

**Name:** Parth Pareek

**UNI:** PP2547

**Date:** 2/10/2016

**Assignment:** HW2

1.
  - a. 95% CI is varying and is in the range of 0 to 0.3
  - b. When mean is unknown, chi-squared distribution can be used. 95% CI is 0.4731 to 3.3328

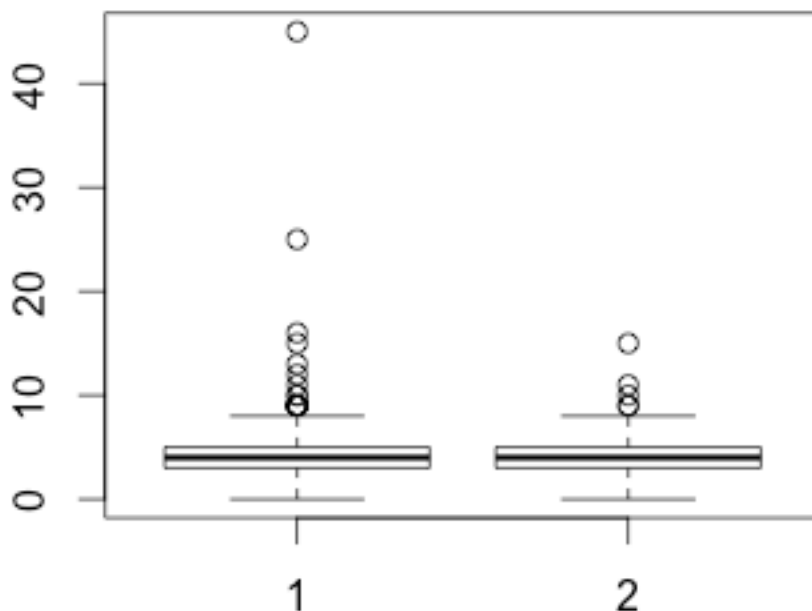
When 1000 samples are drawn, the coverage varies between 87.5% to 97.5%

When 10000 samples are drawn, the coverage varies between 90% to 97%

When 100000 samples are drawn, the coverage is about 95% to 98%

2. On paper (separate)

- 3.



*One sided p-values in different scenarios*

With all observations: 0.3963

Removing #646: 0.4805

Removing #646 and #645: 0.5386

Since removing data points does not affect the result of the analyses, we will report data including outliers.

95% Confidence Interval: -0.4815 to 0.6305

- 4.

*One sided p-values in different scenarios*

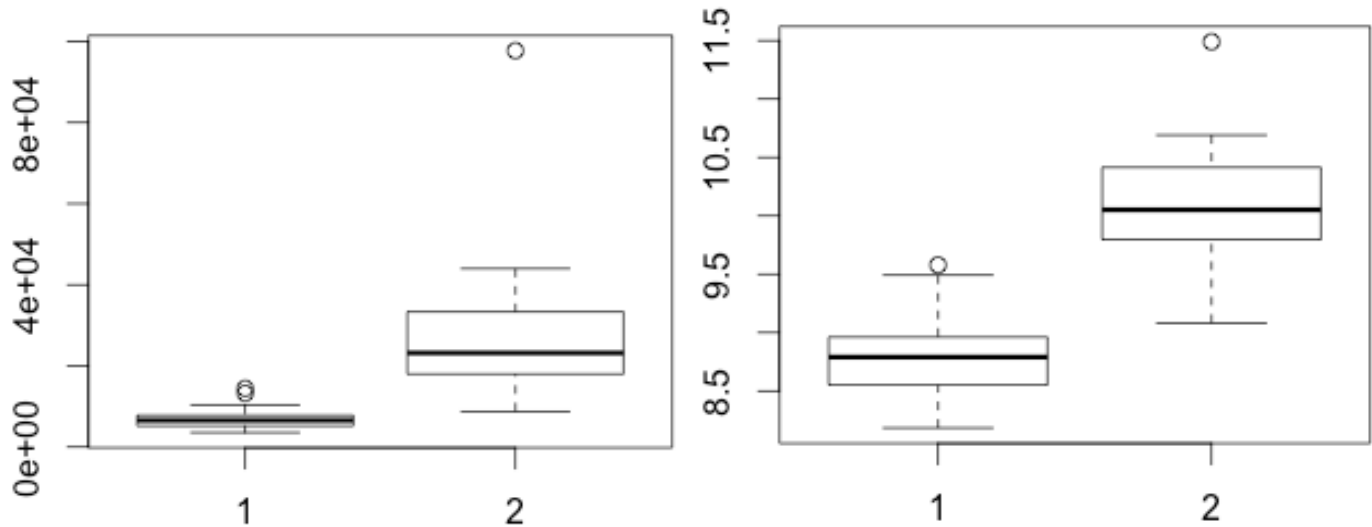
With all observations: 0.0405

Removing lowest observation (length = 0.659 in): 0.09

Yes, conclusions depend on this observation. We notice that the result changes when the lowest observation is excluded from the analyses.

Hence, we can use resistant analyses or report findings from both data sets.

5.



The left boxplot between tuitions of public and private institutes shows there is a huge variation. We will take log-transformation and work on that data set (shown in right)

	<b>Public</b> <b>In state vs. Out of state</b>		<b>In State</b> <b>Private vs. Public</b>		<b>Out of State</b> <b>Private vs. Public</b>	
P-Value	2.132e-12		1.262e-14		0.0002	
95% CI (log transformed)	0.6970	0.9588	1.0660	1.5464	0.2382	0.7184
<b>95% CI Medians' Ratio (back transformation)</b>	<b>2.01</b>	<b>2.61</b>	<b>2.90</b>	<b>4.69</b>	<b>1.27</b>	<b>2.05</b>
Means (log transformed)	9.6467	8.8188	10.1250	8.8188	10.1250	9.6467
Mean Difference (log transformed)	0.8279		1.3062		0.4783	
<b>Median Ratio (back transformation)</b>	<b>2.29</b>		<b>3.69</b>		<b>1.61</b>	

- Low p-value suggests there is a difference between public out of state vs. in state tuitions. Analyses estimates that ratio of medians of public out-of-state to in-state tuition is 2.29 (95% CI is 2.01 to 2.61)
- Low p-value suggests there is a difference between in state public vs private tuitions. Analyses estimates that ratio of medians in-state private to public tuition is 3.69 (95% CI is 2.9 to 4.69)
- Low p-value suggests there is a difference between in state public vs private tuitions. Analyses estimates that ratio of medians of out-of-state private to public tuition is 1.61 (95% CI is 1.27 to 2.05)

6.

- a. Using rank sum test, two-sided p-value is 0.1719
  - b. Based on Normal Approximation
  - c. Continuity correction used (Reverse calculation suggests, observed rank mean was  $631.5 = 631 + 0.5$ )
  - d. The p-value calculated using rank-sum test is 0.1719. Using two-sample test p-value is 0.0809; excluding one observation it is 0.18. Rank-sum test gives p-value closer to one where outlier was excluded
  - e. Two-sample t-test removing outliers let's us experiment with each observation. This way, we can also question if the observation is correct and revisit the experiment. The rank sum test, on the other hand, is more resistant to changes. It makes use of all data points and removes the need to exclude any outliers.
- 

## **CODES**

### **#Question 1**

```
a <- rnorm(10, mean = 0, sd = 1)
variance <- vector(mode="numeric", length = 0)
```

```
mean = 0
s = 1
n = 10
```

```
#mu = 0
l1 <- 0
u1 <- ((sqrt(10)*mean(a))/1.96)^2
CI1 <- c(l1,u1)
```

```
#mu unknown
l2 <- ((n-1)*1/qchisq(0.975, n-1))
u2 <- ((n-1)*1/qchisq(0.025, n-1))
CI2 <- c(l2,u2)
```

```
for (i in 1:1000){
  set.seed(i)
  b <- rnorm(10, , sd = 1)
  variance[i] = var(b)
}
```

```
mean(variance <= l1 | variance >= u1)
```

### **#Question 3**

```
library(Sleuth3)
attach(case0302)
group1 <- vector(mode="numeric", length = 0)
group2 <- vector(mode="numeric", length = 0)
for (i in 1:nrow(case0302)){
  ifelse(case0302$Veteran[i] == "Vietnam", group1 <-
c(case0302$Dioxin[i],group1),group2 <- c(case0302$Dioxin[i],group2))
```

```

}
boxplot(group1,group2)
group1 <- sort(group1)
group1_1 <- vector(mode="numeric", length = 0)
group1_2 <- vector(mode="numeric", length = 0)
for (j in 1:(length(group1)-1)){
  group1_1[j]=group1[j]
}
for (k in 1:(length(group1)-2)){
  group1_2[k]=group1[k]
}

t.test(group1_1,group2,alternative = "greater", var.equal=TRUE) #removing
one data point
t.test(group1_2,group2,alternative = "greater", var.equal=TRUE) #removing
both data points
t.test(group1,group2,alternative = "greater", var.equal=TRUE) #for p-value
t.test(group1,group2, var.equal=TRUE) #for CI

```

#### **#Question 4**

```

library(Sleuth3)
attach(ex0221)
group1 <- vector(mode="numeric", length = 0)
group2 <- vector(mode="numeric", length = 0)
for (i in 1:nrow(ex0221)){
  ifelse(ex0221$Status[i] == "Perished", group1 <-
c(ex0221$Humerus[i],group1),group2 <- c(ex0221$Humerus[i],group2))
}
boxplot(group1,group2)
group1 <- sort(group1)
group1_1 <- vector(mode="numeric", length = 0)
for (j in 2:length(group1)){
  group1_1[j-1]=group1[j]
}

t.test(group1,group2,alternative="less", var.equal=TRUE) #with all
observations
t.test(group1_1,group2,alternative="less", var.equal=TRUE) #without lowest
observation

```

#### **#Question 5**

```

library(Sleuth3)
attach(ex0332)

group_Public <- subset(ex0332,Type == "Public")
group_Private <- subset(ex0332,Type == "Private")

boxplot(group_Public$InState,group_Private$InState)

```

```
t.test((log(group_Public$OutOfState)-log(group_Public$InState)),
var.equal=TRUE)
t.test(log(group_Private$InState),log(group_Public$InState),
var.equal=TRUE)
t.test(log(group_Private$OutOfState),log(group_Public$OutOfState),
var.equal=TRUE)
```

### **#Question 6**

```
library(exactRankTests)
library(Sleuth3)
attach(ex0221)
group1 <- vector(mode="numeric", length = 0)
group2 <- vector(mode="numeric", length = 0)
group <- vector(mode="numeric", length = 0)

group = ex0221
group$rank <- rank(group$Humerus, ties.method = c("average"))

for (i in 1:nrow(ex0221)){
  ifelse(group$Status[i] == "Perished", group1 <-
c(group$Humerus[i],group1),group2 <- c(group$Humerus[i],group2))
}
wilcox.exact(group1,group2,alternative="two.sided")

#b - uses Normal approximation

rbar <- mean(group$rank)
sr <- sqrt(var(group$rank))
mean_t <- rbar * length(group1)
sd_t <- sr *
sqrt((length(group1)*length(group2))/(length(group1)+length(group2)))
qnorm(0.1719/2, mean = mean_t, sd = sd_t)
```