HW7

Natalie Brewer

2023-10-23

Problem 7A

```
pval <- 1 - pchisq(2.015, 2)
pval
```

```
## [1] 0.3651307
```

Problem 7B

```
men_data <- c(3755, 3251, 3777, 3706, 3717, 3660, 3669, 3626, 3481, 3590, 3605, 3392)
ts_men <- sum(((men_data - 3602.42)^2)/3602.42)
ts_men
```

```
## [1] 74.56013
```

```
pval_men <- 1 - pchisq(ts_men, 11)
pval_men
```

```
## [1] 1.645983e-11
```

```
wom_data <- c(1362, 1244, 1496, 1452, 1448, 1376, 1370, 1301, 1337, 1351, 1416, 1226)
ts_wom <- sum(((wom_data - 1364.92)^2)/1364.92)
ts_wom
```

```
## [1] 53.78551
```

```
pval_wom <- 1 - pchisq(ts_wom, 11)
pval_wom
```

```
## [1] 1.291604e-07
```

Problem 7C

```
matrix <- matrix(c(790, 56, 21,
                    98, 11, 7,
                    209, 27, 12), nrow = 3, byrow = TRUE)
rownames(matrix) <- c("employed", "unemployed", "not in labor force")
colnames(matrix) <- c("married", "once married", "never married")
matrix
```

```
##                      married once married never married
## employed                 790           56            21
## unemployed                98           11             7
## not in labor force       209           27            12
```

```
chisq_test <- chisq.test(matrix)
```

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

```
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix
## X-squared = 13.369, df = 4, p-value = 0.009609
```

```
print(chisq_test$expected)
```

```
##                       married once married never married
## employed             772.6231    66.204712     28.172218
## unemployed           103.3729     8.857839      3.769293
## not in labor force 221.0041    18.937449      8.058489
```

Problem 7D

```
# Calculate the TS using the first technique
first_TS <- sum((matrix - chisq_test$expected)^2/chisq_test$expected)
first_TS
```

```
## [1] 13.36855
```

```
# Calculate the TS using the second technique
second_TS <- 2*sum(matrix*log(matrix/chisq_test$expected))
second_TS
```

```
## [1] 12.38856
```

Problem 7E

```
n <- sum(matrix)
n
```

```
## [1] 1231
```

```
prop_unemp <- (56 + 11 + 27)/n
prop_unemp
```

```
## [1] 0.07636068
```

```
est_sd <- sqrt(prop_unemp*(1 - prop_unemp)/n)
est_sd
```

```
## [1] 0.007569324
```

```
CI <- c(prop_unemp - (1.96 * est_sd), prop_unemp + (1.96 * est_sd))
CI
```

```
## [1] 0.06152481 0.09119656
```

Problem 7F

```
prop_employed <- (790+98+209)/n
prop_employed
```

```
## [1] 0.8911454
```

```
diff <- prop_employed - prop_unemp
diff
```

```
## [1] 0.8147847
```

```
s <- sqrt((prop_unemp*(1 - prop_unemp) + prop_unemp*(1 - prop_unemp))/n)
s
```

```
## [1] 0.01070464
```

```
CI_diff <- c(diff - (1.96 * s), diff + (1.96 * s))
CI_diff
```

```
## [1] 0.7938036 0.8357658
```

Problem 7H

```
set.seed(34)
sample <- rbinom(1000, 5, 0.4)

p_hat <- mean(sample)/5 # This is the MLE for binomial

obs_counts <- table(sample)
obs_counts
```

```
## sample
##   0   1   2   3   4   5
##  70 260 342 242  73  13
```

```
exp_counts <- 1000 * dbinom(0:5, 5, p_hat) # n * P(p_hat)
exp_counts
```

```
## [1]  74.32322 253.36894 345.49535 235.55973  80.30265  10.95012
```

```
test <- chisq.test(obs_counts, p = exp_counts/sum(exp_counts))
test
```

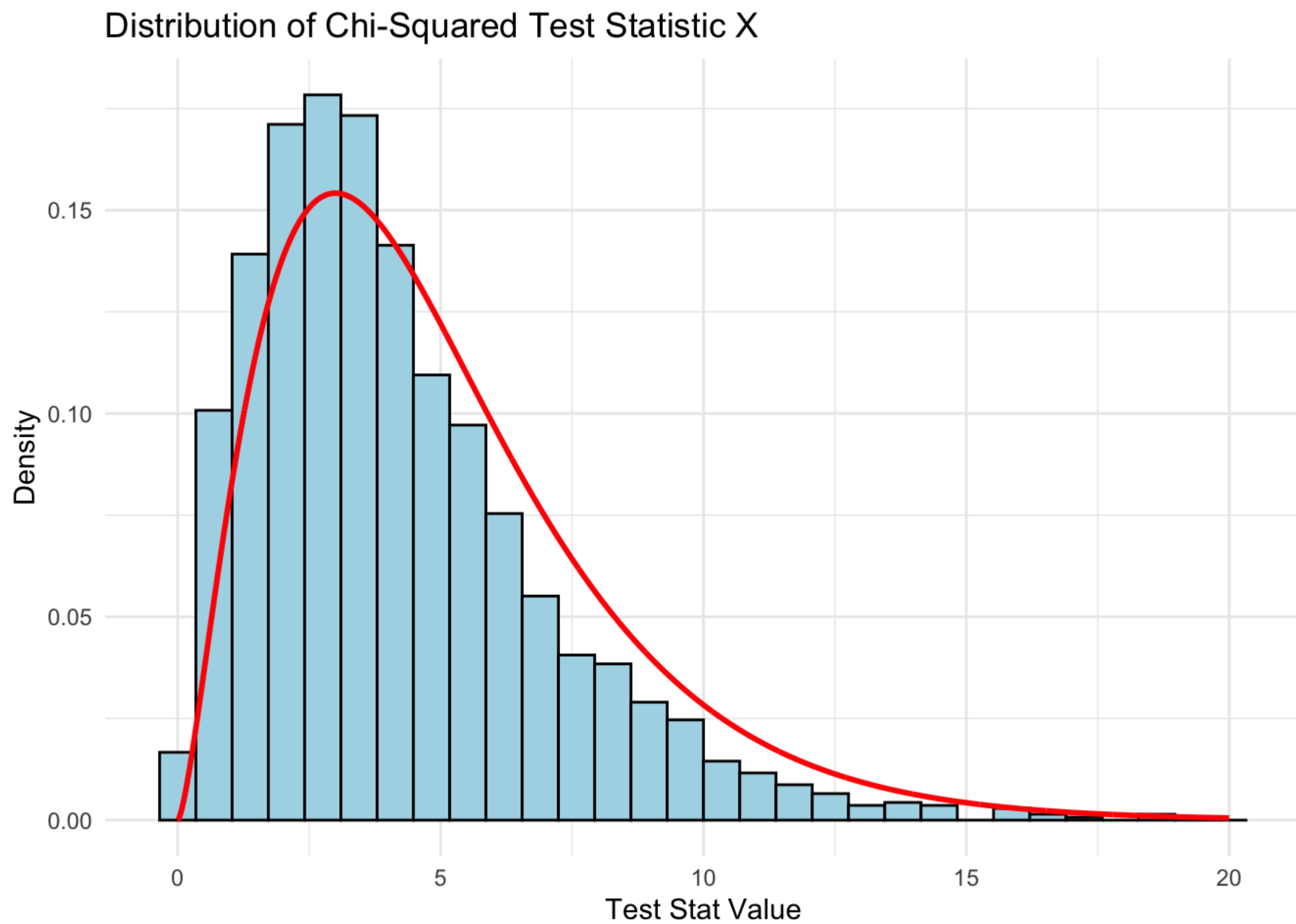
```
##
##  Chi-squared test for given probabilities
##
## data:  obs_counts
## X-squared = 1.6843, df = 5, p-value = 0.8909
```

Problem 7l

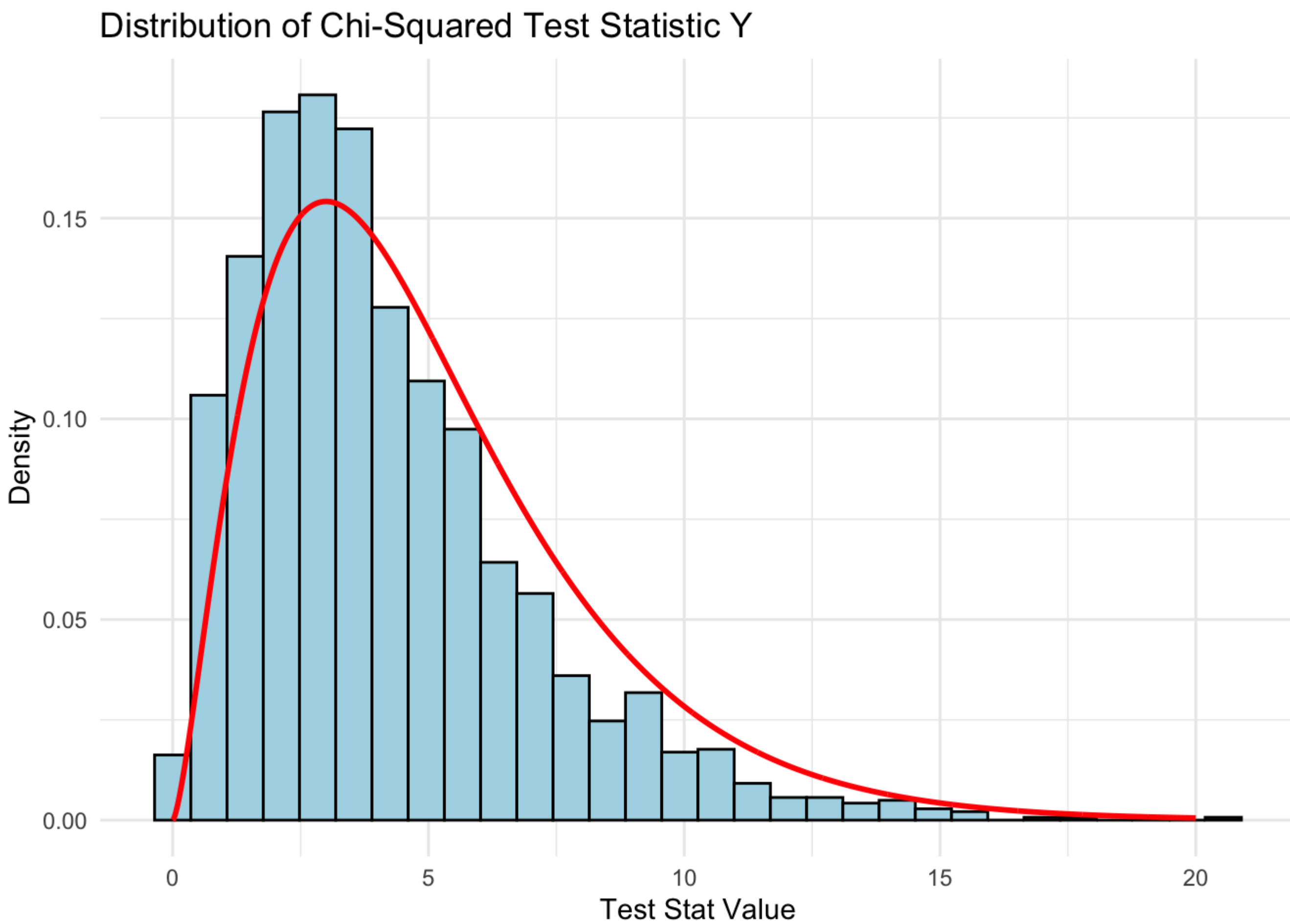
```
repeat_test <- function() {
  new_sample <- rbinom(1000, 5, 0.4)
  new_p_hat <- mean(new_sample)/5

  new_obs_counts <- table(new_sample)
  new_exp_counts <- 1000 * dbinom(0:5, 5, new_p_hat)

  new_test_X <- 2*sum(new_obs_counts*log(new_obs_counts/new_exp_counts))
  new_test_Y <- unname(chisq.test(new_obs_counts, p = new_exp_counts/sum(new_exp_count
s))$statistic)

  return(c(new_test_X, new_test_Y))
}

results <- replicate(2000, repeat_test())

df <- data.frame(X = results[1,], Y = results[2,])
head(df)
```

```
##          X        Y
## 1 4.797292 4.526754
## 2 1.393648 1.386834
## 3 5.153180 5.047197
## 4 4.482966 4.542582
## 5 1.931961 1.866721
## 6 7.163394 6.971668
```

```
x_values <- seq(0, 20, length.out = 2000)
y_values <- dchisq(x_values, 5)

ggplot(df, aes(x=X)) +
  geom_histogram(aes(y=after_stat(density)), fill="lightblue", color="black", bins=30) +
  geom_line(aes(x=x_values, y=y_values), color="red", linewidth=1) +  # Adding the chi-s
quare curve
  labs(title="Distribution of Chi-Squared Test Statistic X",
       x="Test Stat Value",
       y="Density") +
  theme_minimal()
```




```
ggplot(df, aes(x=Y)) +
  geom_histogram(aes(y=after_stat(density)), fill="lightblue", color="black", bins=30) +
  geom_line(aes(x=x_values, y=y_values), color="red", linewidth=1) +  # Adding the chi-s
quare curve
  labs(title="Distribution of Chi-Squared Test Statistic Y",
       x="Test Stat Value",
       y="Density") +
  theme_minimal()
```


Distribution of Chi-Squared Test Statistic Y
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
Test Stat Value