Lab 4

Lily Heidger

2024-02-07

Lab 4

1. Truffles are a great delicacy, sending thousands of mushroom hunters into the forest each fall to find them. A set of plots of equal size in an old-growth forest in Northern California was surveyed to count the number of truffles (Waters et al., 1997). The resulting distribution is presented in the following table. Are truffles randomly located around the forest? How can you tell? (The mean number of truffles per plot, calculated from these data, is 0.60)

H0: truffles are random (poisson) HA: truffles are not random (not poisson) Running a chi-squared test on the observed and expected frequencies results in a p-value of <0.05. This means that we reject the null, indicating that the truffles are not randomly located around the forest.

```
count <- c("0","1","2",">3")
observ <- c(203,39,18,28)
probs <- c(0.548,0.329,0.098,0.025)
exp <- c(158.7,94.83,28.45,6.65)



truffles <- data.frame(truffle_count = count, observed = observ, probs = probs, expected = exp)

chi_sq <- sum(((truffles$observed - truffles$expected)^2)/(truffles$expected))
#chi-squared test stat: 117

p_value <- chisq.test(x = truffles$observed, p = truffles$probs)

#p-value <2.2e-16
```

2. The following list gives the number of degrees of freedom and the chi-squared test statistic for several goodness-of-fit test. Find the P-value for each test.

The p-values are 0.042379078, 0.906748324, 0.008651695, 0.259177369, and 0.115302253.

```
column1 <- c(1,4,2,10,1)
column2 <- c(4.12,1.02,9.5,12.4,2.48)

deg_freedom <- data.frame(df = column1, chi_sq = column2)


p_values <- 1 - pchisq(deg_freedom$chi_sq, deg_freedom$df)


deg_freedom$p_value <- p_values
```

```
print(deg_freedom)
```

```
##   df chi_sq     p_value
## 1  1   4.12 0.042379078
## 2  4   1.02 0.906748324
## 3  2   9.50 0.008651695
## 4 10  12.40 0.259177369
## 5  1   2.48 0.115302253
```

3. Soccer reaches its apex every four years at the World Cup, attracting worldwide attention and fanatic devotion. The World Cup is widely thought to be the event that decides the best soccer team in the world. But how much do skill differences determine the outcome? If the probability of a goal is the same for all teams and games, and if goals are independent, then we would expect the frequency distribution of goals per game to approximate a Poisson distribution. In contrast, if skill differences really matter, then we would expect more high scores and more low scores than predicted from the Poisson distribution. The following table tallies the number of goals scored by one of the two teams (randomly chosen) in every game of the knockout round of the World Cup from 1986 through 2010.

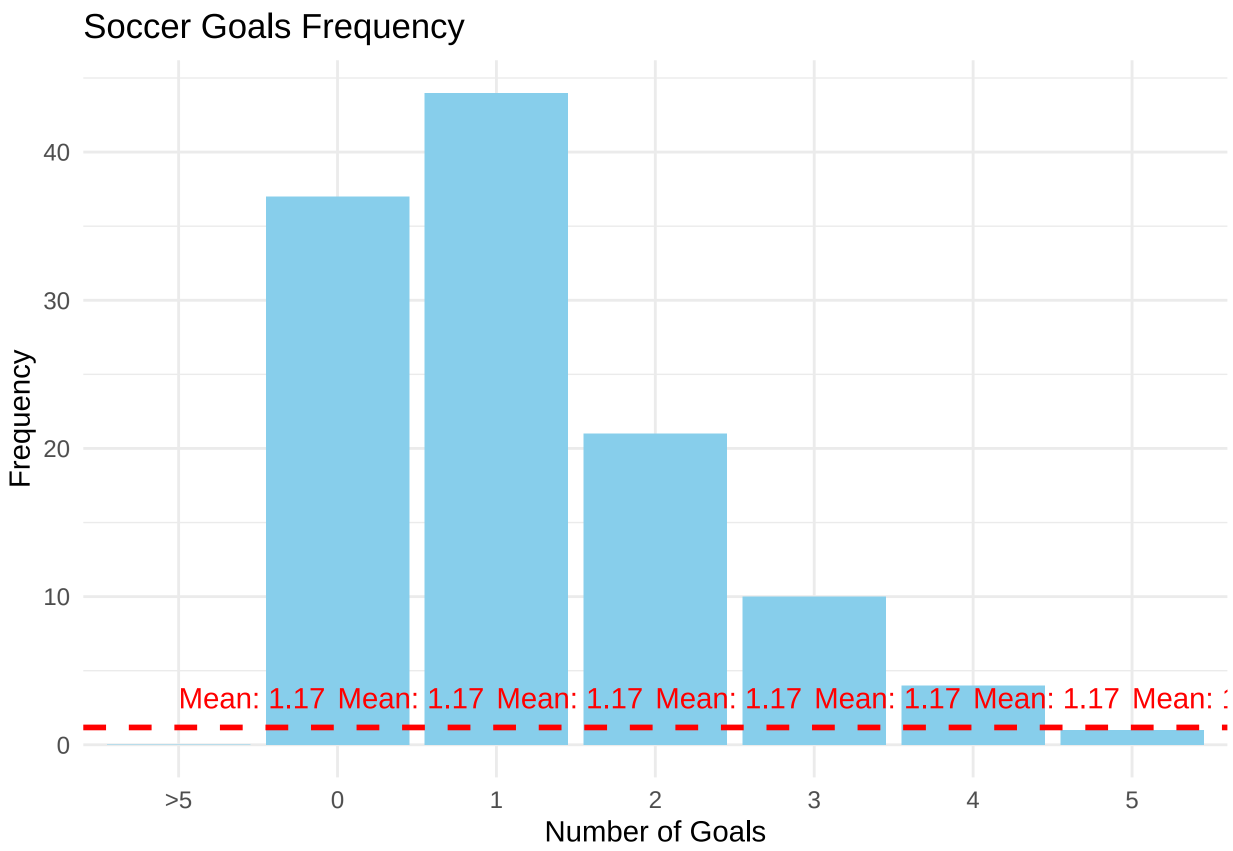
```
column1 <- c("0","1","2","3","4","5",">5")
column2 <- c(37,44,21,10,4,1,0)

soccer <- data.frame(goals = column1, frequency = column2)

column1_2 <- c(0,1,2,3,4,5,5)
column2_2 <- c(37,44,21,10,4,1,0)

soccer_2 <- data.frame(goals = column1_2, frequency = column2_2)
```

a. Plot the frequency distribution of goals per team using the data in the table.

b. What is the mean number of goals per game?

The mean number of goals is 1.17 goals.

```
# Calculate the mean (excluding "total")
mean_value <- sum(soccer_2$goals * soccer_2$frequency) / sum(soccer_2$frequency)

# Create the plot
soccer2_plot <- ggplot(soccer, aes(x = goals, y = frequency)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_hline(yintercept = mean_value, color = "red", linetype = "dashed", size = 1) +
  geom_text(aes(label = paste("Mean:", round(mean_value, 2))),
             y = mean_value + 2, color = "red", size = 4, hjust = 0) +
  labs(title = "Soccer Goals Frequency", x = "Number of Goals", y = "Frequency") +
  theme_minimal()
soccer2_plot
```



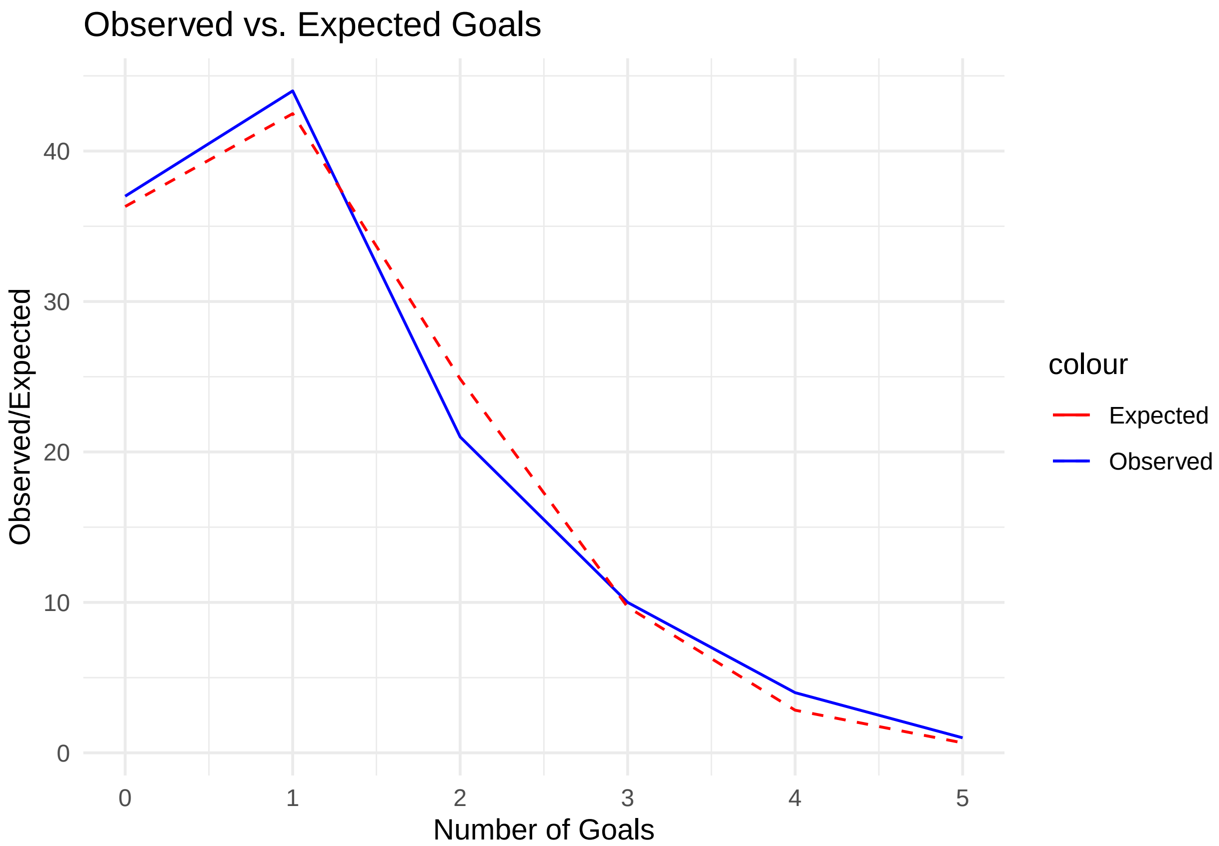
c. Using the Poisson distribution, calculate the expected frequency of games and teams with 0, 1, 2, ….,5 goals, assuming independence and equal probability of scoring. The expected frequencies of goals 0-5 are respectively 36.31, 42.49, 24.85, 9.69, 2.83, 0.66.

```
column1 <- c(1, 2, 3, 4, 5, "<5")
column2 <- c(37, 44, 21, 10, 4, sum(c(37, 44, 21, 10, 4)))

soccer <- data.frame(goals = column1, observed = column2)

total_games <- sum(soccer$observed)



lambda <- 1.17

## Updated code

column1_3 <- c(0,1,2,3,4,5)
column2_3 <- c(37,44,21,10,4,1)
soccer3 <- data.frame(goals = column1_3, observed = column2_3, probs = NA, expec = NA)

prob <- dpois(soccer3$goals, lambda)
total_freq <- sum(soccer3$observed)

soccer3<- soccer3 %>%
  mutate(probs = round((prob),7),
         expec = probs * total_freq)
soccer3
```

```
##   goals observed     probs      expec
## 1     0       37 0.3103669 36.3129273
## 2     1       44 0.3631293 42.4861281
## 3     2       21 0.2124307 24.8543919
## 4     3       10 0.0828480  9.6932160
## 5     4        4 0.0242330  2.8352610
## 6     5        1 0.0056705  0.6634485
```

d. Overlay the expected frequencies calculated in part c on the graph you created in part a. Do they appear similar? The plots are extremely similar.

```
ggplot(soccer3, aes(x = goals)) +
  geom_line(aes(y = observed, color = "Observed")) +
  geom_line(aes(y = expec, color = "Expected"), linetype = "dashed") +
  labs(title = "Observed vs. Expected Goals",
       x = "Number of Goals",
       y = "Observed/Expected") +
  scale_color_manual(values = c("Observed" = "blue", "Expected" = "red")) +
  theme_minimal()
```



e. If skill differences do not matter, would you expect the variance in the number of goals per team and side to be less than, equal to, or greater than the mean number of goals? Calculate the variance in the number of goals per team and side. How similar is this to the mean? If skill differences do not matter, then we'd expect the mean and variance to be equal. This is because a random outcome would resemble a poisson distribution where the mean and variance are equal. The calculated variance is 1.26 goals squared which is close to the mean of 1.17 goals, but slightly greater than it.

```
var <- (sum(((soccer3$goals-lambda)^2)*soccer3$frequency))/116
#1.26
```

4. One thousand coins were each flipped eight times, and the number of heads was recorded for each coin. The results are as follows:

```
column1 <- c(0,1,2,3,4,5,6,7,8)
column2 <- c(6,32,105,186,236,201,98,33,103)

coin_flips <- data.frame(heads = column1, coins = column2)

coin_flips$expected_prob <- (choose(8, coin_flips$heads)*(0.5^8))

coin_flips$expected_coins <- (choose(8, coin_flips$heads)*(0.5^8)*1000)

#coin_flips$observed_prob <- coin_flips$coins/1000
```

a. . Test whether the distribution of coin flips matches the expected frequencies from a binomial distribution assuming all fair coins. The Chi-squared test statistic is 2527.9, and with 8 degrees of freedom, the p-value is less than 0.05. This indicates that the distribution does not match a binomial distribution.

```
chisq.test(coin_flips$coins, coin_flips$expected_prob)
```

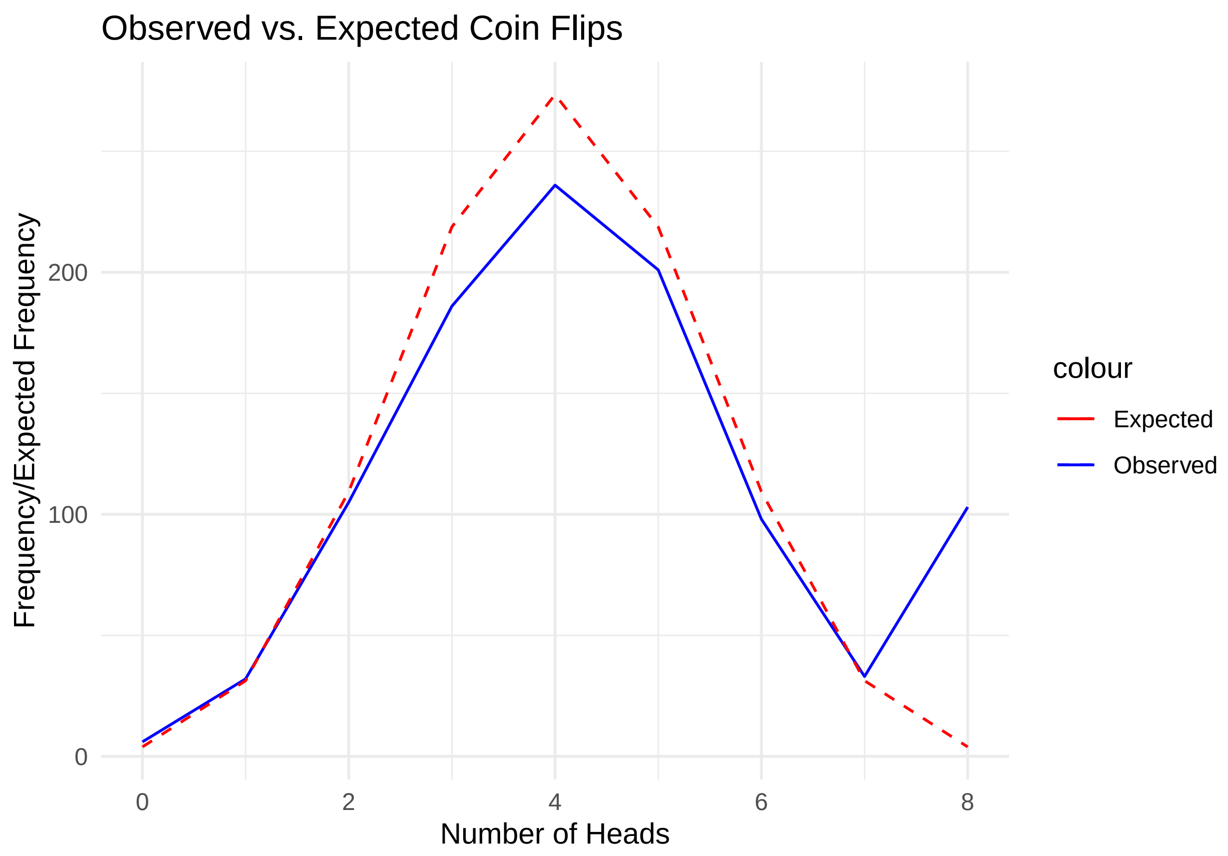
```
##
##  Pearson's Chi-squared test
##
## data:  coin_flips$coins and coin_flips$expected_prob
## X-squared = 36, df = 32, p-value = 0.2867
```

```
chi_sq <- sum(((coin_flips$coins - coin_flips$expected_prob)^2)/(coin_flips$expected_prob))

chisq.test(x = coin_flips$coins, p = coin_flips$expected_prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  coin_flips$coins
## X-squared = 2527.9, df = 8, p-value < 2.2e-16
```

b. . If the binomial distribution is a poor fit to the data, identify in what way the distribution does not match the expectation. The data shows a larger number of heads flipped than the expected binomial distribution. This could potentially indicate a trick coin.

```
ggplot(coin_flips, aes(x = heads)) +
  geom_line(aes(y = coins, color = "Observed")) +
  geom_line(aes(y = expected_prob * 1000, color = "Expected"), linetype = "dashed") +
  labs(title = "Observed vs. Expected Coin Flips",
       x = "Number of Heads",
       y = "Frequency/Expected Frequency") +
  scale_color_manual(values = c("Observed" = "blue", "Expected" = "red")) +
  theme_minimal()
```



5. Hurricanes hit the United States often and hard, causing some loss of life and enormous economic costs. They are ranked in severity by the Saffir-Simpson scale, which ranges from Category 1 to Category 5, with 5 being the worst. In some years, as many as three hurricanes that rate a Category 3 or higher hit the US coastline. In other years, no hurricane of this severity hits the United States. The following table lists the number of years that had 0, 1, 2, 3, or more hurricanes of at least Category 3 in severity, over the 100 years of the 20th century (Blake et al., 2005).

```
column1 <- c(0,1,2,3,">3")
column2 <- c(50,39,7,4,0)

hurricanes <- data.frame(count = column1, years = column2)

column1 <- c(0,1,2,3)
column2 <- c(50,39,7,4)

hurricanes2 <- data.frame(count = column1, years = column2)
```

a. What is the mean number of severe hurricanes to hit the United States per year? The mean number of severe hurricanes to hit the US per year is 0.65 hurricanes.

```
mean_hurricanes <- (sum(as.numeric(hurricanes2$count*hurricanes$years)))/sum(hurricanes2$years)
#0.65 hurricanes per year
```

b. What model would describe the distribution of hurricanes per year, if they were to hit independently of each other and if the probability of a hurricane were the same every year? A poisson model describes a distribution where sample events are independent of each other and the probability of events are the same every year.

```
# Given data
column1 <- c(0, 1, 2, 3)
column2 <- c(50, 39, 7, 4)

hurricanes2 <- data.frame(count = column1, years = column2)

# Calculate the mean/lambda
lambda <- 0.65

# Calculate the expected frequencies assuming a Poisson distribution
expected_freq <- dpois(hurricanes2$count, lambda)
expected_freq
```

c. Test the fit of the model for part (b) to the data.

```
## [1] 0.52204578 0.33932975 0.11028217 0.02389447
```

```
1-(0.52204578+0.33932975+0.11028217)
```

```
## [1] 0.0283423
```

```
hurricanes2$expec <- c(0.52204578, 0.33932975, 0.11028217, 0.0283423)

# Perform chi-squared goodness-of-fit test
chi_sq_test <- chisq.test(x = hurricanes2$years, p = hurricanes2$expec)

# Print the results
print(chi_sq_test)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  hurricanes2$years
## X-squared = 2.8006, df = 3, p-value = 0.4234
```

The p-value of our goodness-of-fit test is not less than 0.05, so we fail to reject the null. This confirms that this is a poisson distribution.

6. Do people typically use a particular ear preferentially when listening to strangers? Marzoli and Tomassi (2009) had a researcher approach and speak to strangers in a noisy nightclub. An observer scored whether the person approached turned either the left or the right ear toward the questioner. Of 25 participants, 19 turned the right ear toward the questioner and 6 offered the left ear. Is this evidence of population difference from 50% for each ear? Use the following steps to help answer this question with a binomial test. Consider that the assumptions of the binomial test are met in this study.

a. State the null and alternative hypotheses. Null hypothesis: 50% of people will prefer the use of their right ear when listening to strangers.

Alternative hypothesis: A percentage different than 50% of people will prefer the use of their right ear.

b. What is the observed value of the test statistic? The observed value of the test statistic is 19 people.

c. Under the null hypothesis, calculate the probability of getting exactly 19 right ears and six left ears.** 0.0053

```
choose(25,19)*(0.5^25)
```

```
## [1] 0.005277991
```

d. List all possible outcomes in which the number of right ears is greater than the 19 observed. 20, 21, 22, 23, 24, 25 #### e. Calculate the probability under the null hypothesis of each of the extreme outcomes listed in (d). 0.001583397, 0.0003769994, 6.854534e-05, 8.940697e-06, 7.450581e-07, 2.980232e-08

```
choose(25, 20)*(0.5^25)
```

```
## [1] 0.001583397
```

```
choose(25, 21)*(0.5^25)
```

```
## [1] 0.0003769994
```

```
choose(25, 22)*(0.5^25)
```

```
## [1] 6.854534e-05
```

```
choose(25, 23)*(0.5^25)
```

```
## [1] 8.940697e-06
```

```
choose(25, 24)*(0.5^25)
```

```
## [1] 7.450581e-07
```

```
choose(25, 25)*(0.5^25)
```

```
## [1] 2.980232e-08
```

f. Calculate the probability of 19 or more right-eared turns under the null hypothesis. The probability of 19 or more right-eared turns is 0.0073.

```
0.001583397+0.0003769994+6.854534e-05+8.940697e-06+7.450581e-07+2.980232e-08+0.0053
```

```
## [1] 0.007338657
```

g. Give the two-tailed P-value based on your answer to (f). The two-tailed P-value is 0.01467731.

```
0.007338657*2
```

```
## [1] 0.01467731
```

h. Interpret this P-value. What does it indicate? The p-value less than our alpha, indicating that we reject the null in favor of the alternative. It's likely that the percentage of people who use their right ear to listen is not 0.5. #### i. State the conclusion from your test. It's likely that the percentage of people who use their right ear to listen is not 0.5.

7. Assume that a null hypothesis is true. Which one of the following statements is true?

a. A study with a larger sample is more likely than a smaller study to get the result that P < 0.05.

b. A study with a larger sample is less likely than a smaller study to get the result that $P < 0.05$.

c. A study with larger sample is equally likely compared to a smaller study to get the result that $P < 0.05$. Statement b is true.

8. Assume a random sample. What effect does increasing the sample size have on:

a. The probability of committing a Type I error? Increasing the sample size does not affect the probability of committing a type 1 error. #### b. The probability of committing a Type II error? Increasing the sample size decreases the probability of committing a type 2 error. #### c. The significance level? The significance level is independent of sample size, so it is not affected.

9. In the toad experiment we looked at in the class, what would the P-value have been if

a. 15 toads were right-handed and the rest were left-handed? P-value: 0.007538

```r
prob <- 0.5
trials <- 18

binom.test(15, 18, p = 0.5, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  15 and 18
## number of successes = 15, number of trials = 18, p-value = 0.007538
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5858225 0.9642149
## sample estimates:
## probability of success
##              0.8333333
```

b. 13 toads were right-handed and the rest were left-handed? P-value: 0.09625

```r
binom.test(13, 18, p = 0.5, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  13 and 18
## number of successes = 13, number of trials = 18, p-value = 0.09625
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4651980 0.9030508
## sample estimates:
## probability of success
##              0.7222222
```

c. 10 toads were right-handed and the rest were left-handed? P-value: 0.8145

```r
binom.test(10, 18, p = 0.5, alternative = "two.sided")
```

```
##
##  Exact binomial test
```

```
##
## data:  10 and 18
## number of successes = 10, number of trials = 18, p-value = 0.8145
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3075717 0.7846985
## sample estimates:
## probability of success
##              0.5555556
```

d. 7 toads were right-handed and the rest were left-handed? P-value: 0.4807

```
binom.test(7, 18, p = 0.5, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  7 and 18
## number of successes = 7, number of trials = 18, p-value = 0.4807
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1729859 0.6425488
## sample estimates:
## probability of success
##              0.3888889
```