Chlorophyll-a index is affected by the presence of Trichodesmium

Code ▾

Project 3

500371329

University of Sydney | DATA1001 | October 2021

1 Recommendation

There is change in chlorophyll-a level during the presence of Trichodesmium blooms, resulting in changing water quality in the ocean. This finding will be benefit for AIMS (Australian Institute of Marine Science) to monitor water quality.

2 Domain knowledge

Concentrations of the plant pigment chlorophyll-a (occurs in all marine phytoplankton) are a useful proxy indicator of the amount of nutrients incorporated into phytoplankton biomass. Chlorophyll-a is today the most commonly used parameter for the monitoring of phytoplankton biomass and nutrient status, as an index of water quality. High levels often indicate poor water quality and low levels often suggest good conditions.

The presence of Trichodesmium can have an impact to chlorophyll-a level.

3 Initial Data Analysis (IDA)

Hide

```
library(tidyverse)
library(RColorBrewer)

# Read in your data

## Option 1: International Airlines operating from Australia
flights = read.csv("http://www.maths.usyd.edu.au/u/UG/JM/DATA1001/r/current/projects/2020data/flights.csv")

## Option 2: Penalty Notices in Australia
penalties = read.csv("http://www.maths.usyd.edu.au/u/UG/JM/DATA1001/r/current/projects/2020data/penalities.csv")

## Option 3: Great Barrier Reef Chlorophyll Monitoring
gbr = read.csv("http://www.maths.usyd.edu.au/u/UG/JM/DATA1001/r/current/projects/2020data/GBR.csv")
```

Hide

```
gbr_cleaned <- gbr %>%
    mutate(TRICHODESMIUM = case_when(
      tolower(TRICHODESMIUM) == 'a' ~ 'Absent', # get A or a and convert to Absent
      tolower(TRICHODESMIUM) == "p" ~ 'Present', # get P or p and convert to Present
      TRICHODESMIUM == 'No Record' ~ 'Absent', # get No record and convert to Absent
      TRICHODESMIUM == '' ~ 'Absent', # get blank and convert to Absent
      TRUE ~ TRICHODESMIUM # keep everything else
  ))
```

Hide

```
gbr_cleaned$TRICHODESMIUM <- as.factor(gbr_cleaned$TRICHODESMIUM) # Reclassify column TRICHODESMIUM to factor
```

The variables in the data are explored below.

Hide

```
# structure
str(gbr_cleaned)
```

```
## 'data.frame':    19174 obs. of  14 variables:
##  $ STATION_NAME  : chr  "Hook Passage" "Hook Passage" "Hook Passage" "Hook
Passage" ...
##  $ TRANSECT      : chr  "Whitsundays" "Whitsundays" "Whitsundays" "Whitsun
days" ...
##  $ STATION_ID    : int  228 228 228 228 229 229 229 229 227 227 ...
##  $ SAMPLE_TIME   : chr  "15/1/09 9:39" "15/1/09 9:39" "15/1/09 9:39" "15/
1/09 9:39" ...
##  $ LATITUDE      : num  -20.2 -20.2 -20.2 -20.2 -19.8 ...
##  $ LONGITUDE     : num  149 149 149 149 149 ...
##  $ SECCHI_DEPTH  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ACOUSTIC_DEPTH: num  25 25 25 25 64 64 64 64 25 25 ...
##  $ REPLICATE_DESC: chr  "A1" "A2" "B1" "B2" ...
##  $ TEMPERATURE   : num  29 29 29 29 29.5 29.5 29.5 29.5 29 29 ...
##  $ SALINITY      : num  38 38 38 38 38 38 38 38 38 38 ...
##  $ SAMPLE_DEPTH  : num  25 25 25 25 64 64 64 64 25 25 ...
##  $ CHL_A         : num  0.61 0.67 0.64 0.69 0.55 0.47 0.47 0.5 0.49 0.53
...
##  $ TRICHODESMIUM : Factor w/ 2 levels "Absent","Present": 1 1 1 1 1 1 1 1
1 1 ...
```

4 Evidence

> **Question : Is there a difference in Chlorophyll-a level if we compare the absence and the presence of Trichodesmium?**

4.1 Hypothesis

We can address this question by performing a two sample T-test [2 sided].

- H_0: The Chlorophyll-a level is not different between absent group and present group
- H_1: The Chlorophyll-a level is different between 2 groups

4.2 Assumptions

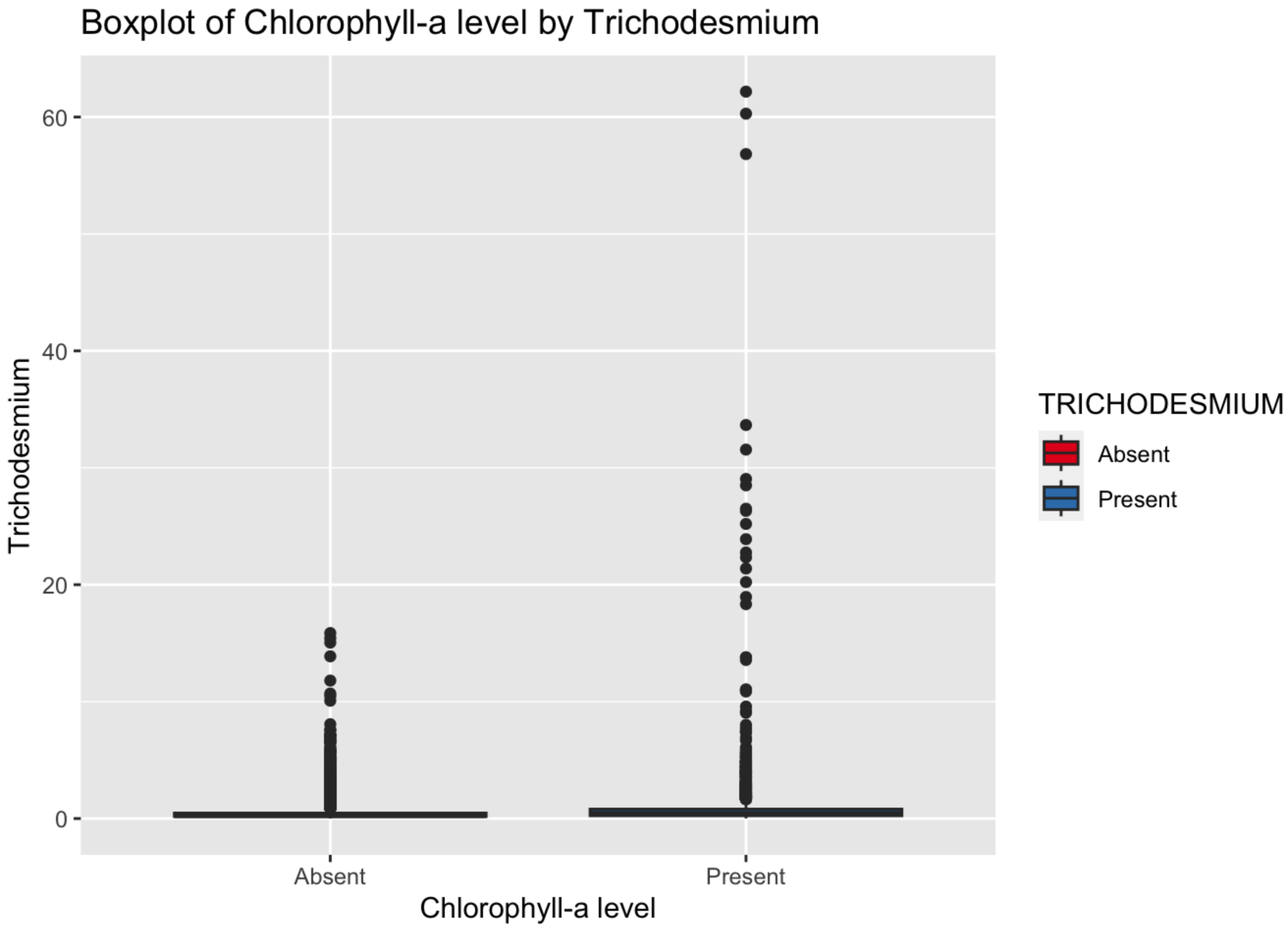
4.2.1 Assumptions (original data)

Boxplot

The boxplot can be used to simultaneously check for **normality** and **equal variance** assumptions.

Hide

```
ggplot(data = gbr_cleaned , aes(x = TRICHODESMIUM, y=CHL_A, fill = TRICHODESM
IUM )) +
  geom_boxplot()+
  scale_fill_brewer(palette = "Set1")+
  xlab ("Chlorophyll-a level")+
  ylab ("Trichodesmium")+
  ggtitle("Boxplot of Chlorophyll-a level by Trichodesmium")
```

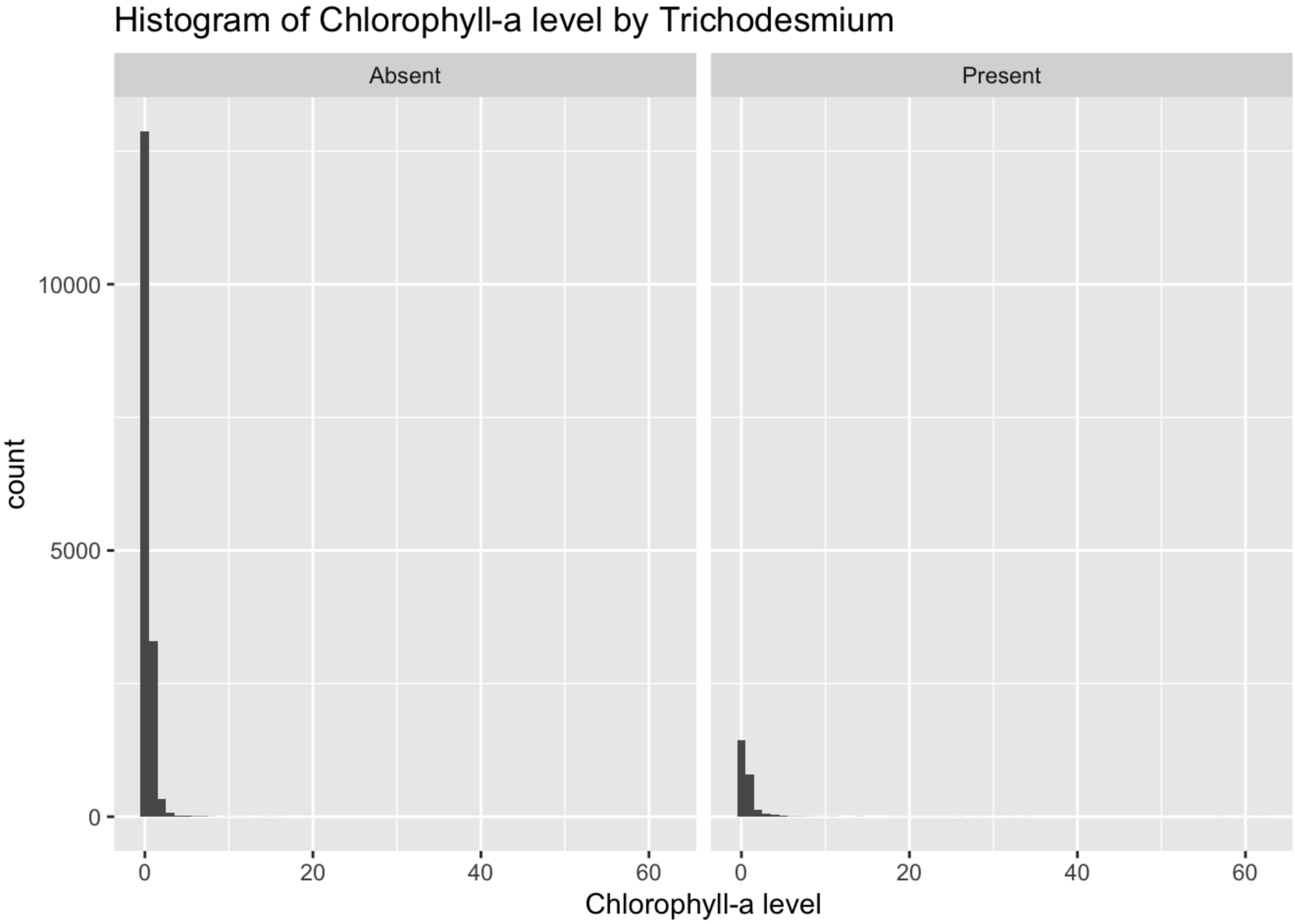


Histogram

The histogram can be used to check for the assumption of **normality** in each sample.

Hide

```
ggplot(gbr_cleaned, aes(CHL_A)) +
geom_histogram(binwidth = 1)+
  facet_wrap(~ TRICHODESMIUM) +
  xlab("Chlorophyll-a level")+
  ggtitle("Histogram of Chlorophyll-a level by Trichodesmium")
```

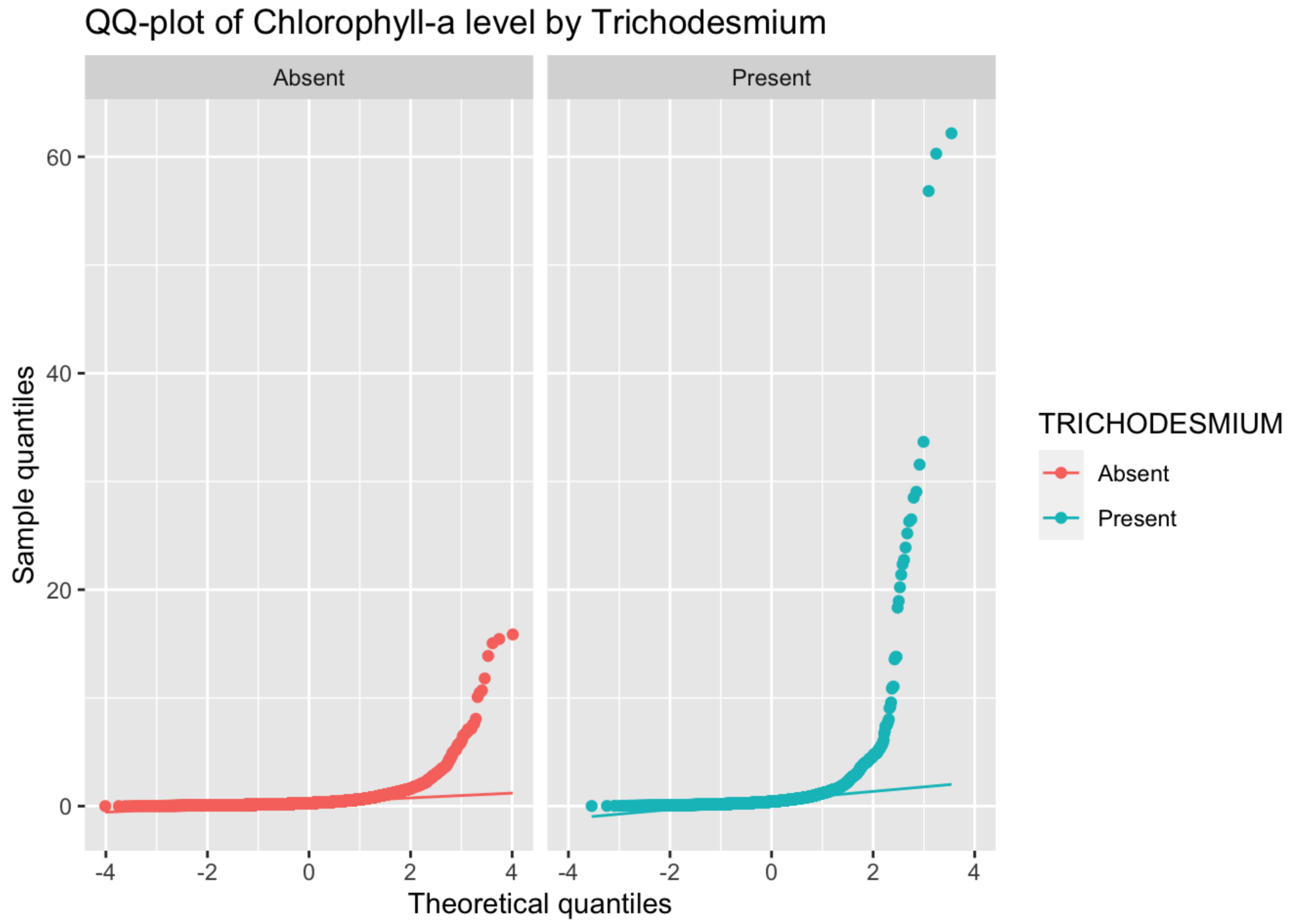


QQ-plot

The QQ-plot can be used to check for the assumption of **normality** in each sample.

Hide

```
ggplot(gbr_cleaned, aes(sample = CHL_A, colour = TRICHODESMIUM)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ TRICHODESMIUM) +
  xlab("Theoretical quantiles") +
  ylab("Sample quantiles")+
  ggtitle("QQ-plot of Chlorophyll-a level by Trichodesmium")
```



Both groups look asymmetric and not normal. Hence we transform the non-normal data to check assumptions again before performing a T-test

Levene's Test (F-test)

The F-test can be used to check for the assumption of **equal variances** in each sample.

Hide

```
var.test (CHL_A ~ TRICHODESMIUM, gbr_cleaned)
```

```
##
##  F test to compare two variances
##
## data:  CHL_A by TRICHODESMIUM
## F = 0.035722, num df = 16648, denom df = 2524, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.03364606 0.03787650
## sample estimates:
## ratio of variances
##         0.03572205
```

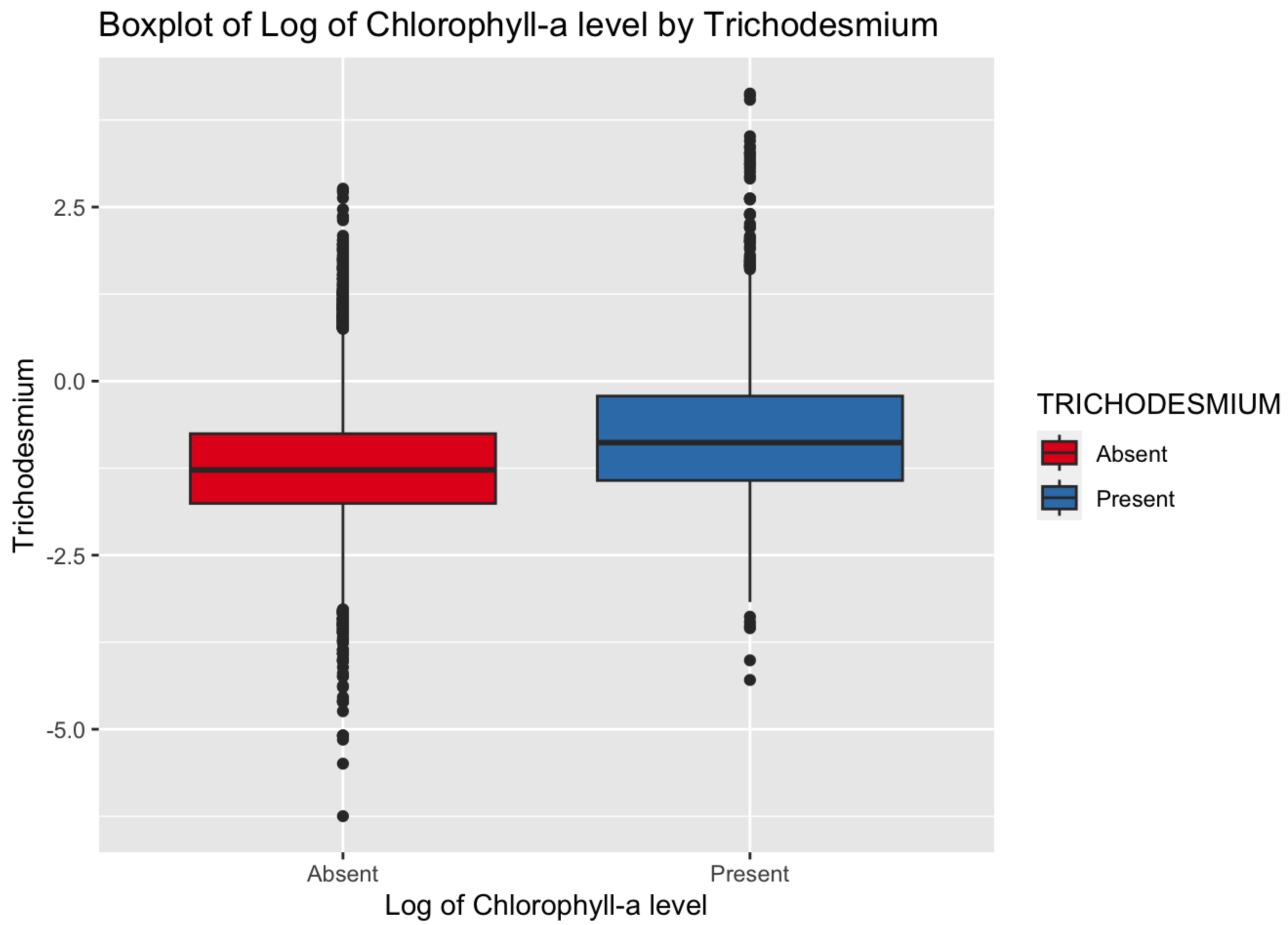
The F-test gives p-value < 0.05 suggesting that data does not have equal variance.

4.2.2 Assumptions (transformed data)

Boxplot

Hide

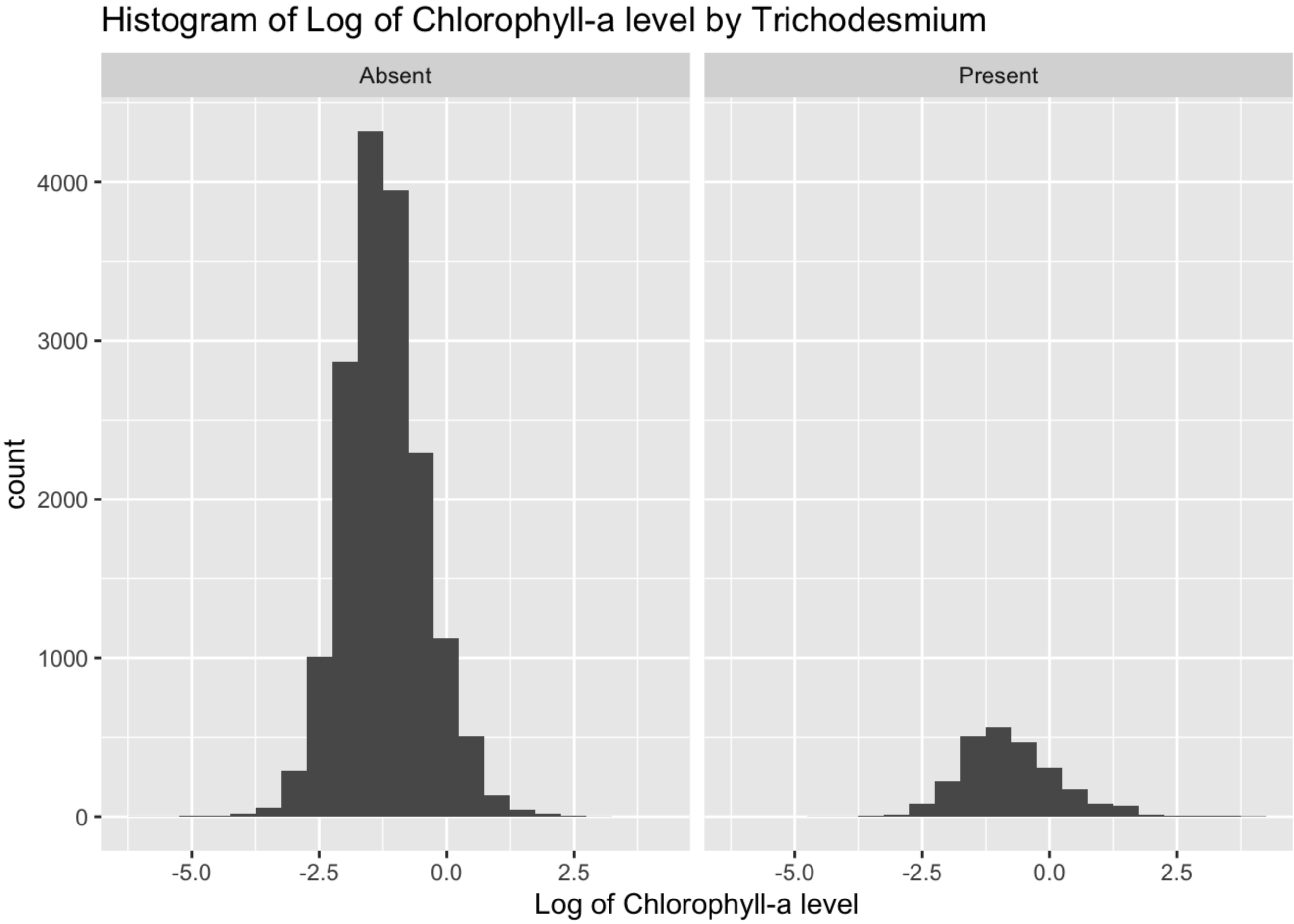
```
ggplot(data = gbr_cleaned, aes(x = TRICHODESMIUM, y=log(CHL_A), fill = TRICHO
DESMIUM )) +
  geom_boxplot()+
  scale_fill_brewer(palette = "Set1")+
  xlab ("Log of Chlorophyll-a level")+
  ylab ("Trichodesmium")+
  ggtitle("Boxplot of Log of Chlorophyll-a level by Trichodesmium")
```



Histogram

Hide

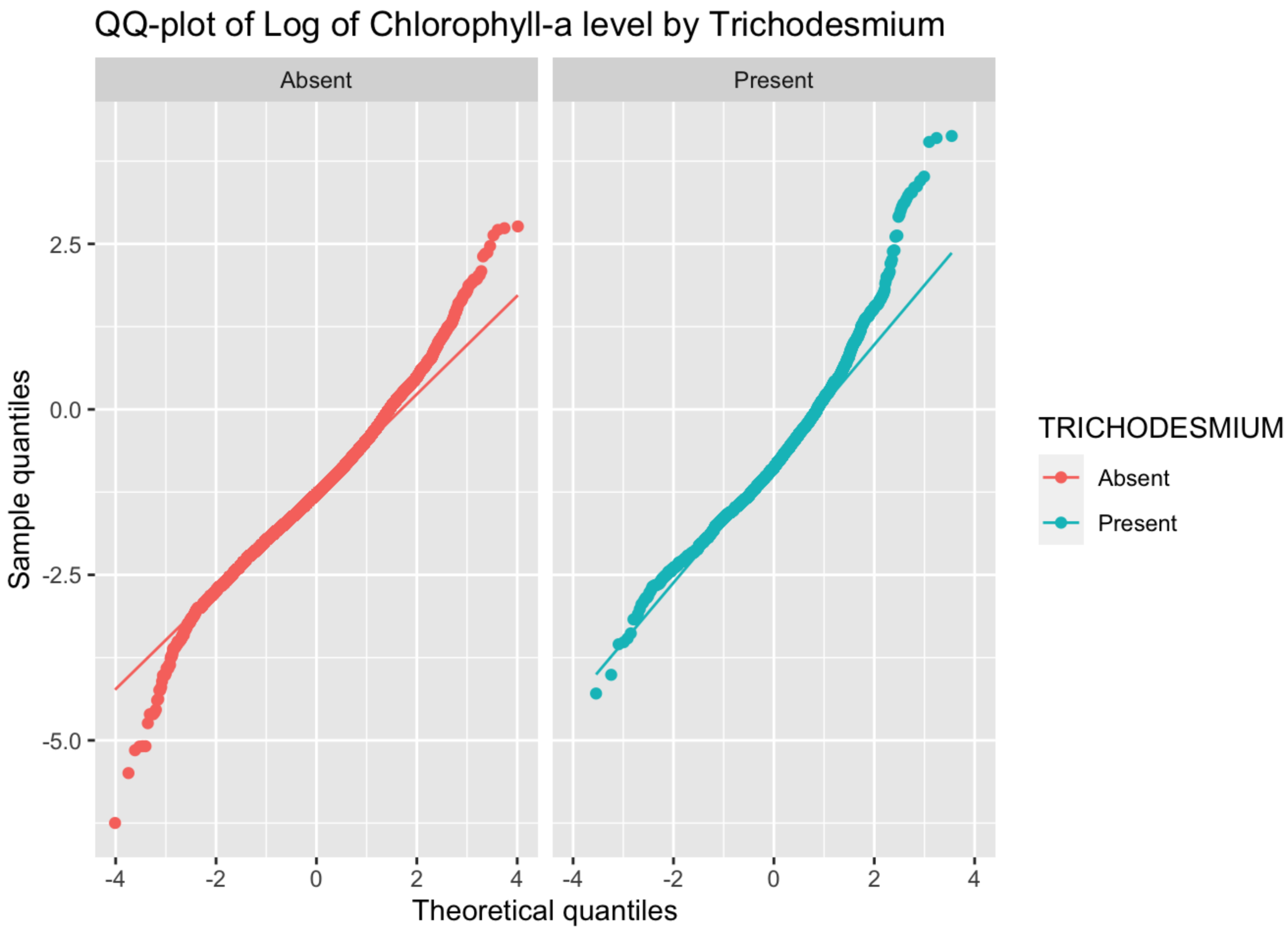
```
ggplot(gbr_cleaned, aes(log(CHL_A)), ) +
geom_histogram(binwidth = 0.5)+
  facet_wrap(~ TRICHODESMIUM) +
  xlab("Log of Chlorophyll-a level")+
  ggtitle("Histogram of Log of Chlorophyll-a level by Trichodesmium")
```



QQ-plot

Hide

```
ggplot(gbr_cleaned, aes(sample = log(CHL_A), colour = TRICHODESMIUM)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ TRICHODESMIUM) +
  xlab("Theoretical quantiles") +
  ylab("Sample quantiles")+
  ggtitle("QQ-plot of Log of Chlorophyll-a level by Trichodesmium")
```



The data appears to be right-skewed in both groups, but can reasonably fit the assumption of normality as most of the sample data are close to the diagonal line.

- **Independent**: 2 samples are independent as data were collected in different time and locations.
- **Normality**: Both groups appear symmetrical, which is consistent with normality even some outliers are observed.
- **Equal variances**: Both groups seem to have have similar spread, but we will use Levene's Test (F-test) to check if the assumption of equal variances is met.

Hide

```
var.test (log(CHL_A) ~ TRICHODESMIUM, gbr_cleaned)
```

```
## 
##  F test to compare two variances
## 
## data:  log(CHL_A) by TRICHODESMIUM
## F = 0.6535, num df = 16648, denom df = 2524, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6155236 0.6929156
## sample estimates:
## ratio of variances 
##          0.6535019
```

The F-test gives p-value < 0.05 suggesting that transformed data does not have equal variance.

Since normality assumption is met, but equal variances assumption is not met, we run a Welch two-sample T-test:

4.3 Test statistic and p-value

Hide

```
t_test_results <- t.test(log(gbr_cleaned$CHL_A) ~ TRICHODESMIUM, gbr_cleaned,
var.equal = FALSE)
t_test_results
```

```
## 
##  Welch Two Sample t-test
## 
## data:  log(gbr_cleaned$CHL_A) by TRICHODESMIUM
## t = -23.027, df = 3044.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Absent and
group Present is not equal to 0
## 95 percent confidence interval:
##  -0.5191714 -0.4376941
## sample estimates:
##  mean in group Absent mean in group Present 
##            -1.2326532            -0.7542204
```

Hide

```
t_test_results_TS <- round(t_test_results$statistic,2)
t_test_results_pvalue <- 2*round(t_test_results$p.value,2)
```

- T: The **observed test statistic** is -23.03
- P: The p-value is 0 < 0.05.

4.4 Conclusions

Statistical conclusion: We reject the null hypothesis that there is no difference between absent group and present group of Trichodesmium.

Scientific conclusion: The evidence suggests that the presence of Trichodesmium does impact Chlorophyll-a level, as a result impact water quality index .

5 References

G. B. Jones. Department of Chemistry and BiochemistryJames Cook University of North Queensland Townsville Australia 1992 'Effect of Trichodesmium Blooms on Water Quality in the Great Barrier Reef Lagoon' https://link.springer.com/chapter/10.1007/978-94-015-7977-3_18 (https://link.springer.com/chapter/10.1007/978-94-015-7977-3_18)

Australian Institute of Marine Science (AIMS) 'The Great Barrier Reef Long-term Chlorophyll Monitoring (1992-2009)' https://researchdata.edu.au/great-barrier-reef-1992-2009/677311 (https://researchdata.edu.au/great-barrier-reef-1992-2009/677311)

Australian Online Coastal Information 'Chlorophyll a concentrations' https://ozcoasts.org.au/indicators/biophysical-indicators/chlorophyll_a/ (https://ozcoasts.org.au/indicators/biophysical-indicators/chlorophyll_a/)

Dr Katharina Fabricius. Australian Institute of Marine Science 'Water quality guidelines for the Great Barrier Reef' https://eatlas.org.au/content/water-quality-guidelines-great-barrier-reef (https://eatlas.org.au/content/water-quality-guidelines-great-barrier-reef)