STAC33-TUT5

Uzair Mirza

13/02/2022

Introduction

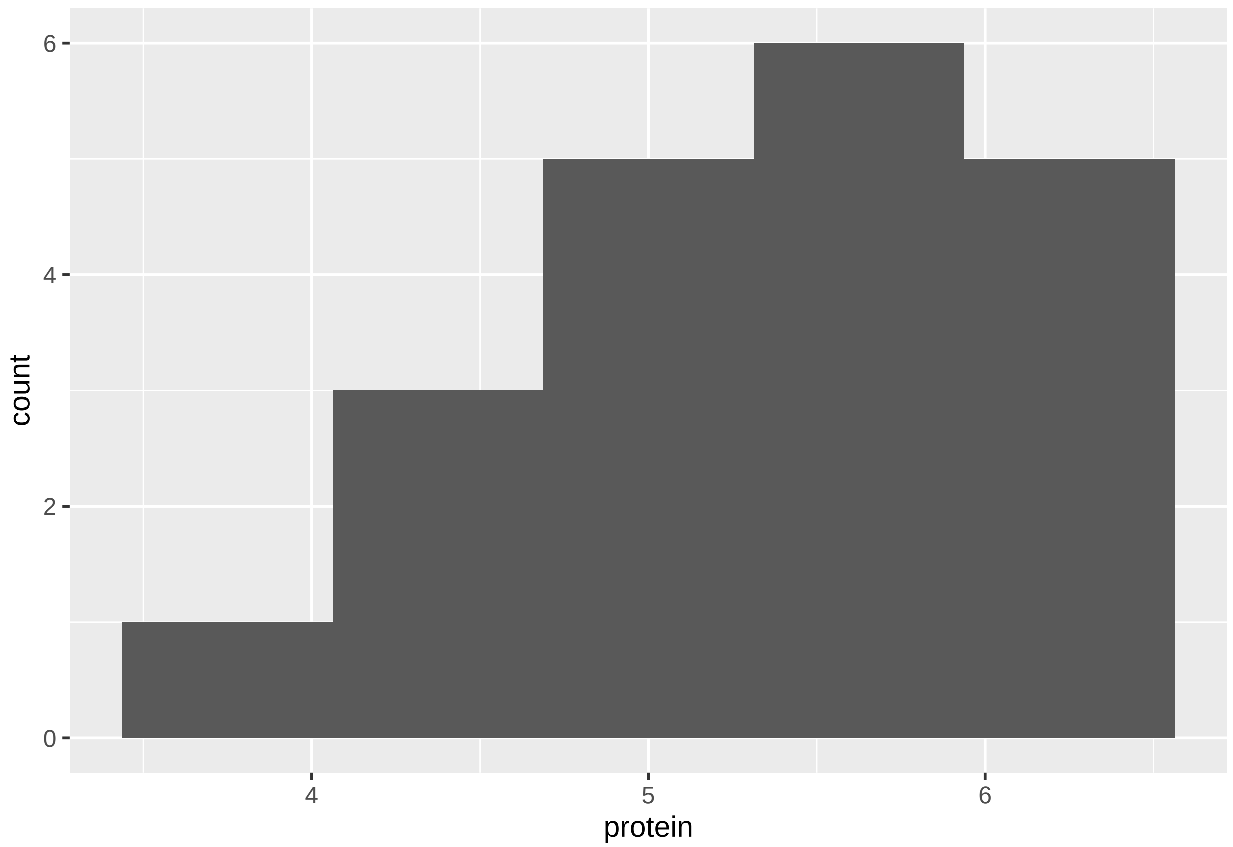
This week's lecture we will be discussing topics from CH:9, Ch:10 and Ch:12. Chapter is about **Sign test** and Chapter 10 is about **Mood Median test** and Chapter 12 is about **QQ-plot**(aka normal quantile plots). All the problems being discussed can be found on the PASIAS here

Question 9.10

Data is about amount of protein in the package. The advertisement claims to have **6 ounces/pack**(our assumption about the package H_0). Goal is to test this claim by sampling(n=20) and conducting some tests.

a + b. Read the data and make sutiable plot

```r
# a
my_url <- "http://ritsokiguess.site/datafiles/protein.txt"
## Data is not seperated by a space/chracter or tab so this works
meals <- read_table(my_url)

# b
## 1 quantitative variable(Can make box-plot/histogram)
## Interested in distribution rather than range so histogram
ggplot(meals, aes(x=protein)) + geom_histogram(bins = 5)
```



Note we can observe?
- (Not symmetric)
- Skewed left

c. Why might a sign test be better than a t-test for assessing the average amount of protein per package? Explain briefly. ("Average" here means any measure of centre.)

Assumptions for t-test:
- Normality

Here we can see the assumption is violated(as not normal from Histogram).
- We testing μ in t-test and here it is not a reliable measure(think about the sample not being accurate and μ) as it is sensitive.
- Sample size not good enough for LLN(n=20) here

Hence a sign test on Median is a better test to use.
Sign test has no dependencies/assumptions about the population parameters or the distributions
Can perform test on Median.
Median, Mode not sensitive to odd values(outliers).

d. Run a suitable sign test for these data. What do you conclude?

Setup, !!!THIS IS WHAT GOOGLE WOULD SUGGEST, DON'T DO THIS!!!

```r
install.packages('smmR')
```

Follow this from lecture slides.
github source: here

```r
# One time setup
install.packages('devtools')
library(usethis) # this step is specific to my machine, might not be required for you
library(devtools)
install_github("nxskok/smmr")
```

```r
library(smmr)
```

This is how you run the test.
Now note we are interested in the the Median being $= 6$, so we state that. Furthermore think about whether it will be 2-sided or 1 sided?
- 2-sided bc the value can lie on either side.

`sign_test(data, variable, $H_{0}$)` $\alpha = 0.05$

```r
sign_test(meals, protein, 6)
```

```
## $above_below
## below above 
##    15     5 
## 
## $p_values
##   alternative    p_value
## 1       lower 0.02069473
## 2       upper 0.99409103
## 3   two-sided 0.04138947
```

The P-value, 0.0414, is less than $\alpha = 0.05$ 0.05, so we reject the null hypothesis and conclude that the median is different from 6 ounces. The advertisement by the company is not accurate.

e. In your sign test, how could you have deduced that the P-value was going to be small even without looking at any of the P-values themselves?

- How is Sign test conducted ?
  - Compare the sampled values against the H_0
    - What happens when values are EQUAL?
      Don't include those values as currently interested in `>` or `<`

```r
meals %>%
  group_by(protein > 6) %>%
  summarise(n = n())
```

```
## # A tibble: 2 x 2
##   `protein > 6`     n
##   <lgl>         <int>
## 1 FALSE            15
## 2 TRUE              5
```

- Note 15 are less than 6. Hence can reject and assume p-value will be less without conducting the test.

f. Obtain a 90% confidence interval for the population median protein content. What does this tell you about the reason for the rejection or non-rejection of the null hypothesis above?

`ci_median(data, variable,conf.level = $1-\alpha$ )` by default $\alpha = 0.05$

```r
alpha = .1
ci_median(meals, protein, conf.level = 1-alpha)
```

```
## [1] 4.905273 5.793750
```

- Does our H_0 value lie withing the range at $\alpha = 0.1$?
  - No! 6 not $\in$ CI.

10.8 Handspans revisited

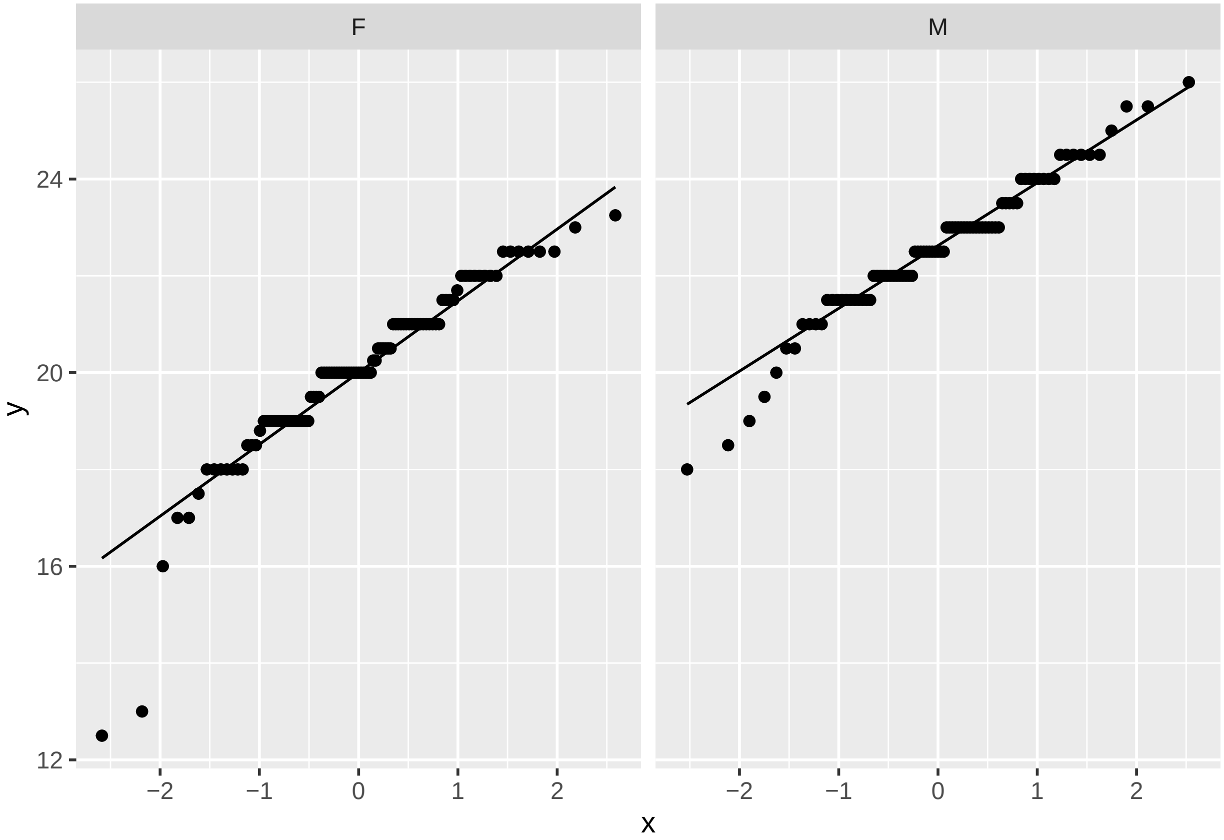
Recall the original study was to compare the handspan between males and females.
So we have 2 groups in our study.

a + b. Load data make plots to investigate and compare normality between the sample group

```r
# data
my_url <- "http://ritsokiguess.site/datafiles/handspan.txt"
span <- read_delim(my_url, " ")
```

```
## Rows: 190 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: " "
## chr (1): sex
## dbl (1): handspan
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# QQ-Plot
ggplot(span, aes(sample=handspan)) + stat_qq() + stat_qq_line() + facet_wrap(~sex)
```



- What is QQ plot?
- What is a quantile ? Can think **similar**(but not really) to bins.
- How is it plotted?
Compare the quantiles of Normal(0,1) against the sample's quantiles
- Our plot description:
- Males:
Not normal as the values to the left are deviating away from the normal line(SLIGHTLY left skewed)
- Females:
Same as males + seem to have outliers aswell on the left side

c. Discuss briefly whether you might prefer to use Mood's median test to compare the handspans of the male and female students, compared to a two-sample t-test.

- T-test assumptions about the normality of the sample is not satisfied.

- Mood-Median is the move as does not rely on μ + Normality.

d. Run the Mood-Median test

$$H_0 : \mu_{\frac{1}{2}} = 0 \qquad \implies \mu_{\frac{1}{2}} = median$$

The median between both the groups is the same ie the difference is 0.
Setup *the call for installing the package is covered above.*

```
# library call
library(smmr)
```

To run the test. `median_test(data, variable of the median, group)`

```
median_test(span, handspan, sex)
```

```
## $table
##      above
## group above below
##     F    17    82
##     M    65    11
##
## $test
##        what        value
## 1 statistic 8.06725e+01
## 2        df 1.00000e+00
## 3   P-value 2.66404e-19
```

The P-value of `2.66404e-19`is extremely small, so we can conclude that males and females have different median handspans. Remember that we are now comparing medians, and that this test is **two-sided**.

12.3 Lengths of heliconia flowers

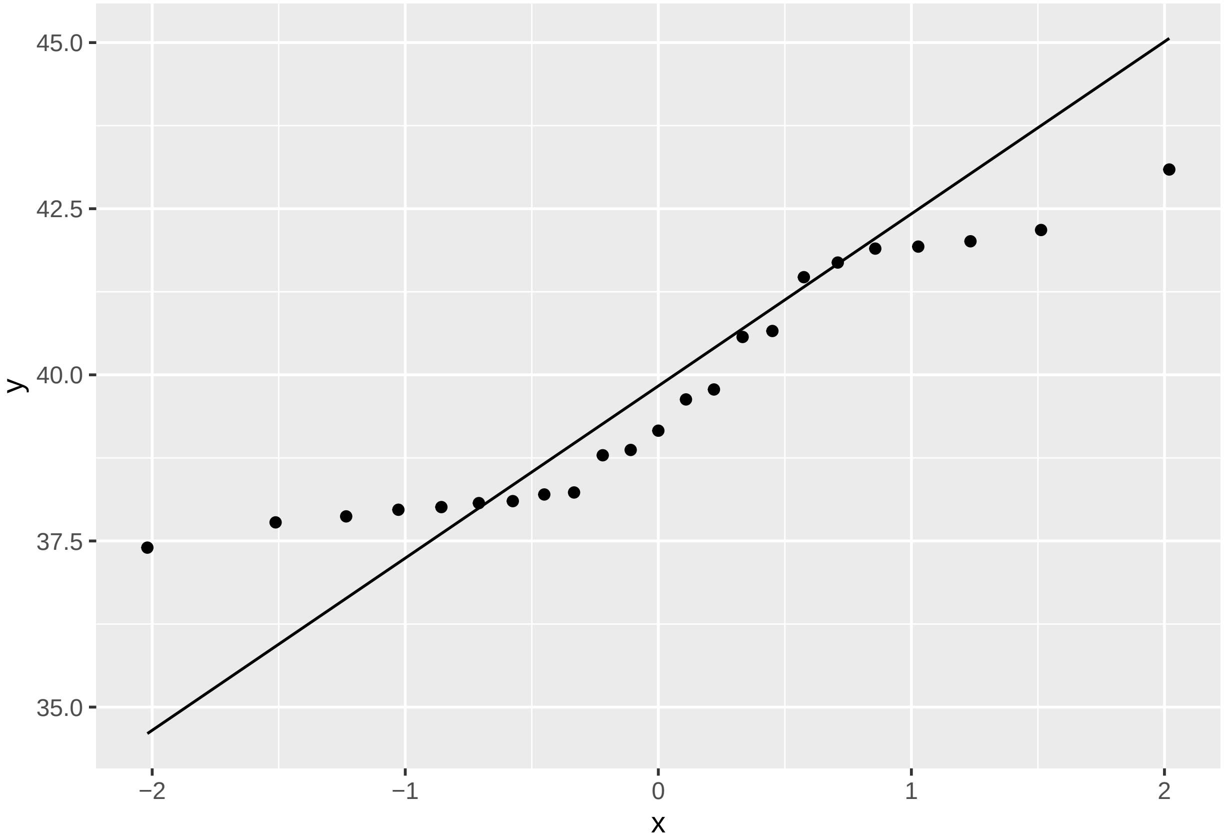
Flower length and the beak of the humming-bird has been evolved according to the specie

a,b,c,d. Read data and make QQ-Plot for each specie

```
# a
my_url <- "http://ritsokiguess.site/datafiles/heliconia.csv"
heliconia = read_csv(my_url)
```

```
## Rows: 23 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): bihai, caribaea_red, caribaea_yellow
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

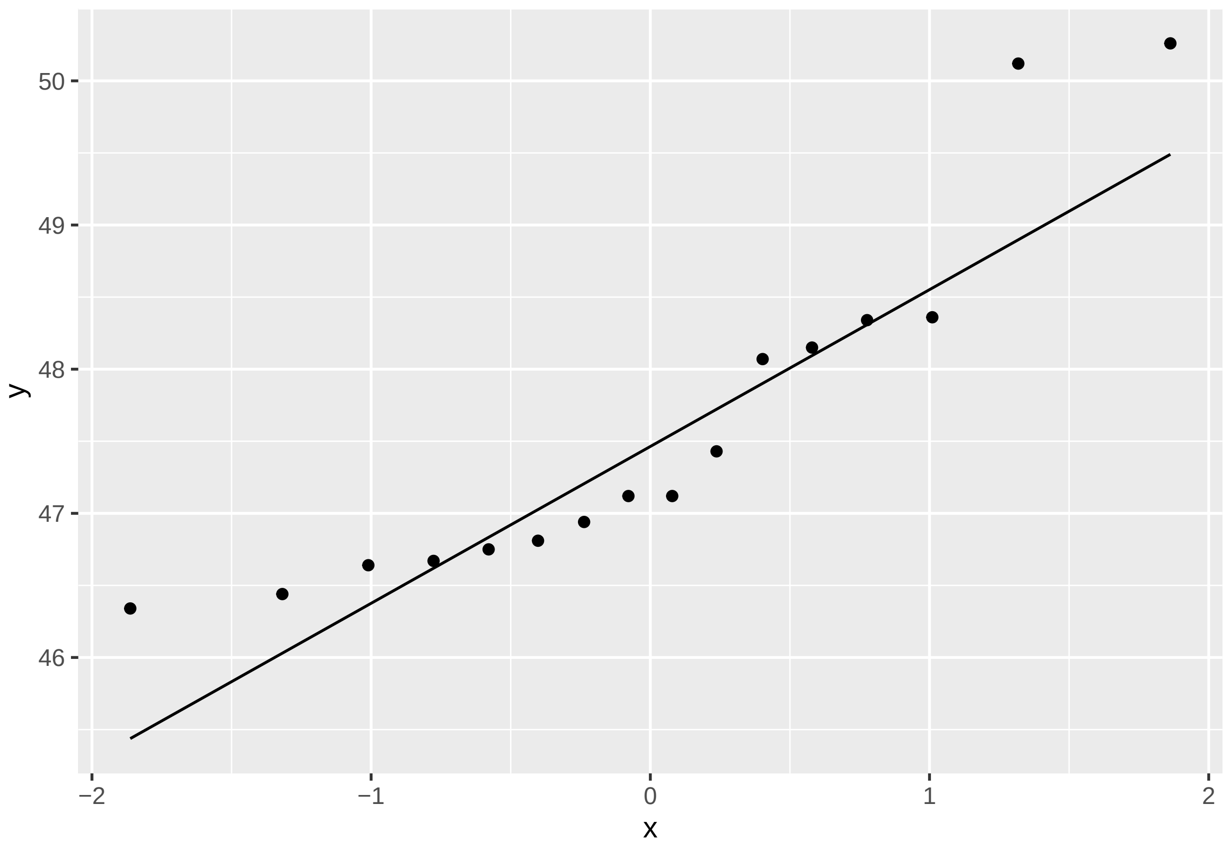
```
# bcd
## way 1
ggplot(heliconia,aes(sample=caribaea_red)) + stat_qq() + stat_qq_line()
```



```
## way 2
heliconia %>% ggplot(aes(sample=bihai)) + stat_qq() + stat_qq_line()
```

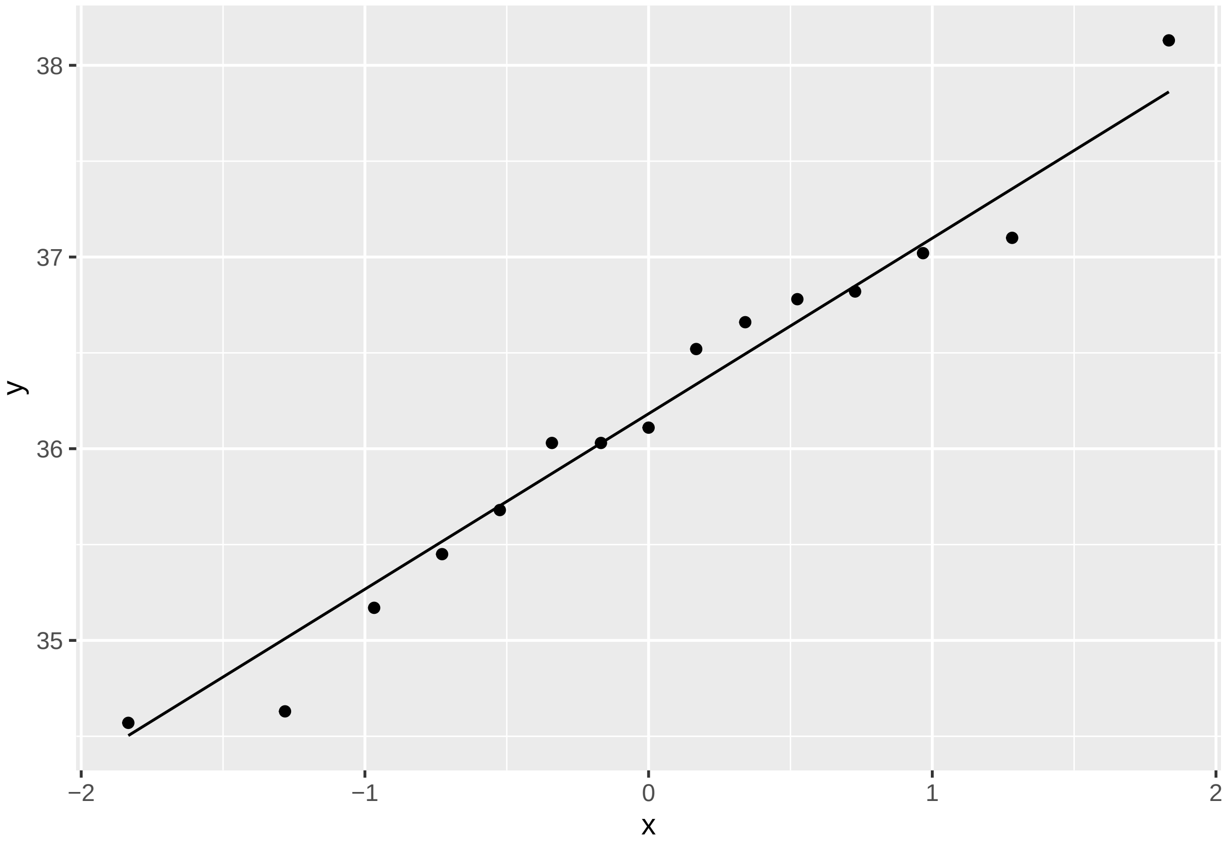
```
## Warning: Removed 7 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 7 rows containing non-finite values (stat_qq_line).
```



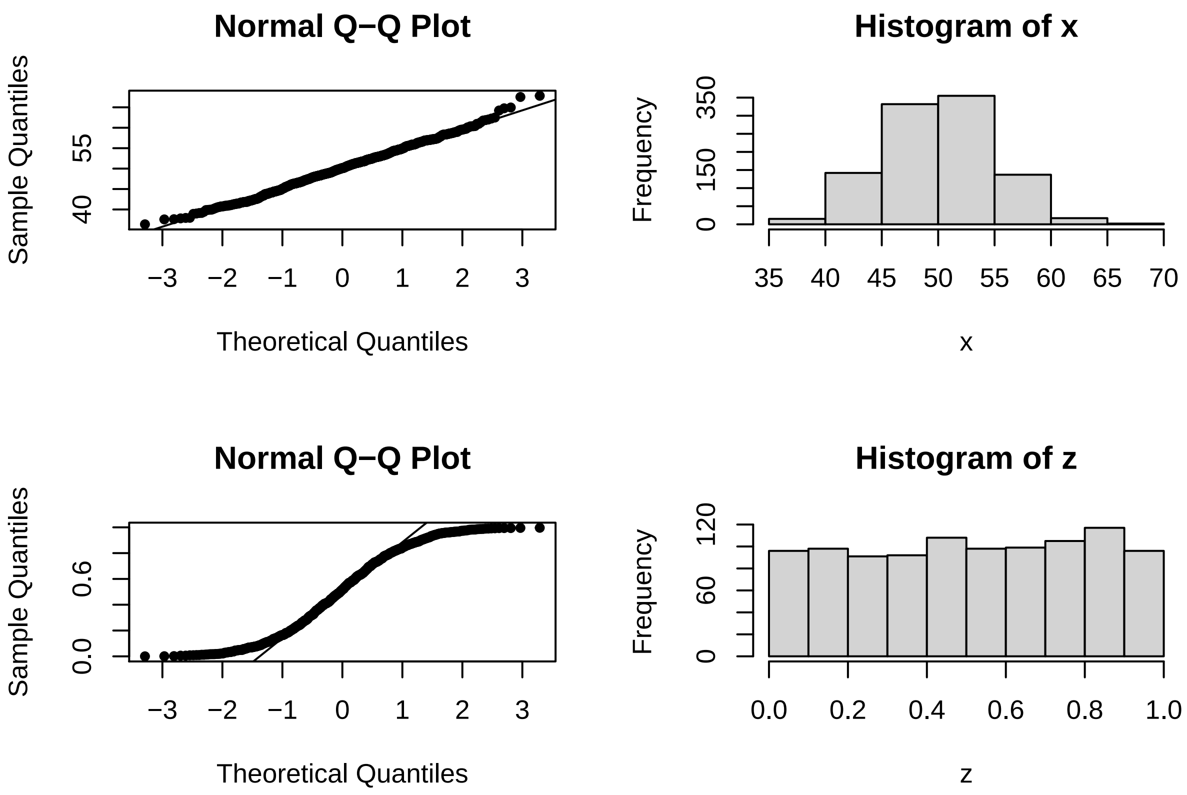
```
ggplot(heliconia,aes(sample=caribaea_yellow))+stat_qq()+stat_qq_line()
```

```
## Warning: Removed 8 rows containing non-finite values (stat_qq).

## Warning: Removed 8 rows containing non-finite values (stat_qq_line).
```



### e, f. Which one is the clossest to the normal and of the other explain why they are not. - `caribaea_yellow` closest to normal as the quantiles seems to match with Normal QQ. - `bihai` seems to have a S-shaped if we negate the outliers at the top + left skewed with removal of outliers and without the removal seems to be right skewed. - `caribaea_red`
Short tailed Initially points seem too bunched up on the line then they seem to tail away from the QQ-line. Looks Uniform distribution.

```
### Source: https://bioinfo.iric.ca/permutations/

x <- rnorm(1000, mean=50, sd=5) # normal distribution
z <- runif(1000) # uniform distribution
par(mfrow=c(2,2))
qqnorm(x, pch=20)
qqline(x)
hist(x)
qqnorm(z, pch=20)
qqline(z)
hist(z)
```



Check more examples of QQ x Distribution relation. here