Coursera : Inferential Analysis Peer assignment

Setting the working directory and load the gss data

```
setwd("C:/Users/ddddd/Inferential")
load("C:/Users/ddddd/Inferential/gss.R")
```

Install the packages and load the library

```
library(devtools)
devtools::install_github("statswithr/statsr")
```

```
## Skipping install of 'statsr' from a github remote, the SHA1 (7b65cd22) has not changed since last install.
##   Use `force = TRUE` to force installation
```

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Part 1 About the Data

According to the given information Gss, the General Social Survey work is to moniter the social activities and to study the growing complexity of American Society. GSS has been working for this since 1972. The basic purpose of this project work is to make us understand the GSS functionality and it's services. And to be able to understand the statistical Problems. The main topics for research are the emerging issues given in the information provided by coursera, it includes national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior. THe GSS conducted a face to face survey to the people of America then asked them and collect the information of their experiences, behaviours etc. THe scope of inference is to research about this data . Their collected data is randomly sampled. In this data the variables are related and the survey is randomly sampled as it is not neccessory that every resident of country will join this survey and ready to answer the questions.They, basically choose the adult people for survey

Part 2 what is the Research question

According to GSS data , I just study about the income difference among the people of United States accoring to the race. So, the research question is to find out the that #who earns more and who earns less?

For this solution, we need to choose the related varibles provided in the dataset. So the two variables are: race and coninc Here race gives information of category of people ,i.e white , black and others whereas coninc gives information about their respective incomes.

Part 3 Perform Exploratory Data Analysis

Plotting the incomes of diffrent races in United States

```
#Removing the missing values
income_race<- gss[ which(!is.na(gss$race) & !is.na(gss$coninc)), ]
income_race <- income_race %>% select(race, coninc)
#summary of incomes
summary(income_race)
```

```
##     race           coninc
##  White:41824   Min.   :   383
##  Black: 6956   1st Qu.: 18445
##  Other: 2452   Median : 35602
##                Mean   : 44503
##                3rd Qu.: 59542
##                Max.   :180386
```

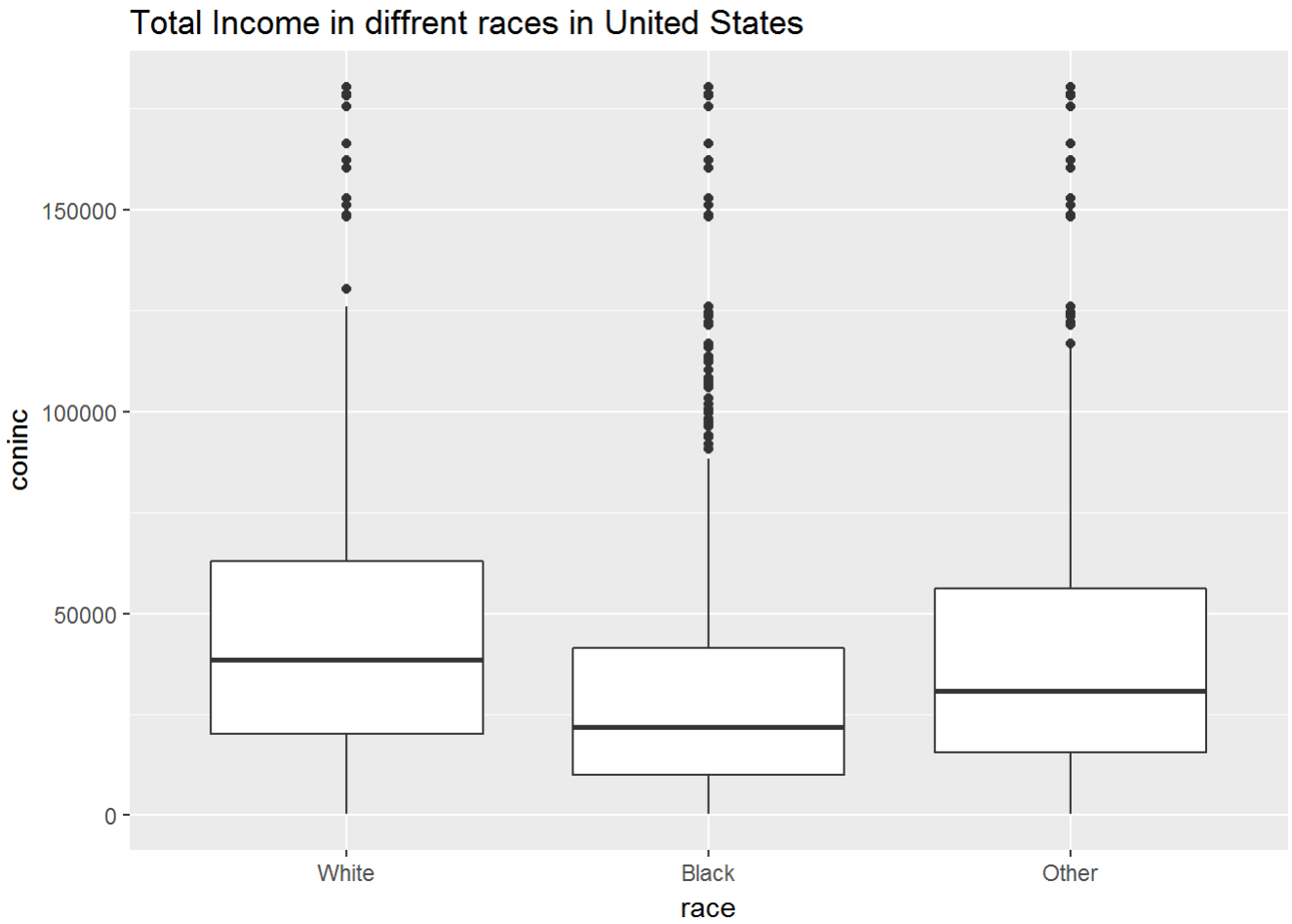
```
dim(income_race)
```

```
## [1] 51232     2
```

Here, we can see that there are total 41824 white, 6956 black and 2452 others races. and also given the detail of conincs, i.e. constant income of family. The data set having total 51232 observation along with 2 varibles. Now, we will see the boxplot to know , who has more income.

Plotting the boxplot to compare the total income of diffrent races in United States

```
ggplot(aes(x= race, y= coninc), data=income_race) + geom_boxplot() + ggtitle('Total Income in diffrent races in United State
s')
```



Here, the boxplot describes that the white race is the maximum income holder and the black race is the minimum income holder and the other races are in between both.

Part 4 Performing Inference Analysis

For the above research question the inferential analysis is processed below with 6 steps which are given in the “stat_inf_project.Rmd” file provided in the coursera’s Inferential Analysis peer assignment: (i)State hypotheses (ii)Check conditions (iii)State the method(s) to be used and why and how (iv)Perform inference (v)Interpret results (vi)If applicable, state whether results from various methods agree

Step (i)State hypotheses tests:

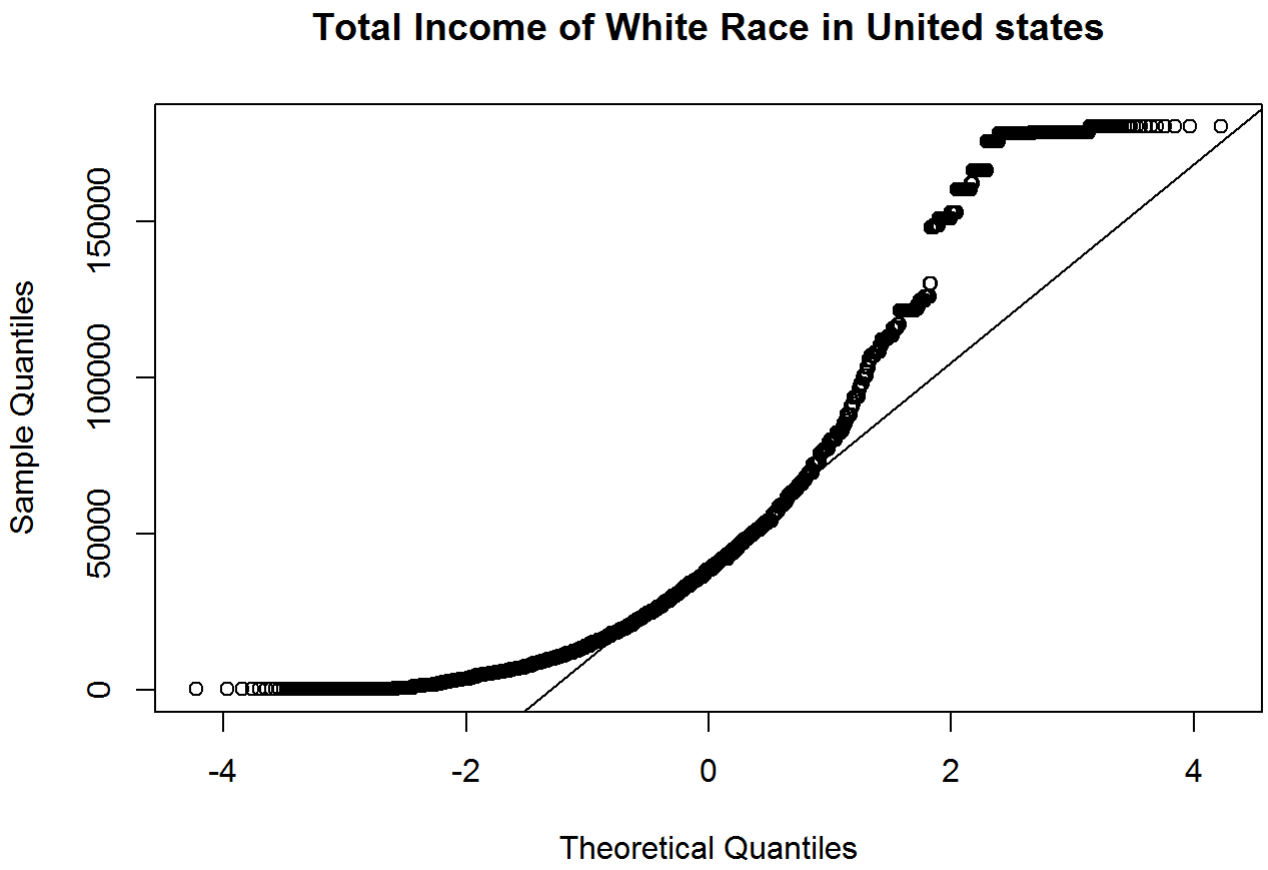
There are two hypotheses tests : 1. Alternative hypotheses : The alternative hypothesis could be less than, greater than , or twosided.In the above research question, When the income of a family doesn’t same,i.e.it should be greater than , less than. 2. Null hypotheses : The null hypothesis Puts equal the two population.In the above research question, when the income of a family is same in all three races.

Step(ii)Checking the conditions:

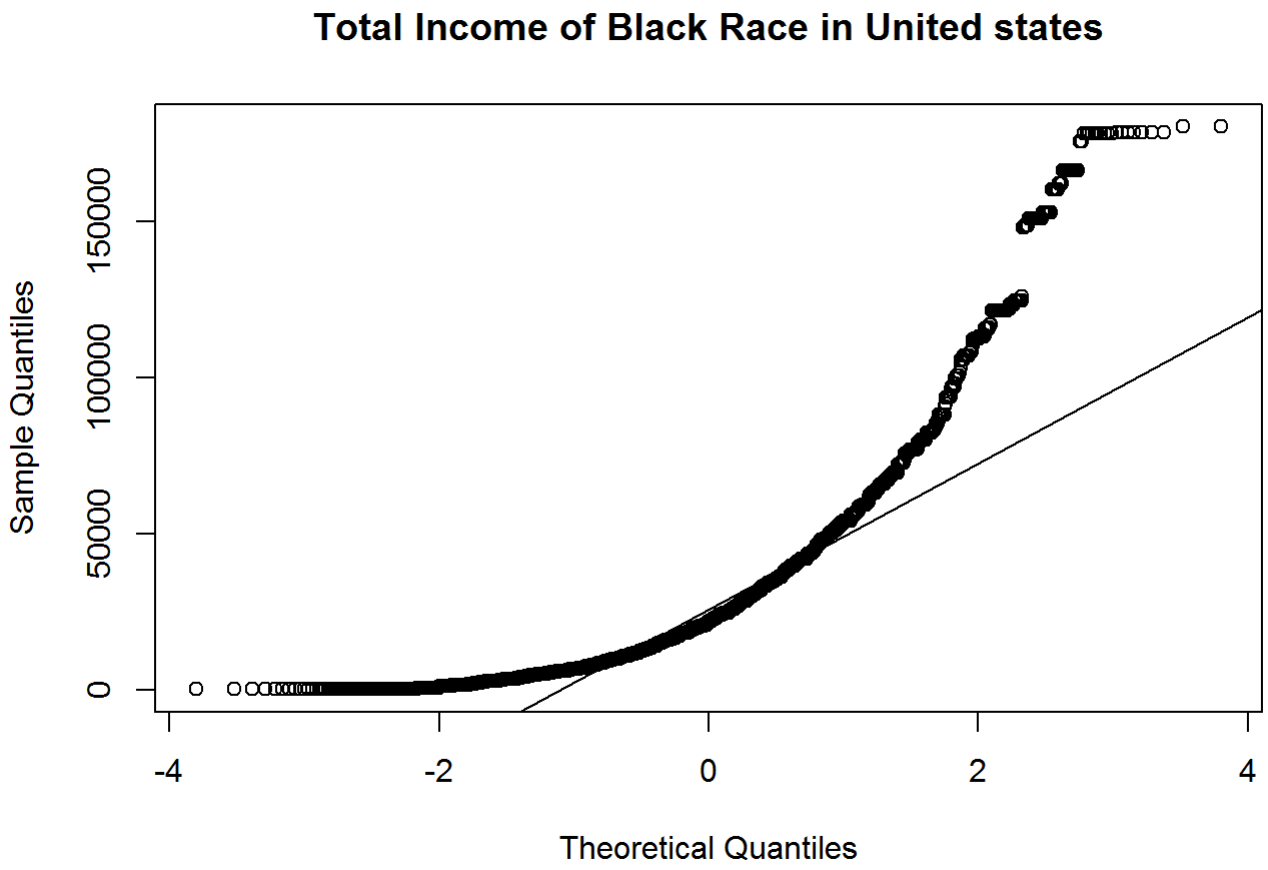
All of the three races white, black and others do not depend on each other ,i.e. income of one race does not effect the other one’s income or we can say independent.They are the random samples of observation.

Now, checking the Total income of each race by qqplot obseving theoritical quantiles verses sample quantiles and then observe the results.

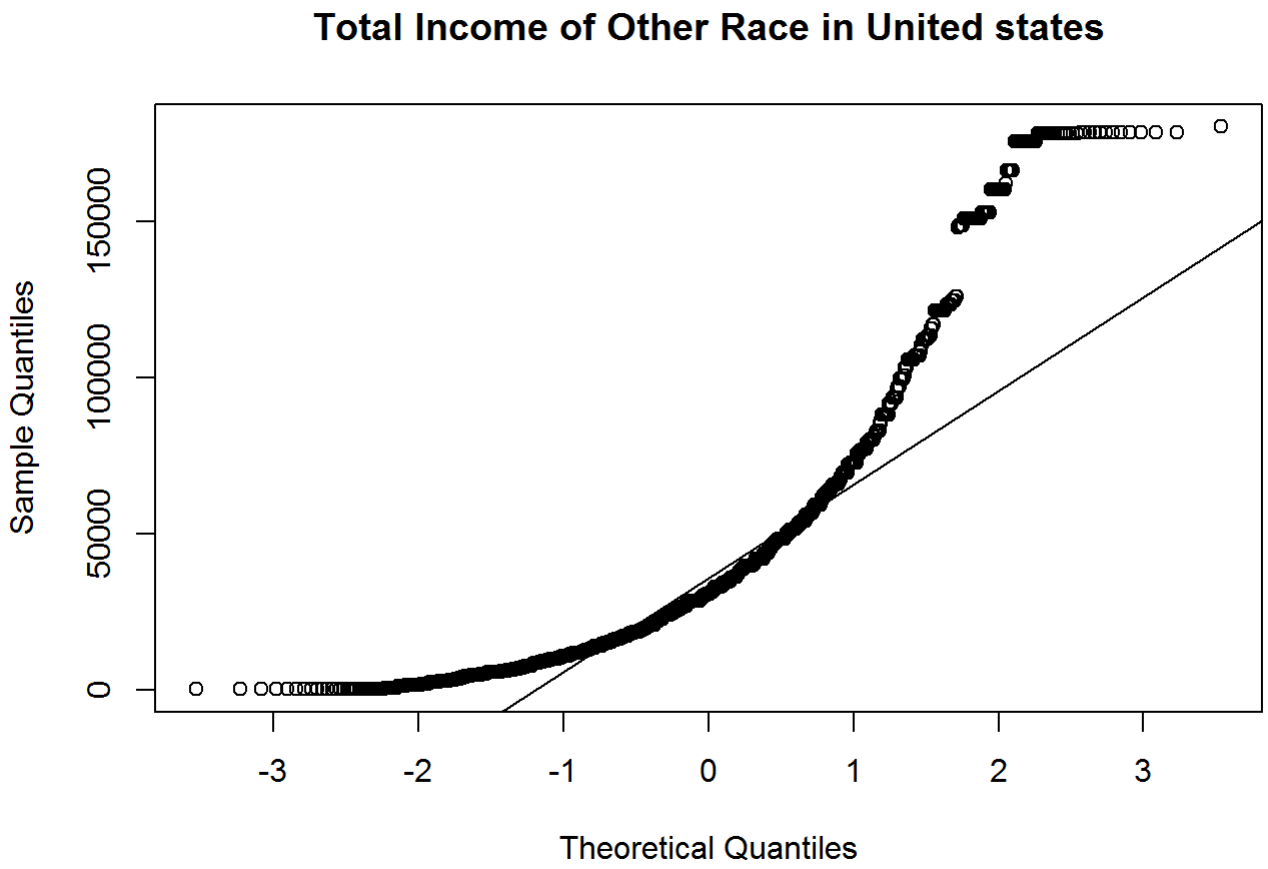
```
#Total Income of White Race in United states
qqnorm(income_race$coninc[income_race$race=='White'], main="Total Income of White Race in United states")
qqline(income_race$coninc[income_race$race=='White'])
```




```
qqnorm(income_race$coninc[income_race$race=='Black'], main="Total Income of Black Race in United states")
qqline(income_race$coninc[income_race$race=='Black'])
```




```
qqnorm(income_race$coninc[income_race$race=='Other'], main="Total Income of Other Race in United states")
qqline(income_race$coninc[income_race$race=='Other'])
```



The idea of income difference is quite visible that white race has more income and black has less . But we also use ANOVA for the next part because we have more than two categorical varibles.

Step 3 Using ANOVA for: State method to be used and why and how

As mentioned in the rubric file (html file) provided by coursera Inferential analysis Peer assignment that: as in this process we are using more than two ,i.e. three catogorical variables and we need to observe that if our observation is worth it ,i.e. the values are not so much far so that it will not create problem. so we need to use ANOVA for a significant result.

```
income_race_method <- lm(coninc ~ race, data = income_race)
anova(income_race_method)
```

```
## Analysis of Variance Table
##
## Response: coninc
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## race          2 1.6989e+12 8.4944e+11  675.08 < 2.2e-16 ***
## Residuals 51229 6.4461e+13 1.2583e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the P-value. If the P-value is smaller than 0.05 then reject the null hypothesis. Here, the p value is very small,.i.e. less than 0.05 so we need to avoid the null hypotheses and be in the favor of alternative hypothesis.

Step 4 Now , interpreting the required results

Although, we have applied Anova and also get the result that there are diffrence between the incomes of races,i.e. significant result, yet we want to know which pair of races has the income variations. so,now test them pairwise. For this we test the variables by pairing them for example: white-black, white-other and black other and see their income differences. For this , we'll get a p value.

```
pairwise.t.test(income_race$coninc, income_race$race, p.adj ='bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  income_race$coninc and income_race$race
##
##       White   Black
## Black < 2e-16 -
## Other 1.4e-09 < 2e-16
##
## P value adjustment method: bonferroni
```

Pairwise comparison of incomes of different races with the use of Bonferroni adjustment. Bonferroni adjustment working is to divide the error by the number of tests performed.Here, the p value is very smaller or nearest to zero so we need to avoid the null hypotheses.

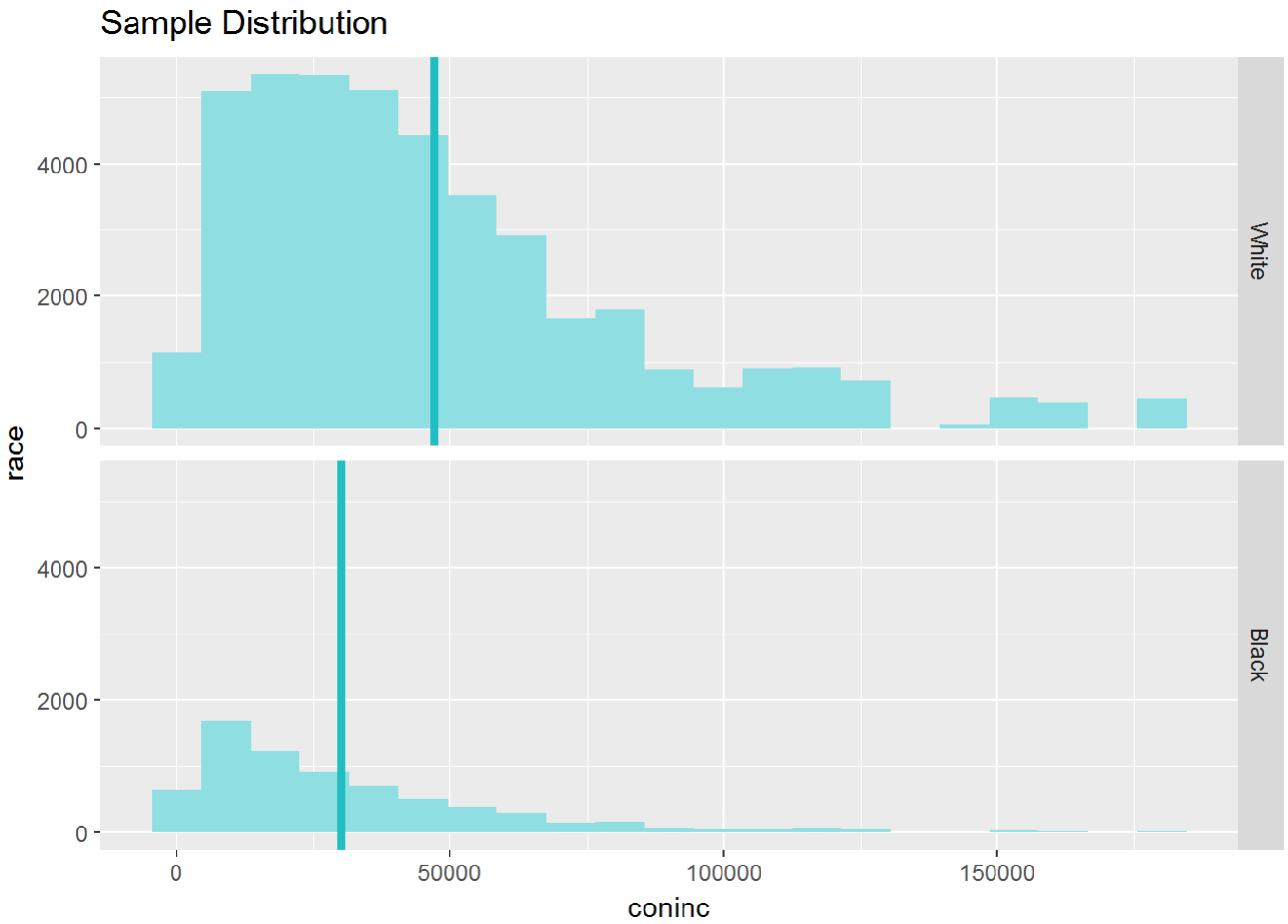
Step 6 Finding the confidence interval

We find the difference between all three pairs. Here, the details are: 1. Explanatory variable is catagorical 2.Response varible is numerical 3.Confidence level is 94 % 4 Method is theoritical Now, compare each of the pairs

Comparing White-Black pair for the sample distribution

```
income_race_new <- subset(income_race, race== 'White' | race=='Black', select=c(race, coninc))
income_race_new <- droplevels(income_race_new)
inference(y = coninc, x = race, data = income_race_new, statistic = "mean", type = "ci", conf_level = 0.94, method = "theore
tical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_White = 41824, y_bar_White = 47006.7433, s_White = 36405.4758
## n_Black = 6956, y_bar_Black = 30185.0203, s_Black = 28047.6414
## 94% CI (White - Black): (16105.9618 , 17537.4842)
```



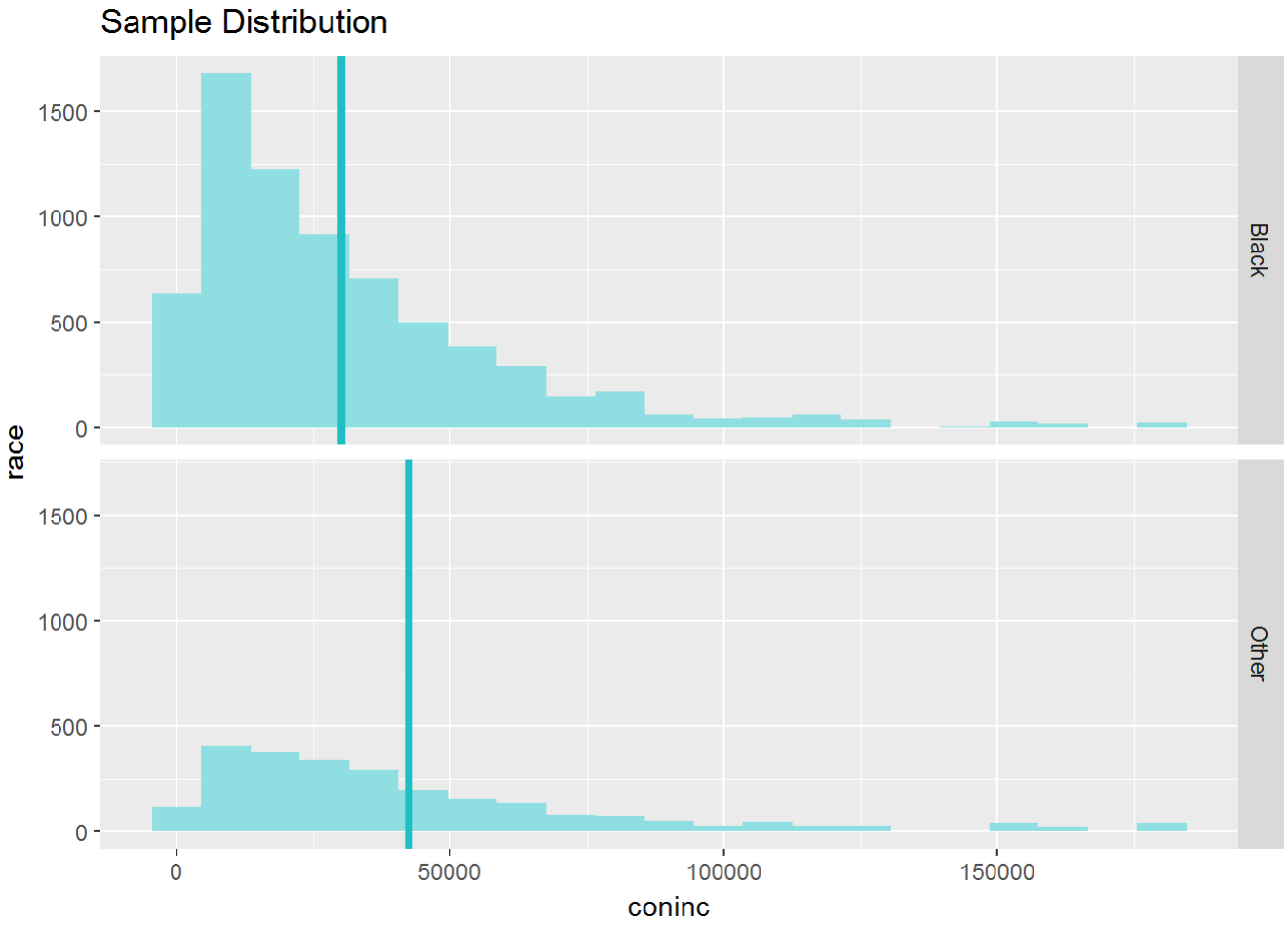
Response variable: numerical, Explanatory variable: categorical (2 levels) n_White = 41824, y_bar_White = 47006.7433, s_White = 36405.4758 n_Black = 6956, y_bar_Black = 30185.0203, s_Black = 28047.6414 98% CI (White - Black): (15936.3404 , 17707.1055)

Result: Thus, we are 94% confident that the income of white race is from 15936 to 17707 greater than black race.

Comparing Black-Other pair for the sample distribution

```
income_race_new <- subset(income_race, race== 'Other' | race=='Black', select=c(race, coninc))
income_race_new <- droplevels(income_race_new)
inference(y = coninc, x = race, data = income_race_new, statistic = "mean", type = "ci", conf_level = 0.94, method = "theore
tical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_Black = 6956, y_bar_Black = 30185.0203, s_Black = 28047.6414
## n_Other = 2452, y_bar_Other = 42415.4274, s_Other = 38105.4353
## 94% CI (Black - Other): (-13810.6377 , -10650.1766)
```



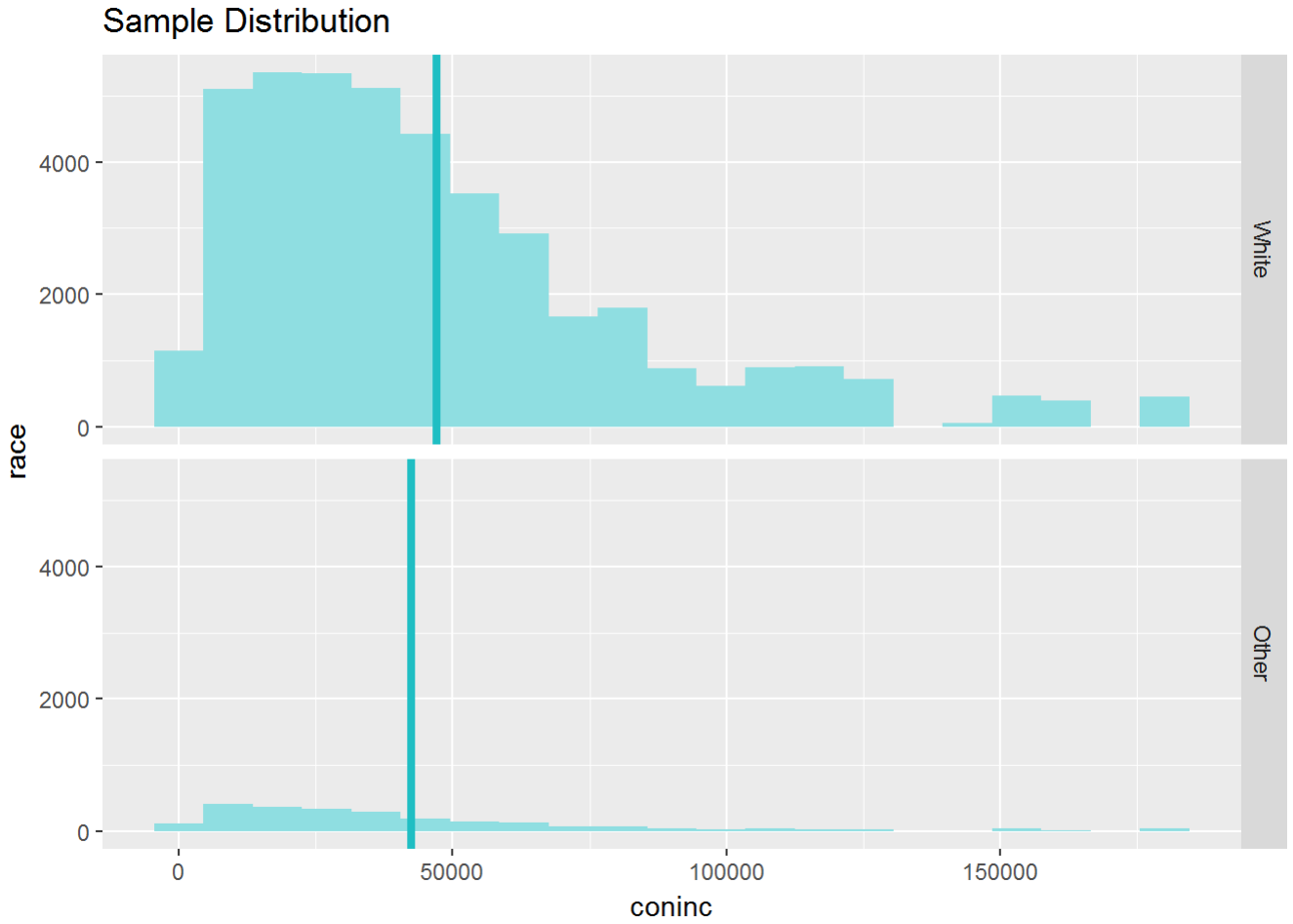
Response variable: numerical, Explanatory variable: categorical (2 levels) n_Black = 6956, y_bar_Black = 30185.0203, s_Black = 28047.6414 n_Other = 2452, y_bar_Other = 42415.4274, s_Other = 38105.4353 98% CI (Black - Other): (-14185.3634 , -10275.4509)

Results: Thus, we are 94% confident that the income of black race is less than black race.

Comparing White-Other pair for the sample distribution

```
income_race_new <- subset(income_race, race== 'Other' | race=='White', select=c(race, coninc))
income_race_new <- droplevels(income_race_new)
inference(y = coninc, x = race, data = income_race_new, statistic = "mean", type = "ci", conf_level = 0.94, method = "theore
tical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_White = 41824, y_bar_White = 47006.7433, s_White = 36405.4758
## n_Other = 2452, y_bar_Other = 42415.4274, s_Other = 38105.4353
## 94% CI (White - Other): (3105.0767 , 6077.555)
```



Response variable: numerical, Explanatory variable: categorical (2 levels) n_White = 41824, y_bar_White = 47006.7433, s_White = 36405.4758 n_Other = 2452, y_bar_Other = 42415.4274, s_Other = 38105.4353 94% CI (White - Other): (3105.0767 , 6077.555)

Result: Thus, we are 94% confident that the income of white race is from 3105 to 6077 greater than black race.