Exploring the wage comparison by gender and age in Seattle

STAT 5303

instructor: Kazeem Adepoju

Group member: Chenghan Yue, Wenzhi Dong, Ruohan Wang

Instruction

Before the experiment, we want to introduce why we choose this topic. We all know that gender is one of the biggest issues in global. Too many people are working for balanced the difference between male and female. And, for preparing this project, we searched a lot of articles. Then, we found an interesting article which called “Moms Work 5 Extra Months to Earn as Much as Dads, Research Says”. The title tells us a truth about real life, the wage of female always lower than male. But we cannot just trust author’s words, we need to analysis by ourselves. And, unifies own actual situation, we are junior and senior student, which means we will graduate in next year or next semester. Thus, we consider about a lot of things around jobs. Wage is one of the most important elements, then, we search the data from Seattle government website which offer us a set of data about wage in different ages and different gender. And, because, in our group we male group member and female group member. We are interested in the different between the wage of female and the wage of male. We will use the randomly completely block design to help us finish this experiment.

Aim and Objective

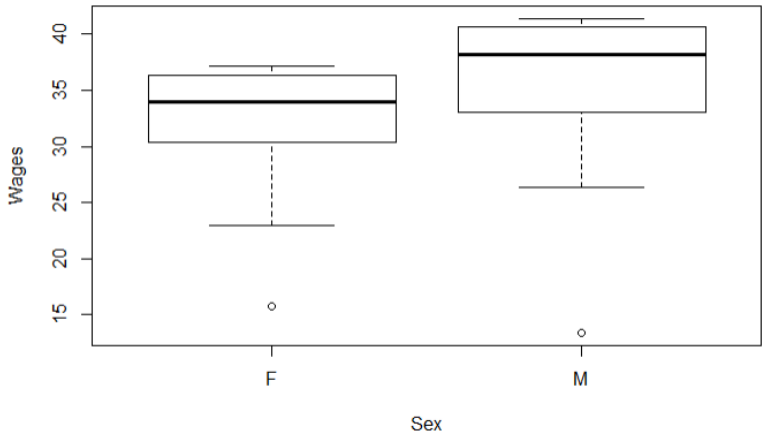
According to our research question is “does male and female get equal wages in Seattle Area”. And, we also study about the wages of different ages. We are willing to know how salaries varied by ages. Then, we expect to see the data in Seattle area, female and

male get comparatively equal wages. We also predict that as people get older, people get more salaries.

Exploring the data

Our data was from government website of Seattle (appendix 1), it has three parts which are ages, wages, and sex. For the part of age, we will concentrate on 12 different age which are “under 20”, “20-25”, “25-30”, “30-35”, “35-40”, “40-45”, “45-50”, “50-55”, “55-60”, “60-65”, “65-70” and “above 70”. For these ages, we also will consider about different gender people’s wage. In Seattle area, investigators randomly select a group of people at each specific age and in dived them into two groups: female and female, calculating their average hourly wages separately.

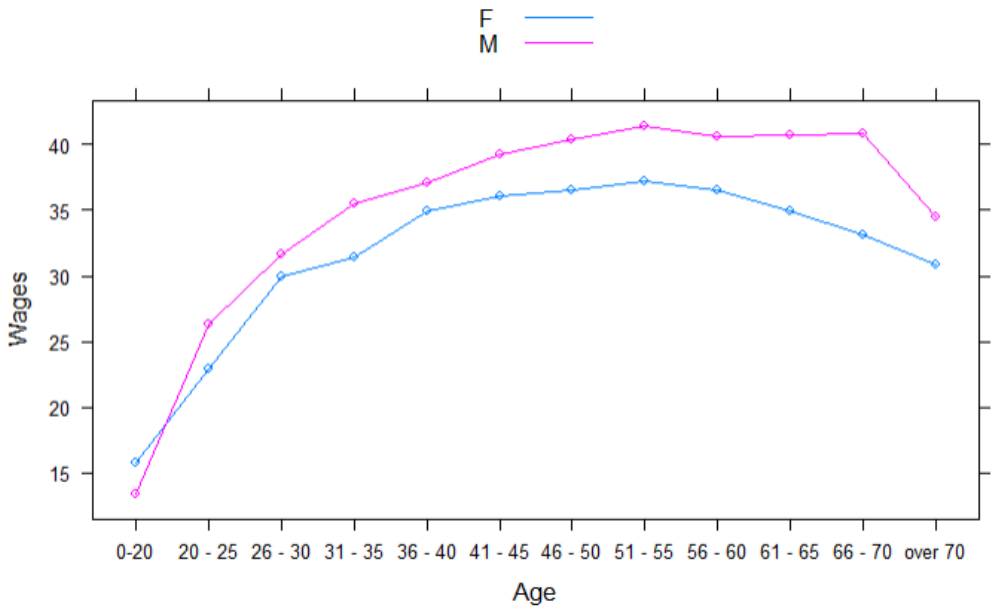
Then, I used the Rstudio to draw a boxplot which looks like following picture.



From the boxplot, the range of female wages is from about 23-37 and male range is from about 26-41. And, it is obviously to find that the male’s wage is higher than female’s wage. And, the box of male’s wage is bigger than the box of female’s wage.

After that I still used Rstuio to produce a linear plot which can help me to find the correlation between wage and age. The pink line represents the correlation between

male's wage and age. And, the blue line represents the correlation between female's wage and age. According to the figure 2, the hourly wages of female and male has a similarly tendency. The lowest hourly wages were around 15 dollars under age 20. The wages increased fast at the beginning and the slope got flatter gradually to the peak at age 51-55 and then decreased. Female only got higher wages than male at the age under 20. After that female always had lower wages than male.



Both of these two graphs can help us get a conclusion, the male's wage is higher than female's wage generally. Furthermore, I need to test the significance of the effect of sex and age on wages.

Methodology

In this data, each treatment only replicated once in each block. Namely, at each age level, there was only one average value.

The data was chosen randomly, but randomization was restricted that it only took place

within each block of ages and sex. In other words, the people in different sex at each specific age were random select separately.

Hence, the design we chosen was single replicated Randomized Complete Block Design (RCBD). There were two levels for treatment (Sex): female and male. Twelve groups of age were the block.

The model was $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

i = 1, 2 , j = 1,2,...,12

α_i the fix effect of sex, constraint: $\sum_{i=1}^{2} \alpha_i = 0$

β_j the fix effect of age, constraint: $\sum_{j=1}^{12} \beta_j = 0$

$\varepsilon_{ij} \sim N(0, \sigma^2)$

Analysis

We want to test the effects of sex and age on wages. The two null hypothesis were:

- H0: $\alpha_i = 0$ vs. Ha: $\alpha_i \neq 0$
- H0: $\beta_j = 0$ vs. Ha: $\beta_j \neq 0$

We set the significant level α of this study equal to 0.05.

In order to do the Anova, the assumptions of residual normality, homoscedasticity and independent individuals had to be met.

Regarding to qqplot (appendix 2), the points did not look straight. Two tails were out

of lines. However, Shapiro test delivered p-value 0.08 (appendix 3). We failed to reject the null hypothesis of normality at significant level of 0.05. We can consider the residuals as normal distribution.

According to residual plot (appendix 4), the points were around the horizontal line. The assumption of equal variance was met.

The assumption of independent individuals was also met. The data was released by Seattle government and investigators selected people randomly in Seattle area. Durbin-Watson test showed p-value of 0.08, which was greater than 0.05. We failed to reject the null hypothesis of uncorrelated data (appendix 5).

Then, Anova was conducted. According to the Anova table (appendix 6), the p-value

of age was near 0 and the p-value of sex was 0.0003. We rejected $\alpha_i = 0$ and $\beta_j = 0$. Sex and age both had significant effect on wages. Female and male got different wages. At different age level, at least one level was different from others.

Tukey interval was used next for deeper analysis. Overall, the hourly wages of male was 4.47 dollars more than female's. The Tukey interval of wage difference between male and female was (1.95, 4.99). The p-value was much less than 0.05. We rejected that male and female got equal wages.

According to Tukey interval for age (appendix 7), the difference of two groups were significant at the beginning and got non-significant, which showed that the wages increased fast at the beginning and gradually the increasing rate got smaller. The wages went to the peak at the age 50-55 and then decreased.

Conclusion

In our study, sex and ages are two significant factors which will affect wages. According to the result of Anova and Tukey interval, female and male get different wages and female get lower wages than male in Seattle area.

In my opinion, female and male get are not get equal wages in Seattle area. And, we all know Seattle is a city where people got higher education and more opened thought. However, different genders are still treated unequally. We assumed that in other areas, the condition may be even worse.

From results, we can conclude that the wages has a peak and then decreased, it has some other thins influenced result which out of my expectation.

According to the study, Tukey interval was only used for age in chronological order. And, we do not compare wage difference among every age level because that would be too many comparisons (12*11/2 = 66 comparisons) that we cannot handle. However, by the chronological order, we will see the tendency of how wages changed during human's lives, which achieved our objective.

On the other hands, the data was not perfect. Just like, for some age levels, investigators calculated the mean hourly wages from groups only with a little amount of people (appendix 9). The investigators only collected one female and one male under 20 age. They neither found more people at age 20 to 25 or over 70. Thus, the mean of hourly wages at these age levels might be not accurate.

And another limitation is that normality assumption was not very ideally, it would cause some problems. The qq-plot is not a completely straight line and Shapiro test shows

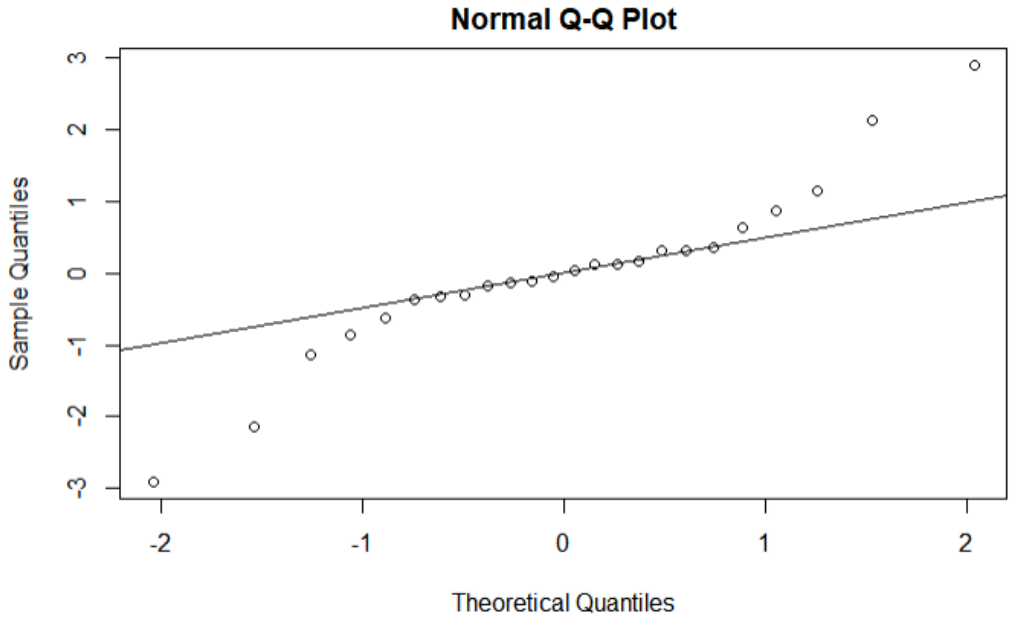
that the p-value is 0.08, it's a small p-value. Because, it is larger than 0.05, we will fail to reject the null hypothesis. However, 0.08 is an extreme number that if the significant level was set to be 0.1, we would reject normality. According to the Durbin-Watson test for uncorrelated data also showed this problem. if the p-value is larger, the result would be reliable.

Appendix

1. **data**

	Average of FEMALE HOURLY RATE	Count of FEMALE EMPLID	Average of MALE HOURLY RATE	Count of MALE EMPLID
0 - 20	15.76	1	13.43	1
20 - 25	22.92	26	26.3	63
26 - 30	29.9	163	31.64	325
31 - 35	31.43	292	35.54	530
36 - 40	34.9	376	37.1	726
41 - 45	36.05	498	39.28	985
46 - 50	36.57	576	40.38	1028
51 - 55	37.19	651	41.39	1036
56 - 60	36.57	562	40.66	913
61 - 65	34.91	369	40.67	521
66 - 70	33.09	75	40.82	131
70 -	30.8	11	34.52	26

2. **qq-plot**

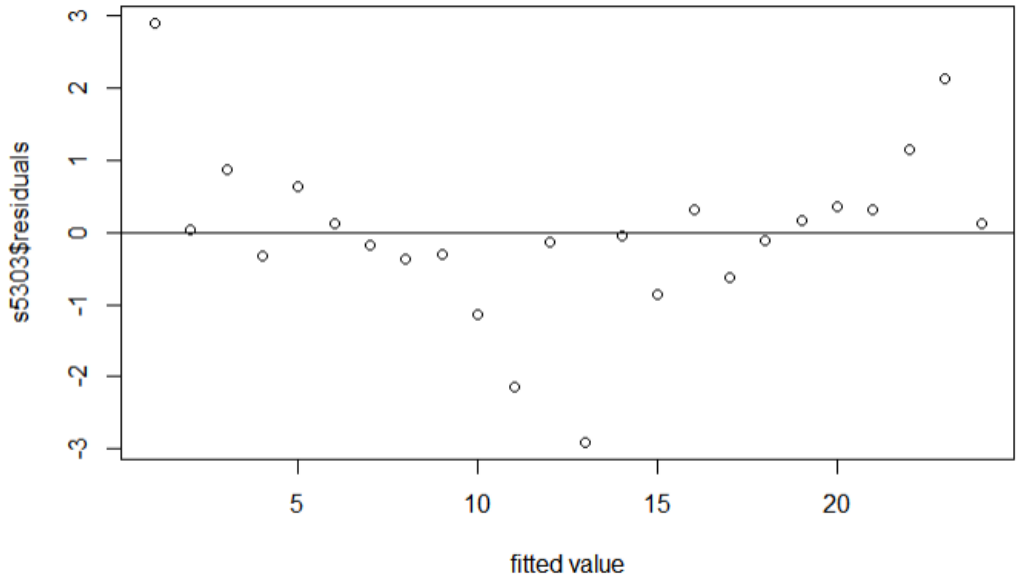


3. Shapiro test of normality

```
        Shapiro-Wilk normality test

data:  s5303$residuals
W = 0.92716, p-value = 0.08422
```

4. Residual plot



5. Durbin-Watson test

```
        Durbin-Watson test

data:  Wages ~ Age + Sex
DW = 1.3332, p-value = 0.08097
alternative hypothesis: true autocorrelation is greater than 0
```

6. Anova table

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Age         11 1165.7  105.98   36.89 4.53e-07 ***
Sex          1   72.2   72.25   25.14 0.000393 ***
Residuals   11   31.6    2.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7. Tukey interval for Sex

	diff	lwr	upr	p adj
M-F	3.47	1.946941	4.993059	0.0003934208

8. Tukey interval for Age

	diff	lwr	upr	p adj
20 - 25-0-20	10.015	3.1676068	16.862393	0.003309311
26 - 30-20 - 25	6.160	-0.6873932	13.007393	0.091373154
31 - 35-26 - 30	2.715	-4.1323932	9.562393	0.877382507
36 - 40-31 - 35	2.515	-4.3323932	9.362393	0.917761206
41 - 45-36 - 40	1.665	-5.1823932	8.512393	0.994870927
46 - 50-41 - 45	0.810	-6.0373932	7.657393	0.999993048
51 - 55-46 - 50	0.815	-6.0323932	7.662393	0.999992603
56 - 60-51 - 55	-0.675	-7.5223932	6.172393	0.999998923
61 - 65-56 - 60	-0.825	-7.6723932	6.022393	0.999991636
66 - 70-61 - 65	-0.835	-7.6823932	6.012393	0.999990559
over 70-66 - 70	-4.295	-11.1423932	2.552393	0.406032027

9. Distribution of people in this dataset

	Count of FEMALE EMPLID	Count of MALE EMPLID
0 - 20	1	1
20 - 25	26	63
26 - 30	163	325
31 - 35	292	530
36 - 40	376	726
41 - 45	498	985
46 - 50	576	1028
51 - 55	651	1036
56 - 60	562	913
61 - 65	369	521
66 - 70	75	131
70 -	11	26

Reference

Data from:

https://data.seattle.gov/City-Business/City-of-Seattle-Wages-Comparison-by-Gender-

Average/24tp-6m99

Moms Work 5 Extra Months to Earn as Much as Dads, Research Says:

https://www.caseygrants.org/evn/moms-work-5-extra-months-to-earn-as-much-as-dads-research-says/?gclid=CjwKCAjwqfDlBRBDEiwAigXUaKywqrq1RlPwuh7xOR6BUV4yjRbN6mEfpaRgPtpsKpCbFrIYhgYvThoC1gkQAvD_BwE