Chapter 3

October 25, 2019

R by Example

Alex Chi

```
In [43]: library(repr)
         options(repr.plot.width=10, repr.plot.height=8)
```

Chapter 3 Exercises

3.1 (Fast food eating preference).

Fifteen students in a statistics class were asked to state their preference among the three restaurtants Wendys, McDonalds, and Subway. The responses for the students are presented below.

```
Wendys McDonalds Subway Subway Subway Wendys
Wendys Subway Wendys Subway Subway Subway
Subway Subway Subway
```

a. Use the scan function to read these data into the R command window.

b. Use the table function to find the frequencies of students who prefer the three restaurants.

c. Compute the proportions of students in each category.

d. Construct two different graphical displays of the proportions.

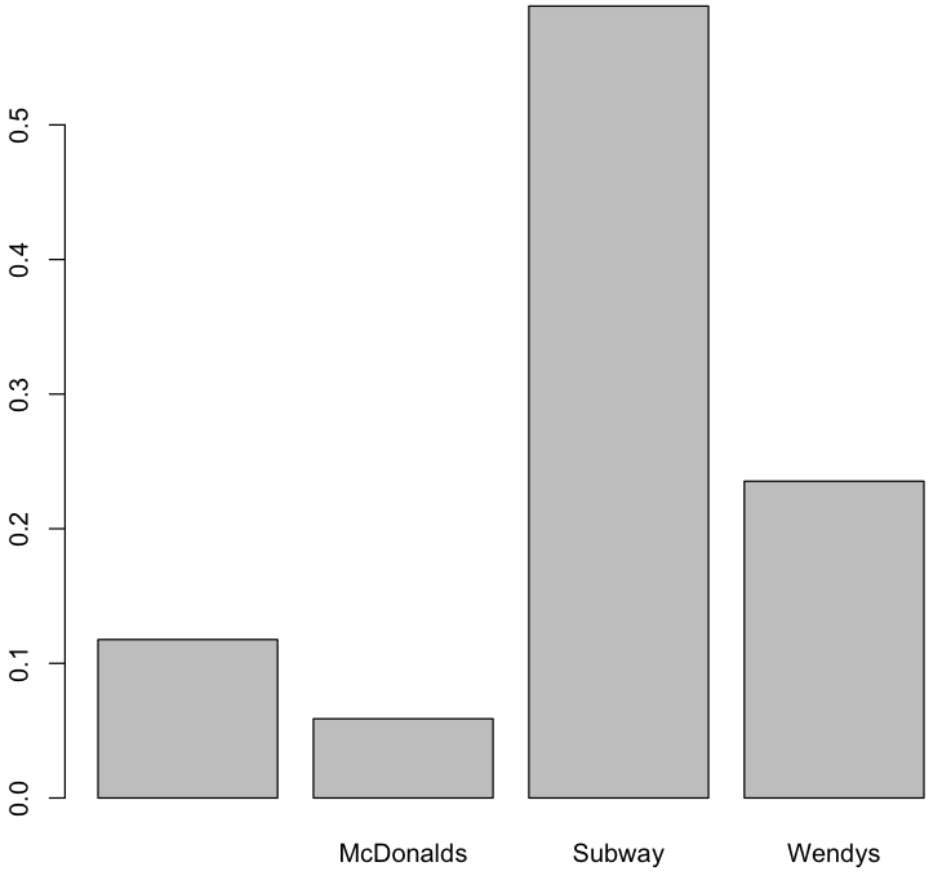
```
In [11]: # scan cannot be used in Jupyter R, so I just construct the table
         restaurants = unlist(strsplit(
             paste("Wendys McDonalds Subway Subway Subway ",
                   "Wendys Wendys Subway Wendys Subway ",
                   "Subway Subway Subway Subway Subway")
             , split=' '))
         freq = table(factor(restaurants))
         freq
```

```
          McDonalds     Subway     Wendys
        2          1         10          4
```

In [12]:
```
freq / length(restaurants)
```

```
            McDonalds     Subway     Wendys
0.11764706 0.05882353 0.58823529 0.23529412
```

In [13]:
```
barplot(prop.table(freq))
```



3.2 (Dice rolls)

Suppose you roll a pair of dice 1000 times a. One can simulate 1000 rolls of a fair die using the R function sample(6, 1000, replace=TRUE). Using this function twice, store 1000 simulated rolls of the first die in the variable die1 and 1000 simulated rolls of the second die in the variable die2.

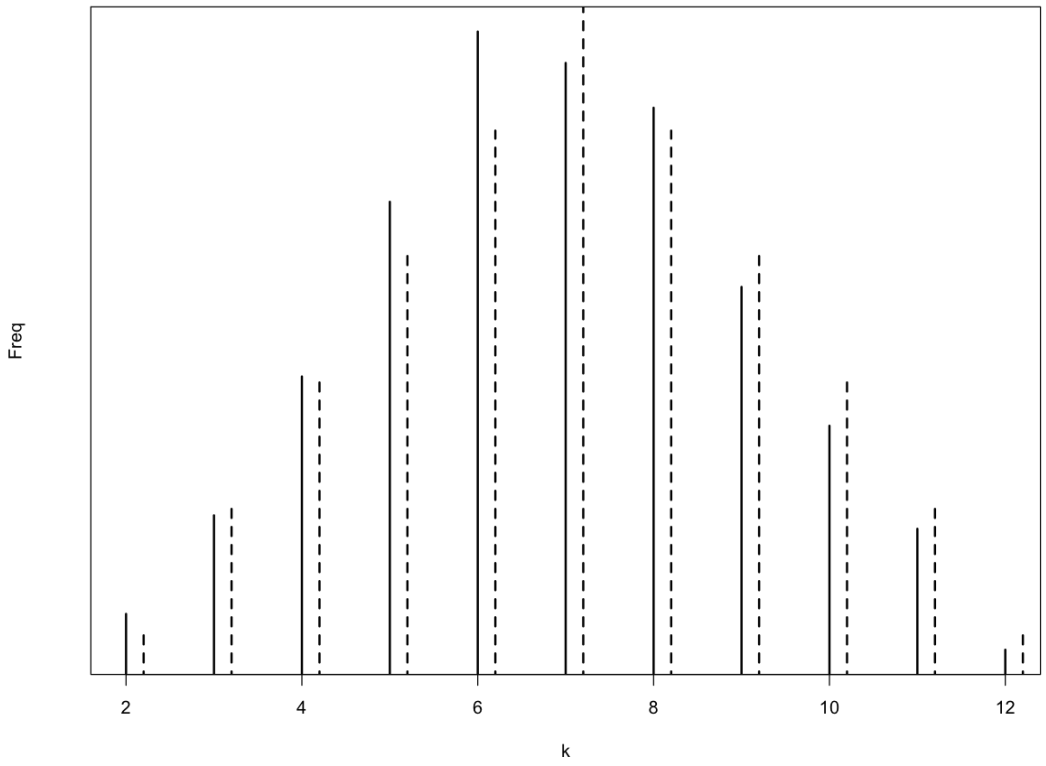
b. For each pair of rolls, compute the sum of rolls, and store the sums in the variable die.sum.

c. Use the table function to tabulate the values of the sum of die rolls. Com- pute the proportions for each sum value and compare these proportions with the exact probabilities of the sum of two die rolls

```
In [33]: die1 = sample(6, 1000, replace=TRUE)
         die2 = sample(6, 1000, replace=TRUE)
         die.sum = die1 + die2
         table(die.sum) / length(die.sum)
```

```
die.sum
    2     3     4     5     6     7     8     9    10    11    12
0.031 0.053 0.084 0.123 0.161 0.154 0.144 0.104 0.073 0.050 0.023
```

```
In [49]: k = 2:12
         die.expected = (6 - abs(k - (6 + 1))) / (6 * 6)
         die.freq = table(die.sum) / length(die.sum)
         plot(k, die.freq, type="h", lwd=2, lty=1, ylab="Freq")
         lines(k + .2, die.expected, type="h", lwd=2, lty=2)
```



3.3 (Does baseball hitting data follow a binomial distribution?)

Albert Pujols is a baseball player who has n opportunities to hit in a single game. If y denotes the number of hits for a game, then it is reasonable to assume that y has a binomial distribution with sample size n and probability of success p = 0.312, where 0.312 is Pujols' batting average (success rate) for the 2010 baseball season.

a. In 70 games Pujols had exactly n = 4 opportunities to hit and the num- ber of hits y in these 70 games is tabulated in the following table. Use the dbinom function to compute the expected counts and the chisq.test function to test if the counts follow a binomial(4, 0.312) distribution.

b. In 25 games Pujols had exactly n = 5 opportunities to hit and the number of hits y in these 25 games is shown in the table below. Use the chisq.test function to test if the counts follow a binomial(5, 0.312) distribution

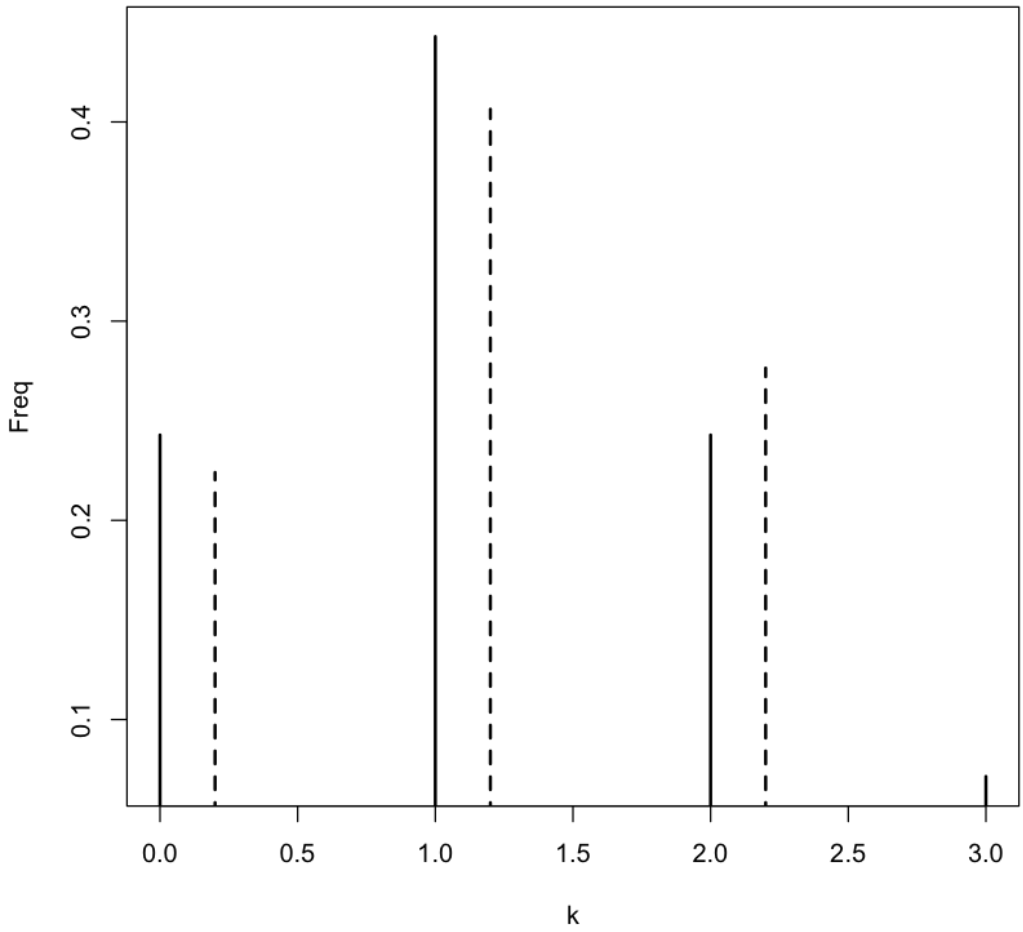
在这里，我在 chisq.test 中使用了 `simulate.p.value` 参数，以防止 "Chi-squared approximation may be incorrect" 的提示。

```
In [154]: n = 4
          k = 0 : (n - 1)
```

```r
p = dbinom(k, n, 0.312)
baseball.expected = c(p[1:3], 1-sum(p[1:3]))
baseball.game = c(17, 31, 17, 5)

plot(k, prop.table(baseball.game), type="h",
     lwd=2, lty=1, ylab="Freq")
lines(k + .2, baseball.expected, type="h",
      lwd=2, lty=2)

result = chisq.test(baseball.game,
                    p=baseball.expected,
                    simulate.p.value=TRUE)
result
cat("p-value = ", result$p.value, ", p-value 太小代表不服从分布")
```

```
Chi-squared test for given probabilities with simulated p-value (based
on 2000 replicates)

data:  baseball.game
X-squared = 0.97692, df = NA, p-value = 0.8241
```

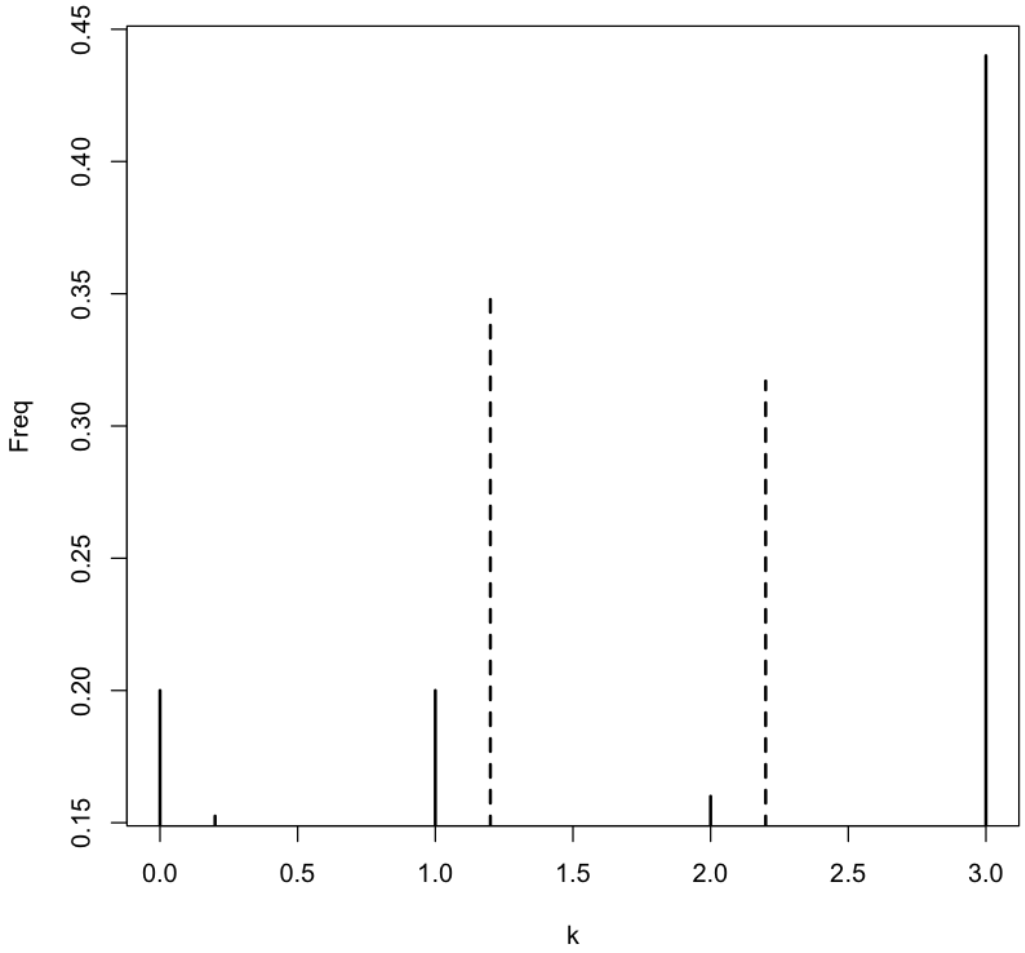
```
p-value =  0.824088 , p-value 太小代表不服从分布
```



In [155]:
```
p = dbinom(0:4, 5, 0.312)
baseball.expected = c(p[1:3], 1 - sum(p[1:3]))
baseball.game = c(5, 5, 4, 11)

plot(k, prop.table(baseball.game), type="h",
     lwd=2, lty=1, ylab="Freq")
lines(k + .2, baseball.expected, type="h",
      lwd=2, lty=2)

result = chisq.test(baseball.game,
                    p = baseball.expected,
                    simulate.p.value=TRUE)
result
cat("p-value = ", result$p.value, ", p-value 太小代表不服从分布")
```

```
Chi-squared test for given probabilities with simulated p-value (based
on 2000 replicates)

data:  baseball.game
X-squared = 13.359, df = NA, p-value = 0.002499
```

```
p-value =  0.002498751 , p-value 太小代表不服从分布
```



3.4 (Categorizing ages in the twins dataset).

The variable AGE gives the age (in years) of twin 1.

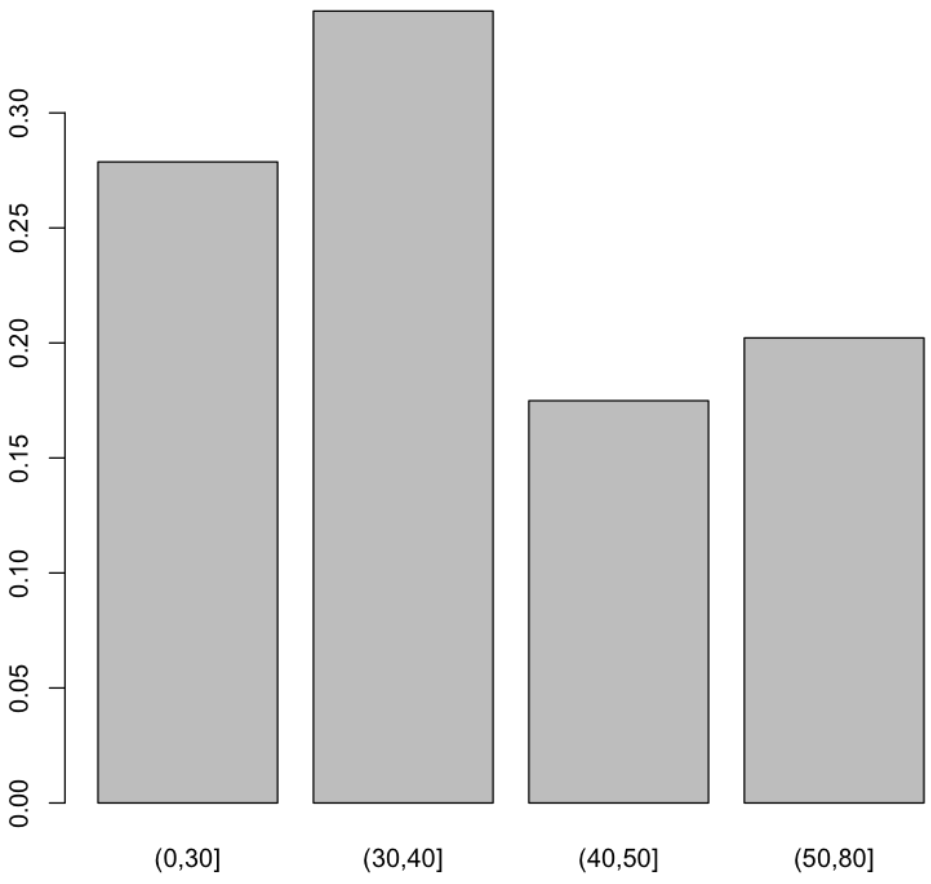
a. Use the cut function on AGE with the breakpoints 30, 40, and 50 to create a categorized version of the twin's age.

b. Use the table function to find the frequencies in the four age categories.

c. Construct a graph of the proportions in the four age categories

```
In [112]: TWINS_DATASET_URL = "http://personal.bgsu.edu/~mrizzo/Rx/Rx-data/twins
          twn = read.table(TWINS_DATASET_URL, header=TRUE,
                           na.strings=".", sep=",")
```

```
In [127]: cut_age = cut(twn$AGE, breaks=c(0, 30, 40, 50, 80))
          freq = table(cut_age)
          barplot(prop.table(freq))
```



3.5 (Relating age and wage in the twins dataset).

The variables AGE and HRWAGEL contain the age (in years) and hourly wage (in dollars) of twin 1.

a. Using two applications of the cut function, create a categorized version of AGE using the breakpoints 30, 40, and 50, and a categorized version of HRWAGEL using the same breakpoints as in Section 3.3.

b. Using the categorized versions of AGE and HRWAGEL, construct a contin- gency table of the two variables using the function table.

c. Use the prop.table function to find the proportions of twins in each age class that have the different wage groups.

d. Construct a suitable graph to show how the wage distribution depends on the age of the twin.

e. Use the conditional proportions in part (c) and the graph in part (d) to explain the relationship between age and wage of the twins.

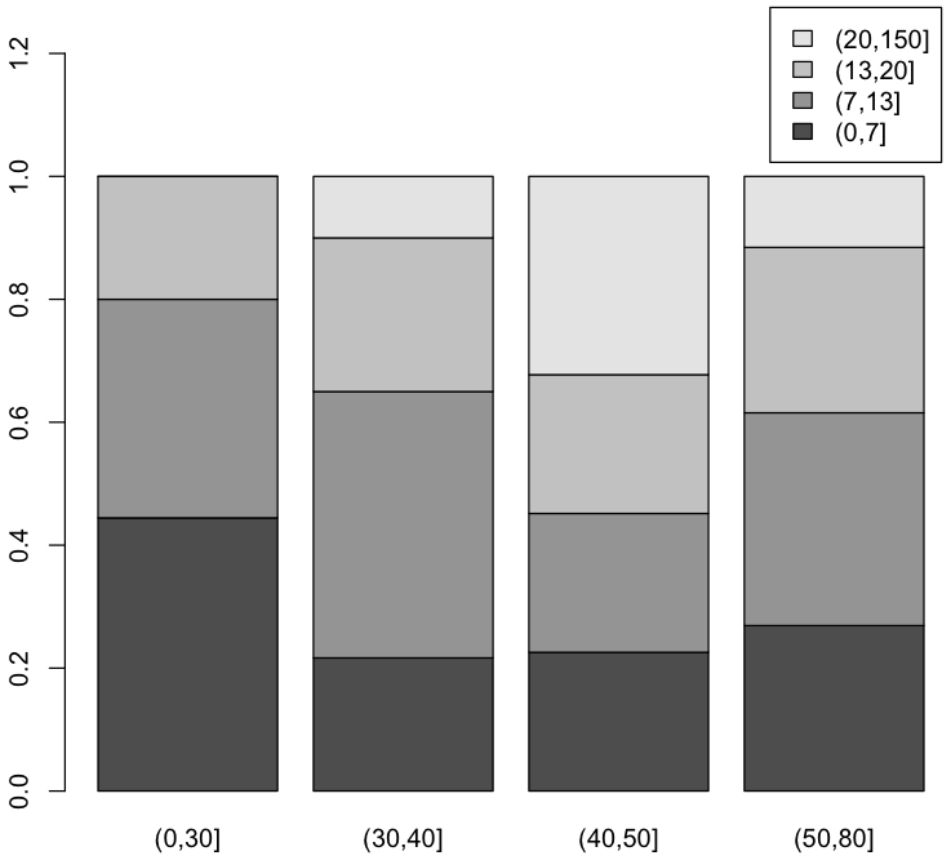
```
In [156]: cut_age = cut(twn$AGE, c(0, 30, 40, 50, 80))
          cut_hrwagel = cut(twn$HRWAGEL, c(0, 7, 13, 20, 150))
          ct_table = table(cut_age, cut_hrwagel)
          ct_table
```

```
          cut_hrwagel
cut_age    (0,7] (7,13] (13,20] (20,150]
  (0,30]      20     16       9        0
  (30,40]     13     26      15        6
  (40,50]      7      7       7       10
  (50,80]      7      9       7        3
```

```
In [157]: prop_table = prop.table(ct_table, margin=1)
          prop_table
```

```
          cut_hrwagel
cut_age        (0,7]    (7,13]   (13,20]  (20,150]
  (0,30]  0.4444444 0.3555556 0.2000000 0.0000000
  (30,40] 0.2166667 0.4333333 0.2500000 0.1000000
  (40,50] 0.2258065 0.2258065 0.2258065 0.3225806
  (50,80] 0.2692308 0.3461538 0.2692308 0.1153846
```

```
In [158]: barplot(t(prop_table),
                  ylim=c(0, 1.3),
                  legend.text=dimnames(prop_table)$cut_hrwagel)
```



3.6 (Relating age and wage in the twins dataset, continued).

a. Using the contingency table of the categorized version of AGE and HRWAGEL and the function chisq.test, perform a test of independence of age and wage. Based on this test, is there significant evidence to conclude that age and wage are dependent?

b. Compute and display the Pearson residuals from the test of independence. Find the residuals that exceed 2 in absolute value.

c. Use the function mosaicplot with the argument shade=TRUE to construct a mosaic plot of the table counts showing the extreme residuals.

d. Use the numerical and graphical work from parts (b) and (c) to explain how the table of age and wages differs from an independence structure.

```
In [161]: result = chisq.test(ct_table)
          result
          cat("p-value = ", result$p.value, ", p-value 太小代表两个变量不独立")
```

```
Warning message in chisq.test(ct_table):
“Chi-squared approximation may be incorrect”


Pearson's Chi-squared test

data:  ct_table
X-squared = 24.771, df = 9, p-value = 0.003235



p-value =  0.003235285 , p-value 太小代表两个变量不独立
```

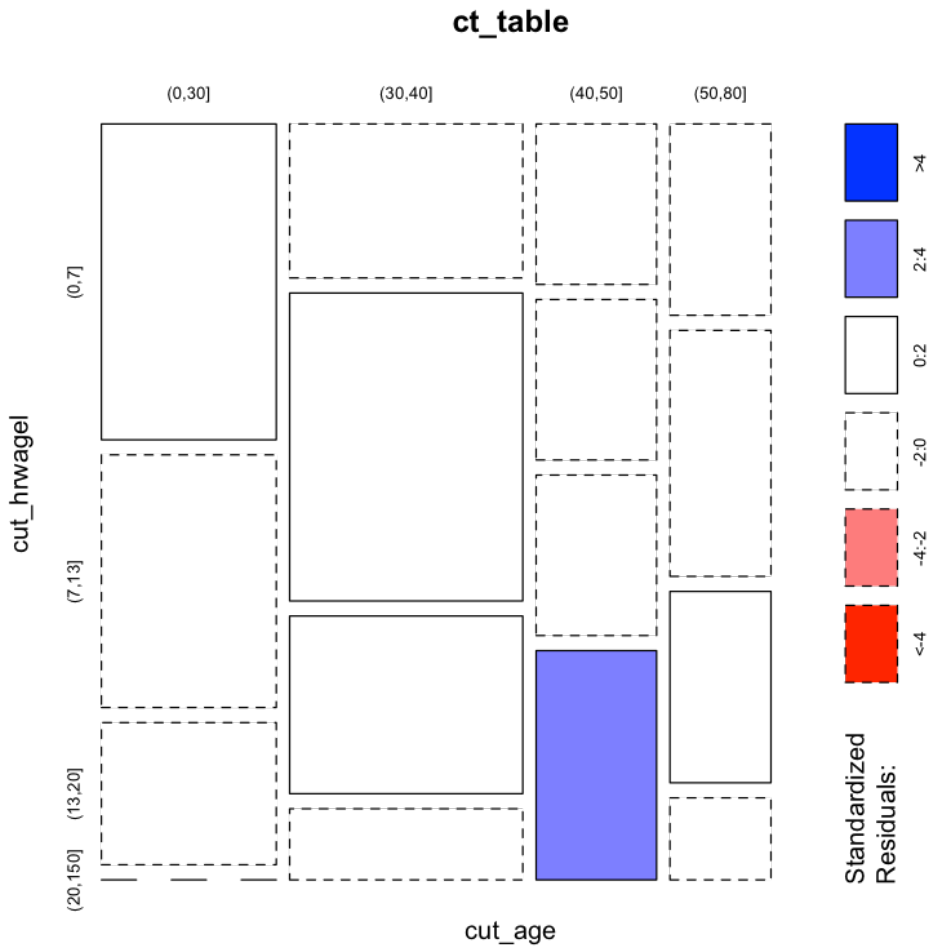
```
In [167]: result$residuals
          abs(result$residuals) > 2
```

```
         cut_hrwagel
cut_age         (0,7]      (7,13]     (13,20]    (20,150]
  (0,30]   1.92194002 -0.02768183 -0.47878990 -2.29734146
  (30,40] -1.05637022  0.97490871  0.24681203 -0.39093031
  (40,50] -0.66483709 -1.23031333 -0.10072158  3.33767089
  (50,80] -0.19778327 -0.10116070  0.36493614 -0.02827932
```

A matrix: 4 × 4 of type lgl

	(0,7]	(7,13]	(13,20]	(20,150]
(0,30]	FALSE	FALSE	FALSE	TRUE
(30,40]	FALSE	FALSE	FALSE	FALSE
(40,50]	FALSE	FALSE	FALSE	TRUE
(50,80]	FALSE	FALSE	FALSE	FALSE

```
In [179]: mosaicplot(ct_table, shade=TRUE)
```



(40, 50] 岁的 twins 很有可能拿到 (20, 150] 区间的工资。这两个变量之间不是独立的。

3.7 (Dice rolls, continued).

Suppose you roll a pair of dice 1000 times and you are interested in the relationship between the maximum of the two rolls and the sum of the rolls.

a. Using the sample function twice, simulate 1000 rolls of two dice and store the simulated rolls in the variables die1 and die2.

b. The pmax function will return the parallel maximum value of two vectors.

Using this function, compute the maximum for each of the 1000 pair of rolls and store the results in the vector max.rolls. Similarly, store the sum for each pair of rolls and store the sums in the vector sum.rolls.

c. Using the table function, construct a contingency table of the maximum roll and the sum of rolls.

d. By the computation of conditional proportions, explore the relationship between the maximum roll and the sum of rolls.

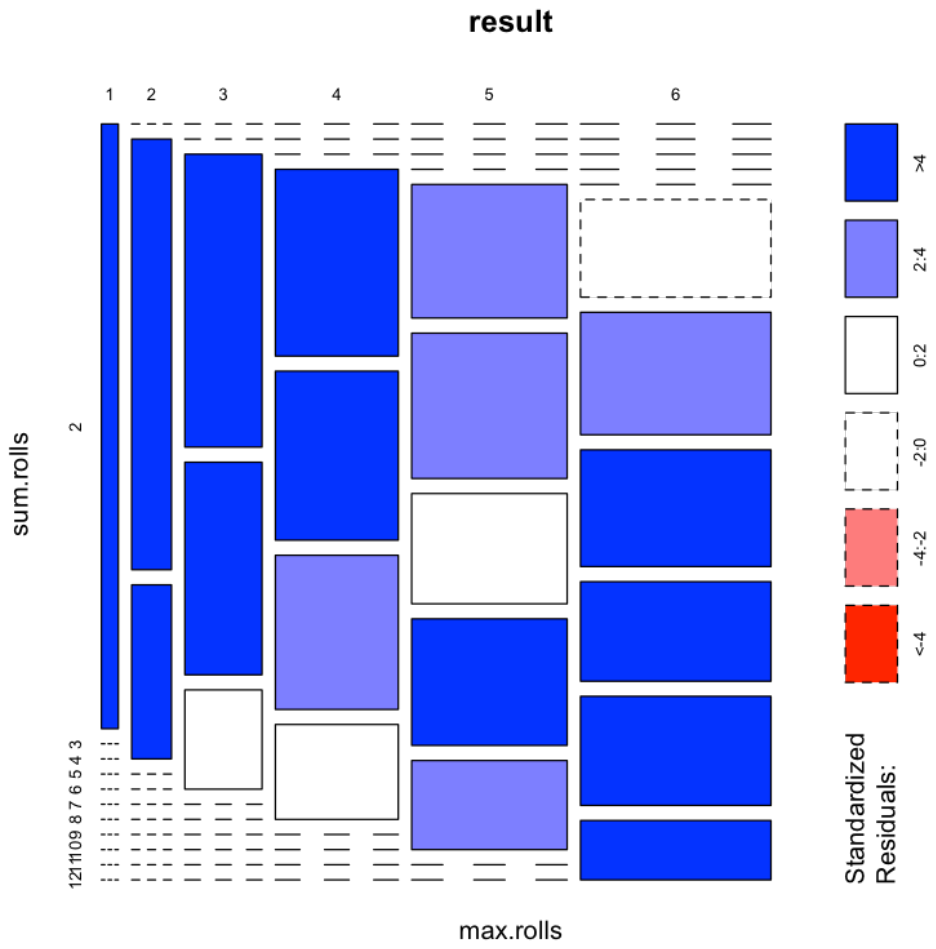
```
In [180]: die1 = sample(6, 1000, replace=TRUE)
          die2 = sample(6, 1000, replace=TRUE)
```

```
In [188]: max.rolls = pmax(die1, die2)
          sum.rolls = die1 + die2
```

```
In [190]: result = table(max.rolls, sum.rolls)
          result
```

```
         sum.rolls
max.rolls  2  3  4  5  6  7  8  9 10 11 12
        1 28  0  0  0  0  0  0  0  0  0  0
        2  0 47 19  0  0  0  0  0  0  0  0
        3  0  0 62 45 21  0  0  0  0  0  0
        4  0  0  0 63 57 52 32  0  0  0  0
        5  0  0  0  0 57 62 47 54 38  0  0
        6  0  0  0  0  0 51 64 61 52 57 31
```

```
In [192]: mosaicplot(result, shade=TRUE)
```

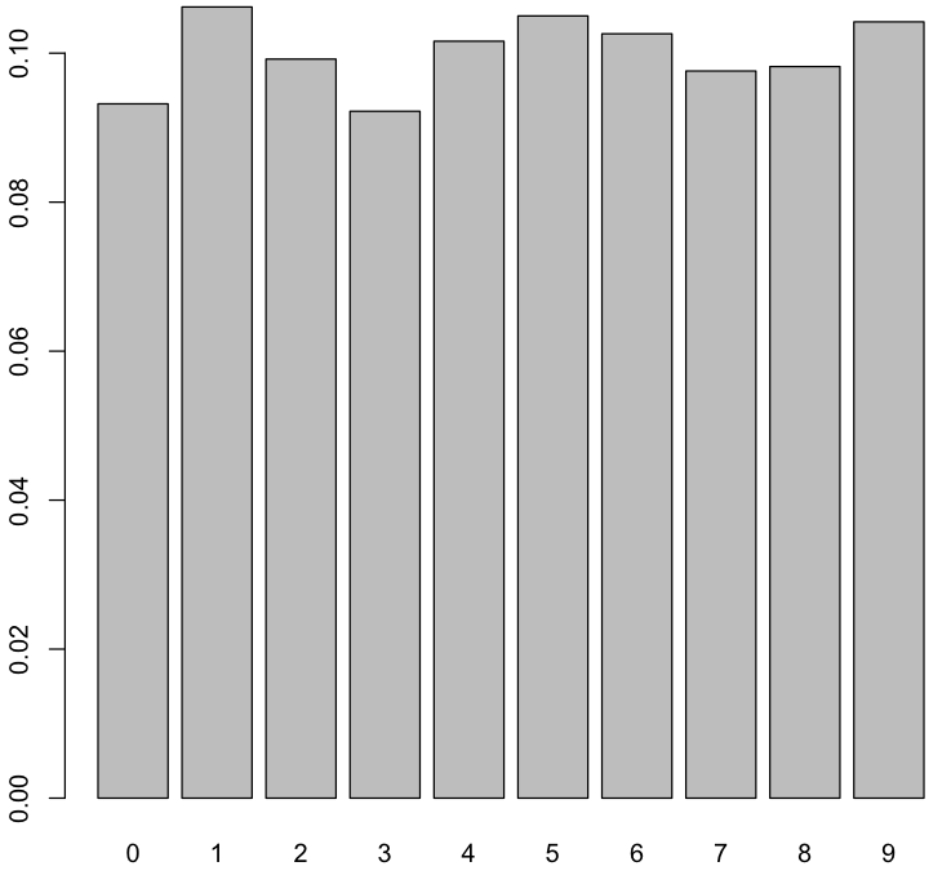


很容易发现，骰子的最大值和骰子的和之间有很强的相关性。

3.8 (Are the digits of π random?).

The National Institute of Standards and Technology has a web page that lists the first 5000 digits of the irrational number π. One can read these digits into R by means of the script

```
In [3]: pidigits = read.table("http://www.itl.nist.gov/div898/strd/univ/data/PiD

In [4]: barplot(prop.table(table(pidigits)))
```




```
In [160]: result = chisq.test(table(pidigits))
          result
          cat("p-value = ", result$p.value, ", p-value 太小代表不服从分布")
```

```
Chi-squared test for given probabilities

data:  table(pidigits)
X-squared = 10.356, df = 9, p-value = 0.3224
```

```
p-value =  0.322441 , p-value 太小代表不服从分布
```