(PA) 9.34: Ch 8 PSS

- Find MLE and probabilities:

$$lik(\theta) = \binom{3839}{1997, 906, 904, 32} .25 (2+\theta)^{1997} (1-\theta)^{906} (1-\theta)^{904} \theta^{32}$$

$$l(\theta) = \log\left(\binom{3839}{1997,906,904,32}.25\right) + 1997 \log(2+\theta) + 906 \log(1-\theta) + 904 \log(1-\theta) + 32 \log\theta$$

$$l'(\theta) = \frac{1997}{2+\theta} + \frac{906+904}{1-\theta} + \frac{32}{\theta} = 0$$

$$\frac{-3839\theta^2 - 1655\theta + 64}{(2+\theta)(1-\theta)\theta} = 0$$

$$-3839\theta^2 - 1655\theta + 64 = 0$$

$$\Rightarrow \hat{\theta}_{ML} = 0.0357 \quad \Rightarrow \quad \begin{aligned} P_1(\hat{\theta}) &= 0.509 \\ P_2(\hat{\theta}) &= 0.241 \\ P_3(\hat{\theta}) &= 0.241 \\ P_4(\hat{\theta}) &= 0.009 \end{aligned}$$

- Expected vs Observed:

i	O_i	$E_i = 3839 P_i(\hat{\theta})$
1	1997	1953.67
2	906	925.58
3	904	925.58
4	32	34.17

- Test Stat:

$$TS = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{1877.49}{1953.67} + \frac{383.49}{925.58} + \frac{465.70}{925.58} + \frac{4.696}{34.17} = 2.015$$

$$TS \sim \chi^2_{4-1-1=2}$$

p value = 1 − pchisq(2.015, 2) = 0.365 $> \alpha$ (for <u>any</u> typical α)

So the model is a good fit for the data.

7B) 9.38

$H_0: p_{Jan} = p_{Feb} = \dots = p_{dec} = \frac{1}{12}$

H_1: not all month's probabilities are equal

n = number of men = 43229 $\Rightarrow E_i = 43229 \cdot \frac{1}{12} = 3602.42 \quad \forall i = 1, \dots, 12$

m = number of women = 16379 $\Rightarrow E_i = 16379 \cdot \frac{1}{12} = 1364.92 \quad \forall i = 1, \dots, 12$

$$TS_{men} = \sum_{i=1}^{12} \frac{(O_i - 3602.42)^2}{3602.42} = 74.56 \quad \text{(in R)}$$

$$TS_{wom} = \sum_{i=1}^{12} \frac{(O_i - 1364.92)^2}{1364.92} = 53.79 \quad \text{(in R)}$$

$$TS \sim \chi^2_{12-1=11}$$

$\text{pvalue}_{men} = 1 - \text{pchisq}(74.56, 11) = 1.65 \times 10^{-11}$

$\text{pvalue}_{wom} = 1 - \text{pchisq}(53.79, 11) = 1.29 \times 10^{-7}$

$\Big\}$ less than any typical α

So we reject H_0 and conclude that the suicide rates are seasonal.

7C

a) H_0: these are 2 independent RVs of size 3,3, ie. $\pi_{ij} = \pi_i \pi_j$

"marital status and employment status are independent"

H_1: these are 2 dependent RVs...

"marital status and employment status are not independent"

Degrees of freedom $= \dim \Omega - \dim \omega_0$

$= (3 \cdot 3 - 1) - (2 + 2) = 8 - 4 = 4$

They are dependent. (low p-value)

b) R gives a warning message because their are very low values in the expected table.

7D

$$\sum \frac{(O_i - E_i)^2}{E_i} = 13.369 \quad \text{(in R)}$$

$$2 \sum O_i \log\left(\frac{O_i}{E_i}\right) = 12.389$$

So the first technique was used by R to do the test.

7E

$n = 1231$

$\hat{p}_1 = 0.076$

$z = 1.96$

95% CI: $\hat{p}_1 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}} = 0.076 \pm 1.96(.0076)$

$= [0.062, 0.091]$

7F

$\hat{p}_2 = 0.891$

$\hat{p}_1 - \hat{p}_2 = 0.815$

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}{n}} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{n}} = 0.011$$

95% CI $= [0.794, 0.836]$

7G

$$\sum_{i=1}^{2}\frac{(x_i-np_i)^2}{np_i}=\frac{(X_1-np_1)^2}{np_1}+\frac{(X_2-np_2)^2}{np_2}=\frac{(X_1-np_1)^2}{np_1}+\frac{(\overbrace{n-X_1}^{=X_2}-n\overbrace{(1-p_1)}^{=p_2})^2}{n(1-p_1)}$$

$$=\frac{(1-p_1)(X_1-np_1)^2+p_1(n-X_1-n(1-p_1))^2}{np_1(1-p_1)}$$

$$=\frac{(1-p_1)(X_1-np_1)^2+p_1(-X+np_1)^2}{np_1(1-p_1)}$$

$$=\frac{(1-p_1)(X_1-np_1)^2+p_1(X_1-np_1)^2}{np_1(1-p_1)}=\frac{(X_1-np_1)^2}{np_1(1-p_1)}$$

7H

We have six categories : 0, 1, 2, 3, 4, 5 so the df $= 6-1=5$, which what R used for the test.

The p-value is 0.9641, which is too large to reject the null hypothesis at any standard significance level. Thus, the data appears to follow a binomial distribution. (which we know to be true!)

7I

yes, these histograms agree with the conclusions on 341-343 of the text. They show that the two test statistics are approximately equal. We can see that they are very similar density histograms compared to the χ^2_5 distribution.

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
```

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

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

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

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

Problem 7H

```r
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
```

```r
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
```

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

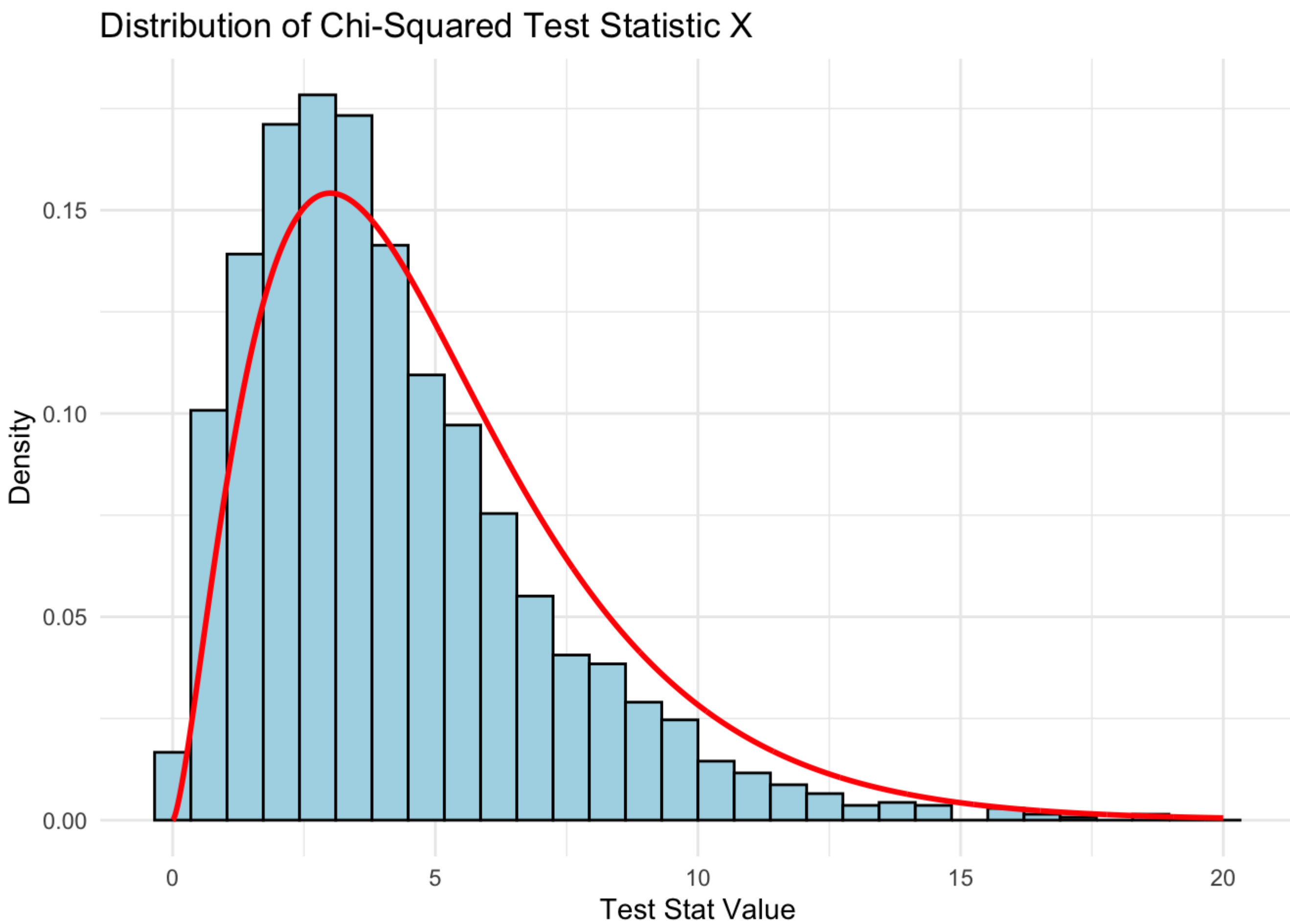
```
##
##  Chi-squared test for given probabilities
##
## data:  obs_counts
## X-squared = 1.6843, df = 5, p-value = 0.8909
```

Problem 7I

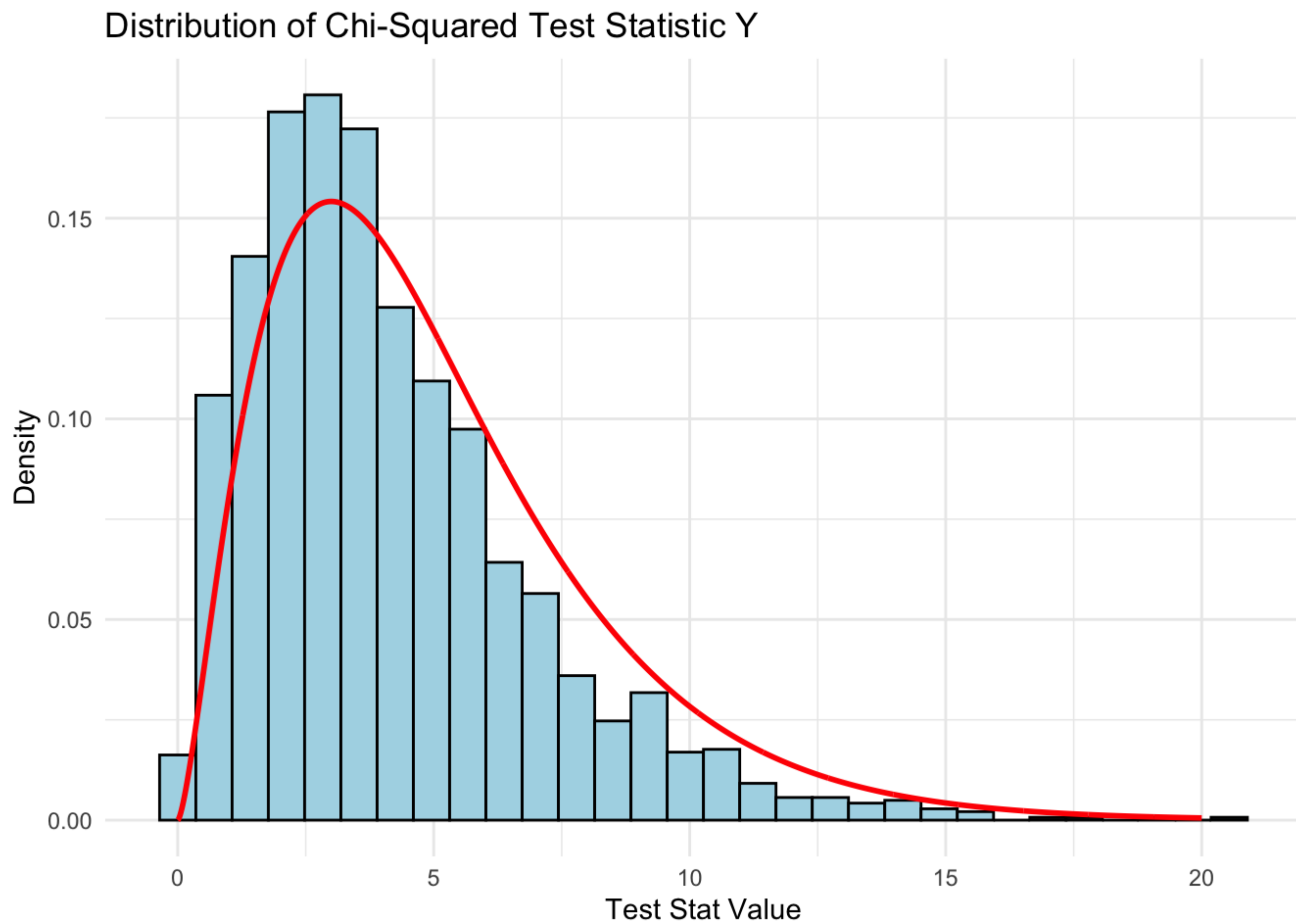
```
repeat_test <- function() {
  new_sample <- rbinom(1000, 5, 0.4)
  new_p_hat <- mean(new_sample)/5

  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
0.15
0.10
0.05
0.00
0
5
10
15
20
Test Stat Value