Statistical inference with the GSS data

Setup

Load packages

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

Load data

```
load("C:/Users/User/Desktop/Rstudio/coursera lab/inference statistic/final project/_5db435f06
  000e694f6050a2d43fc7be3_gss.Rdata")
```

Part 1: Data

The study cannot be generalize to the entire population of United states.

The GSS gather data from survery through personal-interview. Although the samples may be randomly selected, not all people are accessible for interview. There might be possibility of convinience samples being taken.

Since it is an observational study and there is no randomise assignment, it implies correlation relationship instead of causation. To imply a causation relationship, study need to be done in randomize assignment.

Part 2: Research question

analysis of correlation between equality of opportunity and different races (is there an equality amongst races)

1. correlation study between education and different races. Is the rate educated (at least a bachelor degree) are associated with races?
2. correlation study between how easy to find the equivalent job and different races amongst those with Bachelor degree. Is the rate of easiness to find equavalent job are associated with races amongst the bachelor degree?
3. correlation study between uneployment and different race amongst those with Bachelor degree. Is the rate of unemployment are associated with races amongst the bachelor degree?

note: apart from analysing the easiness to find equavalent job and unemployment and their education qualification, more topics are need to be analysed to establish stronger correlation of equality of opportunity and different races.

Part 3: Exploratory data analysis

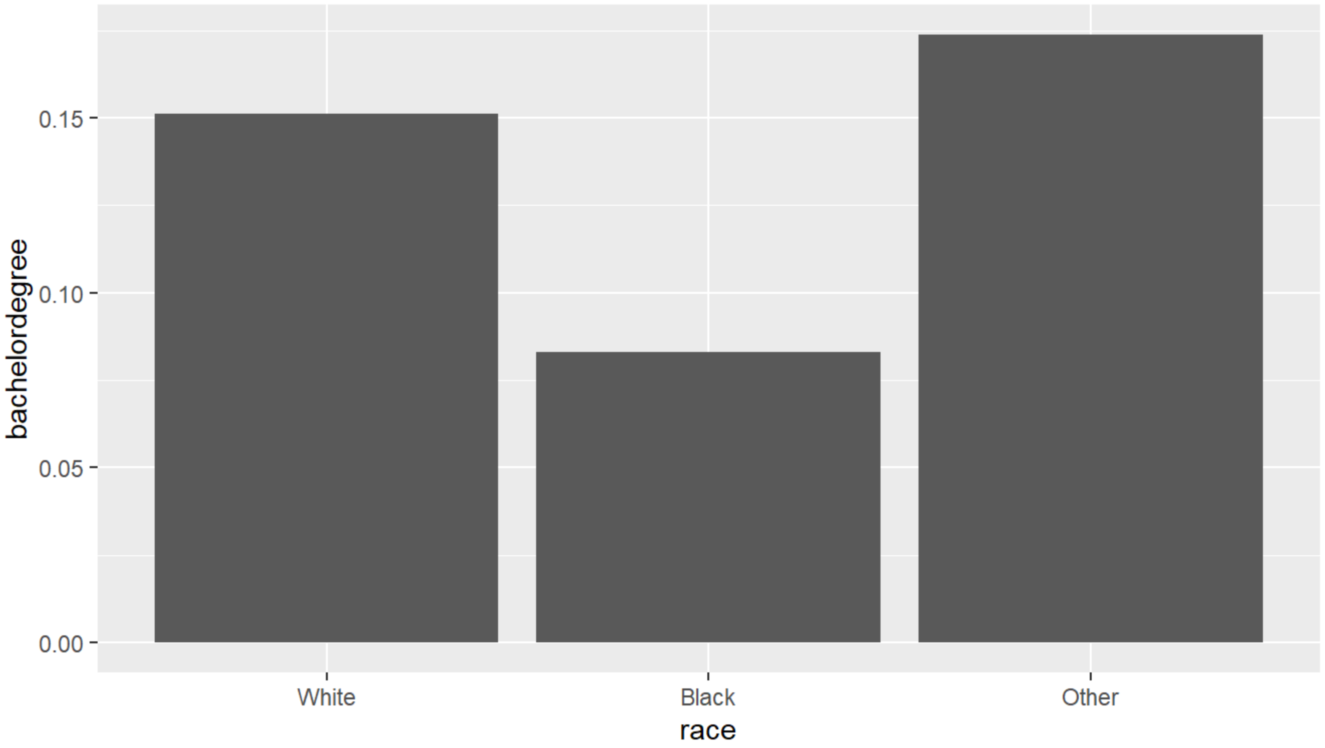
1. Is the rate of higher qualification (at least a bachelor degree) are associated with races?

```
educated_race<- gss%>%
  select(race, degree, age)%>%
  filter (age >=30, !is.na(race), !is.na(degree))
```

Note: to exclude those who are below 30 years old in the observation as they may not completed their highest education

```
educated_race_rate<- educated_race%>%
  group_by(race)%>%
  summarise (bachelordegree=sum(degree=="Bachelor")/n())%>%
  arrange(bachelordegree)
```

```
ggplot (educated_race_rate, aes (x=race, y=bachelordegree)) +
        geom_bar(stat="identity")
```

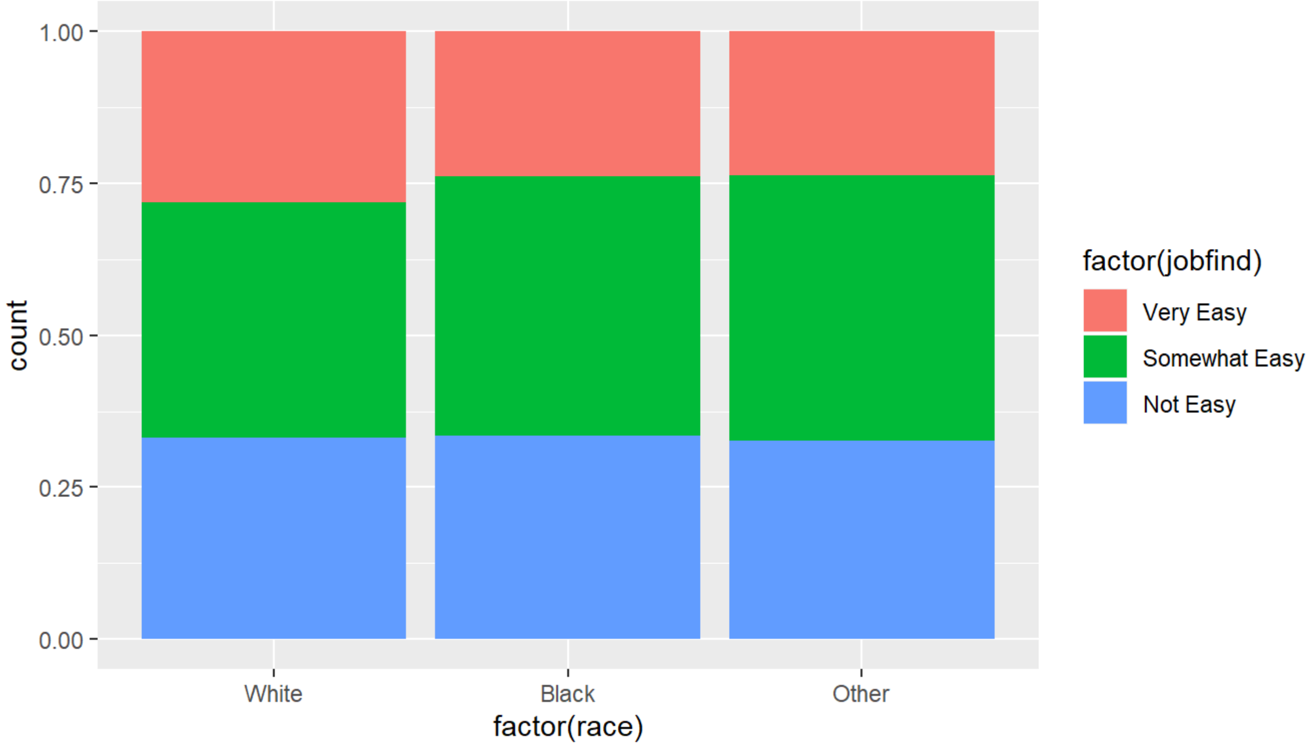


from the graph analysis, it is likely that higher education qualification rate are associated with races. black race are likely to have lower education qualification rate than other races.

2)Is the rate of easiness to find equavalent job are associated with races amongst the bachelor degree?

```
jobfind_race<- gss%>%
  select(race, jobfind, degree)%>%
  filter (degree == "Bachelor", !is.na(race), !is.na(jobfind) )
```

```
ggplot (jobfind_race, aes (factor(race), fill=factor(jobfind))) +
  geom_bar(position="fill")
```

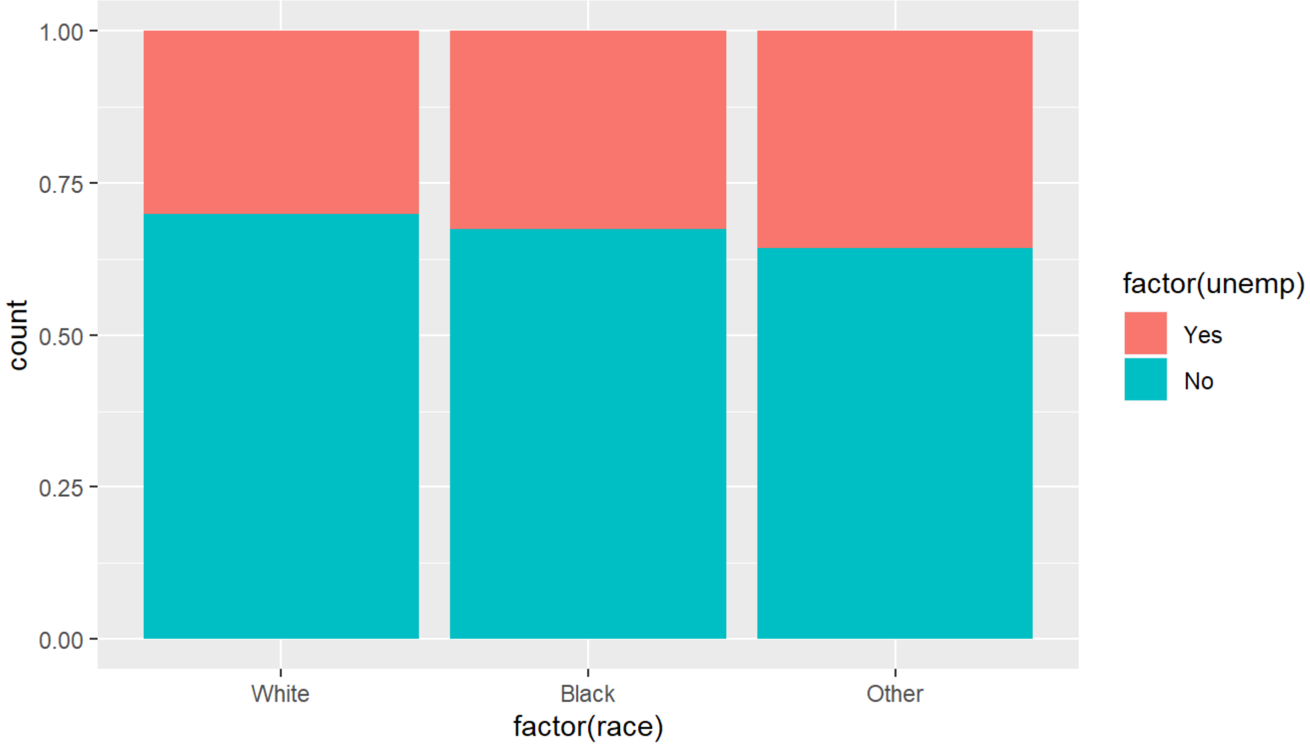


from the graph analysis, there is not significant difference for easiness to find equavalent jobs being observed between races amongst the bachelors

3. Is the rate of unemployment are associated with races amongst the bachelor degree?

```
unemp_race<- gss%>%
  select(race, unemp, degree)%>%
  filter (degree == "Bachelor", !is.na(race), !is.na(unemp) )
```

```
ggplot (unemp_race, aes (factor(race), fill=factor(unemp))) +
  geom_bar(position="fill")
```



from the graph analysis, it is likely that unemployment rate are associated with races. white race are likely to have lesser unemployment rate than other races

Part 4: Inference

1. conducting hypothesis testing. is there any relationship between different races and rate of higher education qualification?

-Hypothesis:

H0= race and higher education qualification are independent. higher education qualification do not vary with race.

Ha= race and higher education qualification are dependent. higher education qualification vary with race.

-Method to be used: chi-square independant test.

As we are testing the hypothesis testing in which involve teting 2 categorical variable with where at least one have more than 2 cateogry,chi-square independant test is the most suitable method.

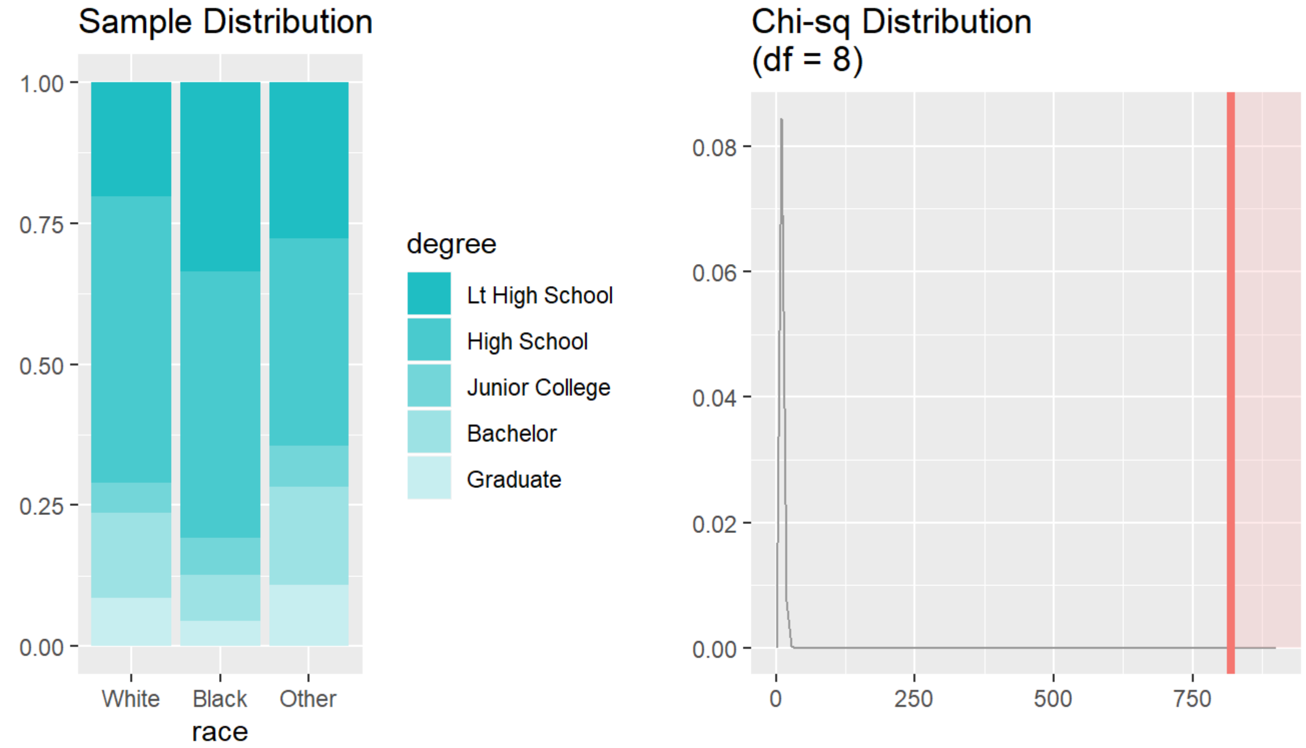
-condition check: before utilizing the method, we have ensure our samples meet the conditions.

1. were the samples randomly selected or assigned? Yes, the personal-interview is randomly selected.
2. were the sampling taken without replacement, n <10% of population? with the population of united stated was more thans 100 millions since 1972, the samples used for this hypothesis test is definetely below 10%
3. was each case only contributes to one cell in the table?yes, it cell is only contributes to one cell in the tables.
4. was each particular scenario contain more than 5 expectred cases? Yes, it is more than 5 expected cases for each of the scenario

- performing inference:

```
inference(data= educated_race, y=degree, x=race, type="ht",
          statistic = "proportion",
          success = "Bachelor", method="theoretical",
          alternative="greater")
```

```
## Response variable: categorical (5 levels)
## Explanatory variable: categorical (3 levels)
## Observed:
##        y
## x        Lt High School High School Junior College Bachelor Graduate
##   White            7339       18358           1880     5467     3103
##   Black            1936        2732            375      479      255
##   Other             525         701            138      330      206
##
## Expected:
##        y
## x        Lt High School High School Junior College  Bachelor  Graduate
##   White       8083.2558   17973.697      1973.7991 5176.5830 2939.6657
##   Black       1291.8629    2872.549       315.4518  827.3196  469.8163
##   Other        424.8813     944.754       103.7491  272.0975  154.5181
##
## H0: race and degree are independent
## HA: race and degree are dependent
## chi_sq = 817.9964, df = 8, p_value = 0
```



- interpret the result:

the p-value= 0. reject the H0 at 5% significant level in which there is relationship exist in the population in which race and higher education qualification are dependant.

2. conducting hypothesis testing. is there any relationship between different races and the easiness to find equavalent job?

-Hypothesis:

H0= race and easiness to find equavalent job are independent. easiness to find equavalent job do not vary with race.

Ha= race and easiness to find equavalent job are dependent. easiness to find job equavalent vary with race.

-Method to be used: chi-square independant test.

As we are testing the hypothesis testing in which involve teting 2 categorical variable with where at least one have more than 2 cateogry,chi-square independant test is the most suitable method.

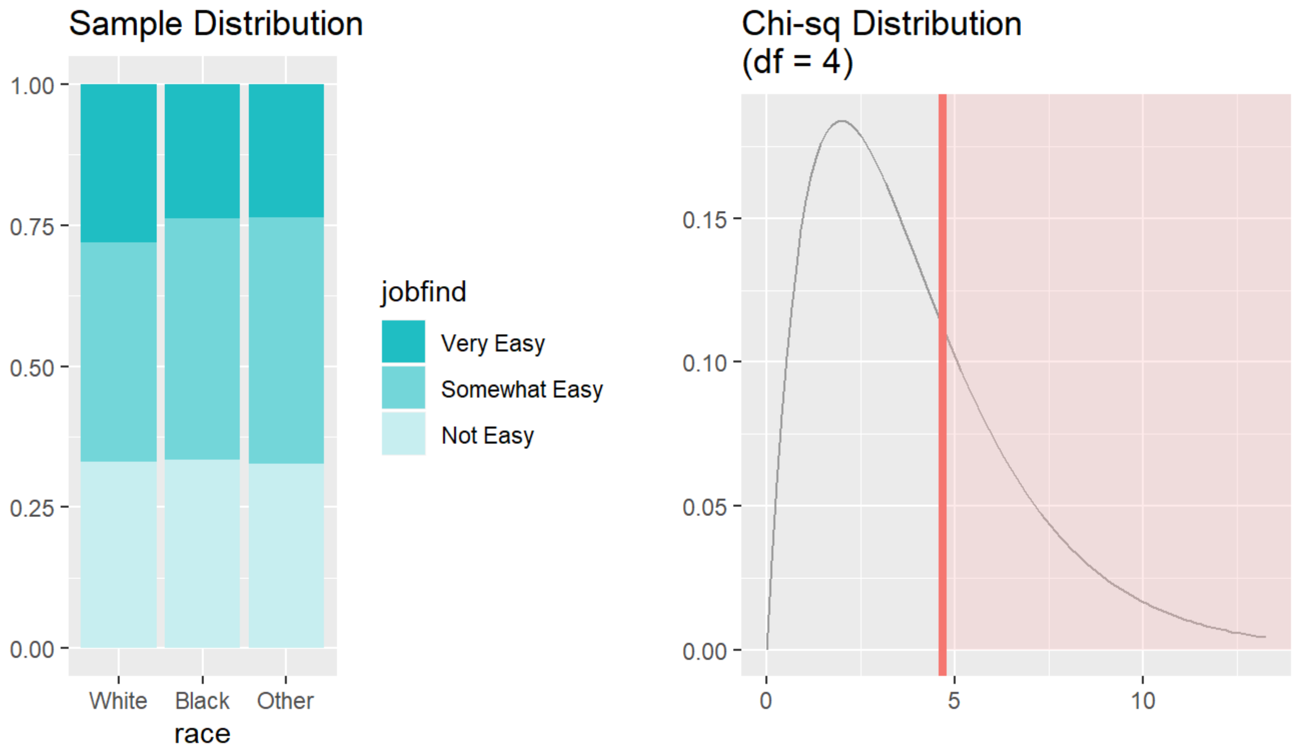
-condition check: before utilizing the method, we have ensure our samples meet the conditions.

1. were the samples randomly selected or assigned? Yes, the personal-interview is randomly selected.
2. were the sampling taken without replacement, n <10% of population? with the population of united stated was more thans 100 millions since 1972, the samples used for this hypothesis test is definetely below 10%
3. was each case only contributes to one cell in the table?yes, it cell is only contributes to one cell in the tables.
4. was each particular scenario contain more than 5 expectred cases? Yes, it is more than 5 expected cases for each of the scenario

- performing inference:

```
inference(data= jobfind_race, y=jobfind, x=race, type="ht",
          statistic = "proportion",
          success = "Very Easy", method="theoretical",
          alternative="greater")
```

```
## Response variable: categorical (3 levels)
## Explanatory variable: categorical (3 levels)
## Observed:
##        y
## x       Very Easy Somewhat Easy Not Easy
##   White       793          1094      935
##   Black        62           111       87
##   Other        47            87       65
## 
## Expected:
##        y
## x       Very Easy Somewhat Easy  Not Easy
##   White 775.81347    1111.25389 934.93264
##   Black  71.47821     102.38342  86.13837
##   Other  54.70832      78.36269  65.92899
## 
## H0: race and jobfind are independent
## HA: race and jobfind are dependent
## chi_sq = 4.6905, df = 4, p_value = 0.3206
```



- interpret the result:

the p-value= 0.3206. At 5% significant level, it is fail to reject the H0, there is no relationship exist in the population in which race and easinest to find equavalent job are independant.

3. conducting hypothesis testing. is there any relationship between different races and the unemployment?

-Hypothesis:

H0= race and unemployment are independent. unemployment do not vary with race.

Ha= race and unemployment are dependent. unemployment vary with race.

-Method to be used: chi-square independant test.

As we are testing the hypothesis testing in which involve teting 2 categorical variable with where at least one have more than 2 cateogry,chi-square independant test is the most suitable method.

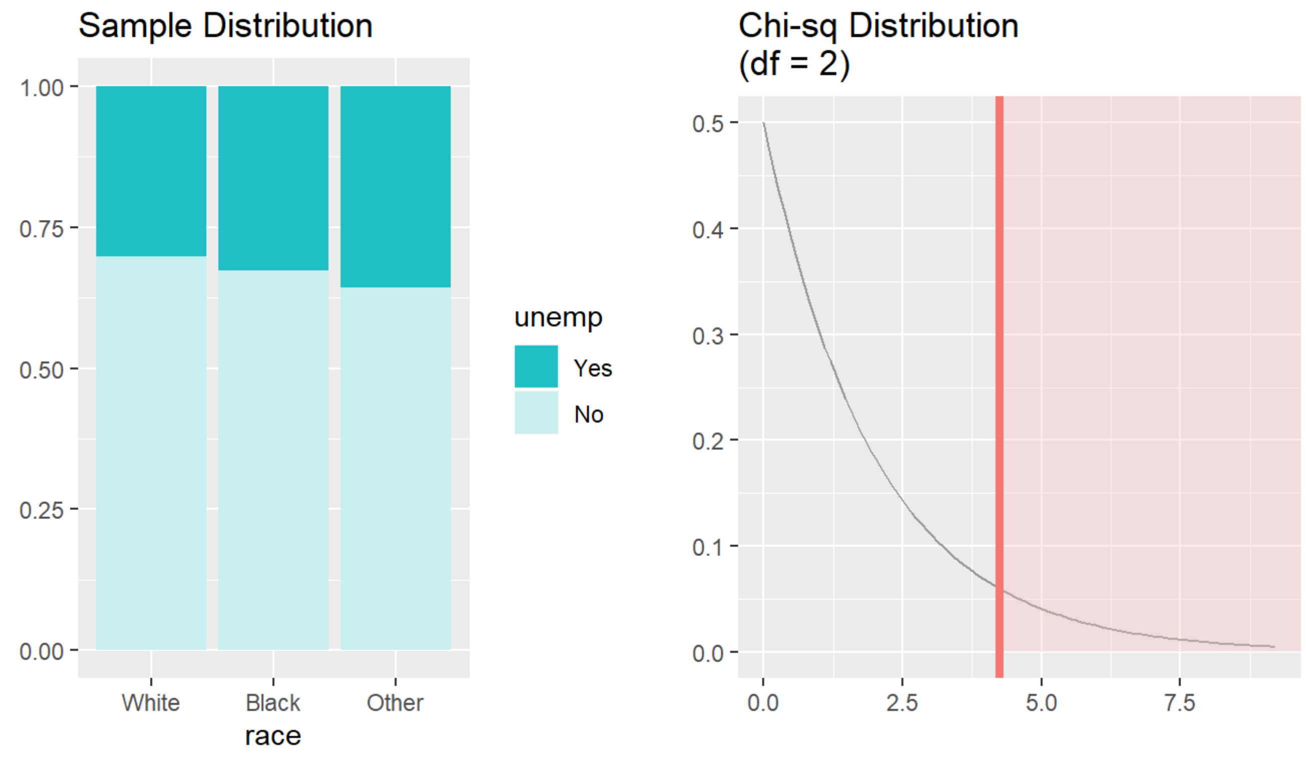
-condition check: before utilizing the method, we have ensure our samples meet the conditions.

1. were the samples randomly selected or assigned? Yes, the personal-interview is randomly selected.
2. were the sampling taken without replacement, n <10% of population? with the population of united stated was more thans 100 millions since 1972, the samples used for this hypothesis test is definetely below 10%
3. was each case only contributes to one cell in the table?yes, it cell is only contributes to one cell in the tables.
4. was each particular scenario contain more than 5 expectred cases? Yes, it is more than 5 expected cases for each of the scenario

- performing inference:

```
inference(data= unemp_race, y=unemp, x=race, type="ht",
          statistic = "proportion",
          success = "Very Easy", method="theoretical",
          alternative="greater")
```

```
## Response variable: categorical (2 levels)
## Explanatory variable: categorical (3 levels)
## Observed:
##          y
## x         Yes   No
##   White 1307 3034
##   Black  109  225
##   Other   90  162
##
## Expected:
##          y
## x               Yes        No
##   White 1326.88167 3014.1183
##   Black  102.09133  231.9087
##   Other   77.02699  174.9730
##
## H0: race and unemp are independent
## HA: race and unemp are dependent
## chi_sq = 4.2492, df = 2, p_value = 0.1195
```



- interpret the result:

the p-value= 0.1195. although from graph analysis shown that there are variation in unemployment rate across races, it is fail to reject the H0 at 5% significant level in which there is no relationship exist in the population in which race and unemployment are independant.

Conclusion:

Based on hypothesis testing, there is no significant evidence that there are inequality exist amongst different races in the america population. Given that more black races completed the higher education qualification and the rate are comparable with the white, there will be likely an equal rate of success during employment.