Hao-HW6

```
#setwd("~/Google Drive/CUNY/git/DATA606/Lab6")
library(dplyr)
library(ggplot2)

# a function to return confidence interval for a proportion based on p, n and cl
prop.ci = function(p, n, cl) {
  moe = qnorm(mean(c(cl, 1))) * sqrt(p * (1 - p) / n)
  return(paste("(", round(p - moe, 4), ",", round(p + moe, 4), ")"))
}

# a function to return confidence interval for the difference between two proportions based on p, n and
prop.diff.ci = function(p1, n1, p2, n2, cl) {
  moe = qnorm(mean(c(cl, 1))) * sqrt(p1 * (1 - p1) / n1 + p2 * (1 - p2) / n2)
  return(paste("(", round(p1 - p2 - moe, 4), ",", round(p1 - p2 + moe, 4), ")"))
}
```

6.6

a) False. The CI is constructed to estimate the population proportion, not the sample proportion.

b) True.

c) True.

d) False. At a 90% confidence level, the margin of error would decrease.

6.12

a) 48% is a sample statistic because it is calculated based on the sample; the population parameter would require surveying the entire population of Americans.

b) (0.4524 , 0.5076)

c) Assuming that the sample was representative of the general population, then the CI should be accurate as the sample size is large and the central limit theorem should apply.

d) No, the confidence interval straddles 50%, so it is unclear whether a majority support or do not support the legalization of marijuana.

```
prop.ci(0.48, 1259, 0.95)
```

```
## [1] "( 0.4524 , 0.5076 )"
```

6.20

599 people would need to be sampled.

```
p = 0.48
moe = 0.02
cl = 0.95
n = (p * (1-p))^2 / moe^2 * qnorm(mean(c(cl, 1)))^2
n
```

```
## [1] 598.3087
```

6.28

The 95% CI is (-0.0175 , 0.0015) which includes zero. As such, at the 95% confidence level, we cannot be sure if there is a difference between the two proportions.

```
prop.diff.ci(0.08, 11545, 0.088, 4691, 0.95)
```

```
## [1] "( -0.0175 , 0.0015 )"
```

6.43 Practice

H_0: No option is favored above any others
H_A: At least one option is favored above one or more others

Since the p-value is below 5%, we reject the null hypothesis and accept the alternative that at least one option is favored above one or more others.

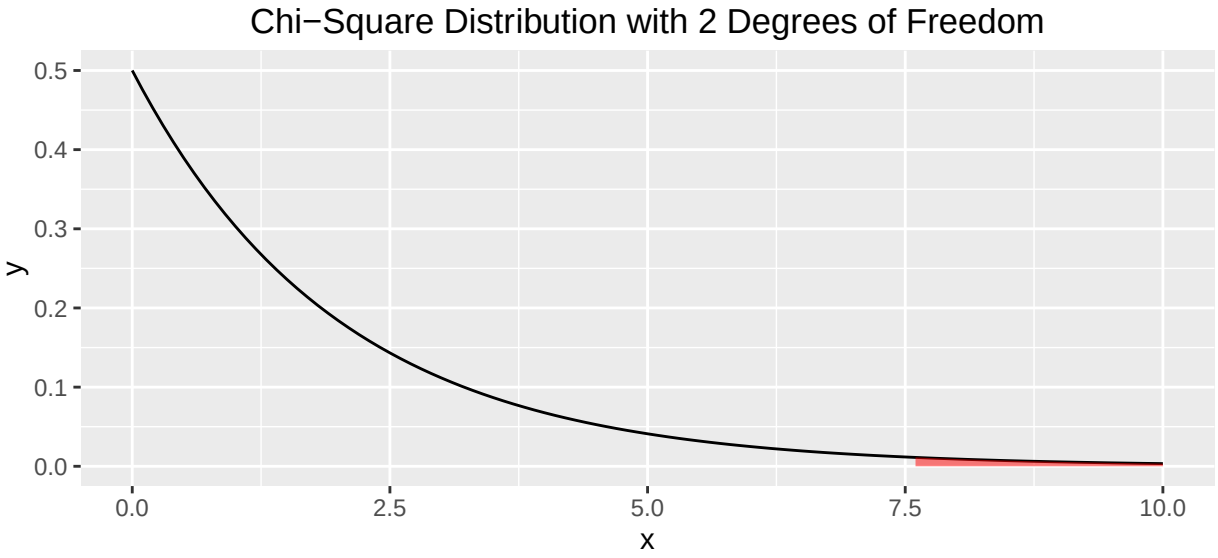
```
# create data frame to house actual and expected results
rps = data.frame(actual = c(43, 21, 35), expected = c((43 + 21 + 35)/3, (43 + 21 + 35)/3, (43 + 21 + 35)
                 row.names = c('Rock', 'Paper', 'Scissor'))
chi.sq = as.numeric(rps %>% mutate(Z2 = (actual - expected)^2 / expected) %>%
  summarise(chi.sq = sum(Z2)))
df = nrow(rps) - 1
1 - pchisq(chi.sq, df)
```

```
## [1] 0.02334025
```

```
chisq.test(rps$actual)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  rps$actual
## X-squared = 7.5152, df = 2, p-value = 0.02334
```

```
chi.sq.stat = chisq.test(rps$actual)
x = seq(0, 10, 0.1)
xy = data.frame(x = x, y = dchisq(x, df))
shade = xy[x > chi.sq.stat$statistic, ]
xy %>% ggplot(aes(x =  x, y = y)) +
  geom_line() +
  geom_ribbon(data = shade, aes(ymax = y, ymin= 0), fill = 'red', alpha = 0.5) +
  ggtitle(paste('Chi-Square Distribution with', df, 'Degrees of Freedom'))
```



6.44

a) H_0: Barking deer have no preference as it relates to foraging in certain habitats
H_A: Barking deer do have a preference as it relates to foraging in certain habitats

b) Chi-square test for one-way table

c) Independence: Met, the type of habitat for one site should not affect the others
Sample size / distribution: Met, each habitat type has at least 5 expected cases

d) The p-value is essentially zero, so we can reject the null hypothesis and accept the alternative that Barking deer do indeed have a preference as it relates to foraging in certain habitats

```
habitat = data.frame(actual = c(4, 16, 61, 345), expected = 426 * c(0.048, 0.147, 0.396, (1-0.048-0.147
                     row.names = c('Woods', 'Cultivatd grassplot', 'Deciduous forests', 'Other'))
chi.sq = habitat %>% mutate(Z2 = (actual - expected)^2 / expected) %>%
  summarise(chi.sq = sum(Z2))
df = nrow(habitat) - 1
1 - pchisq(as.numeric(chi.sq), df)
```

```
## [1] 0
```

```
chisq.test(habitat$actual, p = c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  habitat$actual
## X-squared = 284.06, df = 3, p-value < 2.2e-16
```

6.47 Practice

```
# first reproduce example on pg 298 in book
algo = data.frame(no.new = c(3511, 1749, 1818), new = c(1489, 751, 682), row.names = c('current', 'test
algo.expected.prop = algo %>%
  mutate(total = no.new + new) %>%
  summarise(no.new.prop = sum(no.new)/sum(total), new.prop = sum(new)/sum(total))
algo2 = algo %>%
  mutate(total = no.new + new) %>%
```

```
  mutate(no.new.expected = algo.expected.prop$no.new.prop * total, new.expected = algo.expected.prop$new
chi.sq = algo2 %>% mutate(Z2a = (no.new - no.new.expected)^2 / no.new.expected, Z2b = (new - new.expecte
  summarise(chi.sq = sum(Z2a) + sum(Z2b))
df = (nrow(algo) - 1) * (ncol(algo) - 1)
1 - pchisq(as.numeric(chi.sq), df)
```

```
## [1] 0.04688013
```

```
chisq.test(t(algo))
```

```
##
##  Pearson's Chi-squared test
##
## data:  t(algo)
## X-squared = 6.1203, df = 2, p-value = 0.04688
```

```
# attempt 6.47 using the same logic
offshore = data.frame(support = c(154, 132), oppose = c(180, 126), unsure = c(104, 131),
                      row.names = c('grad', 'notGrad'))

offshore.expected.prop = offshore %>%
  mutate(total = support + oppose + unsure) %>%
  summarise(support.prop = sum(support)/sum(total), oppose.prop = sum(oppose)/sum(total), unsure.prop =

offshore2 = offshore %>%
  mutate(total = support + oppose + unsure) %>%
  mutate(support.expected = offshore.expected.prop$support.prop * total,
         oppose.expected = offshore.expected.prop$oppose.prop * total,
         unsure.expected = offshore.expected.prop$unsure.prop * total)

chi.sq = offshore2 %>% mutate(Z2a = (support - support.expected)^2 / support.expected,
                              Z2b = (oppose - oppose.expected)^2 / oppose.expected,
                              Z2c = (unsure - unsure.expected)^2 / unsure.expected) %>%
  summarise(chi.sq = sum(Z2a) + sum(Z2b) + sum(Z2c))
chi.sq
```

```
##     chi.sq
## 1 11.46082
```

```
df = (nrow(offshore) -1)* (ncol(offshore) - 1)
1 - pchisq(as.numeric(chi.sq), df)
```

```
## [1] 0.003245752
```

```
chisq.test(t(offshore))
```

```
##
##  Pearson's Chi-squared test
##
## data:  t(offshore)
## X-squared = 11.461, df = 2, p-value = 0.003246
```

6.48

a) Chi-square test for two-way table

b) H_0: There is no association between coffee intake and clinical depression
H_A: There is an association between coffee intake and clinical depression

c) 5.13% of women in the study suffer from depression; 94.86% do not

d) The expected counts appear under the yes.exp and no.exp columns in the data frame below

e) The p-value is 0.0003

f) The conclusion of the hypothesis test is to reject the null and accept the alternative that there is an association between coffee intake and clinical depression

g) Yes, because this was an observational study, no causal relationship can be established. As such, the observed association does not necessarily imply causation.

```
coffee = data.frame('1perWeek' = c(670, 11545), '2-6perWeek' = c(373, 6244), '1perDay' = c(905, 16329),
                    '2-3perDay' = c(564, 11726), '4perDay' = c(95, 2288), row.names = c('yes', 'no'))
coffee.expected.prop = coffee %>%
  mutate(total = X1perWeek + X2.6perWeek + X1perDay + X2.3perDay + X4perDay) %>%
  mutate(prop = total / sum(total)) %>%
  select(prop)
coffee.expected.prop
```

```
##         prop
## 1 0.05138059
## 2 0.94861941
```

```
coffee2 = data.frame(intake = colnames(coffee), t(coffee)) %>%
  mutate(total = yes + no) %>%
  mutate(yes.exp = coffee.expected.prop$prop[1] * total, no.exp = coffee.expected.prop$prop[2] * total)
coffee2
```

```
##        intake yes    no total  yes.exp    no.exp
## 1   X1perWeek 670 11545 12215 627.6140 11587.386
## 2 X2.6perWeek 373  6244  6617 339.9854  6277.015
## 3    X1perDay 905 16329 17234 885.4932 16348.507
## 4  X2.3perDay 564 11726 12290 631.4675 11658.532
## 5    X4perDay  95  2288  2383 122.4400  2260.560
```

```
chi.sq = coffee2 %>% mutate(Z2a = (yes - yes.exp)^2 / yes.exp, Z2b = (no - no.exp)^2 / no.exp) %>%
  summarise(chi.sq = sum(Z2a) + sum(Z2b))
chi.sq
```

```
##     chi.sq
## 1 20.93161
```

```
df = (nrow(coffee) - 1) * (ncol(coffee) -1)
1 - pchisq(as.numeric(chi.sq), df)
```

```
## [1] 0.0003267104
```

```
chisq.test(coffee)
```

```
##
##  Pearson's Chi-squared test
##
## data:  coffee
## X-squared = 20.932, df = 4, p-value = 0.0003267
```

```
chi.sq.stat = chisq.test(coffee)
x = seq(0, 50, 0.2)
```

```r
xy = data.frame(x = x, y = dchisq(x, df))
shade = xy[x > chi.sq.stat$statistic, ]
xy %>% ggplot(aes(x =  x, y = y)) +
  geom_line() +
  geom_ribbon(data = shade, aes(ymax = y, ymin= 0), fill = 'red', alpha = 0.5) +
  ggtitle(paste('Chi-Square Distribution with', df, 'Degrees of Freedom'))
```