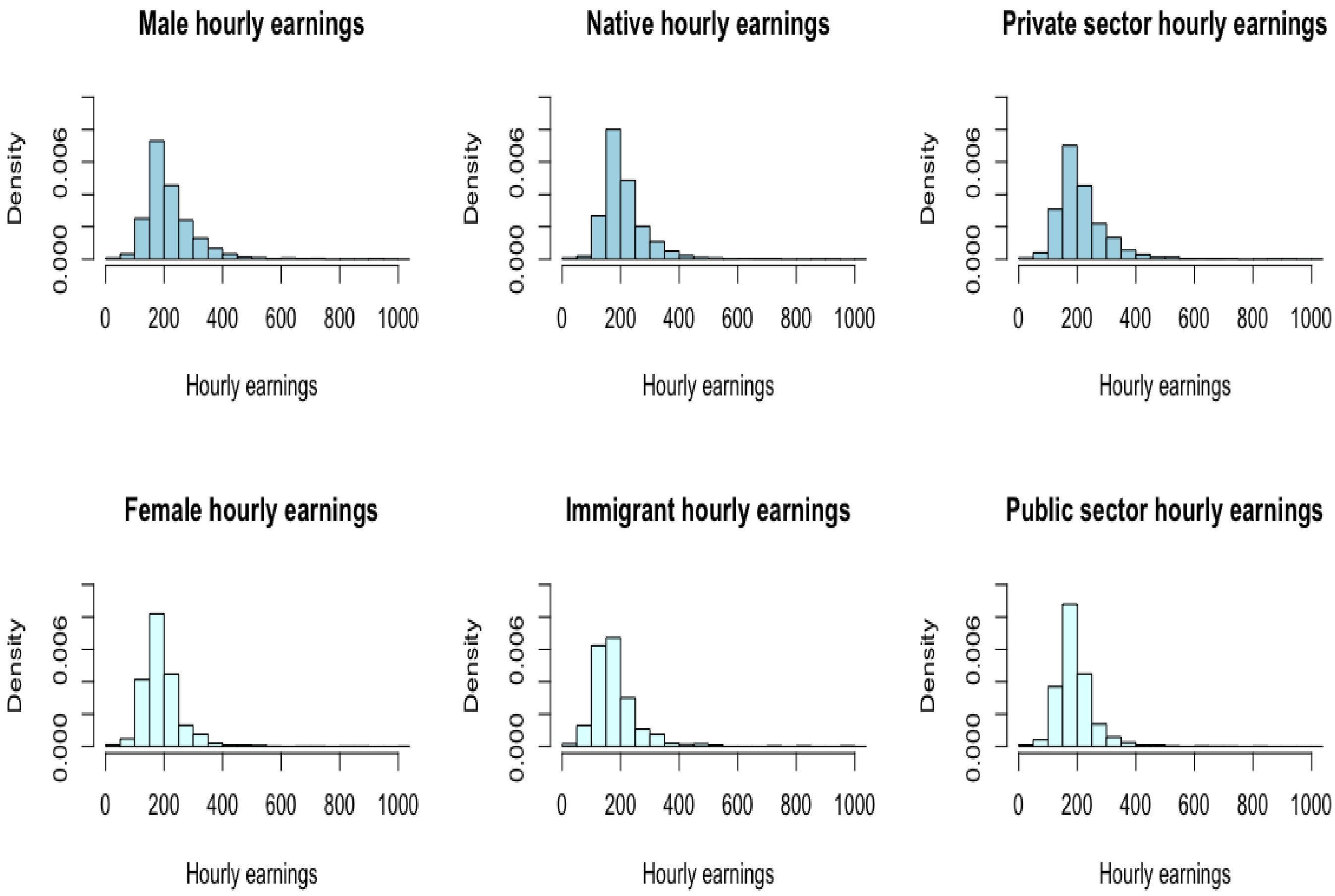

Figure 5: Histogram of hourly earnings for different qualitative variables

Similarily, when looking at the difference in mean earnings between the private- and public sector in Table 3, the mean earnings in private sector seems to be higher than in public sector. Also here, the histogram in Figure 5 indicate that the earnings in public sector is more positively skewed. Both sectors show a high CV, however the CV in public sector is higher than in private sector. This may indicate a higher deviation in earnings in the public sector.

4.3 Hourly Earnings and Other Quantitative Variables

In this section we investigate the relation between hourly earnings and other quantitative variables. Hourly earnings is the target variable investigated by using experience, years of schooling and age as explanatory variables.

Hourly earnings and experience Figure 6 presents scatter plots for experience and hourly earnings. In the first plot, all outliers are included, while in the second the axis are limited for better visualisation. The red line represent the regression line, which indicate a slight positive correlation between hourly earnings and years of experience.

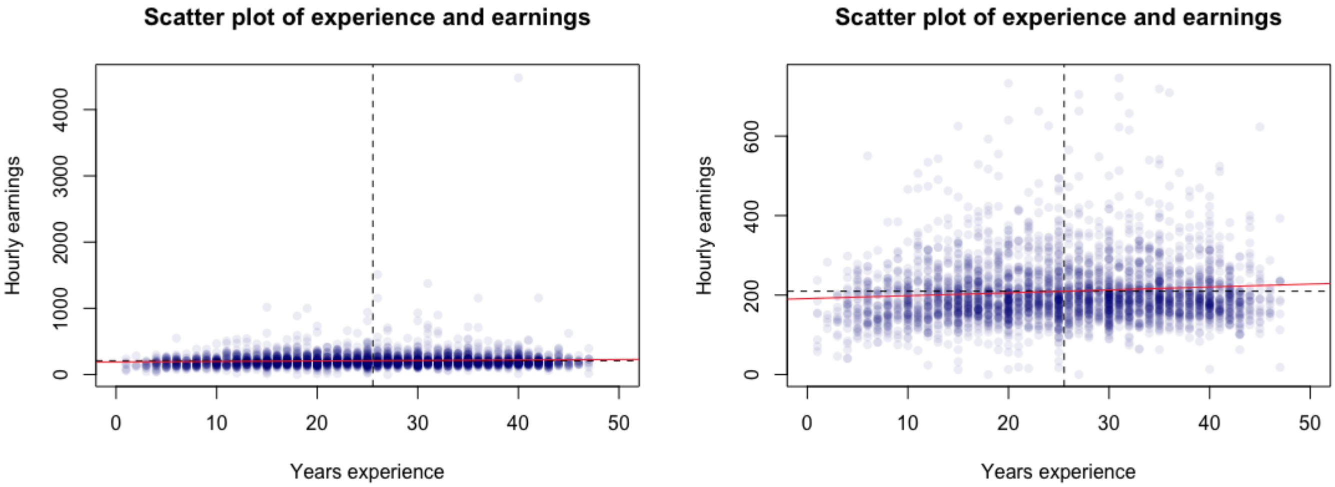

Figure 6: Scatter plot of hourly earnings and years of experience

Table 4: Comparing covariance and correlation for earnings and experience

	$y = x$	$y = log(x)$	$log(y) = x$
Covariance	82.51171	5.674462	0.3855437
Correlation	0.06408651	0.08433568	0.07798125

In Table 4 the covariance and Pearson's correlation coefficient is calculated for different regressions, where y is hourly earnings and x is years experience. According to this, the correlation is 6.4 %. The correlation coefficient in Table 4 further suggests that applying a logarithmic transformation yield a slight improvement in the correlation. Expressing hourly earnings as a function of log experience yield the best improvement in the correlation coefficient. In Figure 9 the two logarithmic transformations are presented. Notice that the y axis, representing hourly earnings have been limited for better visualisation of the data.

Hourly earnings and years of schooling In Figure 8 scatter plots for experience and hourly earnings is presented. In the first plot, all outliers are included, while in the second the axis are limited for better visualisation. The regression line indicate a positive correlation between hourly earnings and years of experience.

In Table 6 the covariance and correlation coefficients for different regressions, where y is hourly earnings and x is years of schooling. According to this, the correlation is 25.6 %. When applying a logarithmic transformation to hourly earnings the correlation increase to 30.2 %. This may some a regression function including an exponential element. In Figure 9 the two logarithmic transformations are presented.

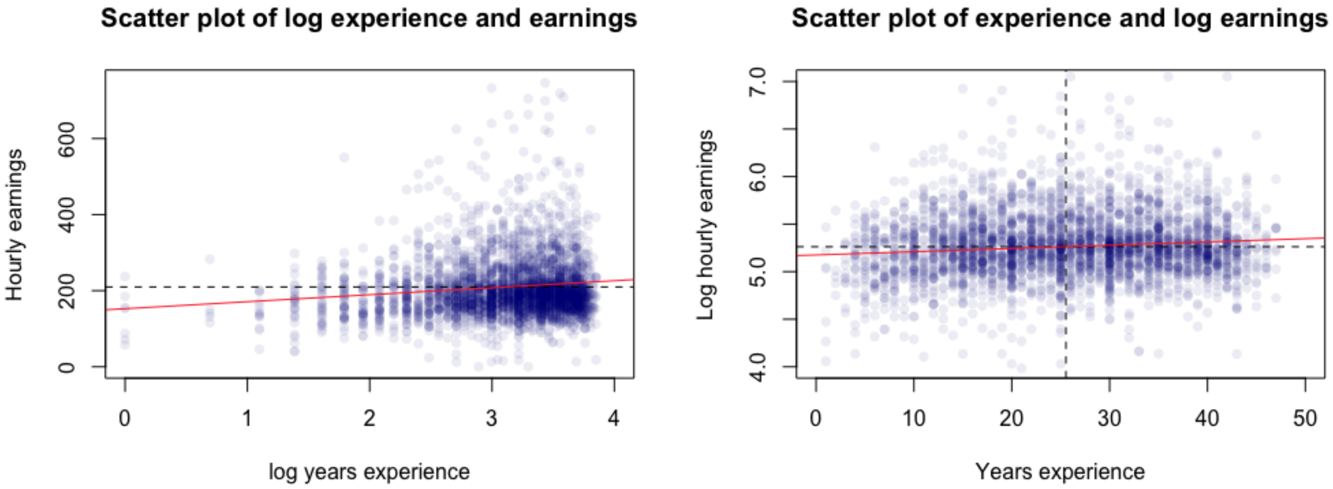

Figure 7: Scatter plot of hourly earnings and years of experience using a logarithmic transformation

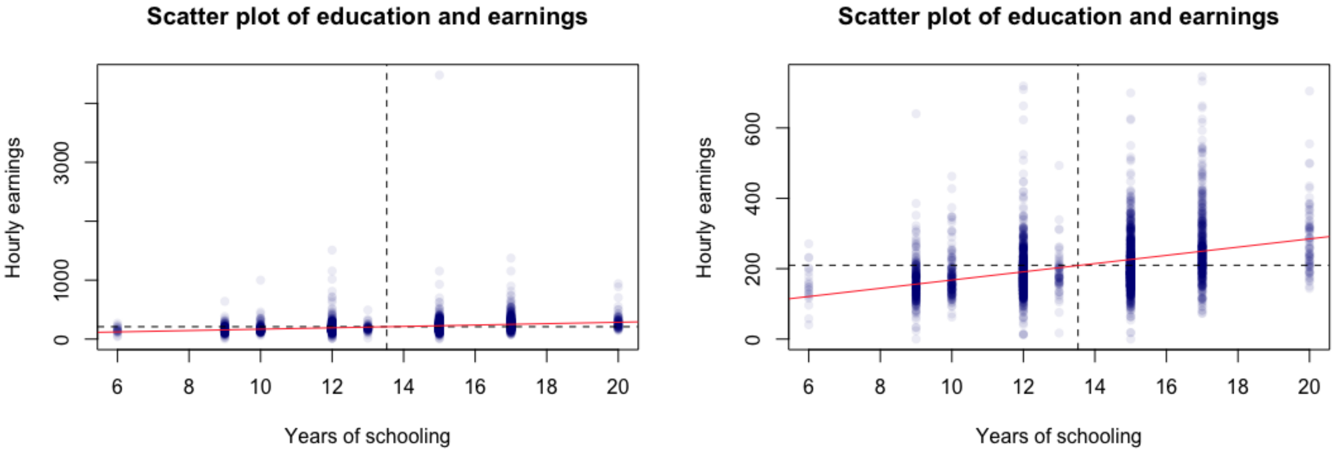

Figure 8: Scatter plot of hourly earnings and years of schooling using

Table 5: Comparing covariance and correlation for earnings and education

	$y = x$	$y = log(x)$	$log(y) = x$
Covariance	82.0481	6.098998	0.3717506
Correlation	0.2560064	0.2451026	0.3020967

Hourly earnings and age In Figure 10 scatter plots for age and hourly earnings are presented. In the first plot, all outliers are included, while in the second the axis are limited for better visualisation. The correlation coefficients in Table 6 indicate a slight positive correlation between hourly earnings and age, with a slight improvement by using the logarithm of age.

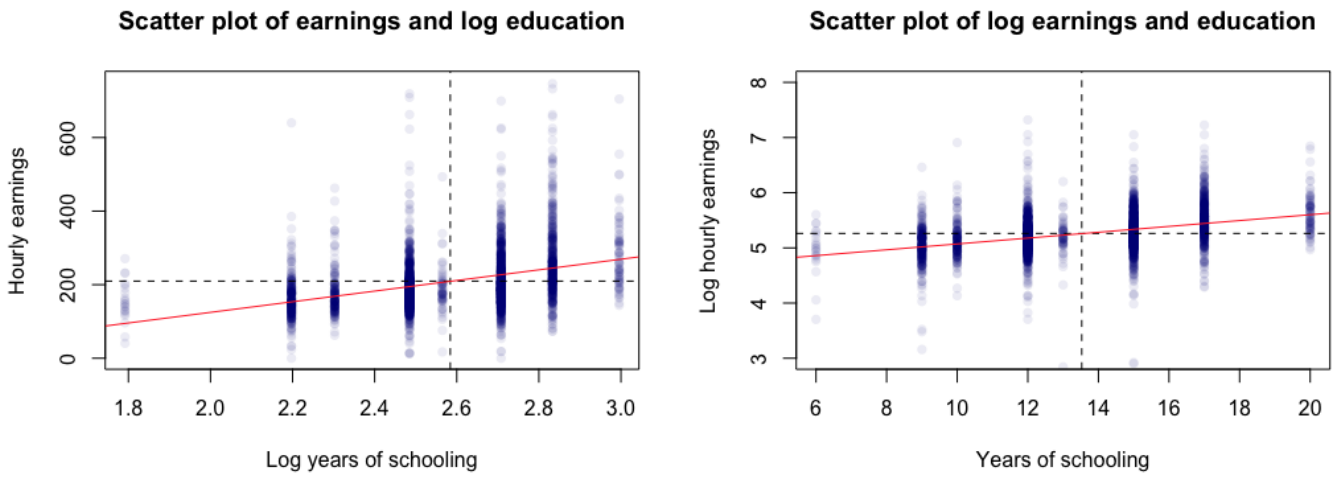

Figure 9: Scatter plot of hourly earnings and years of schooling using a logarithmic transformation

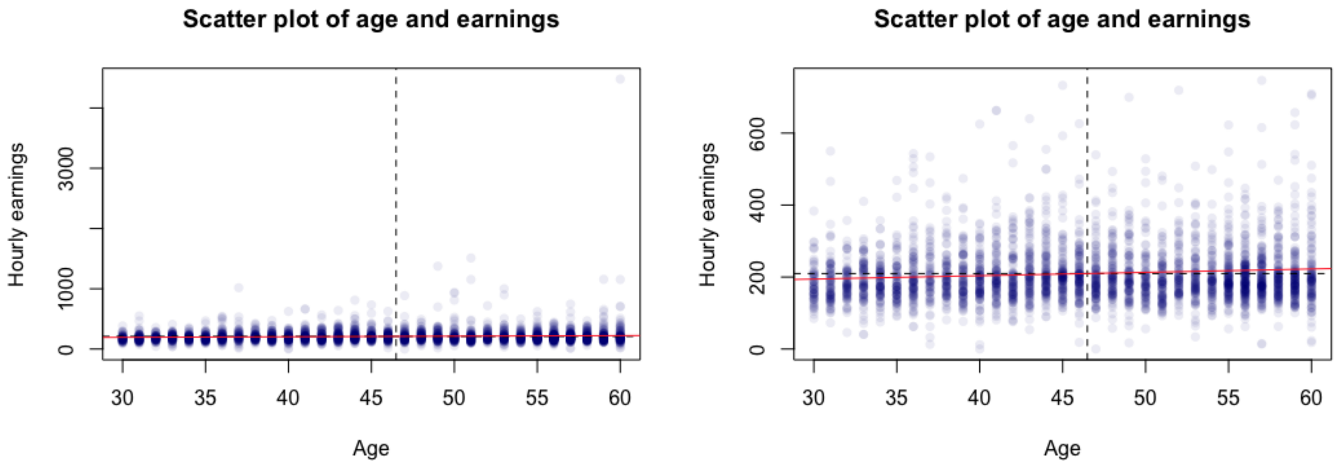

Figure 10: Scatter plot of hourly earnings and age

Table 6: Comparing covariance and correlation for earnings and age

	$y = x$	$y = log(x)$	$log(y) = x$
Covariance	73.82891	1.724068	0.2348142
Correlation	0.06870371	0.07121625	0.05690002

5 Inference

5.1 Confidence Interval

Confidence interval for the difference of males in private and public sector In this section we want to estimate the difference between the proportion of males working in public and private sector, where the two population proportions is referred to as p_1 and p_2, respectively. The two proportions are independent, and a confidence interval for $p_1 - p_2$ can therefore be established by considering the estimated difference, given by $\hat{p_1} - \hat{p_2}$. By the central limit theorem, the distribution of $p_1 - p_2$ is considered approximately normally distributed, with mean $\mu_{\hat{p_1}-\hat{p_2}} = p_1 - p_2$ and variance $\sigma^2_{\hat{p_1}-\hat{p_2}} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$, where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$ and n_1 and n_2 is the sample size for the private and public sector, respectively. By using the sample sizes given in Table 7 we obtain the 95 % confidence interval:

$$0.3167110 < p_1 - p_2 < 0.3783373 \tag{2}$$

Because the interval does not include the value 0, there is reason to believe that there is a significant difference between the two proportions. The confidence interval is positive and this indicate that the true proportion of males in private sector is larger than the proportion of males in public sector.

Table 7: Absolute frequencies of male and female in private and public sector

	Male	Female	Total (n_i)
Private	1510	851	2361
Public	429	1040	1469
Total	1939	1891	3830

Confidence interval for the difference of mean hourly earnings between males and females We further obtain an interval estimate for the difference to the mean between males and females $\mu_1 - \mu_2$. The two variables are considered independent and the population mean and variance are unknown. By computing an F-test, we obtained a 95 % confidence interval for the ratio between the variance of male and female earnings of [0.6907,0.8388]. The confidence interval does not include 1 and the variance is therefore assumed unequal between the two populations. By the Central Limit Theorem, the population is assumed normally distributed. We use the t-distribution with v degrees of freedom. Because the sample sizes are different, we have that v is given by (Walpole et al., 1993):

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \tag{3}$$

We then obtain the 95 % confidence interval:

$$19.83834 < \mu_1 - \mu_2 < 36.29324 \tag{4}$$

Because the interval does not includes the value zero, there is reason to believe that there is a significant difference between the earnings of male and female. The positive interval further suggest that the hourly earning for males is higher than for females, with a difference between 19.8 % - 36.3 %.

5.2 Hypothesis Testing

Hypothesis testing for the equality of males in private and public sector We extend our study from the previous section by testing the equality between the proportion of males working in private and public sector, where the two population proportions is referred to as p_1 and p_2, respectively. In the construction of the confidence interval we concluded that the we noted that the confidence interval for $p_1 - p_2$ by the point estimator $\hat{p_1} - \hat{p_2}$ is approximately normally distributed with with mean $\mu_{\hat{p_1}-\hat{p_2}} = p_1 - p_2$ and variance $\sigma^2_{\hat{p_1}-\hat{p_2}} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$. The constructed confidence interval suggested that the proportion of males in private sector p_1 is larger than the proportion of males in public sector p_2. From this we construct a one sided hypothesis for the two proportions:

$$H_0 : p_1 = p_2 \tag{5}$$

$$H_1 : p_1 > p_2 \tag{6}$$

The obtained confidence interval for the difference of the two proportions are [0.3221283,1], which suggests with 95 % confidence that the difference between the two proportions is larger than 32 %. The p-value given from the two sample test is $2.2 \times 10^{-}16$, which is less than $\alpha = 0.05$. This leads us to reject the hypothesis H_0 of equal proportions between males in private and public sector.

Hypothesis testing for equality of mean hourly earnings between males and females in private sector We extend our study of the confidence interval, by performing a hypothesis test for the equality of mean hourly earnings between males and females working in private sector, referred to as μ_1 and μ_2. By computing an F-test, we obtained a 95 % confidence interval for the ratio between the variance of male and female earnings in private sector of [1.3077, 1.6009]. The confidence interval does not include 1 and the variance is therefore assumed unequal between the two populations. The two sided hypothesis of the two means is written as:

$$H_0 : \mu_1 = \mu_2 \tag{7}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{8}$$

We use the t-distribution with significance level $\alpha = 0.05$ and obtain the following critical region:

$$-12.472526 < \mu_1 - \mu_2 < 3.974021 \tag{9}$$

We do not reject H_0 because $t = -1.0132$ is within our critical region and the p-value is 0.311, which is greater than $\alpha = 0.05$. Thus we are not able to conclude that there is a significant difference between the mean hourly earnings for male and female in private sector.

6 Sampling

6.1 Sampling Strategy

In simple random sampling, every member of the population have an equal chance of being selected and the selection of a member is independent of the selection of every other member. Thus, simple random sampling chooses a sample by pure chance. However, when the sample size is small, there is nontrivial probability that the sample does not represent the target population. Therefore the sample size needs to be taken into account when performing inferential statistics. To ensure a representative sample when simple random sampling is inappropriate, stratified sampling is an alternative approach that can be used. In this case we have seen that gender, immigrant status and working sector all show differences to mean hourly earnings. By using stratified sampling, these differences can be investigated further.

When determining the total sample size you normally consider the desired precision of the estimation and the sampling cost. In this case we disregard the sampling cost and therefore only consider the desired precision, by deciding an appropriate confidence level, a margin of error and the expected standard deviation. The formula for calculating the sample size is presented in Equation 10.

$$n = \frac{(Z_{0.05}CV)^2}{E^2} = \frac{(1.96 \times 0.58)^2}{0.05^2} = 516.93 \approx 517 \tag{10}$$

For this task only we consider our sample data as the true population. In this section we therefore will refer to the data set as the population. We already have information about the variability in the data set. From Table 1 we have that the coefficient of variance for hourly earnings is about 58 %. Using this and a confidence level of 95 %, we obtain a sample size of 517.

In Table 3 location and dispersion measures in hourly earnings for the different qualitative variables are presented. The confidence intervals calculated in Section 5 suggests that the difference in the mean between males and females is significant, but in the hypothesis test we were not able to reject the hypothesis of equal mean earnings between males and females in private sector. In this section we investigate this further by selecting the combination of gender and working sector, giving four different stratas. In Table 8 the relative frequencies for the different stratas are presented.

Table 8: Cross-classified distribution table for gender and working sector

Relative frequencies (%)		
	Private sector	Public sector
Male	39.43 %	11.20 %
Female	22.22 %	27.15 %

One of the major tasks when designing a stratified sampling design is choosing an appropriate allocation method. There exist several different types of allocation methods, such as equal allocation, proportional allocation, optimal allocation and cost based allocation. Equal allocation implies that the total number of sample units are equally divided between the different stratas. This is a simple allocation method, but it requires a weighted analysis to compute the mean. In proportional allocation each stratum is sampled according to the proportion of the population. This method is self weighted, but when your sample size is small or the number of stratas are high, you may get unreliable results for the stratum having a small sample size. The optimal allocation, also referred to as Neyman allocation is an allocation method sampling according to the variance of the stratas, where those having a larger variance are more intensively sampled. Similarly to the equal allocation method, this also require weighting as it is disproportional to the population.

According to Table 8 it appears that males working in public sector have a small relative proportion p. Using a total sample size n of 517 will give this stratum a low sample size. As this may provide unreliable results for the given stratum, using equal allocation is therefore considered appropriate. By adjusting the total sample size n to 520, each stratum i is given a sample size n_i of 130. Because the equal allocation method is selected, we have chosen to disregard rows containing missing values in each sample. This is to ensure that the date calculated within each stratum is actually based on an equal sample size. Within each stratum, each sample is obtained by simple random sampling without replacement.

6.2 Estimation of Mean Hourly Earnings

By using the sample drawn in Section 6.1, an estimation of the mean hourly earnings $\hat{\delta}$ is calculated for the total population U, which in this case is the initial data set. The size of the total population N is 3830. Because we have chosen equal allocation, the different stratas are weighted according to the proportions, where each proportion p_i is given by $p_i = \frac{N_i}{N}$. The formula for calculating the estimated mean is presented in Equation 11.

$$\hat{\delta} = \frac{1}{N}\sum_{i=1}^{L} N_i Y_i = \sum_{i=1}^{L} p_i Y_i \tag{11}$$

where L is the total number of stratas, p_i is the proportion for strata i in the true population and Y_i is the calculated mean within each strata in the sample population. The estimated mean $\hat{\delta}$ obtain is 215.11 DKK. We recall that the true mean, obtained from the original data set is 209.73. This imply there is an error inferred with the obtained sample.

In Table 9 the estimated means is compared to the true means for each strata. The difference is the percentage over the true mean. When looking at the estimated mean for each stratum, we have that for all stratas except males working in private sector, the true mean is higher than the estimated mean. Other than that we observe that the mean earnings for males are higher than for females in both private and public sector. This includes both the true and estimated means.

Table 9: Comparison of true and estimated mean for the different stratas

	Male		Female	
	Private sector	Public sector	Private sector	Public sector
True mean	234.3694	214.1462	209.596	194.2858
Estimated mean	244.521	208.9982	203.4757	184.442
Difference	-4.331454	2.40397	2.920051	5.066681

6.3 Estimation of Population Proportion

To investigate the sample further, the population proportion between natives and immigrants is estimated. When calculating the estimated proportion, no weighting scheme is applied. The frequencies obtained from the sample is presented in Table 10.

Compared to the true population, the proportion of natives is higher in the sample. In the original data set the proportion of natives and immigrants are 0.81 % and 0.19 %, respectively.

When looking at each stratum in isolation we have that the estimation of the proportion between native and immigrants are vary low for males working in public sector. The conditional distributions are pre-

Table 10: Frequency table for estimated distribution of immigrant status

Immigrant status	Abs. freq. (n_j)	Rel. freq. (f_j)
Native	443	0.85
Immigrant	77	0.15
Total	520	1.00

sented in Table 11. This may represent a potential source of error when estimating the mean earnings in the different stratas.

Table 11: Estimated conditional distribution of immigrant status with relative frequencies (%)

	Male		Female	
Immigrant status	Private sector	Public sector	Private sector	Public sector
Native	93.08	99.88	83.85	77.69
Immigrant	6.92	0.12	16.15	22.31
Total	100.00	100.00	100.00	100.00

7 Model Selection

7.1 Selection of Probability Distribution

In this section we want to investigate the distribution for hourly earnings. We know from previous sections that the distribution is right skewed. From Figure 11 we have the histogram, cumulative distribution and QQ-plot for hourly earnings.

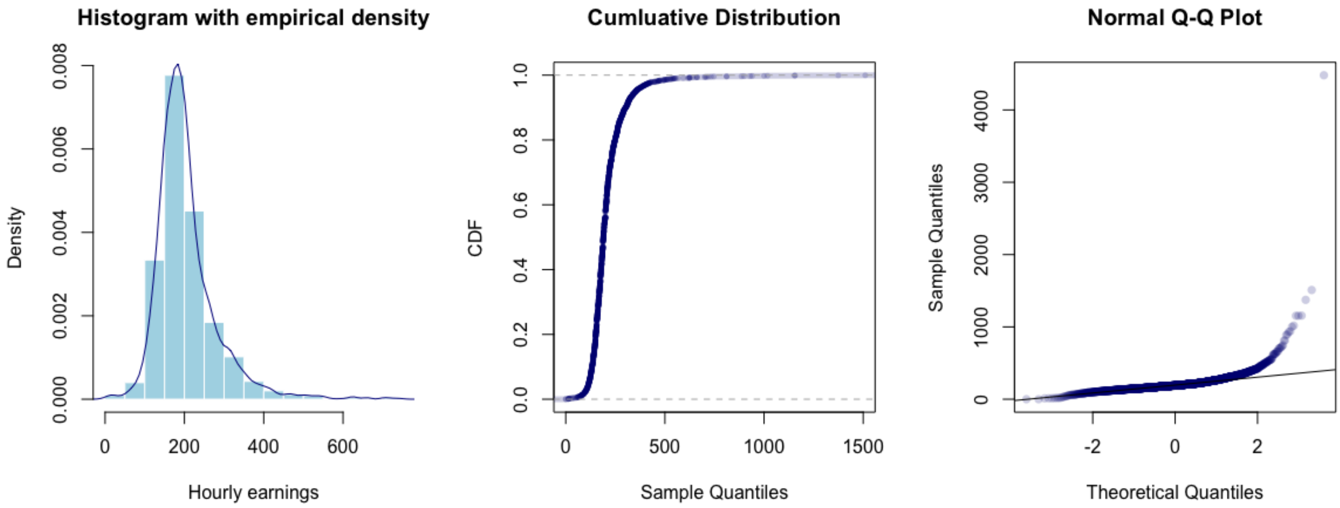

Figure 11: Histogram, cumulative distribution and Q-Q plot for hourly earnings

By looking at the ditribution in Figure 11, a lognormal distribution might provide a good fit. From distribution theory we also have two other possible distributions that may fit our distribution of hourly earnings: the Weibull- and gamma distribution. The Weibull-, lognormal- and gamma distribution functions are given in Equation 12, 13 and 14, respectively, where x correspond to hourly earnings.

$$f(x) = \left(\frac{a}{b}\right)\left(\frac{x}{b}\right)^{(a-1)} e^{-(x/b)^a} \tag{12}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(logx-\mu)^2/2\sigma^2} \tag{13}$$

$$f(x) = \frac{1}{(b^a\Gamma(a))} x^{(a-1)} e^{-(x/b)} \tag{14}$$

7.2 Estimation using Maximum Likelihood

Using maximum likelihood, we obtain the parameters for the three distributions as presented in Table 12. The estimated parameters α and β represent the scale a and shape or rate b in the Weibull and Gamma distribution. For the lognormal distribution we have that $\alpha = \mu$ and $\beta = \sigma$ which is the mean and standard deviation of the logarithm of hourly earnings.

In Figure 12 a histogram with the theoretical densities, a Q-Q plot and the cumulative distribution function (CDF) is presented. From the histogram it appears that none of the distribution manage to

Table 12: Estimated parameters and standard error for theoretical distributions using maximum likelihood

	Weibull		Lognormal		Gamma	
	Estimate	SE	Estimate	SE	Estimate	SE
α	1.8366	0.01688	5.2620	0.0081	6.1303	0.1465
β	234.0051	2.3543	0.4649	0.0058	0.0292	0.0007

correctly describe the center of the distribution. The Q-Q plot emphasize on the lack of fit in the tails of the distribution. From this plot, the lognormal distribution provides the best description of the right tail of the distribution. Finally the cumulative distribution function suggests that lognormal- and gamma distribution best describes the cumulative distribution, which is also apparent in the histogram.

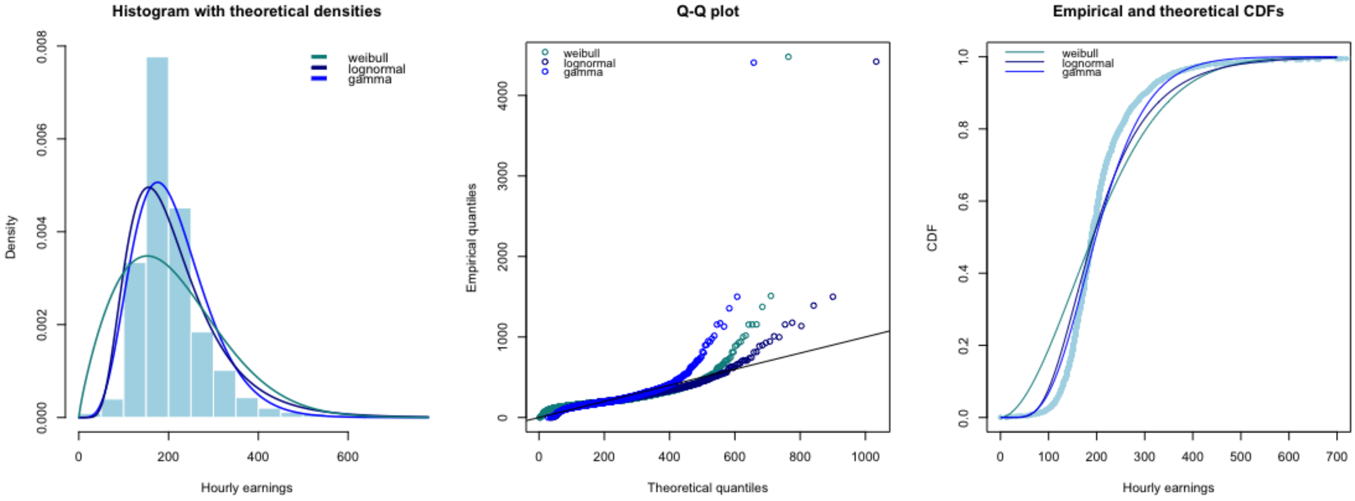

Figure 12: Histogram, cumulative distribution and Q-Q plot for hourly earnings with theoretical densities

8 Conclusion

In this study, data from the PIAAC survey in Denmark is used to investigate several variables. The main purpose of the analysis conducted was to investigate the distribution of earnings in the adult population and to look at different factors that may affect the population's earnings.

The distribution of hourly earnings from the sample used show a large positive skew. This may indicate that a large part of the population have a lower income, while only a few have a very high income. Some differences among gender, working sector and immigrant status are identified. However when comparing males and females working in the same sector, the difference to the mean hourly earnings is not significant. This indicate that the differences in mean hourly earnings is more complex and is not accurately described applying only one qualitative variable.

When related to the quantitative variables in the data set hourly earnings show a slight positive correlation between between age and years of working experience. Applying a logarithmic transformation yield a small increase in the correlation coefficient. When comparing hourly earnings to years of schooling we obtain a correlation of 25.6 %, which is further increased be applying a logarithmic transformation in the hourly earnings variable.

Three different theoretical density functions were tested for the hourly earnings distribution, the Weilbull-, lognormal- and gamma distribution. The results indicate that none of them are able to accurately describe the center of the distribution. However, the gamma- and lognormal distribution seem to be more appropriate for describing the hourly earnings distribution than the Weibull distribution. However, for further studies more advanced testing should be emphasized.

The data set analysed in this study is collected from the PIAAC survey in Denmark. The study was implemented to measure and compare cognitive and workplace skills and not to investigate variables affecting hourly earnings. Therefore the may show some limitations. First some stratas in the data set is oversampled. In order to provide accurate results in the estimations, the oversampling should be accounted for. Due to the limitation of this study, this is not emphasized in this report, but is suggested for further research.

There may be other factors affecting an individuals income, not included in the data set. The data set may show limitations due to the fact that only years of schooling is included, not giving any qualitative information about the studies conducted. It is reasonable to believe that type of education may affect income, as some occupations require a certain education. In addition, the labor demand vary according to fields of study, which may lead to wage differences. Information about whether or not a degree was obtained, may also be relevant. As the data set only provide years of schooling, a student taking 5 years in different fields of study is considered equivalent to a student with a graduate diploma. In addition to type of study, applying categories for elementary school, high school, bachelor degree, master degree and phd studies would be interesting for further investigation of the relation between earnings and education.

Similarily, in addition to years of work experience, investigating the type of experience would possibly provide more insight that could be included in a further investigation of hourly earnings. Further, more details on the type of industry is a natural extension of working sector, and may provide interesting differences in mean hourly earnings.

Further, the data set show some limitations associated in the hourly earnings variable. Firstly, the earnings conducted are pre-tax and disregards bonuses, overtime compensations and pension payments. Thus, in order to get the complete income profile, these have to be estimated. Secondly, as the sample population includes employees between the age of 30 and 60, it is reasonable to assume a large proportion are working full-time. Normally full-time employees receive a fixed annual salary. In this case, the hourly earnings have to be estimated by dividing the monthly or weekly earnings by the number of working hours. The individual may experience variations in the number of working hours. As a result this may create a potential source of error.

The data is limited to one country. The distribution of hourly earnings may show cross-country variations according to differences in labor market and social institutions. As a result, this study only provide insight to the population of Denmark, where the sample data is collected. For further studies, applying the same test across different countries may yield interesting findings.

As income develop throughout the lifetime, having time series data would enabled calculation of income profiles. By estimating the net present value of future income streams, one would be able to explore the variables affecting earnings more closely.

For further research inclusion of more detailed variables is suggested. Including both level of education, type of work experience and working industry would be interesting variables to investigate. Studying the hourly earnings by applying several explanatory variables may also yield interesting findings. Testing the goodness-of-fit for different earnings regressions including several different explanatory variables is highly recommended for further studies. Performing an ANOVA analysis is also recommended for investigating the interactions between the variables in more detail.

References

Mohadjer, L., Krenzke, T. and Van de Kerckhove, W. (2013). Piaac technical report. http://www.oecd.org/skills/piaac/Technical%20Report_Part%204.pdf.

Rosdahl, Anders, F. T. J. V. and Jørgensen, M. (2016). Abilities in reading, calculus and problem solving in denmark. http://www.oecd.org/skills/piaac/Denmark_1328-Danskernes-kompetencer.pdf.

Statistics, D. (2016). Population at the first day of the quarter by time and ancestry. http://www.statbank.dk/statbank5a/selectvarval/saveselections.asp.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (1993). *Probability and statistics for engineers and scientists*, Vol. 5, Macmillan New York.

A Appendix - Tables

The tables below are freqency tables for the different quantitative variables. For the grouped frequency distributions the Interval column represent the range of the different groups. Further the absolute frequency n_j, the relative frequency f_j and the amplitude a_j is given. The amplitude is simply a measure of the size of a given interval $[L_j, L_{j+1}]$ and is calculated as $a_j = L_{i+1} - L_i$. The height h_j for an interval is the relative frequency scaled by the amplitude $h_j = f_j/a_j$. In some tables the Cumulative absolute frequency N_j and Cumulative relative frequency F_j is also given.

Table 13: Grouped frequency distribution of age

Interval	Abs. freq. (n_j)	Rel. freq. (f_j)	Amp. of int. (a_j)	Height (h_j)
[30,34]	460	0.12	4	0.0312
[35,39]	513	0.14	4	0.0348
[40,44]	635	0.17	4	0.0430
[45,49]	653	0.18	4	0.0442
[50,54]	569	0.15	4	0.0386
[55,60]	860	0.23	5	0.0466
Total	3690	1.00	-	-

Table 14: Frequency table for distribution of hourly earnings

Interval	Abs. freq. (n_j)	Rel. freq. (f_j)	Amp. of int. (a_j)	Height (h_j)
[0,100)	79	0.02	100	0.000242
[100,200)	1804	0.55	100	0.005525
[200,300)	1047	0.32	100	0.003207
[300,400)	236	0.07	100	0.000723
[400,4480]	99	0.03	4080	0.000007
Total	3265	1.00	-	-

Table 15: Frequency table for distribution of education

Years	Abs. freq. (n_j)	Cum. abs. freq. (N_j)	Rel. freq. (f_j)	Cum. rel. freq. (F_j)
6	34	34	0.01	0.01
9	310	344	0.08	0.09
10	195	539	0.05	0.14
12	1300	1839	0.34	0.48
13	75	1914	0.02	0.50
15	1268	3182	0.33	0.83
17	560	3742	0.15	0.98
20	87	3829	0.02	1.00
Total	3829	-	1.00	-

Table 16: Frequency table for distribution of experience

Interval	Abs. freq. (n_j)	Rel. freq. (f_j)	Amp. of int. (a_j)	Height (h_j)
[0,5)	62	0.02	5	0.00324
[5,10)	215	0.06	5	0.01124
[10,15)	367	0.10	5	0.01918
[15,20)	513	0.13	5	0.02681
[20,25)	566	0.15	5	0.02958
[25,30)	582	0.15	5	0.03042
[30,35)	600	0.16	5	0.03136
[35,40)	493	0.13	5	0.02576
[40,45)	371	0.10	5	0.01939
[45,47]	58	0.02	2	0.00758
Total	3827	1.00	-	-