E80 Data Analysis: Temperature at Second Depth

Elizabeth Poss

April 26, 2017

Load, separate data into desired vectors.

```
# load data
location1temps = read.csv("~/Downloads/Temp Data Analysis - Depth 2.csv")

# identify columns
colnames(location1temps) = c("t1145","t1","t145", "t210", "t245")

# assign vectors for desired times- isolate nonzero values from matrix
# exported from matlab
t1145 = (location1temps$t1145[1:163])
t1 = (location1temps$t1[1:111])
t145 = (location1temps$t145[1:178])
t210 = (location1temps$t210[1:192])
t245 = (location1temps$t245[1:191])
```

Means and basic plot.

Next we average the data sets, and plot the temperature voltage over time of day.

```
mean(t1145)
```

```
## [1] 0.7528829
```

```
mean(t1)
```

```
## [1] 0.7438777
```

```
mean(t145)
```

```
## [1] 0.8344381
```

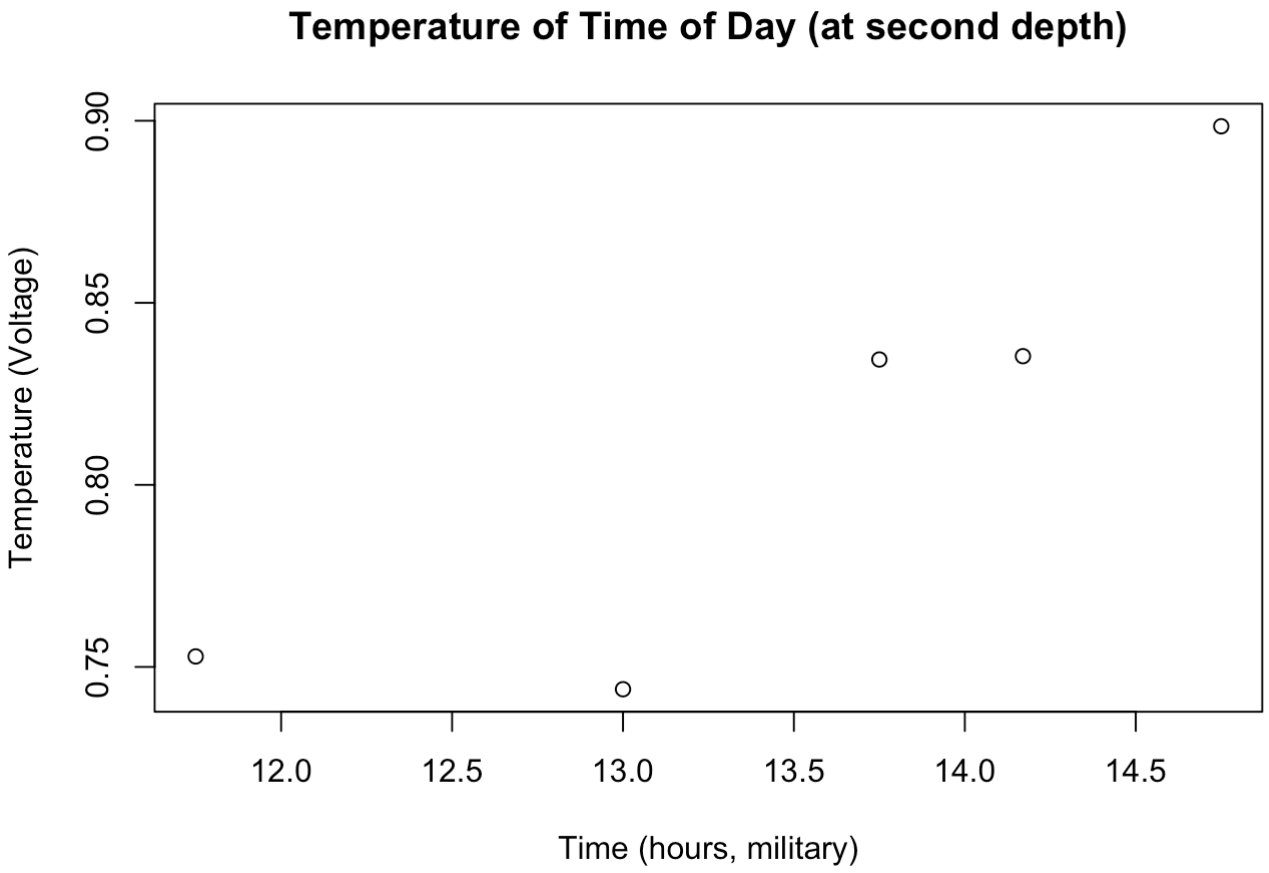
```
mean(t210)
```

```
## [1] 0.8353527
```

```
mean(t245)
```

```
## [1] 0.8984849
```

```
averages = c(mean(t1145),mean(t1),mean(t145),mean(t210),mean(t245))
avgtimes = c(11.75, 13, 13.75, 14.17, 14.75)
plot(avgtimes,averages,main = "Temperature of Time of Day (at second depth)", ylab='Temp
erature (Voltage)', xlab='Time (hours, military)')
```



Next we check for normality- the ANOVA test is ideal for comparing more than two data sets, but makes three assumptions. First, that the data is randomly sampled from the population. We sampled at a consistent 10Hz over data with random noise, so the data would be consistently varying in a random manner. Second, the data must have similar variances. For the ANOVA test, this means variances within an order of magnitude. Third, the data must be normally distributed. However, the ANOVA is relatively strongly resistant for variations from normality, in the case of large sample sizes (which we have here). Method: We calculate and compare the variance of each set. For normality, we use a qq-plot. In this display, the data is graphed as points above a line representing perfect normality. The closer the data is to the line, the better its level of normality.

```
var(t1145)
```

```
## [1] 9.396736e-05
```

```
var(t1)
```

```
## [1] 0.004695632
```

```
var(t145)
```

```
## [1] 6.53981e-05
```

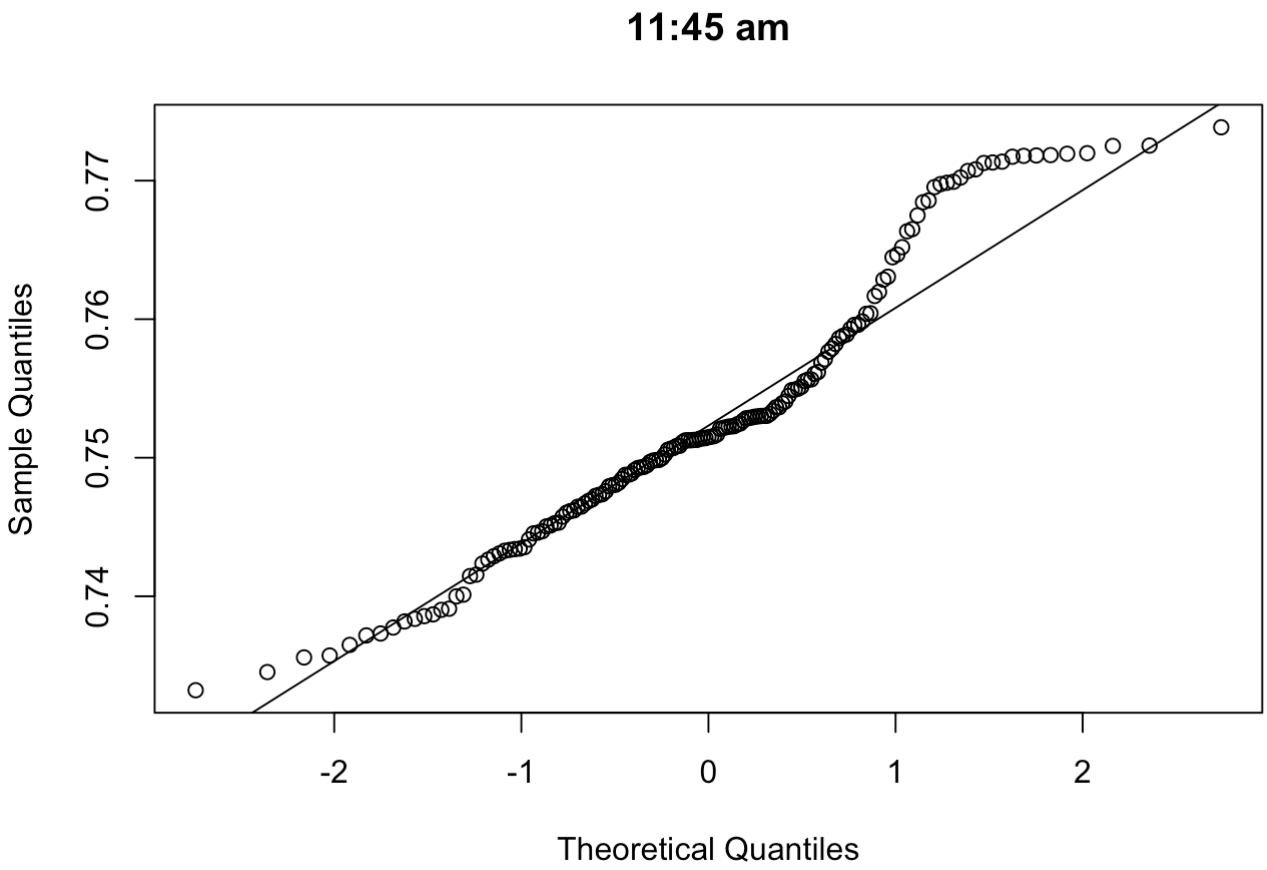
```
var(t210)
```

```
## [1] 1.930824e-05
```

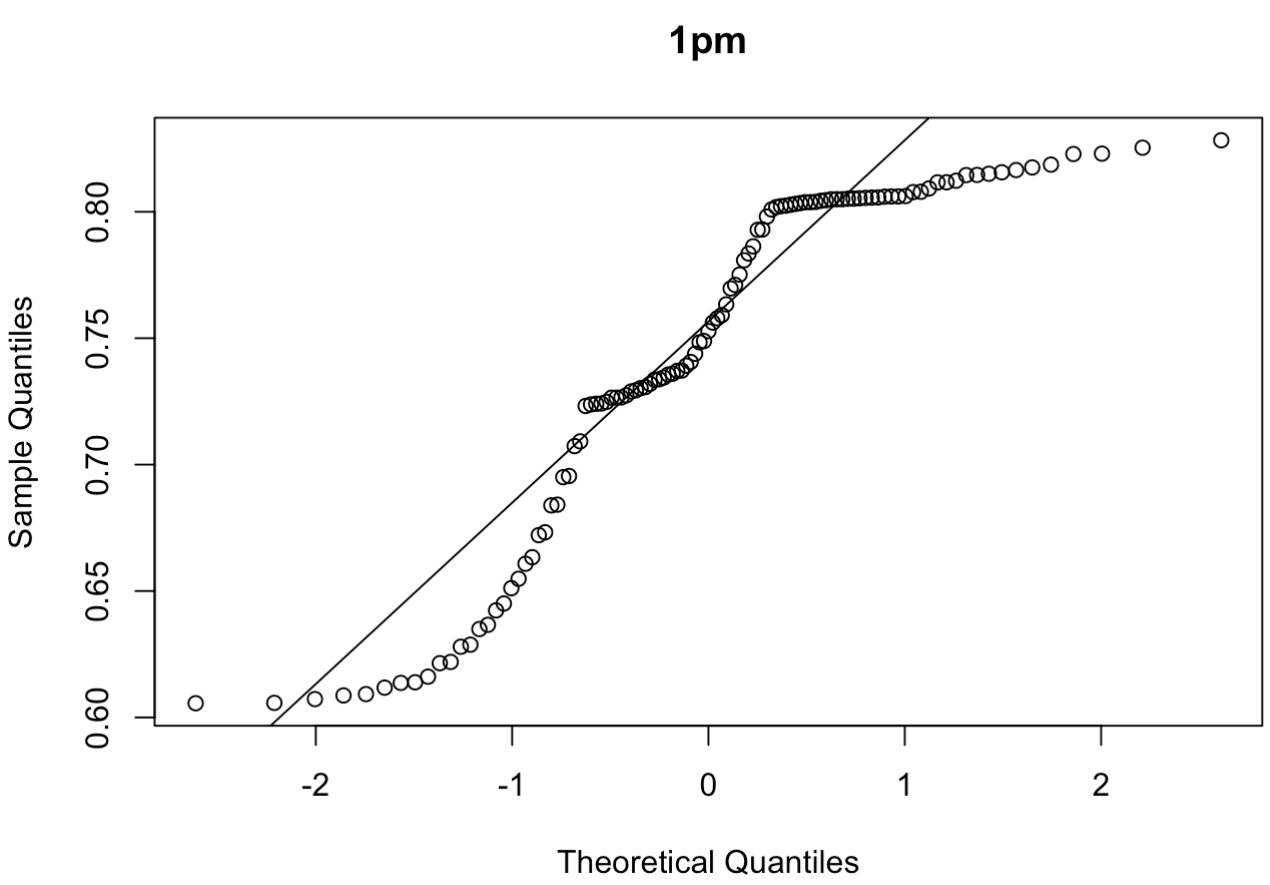
```
var(t245)
```

```
## [1] 0.001281958
```

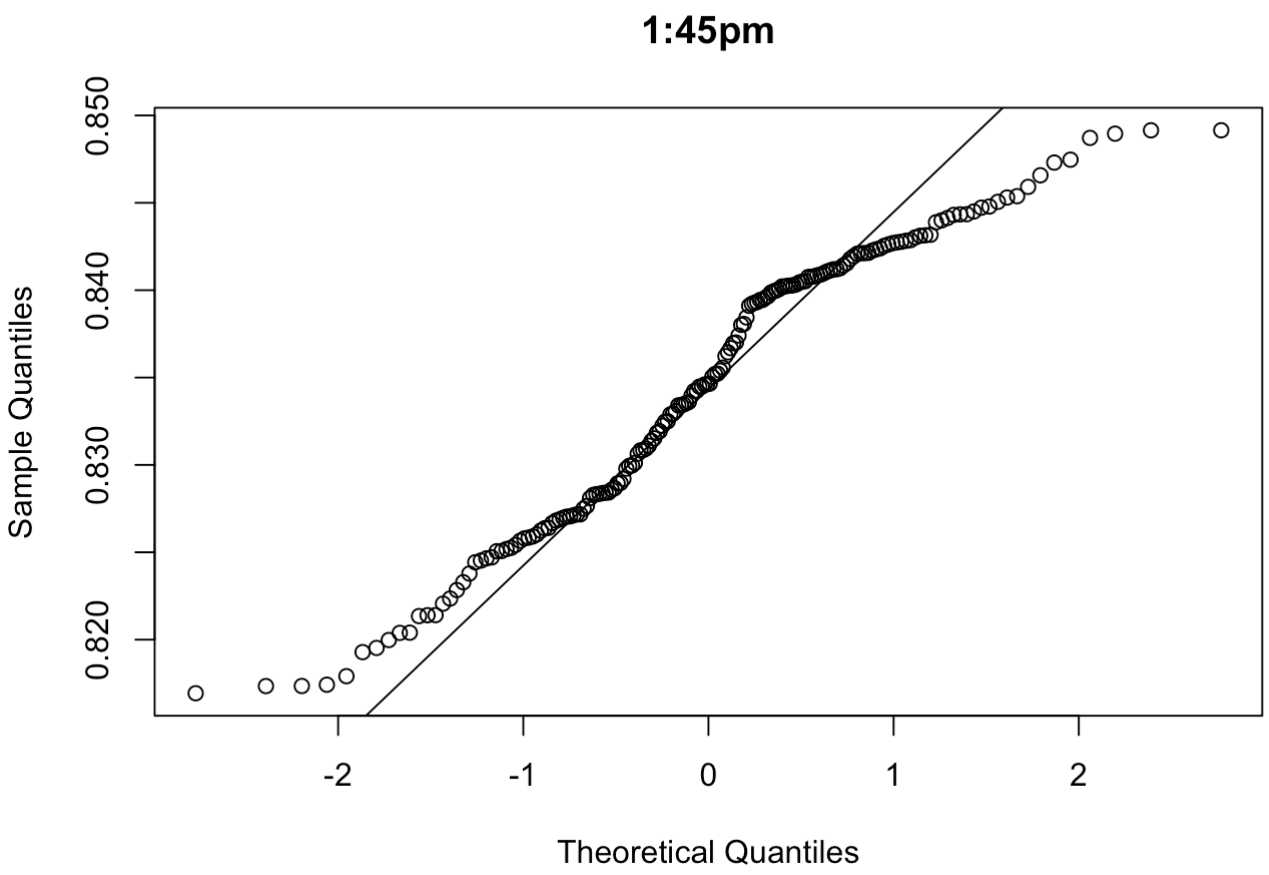
```
# for each dataset make a qq plot to determine normality
qqnorm(t1145, main = "11:45 am")
qqline(t1145)
```



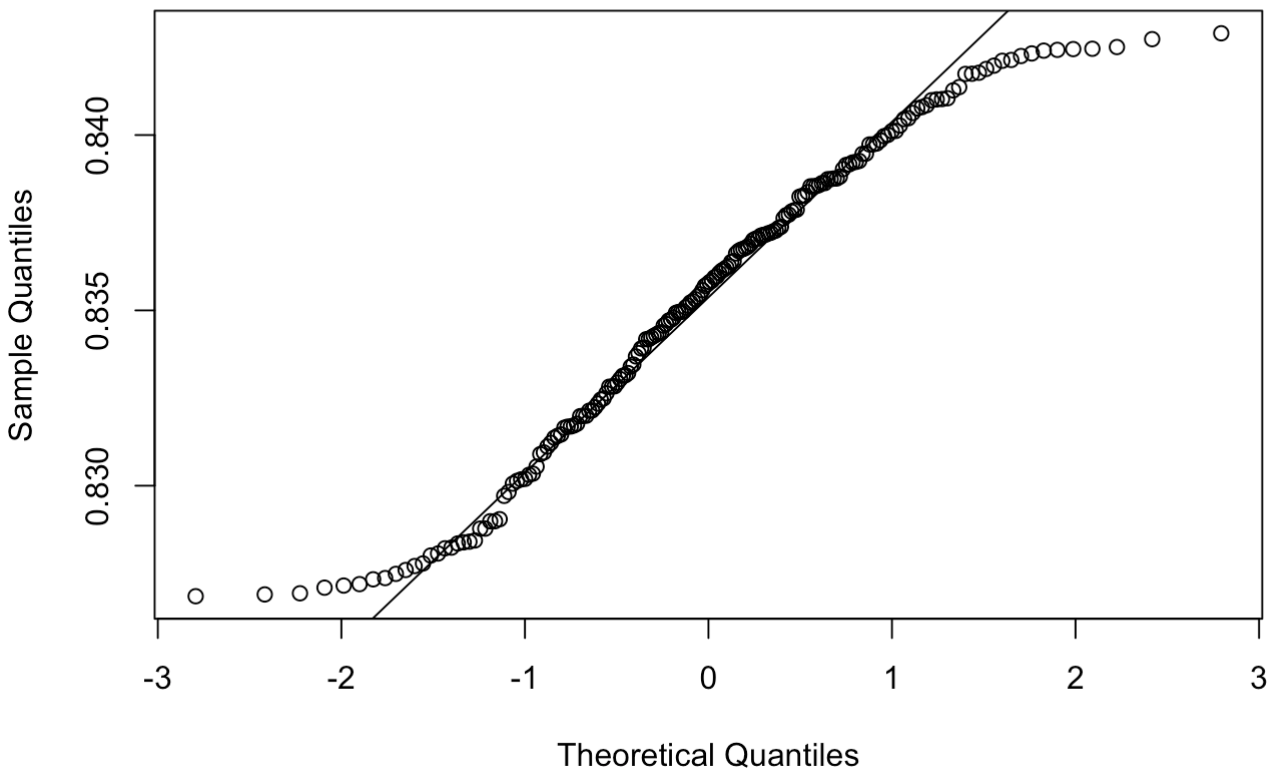

```
qqnorm(t1, main = "1pm")
qqline(t1)
```



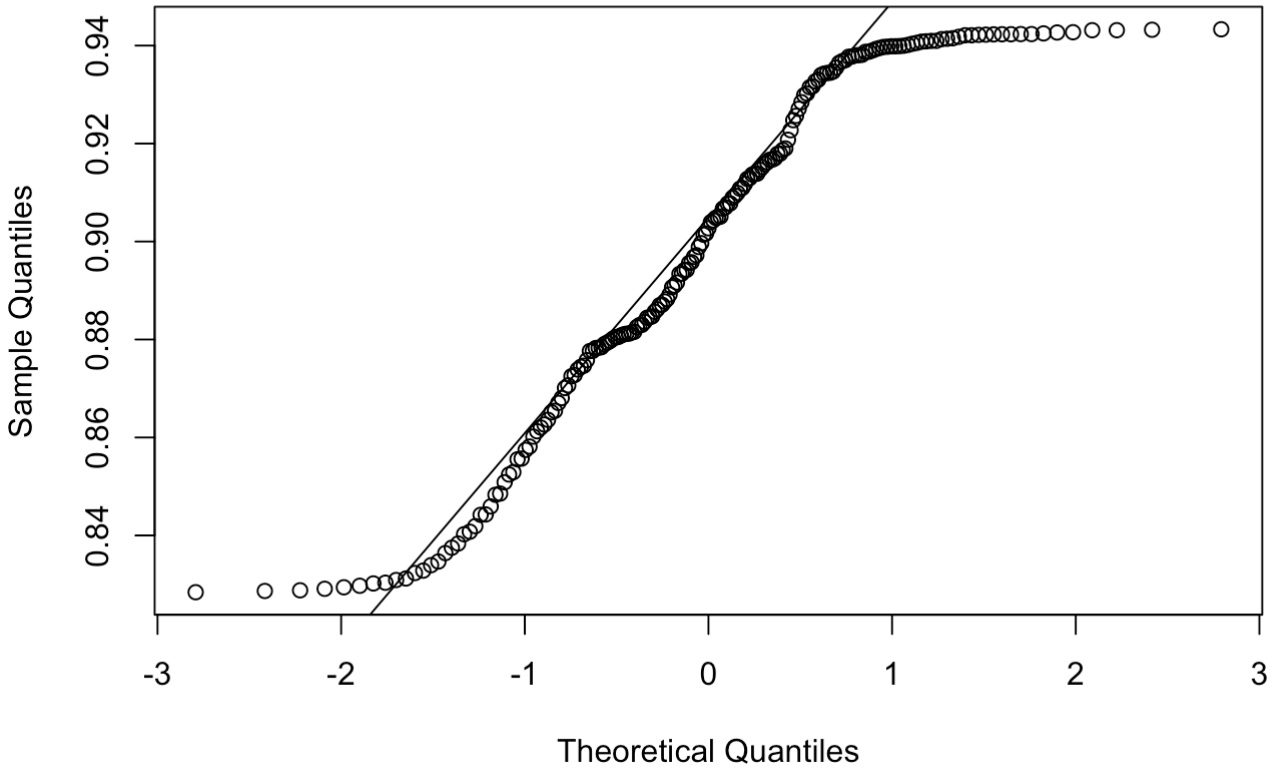

```
qqnorm(t145, main = "1:45pm")
qqline(t145)
```




```
qqnorm(t210, main = "2:10pm")
qqline(t210)
```




```
qqnorm(t245, main = "2:45pm")
qqline(t245)
```



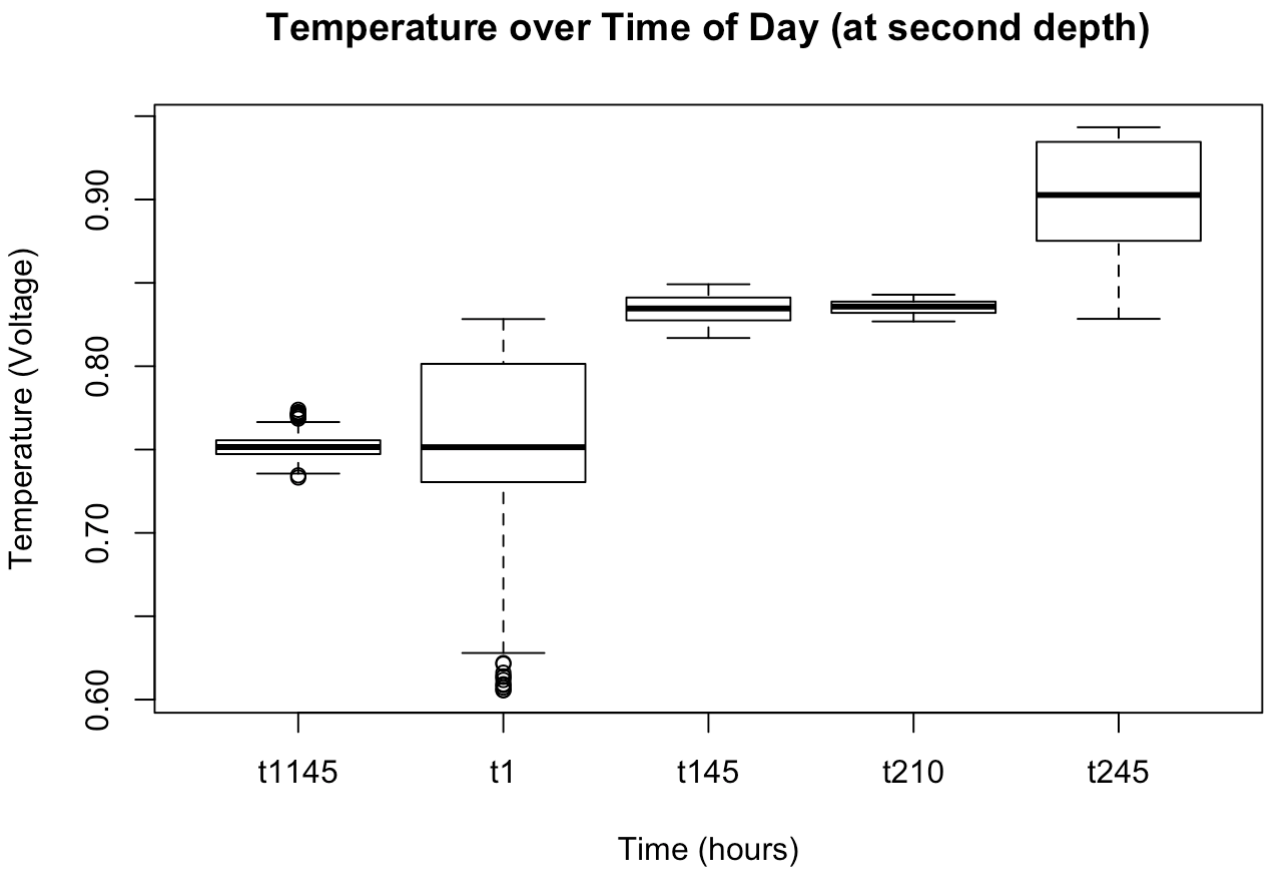
Results: The qq-plots show a moderate level of normality. For the ANOVA, which is resistant to variations in normality, it is good enough. However, the variances are more than an order of magnitude in difference. An option here is to attempt to transform the data- this is an operation taken on each data point, usually sqrt, arcsin, or log/ln, which manipulates the data in a consistent way such that it is more applicable to the needed assumptions. However, each of these transformations failed to get the variance within the necessary order of magnitude difference, so we cannot use the ANOVA test.

Instead, we will use its nonparametric alternative, the Kruskal-Wallace Test. For large sample sizes, there is little decrease in power (very roughly, accuracy) of the test compared to an ANOVA. This test does not assume a normal distribution or equal variance, but it does require that each category of data have a similar shape. We can confirm this by noticing the consistent vague S-shape throughout the qq-plots, or notice that the noise in the data throughout our main temp v depth graph is shaped similarly throughout.

Before we do our test, we choose our null hypothesis. This is usually that there is no difference between the groups. More specifically, our null hypothesis is that there is no statistically significant difference in the means of our time of day temperature groups measured by the robot at its first hovered depth.

```
# this is just rearranging the data such that the anova and kruskal.test functions like
 them- the qualitative aspect is included as a second comment to distinguish sets of dat
a points
timequality = c(rep('t1',163), rep('t1145', 111), rep('t145', 178), rep('t210', 192), re
p('t245', 191))
tempvolts = c(t1,t1145,t145, t210,t245)
anovaframe = data.frame(timequality, tempvolts)
anovaframe$timequality = factor(anovaframe$timequality, c('t1145','t1','t145','t210','t2
45'))

boxplot(tempvolts~timequality, data = anovaframe, main = "Temperature over Time of Day
 (at second depth)", ylab='Temperature (Voltage)', xlab='Time (hours)', horizontal = FAL
SE)
```




```
# either an anova or kruskal-wallace test
#anova = aov(tempvolts~timequality, data = anovaframe)
results = kruskal.test(tempvolts~timequality, data = anovaframe)

# display the results
#summary(anova)
results
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  tempvolts by timequality
## Kruskal-Wallis chi-squared = 680.34, df = 4, p-value < 2.2e-16
```

```
# for an anova test, we can see the differences between all the groups included
#TukeyHSD(anova)
```

For our non-parametric test, we cannot use the Tukey HSD test, which requires the same assumptions as an ANOVA, so there is no ready test to evaluate the difference between each pair of means. It is considered bad practice to attempt to characterize the differences in means by doing every combination of two-group tests (ie, just t-tests or nonparametric equivalents), so we will stop at the conclusion that there is or is not a statistically significant difference between at least two of the means.

Interpretation of the test- we are given a p-value result, and we compare it to the significance level (usually .05). If it is larger, we fail to reject our null hypothesis. If it is smaller, we reject the null. Our null hypothesis is that there is no change between the means of the different time of day groups. In this case we see that there is a 2e-16 p value, so we soundly reject the null hypothesis at any usual significance level.