Data-X

Katherine Zhang

11/26/2018

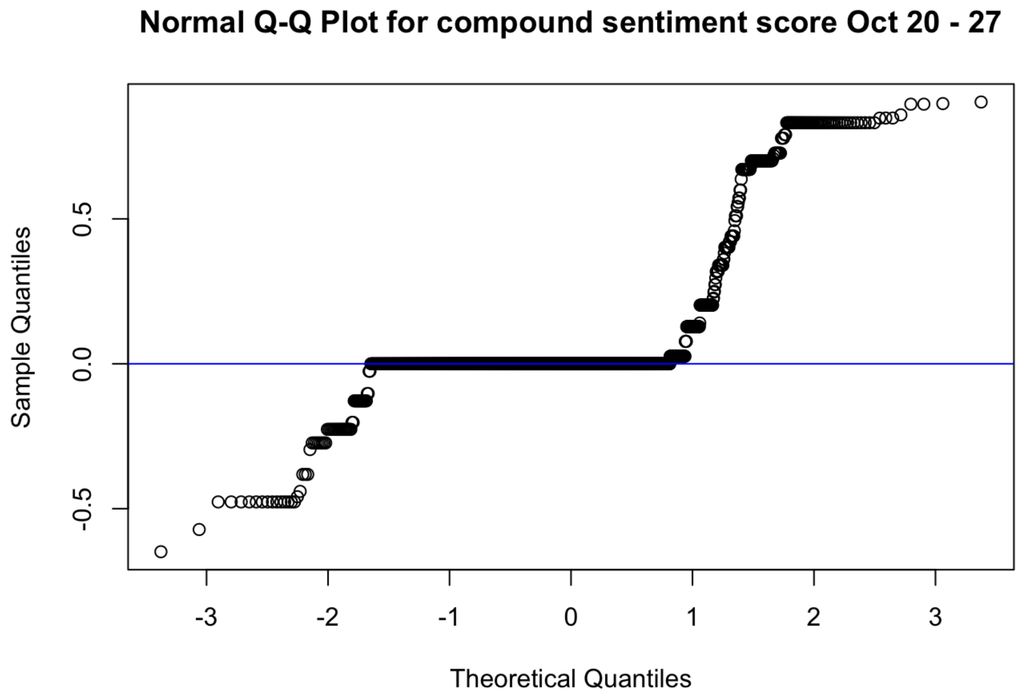
```
library(ggplot2)
library(DataComputing)
library(dplyr)
library(ggpubr)
```

Read the data into table. x1 - 1358 tweets data from Oct 20 to Oct 27. x2 - 2280 tweets data from Nov 1 to Nov 9.

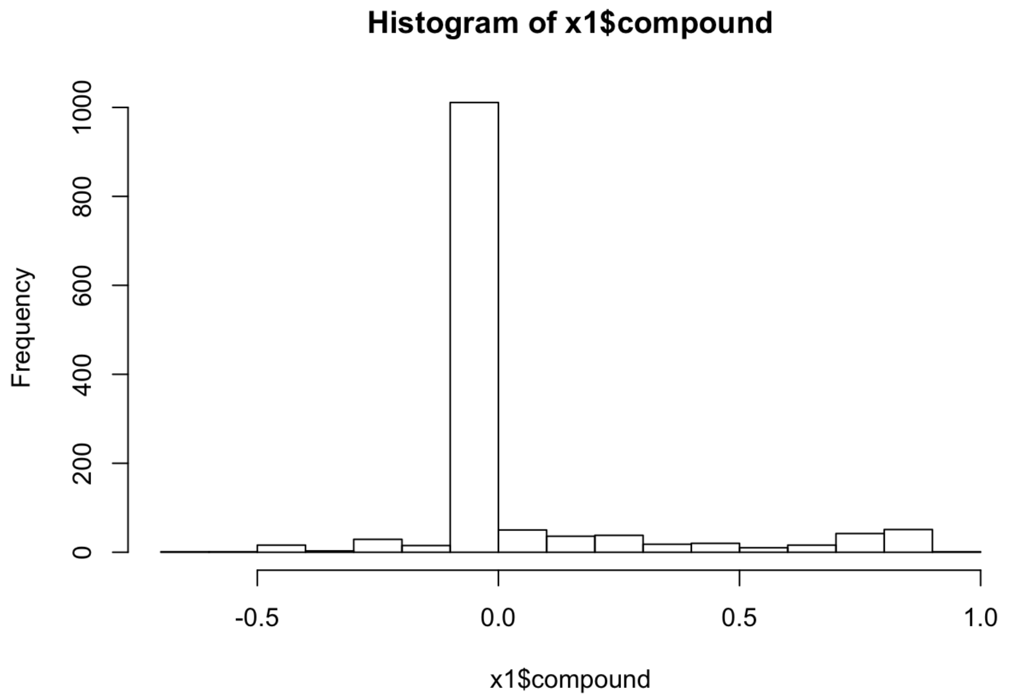
```
x1 <- read.csv("x1.txt", header = TRUE)
x2 <- read.csv('x2.txt', header = TRUE)
```

We plot the QQ-normal plot and the gaussian distribution of x1’s and x2’s compound and text_len to check data normality.

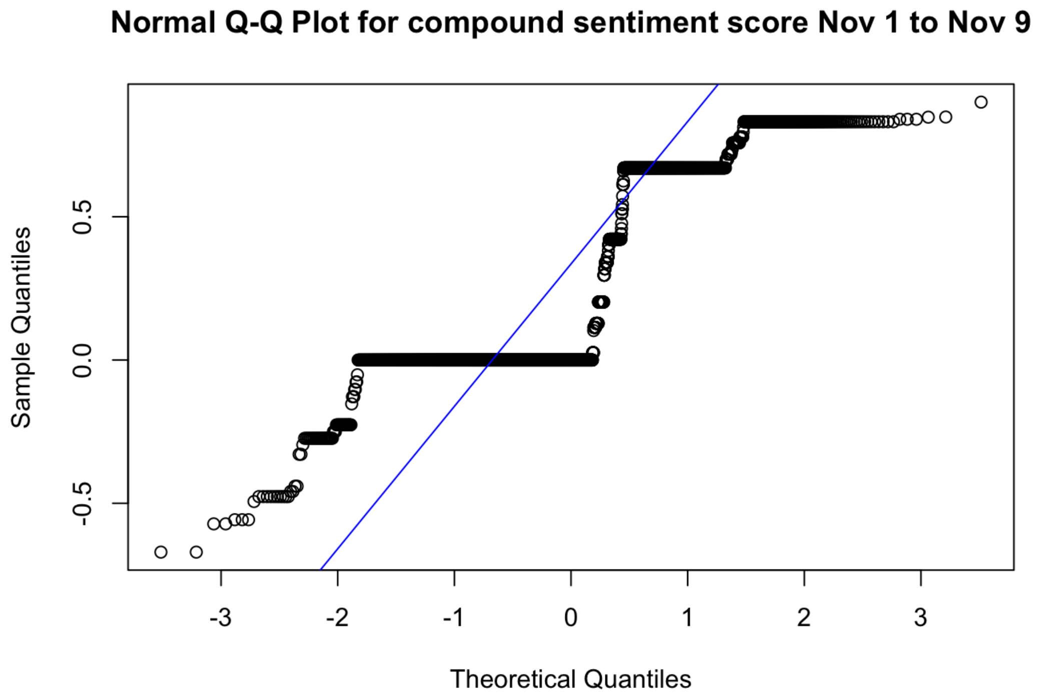
```
qqnorm(x1$compound, main = 'Normal Q-Q Plot for compound sentiment score Oct 20 - 27');
qqline(x1$compound, col = 'blue')
```



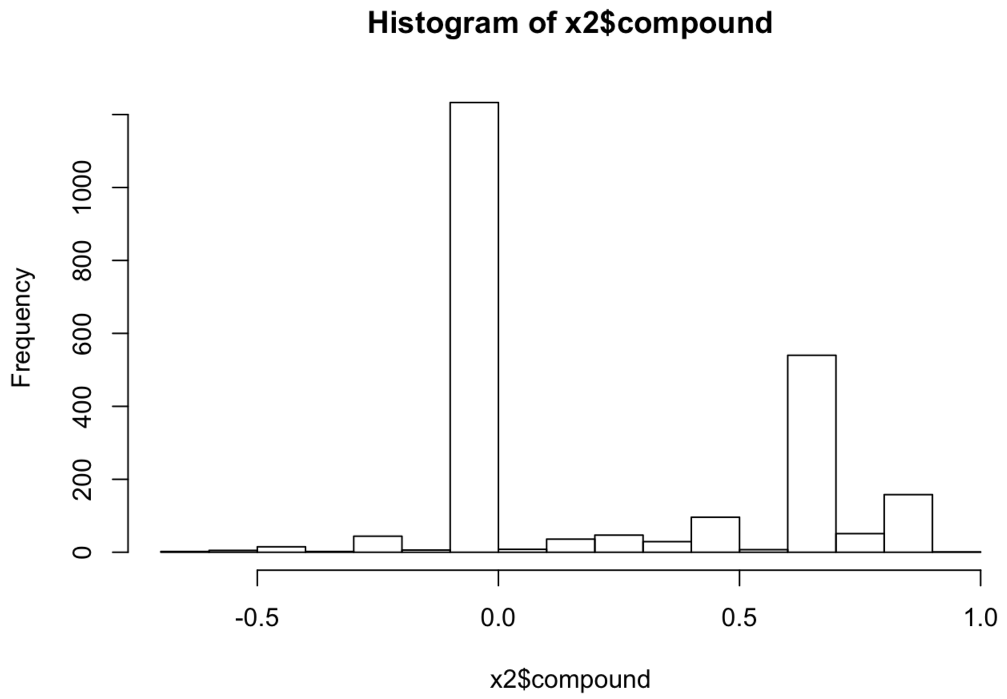
```
hist(x1$compound)
```



```
qqnorm(x2$compound, main = 'Normal Q-Q Plot for compound sentiment score Nov 1 to Nov 9');
qqline(x2$compound, col = 'blue')
```

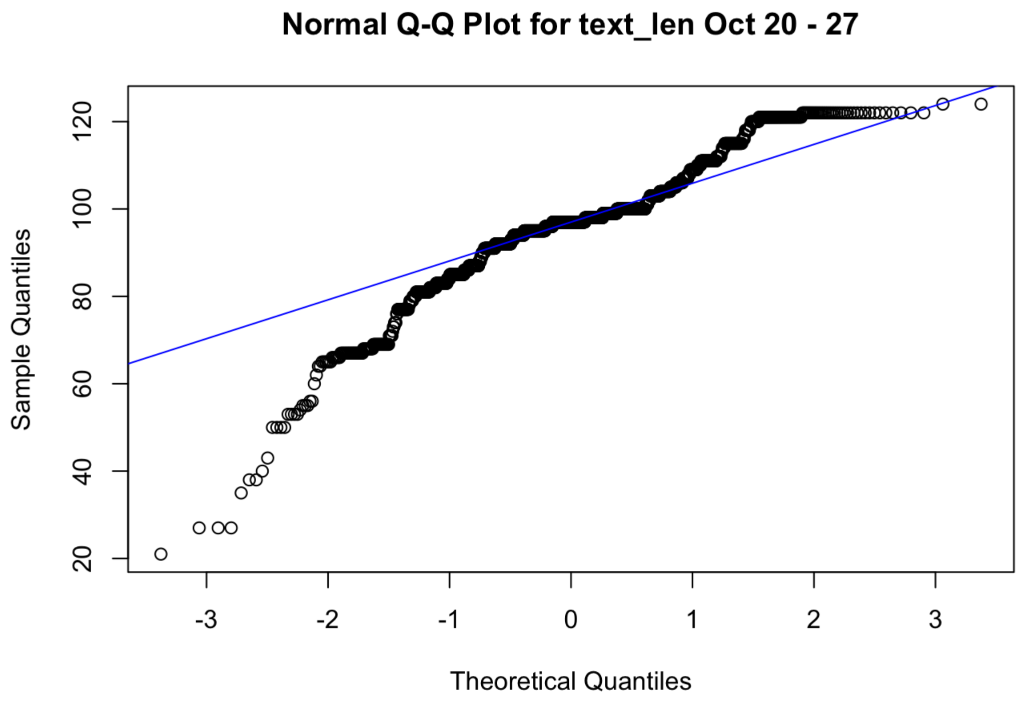


```
hist(x2$compound)
```

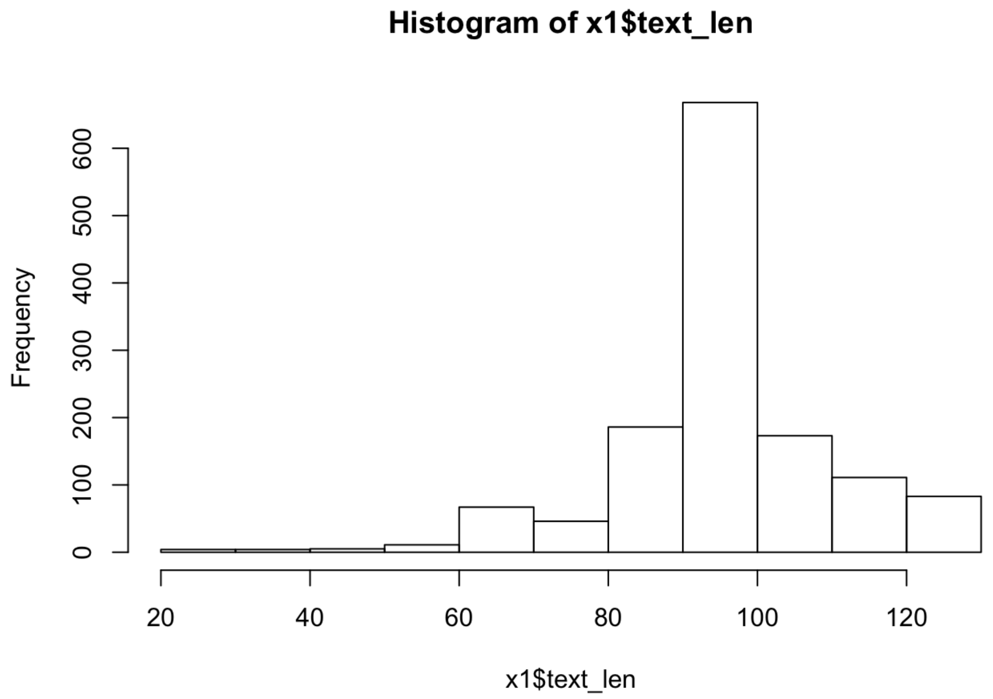


For x1, we can see a nonlinear distribution with a spike of identical values: -1, 0, 1. For x2, we observe more positive identical values: 1, 2, 3. Thus, the assumption of data normality for both distribution is invalidate here.

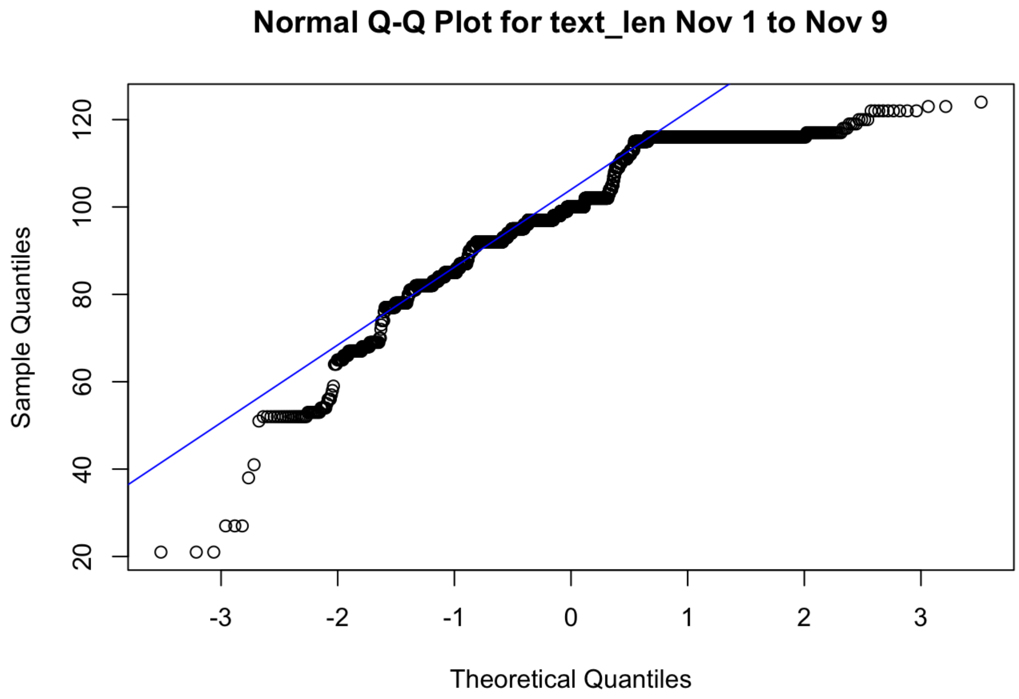
```
qqnorm(x1$text_len, main = 'Normal Q-Q Plot for text_len Oct 20 - 27');
qqline(x1$text_len, col = 'blue')
```



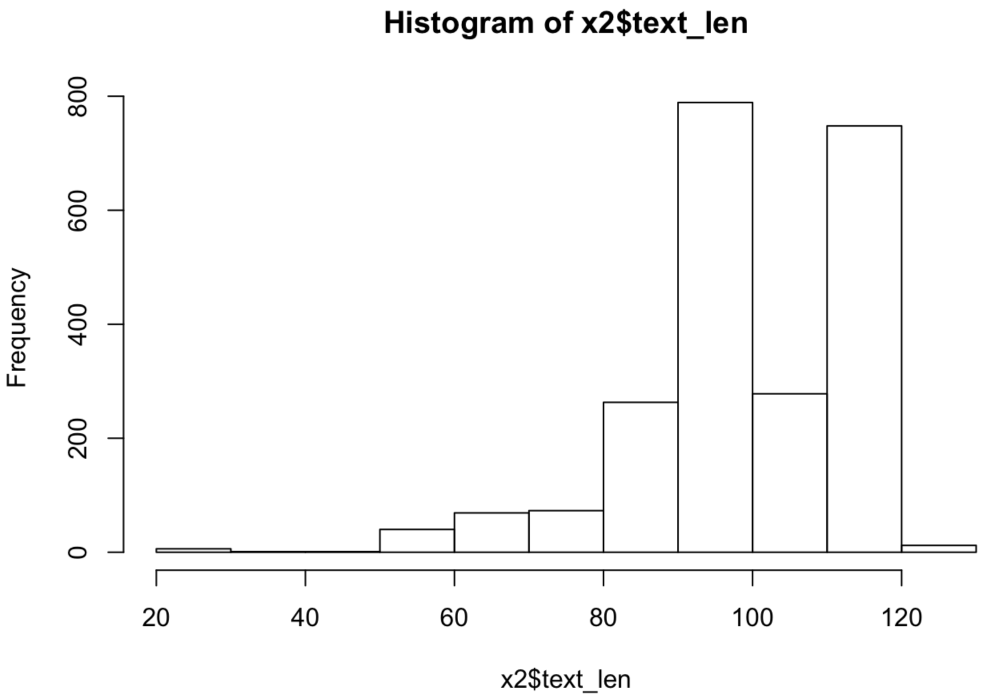
```
hist(x1$text_len)
```



```
qqnorm(x2$text_len, main = 'Normal Q-Q Plot for text_len Nov 1 to Nov 9');
qqline(x2$text_len, col = 'blue')
```

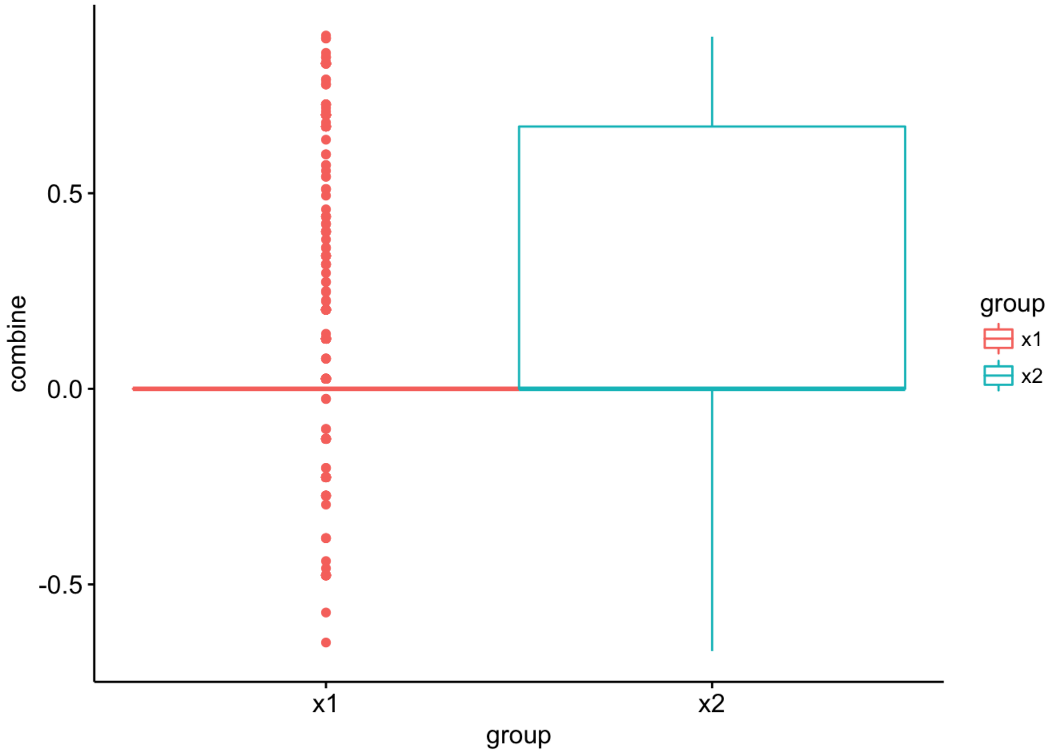


```
hist(x2$text_len)
```



Both of the distributions for tweets length is skewed to the left, so the assumption of data normality for both distribution is invalidate as well.

```
combine <- c(x1$compound, x2$compound)
group <- c(rep('x1', 1358), rep('x2', 2280))
d1 <- data.frame(combine, group)
d1 %>% ggboxplot(x = 'group', y = 'combine', color = 'group')
```



Our hypothesis for the sentiment compound score from Oct 20 to 27(noted as period x1) and from Nov 1 to Nov 9(noted as period x2) are stated as follows.

Null: There is no statistical significant difference between the distribution of compound score in x1 and x2 that they have the same overall distribution. Any deviation is due to chance alone.

Alternative: There exists a statistical significant difference between the distribution of compound score in x1 and x2

We will perform nonparametric 2 sample test – ‘Rank Sum Test’, also called Wilcox test in R, since two samples are randomly and independently selected from the populations and their distributions are not normal.

```r
wilcox.test(x1$compound, x2$compound, alternative = 'two.sided')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x1$compound and x2$compound
## W = 1168700, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(x1$text_len, x2$text_len, alternative = 'two.sided')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x1$text_len and x2$text_len
## W = 1290600, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Given that p-values are extremely small in two cases, we reject the null and conclude that sentiment compound score and tweets length in two different period, Oct 20-27 and Nov 1-9 are statistically significantly different.

We can also construct a 95% confidence interval for sentiment compound score ad tweets length in these two period.

```r
x1_compound_mean <- mean(x1$compound)
x2_compound_mean <- mean(x2$compound)
x1_compound_sd <- sd(x1$compound)
x2_compound_sd <- sd(x2$compound)

x1_margin_error <- qnorm(0.975) * x1_compound_mean/sqrt(1358)
x2_margin_error <- qnorm(0.975) * x2_compound_mean/sqrt(2280)

x1_compound_mean + x1_margin_error
```

```
## [1] 0.07737878
```

```r
x1_compound_mean - x1_margin_error
```

```
## [1] 0.06956349
```

```r
x2_compound_mean + x2_margin_error
```

```
## [1] 0.2634249
```

```r
x2_compound_mean - x2_margin_error
```

```
## [1] 0.242652
```

95% CI for sentiment compound score in x1: [0.0696, 0.0774]

95% CI for sentiment compound score in x2: [0.263, 0.243]

We observe a sligtly more positive sentiment compound score in x2 period.

```r
x1_len_mean <- mean(x1$text_len)
x2_len_mean <- mean(x2$text_len)
x1_len_sd <- sd(x1$text_len)
x2_len_sd <- sd(x2$text_len)

x1_margin_error <- qnorm(0.975) * x1_len_mean/sqrt(1358)
x2_margin_error <- qnorm(0.975) * x2_len_mean/sqrt(2280)

x1_len_mean + x1_margin_error
```

```
## [1] 101.0058
```

```
x1_len_mean - x1_margin_error
```

```
## [1] 90.80419
```

```
x2_len_mean + x2_margin_error
```

```
## [1] 103.7805
```

```
x2_len_mean - x2_margin_error
```

```
## [1] 95.59669
```

95% CI for sentiment compound score in x1: [90.804, 101.006]

95% CI for sentiment compound score in x2: [95.597, 103.781]

Next we perform linear regression analysis. We regress sentiment compound score of each tweet on its length to see if there exists a linear relationship between explanatory variable(tweet length) and respnose variable(sentiment compound score of tweet).

In period x1, since the p-value for x coefficient is smaller than 0.05 and the x coefficient is negative, we can conclude that there exists a negative relationship between tweet length and compound score of tweet that the linear regression model can be written as: y = -0.015x + 0.2169.

```
summary(lm(compound ~ text_len, data = x1))
```

```
##
## Call:
## lm(formula = compound ~ text_len, data = x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75483 -0.07632 -0.06884 -0.04491  0.85827
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2169136  0.0436099   4.974  7.4e-07 ***
## text_len    -0.0014957  0.0004498  -3.325 0.000908 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2353 on 1356 degrees of freedom
## Multiple R-squared:  0.008087,   Adjusted R-squared:  0.007356
## F-statistic: 11.06 on 1 and 1356 DF,  p-value: 0.0009076
```

However in period x2, we observe a positive relationship between tweet length and compound score of tweet that the linear regression model can be written as: y = 0.01x - 0.803

```
summary(lm(compound ~ text_len, data = x2))
```

```
##
## Call:
## lm(formula = compound ~ text_len, data = x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04681 -0.23515 -0.06564  0.24465  0.78476
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8030975  0.0427183   -18.8   <2e-16 ***
## text_len     0.0105944  0.0004237    25.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3044 on 2278 degrees of freedom
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.215
## F-statistic: 625.2 on 1 and 2278 DF,  p-value: < 2.2e-16
```