EEMB146 Lab 4 Homework

Written by Rafael Romero

2025-02-05

Contents

Homework 4 – Comparing Mean with Parametric Tests

This homework will apply your data visualization skills (Lab 3), ability to determine normality and appropriate t-test (Lab 4). If you are having trouble with RStudio or knitting your .Rmd file please speak with a TA before the due date. **All of the information and code needed to answer homework questions can be found in Lab 4 Exercise file.** You will be graded on completeness and correctness.

Homework Questions

Question 1

Please answer **True or False** for 1a-e.

a) If you are comparing the mean of three groups (i.e. cats, dogs, fish), can you appropriately apply a t-test?

FALSE

b) If the distribution of the data looks normal (i.e. bell curve) and the variances between your groups are equal, can you use a parametric t-test to test the means between the groups?

TRUE

c) If the distribution of the data looks not-normal (i.e. not a bell curve), but with a transformation (log or sqrt) the data distribution becomes normal, can you run a parametric test? In this example, the assumption is that your variances are equal.

TRUE

d) Can the Central Limit Theorem be applied to a sample size of 20 (n = 20)?

TRUE

e) The null hypothesis of the Shaprio-Wilk test is that your data are not normally distributed.

FALSE

Question 2

Apply the correct t-test (**one-sample, paired, two sample**) to the situation

a) If the the null hypothesis of the test is $H_0 : \mu_d = 0$

paired t-test

b) If the data are normal and have equal variances

Two-sample t-test

c) If you don't know the true variance of the population

one-sample, paired, two sample t-test

d) If you are comparing whether the mean of one group differs from the mean of another group, assuming equal variances and normally distributed data.

two sample t-test

For Question 3 & 4, it will be very important that you choose the most appropriate t-test for each data set. If you are unsure which t-test to perform please look through Lab 4 Exercise

Question 3

For Question 3, you will use the data set `pine_trees` which is a subset of data from Bren Professor Dr. Joan Dudney's research paper titled, "Nonlinear shifts in infectious rust disease due to climate change", which is also an inspo for figure visualization - go check it out!

White blister rust (referred to as Rust) is an infectious disease that impacts trees in the southern Sierra Nevada. For homework, you will need to compare the `mean_length` (mm) of pine needles from trees infected or uninfected with `Rust`. In this data set `Rust` is the grouping column with "Yes" referring to a tree being infected and "No" referring to a tree being uninfected.

a) Format `pine_tree` data into two data frames `Rust` and `NoRust`

- hint: use filter()

```
Rust <- pine_trees %>%
  filter(Rust == "Yes")

NoRust <- pine_trees %>%
  filter(Rust == "No")
```

b) State the null and alternative hypotheses of your chosen t-test. Must state which t-test you are choosing for full credit.

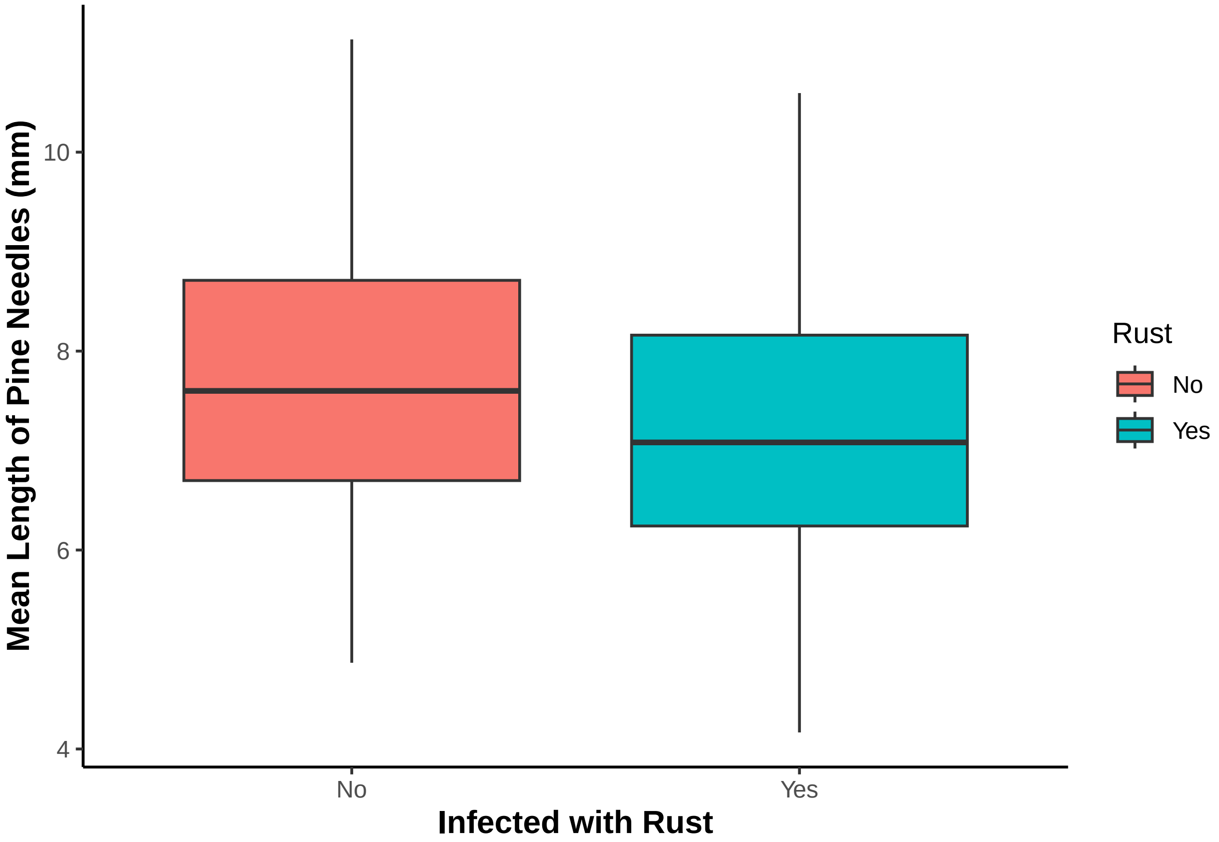
Answer: The t-test we are using is a *two-sample t-test*

$H_0 = \mu$ Rust $= \mu$ NoRust

$H_1 = \mu$ Rust != μ NoRust

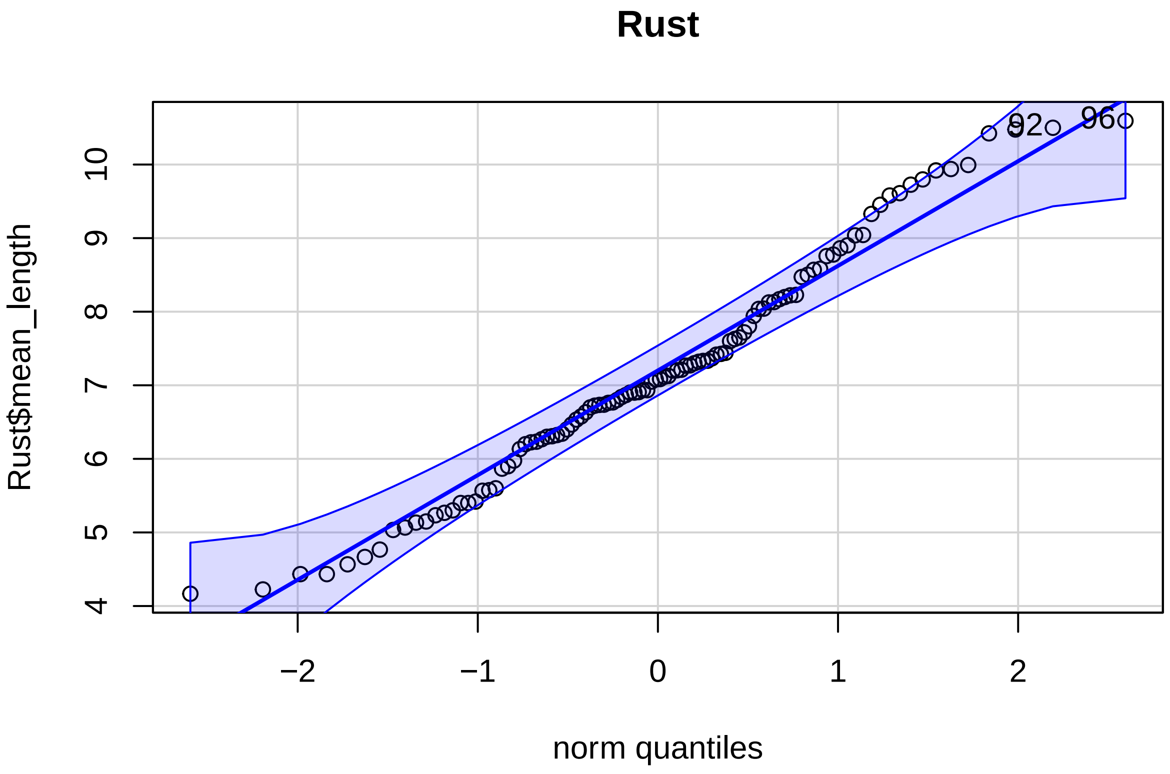
c) Determine the normality of `pine_tree` dataset with a boxplot(). Must include visualization of boxplot and include informative labels to receive full credit.

```
ggplot(pine_trees,
       aes(x=Rust, y=mean_length,fill=Rust)) +
  geom_boxplot() +
  theme_classic() +
  labs(x="Infected with Rust",y="Mean Length of Pine Needles (mm)") +
  theme(axis.title = element_text(size = 12, face = "bold"))
```



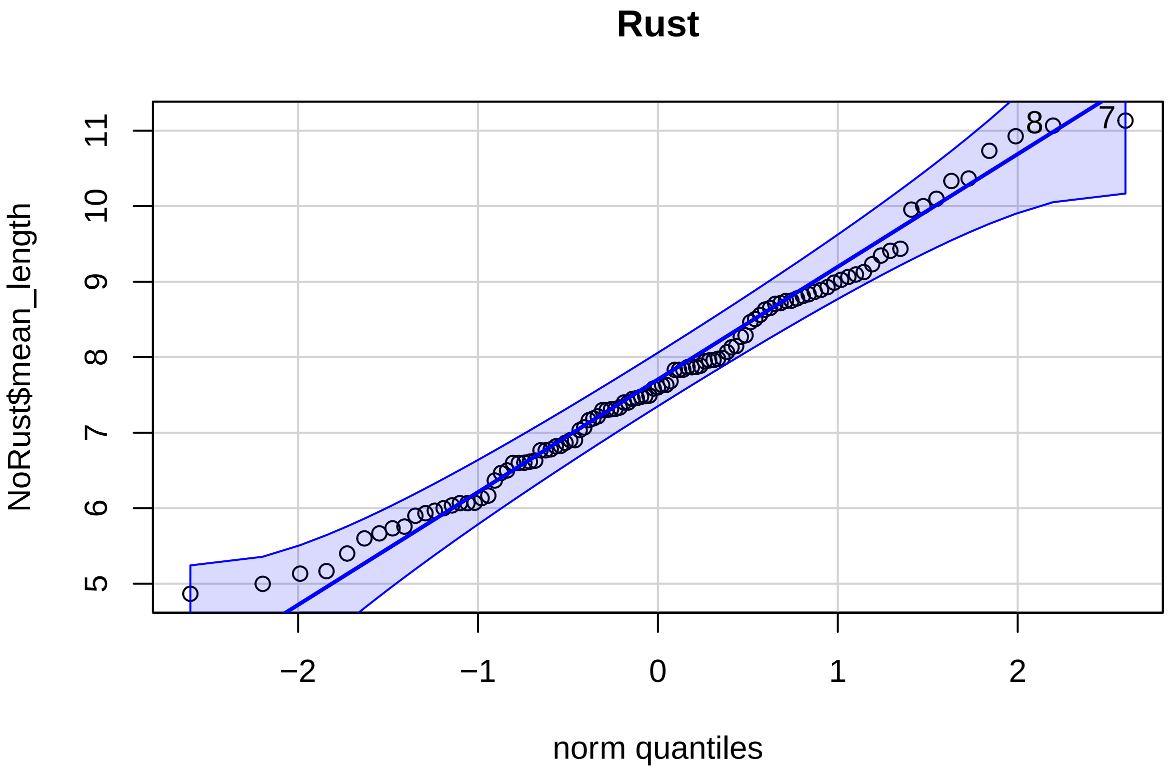
d) Determine the normality of `Rust` and `NoRust` with a qqPlot(). Must include visualization of both qqPlots to receive full credit.

```
qqPlot(Rust$mean_length, main = "Rust")
```



```
## [1] 96 92
```

```
qqPlot(NoRust$mean_length, main = "Rust")
```



```
## [1] 7 8
```

The qqPlots both look normal since they follow the 45 degree line well

e) Test normality with a Shapiro-Wilk test for the Rust and NoRust groups. Report the p-value for each group (infected vs. uninfected). What does each p-value mean in terms of the null hypothesis (ie normal or not)?

```
shapiro.test(Rust$mean_length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Rust$mean_length
## W = 0.97943, p-value = 0.09904
```

```
shapiro.test(NoRust$mean_length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NoRust$mean_length
## W = 0.98455, p-value = 0.2519
```

Answer P-value for Rust is 0.09904, P-value for NoRust is 0.2519. A p-value greater than 0.05 means that the data is normal since we fail to reject our null hypothesis of normality.

f) Base on steps c-e, do you consider the data to be normally distributed and why?

Answer: The data is considered approximately normal because the p values of both are greater than 0.05 in the Shapiro-Wilk normality test, the data closely follows the reference lines in both qqPlots, and in the boxplot we can see the median line centered in the middle.

g) Test for equal variances between rust and no rust groups with a Levene Test. What does each p-value mean in terms of the null hypothesis (ie equal or not equal)?

```
leveneTest(pine_trees$mean_length, pine_trees$Rust) #If p < 0.05 we reject the null hypothesis, if p >
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.5733 0.4498
##       211
```

Answer P-value is 0.4498 which is greater than 0.05, so we can fail to reject the null hypothesis and conclude that the variance between the two groups are equal.

h) Run the correct code for your chosen t-test

We use a two-sample t-test since data is normal and variances are equal between the two groups

$H_0 = \mu$ Rust $= \mu$ NoRust

$H_1 = \mu$ Rust $!= \mu$ NoRust

```
t.test(Rust$mean_length, NoRust$mean_length, var.equal = TRUE) #variance equal to true because levene t
```

```
##
##  Two Sample t-test
##
## data:  Rust$mean_length and NoRust$mean_length
## t = -2.4713, df = 211, p-value = 0.01425
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.8976090 -0.1010337
## sample estimates:
## mean of x mean of y
##  7.183783  7.683104
```

i) Based on the results from your t-test, what can you conclude about the difference in mean needle length (mm) between infected and uninfected trees? To receive full credit, your response must include the p-value from your t-test, report the 95% confidence interval, and state the mean length (mm) for infected and uninfected groups.

Answer: Because our data met the normality (Shapiro-Wilk p-value > 0.05) and variance (Levene's Test p-value > 0.05) assumptions we can run a two sample t-test to test the difference in mean length of pine needles between trees with Rust and No Rust. With our hypotheses

$H_0 = \mu$ Rust $= \mu$ NoRust

$H_1 = \mu$ Rust $!= \mu$ NoRust

and a p-value of 0.01425, we can reject the null hypothesis since it is less than our significance level of 0.05. This tells us that there is a significant difference in the mean length of pine needles between trees with rust and no rust. The 95 percent confidence interval also further supports our claim since it does not contain the value 0 meaning that there exists a difference between the means. In conclusion, there exists a signficant difference in the mean length of trees with Rust and No Rust.

Question 4

For Question 4, you will use the data set `greenhouse`. This data measures the photosynthetic performance (FvFm) of 10 plants from two environments (sunny and shady) in a greenhouse. To reduce variation among measurements, performance of each plant was measured twice, once in the sun and once in the shade, which can be considered paired data points. The measurements are belonging to the same individual plant.

a) Format 'greenhouse' data into two data frames `shady` and `sunny`

- hint: use filter()

```
shady <- greenhouse%>% filter(treatment == "Shady")
sunny <- greenhouse%>% filter(treatment == "Sunny")
```

b) State the null and alternative hypotheses of your chosen t-test. Must state which t-test you are choosing for full credit.

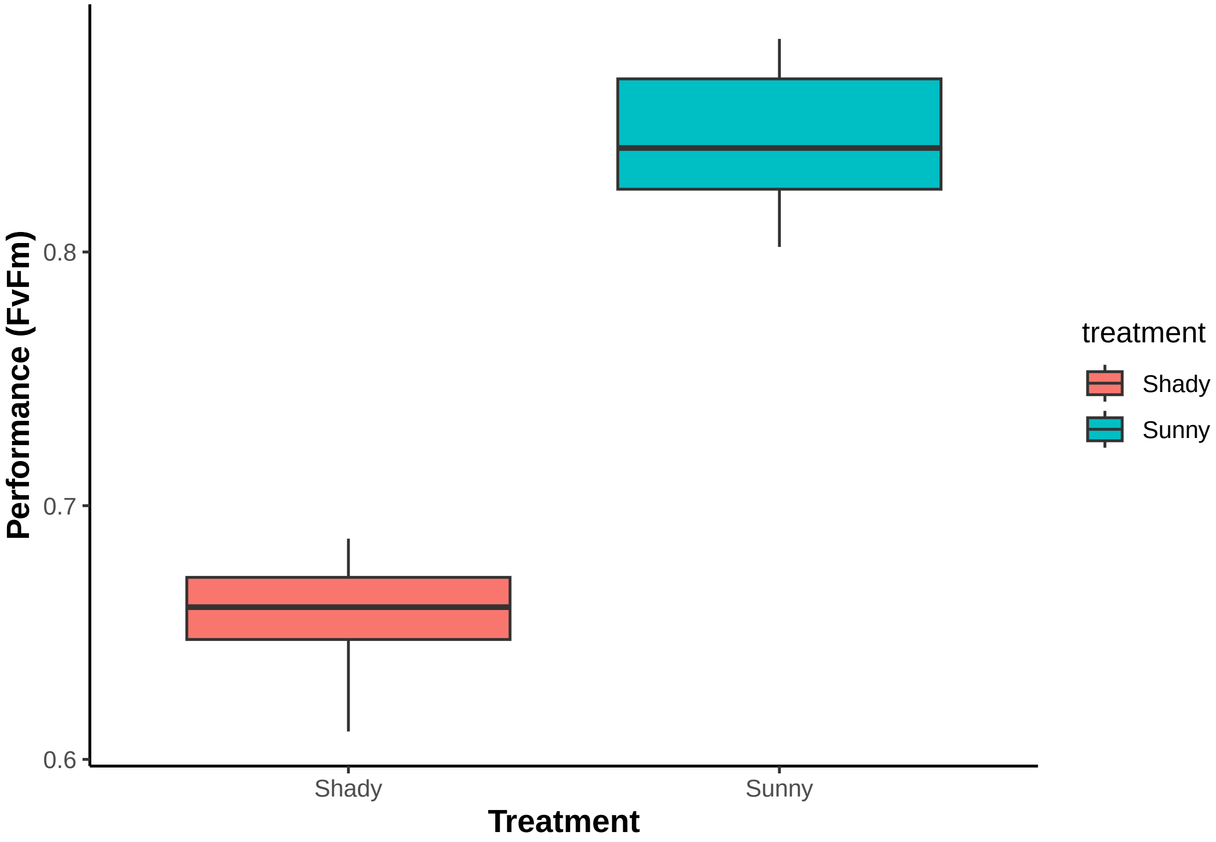
Answer: The t-test that we are using is a paired t-test

$H_0 = \mu$ Shady $= \mu$ Sunny

$H_1 = \mu$ Shady $!= \mu$ Sunny

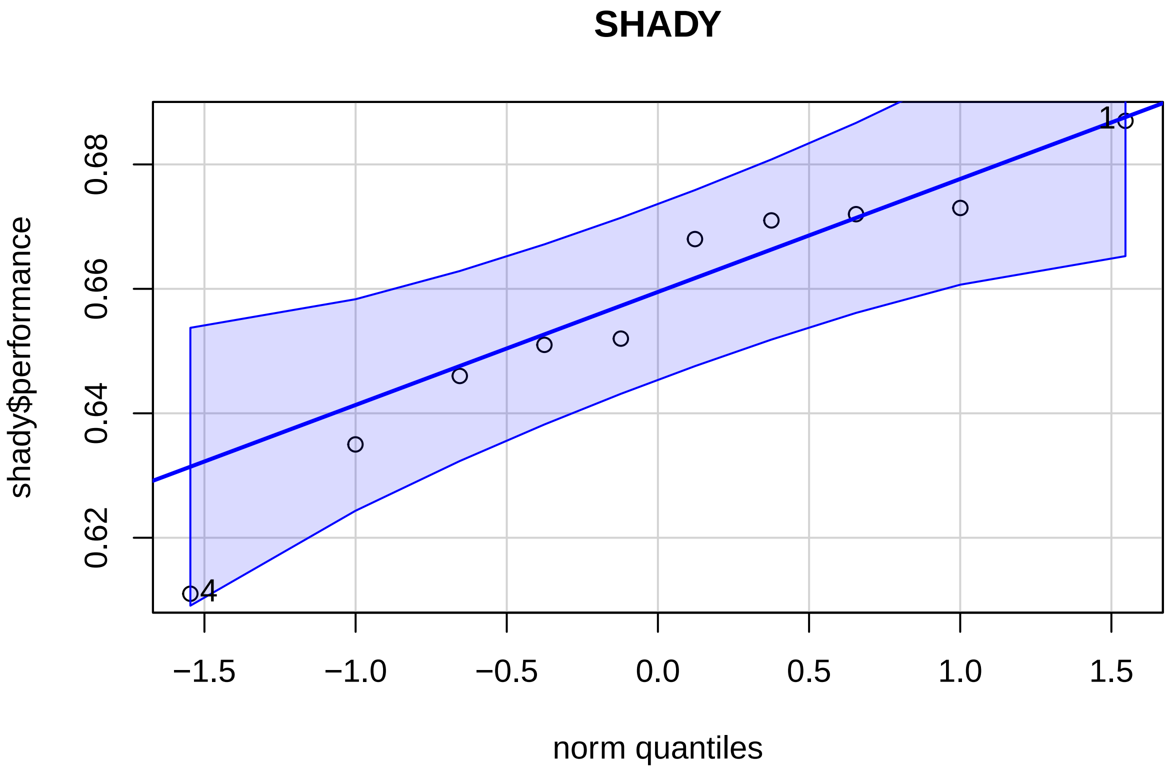
c) Determine the normality of `greenhouse` data set with a boxplot(). Must include visualization of boxplot and include informative labels to receive full credit.

```
ggplot(greenhouse,
       aes(x=treatment, y=performance,fill= treatment)) +
  geom_boxplot() +
  theme_classic() +
  labs(x="Treatment",y="Performance (FvFm)") +
  theme(axis.title = element_text(size = 12, face = "bold"))
```



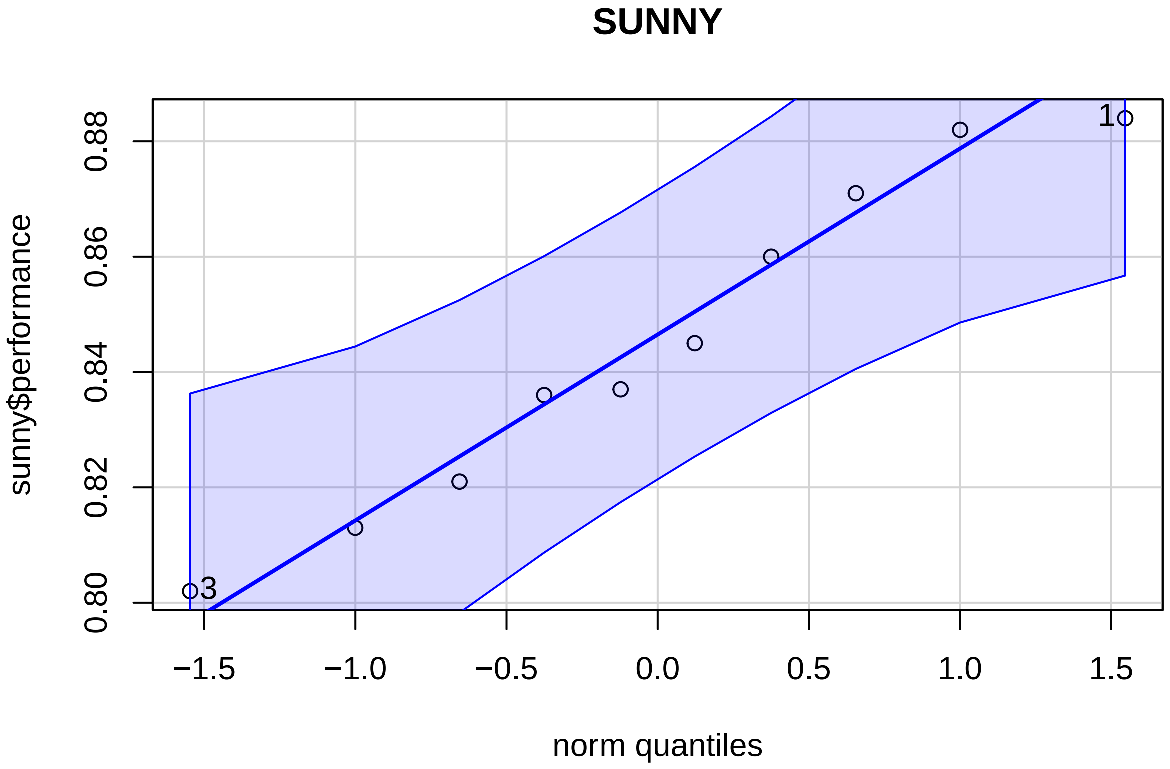
d) Determine the normality of `shady` and `sunny` with a qqPlot(). Must include visualization of both qqPlots to receive full credit.

```
qqPlot(shady$performance, main = "SHADY")
```



```
## [1] 4 1
```

```
qqPlot(sunny$performance, main = "SUNNY")
```



```
## [1] 3 1
```

The qqPlots both look normal since they follow the 45 degree line well

e) Test normality with a Shapiro-Wilk test. Report the p-value for each grouping (shady or sunny). What does each p-value mean in terms of the null hypothesis (ie normal or not)?

```r
shapiro.test(shady$performance)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shady$performance
## W = 0.93685, p-value = 0.5186
```

```r
shapiro.test(sunny$performance)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sunny$performance
## W = 0.9493, p-value = 0.6602
```

Answer: P-value for shady is 0.5186, P-value for Sunny is 0.6602. A p value greater than 0.05 means that the data is normal since we fail to reject our null hypothesis of normality.

f) Base on steps c-e, do you consider the data to be normally distributed and why?

Answer: The data is considered approximately normal because the p values of both are greater than 0.05 in the Shapiro-Wilk normality test, the data closely follows the reference lines in both qqPlots, and in the boxplot we can see the median line centered in the middle of both box plots.

g) Run the correct code for your chosen t-test

```r
t.test(shady$performance, sunny$performance, paired= TRUE)
```

```
##
##  Paired t-test
##
## data:  shady$performance and sunny$performance
## t = -18.812, df = 9, p-value = 1.557e-08
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.2111674 -0.1658326
## sample estimates:
## mean difference
##         -0.1885
```

h) Based on the results from your t-test, what can you conclude about photosynthetic performance for the shady and sunny group? To receive full credit, in your response must include the p-value from your t-test and the sample mean difference.

Answer: We did a paired t-test since the data was approximately normal. Our p-value was 1.557e-08 and the sample mean difference was -0.1885. Since the p-value is less than 0.05 we can reject the null hypothesis of there being no difference between the two groups. Therefore, there is strong proof that a significant difference in the means between the group in the shade and in the sun. The difference of -0.1885 tells us that performance in shady conditions is signfically lower than in sunny conditions.

End of Homework 4