Homework Assignment 03

Juan Cruz Ferreyra

2024-09-29

Section 1

a. What's the chance of getting a sequential pair on two rolls of a die (eg, a 3 then a 4 counts, but a 4 then a 3 does not). (Hint: you can calculate this manually if you like, by counting up the sample space and finding the fraction of that sample space that consists of ordered pairs.)

```
die1 = c("1-","2-","3-","4-","5-","6-")
die2 = c("1","2","3","4","5","6")
twodice = outer(die1, die2, "paste0")
twodice
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
## [1,] "1-1" "1-2" "1-3" "1-4" "1-5" "1-6"
## [2,] "2-1" "2-2" "2-3" "2-4" "2-5" "2-6"
## [3,] "3-1" "3-2" "3-3" "3-4" "3-5" "3-6"
## [4,] "4-1" "4-2" "4-3" "4-4" "4-5" "4-6"
## [5,] "5-1" "5-2" "5-3" "5-4" "5-5" "5-6"
## [6,] "6-1" "6-2" "6-3" "6-4" "6-5" "6-6"
```

As we can see, we have 5 pairs that match the condition of getting a sequential pair after two consecutive rolls of a die. This pairs are "1-2", "2-3", "3-4", "4-5", "5-6". The total sample space is 36, so the probability that we get is 5 out of 36.

```
paste0(
  "The probability of getting a sequential pair on two rolls of a die is: ",
  round(5/36, 3)
  )
```

```
## [1] "The probability of getting a sequential pair on two rolls of a die is: 0.139"
```

b. Given a dartboard with a inner circle that is 2/3 of the total area, and a bulls-eye that is 5% of the total area (and entirely within the inner circle): if you are throwing a random dart (that is guaranteed to hit somewhere on the board, but everywhere inside is equally likely), what is the chance of hitting the bulls-eye conditional on knowing your dart is somewhere inside the innner circle?

We calculate the probability by using the conditional probability theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

In our case, we have $P(A) = 1/20$, being A the probability of hitting the bulls-eye; and $P(B) = 2/3$, being B the probability of hitting the inner circle. Since the bulls-eye is completely inside the inner circle, the event $A \cap B$ is equal to the event A, because there is no way of hitting the bulls-eye without hitting the inner circle. Then, we replace in the equation:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{20}}{\frac{2}{3}} = \frac{3}{40}$$

```
paste0(
  "The probability of hitting the bull-eye given that we hit the inner circle is: ",
  round(3/40, 3)
  )
```

```
## [1] "The probability of hitting the bull-eye given that we hit the inner circle is: 0.075"
```

c. You take a test for a scary disease, and get a positive result. The disease is quite rare ~ 1 in 1000 in the general population. The test has a sensitivity of 95%, and a false positive rate of only 5%. What is the chance you have the disease?

First we model the events as follow:

A: is the event of having the disease, where $P(A) = 1/1000$.

B: is the event of having a positive test.

$B|A$: is the event of having a positive test given that the person is sick, where $P(B|A) = 0.95$

$B|A^C$: is the even of having a positive test given that the person is not sick, where $P(B|A^C) = 0.05$

Then, we use the bayes theorem to calculate $P(A|B)$, that is the probability of actually having the disease given that the person get a positive result.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

$$P(A|B) = \frac{0.95 * 0.001}{0.95 * 0.001 + 0.05 * 0.999}$$

```
prob_a_given_b = (0.95*0.001) / ((0.95*0.001)+(0.05*0.999))
paste0(
  "The probability of actually being sick given that the test is positive is: ",
  round(prob_a_given_b, 3)
  )
```

```
## [1] "The probability of actually being sick given that the test is positive is: 0.019"
```

d. What is the chance you have the disease if everything remains the same, but the disease is even rarer, 1 in 10,000?

We just replace the values corresponding to $P(A)$ and $P(A^C)$ in our equation.

$$P(A|B) = \frac{0.95 * 0.0001}{0.95 * 0.0001 + 0.05 * 0.9999}$$

```r
prob_a_given_b = (0.95*0.0001) / ((0.95*0.0001)+(0.05*0.9999))
paste0(
  "The probability of actually being sick given that the test is positive is: ",
  round(prob_a_given_b, 3)
  )
```

```
## [1] "The probability of actually being sick given that the test is positive is: 0.002"
```

e. What does this tell you about the dangers of tests for rare diseases?

Tests for rare diseases can be misleading, and this can easily confound our interpretation of results. To truly understand the chances of having the disease after a positive result, it's crucial to consider the precision (the probability of truly having the disease given a positive result) and recall (the test's ability to correctly identify those with the disease). These metrics help clarify the real probability of having the disease.

This becomes clear when we think about it in numbers. Imagine a population of 100,000 people, where only 100 have the disease (1 in 1,000). If the test has 95% recall, 95 out of those 100 will test positive. However, with a false positive rate of 5%, 5% of the remaining 99,900 healthy individuals (around 4,995) will also test positive. So, out of all the people who tested positive (95 true positives + 4,995 false positives), only 95 actually have the disease. This means that the vast majority of positive results are false alarms, and only a small fraction (about 1.87%) of those who test positive actually have the disease.

This highlights the importance of context. Even though a 5% false positive rate might sound low, when applied to a rare disease, it results in many more false positives than true positives. Understanding this helps us avoid overestimating the seriousness of a positive result for rare conditions.

Section 2

```r
library(ggplot2)
library(tidyverse)
```

a. You have a 20-side die. Using sample, roll it 10,000 times and count the number of rolls that are 10 or less.

```r
die20 <- 1:20
rolls <- sample(die20, 10000, replace = TRUE)

paste0(
  "The number of rolls that are equal to 10 or less is: ",
  sum(rolls <= 10)
  )
```

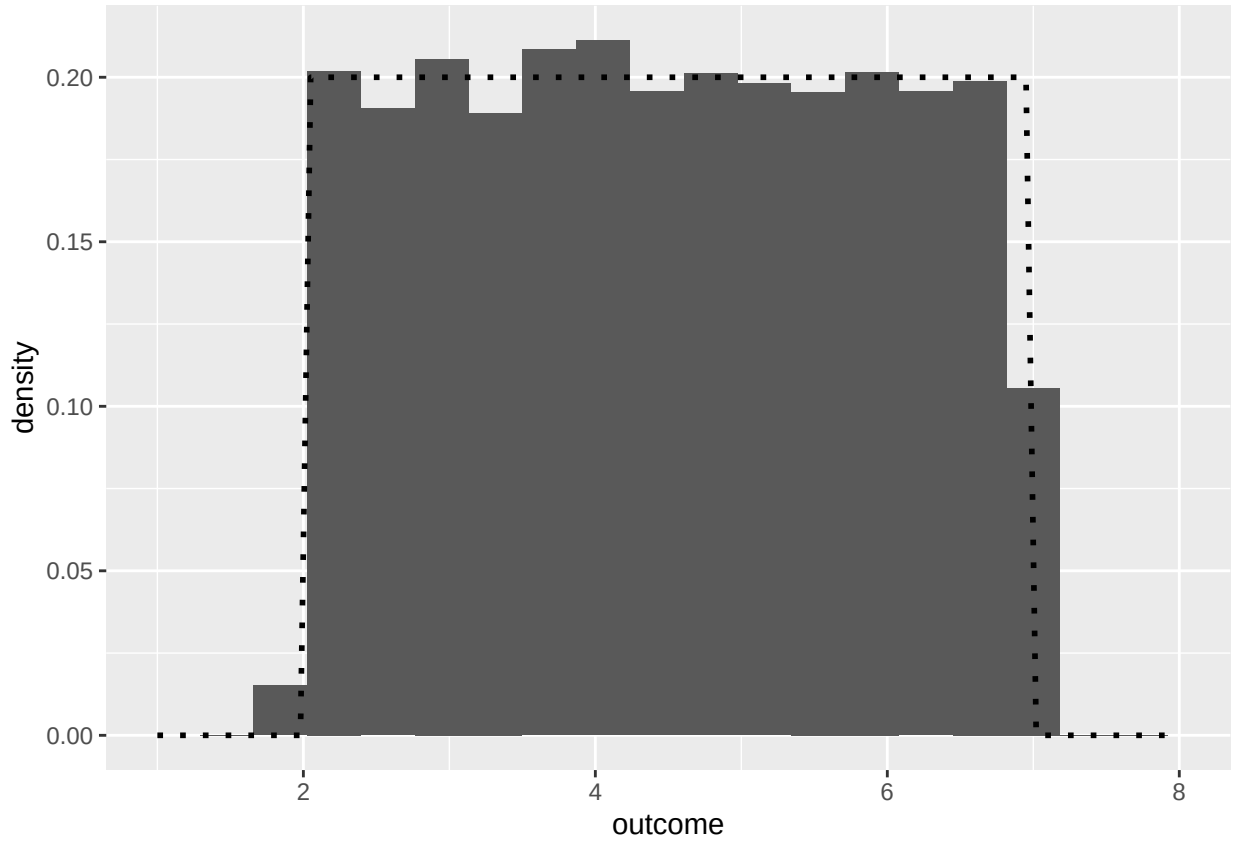
```
## [1] "The number of rolls that are equal to 10 or less is: 4986"
```

b. Generate a histogram using ggplot of 10,000 draws from a uniform distribution between 2 and 7.

```r
min_val <- 2
max_val <- 7
randunifs <- runif(10000, min_val, max_val)

uniformfun <- function(x){ifelse(x >= min_val & x <= max_val, 1 / (max_val - min_val), 0)}

ggplot(data.frame(randunifs)) +
  geom_histogram(aes(x=randunifs, y=..density..), bins=20) +
  xlim(1, 8) + ylab("density") + xlab("outcome") +
  stat_function(fun=uniformfun, color = "black", linewidth = 1, linetype = "dotted")
```



c. Try to write down the equation for this probability density function.

The probability density function (PDF) of the uniform is the following:

$$f_X = \frac{1}{b-a} = \frac{1}{7-2} = \frac{1}{5}$$

for $\{a \leq x \leq b\}$, 0 otherwise.

d. What is the probability that a draw from this distribution will be between 1.5 and 3.2?

We can calculate $P(1.5 \leq x \leq 3.2)$ using the cumulative density function (CDF) of the uniform distribution. We have that $P(1.5 \leq x \leq 3.2) = P(x \leq 3.2) - P(x \leq 1.5)$, then we need to integrate our PDF in boths ranges and get the difference. In this case, since the range of our PDF is $\{2 \leq x \leq 7\}$, that is the lower bound for the calculation is out of the PDF range, we just need to calculate $P(x \leq 3.2)$.

$$\int_2^{3.2} \frac{1}{5}\, dx$$

$$\frac{1}{5} \cdot (3.2 - 2) = \frac{1}{5} \cdot 1.2 = 0.24$$

We can check it using R function *punif()*.

```
paste0(
  "The probability that a draw from the distribution is between 1.5 and 3.2 is: ",
  punif(3.2, 2, 7)
  )
```

```
## [1] "The probability that a draw from the distribution is between 1.5 and 3.2 is: 0.24"
```

Section 3

a. Using R's cdf for the binomial, what is the probability of getting 500 or fewer "20"s when rolling your 20-sided die 10,000 times. Looking back at 2a, how many of your rolls were actually 20s?

The probability mass function (PMF) for the binomial is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, that is the probability of obtaining k number of successes in n trials.

Then, we can define the cumulative distribution function (CDF) for the binomial, that is, the probability of obtaining k or less number of successes in n trials as follow:

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

for $\{0 \leq x\}$, 0 otherwise.

For our current case, p is the probability of getting a 20 in a 20-sided die (we assume it, a fair die), so $p = 1/20$. We also have $k = 500$ and $n = 10,000$. We use R code to solve it, avoiding the large summation.

```
paste0(
  "The probability of getting 500 or fewer 20s is: ",
  round(pbinom(500, 10000, 1/20), 3)
  )
```

```
## [1] "The probability of getting 500 or fewer 20s is: 0.512"
```

```
paste0(
  "The number of 20s obtained in 10,000 simulated rolls of a 20-sided fair die is: ",
  sum(rolls == 20)
  )
```

```
## [1] "The number of 20s obtained in 10,000 simulated rolls of a 20-sided fair die is: 515"
```

b. Using rbinom, roll a 100-sided die 100 times and report the total number of 7s you get.

```
paste0(
  "The number of 7s obtained in 100 simulated rolls of a 100-sided fair die is: ",
  rbinom(1, 100, 1/100)
  )
```

```
## [1] "The number of 7s obtained in 100 simulated rolls of a 100-sided fair die is: 2"
```

c. You are a klutz, and the average number of times you drop your pencil in a day is 1. Using the poisson functions in R, what's the chance of dropping your pencil two or more times in a day? (Hint: calculate the chance of dropping it one or fewer times, and then take 1 minus that.)

The event of dropping k pencils per unit of time is modeled by the poisson distribution with a PDF:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $\{0 \leq x\}$, 0 otherwise.

Where λ is the average number of occurences per unit of time. In this case $\lambda = 1$. Then we can calculate the probability of dropping the pencil 2 or more times in a day in the following way:

$$P(2 \leq k) = \sum_{i=2}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} = 1 - \sum_{i=0}^{1} \frac{\lambda^i e^{-\lambda}}{i!}$$

We use R code to solve it, avoiding the complex equation.

```
paste0(
  "The probability of droppping the pencil 2 or more times in a day is: ",
  round(1 - ppois(1, 1), 3)
  )
```

```
## [1] "The probability of droppping the pencil 2 or more times in a day is: 0.264"
```

d. Because he is lazy, your teacher has assigned grades for an exam at random, and to help hide his deception he has given the fake grades a normal distribution with a mean of 70 and a standard deviation of 10. What is the chance your exam got a score of 85 or above? What is the chance you got a score between 50 and 60?

Since the grades as distributed $N \sim (70, 10)$, to calculate the probability of scoring 85 or above, or getting a score between 50 and 60, we should standarize the normal distribution to use the tables for the CDF of the normal.

Luckily, R allows us to solve directly

```
paste0(
  "The probability of scoring higher than 85 is: ",
  round(1 - pnorm(85, 70, 10), 3)
  )
```

```
## [1] "The probability of scoring higher than 85 is: 0.067"
```

```r
paste0(
  "The probability of scoring between 50 and 60 is: ",
  round(pnorm(60, 70, 10) - pnorm(50, 70, 10), 3)
  )
```

```
## [1] "The probability of scoring between 50 and 60 is: 0.136"
```