Math 138, Section 8.3 – Pushing Dice Off a Ledge

Chi-square Goodness of Fit Test

The chi-square goodness-of-fit test is a theory-based test of significance that compares observed data to a prescribed model. In this course, you will use goodness-of-fit tests to compare the distribution of a categorical variable to a hypothesized distribution. For example, in Chapter 1, you conducted tests to evaluate questions like “When playing the game rock- paper-scissors (see Example 1.2), do novice players play scissors less than one-third of the time in the long run?” These tests evaluated whether sample data provided evidence that the probabilities associated with a categorical variable with two categories (scissors or not scissors) differed from hypothesized probabilities (π and 1 – π). However, at that time you were limited to only evaluating categorical variables with two categories. As you’ve been exploring in Chapter 8, it is often the case that categorical variables have more than two categories. So, for example, maybe you want to test whether the way novice players make choices when playing the game rock-paper-scissors is such that all three options are chosen equally often in the long run (1/3, 1/3, 1/3). Similarly, we could ask: Are birthdays equally distributed across the seven days of the week? Do certain pea plants produce three times as many purple flowers as white flowers? We’ll see how to evaluate these types of questions in this section.

Fair Die roll?

A statistics student wondered whether rolling six-sided dice by pushing them off a 2-inch ledge was a fair way of rolling dice.

In this case, for a die roll to be fair, it means that all six sides are equally likely to occur. In other words, if we were to repeatedly roll the die by pushing it off a 2-inch high ledge, then we would expect each of the 6 numbered faces of the die to appear on top an equal number of times in the long run. If we rolled the die 120 times, we would expect to see each of the 6 different numbers rolled about 20 times. After rolling the die 120 times, if we observe our data to have deviated substantially from what we expected to see from a fair die (~20 times each), we may have evidence that our rolling method is not “fair.”

$$H_0 : \text{The method of rolling is fair. } \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = \frac{1}{6}.$$

H_a : The method of rolling is not fair -- at least one face is likely to appear more than 1/6 of the time.

where π_i is the long-run proportion of rolling the number i, for $i = 1, 2, 3, 4, 5, 6$.

Notice how these hypotheses looks familiar — testing the equality of several probabilities — but a key difference is that we are now specifying a specific numerical value for each π_i (and those values must sum to 1).

Load packages and data

```
#add the other package that we will need
library(ggformula)
library(mosaic)
```

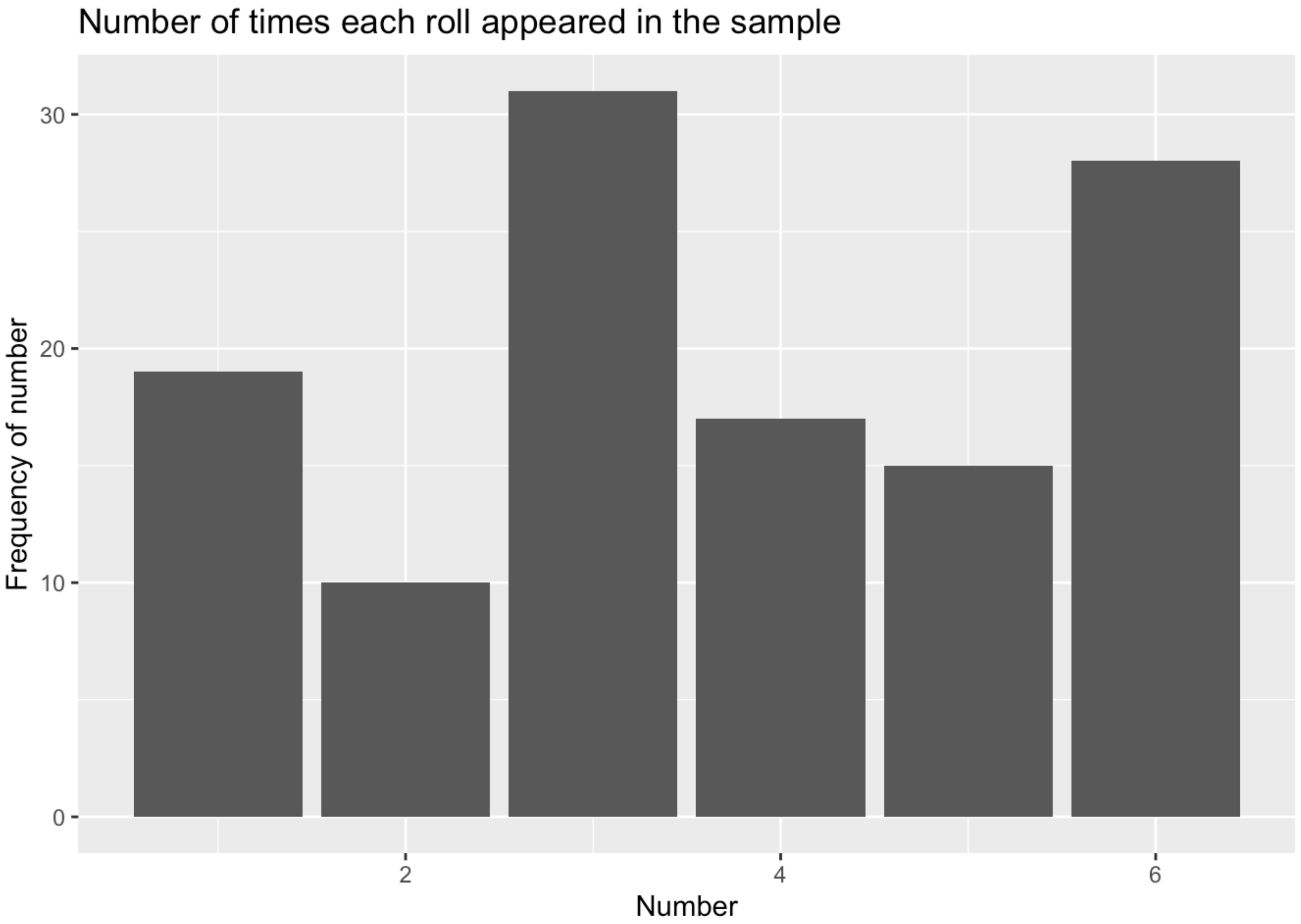
We'll load the data, `Die.csv` , available at this Url: https://raw.githubusercontent.com/IJohnson-math/Math138/main/Die.csv (https://raw.githubusercontent.com/IJohnson-math/Math138/main/Die.csv). We'll use the `read.csv()` function to read in the data.

```
DieData <- read.csv("https://raw.githubusercontent.com/IJohnson-math/Math138/main/Die
        .csv")
```

```
tally(~DieRoll, data=DieData)
```

```
## DieRoll
##  1  2  3  4  5  6
## 19 10 31 17 15 28
```

```
gf_bar(~DieRoll, data=DieData, xlab="Number", ylab="Frequency of number", title = "Nu
        mber of times each roll appeared in the sample")
```



```
tally(~DieRoll, data=DieData)
```

```
## DieRoll
##  1  2  3  4  5  6 
## 19 10 31 17 15 28
```

Calculating the Mean Average Distance, MAD, the observed counts are away from the hypothesized count of 20.

```
MAD = (abs(19-20)+abs(10-20)+abs(31-20)+abs(17-20)+abs(15-20)+abs(28-20))/6
MAD
```

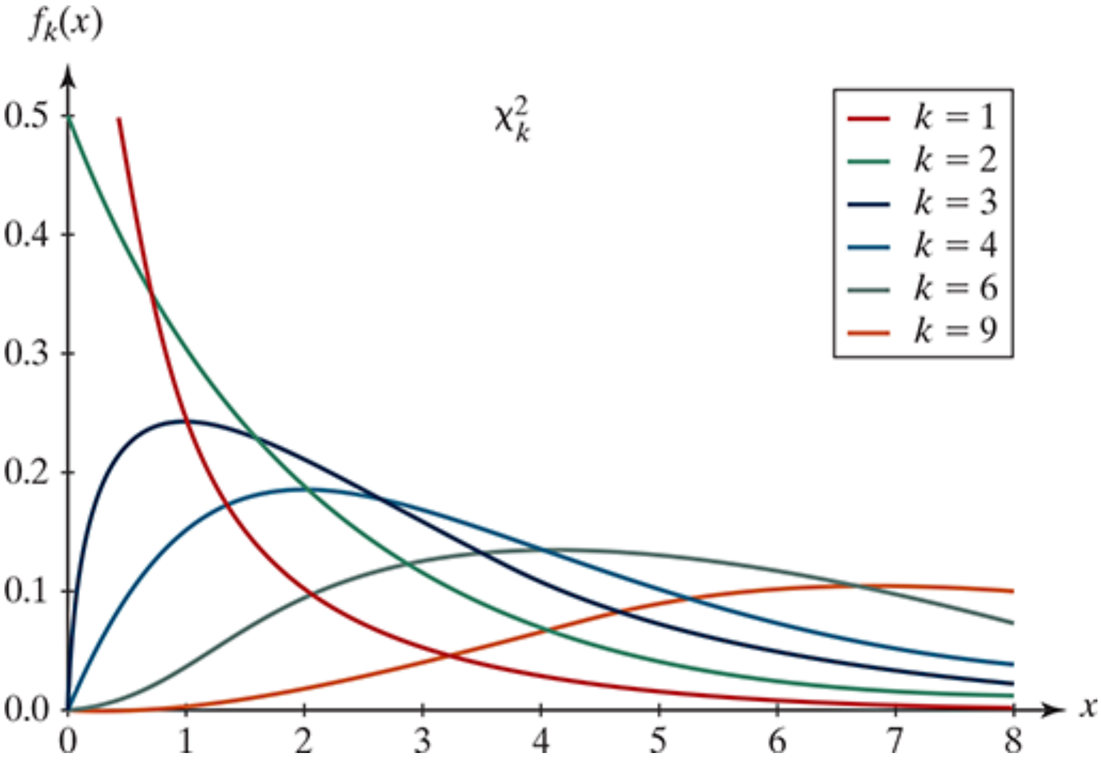
```
## [1] 6.333333
```

Our observed frequencies of the numbers rolled are on average 6.33 units away from expected values.

Go to the Multiple Proportions (https://www.isi-stats.com/isi2nd/ISIapplets2021.html) simulation-based applet to complete the analysis using the MAD statistic.

p-value from Simulation using MAD statistic: 0.005 (25/5000)

p-value from Simulation using χ^2 statistic: 0.0076 (38/5000)



The degrees of freedom, denoted here by k, are computed by multiplying the number of categories in the explanatory variable minus 1 by the number of categories in the response variable minus 1.

Chi-square calculation

$$\chi^2 = \Sigma \frac{(\text{observed count } - \text{ expected count})^2}{\text{expected count}}$$

```
#expected count value is 20
chiSquare <- ((19-20)^2)/20 + ((10-20)^2)/20 + ((31-20)^2)/20 + ((17-20)^2)/20 + ((15
        -20)^2)/20 + ((28-20)^2)/20
chiSquare
```

```
## [1] 16
```

Chi-square goodness of fit test

```
chisq.test(tally(~DieRoll, data = DieData), p = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)) # Go
        odnessFit
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tally(~DieRoll, data = DieData)
## X-squared = 16, df = 5, p-value = 0.006844
```

```
chisq.test(c(19, 10, 31, 17, 15, 28), p = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(19, 10, 31, 17, 15, 28)
## X-squared = 16, df = 5, p-value = 0.006844
```

Conclusion: Our theory-based p-value 0.0068 is very similar to both of our simulation-based p-values. From this we can conclude we have strong evidence that pushing dice off a 2-inch ledge is not a fair way of tossing dice. We will reject the null hypothesis that with this rolling technique all numbers 1, 2, 3, 4, 5, and 6 are equally likely to occur.

The study design is fairly well controlled because the same ledge was used for each trial, the same lineup of dice was used for each trial, and the same person did the pushing of the dice each trial. Thus, it seems reasonable to believe that the method of dice rolling, pushing them off a 2-inch ledge, is the cause of the faces not being rolled equally likely. We weren’t, however, told whether our five dice were a random sample of all dice. It is plausible that they were taken from a generic board game. Then we could argue they are representative of all game board dice of the same size and thus say that this method of dice rolling is not fair for all game board dice of this size. We need more information, though, to extend our conclusions beyond the five dice and 2-inch height used. Additionally, we would need the assurance that these five dice weren’t loaded prior to conducting our study!

Validity Conditions

As in our other theory-based tests, this one comes with validity conditions as well. The validity conditions for a chi-square goodness-of-fit test are that all observed counts are at least 10. Since the values 19, 10, 31, 17, 15, 28 are all larger than 10 the validity conditions have been met.

Post-hoc tests

Our evidence supports the conclusion that at least one number is likely to appear more than 1/6 of the time. Here are a few of the post hoc tests and what the results mean.

```
#post-hoc tests
chisq.test(c(28, 10), p = c(1/2, 1/2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(28, 10)
## X-squared = 8.5263, df = 1, p-value = 0.0035
```

With a p-value of 0.0035 we have very strong evidence against the null. We will reject as highly improbable the null hypothesis that $\pi_2 = \pi_6$. Our evidence supports the alternative hypothesis that $\pi_2 - \pi_6 \neq 0$.

```
chisq.test(c(31, 15), p = c(1/2, 1/2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(31, 15)
## X-squared = 5.5652, df = 1, p-value = 0.01832
```

With a p-value of 0.01832 we have strong evidence against the null. We will reject the null hypothesis that $\pi_3 = \pi_5$. Our evidence supports that probability of “rolling” a 3 is not equal to the probability of rolling a 5.

```
chisq.test(c(10, 15), p = c(1/2, 1/2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(10, 15)
## X-squared = 1, df = 1, p-value = 0.3173
```

A p-value of 0.3173 leads us to conclude that the null hypothesis is plausible. That is, it is plausible that the probability of “rolling” a 2 using the technique is the same as the probability of “rolling” a 5.

```
chisq.test(c(31, 28), p = c(1/2, 1/2))
```

```
## 
##  Chi-squared test for given probabilities
## 
## data:  c(31, 28)
## X-squared = 0.15254, df = 1, p-value = 0.6961
```

A p-value of 0.6961 leads us to conclude that the null hypothesis is plausible. That is, it is plausible that the probability of “rolling” a 3 using the technique is the same as the probability of “rolling” a 6.