**Analysis of IRIS plant species**

BANA 7031

Probability Modelling Project

Pravallika Kalidindi (M12823119)

Anirudh Chekuri (M12909975)

MS Business Analytics

Department of Operations, Business Analytics & Information Systems

Carl H Lindner School of Business: University of Cincinnati

**Abstract:**

The “IRIS” dataset includes three iris species with 50 samples each as well as some properties about each flower. This project includes an analysis of the sepal width of two different species of the iris flower – Setosa, Versicolor . We analyzed the distribution of the sepal width of different species using an ECDF. To calculate the difference of sepal widths of Setosa and Versicolor we used a maximum likelihood estimator and Bootstrap methods and to test if the difference is statistically meaningful, we used Hypothesis testing and Wald test. Finally to validate our result we used a Bayesian approach to calculate the distribution of difference in sepal widths of the two species.

1. **Introduction :**
2. **Data source**

The data set has been taken from the Ecdat package in R.

The columns in this dataset are:

* SepalLength- Length of the sepal in centimeters
* SepalWidth- Width of the sepal in centimeters
* PetalLength-Length of the petal in centimeters
* PetalWidth- Width of the petal in centimeters
* Species – The species to which the flower belongs to. (Setosa, Virginica and Versicolor)

In this project we would be inferring, using statistical inference, about the means of populations of sepal widths of two different species, Serosa and Versicolor.

1. **Summary of the data**

|  |  |  |
| --- | --- | --- |
| Species | Count | Mean |
| Setosa | 50 | 3.428 |
| Versicolor | 50 | 2.770 |

Table 1.A Summary statistics of the sepal width

1. **Distribution**

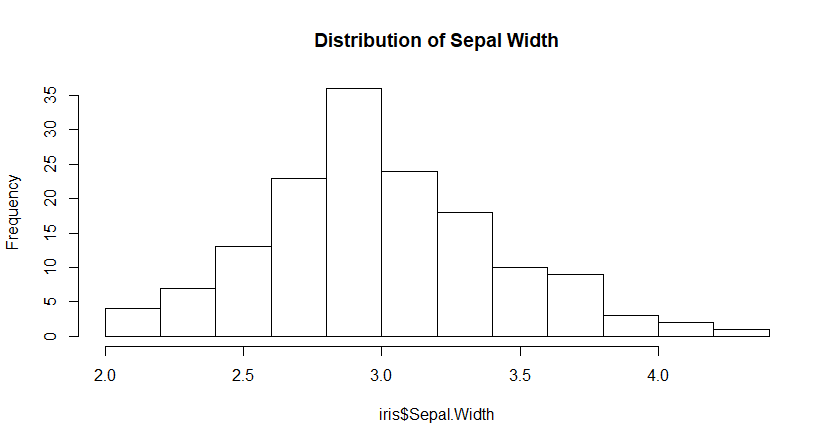


Fig 1.A Distribution of Sepal width

A histogram of sepal width column shows that it follows an approximate normal distribution

1. **Data Analysis**

1. **Empirical Cumulative density function (ECDF)**

To understand the distribution of the sepal width column, although a histogram seems to be more intuitive at the first look and needs less explanation, sometimes information might be lost in binning and because of small sample size. In order to avoid this we will be using the CDF which is a better in our project.

For the purpose of our analysis, we use ECDF as an estimator for the CDF

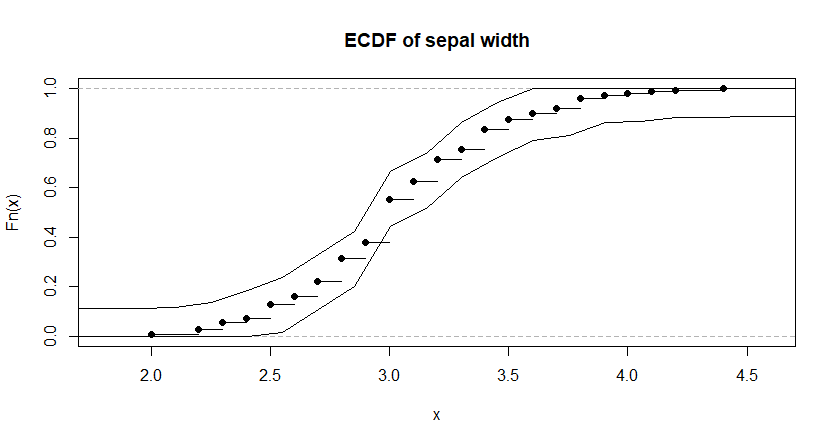


Fig 2.A ECDF of sepal width column

Empirical cumulative distribution function has been plotted for sepal width variable along with a 95% confidence band.

Using the ECDF function in R, it was found that an estimate 45% of the sample flowers had sepal —width higher than the mean sepal width (3.45). Since it is only an estimate, we also calculated a 95% confidence interval for this value to be more accurate with our results. Thus, the true population proportion which can have sepal width higher than 3.42 is between 40% and 52% with 95% confidence.

1. **Maximum likelihood estimator**

To estimate the difference in sepal width of Setosa and Versicolor , we will be using a maximum likelihood estimator.

Parameter of interest = µs - µV, where

µs = Mean sepal width of setosa flowers.

µV= Mean sepal width of versicolor flowers.

MLE (µs - µV) = MLE (µs)-MLE (µV).

MLE (µs - µV) = Sample mean of setosa -Sample mean of versicolor = 0.658

1. **Bootstrap and confidence intervals**

Since MLE is just an estimate , we need to find 95 % confidence interval forµs - µV

We will use bootstrap methods (parametric and non-parametric) to find the standard error to calculate the confidence interval.

Given below are the results -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bootstrap Method | Mean | Standard error | 2.5th Percentile | 95th Percentile |
| Parametric | 0.658 | 0.0701 | 0.5204 | 0.8011 |
| Non-Parametric | 0.657 | 0.0706 | 0.5184 | 0.8009 |

Table 2.1 Summary statistics of Bootstap methods

Below are the distributions using Parametric and Non-Parametric methods

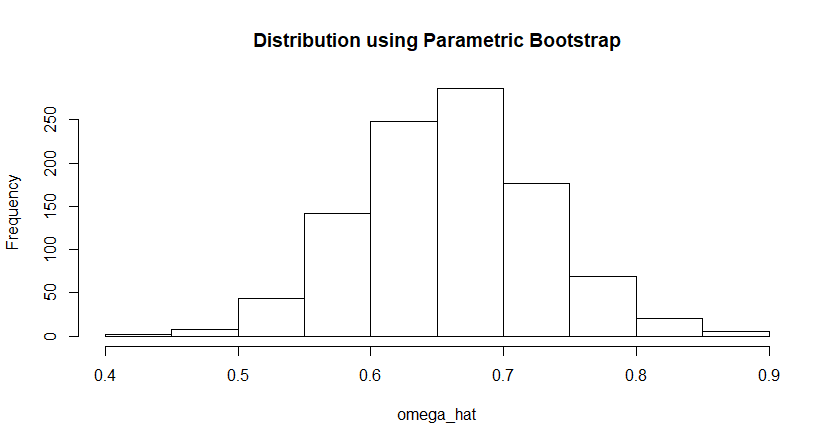


Fig 2.B Distribution of  using parametric bootstrap

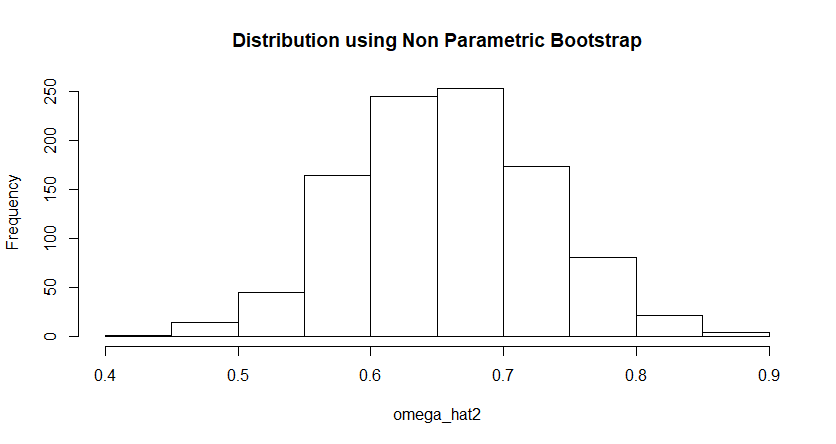


Fig 2.C Distribution of  using non parametric bootstrap

1. **Hypothesis Test**

Given below is the formulation of hypothesis to check if the Setosa and Versicolor have the same sepal widths.

H0 : µs - µV =0 ( Mean sepal width of Setosa equals Mean sepal width of Versicolor)

Ha : µs - µV ≠ 0(Mean sepal width of Setosa not equals Mean sepal width of Versicolor)

Wald test statistic =  = 0.658/0.07 = 9.4

P-Value = 0

At α = 0.05, since the p-value is less than α, we reject the null hypothesis

Conclusion – Mean sepal width of Setosa is significantly different from mean sepal width of Versicolor.

1. **Bayesian Analysis**

The above analysis of the distribution of µs - µV is based on a frequentist approach. We will now find the distribution of µs - µV  using a Bayesian approach.

Assumptions –

* Prior (µs) ~ N(0,1)
* Prior (µV) ~ N(0,1)

Based on the above assumptions, the posterior distribution of µs - µV is -

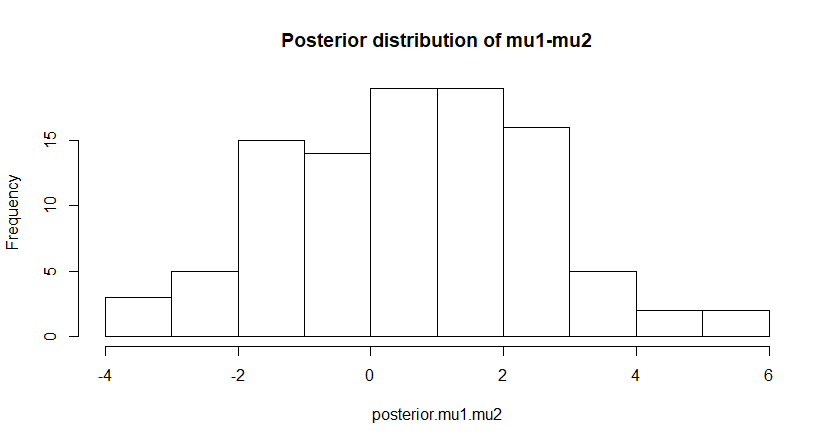


Fig 2.D Posterior distribution of

The mean difference in sepal widths from the above graph is 0.61. which is not very different from the value obtained using a frequentist approach.

**Appendix**

**R Code**

data<-iris  
hist(iris$Sepal.Width,main="Distribution of Sepal Width")

#ecdf  
x<-ecdf(iris$Sepal.Width)  
plot(x,main="ECDF of sepal width")  
mean(iris$Sepal.Width)

Alpha=0.05  
n=length(iris$Sepal.Width)  
Eps=sqrt(log(2/Alpha)/(2\*n))  
grid<-seq(0,150, length.out = 1000)  
lines(grid, pmin(x(grid)+Eps,1))  
lines(grid, pmax(x(grid)-Eps,0))

#MLE Estimate  
#Tau = mu1-mu2 (1-setosa) , (2-versicolor)  
mle\_est = mean(iris[iris$Species == 'setosa',]$Sepal.Width) - mean(iris[iris$Species == 'versicolor',]$Sepal.Width)  
  
#parametric bootstrap  
sample1<- iris[iris$Species == 'setosa',]  
sample2<-iris[iris$Species == 'versicolor',]  
mu1.hat = mean(sample1$Sepal.Width)  
mu2.hat = mean(sample2$Sepal.Width)  
n1=length(sample1$Sepal.Width)  
n2=length(sample2$Sepal.Width)  
sd1.hat = sd(sample1$Sepal.Width)  
sd2.hat = sd(sample2$Sepal.Width)  
   
omega\_hat<-c()  
for (i in 1:1000){  
 x1<-rnorm(n1,mu1.hat,sd1.hat)  
 x2<-rnorm(n2,mu2.hat,sd2.hat)  
 omega\_hat[i]=mean(x1)-mean(x2)  
}  
parametric\_se <- sd(omega\_hat)  
parametric\_CI <- c(mean(omega\_hat)-2\*sd(omega\_hat),mean(omega\_hat)+2\*sd(omega\_hat))  
mean(omega\_hat)

hist(omega\_hat,main = "Distribution using Parametric Bootstrap")

#Non parametric bootstrap  
omega\_hat2<-c()  
for (i in 1:1000){  
 x1<-sample(sample1$Sepal.Width,n1,replace=T)  
 x2<-sample(sample2$Sepal.Width,n2,replace=T)  
 omega\_hat2[i]=mean(x1)-mean(x2)  
}  
  
non\_parametric\_se <- sd(omega\_hat2)  
non\_parametric\_CI <- c(mean(omega\_hat2)-2\*sd(omega\_hat2),mean(omega\_hat2)+2\*sd(omega\_hat2))  
mean(omega\_hat2)

hist(omega\_hat2,main = "Distribution using Non Parametric Bootstrap")

#Hypothesis Testing  
# Null hypothesis = mu1=mu2 , Alternate hypothesis = mu1 not equal to mu2  
## WALD   
sigma.hat<-sqrt((sd(sample1$Sepal.Width)^2/n1)+(sd(sample2$Sepal.Width)^2/n2))  
z.stat<-(mle\_est-0)/(sigma.hat)  
p.value=2\*(1-pnorm(abs(z.stat)))  
#p value is less than 0 and hence fail to reject null hypothosis  
  
#Bayesian analysis  
 #prior of mu1 ~ N(0,1), prior of mu2 ~ N(0,1)  
#posterior of mu1 ~ N(n1\*xbar/1+n1,1/1+n1), posterior of mu2 ~ N(n2\*xbar/1+n2,1/1+n2)  
  
#posterior of mu1  
Ib.1=1  
Ix.1 = n1/var(sample1$Sepal.Width)  
pos1.mean=((mean(sample1$Sepal.Width))\*Ix.1)/(Ib.1+Ix.1)  
pos1.var = 1/(Ib.1+Ix.1)  
posterior.mu1 = rnorm(100,pos1.mean,sqrt(pos1.var))  
  
#posterior of mu2  
Ib.2=1  
Ix.2 = n2/var(sample2$Sepal.Width)  
pos2.mean=((mean(sample2$Sepal.Width))\*Ix.2)/(Ib.2+Ix.2)  
pos2.var = 1/(Ib.2+Ix.2)  
posterior.mu2 = rnorm(100,pos2.mean,sqrt(pos2.var))  
  
#posterior of mu1-mu2  
pos3.mean = pos1.mean-pos2.mean  
pos3.var = posterior.mu1+pos2.var  
posterior.mu1.mu2 = rnorm(100,pos3.mean,sqrt(pos3.var))  
hist(posterior.mu1.mu2,main="Posterior distribution of mu1-mu2")

mean(posterior.mu1.mu2)