Votes on US 2016 presidential race

```r
library(tidyverse) # to be able to use data visual.tools
library(foreign) #reading data in diff.formats
library(ggplot2) # for plots
library(ggrepel) #ggrepel provides geoms for ggplot2
```

Hypothesis Testing

The data set **Votes.csv** comprises data on US 2016 presidential race. The variables are **County:** 159 Georgia counties, **Diversity.Index:** Diversity index, **Percent.Clinton:** Percent voted for Clinton, **Percent.Trump:** Percent voted for Trump.
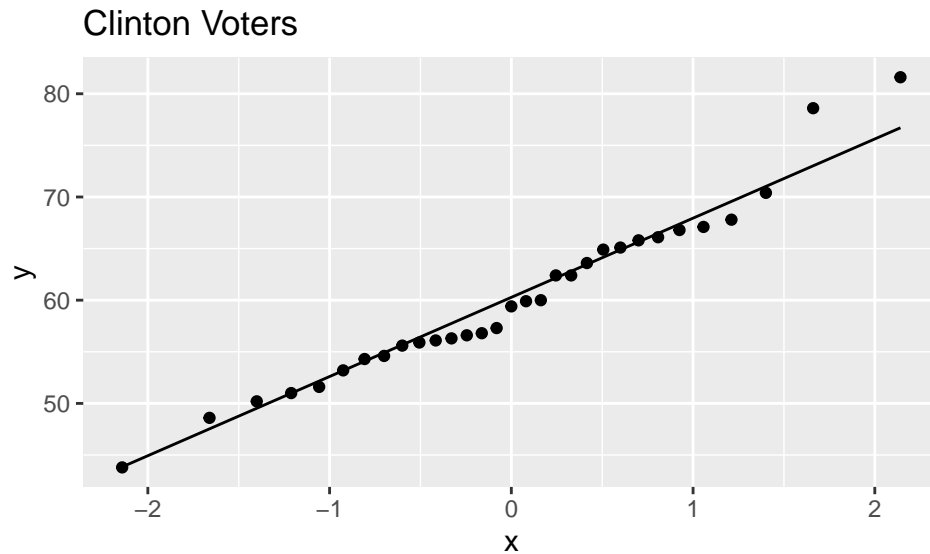
Normal distribution of diversity index.

We will read the data and create a new variable **vote**, indicating if the county voted for Clinton or Trump:

```r
#the procedure for df selection and modification as in 2 ex.
vot <- read.csv("Votes.csv")
votes <- vot[,c("County", "Diversity.Index", "Percent.Clinton", "Percent.Trump")] %>%
  mutate(vote = ifelse(Percent.Clinton > Percent.Trump, "Clinton", "Trump"))
```

We would like to compare diversity indices in both groups. We check that the diversity index in each group is normally distributed using QQ-Plots:

```r
#For Clinton
ggplot(filter(votes, vote == "Clinton"), aes(sample = Diversity.Index)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Clinton Voters")
```

Clinton Voters


```
#For Trump
ggplot(filter(votes, vote == "Trump"), aes(sample = Diversity.Index)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Trump Voters")
```

Trump Voters
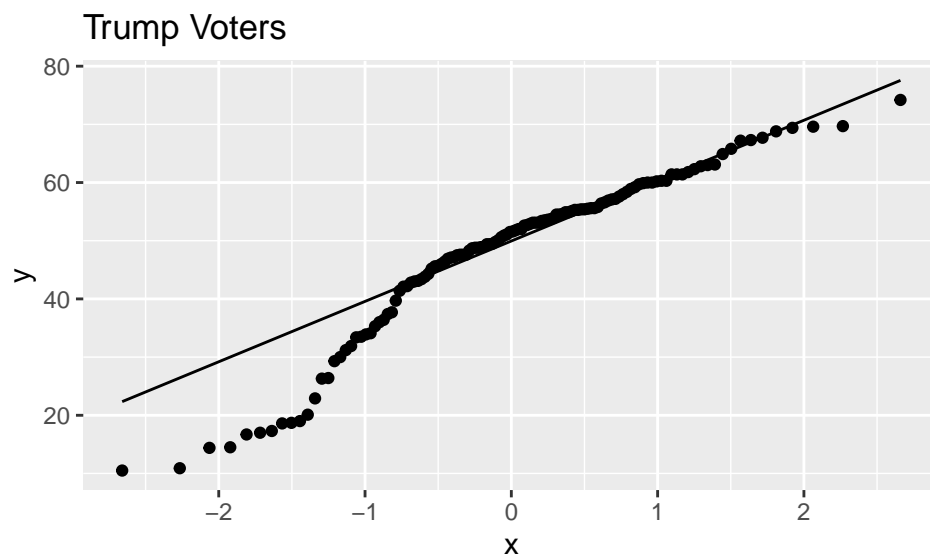
It seems that neither group's diversity index appears to be normally distributed.

t-test vs. Wilcoxon test

Now we test with an appropriate t-test at the significance level $\alpha = 0.05$ if the diversity index in the group that voted for Clinton is significantly larger than that in the group that voted for Trump.

A t-test assumes normality within each group, which is not the case here. I also have to test for equal variances:

```
clinton <- filter(votes, vote == "Clinton")
trump <- filter(votes, vote == "Trump")
var.test(clinton$Diversity.Index, trump$Diversity.Index)
```

```
##
##  F test to compare two variances
##
## data:  clinton$Diversity.Index and trump$Diversity.Index
## F = 0.35265, num df = 30, denom df = 127, p-value = 0.001501
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2095270 0.6567099
## sample estimates:
## ratio of variances
##           0.3526528
```

The variances are not equal! Therefore, I will use the Welch t-test.

```
t.test(clinton$Diversity.Index, trump$Diversity.Index, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  clinton$Diversity.Index and trump$Diversity.Index
## t = 6.1652, df = 76.799, p-value = 3.035e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.078487 15.786987
## sample estimates:
## mean of x mean of y
##  60.12258  48.18984
```

As we see from results, we conclude that the means are not equal.

Now, on to the Wilcoxon test:

```
wilcox.test(clinton$Diversity.Index, trump$Diversity.Index)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clinton$Diversity.Index and trump$Diversity.Index
## W = 3048.5, p-value = 3.729e-06
## alternative hypothesis: true location shift is not equal to 0
```
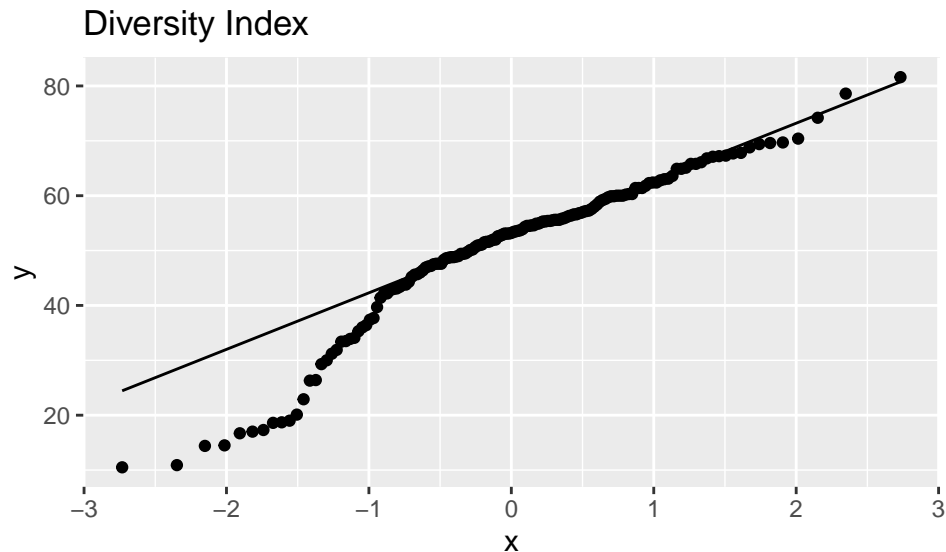
We conclude that the group means are not equal again.

The center of distribution of Diversity.Index

Now we would like to test if the center of distribution of Diversity.Index equals to 50.

Our null hypothesis is: The mean of the diversity index is 50.

```r
ggplot(votes, aes(sample = Diversity.Index)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Diversity Index")
```

Diversity Index

The diversity index is not distributed normally, which would be necessary for a t-test. I will go ahead and compute it anyway:

```r
t.test(votes$Diversity.Index, mu = 50)
```

```
##
##  One Sample t-test
##
## data:  votes$Diversity.Index
## t = 0.46929, df = 158, p-value = 0.6395
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  48.34319 52.68952
## sample estimates:
## mean of x
##  50.51635
```

As we observe from results, we can not reject the null hypothesis.

```r
wilcox.test(votes$Diversity.Index, mu = 50)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  votes$Diversity.Index
## V = 7569.5, p-value = 0.0376
## alternative hypothesis: true location is not equal to 50
```

When we use the Wilcoxon-test we do reject the null hypothesis. The Wilcoxon test does not assume normality and should therefore be more accurate.