Take Home Final

Yigao Li

December 18, 2017

Part I

1

a)

$$\int_{-A}^{A} f(x)dx = 1$$
$$\int_{-A}^{A} c(A^2 - x^2)dx = 1$$
$$\int_{-A}^{A} (cA^2 - cx^2)dx = 1$$
$$cA^2x - \frac{1}{3}cx^3 \Big|_{-A}^{A} = 1$$
$$cA^3 - \frac{1}{3}cA^3 - (-cA^3 + \frac{1}{3}cA^3) = 1$$
$$\frac{4}{3}cA^3 = 1$$
$$c = \frac{3}{4A^3}$$

b)

Sample mean

$$\mu = E[f(x)] = \int_{-A}^{A} x\frac{3}{4A^3}(A^2 - x^2)dx$$
$$= \int_{-A}^{A} (\frac{3x}{4A} - \frac{3x^3}{4A^3})dx$$

Since $\frac{3x}{4A} - \frac{3x^3}{4A^3}$ is an odd function and domain $[-A, A]$ is symmetric,

$$\mu = E[f(x)] = \int_{-A}^{A} (\frac{3x}{4A} - \frac{3x^3}{4A^3})dx = 0$$
$$\mu_{\bar{X}} = \mu = 0$$

Sample Variance

$$
\begin{aligned}
E[(f(x))^2] &= \int_{-A}^{A} x^2 \frac{3}{4A^3}(A^2 - x^2)dx \\
&= \int_{-A}^{A} (\frac{3x^2}{4A} - \frac{3x^4}{4A^3})dx \\
&= \frac{x^3}{4A} - \frac{3x^5}{20A^3} \Big|_{-A}^{A} \\
&= \frac{A^2}{4} - \frac{3A^2}{20} + \frac{A^2}{4} - \frac{3A^2}{20} \\
&= \frac{A^2}{5}
\end{aligned}
$$

$$Var(f(x)) = E[(f(x))^2] - E^2[f(x)] = \frac{A^2}{5}$$

$$\sigma = \sqrt{Var(f(x))} = \frac{A}{\sqrt{5}}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{A}{\sqrt{5n}}$$

By **Central Limit Theorem**, the sample mean $\bar{X}$ for sample size n approximately follows normal distribution with mean 0 and variance $\frac{A^2}{5n}$ ($\bar{X} \sim N(0, \frac{A^2}{5n})$).

2

a)

$$P\{Y_i = x\} = \begin{cases} p + (1-p)P\{X_i = 0\} \text{ when } x = 0 \\ (1-p)P\{X_i = x\} \text{ otherwise} \end{cases}$$

$$P\{Y_i = x\} = \begin{cases} p + (1-p)e^{-\lambda} \text{ when } x = 0 \\ \frac{(1-p)e^{-\lambda}\lambda^x}{x!} \text{ otherwise} \end{cases}$$

b)

When $p = \frac{1}{3}$,

$$P\{Y_i = x\} = \begin{cases} \frac{1+2e^{-\lambda}}{3} \text{ when } x = 0 \\ \frac{2e^{-\lambda}\lambda^x}{3x!} \text{ otherwise} \end{cases}$$

Likelihood function given sample

$$
\begin{aligned}
\mathcal{L}(\lambda|\text{sample}) &= (\frac{1+2e^{-\lambda}}{3})^2(\frac{2}{3}e^{-\lambda}\lambda)^2(\frac{1}{3}e^{-\lambda}\lambda^2)^3(\frac{1}{36}e^{-\lambda}\lambda^4)^2\frac{1}{180}e^{-\lambda}\lambda^5 \\
\ell(\lambda) = \log\mathcal{L}(\lambda) &= \log[(\frac{1+2e^{-\lambda}}{3})^2(\frac{2}{3}e^{-\lambda}\lambda)^2(\frac{1}{3}e^{-\lambda}\lambda^2)^3(\frac{1}{36}e^{-\lambda}\lambda^4)^2\frac{1}{180}e^{-\lambda}\lambda^5] \\
&= 2\log(1+2e^{-\lambda}) - 2\log 3 + 2\log\frac{2}{3} - 2\lambda + 2\log\lambda - 3\lambda + 6\log\lambda - 3\log 3 - 2\lambda + 8\log\lambda - 2\log 36 - \lambda + 5\log\lambda \\
&\quad - \log 180 \\
\ell'(\lambda) &= -\frac{4e^{-\lambda}}{1+2e^{-\lambda}} - 8 + \frac{21}{\lambda} \overset{\text{set}}{=} 0 \\
\lambda &\approx 2.54
\end{aligned}
$$

3

```r
set.seed(1)
sample.median <- c()
median.mean <- c()
n <- 25
N <- 100
for (i in 1:N){
  s3 <- sample(1:100, n, replace = FALSE)
  sample.median <- c(sample.median, median(s3))
  median.list <- replicate(1000, median(sample(s3, n, replace = TRUE)))
  median.mean <- c(median.mean, mean(median.list))
}
t.test(sample.median, median.mean, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sample.median and median.mean
## t = 0.46828, df = 99, p-value = 0.6406
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3933567  0.6363767
## sample estimates:
## mean of the differences
##                0.12151
```

Sample S contains 25 random integers from 1 to 100, and then we obtain 1000 bootstrap samples from S and calculate average of 1000 medians. We save sample medians and median averages to paired vectors. By making a two-sample paired t test, we want to study whether there's difference between them. P-value is 0.6406, which is far larger than 0.0500. We fail to reject null hypothesis. There is not enough evidence to show that $E[X] = m$ is always true.

Furthermore, in this case, m is the estimate computed from sample S. Thus, bootstrap cannot give better parameter estimates.

4

```r
set.seed(2)
N <- 30000
x <- rnorm(N)
y <- 1/x[x>1]
mu <- mean(y)
sigma.sq <- var(y)
cat("E[Y] =", mu)
```

```
## E[Y] = 0.7028457
```

```r
cat("Var(Y) =", sigma.sq)
```

```
## Var(Y) = 0.03013073
```

We want $P\{|\bar{Y}_n - \mu_Y| \leq 0.01\} \geq 0.99$. By **Chebyshev Inequality** for sample mean,

$$P\{|\bar{Y}_n - \mu_Y| \geq 0.01\} \leq \frac{\sigma^2}{0.01^2 n}$$

From the sampling result above, σ^2 is approximately 0.03. Therefore, n must be chosen so that

$$\frac{0.03}{0.01^2 n} \leq (1 - 0.99) = 0.01$$
$$n \geq 30,000$$

It requires the number of simulations to be at least 30,000 so that we can achieve two decimal digits accuracy for 99% of the time.

Part II

5

H_0: Age of mother and length of pregnancy are independent.
H_1: Age of mother and length of pregnancy are dependent.

(a)

```
set.seed(3)
birth <- read.csv("D:/Courses/ANLY 511/NCBirths2004.csv")
mytest.1 <- function(mydf){
  agg <- aggregate(Gestation ~ MothersAge, data = mydf, FUN = mean)
  return(agg$Gestation[1] - agg$Gestation[2])
}
permute.sample.1 <- function(mydf){
  n <- dim(mydf)[1]
  mydf$MothersAge <- mydf$MothersAge[sample(n, n, replace = F)]
  return(mytest.1(mydf))
}
birth.permute <- birth
N <- 1000
test.1 <- replicate(N, permute.sample.1(birth.permute))
cat("P-value =", mean(test.1 > mytest.1(birth)))
```

```
## P-value = 0.671
```

Since p-value is 0.671, greater than 0.05, we fail to reject H_0. There's not enough evidence to show that the age of the mother and the length of pregnancy are dependent.

(b)

```
mytable <- table(birth$MothersAge, birth$Gestation)
mytable
```

```
##
##              37  38  39  40  41  42
##    15-19      9  30  25  33  12   1
##    20-24     25  58  85  75  31   5
##    25-29     16  58 102  68  33   1
##    30-34     25  47  74  55  19   2
##    35-39      6  20  35  27   6   1
```

```
##   40-44       3   6   4   6   2   0
##   45-49       0   0   1   1   0   0
##   under 15    0   0   1   0   1   0
```

```
chisq.test(mytable)
```

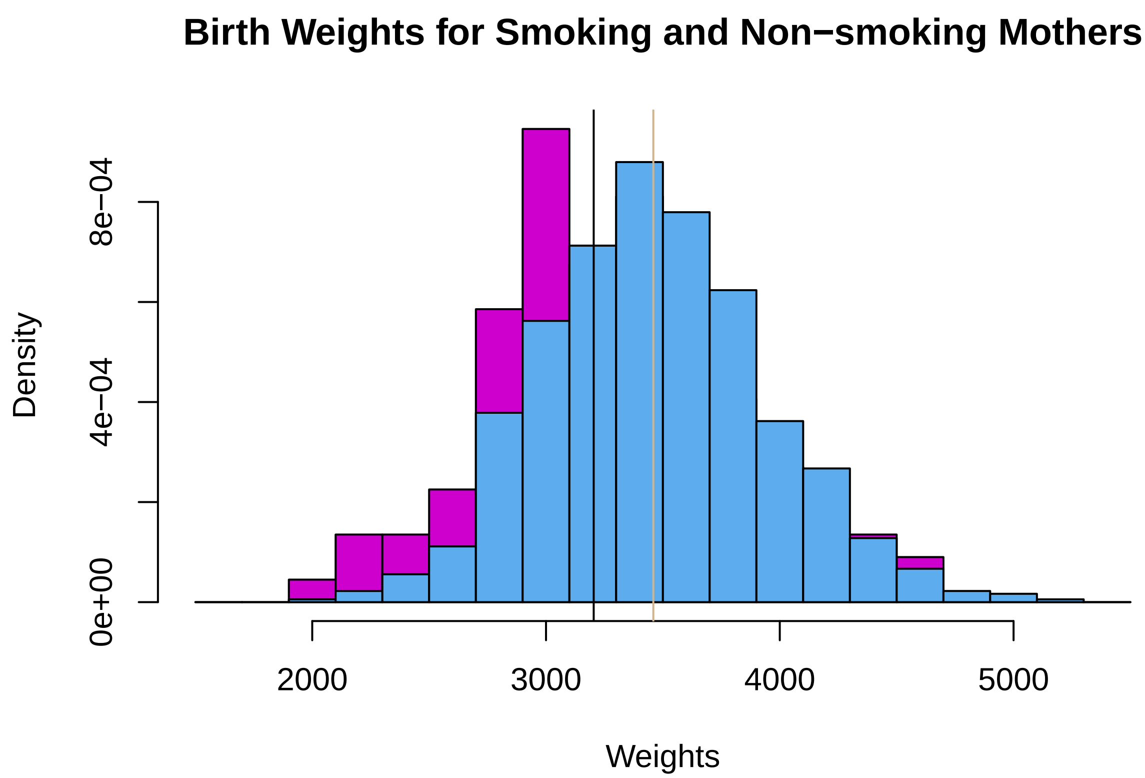
```
## Warning in chisq.test(mytable): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mytable
## X-squared = 28.526, df = 35, p-value = 0.7723
```

χ^2 test statistics is 28.526 with degree of freedom 35
Because the p-value is 0.7723, greater than 0.05, there's still not enough evidence that the age of the mother and the length of pregnancy are dependent.

6

```
set.seed(4)
weight.smoke <- birth$Weight[birth$Smoker == "Yes"]
weight.non <- birth$Weight[birth$Smoker == "No"]
hist(weight.smoke, freq = F, breaks = seq(1500,5500,200), col = "magenta3",
     main = "Birth Weights for Smoking and Non-smoking Mothers", xlab = "Weights")
hist(weight.non, freq = F, breaks = seq(1500,5500,200), col = "steelblue2", add = T)
abline(v = median(weight.smoke), col = "gray0")
abline(v = median(weight.non), col = "tan")
```




```r
ls <- length(weight.smoke)
ln <- length(weight.non)
N <- 1000
weight.smoke.boot <- replicate(N, median(sample(weight.smoke, ls, replace = TRUE)))
weight.non.boot <- replicate(N, median(sample(weight.non, ln, replace = TRUE)))
var.smoke <- var(weight.smoke.boot)/N
var.non <- var(weight.non.boot)/N
test.stat <- (mean(weight.non.boot) - mean(weight.smoke.boot))/sqrt(var.smoke + var.non)
cat("Test statistics is", test.stat)
```

```
## Test statistics is 109.1062
```

```r
degree <- (var.smoke^2 + var.non^2)^2/(var.smoke^2/(N-1) + var.non^2/(N-1))
crit <- qt(0.975, df = degree)
cat("Critical value at 5% significance level is", crit)
```

```
## Critical value at 5% significance level is 1.960059
```

```r
cat("Test statistics greater than critical value is", test.stat > crit)
```

```
## Test statistics greater than critical value is TRUE
```

```r
quantile(weight.non.boot - weight.smoke.boot, 0.05)
```

```
##  5%
## 113
```

Figure above is an overlapping histogram of birth weights for smoking and non-smoking mothers. Purple is for smoking mothers and Blue is non-smoking. Black line is the median of weights for smoking mothers and orange line is for non-smoking.

Based on this histogram, we assume that median birth weight for smoking mothers is less than that for non-smoking mothers. Therefore, we construct a one-sided hypothesis testing with bootstrap.

H_0: Median birth weight is the same for both smoking and non-smoking mothers
H_1: Median birth weight for smoking mothers is less than that for non-smoking mothers

From the R code results above, the test statistics, 109.1062, is greater than the critical value at 5% significance level and null hypothesis value 0 does not lie in the confidence interval ($[113, +\infty)$). Therefore, we conclude that median birth weights for smoking mothers is less than median birth weights for non-smoking mothers.

7

```r
weight37 <- birth$Weight[birth$Gestation == 37]
weight38 <- birth$Weight[birth$Gestation == 38]
weight39 <- birth$Weight[birth$Gestation == 39]
weight40 <- birth$Weight[birth$Gestation == 40]
weight41 <- birth$Weight[birth$Gestation == 41]
t.test(weight38, weight37, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  weight38 and weight37
## t = 4.2934, df = 141.45, p-value = 1.624e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  164.3856      Inf
## sample estimates:
## mean of x mean of y
##  3298.986  3031.417
```

H_0: $\text{Weight}_{38} = \text{Weight}_{37}$
H_1: $\text{Weight}_{38} > \text{Weight}_{37}$
Two-sample one-sided t test
Confidence interval is $[164.3856, +\infty)$

```r
t.test(weight39, weight38, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  weight39 and weight38
## t = 4.1078, df = 459.78, p-value = 2.364e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  98.02006      Inf
## sample estimates:
## mean of x mean of y
##  3462.688  3298.986
```

H_0: $\text{Weight}_{39} = \text{Weight}_{38}$
H_1: $\text{Weight}_{39} > \text{Weight}_{38}$
Two-sample one-sided t test
Confidence interval is $[98.02006, +\infty)$

```r
t.test(weight40, weight39, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  weight40 and weight39
## t = 3.2391, df = 568.32, p-value = 0.0006345
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  58.65858      Inf
## sample estimates:
## mean of x mean of y
##  3582.068  3462.688
```

H_0: $\text{Weight}_{40} = \text{Weight}_{39}$
H_1: $\text{Weight}_{40} > \text{Weight}_{39}$
Two-sample one-sided t test
Confidence interval is $[58.65858, +\infty)$

```r
t.test(weight41, weight40, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  weight41 and weight40
## t = 2.0229, df = 178.6, p-value = 0.02229
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  19.78943      Inf
## sample estimates:
## mean of x mean of y
##  3690.413  3582.068
```

H_0: $\text{Weight}_{41} = \text{Weight}_{40}$
H_1: $\text{Weight}_{41} > \text{Weight}_{40}$
Two-sample one-sided t test
Confidence interval is $[19.78943, +\infty)$

Because we want to study weight gains between each gestation week k and $k + 1$, we construct 4 two-sample one-sided t tests for each group pair to find whether higher gestation week brings gains more weight. Overall from test results, confidence intervals are all positive, which means that null hypothesis value 0 lies outside confidence intervals. We reject H_0 and conclude that weight increases in consecutive 4 weeks from gestation week 37.

Bonus question

```r
gestation.y <- birth$Gestation[birth$Tobacco == "Yes"]
gestation.n <- birth$Gestation[birth$Tobacco == "No"]
t.test(gestation.y, gestation.n, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  gestation.y and gestation.n
```

```
## t = -1.757, df = 139.19, p-value = 0.04056
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01151172
## sample estimates:
## mean of x mean of y
##  38.93694  39.13697
```

H_0: Usage of tobacco by mothers does not change gestation length
H_1: Mother using tobacco will shorten gestation length
Two-sample one-sided t test
P-value $= 0.04056 < 0.05 = \alpha$
Reject H_0
We conclude that tobacco use by mother shortens the gestation length.