Analyzing the gender pay inequality in Europe's IT industry and the role of working experiences

Yiwen Li - 1005749219

4/12/2021

Abstract

Gender inequality has always been an issue that sociologists have worked to address. In the workplace, gender inequality is often reflected in the pay gap between men and women. The purpose of this report is to examine gender pay inequalities in Europe's IT industry. We used the 2020 IT Salary Survey in Europe as sample data, in which 1,253 IT professionals participated. In this report, we use six statistical methods to study gender pay inequality. We first compared the true average salary range of males and females using confidence intervals. We then judged our prediction of the true average salary of female IT employees by hypothesis test. We found that the true average salary range of males is higher than that of females. Besides, by deriving the maximum likelihood estimator, we infer that the salary increase of female IT employees in the EU follows an exponential distribution with lambda equal to 0.151. Then, we analyzed the number of female employees in the high-income group of the IT industry by applying the goodness of fit test and deriving the Bayesian credible interval. The results show that the number of women in the high-income group is much less than that of men. Finally, by the linear regression model, we found a positive relationship between total annual income and work experience, and the average work experience of females is less than that of males. To sum up, gender inequality in Europe's IT industry needs to be taken seriously. Governments and society members can take effective measures to reduce gender discrimination and increase women's work experience, thus improving women's well-being at work.

Introduction

- Wealth is essential for a person to live in this world. More or less, wealth affects the quality of lives and the social resources obtainable to a person. Many people work hard to get into prestigious schools and improve their abilities to earn more wealth to enrich their lives. For people who have a job, their salary largely determines the level of wealth. It is known that the level of salary is directly related to one's ability, but many other factors can affect salary, such as gender. Gender inequality has always been an issue of close concern to society, and the gender pay gap is an important indicator of gender inequality. The gender pay gap calculates the average wage gap between women and men, and the existence of this gap exacerbates the problem of social inequality[1]. Such inequality has a negative impact on women and the economy as a whole, which cannot be ignored. According to the data, the average income of women in Germany was 19% less than that of men in 2019[2]. Therefore, this report will investigate further whether gender pay inequality is even worse in male-dominated work fields.

- This report will focus on the gender pay inequality in the IT sector in the euro area. We predict that gender pay inequality is serious in Europe's IT field. Moreover, the report will analyze the link between salary and working experiences to understand whether the lack of working experiences will lower salary. This report aims to provide some suggestions for closing the gender pay gap, and to raise social awareness of gender inequality.

- To support this analysis, we will use the salary dataset of IT specialists in the EU region in 2020. The data were collected from an anonymous salary survey in 2020, in which 1,253 respondents volunteered to participate.

Data

Data Description and Data Collection Process

- The dataset was collected from the Kaggle website (https://www.kaggle.com/parulpandey/2020-it-salary-survey-for-eu-region), which is publicly available and covers the latest information needed for this report. The dataset contains valuable information about the gender, age, positions, and salary level of IT specialists in Europe in 2020. Since this is a large dataset and we only need some of the information to complete our research, data cleaning is essential.

- The source of these data is an anonymous salary survey, which has been conducted annually since 2015. We will be using data from the 2020 survey to ensure that our research is based on the most up-to-date data. This data is collected and released by the person in charge of the survey. Researchers have already done some studies with this data, so we can assure that this dataset is valid and reliable to a certain extent.

Data Cleaning Process

- While processing the data, We found that among the 1253 respondents, there were individuals with extremely high salaries. Since we will be looking at average salary levels, these very high numbers may make the final results less credible. Therefore, we focused on the data with annual salaries below 300,000 euros. Also, we found several employees whose wages have dropped compared to 2019, and this is not common in real life, so we filtered out these observations.

- We selected four variables from the dataset: Gender, Yearly Gross Salary in 2019, Yearly Gross Salary in 2020, and total years of IT field working experiences. Besides, we used the top 5% as a criterion line to classify people into high-income earners and low-income earners, and calculated the salary rise from 2019 to 2020 for each individual. As per the sample data, workers who earn over 108,000 euros in 2020 rank in the top 5% of IT specialist's salary levels. Also, in order to make the axes of the graph in the following sections more straightforward and more concise, we advanced all the salary values by three decimal places, meaning that the units were changed to thousands of euros.
  After the cleaning process, there are 851 observations and 5 variables in the dataset.

Description of Important Variables

Variable	Description	Types
Gender	Gender of the respondent (Male or Female)	Categorical
Income	Yearly Gross salary(in thousands) without bonus and stocks in the EU Region	Numerical
Working_Experiences	Total years of working experiences as an IT specialist	Numerical
Classification	Classify people into high-income earner and low-income earner (High: earn at least 108K euros per year; Low: otherwise)	Categorical
Payrise	The wage increase in 2020 compared to 2019 (in thousands)	Numerical

- Table.1
  These are the 5 variables that will be used in this report. We will utilize Gender, Payrise, Classification, and Income variables to analyze the gender pay inequality based on different perspectives. And the variable Working Experiences are one potential factor that could affect earnings.

Numerical Summeries

Gender	Number of repondents	Average Salary	Highest Salary	Lowest Salary	Number of pay rise	Average Working Experiences	High-salary earners
Male	724	74.875	250	14.712	537	10	45
Female	127	59.999	108	12.000	96	7	2

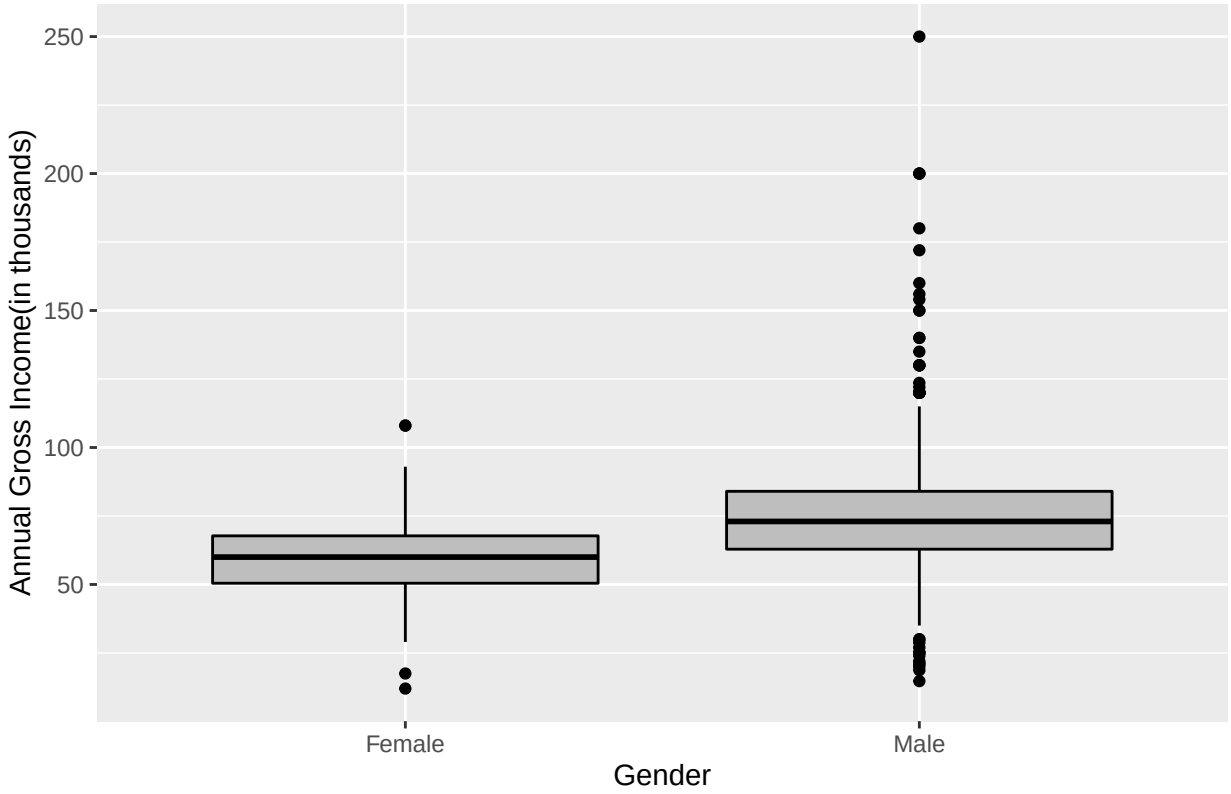
- Table.2
  According to the table, there are 724 males and 127 females in our cleaned dataset. Among these respondents, the male's average annual salary is much higher than that of the women in the same industry. Note that the salary is measured in thousands, so among these respondents, men's average annual salary is 15,000 euros higher than that of women. In other words, the average yearly income of men in the IT industry in the Eurozone is 20% higher than that of women in 2020.

- The highest income of male respondents reached 250,000 euros per year, which is two times higher than the top income of female participants. The lower bound of the annual salary in the survey for both sexes shows little difference, with the lowest-paid men respondents earning slightly more than the lowest-paid women. Also, 537 male IT workers receive a pay rise, and only 96 women experienced a pay rise in 2020.

- As stated, we define IT workers who earn more than 108,000 euros in 2020 as high-salary earners. According to the dataset, people with annual income above 108,000 euros are in the top 5 percent of income earners. Of the top 5 percent of income earners in the survey, 45 were men, and only 2 were women. Besides, the average working experience of men in the IT industry is longer than their female counterparts.

Graphical Summeries

Fig. 1 Annual Gross Income in the IT industry by gender

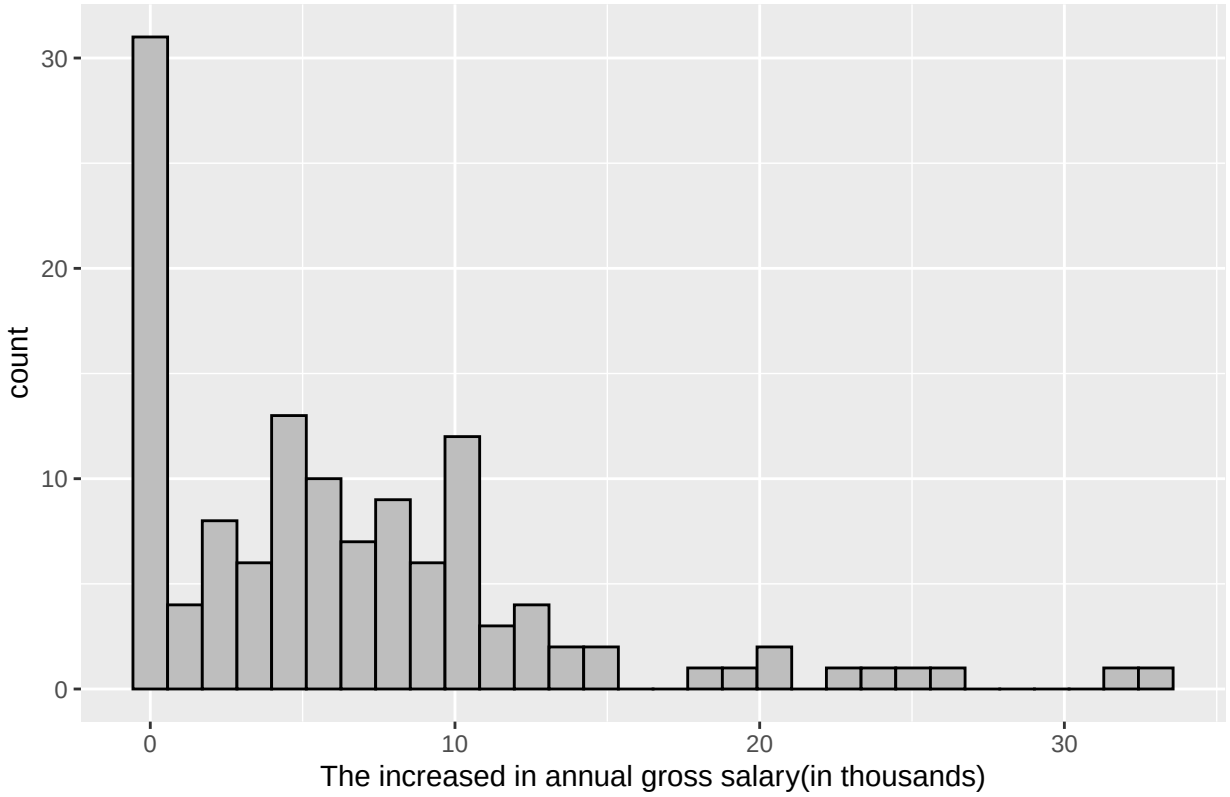


- Fig. 1 shows boxplots of the annual gross income in Europe's IT industry for women and men. From the graph, the average gross income for female IT workers is 60,000 euros per year, while the average annual gross income for male IT workers in Europe is about 75,000 euros. This is undoubtedly a non-negligible gap.

- The distribution of annual gross income for the female is a bit left-skewed, and the distribution of annual gross income for the male is right-skewed. Comparing men and women, the range of their annual income does not differ much; the income range of men will be slightly more extensive than that of women.

- For both genders, there are some outliers. Obviously, there are more male outliers than female outliers, and the male outliers are mainly those earning between 110,000 and 250,000 euros. Recall that we defined people with annual salaries above 108,000 as high-income earners. So, it is clear from the graph that the number of high-income males is greater than that of females.

Fig. 2 Distribution of annual pay rise for female IT workers in 2020



- Fig. 2 demonstrates the annual pay rise distribution for female IT workers who participated in the 2020 survey. It is a bimodal and right-skewed histogram peak at 0, indicating that over 30 women in the survey didn't receive a pay rise in 2020. Since it is right-skewed, the mean of all respondents' annual pay rise should be greater than the median. Therefore, the average gross pay rise for IT specialists in the survey should be around 8,000 euros per year.

- Since salary is a continuous and numerical variable and the highest point of the distribution is on the leftmost side of the plot, we predict that the annual pay rise of female IT workers follows an exponential distribution. Also, we can notice that the annual pay rise of female IT workers in the survey ranged from 0 to 15,000 euros, while most of the data ranged from 0 to 33,000 euros.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

This report will use six statistical methods to examine gender pay inequality in the IT industry in Europe. More specifically, we will focus on the average salary level of men and women, the salary increase of women, and the proportion of women in the high-income group to investigate the well-being of female IT employees in the EU. Finally, we will analyze the relationship between working experience and total annual income to analyze the reasons for women's lower average salary level compared to men. Detailed descriptions of variables can be found in Table.1 in the Data section.

Empirical Bootstrap Sampling and Confidence Intervals

Goals and parameters

- We've already compared the sample average of the annual gross salary for males and females in Table.2 of the Data section. To explore the true gender pay gap in Europe's IT field, we need to use confidence intervals to estimate the true mean of the annual gross salary for males and females in Europe.

Data processing and Empirical Bootstrap

- Since the original data included both male and female information, we first classified the original data. We created two new sample data based on gender, one containing all-male data from the original sample and the other collecting female data. Thus, we conduct the empirical bootstrap sampling for these two samples separately. Here, we care about the gender pay difference between males and females. So the statistics we calculated for each bootstrap sample will be the mean annual salary.

- First, we randomly sample from the two original samples with replacement and keep the sample size unchanged. The newly generated sample is called a Bootstrap sample. Then, we repeat the process by 1000 times to get 1000 bootstrap samples. Finally, we calculated the statistics and built the bootstrap distributions.

Confidence Interval

- A confidence interval provides a range of possible values for the true parameter, given that the data for the population is incomplete. The confidence interval is calculated based on the empirical bootstrap sampling distribution. Since we only have one sample from the population, empirical bootstrap sampling helps generate as many samples as we want from the original sample data. However, it is important to note that bootstrap cannot provide a better estimate than the original sample. Instead, it explores the variability of estimates from the original sample.

- We want to keep the confidence level as high as possible while trying not to make the range too wide, so a 95% confidence level will be a good choice. Also, the 95% confidence interval is commonly used by researchers. A 95% confidence interval indicates that we are 95% confident that the true average annual income for male(or female) IT workers in Europe is within the range.

Hypothesis test

Goals and parameters

- The hypothesis test aims to make appropriate inferences about the unknown true parameter based on the results. To study women's well-being in the workplace in more detail, we will examine the true average salary of female employees in the IT sector over the Eurozone. In the confidence interval section above, we found a range for the true average salary of female IT workers, based on which we wanted to get a more specific estimate.

Hypotheses

- Since the dataset is randomly collected over Europe and covers people with different characteristics, the data sample is representative of the population. Recall from Table.2, the sample average of female's annual salary is about 60,000 euros, which also falls within the confidence interval derived above.

- Therefore, the hypotheses are

$$H_0 : \mu_{female} = 60$$
$$H_A : \mu_{female} \neq 60$$

  The Null hypothesis H_0 states that the female's true average annual gross income is 60,000 euros. The alternative hypothesis H_a is that the female's true average yearly gross income is not 60,000 euros.

Z-test

- We will use a Z-test to calculate the p-value. We have no information about the true variance of the population distribution. However, we assume that all data points are independently and identically distributed (i.i.d.), and the size of our sample data is relatively large. By Central Limit Theorem, we know the mean converges to the Normal distribution. So we will use the Normal(0,1) distribution to generate the p-value for our hypothesis. The detailed calculations of the p-value can be found in section 1 of the Appendix.

The p-value

- We will generate a conclusion from the hypothesis test based on the p-value. A p-value indicates the likelihood of observing our data, given that the null hypothesis is true. Then, by comparing the p-value with a 5% significant level, we conclude whether the data reject the null hypothesis or not. If the p-value is greater than the significant level, the data rejects the null hypothesis test and vice versa.

Maximum Likelihood Estimator

Goals and parameters

- Maximum Likelihood Estimator(MLE) helps us estimate the distribution parameter by maximizing the likelihood function, so that the sample data will be most likely to be observed under the estimated distribution. Once we know what distribution a variable follows, we can obtain more information based on that distribution.

- We've looked at the average salary for men and women. Now, we want to study the salary increase of female employees to learn more about gender inequality in the IT industry. Therefore, I will explore the distribution of the annual salary rise for European females IT workers.

Assumptions

- Based on the histogram of the annual salary rise for European females in 2020 (Fig. 2), we predict that the annual pay rise (in thousands) for female IT workers in Europe in 2020 follows an exponential distribution. Therefore, the assumption for the derivation is that

$$X \sim Exp(\lambda)$$

  X is the annual pay rise(in thousands) for a random female IT worker in the 2020 survey, and X is identically and independently distributed.

Derivations of the MLE

- This section aims to find the MLE of λ, which is the parameter of the exponential distribution. Through derivations, the MLE of λ is $\frac{1}{\bar{X}}$. The detailed derivations regarding the MLE can be found in section 2 of the Appendix. Therefore

$$X \sim Exp(\frac{1}{\bar{x}})$$

  where $\bar{X}$ is the sample mean.

Goodness of Fit Test

Goals and parameters

- A Goodness of Fit(GoF) Test gives us a sense of how well the sample data align with our expectation. In the section above, we have studied the average salary of male and female IT employees in Europe and the salary increase of women in the European IT field. To have a more comprehensive analysis of pay inequality in male-dominated fields, we will study the number of women among the top earners in the IT field.

Data processing

- We previously defined high-wage earners as those with gross annual income greater than 108,000 euros. According to Table.2 in the Data section, there are 47 high-income IT professionals among those who participated in the survey, and only 2 of them are women. High-income earners tend to occupy high positions in companies. One report states that in 2019, female executives accounted for only 18% of all employed persons in the EU.[3] Since our data focus on the IT sector, which is commonly believed as male-dominated, the inequality maybe even worse. Therefore, we predict that women only make up 10% of the top earners in the EU's IT sector.

Hypotheses

$$H_0 : Y \sim BIN(n, 0.1)$$
$$H_A : Y \not\sim BIN(n, 0.1)$$

- The null hypothesis H_0 is that the number of females Y among n high-income IT specialists in the EU's IT sector follows a binomial distribution with a probability of 0.1.
  And the alternative hypothesis states that the number of females Y among n high-income IT specialists in the EU's IT sector does not follow a binomial distribution with a probability of 0.1.

Derivations of p-value

- We first calculate the ratio of the likelihood that the binomial distribution parameter is 0.1 to the likelihood that the parameter is the MLE. Then, based on the chi-squared test, a p-value is derived, and we can make conclusions about the hypothesis by comparing the p-value with a 5% significance level. The derivations process can be found in section 3 of the Appendix.

Bayesian Credible Interval

Goals and parameters

- By conducting the Goodness of Fit Test, we have an inference about the distribution of high-income females in the EU's IT sector. However, the GoF Test result can only provide an idea of the extent to which the sample data fit our estimated distribution; thus, the information we generate from the test is limited. Therefore, to improve the usefulness and reliability of our findings, we want to investigate the confidence intervals of the distribution parameters. Here, we hold the assumption that the number of females among high-income IT specialists in the EU's IT sector randomly follows a binomial distribution with a probability of p. And the parameter we want to study is p, which is the probability that an IT worker who earned more than 108,000 euros in 2020 is female.

Prior distribution

- The Bayesian Credible Interval helps to generate a range for the unknown true parameter within the domain of a posterior distribution. The Bayesian approach assumes that the unknown true parameter is a random variable instead of a fixed number. To derive the posterior distribution for p, we need to find the prior distribution of p.

- We have little information about the prior distribution of p, so we will choose the BETA(α, β) to be the prior since BETA distribution is the conjugate prior of Binomial distribution.[4] This combination can simplify our derivation for the posterior function. We set the BETA distribution parameter based on the 2019 report that only 18% of senior executives were females in Europe.[3] And we predict that the proportion of high-ranking women in the EU's IT sector will be smaller than 18%, which indicates more severe gender inequality in a male-dominated field. Therefore, we set α to be 5 and β to be 30. In this case, the BETA distribution will peak at a value smaller than 0.18, and there is no weight on p that is larger than 0.5.[5]

Posterior distribution

- After derivations, the posterior distribution of p is BETA(x+α, n-x+β), where n is the number of high-income earners, x is the number of females among high-income earners. Therefore, we will find the credible interval based on the BETA(x+α, n-x+β) distribution. All derivations regarding the posterior distribution can be found in section 4 of the Appendix.

- We will find the 95% credible interval. Like the confidence interval, we calculated the 2.5th and the 97.5th percentiles of the BETA(x+α, n-x+β) distribution to derive the range that p has a 95% probability of falling into.

- The posterior distribution of p is

$$p \sim BETA(x + \alpha, n - x + \beta)$$

Simple Linear Regression

Goals and parameters

- Through the simple linear regression model, we can explore the relationship between two numerical variables. More specifically, we learn how a dependent variable changes with the independent variable. We will analyze the relationship between the annual gross salary and the years of working experiences of European IT workers in 2020, in order to find out the possible reasons for gender pay inequality.

The model

The simple linear regression model can be expressed as

$$Y_i = \alpha + \beta x_i + U_i \quad for \quad i = 1, 2, \cdots, n \tag{1}$$

- x_i is the independent variable; in this report, x_i is the total years of working experience in the IT industry in Europe.
- Y_i is the dependent variable; in this report, Y_i is the annual gross salary(in thousands) earned by IT employees in 2020.
- U_i is the residual and has expectation 0, so the points are assumed equally distributed around the line. And U_i have constant variability at every level of x
- α is the intercept of the line; it indicates the annual gross salary(in thousands) of an IT worker in Europe in 2020 when he(or she) has no working experience in the IT field.
- β is the slope of the line; it indicates how the annual gross salary(in thousands) of an IT worker in Europe changes when the respondent's working experience changes by one unit.

The estimated model

To conclude, the estimated model will be

$$\hat{\text{Annual gross income}} = \hat{\alpha} + \hat{\beta} * \text{Years of working experiences} \tag{2}$$

- By applying the sample data, we can calculate the intercept $\hat{\alpha}$, and the slope $\hat{\beta}$ and build the estimated linear model 2.

Results

This section will analyze the results of the six methods separately. All of the findings, whether the average salary level, annual pay rise for females, or the proportion of women in the upper-income group, point to the existence of gender inequality in the IT industry in Europe.

Confidence Intervals

Gender	Parameter	Confidence Level	Lower bound	Upper bound
Male	Average annual gross income(in thousands)	95%	73.439	76.209
Female	Average annual gross income(in thousands)	95%	58.022	62.268

- Table.3
  For males, the 95% confidence interval of the mean annual gross income is [73.439, 76.209]. It means we are 95% confident that the true average annual gross income of male IT specialists in the EU region is between 73,439 euros and 76,209 euros.

- For females, the 95% confidence interval of the mean annual gross income is [58.022, 62.268]. It indicates that we are 95% confident that the true average annual gross income of female IT specialists in the EU region is between 58,022 euros and 62,268 euros.

- Both intervals seem reasonable since they are positive and are not extreme values. Also, the range for the true average of annual gross income for males is much higher than the range for females, indicating that the true average annual gross income for a male would be larger than that of a female in the IT field in Europe. The results are consistent with our hypotheses that gender pay inequality is severe in Europe's IT industry.

Hypothesis Test

Hypotheses

$$H_0 : \mu_{female} = 60$$
$$H_A : \mu_{female} \neq 60$$

The Null hypothesis states that the true average annual gross income for the female is 60,000 euros. The alternative hypothesis is that the true average annual gross income for the female is not 60,000 euros.

The p-value and conclusions

- The p-value for this hypothesis test is 0.99 (Appendix.1).
  So the probability of observing a test result that is at least as extreme as our sample is 0.99, given the assumption that the population mean of female's annual gross income is 60,000 in Europe in 2020.

- This p-value is much larger than the 5% significance level, so there is no evidence against the null hypothesis. In other words, there is very strong evidence that the true average annual gross income for female IT workers is 60,000 euros. This result seems reasonable as it is consistent with our sample average and is falls within the confidence range we derived above. Therefore, we can confirm that our dataset is a representative sample of the population. Also, the lower bound of the confidence range for male's true average salary is higher than 60; this reconfirms our prediction that female's true average salary level is lower than male's

Maximum Likelihood Estimator

- After derivations, the MLE of $\hat{\lambda}$ is $\frac{1}{\bar{X}}$, which is the inverse of the sample mean (Appendix.2). The sample average of the rise in salary for female IT worker in Europe is 6.622 thousand euros. Therefore, the MLE of $\hat{\lambda}$ is 0.151 based on the sample data.

- Assume that X is the annual pay rise(in thousands) received by a random female IT worker in the 2020 survey, and X is identically and independently distributed. Then,

$$X \sim Exp(0.151)$$

- To conclude, the annual pay rise(in thousands) received by female IT workers in 2020 in our dataset is most likely to follow the exponential distribution with the statistics $\hat{\lambda}$ equals to 0.151. With the exact distribution function, we are now able to learn more information from the sample data. For example, we can calculate the probability of a pay rise below a specific cutoff for female IT workers in the 2020 survey.

Goodness of Fit Test

Hypotheses

- Recall that the hypotheses are

$$H_0 : Y \sim BIN(n, 0.1)$$
$$H_A : Y \not\sim BIN(n, 0.1)$$

- Y is the number of females among high-income IT specialists in the EU's IT sector n is the total number of high-income IT specialists in the EU's IT sector.

- For the sample data, 4% of the high-income IT workers are females in the EU. From our derivations, the MLE of p is $\frac{x}{n}$, which represents the proportion of females over the high-income earns. Therefore, 0.04 is the MLE for parameter p of the binomial distribution. Together with the null hypothesis, we conduct the chi-squared test.

The p-value and conclusions

- The p-value is 0.14.
  The p-value is larger than the 5% significance level, meaning that we do not reject the null hypothesis. Therefore, we have strong evidence that the number of females among high-income IT specialists in the EU's IT sector follows a binomial distribution with a probability of 0.1. This means that women and men are not equally likely to be high earners. In contrast, women have a much lower probability of earning a high salary. From this perspective, women are undervalued, and they received fewer opportunities than men.

Bayesian Credible Interval

Recall that we've assumed that the number of females among high-income IT specialists in the EU's IT sector randomly follows a binomial distribution with a probability of p. And the prior distribution of p is BETA(5,30)

Gender	Parameter	Credible Interval	Lower bound	Upper bound
Female	p	95%	0.035	0.154

- Table.4
  From derivations (Appendix.4), the posterior distribution of p is BETA(x+α, n-x+β). Since n is 47, α is 5, and β is 30. The posterior distribution of p is BETA(x+5, 77-x). Therefore, the 95% credible interval of p is [0.035, 0.154]. It means that there is a 95% probability that the true parameter p will fall between 0.035 and 0.154, given that the prior distribution is BETA(5, 30).

- Based on the credible interval, the true parameter p of the binomial distribution should likely be small, indicating that the probability of a random high-income IT employee being female is very low. In other words, women have a much lower probability of earning a high income compared to men. That's why the proportion of women in the high-income group is very small. This is another manifestation of gender inequality.

- The interval is consistent with the results from the Goodness of Fit test section, which states that there is strong evidence that p is 0.1. Also, the sample proportion falls within the interval, where 4% of females earn a high salary. However, it's important to note that the choice of prior will affect the posterior distribution, thus affecting the credible interval. So this analysis has some limitations, which we will discuss in the Conclusion section.

Simple Linear Regression

Numerical output of the model

coefficients of the estimated model	
Intercept($\hat{\alpha}$)	57.165
slope($\hat{\beta}$)	1.667

$$\hat{\text{Annual gross income}} = 57.165 + 1.667 * \text{Years of working experiences} \quad (3)$$

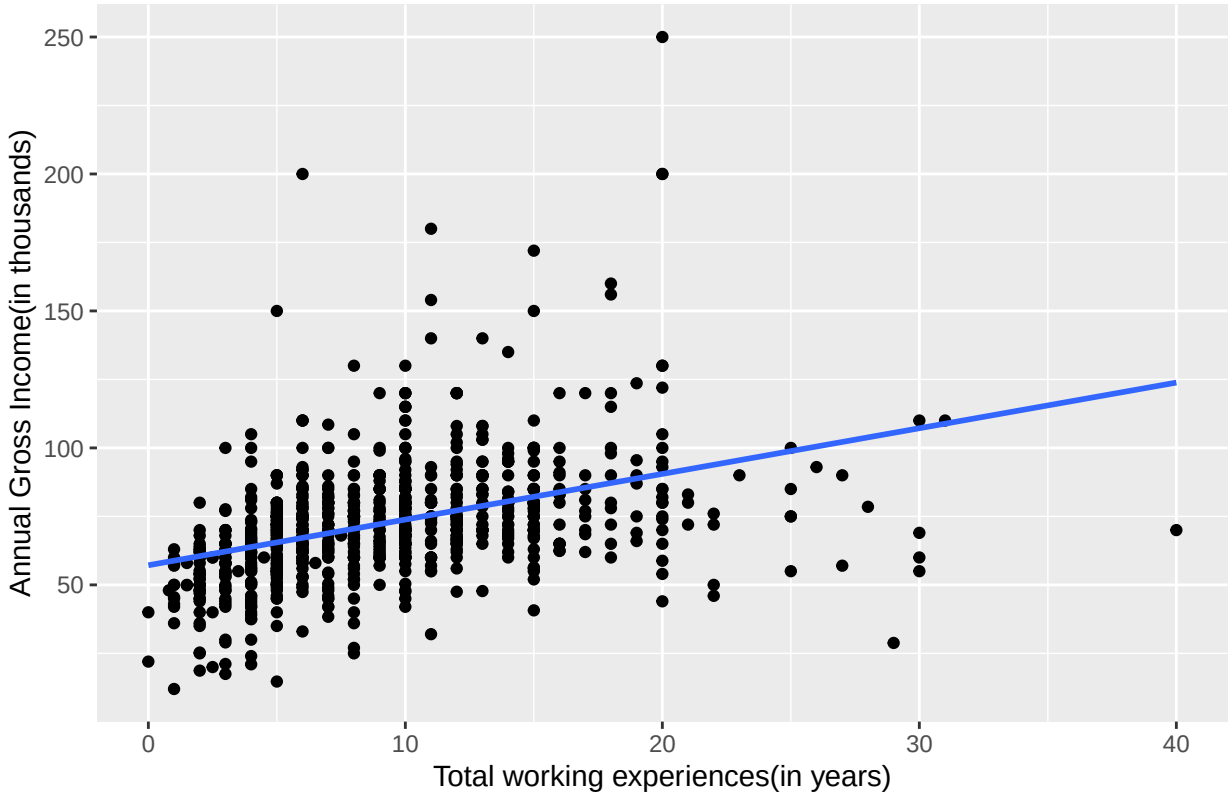
- By substituting the coefficients into model 2 in the Method section, we got the estimated model 3.

- 57.165 is the y-intercept of the estimated model, which means that the expected annual gross income is 57,165 euros when the employee has no working experience in the IT field. In order words, newcomers to the IT sector are expected to earn an annual salary of 57,165 euros.

- 1.667 is the slope of the estimated model, which means that For each increase in the years of working experiences in the IT field, the expected annual gross income will rise by 1,667 euros. Since the slope is positive, there is a positive linear relationship between the working experiences and the annual gross income.

- These results seem reasonable as both coefficients are positive and are not extreme numbers. Employees with extensive work experience tend to be more efficient and capable, so their corresponding salary levels are likely to be higher. Also, the longer a person works in the IT field, the more likely they will be promoted, which means that the relationship between work experience and salary is positive.

Fig. 4 Annual Gross Income by working experiences for both genders



- Fig. 4 is a scatterplot that shows the relationship between the years of working experiences and the annual gross income for the IT workers in Europe. The blue line indicates the estimated linear regression model 3. There is a strong and positive linear relationship between the working experiences and the annual gross income. In other words, total annual earnings are expected to increase with work experience.

- The linear regression model is appropriate for analyzing our data since the assumptions of the model hold. From Fig. 4, although there are several outliers, most points uniformly distribute around both sides of the model. Moreover, the variability of most points to the line is consistent over the data range.

- The positive relationship between the working experiences and the annual gross income could explain gender pay inequality. In the numerical summaries of the Data section, we concluded that, on average, men's working experiences are three years more than that of women counterparts. Women can be fired during their working careers due to a variety of discrimination, the most common of which is maternity discrimination. Many companies are reluctant to keep pregnant women because of their limited mobility during pregnancy and the need to take a period of maternity leave. In the UK, 54,000 women lose their jobs each year due to pregnancy[3]. This results in women having less work experience than men, which in turn makes women's average wages less than men's.

Conclusions

Summary of the Hypotheses, Methods, and Results

- Gender pay inequality is a widespread problem. Based on the gender pay gap in Germany in 2019, we predict that inequality in the EU IT sector will still be severe in 2020. Therefore, we analyzed women's well-being in male-dominated industries from different aspects and drew society's attention to the inequality phenomenon. We used data from the 2020 Salary Survey of IT professionals in the EU as a sample for the study.

- Six statistical methods were used for the analysis. Through confidence intervals, we compared the ranges of the true average salary levels of men and women. Moreover, based on the confidence interval, we further predict the true mean salary of women and evaluate our prediction by applying the hypothesis test. By deriving the MLE, we make inferences about the distribution of salary increase for females. Furthermore, we analyze the distribution of the proportion of women among high earners by using the GoF test and the Bayesian credible interval. Finally, we analyzed the relationship between work experiences and salary using the linear regression model.

Key results and big pictures

- The results show that gender pay inequality was indeed serious in the IT field in 2020. We found that the 95% confidence intervals of the true average salary of male employees in Europe's IT industry are much higher than those of women, which means that we are 95% certain that men's true average annual salary is higher than that of women. In the sample, the annual average wage of females is 60,000 euros. Based on the sample average and the confidence interval, we have strong evidence that female IT employees' true yearly average wage in the EU is 60,000 euros in 2020.

- We also find that the annual salary increase for female IT specialists follows an exponential distribution with a rate equal to 0.151. With this distribution, we can obtain plenty of information. For example, we can calculate the probability that a woman's annual salary increase is less than 20,000 euros, or the 25th percentile of the annual salary increase. Furthermore, we study the distribution of highly paid women. We showed strong evidence that the number of high-salary women in the IT industry follows a binomial distribution with a probability of 0.1. We next predict the true value of parameter p in the binomial. Using the Bayesian credible interval, we found a 95% probability that the true parameter p is between 0.035 and 0.154, which is consistent with GoF test results.

- Finally, we found that working experience and salary are positively related. For new IT employees, they are expected to earn an annual salary of 57,831 euros. People with more work experience tend to be more capable and efficient in handling things, thus making a higher salary. Based on this relationship, we speculate that the reason for the lower salary levels of women compared to men may be the lack of work experience of women. The sample points out that the average work experience of female IT employees in the EU is three years less than that of men. This may be because women are faced with childbirth or household care as they get older. Since pregnant women tend to have difficulty in mobility, they are likely to get fired, resulting in a lack of work experience. It also means that women have fewer opportunities for advancement than men, which may be one of the reasons why women make up a smaller percentage of the high-earning population. Therefore, the average salary level of women will be lower than that of men.

Recommendations

- Based on the analysis results, women's well-being at work needs to be improved compared to men's. The company's administrators need to measure the employee's salaries based on their performance, not gender. So the company can hide the personal information such as gender and age of the employees when evaluating them and only focus on their work accomplishment. At the same time, governments can enhance public education by utilizing social media and schools to draw society's attention to gender inequality. Also, the government can make regulations to strengthen the supervision of companies, ensuring that they treat all employees equally and punish discrimination. Moreover, companies need to protect pregnant women, for example, by giving more paid time off or providing opportunities for pregnant women to work from home. Besides, companies can provide some training opportunities for employees who return to work after maternity leave so that they can quickly readjust to their jobs.

Weaknesses

- There are some limitations to our analysis. When examining the number of women in the high-income group, we set the criterion for high income at 108,000 euros or more per year. This criterion is subjective, and since we do not have access to all employees in the IT sector, we can only set the standard based on the top 5% of the sample data. So the true distribution may be different from the analysis results. However, since the sample is representative, this analysis can still provide helpful information.

- In the calculation of MLE, we assume that the annual salary increase of female IT workers follows the exponential distribution by the sample data histogram. However, because the sample size is still too small compared to the population, the sample's distribution cannot capture all information. Therefore, the population distribution may not be exponential.

- Moreover, the result of the credible interval may be slightly different from the true range. We do not have enough information when choosing the prior distribution, and the distribution of the posterior is affected by the prior. We found that the resulting credible interval will be slightly different when we changed the priors' parameter. However, the degree of change is not significant, and the lower bound and upper bound of the interval are still around 0.035 and 0.154. In the future, we can obtain more information about the priors through other analyses, or use more sophisticated methods to improve the results' reliability.

Next Steps

- Many factors contribute to gender inequality, and this report only focuses on work experiences. In the future, we can study other potential factors, such as age, number of leaves of absence, and job performance. In this way, we can analyze gender inequality more comprehensively and take better measures to solve it. Also, pay inequality exists not only between men and women. Among women, pay levels may vary according to other attributes, such as race and marital status. Therefore, we can further examine how racial pay discrimination or marital status affects the pay an employee receives among women in the future.

Discussion

- To sum up, gender pay inequality is still severe in Europe's IT industry. Women are not only paid less than men, but they also have fewer opportunities to earn higher salaries than men. And discrimination against pregnant women can further widen the pay gap and make female employees worse off. Therefore, society needs to draw attention to gender inequality. The annual gross salary and working experiences are positively related. Governments can adopt policies to improve women's work experience reduce gender discrimination, thus reducing the gender pay gap. Also, women need to defend their rights actively.

Bibliography

1. Gender pay gap. (2021, April 02). Retrieved April 04, 2021, from https://en.wikipedia.org/wiki/Gender_pay_gap

2. Welle, D. (2020, August 12). Germany's gender pay gap shrinks, but still higher than EU Average. Retrieved April 04, 2021, from https://www.dw.com/en/germanys-gender-pay-gap-shrinks-but-still-higher-than-eu-average/a-55860947

3. Women business leaders: Global statistics. (n.d.). Retrieved April 15, 2021, from https://www.catalyst.org/research/women-in-management/

4. Kim, A. (2020, January 16). Conjugate prior Explained. Retrieved April 15, 2021, from https://towardsdatascience.com/conjugate-prior-explained-75957dc80bfb

5. Beta distribution Applet/Calculator. (n.d.). Retrieved April 15, 2021, from https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html

6. All analysis for this report was programmed using `R version 4.0.2`.

7. The package I used was tidyverse. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686/

8. I also use the packages of patchwork. Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. https://patchwork.data- imaginist.com, https://github.com/thomasp85/patchwork.

9. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

10. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

11. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

Appendix

Section 1(Hypothesis test)

Recall that the hypotheses are:

$H_0 : \mu_{\text{female}} = 60$
$H_A : \mu_{\text{female}} \neq 60$

So the test statistics t will follow a Normal(0,1) distribution,

$t = \frac{\bar{x}-60}{\frac{s}{\sqrt{n}}} \sim Normal(0, 1)$

Then the p-value will be,

$$\begin{aligned} \text{p-value} &= P(|Z| > |t|) \\ &= 2 \cdot (1 - \Phi(|t|)) \\ &= 0.9995 \end{aligned}$$

Section 2 (MLE)

We want to explore the distribution of the annual salary rise for European females IT workers. Maximum Likelihood Estimator(MLE) helps us estimate the distribution parameter by maximizing the likelihood function, so that the sample data will be most likely to be observed under the estimated distribution.

Assume that X is the annual pay rise(in thousands) received by a random female IT worker in the 2020 survey and X is identically and independently distributed. Then,

$$X \sim Exp(\lambda)$$

The exponential function is expressed as $f(x) = \lambda e^{-\lambda x}$.
Therefore, the lilklihood function will be

$$\begin{aligned} L(\lambda) &= f(x_1) \cdots f(x_n) \\ &= \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

To find the maximum estimator of λ, we set the derivative of the loglikelihood function equals 0.

$$\begin{aligned} l(\lambda) &= \ln L(x) \\ &= n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

$\frac{dl}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$

$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

To verify that the MLE we derived is the maximum, a second derivative test need to be conducted.

$$\begin{aligned} \frac{d^2 l}{d\lambda^2} &= -\frac{n}{\lambda^2} \\ \left.\frac{d^2 l}{d\lambda^2}\right|_{\hat{\lambda}=\frac{1}{\bar{x}}} &= -\frac{n}{\left(\frac{1}{\bar{x}}\right)^2} \\ &= -n\bar{x}^2 \leq 0 \end{aligned}$$

According to the second derivative rule, x will be a local maximum of the function f(x) is f"(x) < 0. Since both sample size(n) and the square of sample mean are greater than zero, $-n\bar{x}^2$ must be smaller than 0, indicating that $\frac{1}{\bar{X}}$ is indeed the maximum likelihood estimator for λ.

In conclusion, the MLE of λ is

$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

Section 3 (Goodness of Fit Test)

Assume that Y is the number of high-income females over high-income earners in the 2020 survey and Y is identically and independently distributed. The probability that there are y females within the high-income group is p. Then,

$$Y \sim BIN(n, p)$$

Therefore, the lilelihood function for p is

$$L(p) = p(y) = \begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y}$$

$$l(p) = \ln L(p)$$

$$= \ln \begin{pmatrix} n \\ y \end{pmatrix} + y \ln p + (n-y) \ln(1-p)$$

To find the maximum estimator of p, we set the derivative of the likelihood function equals 0.

$\frac{dl}{dp} = \frac{y}{p} + \frac{n-y}{1-p}(-1) = 0$

$\frac{y}{p} - \frac{n-y}{1-p} = 0$

$y(1-p) - (n-y)p = 0$

$\hat{p} = \frac{y}{n}$

To verify that the MLE we derived is the maximum, a second derivative test need to be conducted.

$\frac{d^2 l}{dp^2} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}(-1)(-1)$

$$\left.\frac{d^2 l}{dp^2}\right|_{\hat{p}=\frac{y}{n}} = -\frac{y}{\left(\frac{y}{n}\right)^2} - \frac{n-y}{\left(1-\frac{y}{n}\right)^2}(-1)(-1)$$

$$= -\frac{n^2}{y} - \frac{n-y}{\left(1-\frac{y}{n}\right)^2} \leq 0$$

According to the second derivative rule, y will be a local maximum of the function f(y) is f"(y) < 0. Since y and n are positive numbers, and y is not greater than n. The result of the second derivative test must be smaller than 0, indicating that $\frac{y}{n}$ is indeed the maximum likelihood estimator of p.

Therefore, the likelihood ratio will be

$$\begin{aligned}\text{Likelihood ratio } &= \frac{L(0.1)}{L\left(\frac{y}{n}\right)} \\ &= \frac{\left(\begin{array}{c} n \\ y \end{array}\right) 0.1^y 0.9^{n-y}}{\left(\begin{array}{c} n \\ y \end{array}\right) \left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}} \\ &= \frac{0.1^y 0.9^{n-y}}{\left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}}\end{aligned}$$

Section 4 (Bayesian Credible Interval)

We chose the BETA(α, β) to be the prior since BETA distribution is the conjugate prior of Binomial distribution. Assume that X and p are identically and independently distributed

$X \sim BIN(n, p)$
$p \sim BETA(\alpha, \beta)$

Then the distribution of p can be expressed as

$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

The likelihood function for p is

$$L(p) = p(x) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x (1-p)^{n-x}$$

Next, we calculate the product of the likelihhod function of p and the prior distribution of p

$$L(p) \cdot f(p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x (1-p)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The posterior function of p is proportional to the product above
post $\propto L(p) \cdot f(p)$
$\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$
$\sim BETA(x+\alpha, n-x+\beta)$

Therefore, we conclude that if we choose BETA(α, β) as the prior distribution and p follows a BIN(n,p) distribution. The posterior distribution of p will be BETA(x+α, n-x+β).