

Homework2

Yanyu Zheng yz2690

February 8, 2016

Problem 1

a

Since μ is 0, $\frac{\sum_{i=1}^n X_i^2}{n\sigma^2}$ follows chi-square(n). Suppose (a,b) is the (0.025, 0.975) quantile of chi-square(n). The 95% confidence interval of σ^2 is $(\frac{\sum_{i=1}^n X_i^2}{b}, \frac{\sum_{i=1}^n X_i^2}{a})$

b

Since μ is 0, we estimate μ by $\bar{x} = \frac{\sum_{i=1}^n X_i}{n}$. And $\frac{\sum_{i=1}^n (X_i - \bar{x})^2}{n\sigma^2}$ follows chi-square(n-1). Suppose (c,d) is the (0.025, 0.975) quantile of chi-square(n-1). The 95% confidence interval of σ^2 is $(\frac{\sum_{i=1}^n (X_i - \bar{x})^2}{d}, \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{c})$

c

```
set.seed(1)
outcome = vector()
upper = qchisq(0.975,10)
lower = qchisq(0.025,10)
for(i in 1:1000){
  sample = rnorm(10)
  var = var(sample)
  if(9*var<=upper && 9*var>=lower){
    outcome = c(outcome,1)
  }else{
    outcome = c(outcome,0)
  }
}
perc = sum(outcome)/length(outcome)
```

The percentage of coverage is 0.935.

Problem 2

a

```
## Load the data and library
library("Sleuth3")
library("dplyr")
attach(case0802)
## Clean the data
Y1 = case0802 %>%
  filter(Voltage == 26) %>%
  select(Time) %>%
  transmute(logTime = log(Time))
Y2 = case0802 %>%
  filter(Voltage == 28) %>%
  select(Time) %>%
  transmute(logTime = log(Time))
Y1 = as.matrix(Y1)
Y2 = as.matrix(Y2)
tStatistic = mean(Y1)-mean(Y2)
```

$Y_1 = 1.7561323, 7.3648763, 7.750916$ $Y_2 = 4.2319302, 4.6848128, 4.7031133, 6.0546037, 6.9731684$

b

The difference is 0.2944492

c

The antilogarithm of the estimate is 1.3423868 It's the multiplicative treatment effect on the original scale of measurement.

d

```
n1 = nrow(Y1)
n2 = nrow(Y2)
s1 = var(Y1)
s2 = var(Y2)
sp = sqrt((n1-1)*s1^2 + (n2-2)*s2^2)/sqrt(n1+n2-2)
se = sp*sqrt(1/n1+1/n2)
lb = qt(0.025, n1+n2-2)
ub = qt(0.975, n1+n2-2)
ci = c(tStatistic+lb*se, tStatistic+ub*se)
ratio = exp(ci[2])
#
```

The 95% confidence interval is $-11.4372856, 12.026184$. The antilogarithms of the endpoints is 1.6707264×10^5 . Which means $\frac{\bar{Y}_1}{\bar{Y}_2}$ has a probability of 95% to be less than 1.6707264×10^5

Problem3

```
attach(case0302)
##With the two largest dioxin levels in the Vietnam veterans group
t.test(Dioxin~Veteran,case0302, alternative = "less")
```

```
##
## Welch Two Sample t-test
##
## data: Dioxin by Veteran
## t = -0.29122, df = 136.96, p-value = 0.3857
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 0.3491265
## sample estimates:
## mean in group Other mean in group Vietnam
##      4.185567      4.260062
```

```
case0302[which.max(Dioxin),]
```

```
##      Dioxin Veteran
## 646      45 Vietnam
```

```
case0302D = case0302[-which.max(Dioxin),]
case0302D[which.max(case0302D$Dioxin),]
```

```
##      Dioxin Veteran
## 645      25 Vietnam
```

```
case0302D = case0302D[-which.max(case0302D$Dioxin),]
##Without the two largest dioxin levels in the Vietnam veterans group
t.test(Dioxin~Veteran,case0302D, alternative = "less")
```

```
##
## Welch Two Sample t-test
##
## data: Dioxin by Veteran
## t = 0.0853, df = 117.33, p-value = 0.5339
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 0.4285707
## sample estimates:
## mean in group Other mean in group Vietnam
##      4.185567      4.164596
```

Problem4

```
attach(ex0221)
t.test(Humerus~Status, ex0221)
```

```
##
## Welch Two Sample t-test
##
## data: Humerus by Status
## t = -1.7207, df = 43.824, p-value = 0.09236
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.021894675 0.001728008
## sample estimates:
## mean in group Perished mean in group Survived
## 0.7279167 0.7380000
```

```
ex0221D = ex0221[-which.min(Humerus),]
t.test(Humerus~Status, ex0221D)
```

```
##
## Welch Two Sample t-test
##
## data: Humerus by Status
## t = -1.373, df = 48.967, p-value = 0.176
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.017460170 0.003286257
## sample estimates:
## mean in group Perished mean in group Survived
## 0.730913 0.738000
```

- P-value with outlier: 0.09236
- p-value without outlier: 0.176

Yes. The conclusion does not depends on this one single data.

We should look into this special case to see if there was error in the process of obtaining it. If there is, we can either correct or exclude it from the sample. If not, we can try test methods that's more robust to outliers like rank-sum test. And if we still want to stick with t-test, we have to first examine if the distribution is normal. If it is, draw more samples.

Problem5

```
attach(ex0332)
## Use paired t-test for problem(a)
ex0332D = ex0332[Type=="Public",]
t.test(ex0332D$InState, ex0332D$OutOfState, alternative = "less", paired = TRUE)
```

```
##
## Paired t-test
##
## data: ex0332D$InState and ex0332D$OutOfState
## t = -13.805, df = 24, p-value = 3.264e-13
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -7918.213
```

```
## sample estimates:
## mean of the differences
##                -9038.4

## Use Welch t-test for problem(b)
t.test(InState~Type, ex0332, alternative = "greater")

##
## Welch Two Sample t-test
##
## data: InState by Type
## t = 5.9915, df = 25.081, p-value = 1.455e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  15197.82      Inf
## sample estimates:
## mean in group Private  mean in group Public
##           28409.84           7152.36

## Use Welch t-test for problem(c)
t.test(OutOfState~Type, ex0332, alternative = "greater")

##
## Welch Two Sample t-test
##
## data: OutOfState by Type
## t = 3.3733, df = 27.145, p-value = 0.001124
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6050.507      Inf
## sample estimates:
## mean in group Private  mean in group Public
##           28409.84           16190.76
```

From the p-values, we can see how much they differ.

Problem6

a

```
attach(ex0221)

## The following objects are masked from ex0221 (pos = 4):
##
##      Humerus, Status

wilcox.test(Humerus~Status, ex0221)
```

```
## Warning in wilcox.test.default(x = c(0.659, 0.689, 0.703, 0.702, 0.709, :
## cannot compute exact p-value with ties

##
## Wilcoxon rank sum test with continuity correction
##
## data:  Humerus by Status
## W = 331, p-value = 0.1718
## alternative hypothesis: true location shift is not equal to 0
```

The two-sided p-value is 0.1718

b

It reported the one based on the normal approximation.

c

Yes. It use a continuity correction.

d

The p-value of the rank-sum test is higher compared to t-test with the smallest observation and lower than the t-test without the smallest observation.

e

Two-sample t-test is more powerful but less robust. And the strategy we can use for dealing with outliers are limited, especially if the outlier is sampled from the targeted distribution, obtained without error. Then we can only try to get more samples to diminish the influence of a single outlier. While rank-sum test is more robust but less powerful.