Nonparametric Methods Final Project

Group 11

Due: 5/13/2019

Title:

Wage difference between Black &White People and the reason behind it

Team Members:

Chen Chen *cc*4291

Sixing Hao *sh*3799

Wenwen Shen *ws*2561

Yuting He *yh*3054

Yang Meng *ym*2696

~~~~~~

Contents:

Part I: Introduction

Part II: Main objectives

- Test for Normality
- Exploratory Data Analysis
- Correlation

Part III: Description

- Two Sample Tests for Difference in Wage
- Survival Analysis for Education Years between Races
- Mantel-Haenszel Test for Work Experience between Races

Part IV: Results

Part V: Conclusions

Part VI: Appendix
~~~~~~

Part I: Introduction

The goal of our final project is to study the wage difference between Black and White People and the reason behind it.

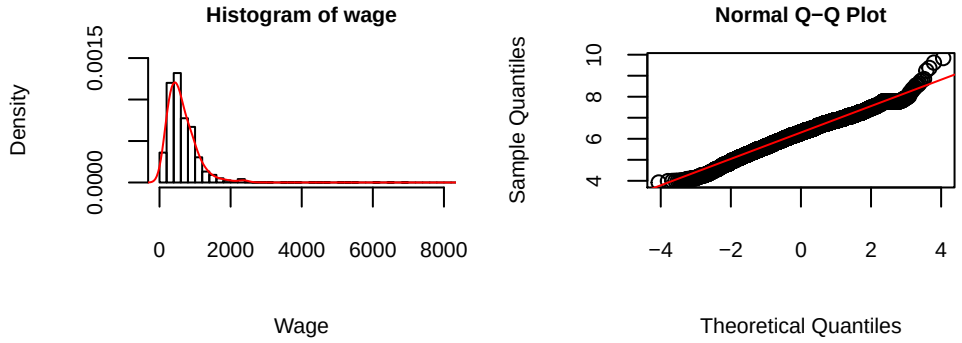
The data is roughly about 25,000 records of people between the ages of 18 and 70, and the data are taken many decades ago so the wages are lower compared to current times. This is how our dataset looks like:

Response	Numerical	Weekly wages (in dollars)	wage
Predictors	Numerical	Years of education	edu
	Numerical	Job experience	exp
	Numerical	Number of employees in a company	emp
	Numerical	Commuting distance	com
	Category	US region (Midwest/Northeast/South/West)	reg
	Category	Race (Black/White)	race

Part II: Main objectives

Before we begin to use non-parametric methods, we want to have a brief idea about our dataset, and this plan includes **test for normality**, **Exploratory Data Analysis** and **Correlation** among variables.

1. Test for Normality



It can be shown that histogram of wage is heavily right skewed, so we may need the log transformation to make it symmetric. Even using Log transformation, wage is still not normal distributed from Q-Q plot and Shapiro-wilk test (Appendix 1:Shapiro-wilk test for wage), we have $P - value_{Shapiro-wilk\ test} = 1.311 * 10^{-13}$, so that is the reason why we go to Non-parametric methods.

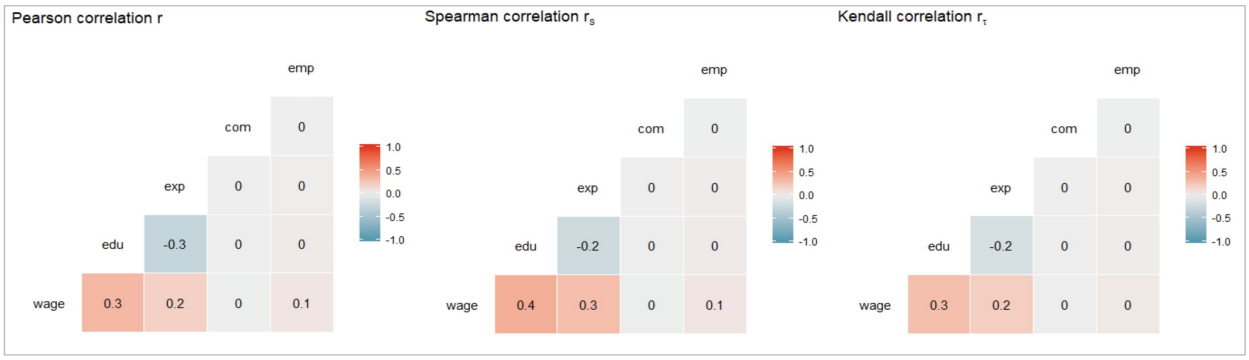
2. Exploratory Data Analysis

From the boxplots which show the differences based on Race (Appendix 2: EDA-Two Sample Tests for Difference in Wage), we can generate an explore analysis table:

Variables	Wage	Education	Employees	Commute Distance	Working Years
Boxplot Difference	Yes	Yes	Not Big	Not Big	Need further research

3. Correlation

As it is hard to tell the linear relationship of our dataset, we used three different methods to measure the strength of correlation between variables (Appendix 2: EDA-Correlation).



$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$t_{corr} = \sqrt{\frac{n-2}{1-r^2}}\, r \sim t_{(n-2)} \quad under\ H_0$$

$$r_s = \frac{\sum_{i=1}^{n}(R_i^x - \frac{n+1}{2})(R_i^y - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R_i^x - \frac{n+1}{2})^2 \sum_{i=1}^{n}(R_i^y - \frac{n+1}{2})^2}}$$

$$Z = \frac{r_s}{\sqrt{Var(r_s)}} = r_s\sqrt{n-1} \sim N(0,1) \quad under\ H_0$$

$$r_\tau = 2\frac{\sum_{i=1}^{n-1} V_i}{C_n^2} - 1$$

$$Z = \frac{r_\tau}{\sqrt{Var(r_\tau)}} = \frac{r_\tau}{\sqrt{\frac{4n+10}{9(n^2-n)}}} \sim N(0,1) \quad under\ H_0$$

where V_i is the number of pairs (X_i, Y_i) that are concordant

Based on the picture above, we can see that the results are very similar. Years of education, job experience and number of employees in a company all have positive correlations with wage while the correlation between years of education and job experience is negative.

Part III: Description

1. Two Sample Tests for Difference in Wage

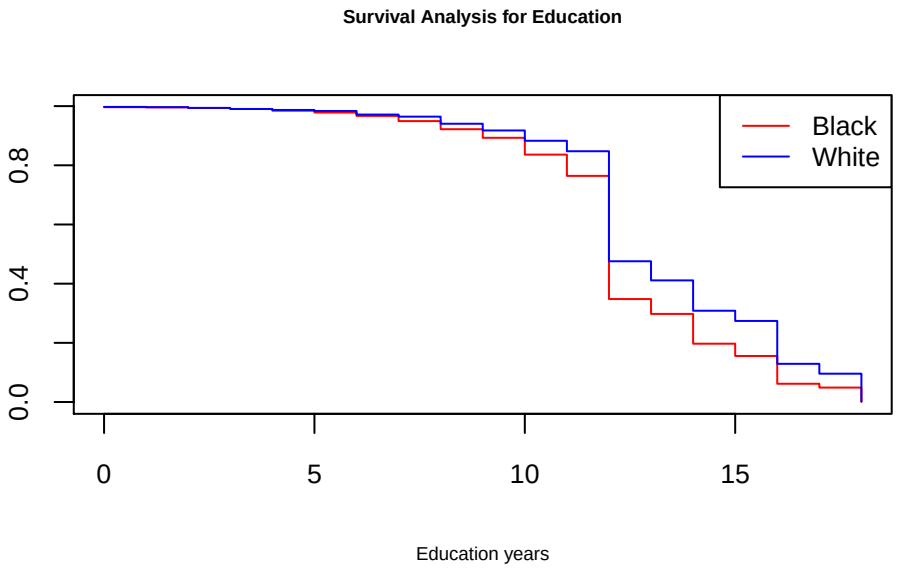
We use altogether 8 tests, including 4 **two-sample tests** and their **permutation versions** to test whether there exists difference in wage between black and white.

- Two sample t-test is applied to compare whether the average difference between two groups.
- The Wilcoxon rank-sum test is a nonparametric approach to the two sample t-test, which is based on the ranks.
- F-test considers a decomposition of the variability (ANOVA) to test the difference between groups.
- The Kruskal-Wallis test is a nonparametric approach to the F-test, which is based on ranks.

P-values for each one of them is **close to 0**, which shows a significant difference in wage between races.

2. Survival Analysis for Education Years between Races

We also construct **Survival analysis** to explore the reason behind the wage difference of education between black and white. We consider those people who start to work as censored, and the other people who are still studying as uncensored. The following plot shows the Kaplan-Meier estimates for the difference of education years between black and white.



In the Cox model:

$$\lambda(t|Z) = e^{\beta Z}\lambda_0(t),\ Z = \begin{cases} 0,\ Black \\ 1,\ White \end{cases}$$

which implies the longer education years white people have, with $\hat{e}^{\beta} = 0.74$ far away from 1.

Three tests are available for testing the difference between two groups (black and white here): **the likelihood ratio test**, **the Wald test** and **the Score test (Log Rank test)**. Based on our tests(Appendix 3: Survival Analysis), we reject the three null hypothesis as p-values are all close to 0, so we can conclude that the education years between black and white is different.

3. Mantel-Haenszel Test for Work Experience between Races

Before the Mantel-Haenszel test, we did **Chi-square test** to see whether the job experience influences the wage between black and white. We regard those people with more than 16 working years as experienced workers because 16 years is the median. $P - value_{Chi-square}$ is larger than 0.1, so we fail to detect the difference in work experience, which is differ from the common sense.

$$MH = \frac{(\sum_{k=1}^{s}[X_k - \frac{r_{1k}c_{1k}}{N_k}])^2}{\sum_{k=1}^{s}\frac{r_{1k}r_{2k}c_{1k}c_{2k}}{N_k^2(N_k-1)}} \sim \chi^2_{(1)}\ under\ H_0$$

Then we stratified the data into four regions, and utilized **Mantel-Haeszel test** to evaluate whether the work experience influences the wage with the consideration of different regions. $P - value_{Mantel-Haeszel\ test}$ is 0.0946< α=0.1 and the result is different from **Chi-square test** (Appendix 4: Chi-square & MH test).

From the boxplot of work experience group by the four regions (Appendix 4: Chi-square & MH test), we can see the **Simpson's paradox**. The difference in midwest and northeast region is significant, but their population is less than a half, so the difference becomes not as significant as it is in the whole dataset.

4. Regression

We finally fit a regression function to wage. We use race, education years, work experience, regions and the interaction term as predictors. After the Box-cox transformation, the model is:

$$log(wage) = 4.042 + 0.240race + 0.100edu + 0.168\sqrt{exp} + 0.194NE - 0.027NE\sqrt{exp}$$

$$where\ NE = \begin{cases} 1,\ Northeast \\ 0,\ Others \end{cases},\ race = \begin{cases} 1,\ White \\ 0,\ Black \end{cases}$$

In this model, all the p-values for coefficients are close to 0 and $R^2 = 0.31$ (Appendix 5: Regression), which is good enough in practice.

Part IV: Results

From part III, we find the following highly influenced predictors to wage:

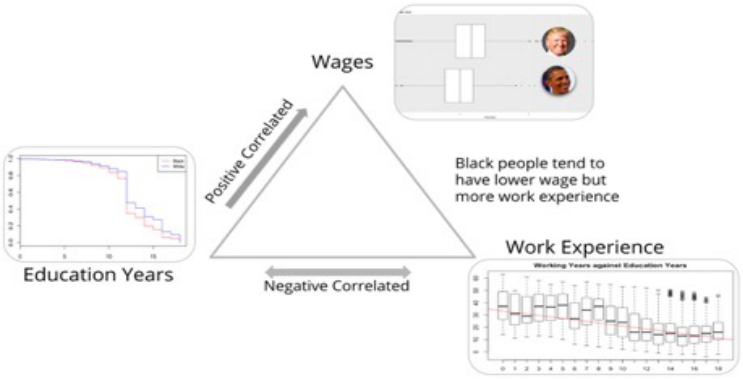
Variables	Results
Races	The wages of black and white is significantly different according to 8 two-sample tests with $p - value << 0$
Education Years	Black people receive less education from the K-M curve. The big jump at 12 years shows that half of black people discontinued their studies after high school, and there are only 10% of them graduated from college
Work Experience	The difference between $P - value_{Mantel-Haenszel}$ and $P - valus_{Chi-square}$ shows the Simpson's paradox and the importance of interaction term

From the regression model, we find out the following reasons behind the wage difference between races.

- Black people have lower wage because of the lack of education and the low salary working. Besides, it may also be caused by racial discrimination according to the coefficient of the dummy variable "race". So black people tend to have lower wage, even holding all the other factors fixed.
- The more education or work experience one has, the higher the salary will be. However, work experience is not as important as Northeast.

Part V: Conclusions

The triangle relationship can be shown below (Appendix 6: Conclusion):



In conclusion, black people have less wage compared to white people. The reasons behind are: Lack of education, regional difference, quality of work, and racial discrimination.

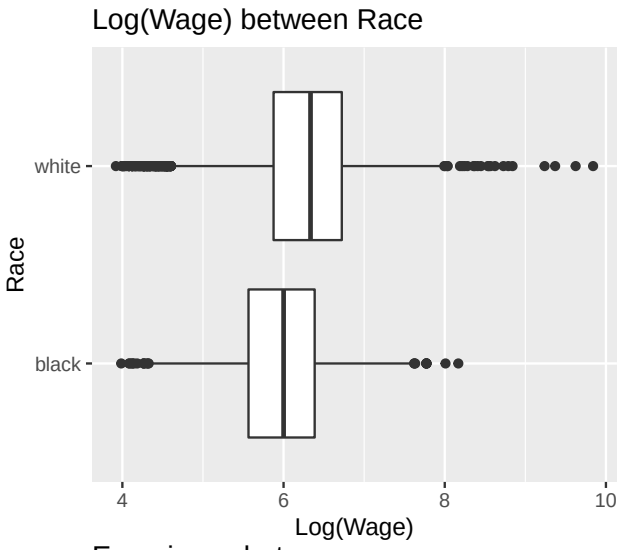
Part VI: Appendix

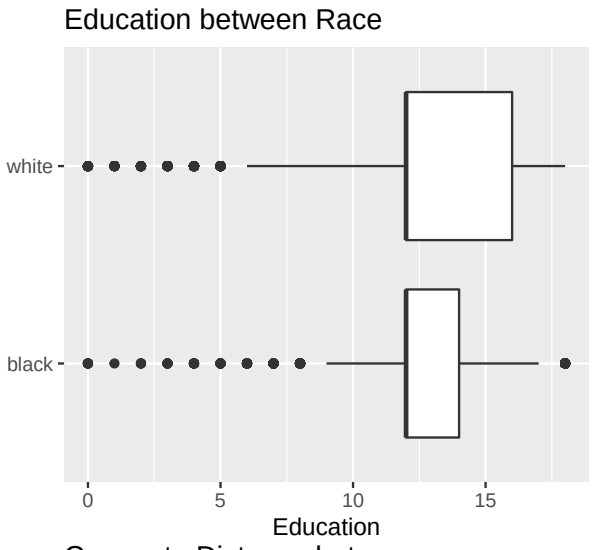
1. Shapiro-wilk test for wage

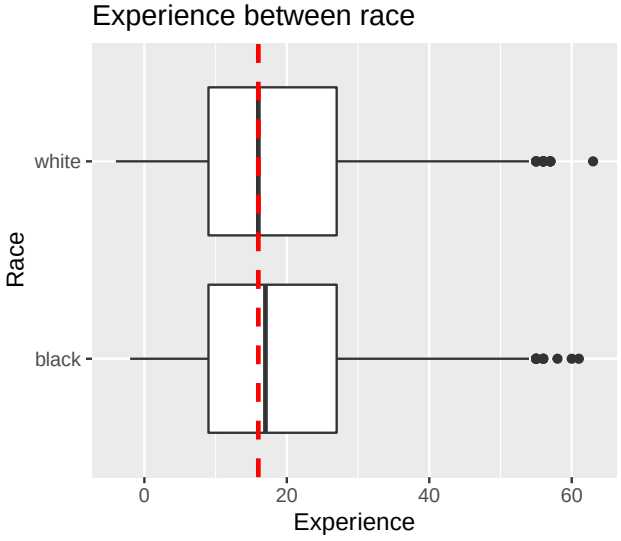
```
wage.test <- sample(wage,5000)
shapiro.test(log(wage.test))
```

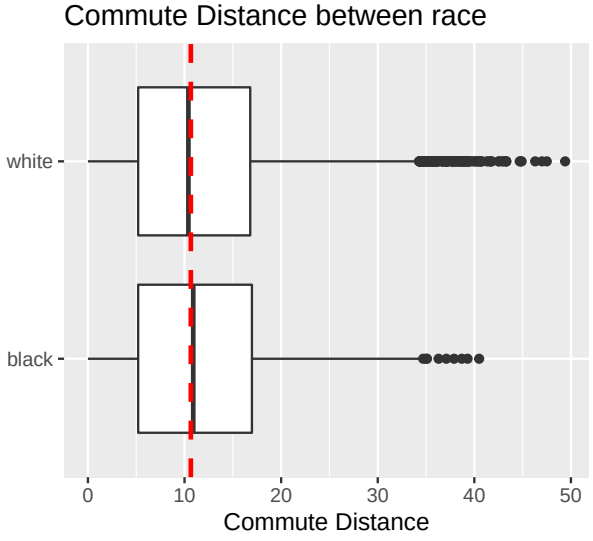
```
##
##  Shapiro-Wilk normality test
##
## data:  log(wage.test)
## W = 0.99493, p-value = 3.157e-12
```

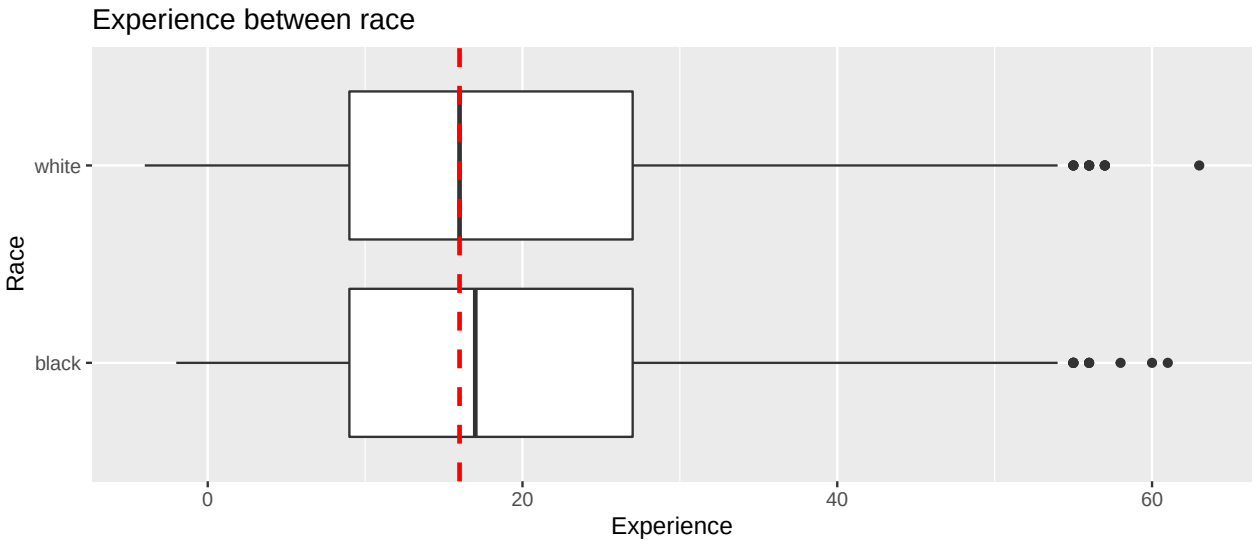
2. EDA



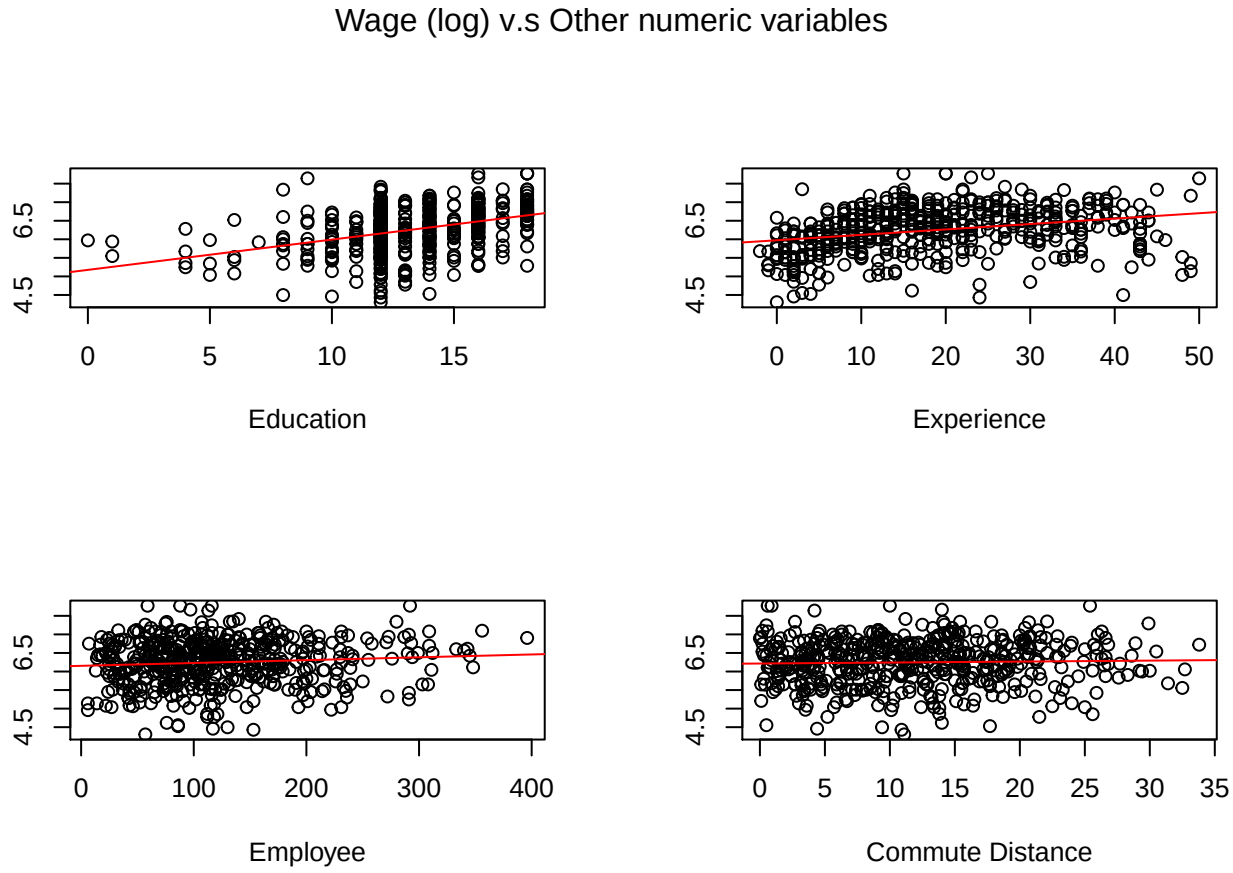








We sample 100 to see the scatter plot and the pattern, which is equivalent to the boxplot above. Correlation tests can also explain this in the following:



Scatter plot and best fitting line displayed similar results.

Two Sample Tests for Difference in Wage

In the following 4 chunks, we show the t-test, Wilcoxon Rank-sum test, F-test and Kruskal-Wallis test, and each test has a standard version and a permutation version.

t-test

```
# t-test
black <- wage[race=="black"]
white <- wage[race=="white"]
t.star <- as.numeric(t.test(black, white)[1])

# Permutation t-test
N <- 10^1 #sample N times
t.perm <- rep(0, N)
for(i in 1:N){
  index <- sample(1:length(wage), 1934)
  black.s <- wage[index]
  white.s <- wage[-index]
  t.perm[i] <- as.numeric(t.test(black.s, white.s)[1])
}

t.test(black, white)
```

```
##
##  Welch Two Sample t-test
##
## data:  black and white
## t = -22.975, df = 2952.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -195.4311 -164.6961
## sample estimates:
## mean of x mean of y
##  472.2070  652.2706
```

```
mean(t.perm <= t.star)
```

```
## [1] 0
```

- $t_{obs} = -22.975$ and $P - value_{t.test} < 2.2e-16$, so we reject H_0 at level α=0.1
- $P - value_{Permutation:t-test}$ is 0 even when $N = 10^4$ (because it is too extreme)

Wilcoxon Rank-sum test

```
# Wilcoxon Rank sum test
w.star <- as.numeric(wilcox.test(black, white)[1])

# Permutation Wilcoxon Rank-sum test
N <- 10^1 #sample N times
w.perm <- rep(0, N)
for(i in 1:N){
  index <- sample(1:length(wage), 1934)
```

```
  black.s <- wage[index]
  white.s <- wage[-index]
  w.perm[i] <- as.numeric(wilcox.test(black.s, white.s)[1])
}

wilcox.test(black, white)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  black and white
## W = 12536774, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
mean(w.perm <= w.star)
```

```
## [1] 0
```

- $W_{obs} = 12536774$ and $P - value_{Wilcoxon} < 2.2e-16$, so we reject H_0 at level $\alpha = 0.1$
- $P - value_{Permutation:Wilcoxon}$ is 0 even when $N = 10^4$ (because it is too extreme)

F test

```
# F-test
f.star <- as.numeric(anova(lm(wage ~ race))[[4]][1])

# Permutation F test
N <- 10^1 #sample N times
f.perm <- rep(0, N)
for(i in 1:N){
  race.s <- race[sample(1:length(wage), length(wage))]
  f.perm[i] <- as.numeric(anova(lm(wage ~ race.s))[[4]][1])
}

anova(lm(wage ~ race))
```

```
## Analysis of Variance Table
##
## Response: wage
##               Df     Sum Sq  Mean Sq F value    Pr(>F)
## race           1   56713844 56713844   275.6 < 2.2e-16 ***
## Residuals 20237 4164404694   205782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mean(f.perm >= f.star)
```

```
## [1] 0
```

- $F_{obs} = 275.6, P - value_{F.test} < 2.2e-16$, so we reject H_0 at level $\alpha = 0.1$
- $P - value_{Permutation:F.test}$ is 0 even when $N = 10^4$ (because it is too extreme)

Kruskal-Wallis test

```
# Kruskal-Wallis test
kw.star <- as.numeric(kruskal.test(wage ~ race)[1])

# Permutation Kruskal-Wallis test
N <- 10^1#sample N times
kw.perm <- rep(0, N)
for(i in 1:N){
  race.s <- race[sample(1:length(wage), length(wage))]
  kw.perm[i] <- as.numeric(kruskal.test(wage ~ race.s)[1])
}

kruskal.test(wage ~ race)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  wage by race
## Kruskal-Wallis chi-squared = 446.68, df = 1, p-value < 2.2e-16
```

```
mean(kw.perm >= kw.star)
```

```
## [1] 0
```

- $P-value_{KW} < 2.2e-16$,so we reject H_0 at level $\alpha = 0.1$
- $P-value_{Permutation:KW}$ is 0 even when $N = 10^4$ (because it is too extreme)

Mantel-Haenszel Test

Correlation

Pearson's correlation (Linear Measure)

$H_0 : \rho = 0 \quad v.s \quad H_a : \rho \neq 0$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$t_{corr} = \sqrt{\frac{n-2}{1-r^2}} r \sim t_{(n-2)} \quad under\ H_0$$

Spearman's correlation (Nonlinear measure)

$H_0 : \rho = 0 \quad v.s \quad H_a : \rho \neq 0$

$$r_s = \frac{\sum_{i=1}^{n}(R_i^x - \frac{n+1}{2})(R_i^y - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R_i^x - \frac{n+1}{2})^2 \sum_{i=1}^{n}(R_i^y - \frac{n+1}{2})^2}}$$

$$Z = \frac{r_s}{\sqrt{Var(r_s)}} = r_s\sqrt{n-1} \sim N(0,1) \quad under\ H_0$$

Kendall's tau (Counts of concordant and discordant pairs)

$H_0 : \tau = 0 \quad v.s \quad H_a : \tau \neq 0$

$$r_\tau = 2\frac{\sum_{i=1}^{n-1} V_i}{C_n^2} - 1$$

$$Z = \frac{r_\tau}{\sqrt{Var(r_\tau)}} = \frac{r_\tau}{\sqrt{\frac{4n+10}{9(n^2-n)}}} \sim N(0,1) \quad under\ H_0$$

where V_i is the number of pairs (X_i, Y_i) that are concordant

3. Survival Analysis

```
## Cox Model
summary(coxph(Surv(edu,status) ~ group))
```

```
## Call:
## coxph(formula = Surv(edu, status) ~ group)
##
##   n= 20239, number of events= 19773
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## groupwhite -0.30170   0.73956  0.02416 -12.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## groupwhite    0.7396      1.352    0.7054    0.7754
##
## Concordance= 0.518  (se = 0.001 )
## Likelihood ratio test= 143.9  on 1 df,   p=<2e-16
## Wald test            = 155.9  on 1 df,   p=<2e-16
## Score (logrank) test = 157.1  on 1 df,   p=<2e-16
```

4. Chi-square & MH test

Chi-square

Recall working years difference between races, it is hard to see difference. Now we define worker with more than 16 years of working experience as experienced. Chi-square test gets the same result here, in this test, the proportion of experienced workers in two races has no difference.

```
table.total <- table(salary$race,salary$exp>16)
table.total
```

```
##
##         FALSE TRUE
##   black   954  980
##   white  9333 8972
```

```
chisq.test(table.total)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table.total
## X-squared = 1.8587, df = 1, p-value = 0.1728
```

- $\chi^2_{obs} = 1.8587, P - value_{chi-square.test} = 0.1728$, so we fail to reject H_0 at level $\alpha = 0.1$

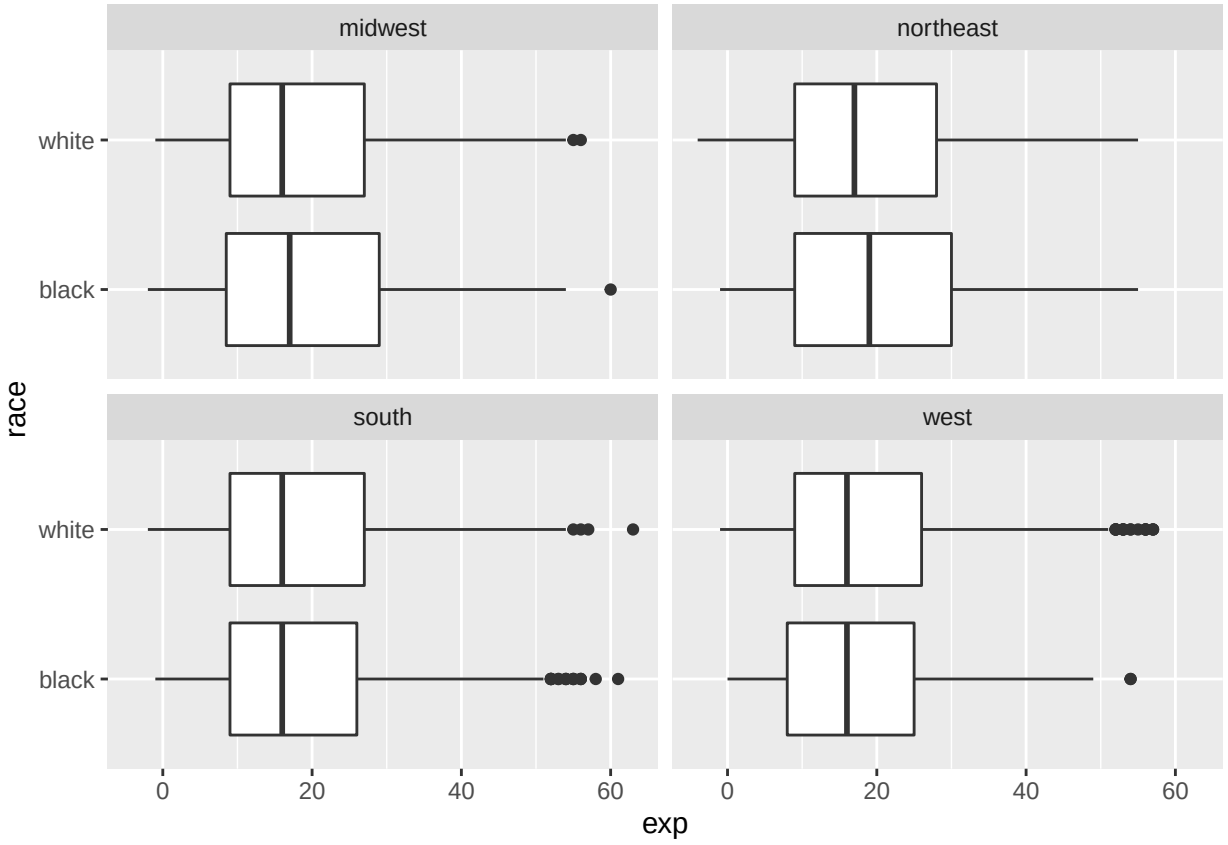
From the 4 boxplots and the population table below, the difference in midwest and northeast region is significant, but the total population is less than a half, so here comes the simpson's paradox. So we carry out the MH test.

MH test

```
table(reg)
```

```
## reg
##   midwest northeast     south      west
##      4852      4650      6427      4310
```

```
reg_box <- ggplot(salary,aes(x=race,y=exp))+geom_boxplot()+facet_wrap(~reg)+coord_flip()
reg_box
```




```
salary.1 <- filter(salary,reg=="midwest")
table.1 <- table(salary.1$race, salary.1$exp>16)
salary.2 <- filter(salary,reg=="northeast")
table.2 <- table(salary.2$race, salary.2$exp>16)
salary.3 <- filter(salary,reg=="south")
```

```r
table.3 <- table(salary.3$race, salary.3$exp>16)
salary.4 <- filter(salary,reg=="midwest")
table.4 <- table(salary.4$race, salary.4$exp>16)

table.1
```

```
##
##         FALSE TRUE
##   black   153  154
##   white  2366 2179
```

```r
table.2
```

```
##
##         FALSE TRUE
##   black   134  183
##   white  2144 2189
```

```r
table.3
```

```
##
##         FALSE TRUE
##   black   577  568
##   white  2673 2609
```

```r
table.4
```

```
##
##         FALSE TRUE
##   black   153  154
##   white  2366 2179
```

```r
tables.reg <- array(dim = c(2, 2, 4))
tables.reg[,,1] <- table.1
tables.reg[,,2] <- table.2
tables.reg[,,3] <- table.3
tables.reg[,,4] <- table.4
mantelhaen.test(tables.reg, exact = T)
```

```
##
##  Exact conditional test of independence in 2 x 2 x k tables
##
## data:  tables.reg
## S = 1017, p-value = 0.09069
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8412172 1.0136470
## sample estimates:
## common odds ratio
##         0.9234377
```

- $p - value = 0.09$, so we reject H_0 at level $\alpha = 0.1$

5. Regression

```r
NE <- reg=="northeast"
summary( lm(log(wage) ~ edu + sqrt(exp) * NE + race) )
```

```
##
## Call:
## lm(formula = log(wage) ~ edu + sqrt(exp) * NE + race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7110 -0.2961  0.0360  0.3393  3.6594
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.042444   0.024795 163.038  < 2e-16 ***
## edu               0.100445   0.001307  76.880  < 2e-16 ***
## sqrt(exp)         0.168118   0.002830  59.403  < 2e-16 ***
## NETRUE            0.193577   0.024463   7.913 2.64e-15 ***
## racewhite         0.240236   0.012588  19.085  < 2e-16 ***
## sqrt(exp):NETRUE -0.026620   0.005599  -4.755 2.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5228 on 20070 degrees of freedom
##   (163 observations deleted due to missingness)
## Multiple R-squared:  0.3094, Adjusted R-squared:  0.3093
## F-statistic:  1799 on 5 and 20070 DF,  p-value: < 2.2e-16
```

6. Conclusion

```
boxplot(exp ~ edu, main="Job Experience against Education Years",
        ylab="Job Experience", xlab="Education")
abline(lm(exp ~ edu), col=2)
```