R handout: Likelihood Ratio Test

*Yuhui Yao**

This handout will show various ways to conduct the Likelihood ratio test in practice. Suppose we are interested in the following hypothesis testing for the mean of the variable, sepal width, of Virginica (from the Iris data set (Fisher (1936))) $H_0 : \mu = \mu_0 = 3$ vs. $H_1 : \mu = \mu_1 \neq 3$ under the assumption that the observations X_i's are following i.i.d. the normal distribution with mean μ and known variance $\sigma^2 = 0.1040$. Also, let the significance level be $\alpha = 0.05$. First of all, we load the data and describe the variable, sepal width, of Virginica.

- R code

```
# set a seed
set.seed(12345)

# specify the address of the data
data.addr <- 'https://raw.githubusercontent.com/
    bolus123/R-handout/master/LikelihoodRatioTest/iris.csv'

# load the data
data <- read.csv(file = data.addr)

# name the columns
names(data) <- c(
        'sl' #sepal lengt
        ,'sw' #sepal width
        ,'pl' #petal length
        ,'pw' #petal width
        ,'class' #class
)

# get the fraction of Virginica
data.virginica <- data[data$class == 'Iris-virginica', ]

# describe the Virginica data
mean(data.virginica$sw)
```

```
## [1] 2.974
```

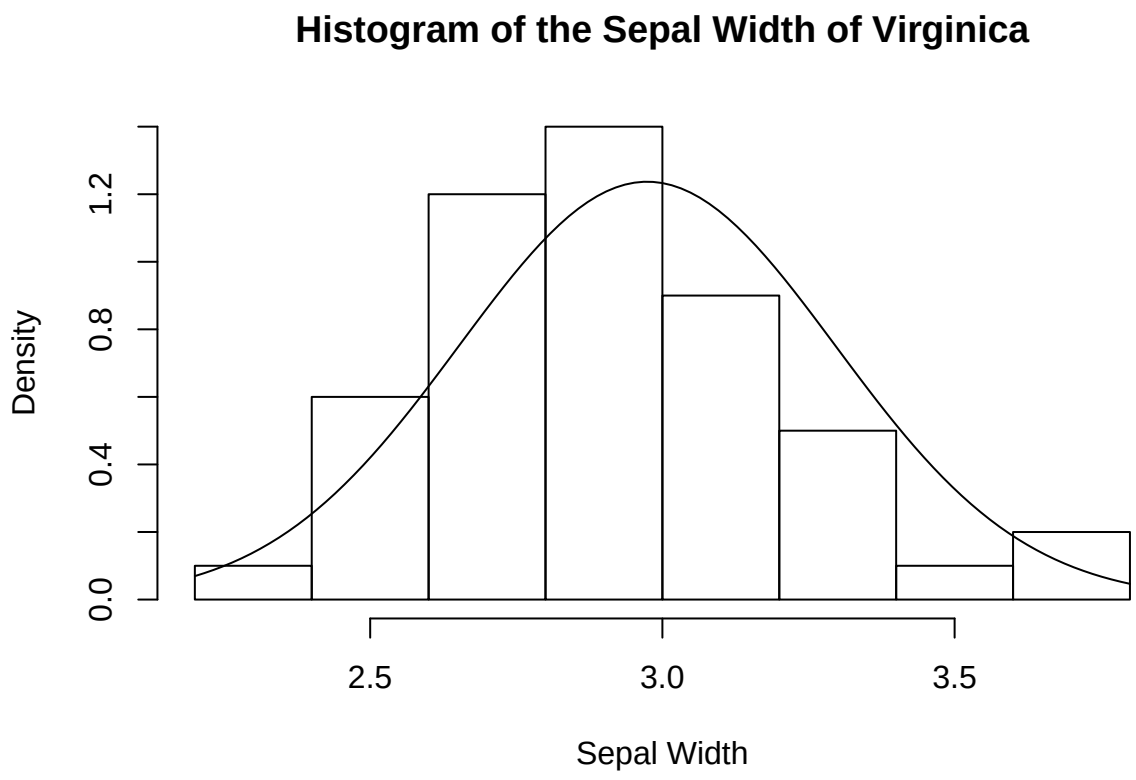
```
var(data.virginica$sw)
```

```
## [1] 0.1040041
```

```
# build a histogram
hist(data.virginica$sw, freq = FALSE, xlab = 'Sepal Width',
  main = 'Histogram of the Sepal Width of Virginica')
```

*Yuhui Yao is a PhD student in the department of ISM in The University of Alabama and his email is yyao17@crimson.ua.edu

```r
# add a normal distribution
curve(dnorm(x, mean(data.virginica$sw), sd(data.virginica$sw)), add = TRUE)
```




```r
# It looks good!
```

1. Exact Method for the Normal distribution with mean μ and known variance σ^2

The likelihood

$$L(\mu; \underline{x}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

The log-likelihood

$$lnL(\mu; \underline{x}) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{1}$$

It is easy to show the m.l.e. be $\hat{\mu} = \bar{x}$. The log-likelihood ratio

$$\begin{aligned} \Lambda &= ln(\frac{L(\mu_0; \underline{x})}{L(\hat{\mu}; \underline{x})}) = lnL(\mu_0; \underline{x}) - lnL(\hat{\mu}; \underline{x}) \\ &= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2 + (\frac{n}{2} ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2) \end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu_0)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu_0)^2 \\ &= \sum_{i=1}^{n}[(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu_0)+(\bar{x}-\mu_0)^2] \\ &= \sum_{i=1}^{n}[(x_i-\bar{x})^2]+n(\bar{x}-\mu_0)^2\end{aligned}$$

So

$$\Lambda = -\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2$$

Because $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \Rightarrow (\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}})^2 \sim \chi_1^2$, according to the equal-tail assumption, the size

$$P(-2\Lambda \geq C|H_0) = P(-2\Lambda \geq C|\mu = \mu_0) = \alpha$$

where under the null hypothesis $-2\Lambda \sim \chi_1^2$.

Under the assumption of normality, -2Λ is exactly following the chi-squared distribution with some degrees of freedom, but, in practice, the normality assumption is not always true. Fortunately, Wilks (1938) proved that, if the sample size is large enough and H_0 is true, for a nested model, -2Λ is asymptotically following the chi-squared distribution with the degrees of freedom approximately equal to the difference between the degrees of freedom under the alternative hypothesis and the degrees of freedom under the null hypothesis. This result is called **Wilks' Theorem**.

- R code

```
# sample size of Virginica
n <- dim(data.virginica)[1]

# sigma2 is known
sigma2 <- 0.1040
sigma <- sqrt(sigma2)

# under the null hypothesis
mu.0 <- 3

# under the alternative hypothesis
mu.1 <- mean(data.virginica$sw)

# calculate the statistic
Lambda <- - n / 2 / sigma2 * (mu.1 - mu.0)^2
Lambda <- -2 * Lambda
Lambda
```

```
## [1] 0.325
```

```
# critical values for alpha = 0.05 under the equal-tailed assumption
C <- qchisq(0.95, 1)
C
```

```
## [1] 3.841459
```

```r
# calculate p-value
p <- pchisq(Lambda, 1)
p.value <- 1 - p
p.value
```

```
## [1] 0.5686182
```

2. Bootstrap Method

Suppose both of the analytical form and the distribution of -2Λ are too complicated and we do not have enough sample size to have the asympotical approximation, but we can do our test by simulating the empirical distribution of the log-likelihood ratio -2Λ.

(a) Parametric Bootstrap (Assuming the distribution is normal)

We will **simulate the data from a specific distribution (normal) under the null hypothesis** and then do the test repeatedly.

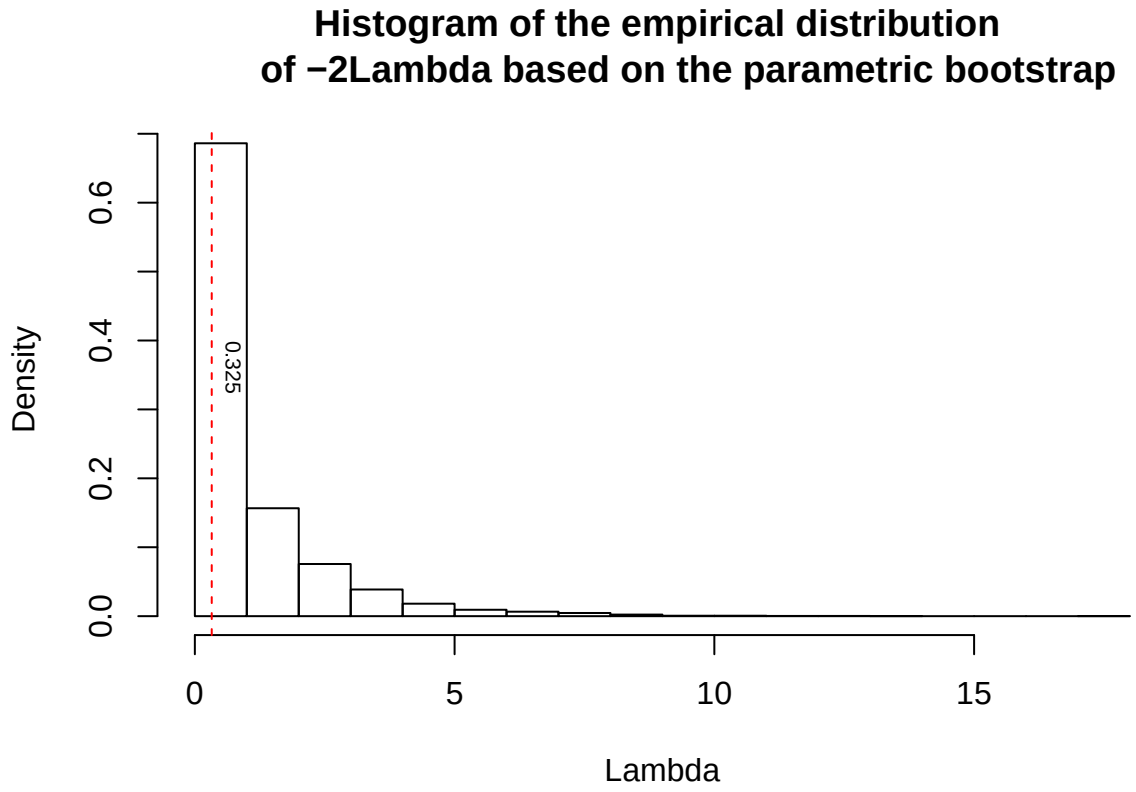
Steps for our problem:

i. **Simulate a sample from the given distribution under the null hypothesis**
ii. Calculate the likelihood under the null hypothesis
iii. Calculate the likelihood under the alternative hypothesis (Notice that the m.l.e. need to be re-estimated)
iv. Calculate and save the value of -2Λ denoted $-2\Lambda_s$ for simulation for the empirical distribution of $-2\Lambda_s$
v. repeat (i) - (iv) until the process reaches the maximum of simulations
vi. Compare the -2Λ calculated from the original data with the empirical distribution of $-2\Lambda_s$

- R code

```r
# maximum number of simulations
sim <- 10000

# sample size of Virginica
n <- dim(data.virginica)[1]

# get the empirical distribution based on the parametric bootstrap
ref.par <- rep(NA, sim)

for (i in 1:sim){

    # simulate data under the null hypothesis
    X <- rnorm(n, mu.0, sigma)
  # mu.0 = 3 and sigma2 = 0.1040

    # calculate Lambda
    lnL.0 <- sum(log(dnorm(X, mu.0, sigma)))
```

```
    lnL.1 <- sum(log(dnorm(X, mean(X), sigma)))
    ref.par[i] <- -2 * (lnL.0 - lnL.1)

}

# describe the empirical distribution of Lambda based on the parametric bootstrap
hist(ref.par, freq = FALSE, main = 'Histogram of the empirical distribution
     of -2Lambda based on the parametric bootstrap', xlab = 'Lambda')
# point out the Lambda from the original data
abline(v = Lambda, lty = 2, col = 'red')
text(Lambda, 0.4, as.character(round(Lambda, 4)), srt = 270, pos = 4, cex = 0.7)
```




```
# critical values for alpha = 0.05 under the equal-tailed assumption
C <- quantile(ref.par, 0.95)
C
```

```
##      95%
## 3.730281
```

```
# calculate p-value
p <- mean(ref.par < Lambda)
p.value.par <- 1 - p
p.value.par
```

```
## [1] 0.5626
```

(b) Nonparametric Bootstrap

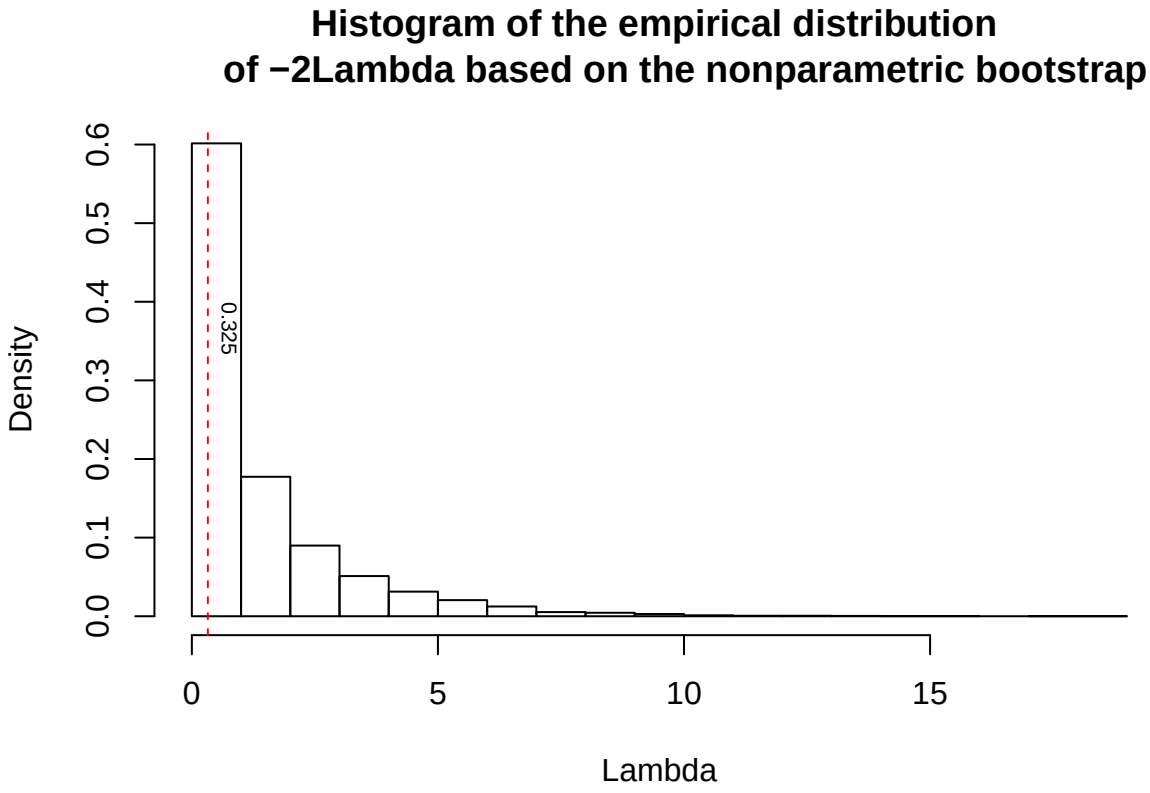
We will **resample the data from the original data set** and then do the test repeatedly.

Steps for our problem:

i. **Resample a sample from the original data**

ii. the rest of steps are exactly same as the ones ii. - vi. for the parametric bootstrap shown above.

```r
# maximum number of simulations
sim <- 10000

# sample size of Virginica
n <- dim(data.virginica)[1]

# get the empirical distribution based on the nonparametric bootstrap
ref.nonpar <- rep(NA, sim)

for (i in 1:sim){

    # resample from the original data set
    X <- sample(data.virginica$sw, size = n, replace = TRUE)

    # calculate Lambda
    lnL.0 <- sum(log(dnorm(X, mu.0, sigma)))
    lnL.1 <- sum(log(dnorm(X, mean(X), sigma)))
    ref.nonpar[i] <- -2 * (lnL.0 - lnL.1)

}

# describe the empirical distribution of Lambda based on the nonparametric bootstrap
hist(ref.nonpar, freq = FALSE, main = 'Histogram of the empirical distribution
     of -2Lambda based on the nonparametric bootstrap', xlab = 'Lambda')
# point out the Lambda from the original data
abline(v = Lambda, lty = 2, col = 'red')
text(Lambda, 0.4, as.character(round(Lambda, 4)), srt = 270, pos = 4, cex = 0.7)
```




```
# critical values for alpha = 0.05 under the equal-tailed assumption
C <- quantile(ref.nonpar, 0.95)
C
```

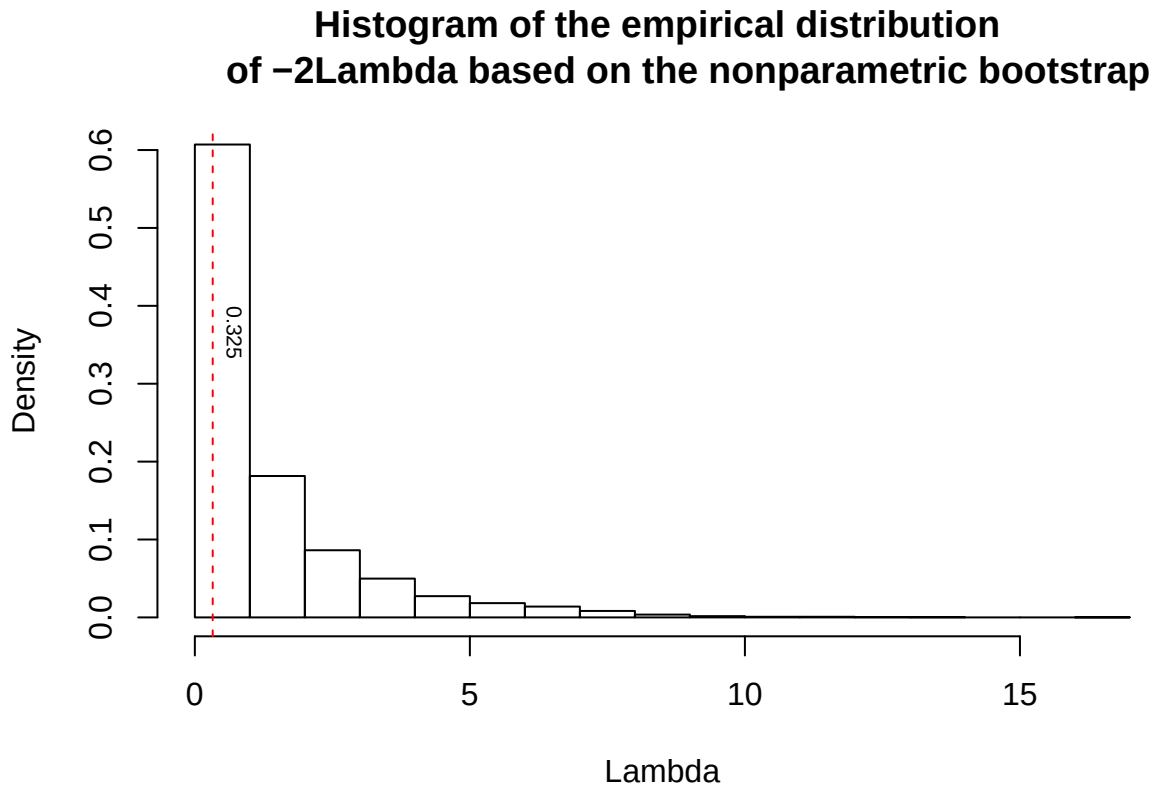
```
##      95%
## 4.807692
```

```
# calculate p-value
p <- mean(ref.nonpar < Lambda)
p.value.nonpar <-  1 - p
p.value.nonpar
```

```
## [1] 0.6401
```

```
# maximum number of simulations
sim <- 10000

# sample size of Virginica
n <- dim(data.virginica)[1]

# get the empirical distribution based on the nonparametric bootstrap
ref.nonpar <- rep(NA, sim)

for (i in 1:sim){
```

```
    # resample from the original data set
    X <- sample(data.virginica$sw, size = n, replace = TRUE)

    # calculate Lambda
    lnL.0 <- sum(log(dnorm(X, mu.0, sigma)))
    lnL.1 <- sum(log(dnorm(X, mean(X), sigma)))
    ref.nonpar[i] <- -2 * (lnL.0 - lnL.1)

}

# describe the empirical distribution of Lambda based on the nonparametric bootstrap
hist(ref.nonpar, freq = FALSE, main = 'Histogram of the empirical distribution
     of -2Lambda based on the nonparametric bootstrap', xlab = 'Lambda')
# point out the Lambda from the original data
abline(v = Lambda, lty = 2, col = 'red')
text(Lambda, 0.4, as.character(round(Lambda, 4)), srt = 270, pos = 4, cex = 0.7)
```




```
# critical values for alpha = 0.05 under the equal-tailed assumption
C <- quantile(ref.nonpar, 0.95)
C
```

```
##      95%
## 4.807692
```

```r
# calculate p-value
p <- mean(ref.nonpar < Lambda)
p.value.nonpar <-  1 - p
p.value.nonpar
```

```
## [1] 0.6352
```

3. Comparison

The following is the table for the p-values from different methods

```
##         ExactMethod ParametricBootstrap NonparametricBootstrap
## p-value      0.5686              0.5626                 0.6352
```

The p-values from the exact method are similar with the the one from the parametric bootstrap method which is less than the one from the nonparametric bootstrap method. All of them show that the null hypothesis $H_0 : \mu = \mu_0 = 3$ fails to be rejected. So why did these three methods give us different p-values? Which one would you like to use?

4. Practice (Goodness-of-fit)

Suppose we are interested in the following hypothesis testing for deciding whether the sample of sepal width of Virginica is from the specific normal distribution with mean $\mu = 3$ and variance $\sigma^2 = 0.1$. Following the three methods of conducting the likelihood ratio test based on the normality assumption, what is your setting for the test and what is your conclusion? (**Hint**: the degrees of freedom for this test is not equal to 1. Also, you need to estimate the m.l.e. of sample mean and sample variance, simultaneously).

Reference

Fisher, Ronald A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Human Genetics* 7 (2). Wiley Online Library:179–88.

Wilks, Samuel S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics* 9 (1). JSTOR:60–62.