Statistical Inference Course Project

Andrew Szwec

23 August 2014

1. Simulation Exercise

Do 1000 simulations of 40 x Exponetials

Run and plot the simulation and explain the properties of the distribution 1. Show where the distribution is centered at and compare it to the theoretical center of the distribution. 2. Show how variable it is and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal. 4. Evaluate the coverage of the confidence interval for 1/lambda: X¯±1.96Sn. (This only needs to be done for the specific value of lambda).

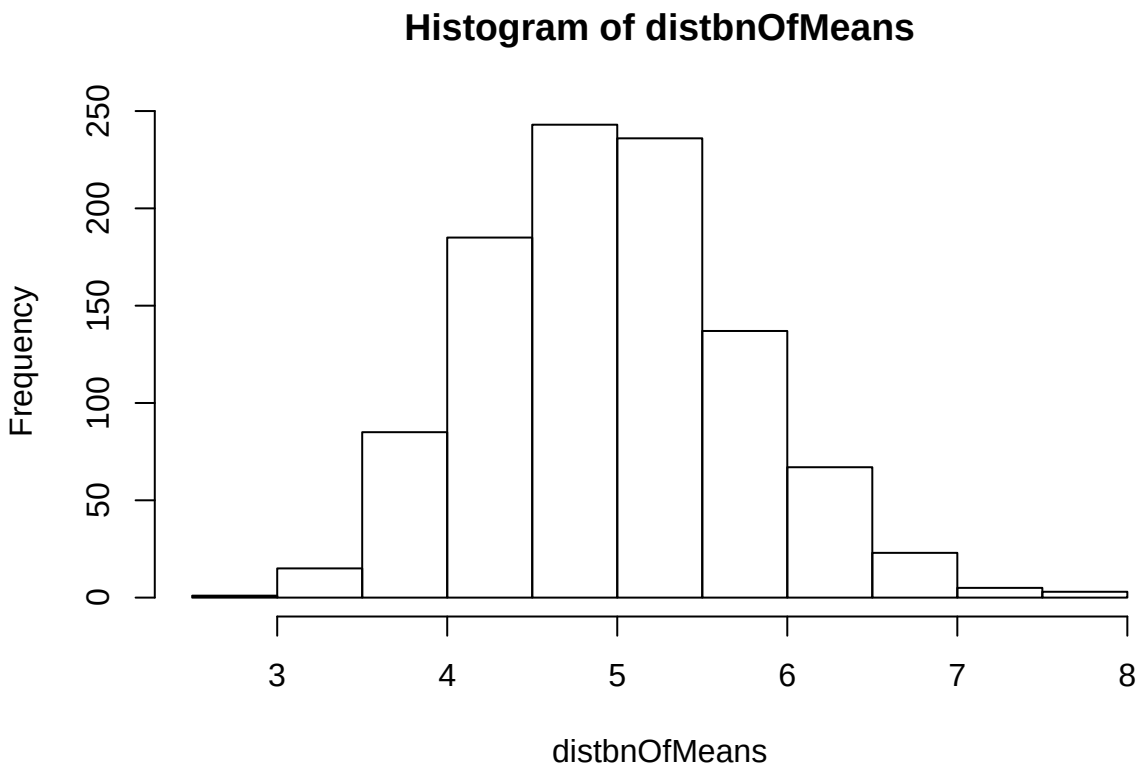
```
# Exponential Distribution
set.seed(1234)
lambda = 0.2
actualMean = 1/lambda
actualstdDev = 1/lambda
actualVariance = actualstdDev^2
n = 40
nosim  <- 1000

# -- This code below creates a 1000 x 40 matrix with 40,000 exponentially
# -- distributed random variables. Then it takes the mean of
# -- each row of 1000 rand numbers and makes a tall vector.

mydata <- matrix(rexp(nosim * n, lambda) , nosim)
distbnOfMeans <- apply(mydata, 1, mean)
# Distribution is centred at
meanOfSamples <- mean(distbnOfMeans)
# Standard Deviation of Means
stdDevOfSampleMean <- sd(distbnOfMeans)

# Optional
stdDevOfSampleStd <- sd(apply(mydata, 1, sd))

# Mean Standard Deviation of Sample
meanOfSampleStd  <- mean(apply(mydata, 1, sd))
varianceOfSample <- meanOfSampleStd^2

rbind(
      c(1, 'Theoretical Centre of distribution', actualMean)
      ,c(1, 'Centre of Simulation Distribution' , meanOfSamples)
      ,c(2, 'Theoretical Variance of Distribution', actualVariance )
      ,c(2, 'Vairance of Simulation Distribution', varianceOfSample)
)
```

```
##      [,1] [,2]                                [,3]
## [1,] "1"  "Theoretical Centre of distribution"   "5"
```

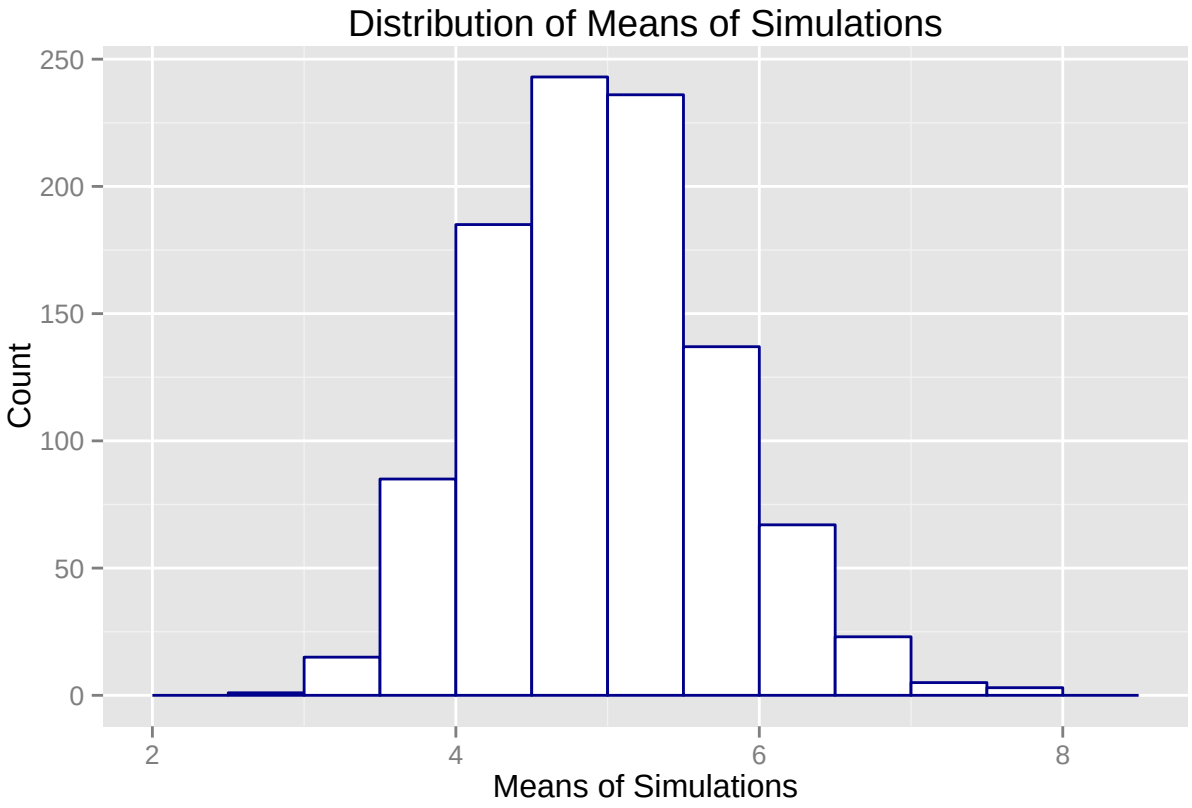
```
## [2,] "1"  "Centre of Simulation Distribution"    "4.97423877125153"
## [3,] "2"  "Theoretical Variance of Distribution" "25"
## [4,] "2"  "Vairance of Simulation Distribution"  "23.3726136368744"
```

```
# See image below: This shows the distribution of means of the 1000 sample
# Simulations is approximately normally distributed
hist(distbnOfMeans)

distbnOfMeansDf <- data.frame(distbnOfMeans)

require(ggplot2)
```

```
## Loading required package: ggplot2
```




```
m <- ggplot(distbnOfMeansDf, aes(x=distbnOfMeans) )
m + geom_histogram(colour = "darkblue", fill = "white", binwidth = 0.5) + ggtitle('Distribution of Means
```




```
# Confidence Interval
# Std Dev = 1/lambda, Mean = 1/lambda
# X_sample +/- 1.96 * S / sqrt(n)

CI <- meanOfSamples + c(-1, 1) * 1.96 * meanOfSampleStd/ sqrt(length(distbnOfMeans))
# CI = 4.702276 5.302208
pnorm(5.302208, mean = meanOfSamples, sd = meanOfSampleStd, lower.tail = TRUE) - pnorm(4.702276, mean =
```

```
## [1] 0.04947
```

```
# 0.04942181 ~ 5%
```

2. Basic Inferential Data Analysis

Load the ToothGrowth data and perform some basic exploratory data analyses Provide a basic summary of the data. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering) State your conclusions and the assumptions needed for your conclusions.

1. Load the Tooth Growth Dataset & Basic Exploratory Analysis

```
require(data.table)
```

```
## Loading required package: data.table
```

```
library(datasets); data(ToothGrowth);
dt <- data.table(ToothGrowth)
# Types of Supplement
unique(dt$supp)
```

```
## [1] VC OJ
## Levels: OJ VC
```

```
# Doses Used in Trial
unique(dt$dose)
```

```
## [1] 0.5 1.0 2.0
```

```
# Number of Each Supp
nrow(dt[ dt$supp == 'OJ' ])
```

```
## [1] 30
```

```
nrow(dt[ dt$supp == 'VC' ])
```

```
## [1] 30
```

```
# Number of Each Dose
nrow(dt[ dt$dose == 0.5 ])
```

```
## [1] 20
```

```
nrow(dt[ dt$dose == 1.0 ])
```

```
## [1] 20
```

```
nrow(dt[ dt$dose == 2.0 ])
```

```
## [1] 20
```

2. Basic Summary of Data

```
summary(ToothGrowth)
```

```
##       len       supp         dose
##  Min.   : 4.2   OJ:30   Min.   :0.50
##  1st Qu.:13.1   VC:30   1st Qu.:0.50
##  Median :19.2           Median :1.00
##  Mean   :18.8           Mean   :1.17
##  3rd Qu.:25.3           3rd Qu.:2.00
##  Max.   :33.9           Max.   :2.00
```

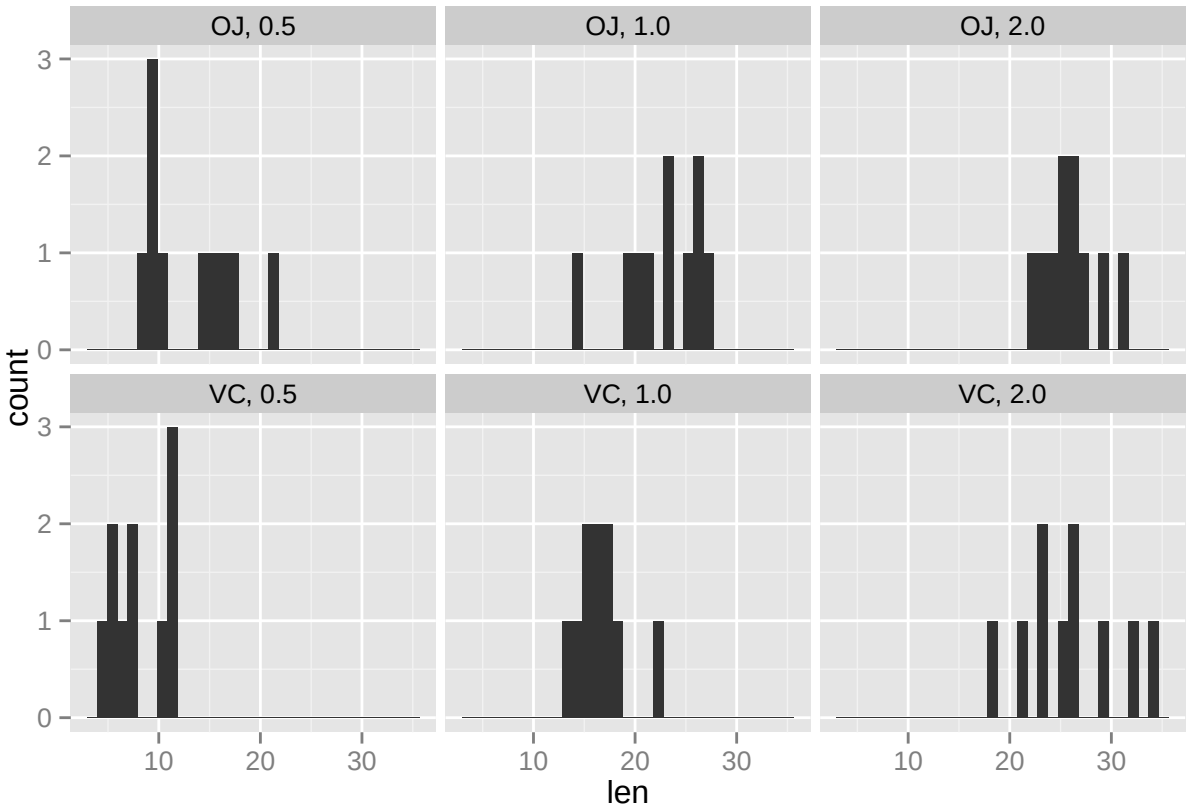
```
dt[, list(meanLength=mean(len), sdLength=sd(len)), by = c('supp', 'dose')]
```

```
##    supp dose meanLength sdLength
## 1:   VC  0.5       7.98    2.747
## 2:   VC  1.0      16.77    2.515
## 3:   VC  2.0      26.14    4.798
## 4:   OJ  0.5      13.23    4.460
## 5:   OJ  1.0      22.70    3.911
## 6:   OJ  2.0      26.06    2.655
```

```
#    supp dose meanLength sdLength
# 1:   VC  0.5       7.98 2.746634
# 2:   VC  1.0      16.77 2.515309
# 3:   VC  2.0      26.14 4.797731
# 4:   OJ  0.5      13.23 4.459709
# 5:   OJ  1.0      22.70 3.910953
# 6:   OJ  2.0      26.06 2.655058
```

```
ggplot(dt, aes(len) ) + geom_histogram() + facet_wrap(supp ~ dose)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



3. Use Confidence Intervals and Hypothesis Tests to compare tooth growth by supplement and dose

a) Confidence Intervals

Comparing by Supp

```
dt[, list(meanLength=mean(len), sdLength=sd(len)), by = supp]
```

```
##    supp meanLength sdLength
## 1:   VC      16.96    8.266
## 2:   OJ      20.66    6.606
```

```
#    supp meanLength sdLength
# 1:   VC   16.96333 8.266029
# 2:   OJ   20.66333 6.605561


OJ_mean = 20.66333
OJ_sd = 6.605561
OJ_n = 30

VC_mean = 16.96333
VC_sd = 8.266029
VC_n = 30
```

```
t_df  <- ( (VC_sd^2 / VC_n) + (OJ_sd^2 / OJ_n) )^2 / ( (VC_sd^2 / VC_n)^2 / (VC_n - 1) + (OJ_sd^2 / OJ_
```

```
OJ_mean - VC_mean + c(-1,1) * t_df * sqrt( (VC_sd^2 / VC_n) + (OJ_sd^2 / OJ_n)  )
```

```
## [1] -103.1  110.5
```

```
# 95% Confidence Interval is...
# -103.1492  110.5492
```

Comparing by Dose

```
dt[, list(meanLength=mean(len), sdLength=sd(len)), by = dose]
```

```
##    dose meanLength sdLength
## 1:  0.5      10.61    4.500
## 2:  1.0      19.73    4.415
## 3:  2.0      26.10    3.774
```

```
#    dose meanLength sdLength
# 1:  0.5     10.605 4.499763
# 2:  1.0     19.735 4.415436
# 3:  2.0     26.100 3.774150
```

b) Hypothesis Test

4. Conclusions and Assumptions

Conclusions

1. 2. 3. 4.

Assumptions

1. Use t interval as not sure if data is normally distributed
2. Assume unequal variances for t distribution confidence interval