Statistical Inference Course Project

Andrew Szwec

24 August 2014

1. Simulation Exercise

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution

```
rbind(
      c(1, 'Theoretical Centre of distribution', actualMean)
      ,c(1, 'Centre of Simulation Distribution' , meanOfSamples)
)
```

```
##      [,1] [,2]                                [,3]
## [1,] "1"  "Theoretical Centre of distribution" "5"
## [2,] "1"  "Centre of Simulation Distribution"  "4.97423877125153"
```

2. Show how variable it is and compare it to the theoretical variance of the distribution.

```
rbind(
      c(2, 'Theoretical Variance of Distribution', actualVariance )
      ,c(2, 'Vairance of Simulation Distribution', varianceOfSample)
)
```

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

3. Show that the distribution is approximately normal. 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)
```

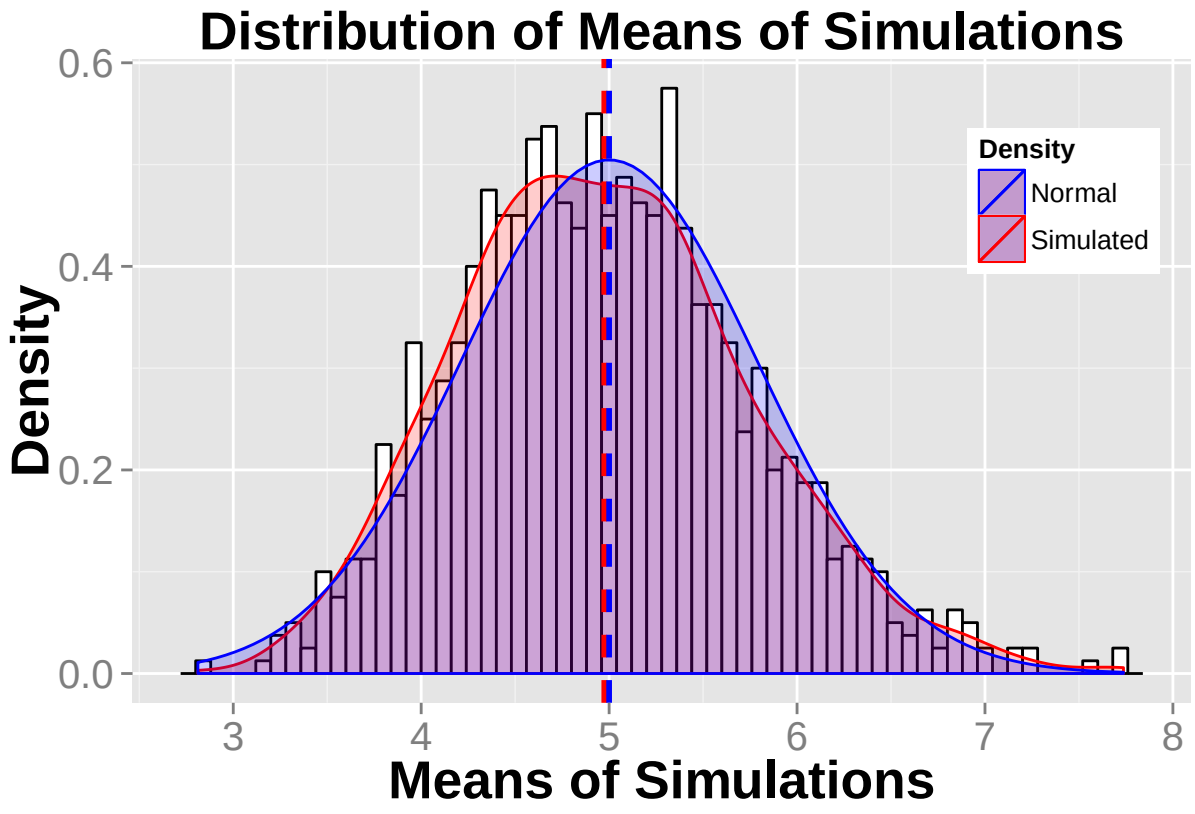
```
## Loading required package: ggplot2
```

```
# m <- ggplot(distbnOfMeansDf, aes(x=distbnOfMeans) )
# m + geom_histogram(colour = "darkblue", fill = "white", binwidth = 0.08) + ggtitle('Distribution of M


m <- ggplot(distbnOfMeansDf, aes(x = distbnOfMeans))
m <- m + geom_histogram(aes(y = ..density..), binwidth = .08 ,color="black", fill="white")
m <- m + geom_density(alpha=.2, fill="red", aes(colour="Simulated"))
m <- m + stat_function(fun = dnorm, arg = list(mean = 5, sd = 5/sqrt(40)), geom='area', alpha=.2, fill=
m <- m + scale_colour_manual("Density", values=c( "blue", "red"))
```

```
m <- m + scale_x_continuous(breaks = 2:9)
m <- m + scale_y_continuous(name = "Density")
m <- m + geom_vline(data=distbnOfMeansDf, aes(xintercept=meanOfSamples,  colour= "Simulated"),
                linetype="dashed", size=1)
m <- m + geom_vline(data=distbnOfMeansDf, aes(xintercept=5,  colour= "Normal"),
               linetype="dashed", size=1)
m <- m + theme(legend.position=c(.88, .78),
           legend.text = element_text(size = 10, colour = "black"),
           axis.text.x = element_text(size=15),
           axis.text.y = element_text(size=15),
           axis.title=element_text(size=20,face="bold"),
           plot.title = element_text(size=20, face="bold"))
m <- m + ggtitle('Distribution of Means of Simulations') + xlab('Means of Simulations')+ ylab('Density')
m
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



4. Evaluate the coverage of the confidence interval for 1/lambda: $\bar{X} \pm 1.96 S_n$. (This only needs to be done for the specific value of lambda). **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
```

```
## [1] 4.675 5.274
```

```
# 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%
```

```
## Tips
# Mean of each of 1000 simulations
# sd of each of 1000 simulations
```

2. Basic Inferential Data Analysis

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. 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)
4. 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

```
require(ggplot2)
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
```

Supplements and Doses Summary Table

```
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
```

Supplements Summary Table

```
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
```

Doses Summary Table

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

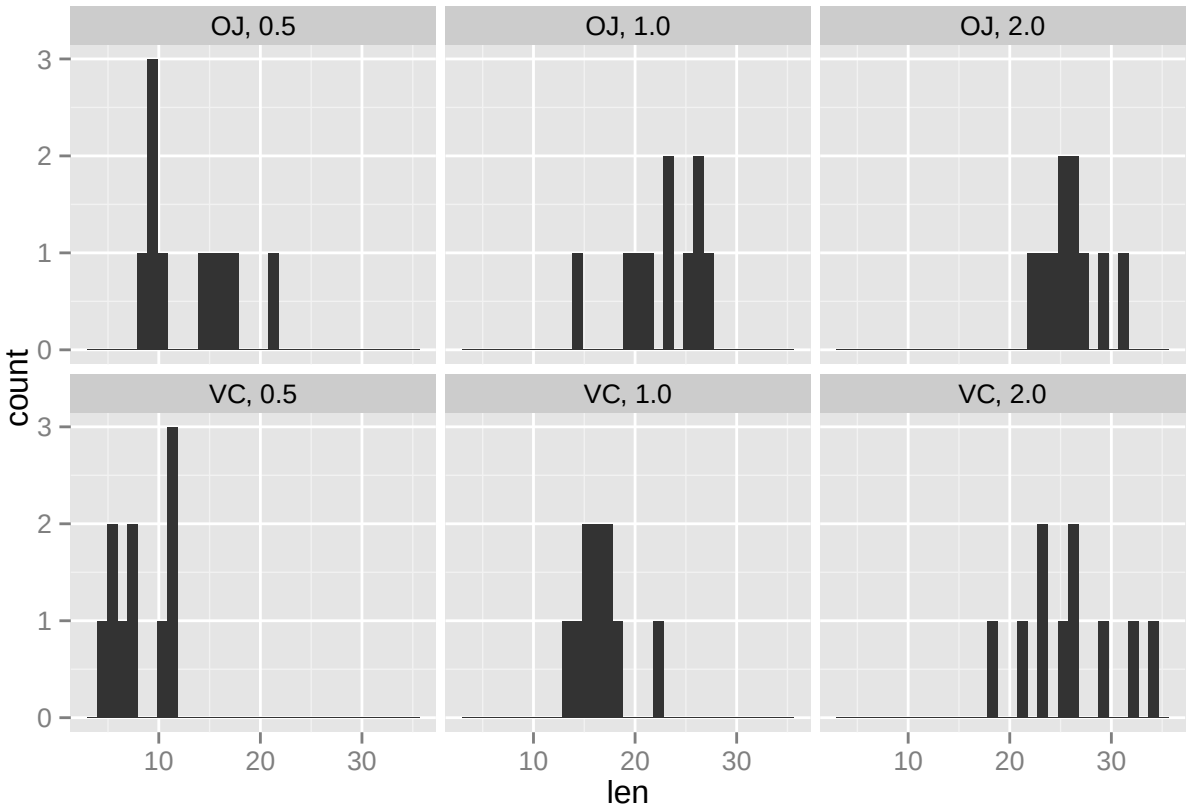
```
##    dose meanLength sdLength
## 1:  0.5      10.61    4.500
## 2:  1.0      19.73    4.415
## 3:  2.0      26.10    3.774
```

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

Graphical Summary of Supplements and Doses

```
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   16.96333 8.266029
# 2:   OJ   20.66333 6.605561

###############
# Individually
###############
n_VC = 30
X_VC = 16.96333
s_VC = 8.266029
degfree = n-1
```

What is 95% CI for the Mean length using VC?

```
# X_bar +/- tn-1 * s / sqrt(n)
round(X_VC + c(-1,1) * qt(0.975, degfree) * s_VC / sqrt(n_VC))
```

```
## [1] 14 20
```

```
# CI = 13, 20

n_OJ = 30
X_OJ = 20.66333
s_OJ = 6.605561
degfree = n-1
```

What is 95% CI for the Mean length using OJ?

```
# X_bar +/- tn-1 * s / sqrt(n)
round(X_VC + c(-1,1) * qt(0.975, degfree) * s_VC / sqrt(n_VC))
```

```
## [1] 14 20
```

```
# CI = 18, 23


################
# Paired
################
# Y
n_OJ = 30
X_OJ = 20.66333
s_OJ = 6.605561

# X
n_VC = 30
X_VC = 16.96333
s_VC = 8.266029

alpha = 0.05

degfree = n_OJ + n_VC - 2

Sp = sqrt( ( (n_VC - 1) * s_VC^2 + (n_OJ - 1) * s_OJ^2 ) /  degfree )
```

Independent t confidence interval for OJ - VC

```
# t for 95%
round(X_OJ - X_VC + c(-1,1) * qt(1 - alpha/2, degfree) * Sp * sqrt(1/n_VC + 1/n_OJ), 2)
```

```
## [1] -0.17  7.57
```

```
# CI = -0.17,  7.57
# Most of the time the OJ out-performs the VC with 95% confidence
```

Most of the time the OJ out performs the VC with 95% confidence

Comparing by Dose

```
# dt[, list(meanLength=mean(len), sdLength=sd(len)), by = dose]
#    dose meanLength sdLength
# 1:  0.5     10.605 4.499763
# 2:  1.0     19.735 4.415436
# 3:  2.0     26.100 3.774150

# X
n_05 = 20
X_05 = 10.605
s_05 = 4.499763

# Y
n_10 = 20
X_10 = 19.735
s_10 = 4.415436

# Z
n_20 = 20
X_20 = 26.100
s_20 = 3.774150

# 95% CI
alpha = 0.05

## Compare 0.5 vs 1.0
degfree = n_10 + n_05 - 2

Sp = sqrt( ( (n_05 - 1) * s_05^2 + (n_10 - 1) * s_10^2 ) /  degfree )
```

Independent t confidence interval comparing doses 0.5 vs 1.0 milligrams

```
# t for 95%
round(X_10 - X_05 + c(-1,1) * qt(1 - alpha/2, degfree) * Sp * sqrt(1/n_05 + 1/n_10), 2)
```

```
## [1]  6.28 11.98
```

```
# CI = 6.28 11.98
# Dose of 1.0 always out performs does of 0.5 in stimulating tooth growth with 95% confidence
```

Dose of 1.0 always out performs does of 0.5 in stimulating tooth growth with 95% confidence

```
## Compare 0.5 vs 2.0
degfree = n_20 + n_05 - 2

Sp = sqrt( ( (n_05 - 1) * s_05^2 + (n_20 - 1) * s_20^2 ) /  degfree )
```

Independent t confidence interval comparing doses 0.5 vs 2.0 milligrams

```
# t for 95%
round(X_20 - X_05 + c(-1,1) * qt(1 - alpha/2, degfree) * Sp * sqrt(1/n_05 + 1/n_20), 2)
```

```
## [1] 12.84 18.15
```

```
# CI = 12.84 18.15
# Dose of 2.0 always out performs does of 0.5 in stimulating tooth growth with 95% confidence
```

Dose of 2.0 always out performs does of 0.5 in stimulating tooth growth with 95% confidence

```
## Compare 1.0 vs 2.0
degfree = n_20 + n_10 - 2

Sp = sqrt( ( (n_10 - 1) * s_10^2 + (n_20 - 1) * s_20^2 ) /  degfree )
```

Independent t confidence interval comparing doses 2.0 vs 1.0 milligrams

```
# t for 95%
round(X_20 - X_10 + c(-1,1) * qt(1 - alpha/2, degfree) * Sp * sqrt(1/n_10 + 1/n_20), 2)
```

```
## [1] 3.74 8.99
```

```
# CI = 3.74 8.99
# Dose of 2.0 always out performs does of 1.0 in stimulating tooth growth with 95% confidence
```

Dose of 2.0 always out performs does of 1.0 in stimulating tooth growth with 95% confidence

b) Hypothesis Test

Comparing by Supp

```
## H0 = VC is a better supplement that OJ
## Ha = OJ is better
## If p-value > 0.05 then reject H0. Therefore OJ is a better supplement than VC


t.test(dt[supp=='OJ']$len , dt[supp=='VC']$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dt[supp == "OJ"]$len and dt[supp == "VC"]$len
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.171  7.571
## sample estimates:
## mean of x mean of y
##     20.66     16.96
```

```
# Alternative notation
# t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data=dt)

# Welch Two Sample t-test
#
# data:  dt[supp == "OJ"]$len and dt[supp == "VC"]$len
# t = 1.9153, df = 55.309, p-value = 0.06063
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -0.1710156  7.5710156
# sample estimates:
# mean of x mean of y
#  20.66333  16.96333

dt[, list(mean=mean(len), sd=sd(len))]
```

```
##     mean    sd
## 1: 18.81 7.649
```

```
X = 16.9333
mu = 18.81333
s = 7.649315
n = nrow(dt)
TS = (X - mu) / (s / sqrt(n))

#let
alpha = 0.05
Z_alpha = qnorm(alpha)
Z_alpha
```

```
## [1] -1.645
```

```
# Reject Null Hypothesis when
if ( TS <= Z_alpha ) 'Reject H0' else 'Fail to Reject H0'
```

```
## [1] "Reject H0"
```

```
Z_1_a_2 = qnorm(1-alpha/2)
Z_1_a_2
```

```
## [1] 1.96
```

```
# Reject Null Hypothesis when
if ( abs(TS) >= Z_1_a_2 ) 'Reject H0' else 'Fail to Reject H0'
```

```
## [1] "Fail to Reject H0"
```

```r
Z_1_a = qnorm(1-alpha)
Z_1_a
```

```
## [1] 1.645
```

```r
# Reject Null Hypothesis when
if ( TS >= Z_1_a ) 'Reject H0' else 'Fail to Reject H0'
```

```
## [1] "Fail to Reject H0"
```

```r
# Question
# How do you know what the mean of the population is?
# E.g. in this project we only know the mean of the sample from the Tooth Growth data
```

Comparing by Dose

```r
## 0.5 vs 1.0
## H0 = dose 1.0
## Ha = dose 0.5
## if p-value < 0.05 then fail to reject H0. Therefore dose 1.0 is better than 0.5
t.test(dt[dose==0.5]$len , dt[dose==1.0]$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dt[dose == 0.5]$len and dt[dose == 1]$len
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean of x mean of y
##     10.61     19.73
```

```r
#       Welch Two Sample t-test
#
# data:  dt[dose == 0.5]$len and dt[dose == 1]$len
# t = -6.4766, df = 37.986, p-value = 1.268e-07
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -11.983781  -6.276219
# sample estimates:
# mean of x mean of y
#    10.605    19.735

## 0.5 vs 2.0
t.test(dt[dose==0.5]$len , dt[dose==2.0]$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  dt[dose == 0.5]$len and dt[dose == 2]$len
## t = -11.8, df = 36.88, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16 -12.83
## sample estimates:
## mean of x mean of y
##     10.61     26.10
```

```
#       Welch Two Sample t-test
#
# data:  dt[dose == 0.5]$len and dt[dose == 2]$len
# t = -11.799, df = 36.883, p-value = 4.398e-14
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -18.15617 -12.83383
# sample estimates:
# mean of x mean of y
#    10.605    26.100
```

```
## 1.5 vs 2.0
t.test(dt[dose==1.0]$len , dt[dose==2.0]$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dt[dose == 1]$len and dt[dose == 2]$len
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean of x mean of y
##     19.73     26.10
```

```
#       Welch Two Sample t-test
#
# data:  dt[dose == 1]$len and dt[dose == 2]$len
# t = -4.9005, df = 37.101, p-value = 1.906e-05
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -8.996481 -3.733519
# sample estimates:
# mean of x mean of y
#    19.735    26.100
```

4. Conclusions and Assumptions

Conclusions

1. Some Conculsions

Assumptions

1. Use t interval as not sure if data is normally distributed
2. Assume unequal variances for t distribution confidence interval
3. Central limit theorm for Z test
4. n must be large enough to be statistically signifcant
5. If n is small then Gossett's T test is used, n is small for each set of tests so use t test
6. Assuming a constant variance between groups of Guinea Pigs receiving difference amounts of treatment and different supplements