[Introduction 2](#_Toc86582244)

[Expected value and proportion estimation 2](#_Toc86582245)

[Hyperthesis testing 3](#_Toc86582246)

[Corelation and covariance 5](#_Toc86582247)

[Conclusion : 8](#_Toc86582248)

## Introduction

The excel file is a dataset for employee s’infomation from a huge-size company . There are 473 observers , with eight variables , including three qualitative variables :gender , jobcat , minority and five quantitative :educ , salary , jobtime , prevexp , age .

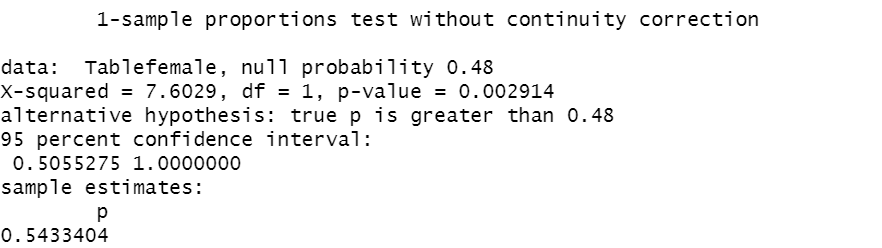
## Expected value and proportion estimation

To test some estimations made beforehand, we need to use mean estimation on samples we have , to expect to derive some meaningful conclusion for population . To test the expected value of two quantitative variables : salary、 prevexp , in the case of 90% confidence interval , we start from deriving margin of error between the mean of that in sample and mean of that in the population, that is . qt(0.95,473-1)\*sd(employee$prevexp)/sqrt(473) . Here 0.95 = 1- alpha/2 (for only one side ) . We know alpha by 1-confidence level , that is 10% . Eventually , in 90% CI , the salary mean in population should be in the interval between 33123.1 dollar to 35713.79 dollar , while the prevexp has a interval between 88.01months to 103.88 months . It implies the most workers in te company have a decent wage and above 50% of workers have long enough pre working period .

To have proportion estimation test for minority works in population by the sample data, we need to find the margin of error for sample proportion compared to population proportion by calculating : z\*sqrt(Propofminority\*(1-Propofminority)/n). We get quantity of minority “yes”/ “no” by table function , then divided by 473 . We get the proportion estimate of minority in population is 17.56% – 26.42% in IC 98% circumstance.

## Hyperthesis testing

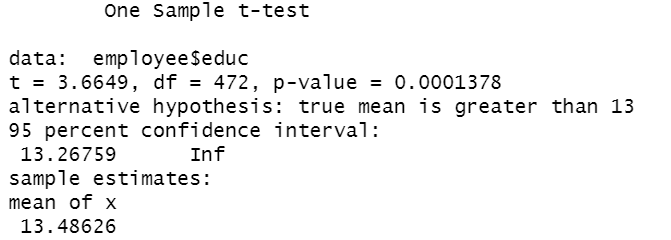
Table 1 proportion test



*(source: author ’s R programming)*

In Table 1 , the statement of “proportion of female workers should be not higher than 48%” , should be tested by proportion test in R . We choose an alternative statement as “greater ” , then we have a P value at 0.29 % . As we always assume alpha value is around 5% . As P value is much less than alpha , in the confidence level of 95 , we reject null hyperthesis , and accept the alternative hyperthesis of proportion is greater than 48% , the estimation for sample is 54.33% . The proportion for population is more likely higher than 48% as well according to proportion test .

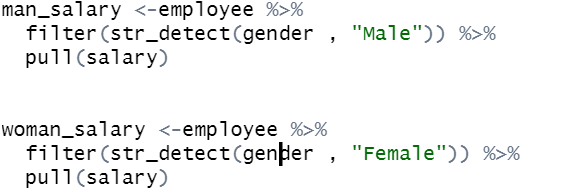
Table one-sample t-test



*(source: author ’s R programming)*

In Table 2 ， the next statement we want to test is that “ average years of education is small or equal to 13 years ” . Our alternative is “average years of education is small or equal to 13 years” . We test if mean would be over 13 . Thus we need to use one sample t-test . As we get p-value of 0.1378 % (much less than 5% ) , we reject null hyperthesis . the estimation for sample is 54.33% . The proportion for population is more likely higher than 48% as well accroding to proportion test .

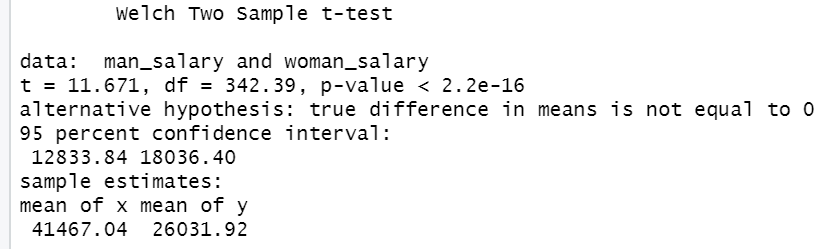
Table 3 Group by salary according to gender



*(source: author ’s R programming)*

In Table 3 , for the statement , “the average salary of men and women is equal “.We test difference between 2 means of man salary and woman salary . As the first step , we group salaries by gender . Taking use of filter function and str-detect . We have divided salaries into two groups .

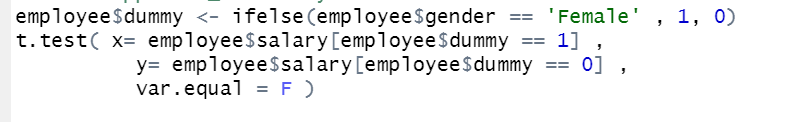
Table 4 Two sample t-test



*(source: author ’s R programming)*

In Table 4 , after t-test between man ’ s salary and woman ’ s salary , according to P-value ,we need to REJECT the null hyperthesis : “true difference in means is eqaul to 0 “. Under circustance of IC 95% , it is coincident that sample estimates shows a very huge gap between mean of man ’s salary and mean of woman ’s salary .

Table 5 test mean of man 's salary and mean of woman' s salary

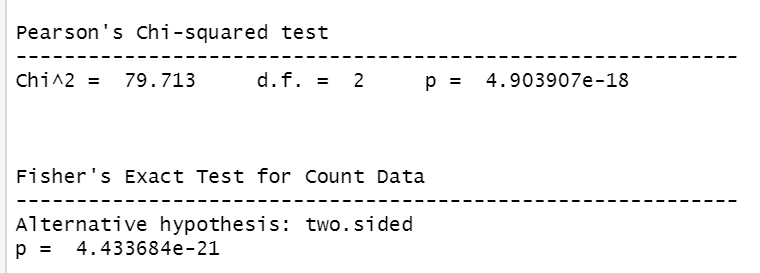


*(source: author ’s R programming)*

Apart from this way , we could set dummy to derive salary for different group as well . And surely it could turn into the same result as group-by does .

## Corelation and covariance

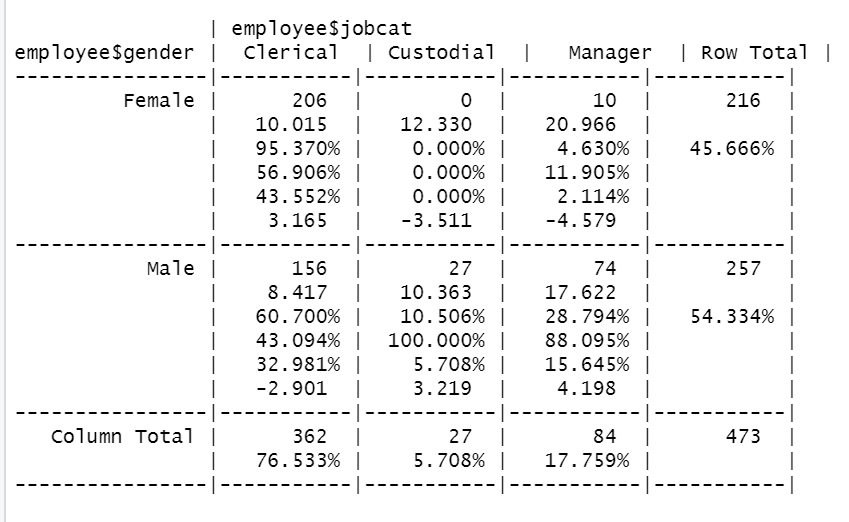
Table 6 crosstable for jobcat and gender



*(source: author ’s R programming)*

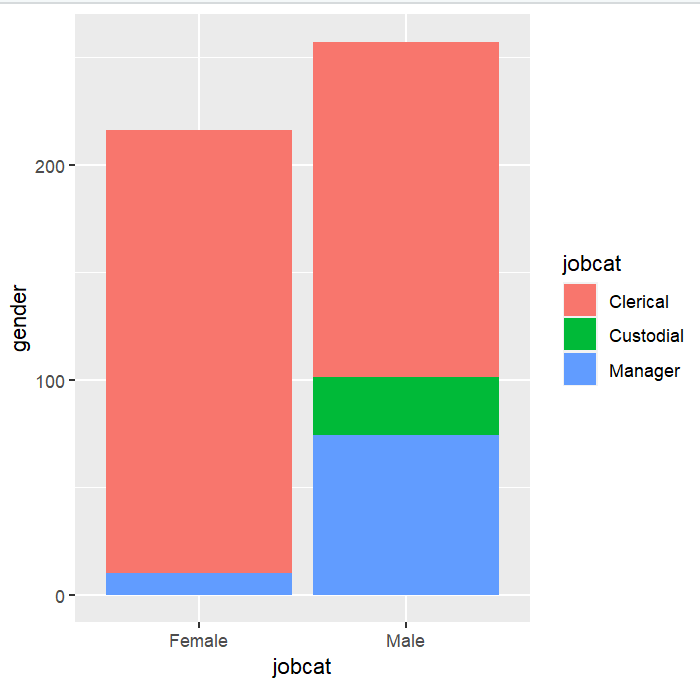
In Table 6 , we use crosstab as a assistant in looking on relations between two qualitative variables . Cross Tabulation could be for qualitative factor independence. It is questionable that weather jobcat and gender would have any weight on each other . For this question , We create crosstable , it have Chi squared test and Fisher ’s Exact test . Both of them show small P-value , much smaller than 5 % . Thus we agree that gender and jobcat has influences on each other .

Table 7 crosstable for gender and jobcat



More concrete relation shows : For female , it can see a greater ratio of clerical than man , and much more small ratio on the level of manager . This could be one of the reason why mean of salary for genders can be seen a huge gap .

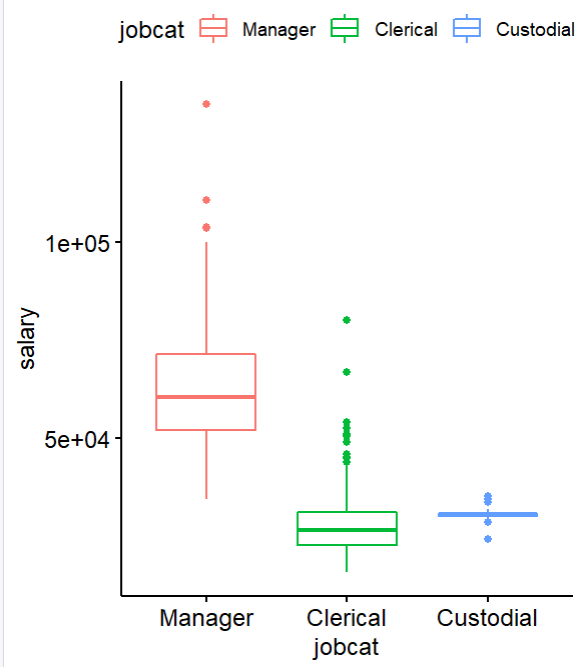
Table 8 ggplot for gender and jobacat



*(source: author ’s R programming)*

It is clear that in Table 8, man have a greater proportion on manager-level while female has more proportion in clerical-level .

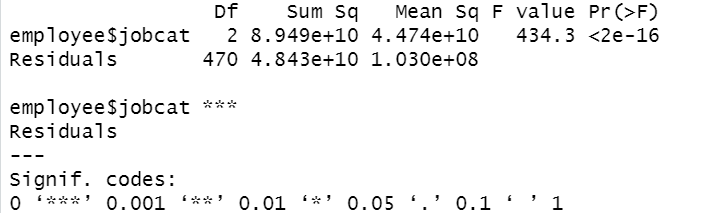
Table 9 ggboxplot for jobacat and salary



*(source: author ’s R programming)*

By boxggplot , we could have a comparison between quantitative variables as well as detect outliers for each group . It is expected that salary would be floated along with different positions and boxplot demonstrates this relation for us . There are many outliers on salary for each level of position . We credit it to normal job market behavior . Manager has the biggest range of salary , but custodial has the smallest range.

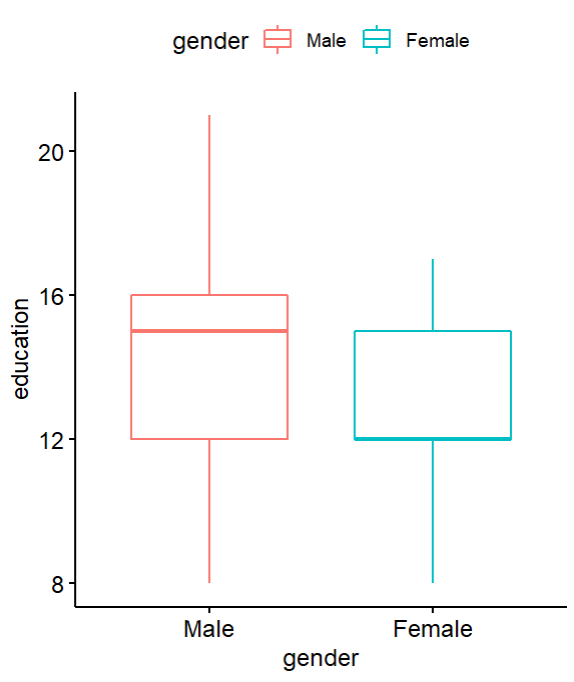
Table 10 Anova for jobacat and salary



*(source: author ’s R programming)*

For P value is samller than 5 % , F value is greater than 5 . we see a strong relation for salary and jobcat .

Table 11 ggboxplot for gender and educ



*(source: author ’s R programming)*

From this boxplot , we see males have better education-condition than woman . It is mostly demostrated by higher education over 12 years . Man could wins obviously by right skewed on higher education

It is valuable to investigate relations among all quatitative variables : salary , educ , jobtime, age and prevexp , which can act a good summary for different factors .

Table 12 heatmap for all quatative variables



*(source: author ’s R programming)*

In Table12 , By plotting heat map, the variation in color may be by red , purple or in between , giving obvious hue-visual cues to the reader about how two variables is positively or negatively related to each other . Seen by the order of colons. prevexy is being effected by age , having same moving directions with age growing , slightly negatively related with education . It implies elder workers in the company ,is more likely to have more prevexp , but possibly are less educated . Jobtime is a relatively independent variable , not having very relation with other variables . Education could probably have a positive influence on salary , accordingly , salary is also showing a exactly same positive relation with education .

## Conclusion :

* Female and male have an relatively equally-proportion count with male employees are slightly higher .
* Mean of salary for different gender could have a huge difference under IC 95 % . It is related to higher level of positions occupied by males . It can be explained by education period differences between different genders . Men overall have better education than women , so that they could overtake more higher positions . No evidence show obvious gender discrimination from dataset .
* Education can be slightly related with age . Youngsters are more tended to have a longer school education while senior people had more work experience due to shorter school education .
* Jobtime is relatively unique, independent variable among this dataset , for having weak correlation and covariance with other variables ,