Hypothesis Testing

Effect of gender on unemployment rate in Australia

Vishwa Gandhi(3714805), Vikas Virani(3715555), Jigar Mangukiya (3715807)

Last updated: 28 October, 2018

RPubs link information

- Rpubs link comes here: http://rpubs.com/jigar/hypothesisTesting

Introduction

- One of the measure to check the health of country's labor market is unemployment rate.
- A lower unemployment rate signifies healthier economy, while high unemployment rate can destabilize a nation.
- In this report, we use unemployment rate percentage to statistical check equality of employment opportunities for males & females of Australia
- We will try to confirm unemployment rate does not vary by gender.

Problem Statement

- Aim - To test if unemployment rate for adults vary significantly by gender in Australia.
- We will use a Two sample t-test on our data, which, if significant, will represent that unemployment rate for males and females are significantly different.
- The findings of the reports can advocate a partial explanation, on any significant bias in employment opportunity to a particular gender.

Data

- Following two datasets are merged to form our operational dataset.
    1. Unemployment rate - female - 25-54
    2. Unemployment rate - male - 25-54
- Each dataset contains unemployment rates for 30 countries from 1981 to 2005. Our focus will be on Australian data from 1986 to 2005.
- We will compare male & female unemployment rates for Australia to check if there is a significant difference between them.
- Unemployment rate represents percentage of subjects without employment from the active & able labor force.
- labor force is restricted to 25-54 age group alternatively termed Adults.

Data ref.

- Original Data is collected by "International labor Organization" and further hosted on https://www.gapminder.org. Following are links to original data collector as well as the download link from gapminder.org.

- Original Data Collector - https://www.ilo.org/global/lang--en/index.htm

    1. Database Location - https://www.ilo.org/ilostat/faces/wcnav_defaultSelection

- Dataset downloaded from - https://www.gapminder.org/data/

    1. Female Unemployment Rate Dataset download location - https://docs.google.com/spreadsheet/pub?key=r9StWVETzyX9Lv-r4-2sh6w&output=xlsx
    2. Male Unemployment Rate Dataset download location - https://docs.google.com/spreadsheet/pub?key=rjkDFSPV2pw9Pbnz2kpiqPQ&output=xlsx

Data Cont.

- Operational Dataset will have three variables, all of which are fairly easy to comprehend. Following are attribute specifications

  1. Year [86-05] - levels - 1985,1986….2005
  2. Unemployment rate - % 0 to 100
  3. Gender [Male, Female] - levels - Male, Female

- To convert the data in a more convenient format, we have converted data into a Long format from the original wide format, by gathering year columns in both Male & Female Dataset.

- We further rowbinded both datasets, merging them into a single database, on which statistical study will be carried out.

```
Unemployement_female <- read_csv("indicator_f 25-54 unemploy.csv")
head(Unemployement_female)
```

Female 25-54 unemployment (%)	1981	1982	1983	1984	1985	1
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<
Australia	*NA*	*NA*	*NA*	*NA*	*NA*	
Canada	7.1	9.1	10.0	10.2	9.7	
Czech Rep.	*NA*	*NA*	*NA*	*NA*	*NA*	
Estonia	*NA*	*NA*	*NA*	*NA*	*NA*	
Finland	3.4	3.9	3.8	3.5	3.2	
France	7.5	7.9	8.1	8.9	9.7	

6 rows | 1-10 of 27 columns

```
Unemployement_male <- read_csv("indicator_m 25-54 unemploy.csv")
head(Unemployement_male)
```

Male 25-54 unemployment (%)	1981	1982	1983	1984	1985	198
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<db
Australia	*NA*	*NA*	*NA*	*NA*	*NA*	5
Canada	5.2	8.7	9.9	9.4	8.8	7
Czech Rep.	*NA*	*NA*	*NA*	*NA*	*NA*	N
Estonia	*NA*	*NA*	*NA*	*NA*	*NA*	N
Finland	4.1	4.5	4.6	4.0	4.4	5

Male 25-54 unemployment (%)	1981	1982	1983	1984	1985	198
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<db
France	3.8	4.3	4.6	5.6	6.2	6

6 rows | 1-10 of 27 columns

```
Unemployement_female <- Unemployement_female %>% gather(`1981`:`2005`,key = "Year",value = "Unemployment Rate")
Unemployement_male <- Unemployement_male %>% gather(`1981`:`2005`,key = "Year",value = "Unemployment Rate")

# Year %in% c('2000':'2005')

Unemployement_female_filtered <- Unemployement_female %>% filter(`Female 25-54 unemployment (%)`=='Australia' & Year

Unemployement_male_filtered <- Unemployement_male %>% filter(`Male 25-54 unemployment (%)`=='Australia' & Year > 198

Combined_data <- rbind(Unemployement_female_filtered,Unemployement_male_filtered)
Combined_data$Year <- factor(Combined_data$Year)
Combined_data$Gender <- factor(Combined_data$Gender)
head(Combined_data)
```

Year	Unemployment Rate	Gender
<fctr>	<dbl>	<fctr>
1986	6.4	Female
1987	6.4	Female
1988	6.0	Female
1989	5.2	Female
1990	5.4	Female
1991	7.0	Female

6 rows

Descriptive Statistics

- Following table gives a quick summary of data
- The mean unemployment rates for male and females are not exactly same. We will check if the difference in the mean is significant or insignificant in the following parts of reports.

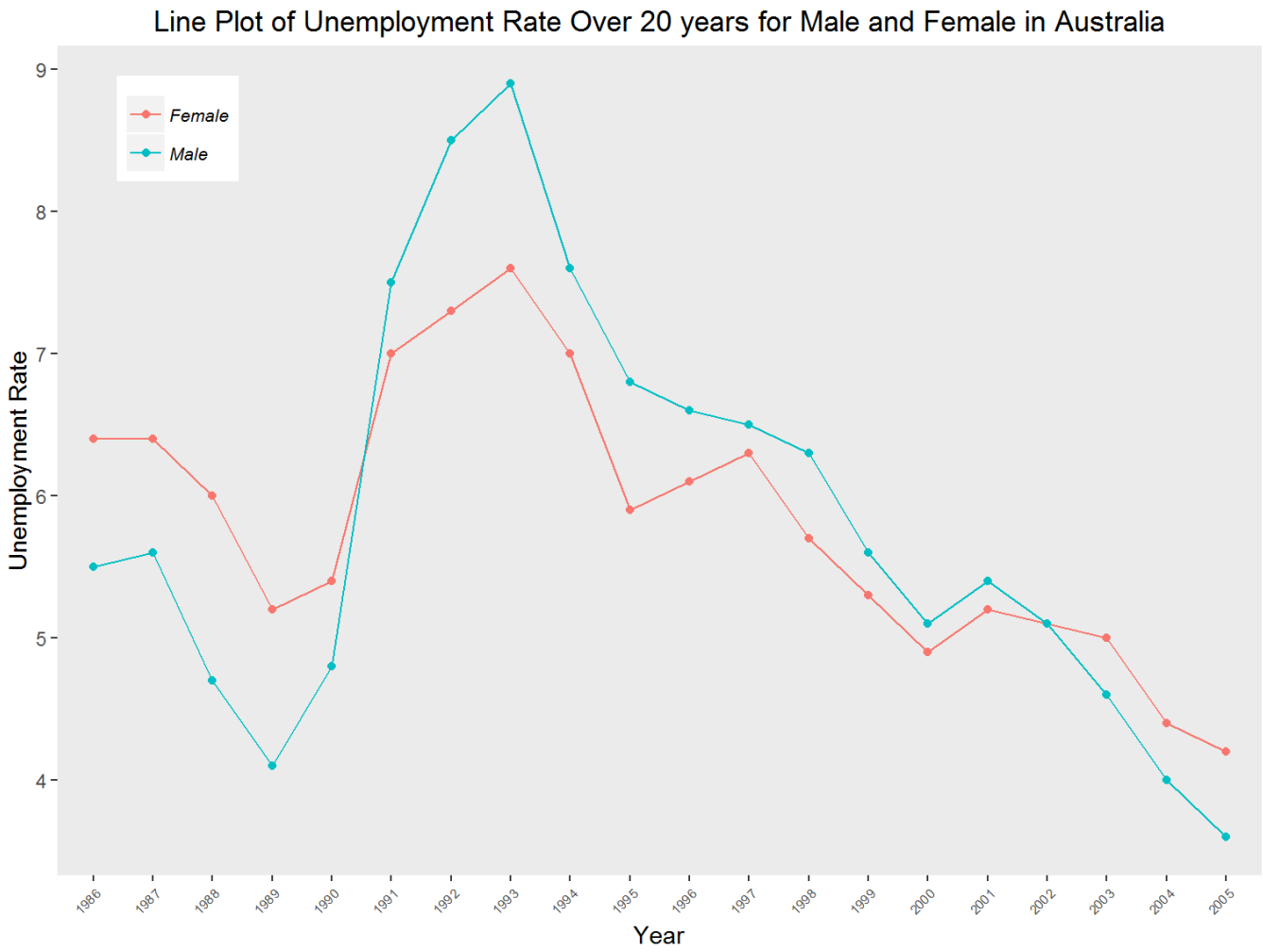
```
Combined_data %>% group_by(Gender) %>% summarise(Min = min(`Unemployment Rate`,na.rm = TRUE),
                                                 Q1 = quantile(`Unemployment Rate`,probs = .25,na.rm = TRUE),
                                                 Median = median(`Unemployment Rate`, na.rm = TRUE),
                                                 Q3 = quantile(`Unemployment Rate`,probs = .75,na.rm = TRUE),
                                                 Max = max(`Unemployment Rate`,na.rm = TRUE),
                                                 Mean = mean(`Unemployment Rate`, na.rm = TRUE),
                                                 SD = sd(`Unemployment Rate`, na.rm = TRUE),
                                                 n = n(),
                                                 Missing = sum(is.na(`Unemployment Rate`))) -> table
knitr::kable(table)
```

Gender	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
Female	4.2	5.175	5.80	6.40	7.6	5.82	0.9479063	20	0
Male	3.6	4.775	5.55	6.65	8.9	5.84	1.4744847	20	0

Descriptive Statistics Visualization

- The following code generates a line chart for time series analysis.
- The chart indicates an alternate relative increase and drop in both Male & Female Unemployment rates in Australia for over years 1986 to 2005.
- For the middle part or middle 10 years, Male unemployment rate was higher than Female unemployment rate. while in beginning & ending 5 years of observation period, the trend was reversed.

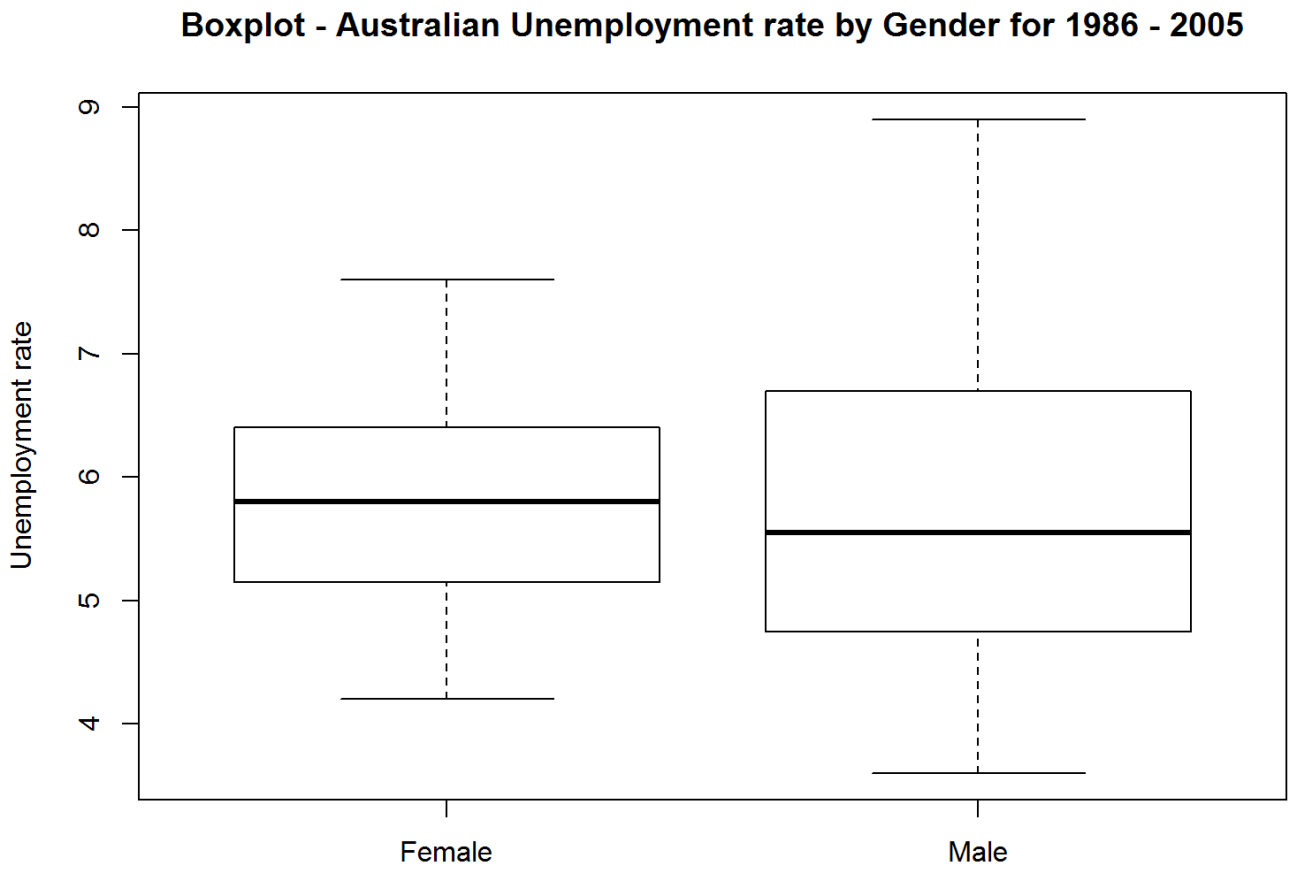
```
ggplot(data=Combined_data,
 aes(x=Year, y=`Unemployment Rate`,group=Gender,colour=Gender)) +
 geom_line() +geom_point() + ggtitle("Line Plot of Unemployment Rate Over 20 years for Male and Female in Australia")

 theme(axis.text.x=element_text(size=6, angle=45, vjust=1, hjust=1),
 panel.grid.major.x=element_blank(),
 panel.grid.minor.x=element_blank(),
 panel.grid.minor.y=element_blank(),
 panel.grid.major.y=element_blank()) +
 theme(legend.text = element_text(size=8, face="italic"),
 legend.title = element_blank(),
 legend.position=c(0.1, 0.9))
```


Line Plot of Unemployment Rate Over 20 years for Male and Female in Australia
Female
Male
Unemployment Rate
9
8
7
6
5
4
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Year

Descriptive Statistics Cont.

- The following boxplot suggests absence of outliers in the data.
- Male unemployment rate can be seen to vary in a broader range than female unemployment rate.

```
Combined_data %>% boxplot(`Unemployment Rate` ~ Gender, data = .,ylab="Unemployment rate", main = "Boxplot - Austral
```



Hypothesis Testing

- We will use two sample t-test to compare the means of our two groups, so that we can conclude if there is statistically significant difference in the two means.

Null Hypothesis - There is no statistically significant difference in the mean unemployment rates of males and females

$$H_0 : \mu_1 - \mu_2 = 0$$

Alternate hypothesis: There is a statistically significant difference in the mean unemployment rates of males and females

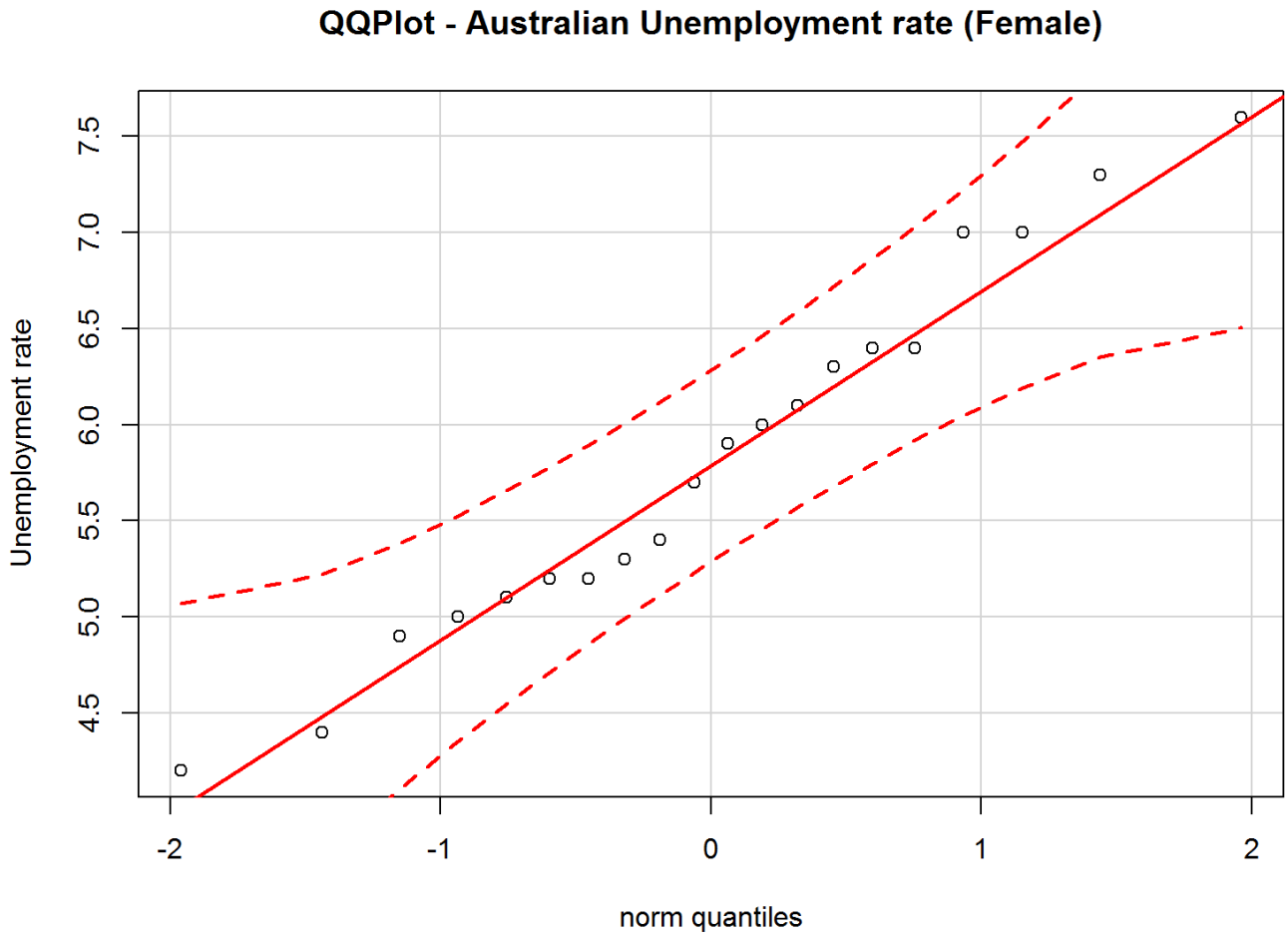
$$H_a : \mu_1 - \mu_2 \neq 0$$

- Two sample t-test expects following characteristics of data
    1. Independence
    2. Normality of data
- We will also need to check for Equal variance for parametric adaption
- Data subgroups being compared here (Male & Female) are intuitively independent of each other, so the independence can be assumed.

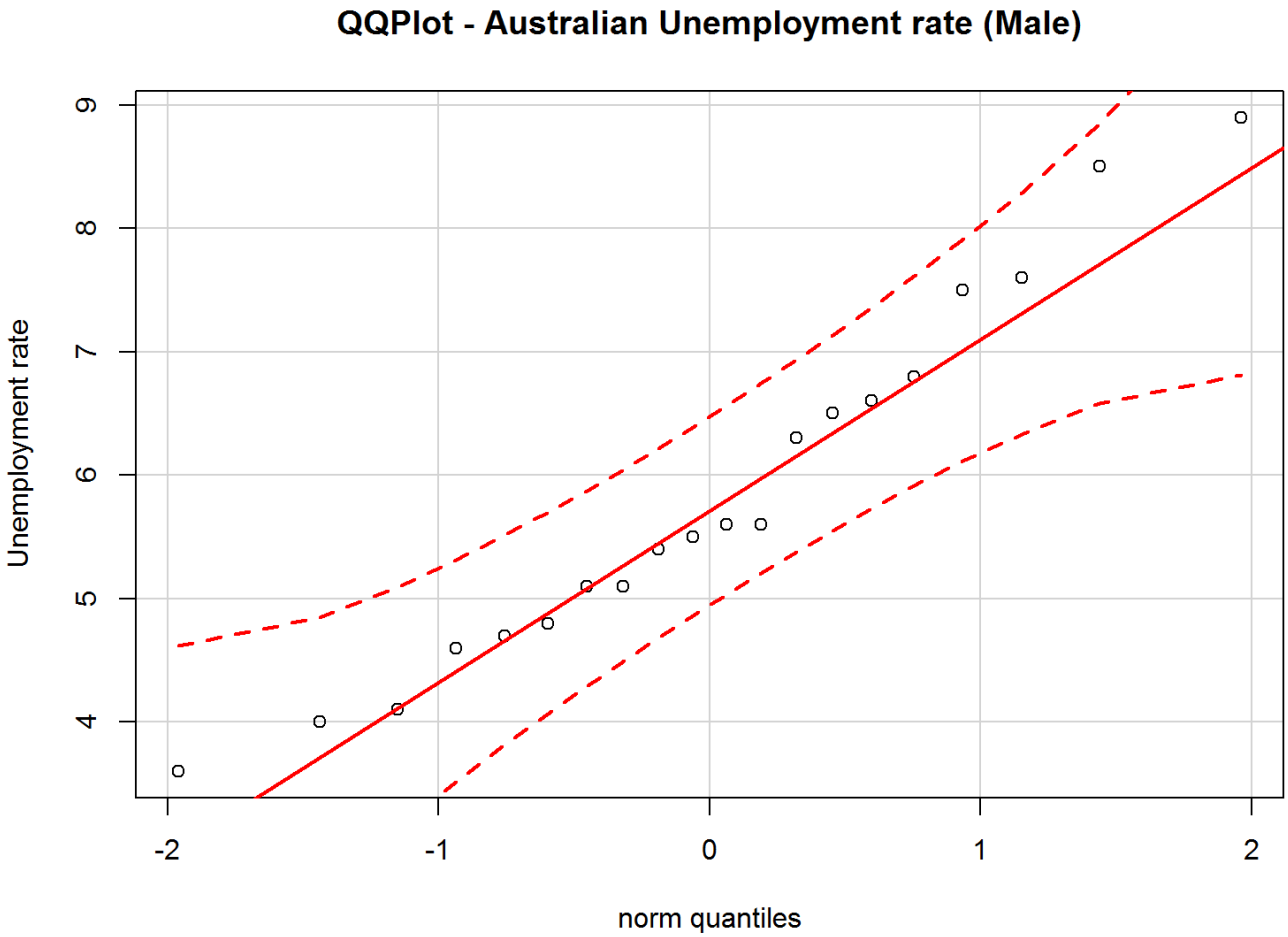
Hypothesis Testing - Checking Normality

- The QQ Plots suggests that all the observations fall within 95% CI of normal distribution. Which is important as the sample size < 30, which restricts us from applying Center Limit Theorem (CLT).
- Normality can be considered and t-test can be performed.

```
Unemployement_female_filtered$`Unemployment Rate` %>% qqPlot(dist="norm",ylab="Unemployment rate", main = "QQPlot - A
```




```
Unemployement_male_filtered$`Unemployment Rate` %>% qqPlot(dist="norm",ylab="Unemployment rate", main = "QQPlot - Aus
```


QQPlot - Australian Unemployment rate (Male)
Unemployment rate
norm quantiles
4
5
6
7
8
9
-2
-1
0
1
2

Hypothesis Testing - Checking equal variance

- Null Hypothesis - There is a statistically significant difference in the variance of unemployment rates of males and females.

$$H_0 : \sigma_1 = \sigma_2$$

- Alternate hypothesis - There is no statistically significant difference in the variance of unemployment rates of males and females.

$$H_a : \sigma_1 \neq \sigma_2$$

- According to the results of levene test, p value is 0.127, which is higher than 0.05.
- The results suggests that equal variance can not be assumed.

```
leveneTest(Combined_data$`Unemployment Rate` ~ Combined_data$Gender)
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	2.425042	0.1277005
	38	*NA*	*NA*

2 rows

Hypothesis Testing Cont.

- As per the result of the levene test, we will perform two sample t-test with unequal variance. We will perform two sided test for 95% CI.

```
t.test(
  `Unemployment Rate` ~ Gender,
  data = Combined_data,
  var.equal = FALSE,
  alternative = "two.sided"
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  Unemployment Rate by Gender
## t = -0.051026, df = 32.414, p-value = 0.9596
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8179941  0.7779942
## sample estimates:
## mean in group Female   mean in group Male
##                 5.82                 5.84
```

Discussion

- As per the results of welch two sample t-test, the p-value is 0.960 (rounded), & 95% CI for difference in means is [-0.051,0.778].
- t value is -0.051, which falls comfortably within the 95% CI & p value is > 0.05.
- The test results can be termed as insignificant & we fail to reject the null hypothesis.
- By which, we can, in the limited scope of the current data context, conclude that unemployment rate in Australia for adults is not significantly affected by gender.
- This investigation can be further enhanced by

    1. considering data for several countries, to check if the result is same for all the countries or they differ by the economic state (first world, third world etc) of country.
    2. considering other factors that affect unemployment rate of a country. e.g. Economic downturn & financial crunches due to some natural or manmade events like calamities or wars. This can be used to adjust the rate to get a more accurate result.