# SectionC

EMATM0061,TB1 2022

YujieWang

2023-01-09

# Contents

# Abstract

In this paper, the case study the unpaired Student's approach to the real-world data and relevant concepts.The results of the experiment indicate that the sample size would make a difference in the results of effect size and different types of samples have different effect sizes.The studies performed showed that the sepal width is better used to distinguish the Iris versicolor and Iris setosa than the sepal length.


# Introduction

Hypothesis testing is a systematic approach to drawing inferences from data.People can justify the inferences from data,understand and control the role of statistical variation by combing hypothesis testing with an appropriate experimental design. The main purpose of this study is to develop an understanding of the unpaired Student's t-test.In this paper,the hypothesis will be tested and this work takes the form of a case-study of the Iris data.

The overall structure of the study takes the form of four parts,including the description of key concepts,the simulation study,the application of the real-world data and the discusstion of results.

# Section C–An unpaired Student's test

## The decription of the hypothesis test's key concepts

The purpose of the Section C is to focus on the unpaired Student's t-test.First of all,a test statistic is an essential part of the hypothesis test,which is some function of the data used within a statistical hypothesis test. The test statistic must have a known distribution under the null hypothesis $H_0$. In addition, the test statistic should emphasize differences between null and alternative hypothesis. It often takes on large or "extreme" values under the alternative hypothesis $H_1$.Then by looking at the value of the test statistic, The conclusions can be drawn on our test:

Case1: The test statistic takes on typical values for $H_0 \rightarrow$ the experiment fails to rejects the null hypothesis $H_0$.

Case2: The test statistic takes on non-typical values for $H_0 \rightarrow$ the experiment rejects the null & accepts the alternative $H_1$.

Therefore,the test statistic would be useful to distinguish between the null and the alternative.

In the unpaired Student's t-test, the test statistic is that:

Suppose that $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d.random variables.

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

where $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1}(X_i - \bar{X})^2$.

Then the random variables is $t$-distributed with $n-1$ degree of freedom.

Apart from the test statistic, the underlying assumptions frames the research question in terms of the parameters of a statistical model. And there are two hypotheses:

(1) $H_0$ :The null hypothesis is the default position in a statistical hypothesis typically declaring an absence of an interesting phenomenon, for example the equality of two statistical parameters.

(2) $H_1$ :The alternative hypothesis is a statistical hypothesis which contradicts the null hypothesis and typically declares the presence of some interesting phenomenon, often consistent with the research hypothesis a scientist is attempting to prove. For example, a difference in the values of two statistical parameters.

The p-value and the significance level are also the important factors to study the hypothesis test.The p-value is probability under the null hypothesis that the test statistic will achieve a value as extreme or more extreme than the value which is actually observed. A very small p-value indicates that the observed data is sufficiently inconsistent with the null hypothesis that the null hypothesis can be reasonably rejected.And the significance level is an upper bound on the size of the test. This should be chosen in advance of seeing the data.Typically, the value is $\alpha = 0.05$.

## Perform the simulation study

In this part, a simulation study will be performed to investigate the probability of `Type I error` under the null hypothesis for the unpaired Student's t-test. It should be noted that the unpaired data means that there are two samples $X_1, \cdots, X_n$ (which are i.i.d. copies of $X$) and $Y_1, \cdots, Y_k$ (which are i.i.d. copies of $Y$) where $n$ may not be equal to $k$.Also, a `Type I error` is a rejection of the null hypothesis in favor of the alternative hypothesis when the null hypothsis is true.The size of a test is the probability of `Type I error`. A important property of valid statistical hypothesis tests with a certain significance level is that the size of the test does not exceed the significance level.

Before the simulation study,the package should be loaded:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

In this simulation study: 1. Firstly, the random data should be generated which includes two samples(sample0,sample1),the sample0 and sample1 are the random Guassian variables.They should be generated by the function `rnorm()`, and the values of $\mu$ and $sigma$ should be set.

2. Secondly, the case apply unpaired Student's t-test with significance level alpha on the sample0 and sample1, and extract the p-value.Due to the Student's t-test, it means that $\sigma_0 = \sigma_1$, the value of `var.equal` should be `TRUE`.

3. Thirdly, compute the value of `Type I error` which should be either 0 or 1, if p-value $< \alpha$, the `Type I error` occurs,whose value would be 1.

4. Then, the function `mean()` is used to compute the probability that `Type I error` occurs.

```
num_trials<-10000
sample_size1<-50
sample_size2<-60
mu_0<-1
mu_1<-1
```

```
sigma_0<-5
sigma_1<-5
alpha<-0.05
set.seed(0) #set random seed for reproducibility

simulation_df<-data.frame(trial=seq(num_trials))%>% # generate random Gaussian samples
  mutate(sample0=map(.x=trial,.f=~rnorm(n=sample_size1,mean=mu_0,sd=sigma_0)),
         sample1=map(.x=trial,.f=~rnorm(n=sample_size2,mean=mu_1,sd=sigma_1)))%>%
  # generate p values with t.test() function
  mutate(p_value=pmap(.l=list(trial,sample0,sample1),
                      .f=~t.test(..2,..3,var.equal=TRUE)$p.value))%>%
  # compute the value of type I error
  mutate(type_1_error=p_value<alpha)

simulation_df%>%
  pull(type_1_error)%>%
  mean() #estimate of converge probability
```

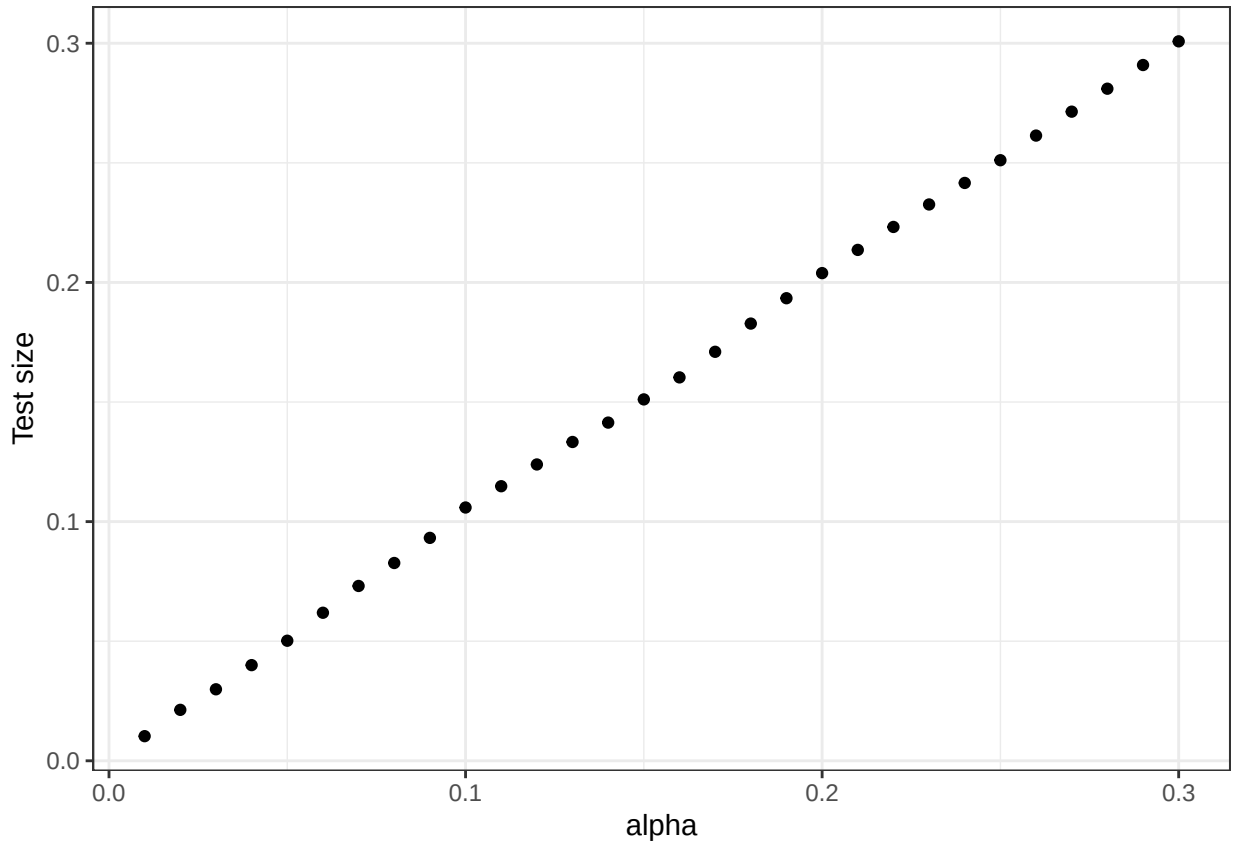
```
## [1] 0.0502
```

5. Finally,the case explores how the size of the test varies as a function of the significance level $\alpha$.

```
alpha_list = seq(0.01, 0.3, 0.01) # generate a list of alpha
compute_test_size <- function(alpha){
# compute the value of type I error
type_1_error = simulation_df$p_value<alpha
return (mean(type_1_error)) # estimate of coverage probability
}

# create the new data frame which includes the alpha
# and the corresponding probabiltiy of type 1 error
multiple_alpha_simulation<-data.frame(alpha=alpha_list)%>%
  mutate(test_size=map_dbl(alpha,compute_test_size))

# draw a plot of the new data frame by using the function geom_point()
multiple_alpha_simulation%>%ggplot(aes(x=alpha,y=test_size))+
  geom_point()+ylab('Test size')+theme_bw()
```




In this case, as the chart shown, the test size is almost equal to the significance level.

### Implementing the unpaired Student's t-test

In this case, the experiment uses the Iris data to implement the unpaired Student's t-test. The data set consists of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.The data set contains a set of 148 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class(Species).And The Iris Setosa and Iris virginca separately consists of 50 samples,the Iris versicolor consists of 48 samples.

The code below shows the structure of the data.

```
iris_data<-read.csv("IRIS.csv") # load the data from the csv
iris_data<-iris_data%>%drop_na() # strip the na values
iris_data%>%head(10)
```

```
##    sepal_length sepal_width petal_length petal_width     species
## 1           5.1         3.5          1.4         0.2 Iris-setosa
## 2           4.9         3.0          1.4         0.2 Iris-setosa
## 3           4.7         3.2          1.3         0.2 Iris-setosa
## 4           4.6         3.1          1.5         0.2 Iris-setosa
## 5           5.0         3.6          1.4         0.2 Iris-setosa
## 6           5.4         3.9          1.7         0.4 Iris-setosa
## 7           4.6         3.4          1.4         0.3 Iris-setosa
```

```
## 8          5.0         3.4          1.5         0.2 Iris-setosa
## 9          4.4         2.9          1.4         0.2 Iris-setosa
## 10         4.9         3.1          1.5         0.1 Iris-setosa
```

**The Hypothesis test-unpaired Student's t-test:**

Suppose that the biologist is investigating the features of the different iris flowers.

Then there is a question that the difference between the average sepal lengths of the Iris-versico and Iris-setosa.Therefore, the case draws a random sample of Iris-versico with measuring their lengths $X_1, \cdots, X_n$,and draws a random sample of Iris-setosa with measuring their lengths $Y_1, \cdots, Y_k$.($n$ and $k$ can be different).

It should be considered that the prerequisites for using Student's t-test is: (1) $X_1, \cdots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ are i.i.d. and $Y_1, \cdots, Y_k \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are i.i.d. (2) $\sigma_1 = \sigma_2$

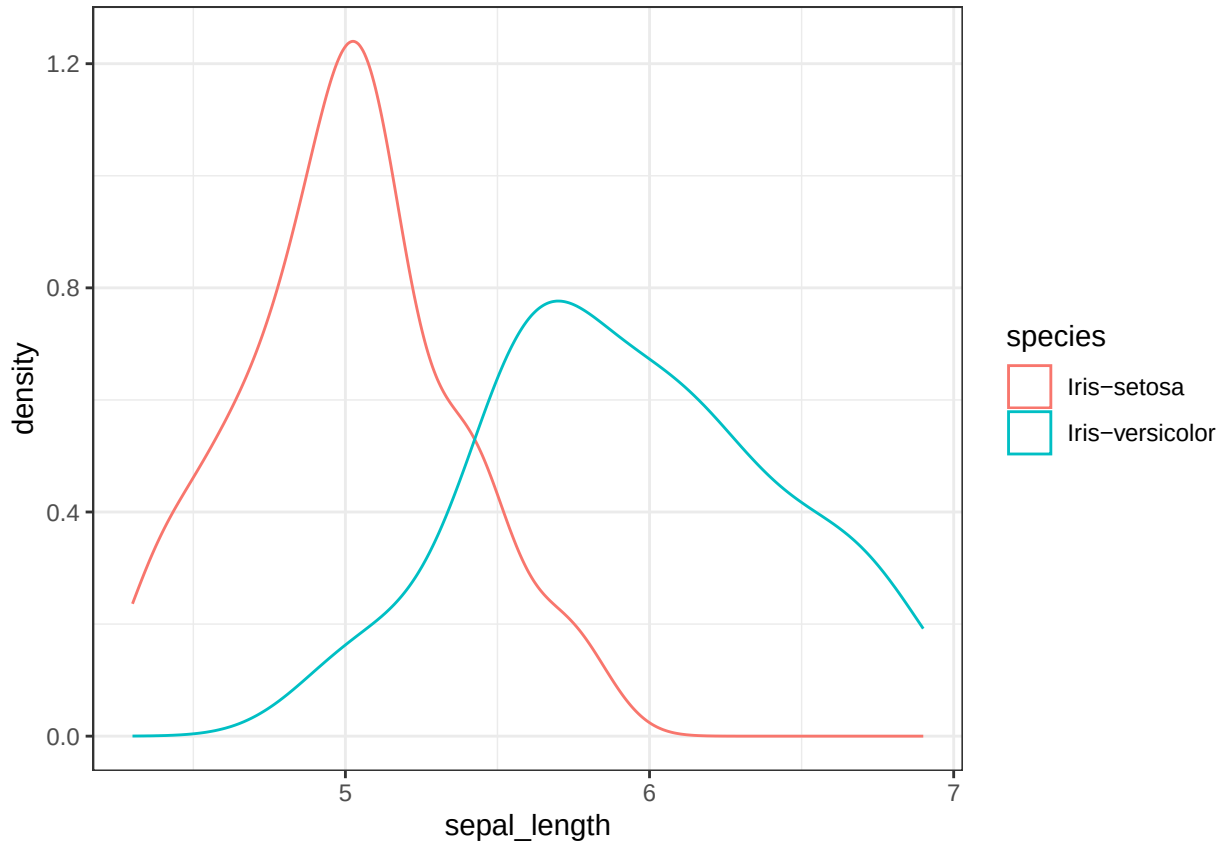
1. Form the null and alternative hypothesis:

   (1) Null hypothesis: $H_0 : \mu_1 = \mu_2$
   (2) Alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

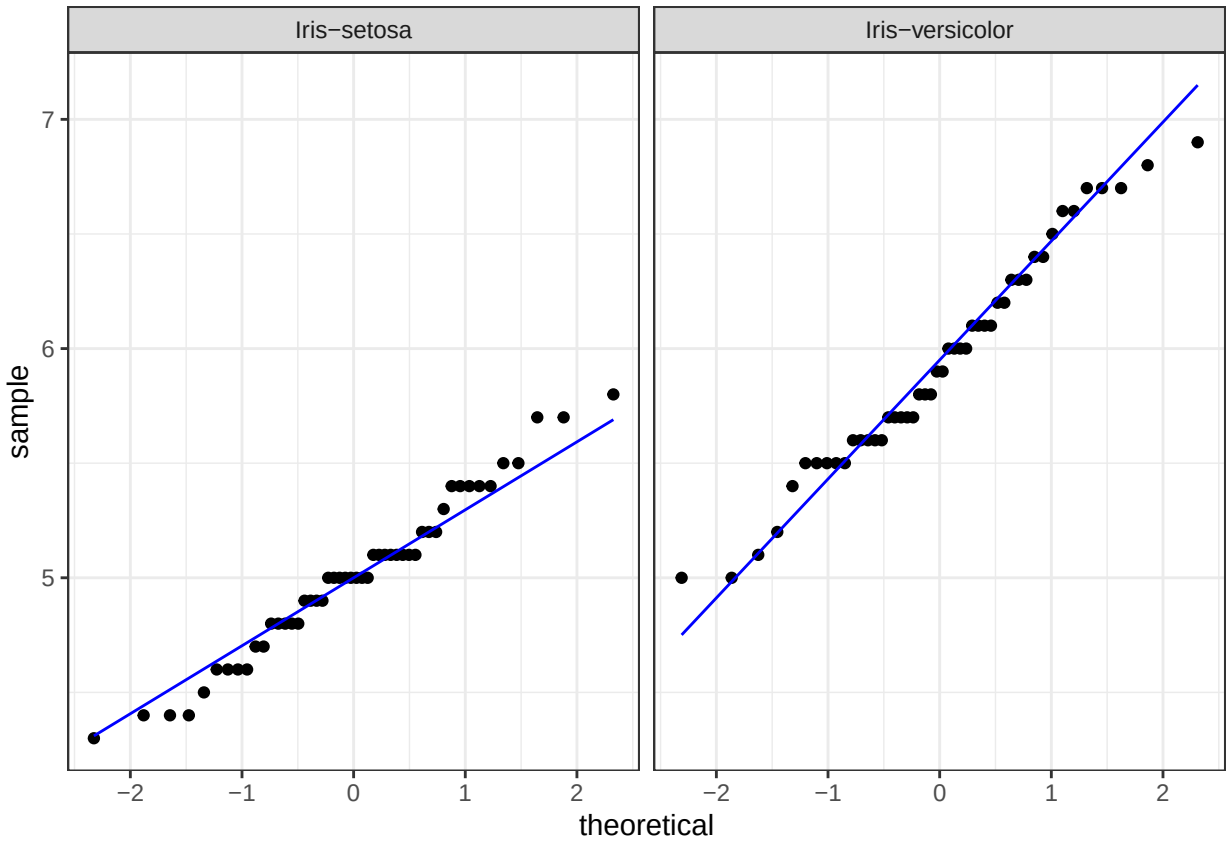
2. Validate modelling assumptions:

```
# create the new data frame including both samples
sepal_length_data<-iris_data%>%
  filter(species=="Iris-versicolor"|species=="Iris-setosa")%>%
  select(species,sepal_length)

# draw a plot of the density distribution of both samples.
ggplot()+
  geom_density(data=sepal_length_data,aes(x=sepal_length,color=species))+
  theme_bw()+xlab("sepal_length")+ylab("density")
```




```
# draw the quantile-quantile plot
sepal_length_data%>%ggplot(aes(sample=sepal_length))+
  stat_qq()+stat_qq_line(color="blue")+
  theme_bw()+
  facet_wrap(~species)+xlab("theoretical")+ylab("sample")
```




According to the chart shown, the distributions for both samples are approximately Gaussian.

The data has n=50,k=48,hence,by the central limit theorem, the sample means will be approximately Gaussian.

3. Select a significance level: $\alpha = 0.05$
4. Select an appropriate statistical test-Student's t-test statistic:
   - For $X_1, \cdots, X_n$, let the sample mean and variance be defined as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ , respectively.
   - Similarly, for $Y_1, \cdots, Y_k$, we can define the sample mean $Y$ and sample variance $S_Y^2$.
   - Student's t-test statistic:

$$\hat{T} := \frac{\bar{X} - \bar{Y}}{S_{X,Y}\sqrt{\frac{1}{n} + \frac{1}{k}}} \text{ where } S_{X,Y}^2 = \frac{(n-1)S_X^2 + (k-1)S_Y^2}{n + k - 2}$$

   Assume that $\mu_1 = \mu_2$. Then $\hat{T}$ is t-distributed with $n + k - 2$ degrees of freedom.

5. Compute the numerical value $\tau$ of $\hat{T}$.

```
mean_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of X
mean_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of Y
```

```
sd_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) # sd of X
sd_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) #sd of Y
n<-length(sepal_length_data%>%filter(species=="Iris-setosa")%>%
            pull(sepal_length)%>%na.omit()) #n.omit: strip the null numbers
k<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%nrow()

s_xy<-sqrt(((n-1)*sd_X^2+(k-1)*sd_Y^2)/(n+k-2))
t_statistic<-(mean_X-mean_Y)/(s_xy*sqrt(1/n+1/k))
# compute the test statistic
t_statistic
```

```
## [1] -10.94984
```

The current study found that the value of test statistic is -10.52099.

6. Compute p-value based on the test statistic:

$$p := \mathbb{P}(|\hat{T}| > |\tau||H_0) = 2 \cdot (1 - F_{n+k-2}(|\tau|))$$

```
p_value<-2*(1-pt(abs(t_statistic),df=n+k-2)) #compute the p-value
n
```

```
## [1] 50
```

```
k
```

```
## [1] 48
```

```
t_statistic
```

```
## [1] -10.94984
```

```
p_value
```

```
## [1] 0
```

The results of this study show lindicate that the values of n,k,test statistic and the p-value.

7. Compute the effect size:a measure of the magnitude of the observed phenomenon which reflects the extent to which the null hypothesis is false.This is interesting if the case rejected the null and established an effect.A large effect size means that a research finding has practical significance, while a small effect size indicates limited practical applications.

Under such circumstances of unpaired data, the effect size is computed by the formula:

$$d := \frac{\bar{X} - \bar{Y}}{S_{X,Y}} \text{ where } S^2_{X,Y} = \frac{(n-1)S^2_X + (k-1)S^2_Y}{n+k-2}$$

```
effect_size<-(mean_X-mean_Y)/s_xy
effect_size
```

```
## [1] -2.212662
```

8. Unpaired Student's t-test with R. Finally, the values of hypothesis test could be computed by the function `t.test()`, the equality of $\sigma$ should be considered in this case.

```
t.test(sepal_length~species,data=sepal_length_data,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sepal_length by species
## t = -10.95, df = 96, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Iris-setosa and group Iris-versicolor
## 95 percent confidence interval:
##  -1.0979011 -0.7609322
## sample estimates:
##     mean in group Iris-setosa mean in group Iris-versicolor
##                         5.006                         5.935417
```

## Conclusion and discussion

The results shows that the p-value is close to 0,which is less than $\alpha$,and the effect size of the data is -2.21.This observation may support the hypothesis that the true difference between the group Iris-setosa and the group Iris-versicolor.

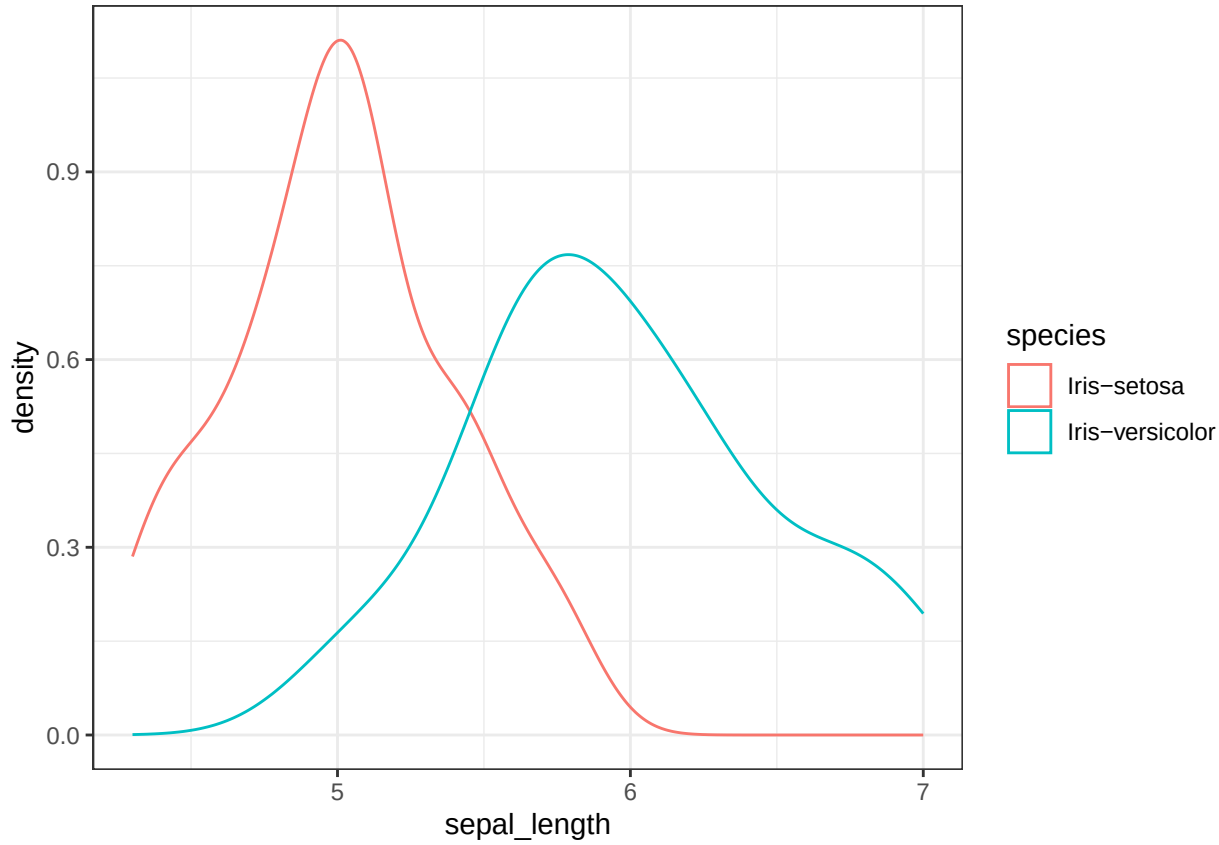
Taking it a step further, the sample size should be considered into the experiment, so the case study loads the data "IRIS1.csv" to get the less data of two samples to compare the results.

```
iris_data<-read.csv("IRIS1.csv") # load the data from the csv
iris_data<-iris_data%>%drop_na() # strip the na values
iris_data%>%head(10)
```

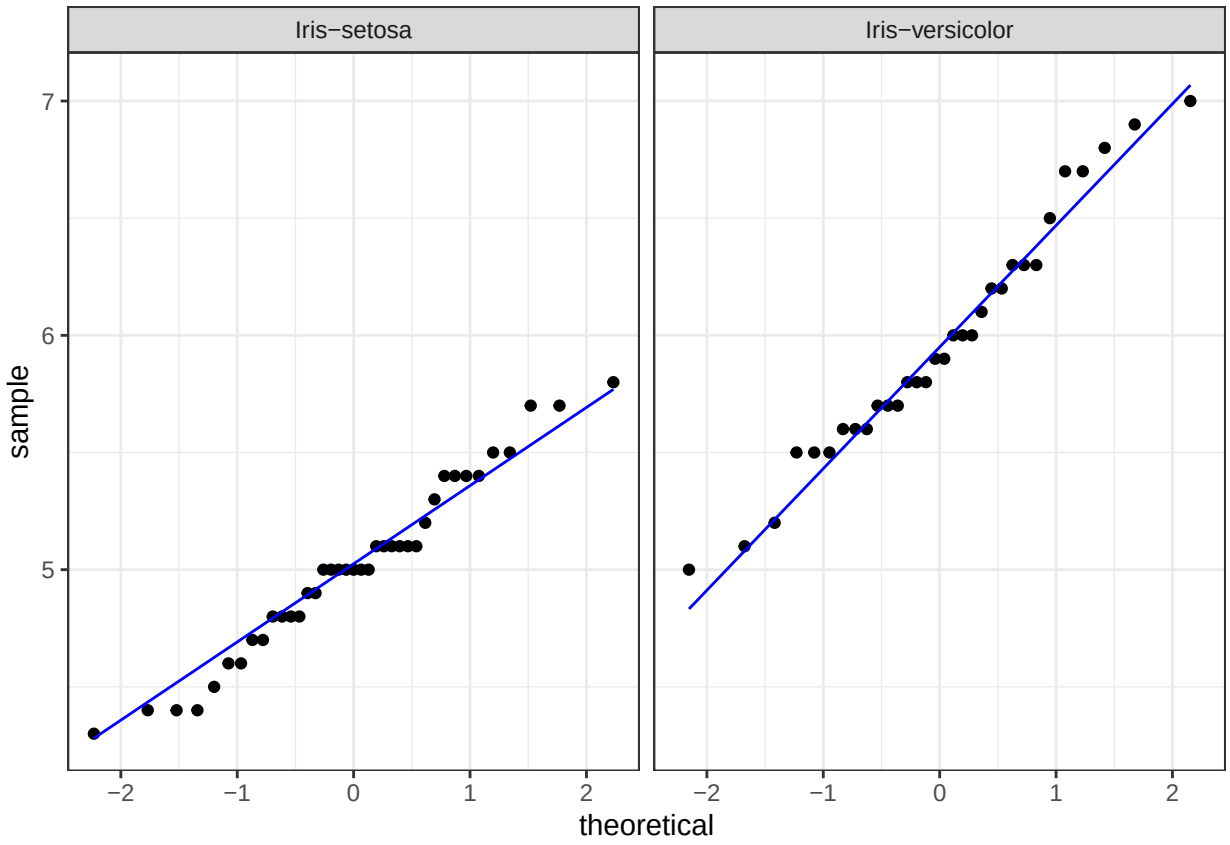
```
##    sepal_length sepal_width petal_length petal_width     species
## 1           5.1         3.5          1.4         0.2 Iris-setosa
## 2           4.7         3.2          1.3         0.2 Iris-setosa
## 3           5.0         3.6          1.4         0.2 Iris-setosa
## 4           5.4         3.9          1.7         0.4 Iris-setosa
## 5           4.6         3.4          1.4         0.3 Iris-setosa
## 6           5.0         3.4          1.5         0.2 Iris-setosa
## 7           4.4         2.9          1.4         0.2 Iris-setosa
## 8           5.4         3.7          1.5         0.2 Iris-setosa
## 9           4.8         3.4          1.6         0.2 Iris-setosa
## 10          4.3         3.0          1.1         0.1 Iris-setosa
```

```
# create the new data frame including both samples
sepal_length_data<-iris_data%>%
  filter(species=="Iris-versicolor"|species=="Iris-setosa")%>%
  select(species,sepal_length)
```

```
# draw a plot of the density distribution of both samples.
ggplot()+
  geom_density(data=sepal_length_data,aes(x=sepal_length,color=species))+
  theme_bw()+xlab("sepal_length")+ylab("density")
```




```
# draw the quantile-quantile plot
sepal_length_data%>%ggplot(aes(sample=sepal_length))+
  stat_qq()+stat_qq_line(color="blue")+
  theme_bw()+
  facet_wrap(~species)+xlab("theoretical")+ylab("sample")
```




```
mean_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of X
mean_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of Y
sd_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) # sd of X
sd_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) #sd of Y
n<-length(sepal_length_data%>%filter(species=="Iris-setosa")%>%
            pull(sepal_length)%>%na.omit()) #n.omit: strip the null numbers
k<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%nrow()

s_xy<-sqrt(((n-1)*sd_X^2+(k-1)*sd_Y^2)/(n+k-2))
t_statistic<-(mean_X-mean_Y)/(s_xy*sqrt(1/n+1/k))
# compute the test statistic
t_statistic
```

```
## [1] -9.019134
```

```
p_value<-2*(1-pt(abs(t_statistic),df=n+k-2))
n
```

```
## [1] 39
```

```
k
```

```
## [1] 32
```

```
t_statistic
```

```
## [1] -9.019134
```

```
p_value
```

```
## [1] 2.753353e-13
```

```
effect_size<-(mean_X-mean_Y)/s_xy
effect_size
```

```
## [1] -2.151229
```

```
t.test(sepal_length~species,data=sepal_length_data,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sepal_length by species
## t = -9.0191, df = 69, p-value = 2.754e-13
## alternative hypothesis: true difference in means between group Iris-setosa and group Iris-versicolor
## 95 percent confidence interval:
##  -1.1635556 -0.7420533
## sample estimates:
##     mean in group Iris-setosa mean in group Iris-versicolor
##                         5.012821                       5.965625
```

It is interesting that when the difference between the number of two samples increases, the effect size increases too, moreover, the p-value also increases.The effect size's increase means that the research finding has more practical significance. It is therefore likely that such connections exist between the sample size and the accuranc of the results.

Next step, this case should increase the number of samples.

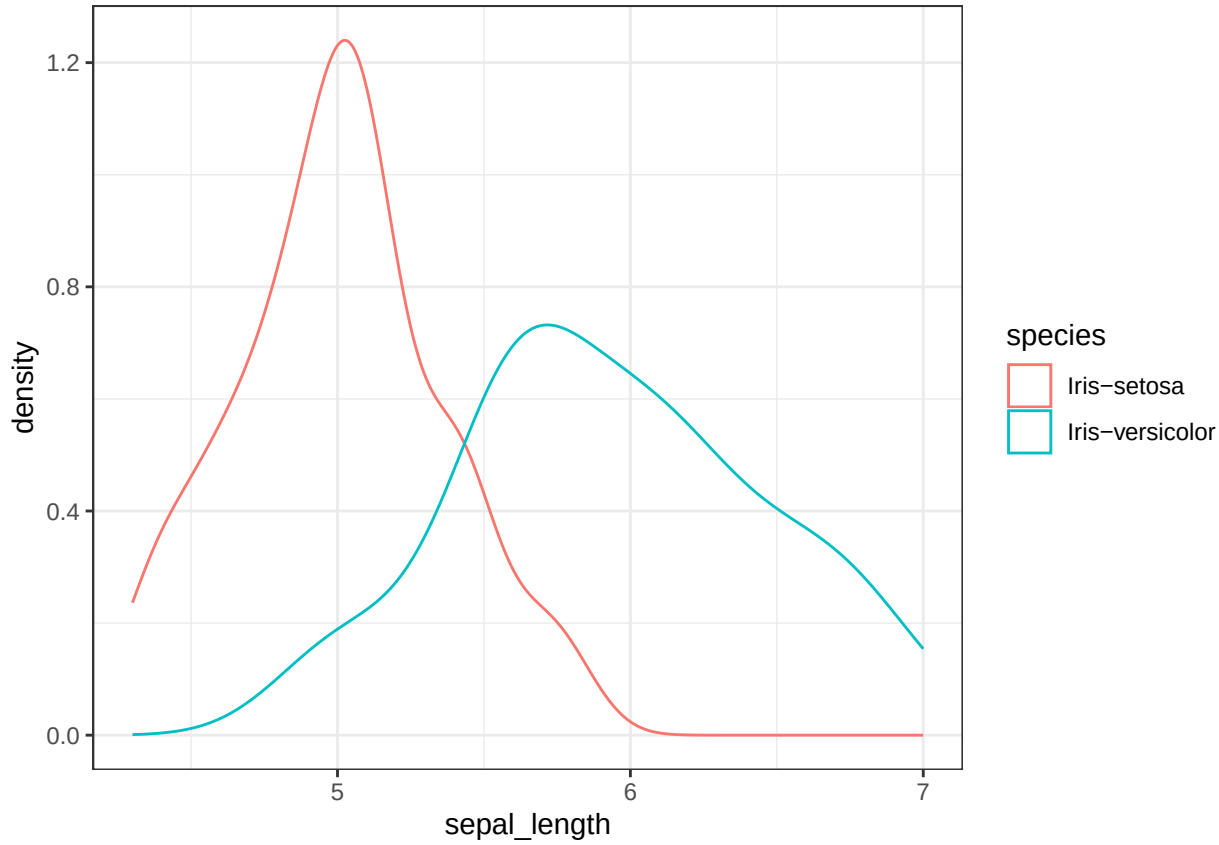
The test load the iris data which consists of the same number of different samples:

```
iris_data<-read.csv("IRIS2.csv") # load the data from the csv
iris_data<-iris_data%>%drop_na() # strip the na values
iris_data%>%head(10)
```

```
##    sepal_length sepal_width petal_length petal_width     species
## 1           5.1         3.5          1.4         0.2 Iris-setosa
## 2           4.9         3.0          1.4         0.2 Iris-setosa
## 3           4.7         3.2          1.3         0.2 Iris-setosa
## 4           4.6         3.1          1.5         0.2 Iris-setosa
## 5           5.0         3.6          1.4         0.2 Iris-setosa
```

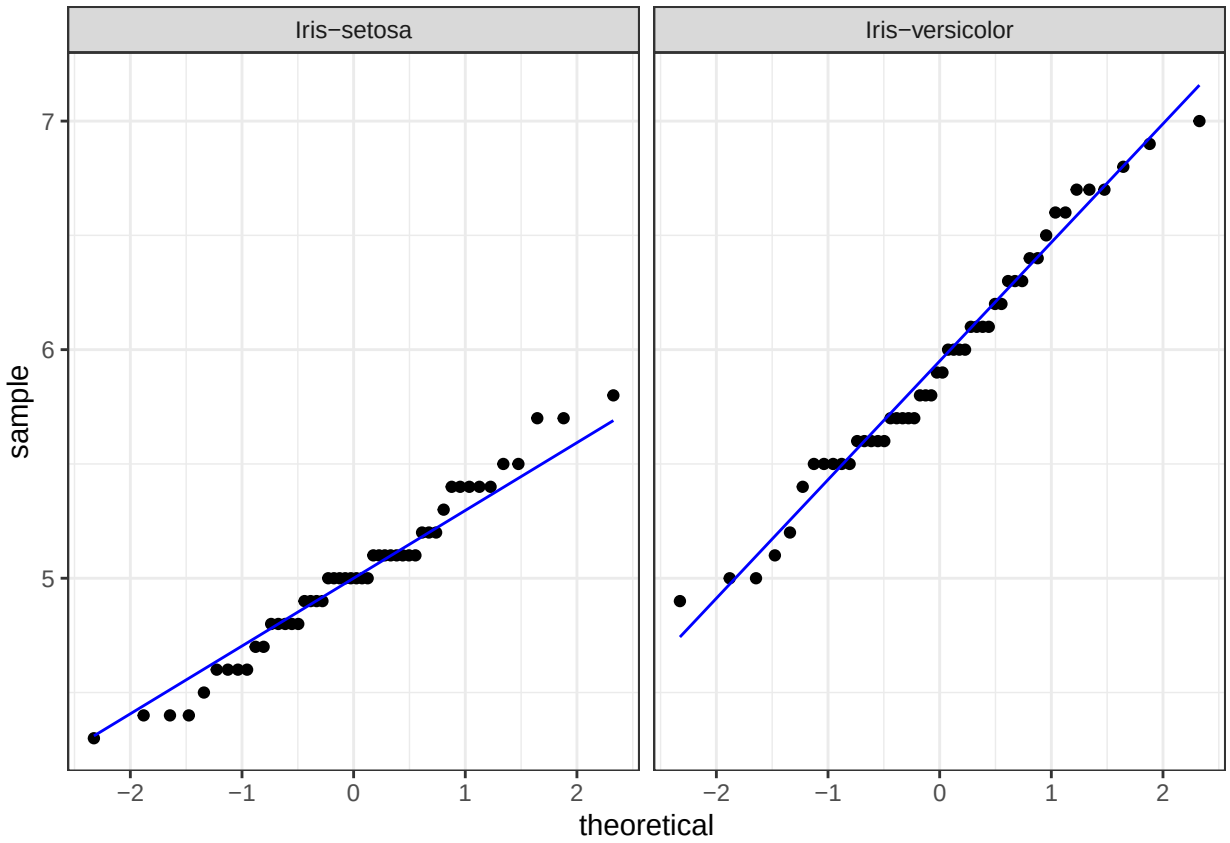
```
## 6           5.4         3.9          1.7         0.4 Iris-setosa
## 7           4.6         3.4          1.4         0.3 Iris-setosa
## 8           5.0         3.4          1.5         0.2 Iris-setosa
## 9           4.4         2.9          1.4         0.2 Iris-setosa
## 10          4.9         3.1          1.5         0.1 Iris-setosa
```

```
sepal_length_data<-iris_data%>%
  filter(species=="Iris-versicolor"|species=="Iris-setosa")%>%
  select(species,sepal_length)
# create the new data frame including both samples
ggplot()+
  geom_density(data=sepal_length_data,aes(x=sepal_length,color=species))+
  theme_bw()+xlab("sepal_length")+ylab("density")
```




```
# draw a plot of the density distribution of both samples.
```

```
# draw the quantile-quantile plot
sepal_length_data%>%ggplot(aes(sample=sepal_length))+
  stat_qq()+stat_qq_line(color="blue")+
  theme_bw()+
  facet_wrap(~species)+xlab("theoretical")+ylab("sample")
```




In this case, the data uses the unpaired statistical hypothesis test.

$$\hat{T}_W := \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{k}}} \text{ where } S_X^2 \text{ and } S_Y^2 \text{ are sample variances}$$

```
mean_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of X
mean_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%mean(na.rm=TRUE) #the mean of Y
sd_X<-sepal_length_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) # sd of X
sd_Y<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_length)%>%sd(na.rm=TRUE) #sd of Y
n<-length(sepal_length_data%>%filter(species=="Iris-setosa")%>%
            pull(sepal_length)%>%na.omit()) #n.omit: strip the null numbers
k<-sepal_length_data%>%filter(species=="Iris-versicolor")%>%nrow()

s_xy<-sqrt(((n-1)*sd_X^2+(k-1)*sd_Y^2)/(n+k-2))
t_statistic<-(mean_X-mean_Y)/(s_xy*sqrt(1/n+1/k))
# compute the test statistic
t_statistic
```

```
## [1] -10.52099
```

```
p_value<-2*(1-pt(abs(t_statistic),df=n+k-2)) # compute the p-value
n
```

```
## [1] 50
```

```
k
```

```
## [1] 50
```

```
t_statistic
```

```
## [1] -10.52099
```

```
p_value
```

```
## [1] 0
```

```
effect_size<-(mean_X-mean_Y)/s_xy #compute the effect_size
effect_size
```

```
## [1] -2.104197
```

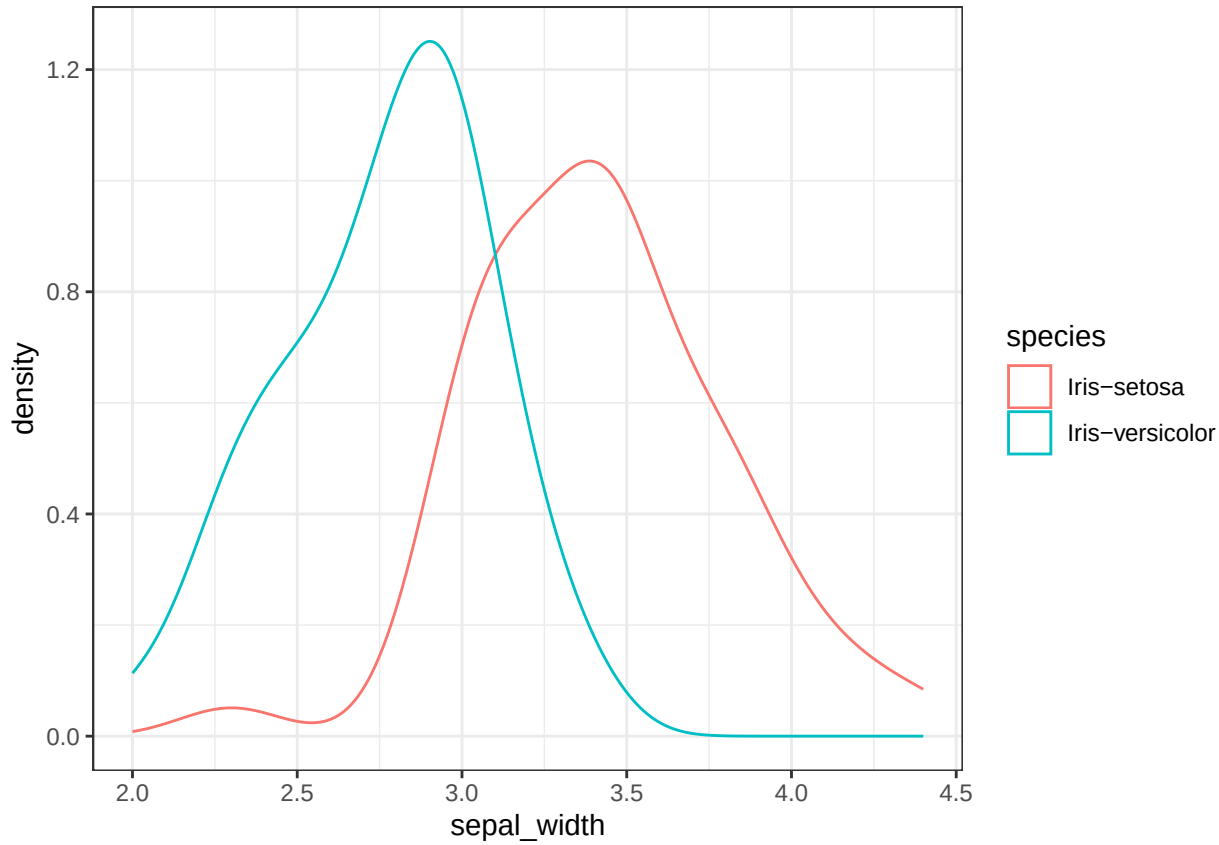
```
t.test(sepal_length~species,data=sepal_length_data,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sepal_length by species
## t = -10.521, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Iris-setosa and group Iris-versicolor
## 95 percent confidence interval:
##  -1.1054165 -0.7545835
## sample estimates:
##     mean in group Iris-setosa mean in group Iris-versicolor
##                         5.006                         5.936
```

The results of this study show lindicate that when the sample size increases,the effect size also would increases.The statistical significance depends on the sample size.

Next step, the case focus on comparing the difference of sepal width between the two samples.

```
sepal_width_data<-iris_data%>%
  filter(species=="Iris-versicolor"|species=="Iris-setosa")%>%
  select(species,sepal_width)
# create the new data frame including both samples
ggplot()+
  geom_density(data=sepal_width_data,aes(x=sepal_width,color=species))+
  theme_bw()+xlab("sepal_width")+ylab("density")
```




```
# draw a plot of the density distribution of both samples.

mean_X<-sepal_width_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_width)%>%mean(na.rm=TRUE) #the mean of X
mean_Y<-sepal_width_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_width)%>%mean(na.rm=TRUE) #the mean of Y
sd_X<-sepal_width_data%>%filter(species=="Iris-setosa")%>%
  pull(sepal_width)%>%sd(na.rm=TRUE) # sd of X
sd_Y<-sepal_width_data%>%filter(species=="Iris-versicolor")%>%
  pull(sepal_width)%>%sd(na.rm=TRUE) #sd of Y
n<-length(sepal_width_data%>%filter(species=="Iris-setosa")%>%
            pull(sepal_width)%>%na.omit()) #n.omit: strip the null numbers
k<-sepal_width_data%>%filter(species=="Iris-versicolor")%>%nrow()

s_xy<-sqrt(((n-1)*sd_X^2+(k-1)*sd_Y^2)/(n+k-2))
t_statistic<-(mean_X-mean_Y)/(s_xy*sqrt(1/n+1/k))
# compute the test statistic
```

```
t.test(sepal_width~species,data=sepal_width_data,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sepal_width by species
## t = 9.2828, df = 98, p-value = 4.362e-15
```

```
## alternative hypothesis: true difference in means between group Iris-setosa and group Iris-versicolor
## 95 percent confidence interval:
##  0.5094708 0.7865292
## sample estimates:
##     mean in group Iris-setosa mean in group Iris-versicolor
##                         3.418                         2.770
```

```
effect_size<-(mean_X-mean_Y)/s_xy #compute the effect_size
effect_size
```

```
## [1] 1.856555
```

From the p-value,it indicates that the true difference in means between group Iris-setosa and group Iris-versicolor is not equal to 0. Furthermore, from these results it is clear that the effect size of sepal width is more than the effect size of sepal length,which is greater than 0.It can be inferred that the the sepal-width's research finding has a more practical significance than sepal-length's. The sepal width could be a key factor to distinguish the different types of Iris flowers.

# Conclusion

This study set out to determine the unpaired Student's t-test.The relevance of sepal width is clearly supported by the current findings and these findings enhance the understanding of the unpaired Student's t-test and the differences of different types of Iris flowers.It shows that there is a true difference between the different types Iris flowers for their sepal length and sepal width.At the same time,the width is better used to distinguish them than the length.However, the scope of this study is also limited in terms of the sample size of the data.