statistics project

Nguyen Ngoc Duy

Part 2: Basic Inferential Data Analysis Instructions

```
#if(!require(devtools)) install.packages("devtools")
#devtools::install_github("kassambara/ggpubr")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
```

2.1 Load the ToothGrowth data and perform some basic exploratory data analyses

```
data("ToothGrowth")
data <- ToothGrowth
head(data)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
str(data)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

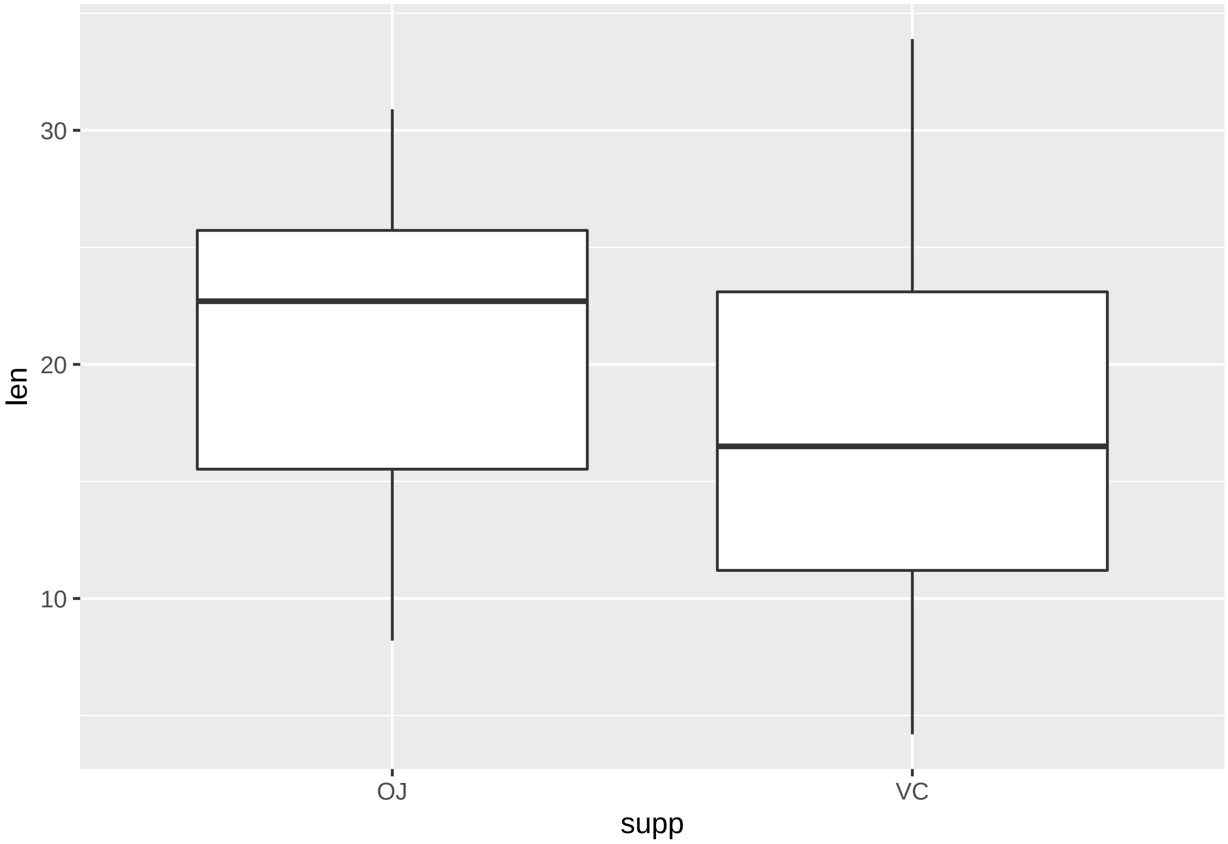
The data has 60 observations and 3 variables (from the str() we get the type of variables): 1. len (numeric) - Tooth length 2. supp (factor) - Supplement type (VC or OJ) 3. dose (numeric) - Dose in milligrams

2.2 Provide a basic summary of the data.

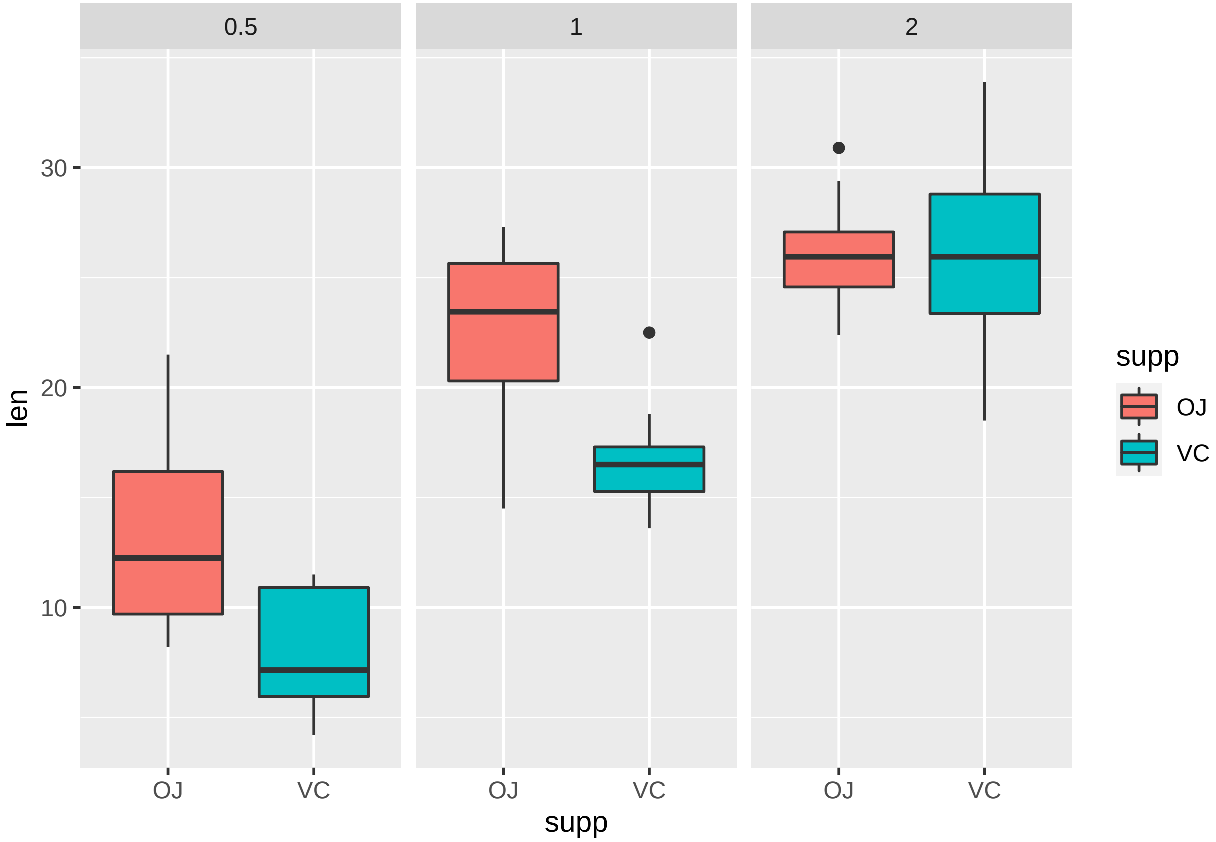
```
summary(data)
```

```
##       len         supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

```
data %>%
  ggplot(aes(x=supp,y=len)) +
  geom_boxplot()
```

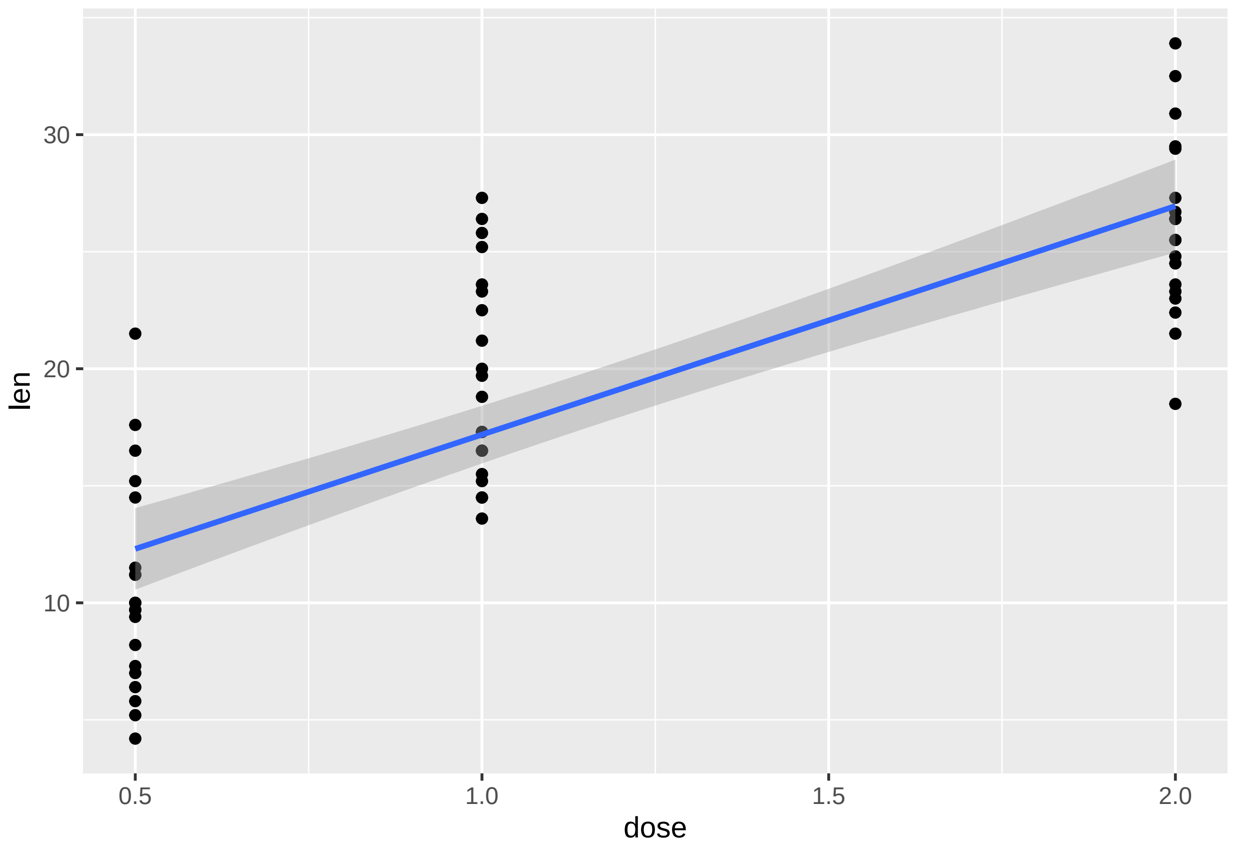


```
data %>%
  ggplot(aes(x=supp,y=len)) +
  geom_boxplot(aes(fill = supp))+
  facet_wrap(~dose)
```



```r
data %>%
  ggplot(aes(x=dose,y=len)) +
  geom_point()+
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



positive effect of the dosage, as the dosage increases the tooth growth increases. In the specific case of the VC, the tooth growth has a linear relationship with dosage. The higher dossage (2.0mg) has less improvement in tooth growth with the OJ supplement. However, the OJ supplement generally induces more tooth growth than VC except at higher dosage (2.0 mg).

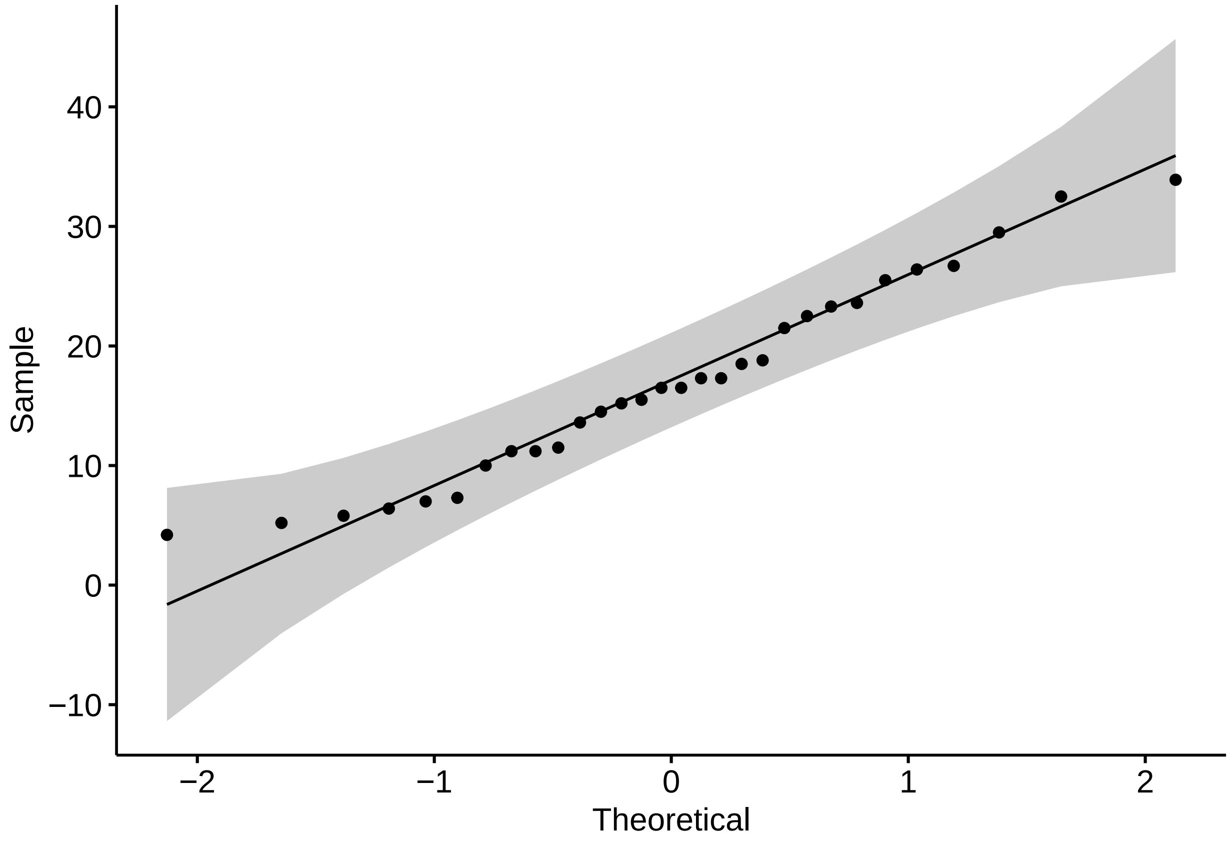
2.3 Hypothesis for the supplement OJ vs VC

Note that, unpaired two-samples t-test can be used only under certain conditions:

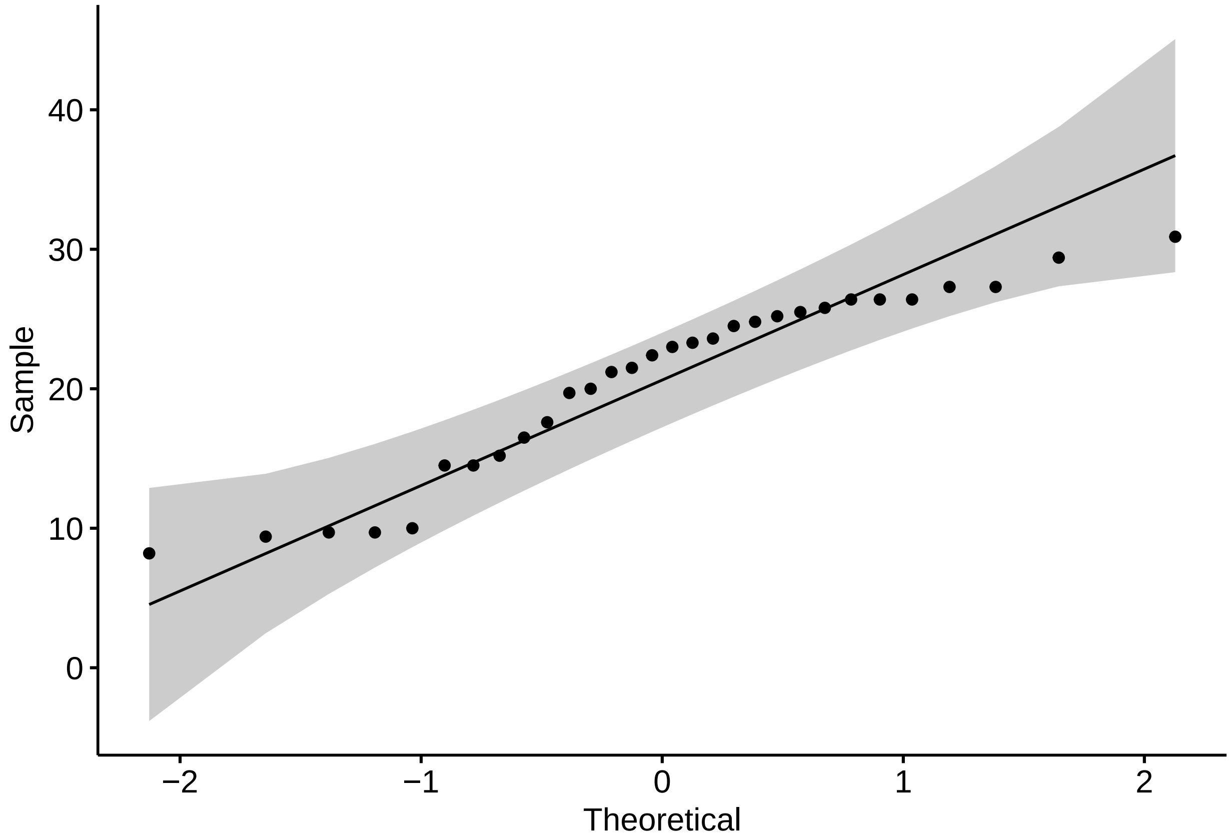
- When the two groups of samples (A and B), being compared, are normally distributed. This can be checked using Shapiro-Wilk test.
- When the variances of the two groups are equal. This can be checked using F-test.

Assumtion : Are the data from each of the 2 groups follow a normal distribution?

```
VC_len <- data$len[data$supp == 'VC']
OJ_len <- data$len[data$supp == 'OJ']
ggqqplot(VC_len)
```



```
ggqqplot(OJ_len)
```



```
shapiro.test(VC_len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  VC_len
## W = 0.96567, p-value = 0.4284
```

```
shapiro.test(OJ_len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OJ_len
## W = 0.91784, p-value = 0.02359
```

len of VC is normal but len from OJ is not normal distribution, we should use Wilcoxon test

Let our null hypothesis to be there is no difference in tooth growth when using the supplement OJ and VC.

OJ_len = VC_len

Let our alternate hypothesis to be there are more tooth growth when using supplement OJ than VC.

OJ_len > VC_len

```
res <- wilcox.test(OJ_len, VC_len,exact = FALSE,alternative = "greater")
res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  OJ_len and VC_len
## W = 575.5, p-value = 0.03225
## alternative hypothesis: true location shift is greater than 0
```

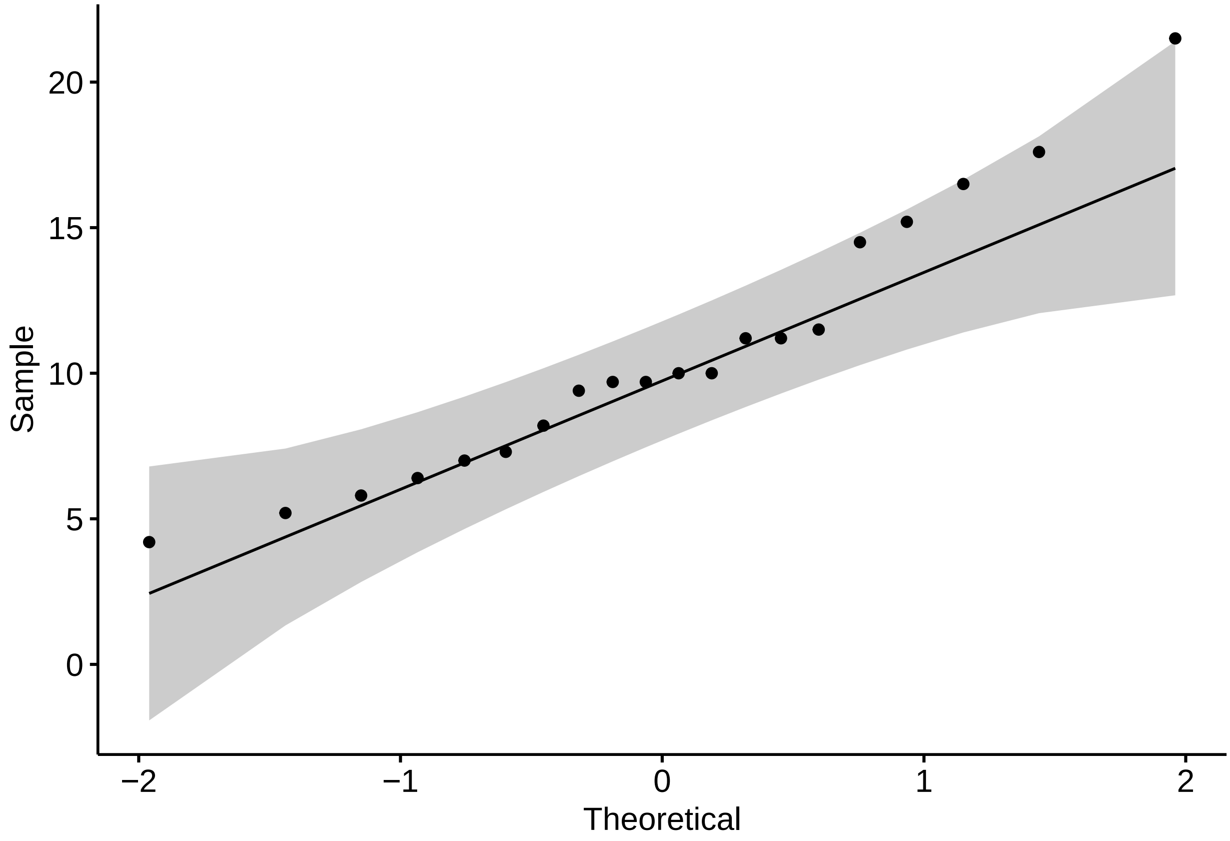
The p-value of the test is 0.03, which is lower than the significance level alpha = 0.05. We can conclude that OJ's median len is higher than VC's median len

2.3 Hypothesis for the dossage

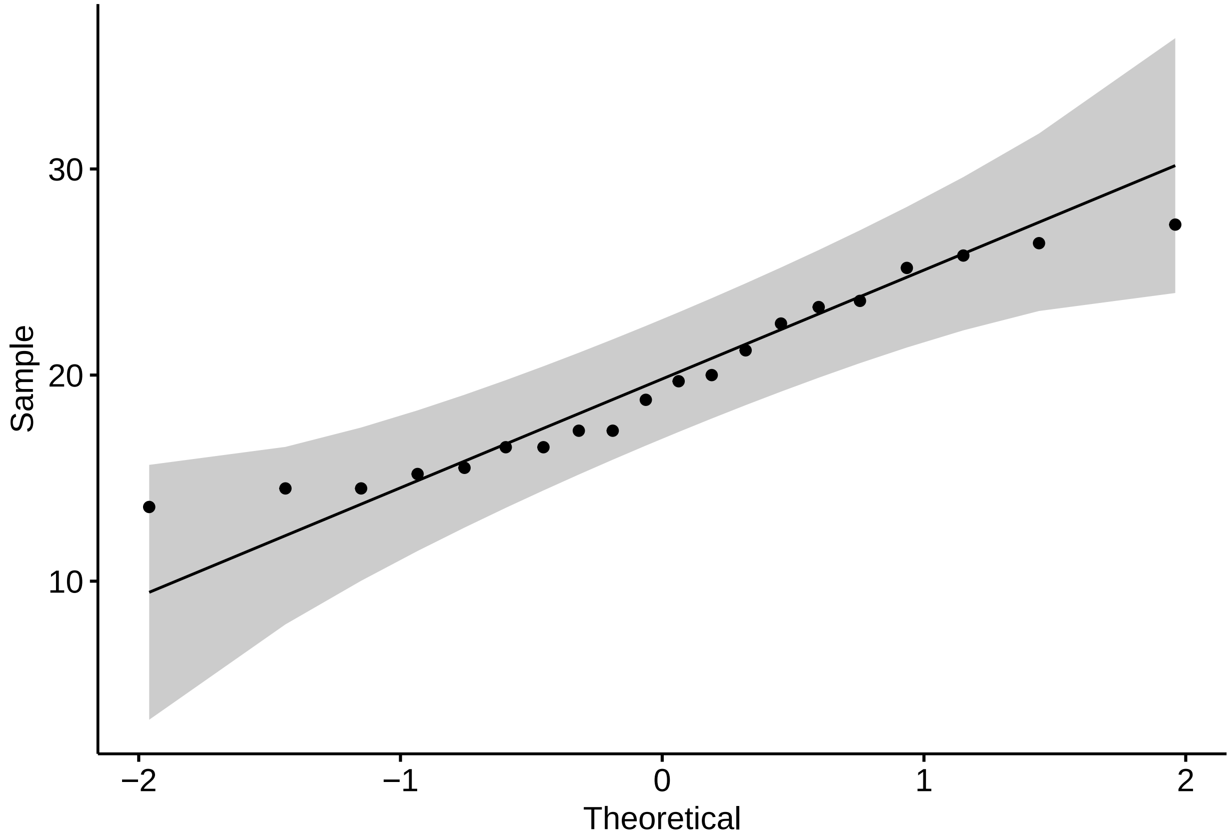
```
doseHalf = data$len[data$dose == 0.5]
doseOne = data$len[data$dose == 1]
doseTwo = data$len[data$dose == 2]
```

Assumtion : Are the data from each of the 2 groups follow a normal distribution?

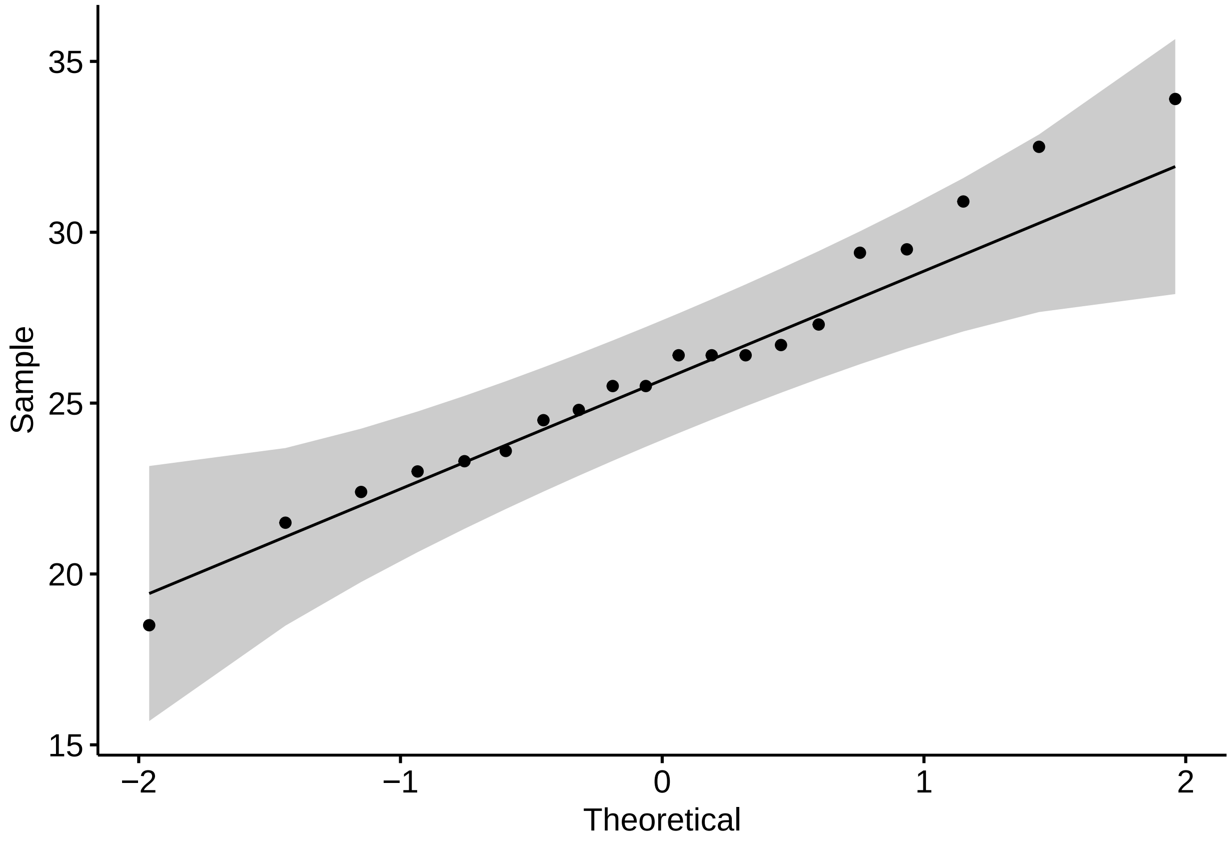
```
ggqqplot(doseHalf)
```



```
ggqqplot(doseOne)
```



```
ggqqplot(doseTwo)
```



```
shapiro.test(doseHalf)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doseHalf
## W = 0.94065, p-value = 0.2466
```

```
shapiro.test(doseOne)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doseOne
## W = 0.93134, p-value = 0.1639
```

```
shapiro.test(doseTwo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doseTwo
## W = 0.97775, p-value = 0.9019
```

From the output, the all p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from the normal distribution. In other words, we can assume the normality.

Assumption Do the two populations have the same variances?

```
res.ftest <- var.test(doseHalf,doseOne)
res.ftest
```

```
##
##  F test to compare two variances
##
## data:  doseHalf and doseOne
## F = 1.0386, num df = 19, denom df = 19, p-value = 0.9351
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4110751 2.6238736
## sample estimates:
## ratio of variances
##           1.038561
```

The p-value of F-test is greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two sets of data. Therefore, we can use the classic t-test witch assume equality of the two variances.

```
t.test(doseHalf, doseOne, alternative = "less", paired = FALSE, var.equal = TRUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  doseHalf and doseOne
## t = -6.4766, df = 38, p-value = 6.331e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -6.753344
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

As the p-value is lower than 0.05 (the default value for the tolerance of the error alpha), then, we reject the null hypothesis. That can be interpreted as there is almost null chances of obtain an extreme value for the difference in mean of those dossages (doseHalf < doseOne) on the tooth growth.

```
res.ftest <- var.test(doseOne,doseTwo)
res.ftest
```

```
##
##  F test to compare two variances
##
## data:  doseOne and doseTwo
## F = 1.3687, num df = 19, denom df = 19, p-value = 0.5005
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5417489 3.4579584
## sample estimates:
## ratio of variances
##           1.368702
```

The p-value of F-test is greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two sets of data. Therefore, we can use the classic t-test witch assume equality of the two variances.

```
t.test(doseOne, doseTwo, alternative = "less", paired = FALSE, var.equal = TRUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  doseOne and doseTwo
## t = -4.9005, df = 38, p-value = 9.054e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.175196
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

As the p-value is lower than 0.05 (the default value for the tolerance of the error alpha), then, we reject the null hypothesis. That can be interpreted as there is almost null chances of obtain an extreme value for the difference in mean of those dossages (doseOne < doseTwo) on the tooth growth.