Assignment 4 for Statistical Computing and Empirical Mathods

Dr. Henry WJ Reeve

Teaching block 1 2021

Introduction

This document describes your fourth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommend that you first watch video lectures 9 and 10.

You are encouraged to discuss these questions with your colleagues.

Begin by creating an Rmarkdown document with html output. You are not expected to hand in this piece of work, but it is a good idea to get used to using Rmarkdown.

Several **optional** extra questions have also been included. These are marked by stars ('*"). It is recommended that you first complete all unstarred questions before proceeding through the starred questions. **Do not** be concerned if you do not have time to complete the starred questions!

1 Bayes theorem

Let A be the event that it rains next week and B the event that the weather forecaster predicts that there will be rain next week. Let's suppose that the probability of rain next week is $\mathbb{P}(A) = 0.9$. Suppose also that the conditional probability that there is a forecast of rain, given that it really does rain, is $\mathbb{P}(B|A) = 0.8$. On the other hand, the conditional probability that there is a forecast of dry weather, given that there really isn't any rain is $\mathbb{P}(B^c|A^c) = 0.75$.

Now suppose that there is a forecast of rain. What is the conditional probability of rain, given the forecast of rain $\mathbb{P}(A|B)$?

2 Conditional probabilities

Suppose we have a probability space $(\Omega, \mathcal{E}, \mathbb{P})$.

1. Suppose that A,$B \in \mathcal{E}$ and $A \subseteq B$. Give an expression for $\mathbb{P}(A|B)$ in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$. What about when $\mathbb{P}(A \setminus B) = 0$?
2. Suppose that A, $B \in \mathcal{E}$ with $A \cap B = \emptyset$. Give an expression for $\mathbb{P}(A|B)$. What about when $\mathbb{P}(A \cap B) = 0$?
3. Suppose that A, $B \in \mathcal{E}$ with $B \subseteq A$. Give an expression for $\mathbb{P}(A|B)$. What about when $\mathbb{P}(B \setminus A) = 0$?

4. Suppose that $A \in \mathcal{E}$. Give an expression for $\mathbb{P}(A|\Omega)$ in terms of $\mathbb{P}(A)$.
5. Show that given three events $A, B, C \in \mathcal{E}$ we have $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C) \cdot \mathbb{P}(B|C) \cdot \mathbb{P}(C)$.
6. Show that given three events $A, B, C \in \mathcal{E}$ we have $\mathbb{P}(A|B \cap C) = \frac{\mathbb{P}(B|A \cap C) \cdot \mathbb{P}(A|C)}{\mathbb{P}(B|C)}$.

3 Sampling with replacement

Recall that for positive integers n and k, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ gives the number of subsets of size k from a set of n objects.

You can compute this number straightforwardly within R via the choose function `choose( )`. For example, if we want to compute the number of different subsets of size 3 from a collection of size 8 we would compute:

```
choose(8,3)
```

```
## [1] 56
```

Suppose we have a bag containing 10 spheres. This includes 3 red spheres and 7 blue spheres.

Let's suppose that we draw a sphere at random from the bag (all spheres have equal probability of being drawn). We record its colour and then return the sphere to the bag. This process is repeated 35 times. This is an example of **sampling with replacement** since the spheres are replaced after each draw.

Write down a mathematical expression for the probability that z out of the 35 selections were red spheres (here $z \in \{1, \ldots, 35\}$).
Try doing this with in laTex, making use of the laTex functions `binom{}{}` and `frac{}{}`.

Next write an R function called `prob_red_spheres()` which takes z as an argument and computes the probability that z out of a total of the 35 balls selected are red.

Test your function as follows:

```
prob_red_spheres(20)
```

```
## [1] 0.0005376555
```

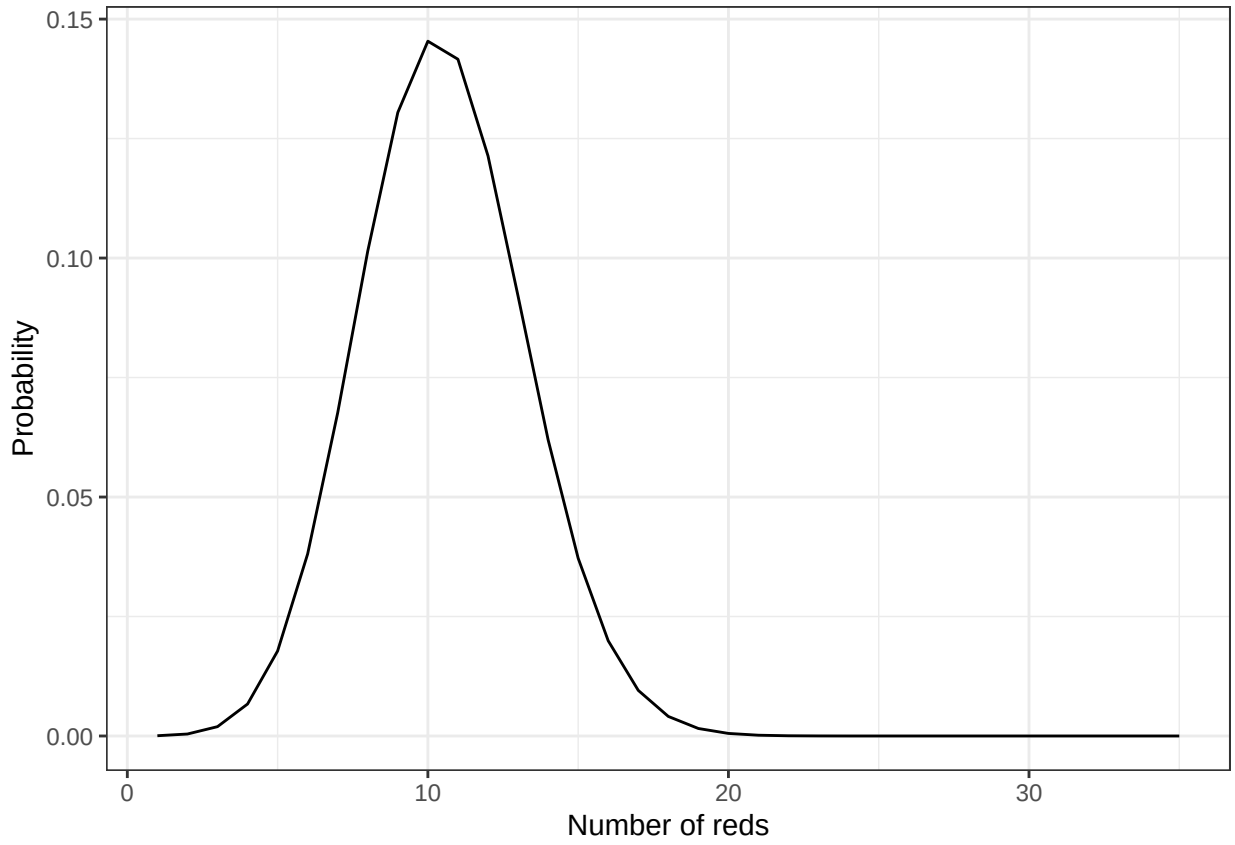
Generate a data frame called `prob_by_num_reds` with two columns `num_reds` and `prob`. The `num_reds` column should contain numbers 1 through 35 and the `prob` column should give the associated probability of selecting that many reds out of a total number of 35 selections.

Display the first 3 rows of your data frame:

```
prob_by_num_reds%>%head(3)
```

```
##   num_reds         prob
## 1        1 0.0000568228
## 2        2 0.0004139947
## 3        3 0.0019516894
```

Now use the `geom_line()` function within the `ggplot2` library, in conjunction with your data frame to display a plot of the probability as a function of the number of reds. Your plot should look as follows:



Next we shall explore the `sample()` function within R. Let's suppose we want to simulate a random experiment in which we sample with replacement from a collection of 10 objects, and repeat this process 35 times. We can do this by calling:

```
sample(10,35,replace=TRUE)
```

Try this out for yourself. The output should be a vector of length 35 consisting entirely of numbers between 1 and 10. Since this is sampling with replacements and the number of samples exceeds the number of elements there will be repetitions.

Try rerunning the function. You probably get a different sample. This is to be expected, and even desirable, since the function simulates a random sample. However, the fact that we get a different answer every time we run the code is problematic from the prespective of reproducibility. To avoid this process we can set a random seed via the function `set.seed()`. By doing so we should get the same output every time. Try the following out for your self:

```
## Setting the random seed just once

set.seed(0)

for(i in 1:5){

  print(sample(100,5,replace=FALSE))
```

```
  # The result may well differ every time

}


## Resetting the random seed every time

for(i in 1:5){

  set.seed(1)
  print(sample(100,5,replace=FALSE))
  # The result should not change

}
```

We shall now use the `sample()` to construct a simulation study to explore the probability of selecting z red balls from a bag of size 10, with 3 red and 7 blue balls, when sampling 35 balls with replacement.

First set a random seed. Then create a data frame called `sampling_with_replacement_simulation` consisting of a two columns. The first is called `trial` and contains numbers 1 through 1000. The second is called `sample_balls` and corresponds to a random sample of size 35 from a bag of size 10, with replacement. We can do this as follows:

```
num_trials<-1000 # set the number of trials
set.seed(0) # set the random seed

sampling_with_replacement_simulation<-data.frame(trial=1:num_trials)%>%
  mutate(sample_balls=map(.x=trial,~sample(10,35,replace = TRUE)))
  # generate collection of num_trials simulations
```

Now add a new column called `num_reds` such that, for each row, `num_reds` contains an integer which gives the number of items within the sample for that row (the entry in the `sample_balls` column) which are less than or equal to three. For example, suppose that some row of the data frame, the `sample_balls` column contains the following list:

```
9, 4, 7, 1, 2, 7, 2, 3, 1, 5, 5, 10, 6, 10, 7, 9, 5, 5, 9, 9,
5, 5, 2, 10, 9, 1, 4, 3, 6, 10, 10, 6, 4, 4, 10
```

Then the corresponding row of the `num_reds` column should contain the number 8, since 8 of these values are less than equal to 3. You may want to use the functions `mutate()`, `map_dbl` and `sum()`.

Next we shall add a new column called `simulation_count` to our existing data frame `prob_by_num_reds` which gives the number of times within our simulation we observed the corresponding number of reds. We can do this as follows:

```
num_reds_in_simulation<-sampling_with_replacement_simulation%>%pull(num_reds)
# we extract a vector corresponding to the number of reds in each trial

prob_by_num_reds<-prob_by_num_reds%>%
  mutate(simulation_count=map_dbl(.x=num_reds,~sum(num_reds_in_simulation==.x)))
  # add a column which gives the number of trials with a given number of reds
```

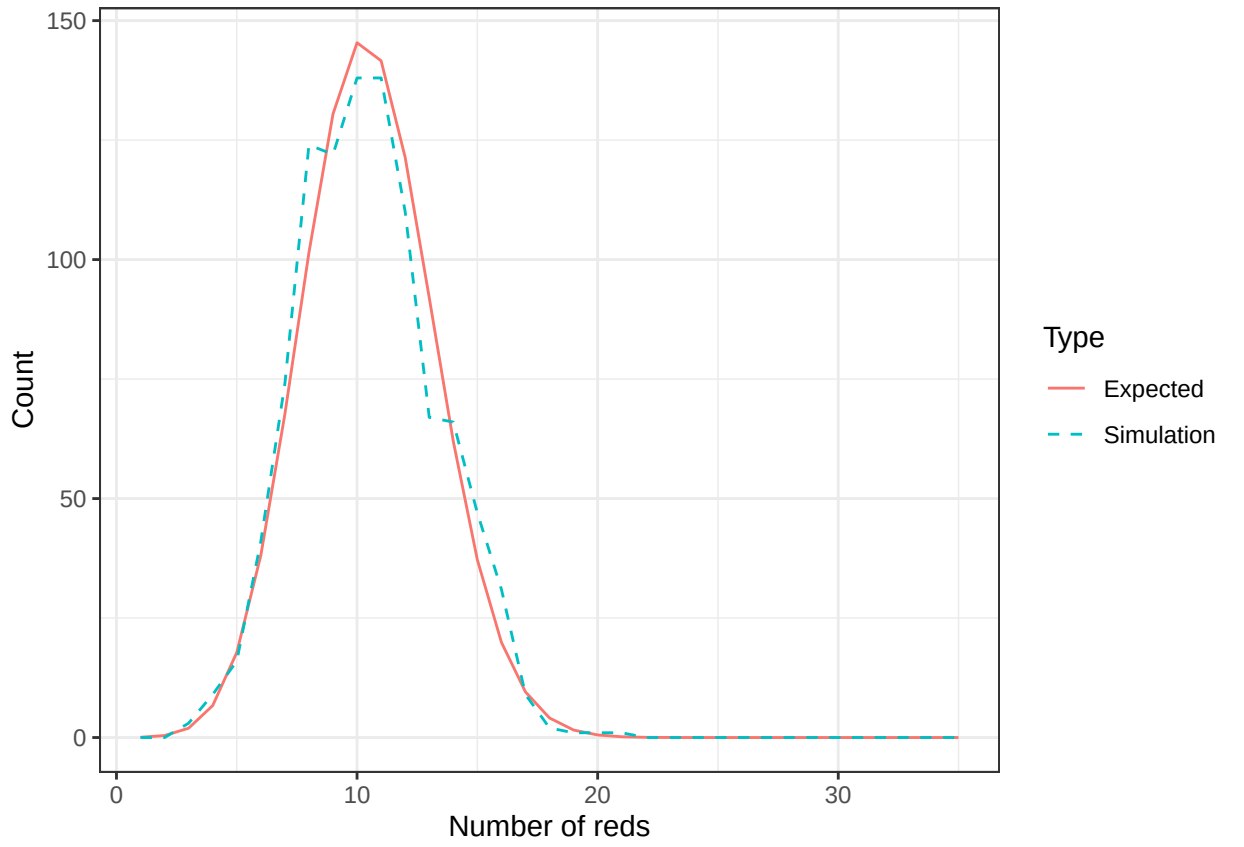
Next we add a column called `expected_count` corresponding to the *expected* number of observed reds in 1000 based upon your probability formula.

```
prob_by_num_reds<-prob_by_num_reds%>%
  mutate(expected_count=num_trials*prob)
  # add a column which gives the expected number of reds
  # based on the probability formula
```

Finally, create a plot which compares the results of your simulation with the expected count based on your probability formula. The concept of expectation will be discussed in Lecture 11.

Your result should look something like the plot below. Of course, since this is a random simulation, your result may well look slightly different.

```
prob_by_num_reds%>%
  rename(Simulation=simulation_count,Expected=expected_count)%>%
  pivot_longer(cols=c("Simulation","Expected"),
               names_to="Type",values_to="count")%>%
  ggplot(aes(num_reds,count)) +
    geom_line(aes(linetype=Type, color=Type)) +
    scale_linetype_manual(values = c("solid", "dashed"))+
  theme_bw()+
  xlab("Number of reds")+
  ylab("Count")
```



Try to make sense of each line in the above code.

Notice the close relationship between probabilities and long-run frequency behaviour. We shall explore this connection over the next few lectures.

4 Sampling without replacement

This question is more challenging. However, you should aim to complete at least the simulation component using ideas from the previous question.

Let's suppose we have large bag containing 100 spheres. There are 50 red spheres, 30 blue spheres and 20 green spheres. Suppose that we sample 10 spheres from the bag **without** replacement.

What is the probability that one or more colours is missing from your selection?

First aim to answer this question via a simulation study using ideas from the previous question.

You may want to use the following steps:

1. First set a random seed;
2. Next set a number of trials, a number of reds, a number of blues, a number of greens and a sample size;
3. Now use a combination of the functions `sample()`, `mutate()` and `map()` to generate your samples. Here you are creating sample of size 10 from a collection of 100 balls - the sampling is done **without** replacement;
4. Now compute the number of "reds", "greens" and "blues" in your sample using the `map_dbl()` and `mutate()` functions;
5. Compute the minimum of the three counts using the `pmin()` function. When this minimum is zero, then one of the three colours is missing. It is recommended that you look up the difference between `pmin()` and `min()` here;
6. Compute the proportion of rows for which the minimum number of the three counts is zero.

Try this initially with a small number of simulations. Increase your number of simulations to about a relatively large number to get a more accurate answer, once everything seems to be working well.

The next part of the question is more challenging and may be omitted if you are short on time.

(*) Once you have a simulation based answer you can try and use "combinations" with $\binom{n}{k}$ to compute the probability directly. First aim and compute the number of subsets of size 10 from 100 which either entirely miss out one of the subsets Reds $= \{1, \ldots, 50\}$, Blues $= \{51, \ldots, 80\}$, Greens $= \{81, \ldots, 100\}$. Be careful not to double count some of these subsets! Once you have computed all such subsets combine with the formula for the total number of subsets of size 10 from a set of 100, to compute the probability of missing a colour.

5 Mutual independence and pair-wise independent

Consider a simple probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with $\Omega = \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$.

Since $(\Omega, \mathcal{E}, \mathbb{P})$ is a simple probability space containing four elements we have

$$\mathbb{P}(\{(0,0,0)\}) = \mathbb{P}(\{(0,1,1)\}) = \mathbb{P}(\{(1,0,1)\}) = \mathbb{P}(\{(1,1,0)\}) = \frac{1}{4}.$$

Consider the events $A := \{(1,0,1),(1,1,0)\}$, $B := \{(0,1,1),(1,1,0)\}$ and $C := \{(0,1,1),(1,0,1)\}$.

Verify that $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, $\mathbb{P}(A \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(C)$ and $\mathbb{P}(B \cap C) = \mathbb{P}(B) \cdot \mathbb{P}(C)$. Hence, we deduce that the events A, B, C are **pair-wise independent**.

What is $A \cap B \cap C$? What is $\mathbb{P}(A \cap B \cap C)$?

Are the events A,B,C **mutually independent**?

6 The Monty hall problem (*)

Consider the following game:

At a game show there are three seemingly identical doors. Behind one of the doors is a car, and behind the remaining two is a goat. The contestant of the game first gets to choose one of the three doors. The host then opens one of the other two doors to reveal a goat. The contestant now get's a chance to either (a) switch their choice to the other unopened door or (b) stick to their original choice. The host then opens the door corresponding to the contestant's final choice. They get to keep whatever is behind their final choice of door.

Does the contestant improve their chances of winning the car if they switch their choice?

For clarity, we make the following assumptions:

1. The car is assigned to one of the doors at random with equal probability for each door.
2. The assignment of the car and the initial choice of the contestant are independent.
3. Once the contestant makes their initial choice, the host always opens a door which (a) has a goat behind it and (b) is not the contestant's initial choice. If there is more than one such door (i.e. when the contestant's initial choice corresponds to the door with a car behind it) the host chooses at random from the two possibilities with equal probability.

To formalise our problem we introduce the following events for $i = 1, 2, 3$:

- A_i denotes event car is placed behind the i-th door;
- B_i denotes event contestant initially chooses the i-th door;
- C_i denotes event the host opens the i-th door to reveal a goat.

Consider a situation in which the contestant initially selects the first door ($B1$) and then the host opens the second door to reveal a goat ($C2$). What is $\mathbb{P}(A_3 | B_1 \cap C_2)$?

What does this suggest about a good strategy? Should we switch choices?

7 A game of marbles (**)

This question is left as an optional challenge for those who have completed all remaining questions.

Suppose there two players - A and B. Player A has n marbles and player B has $k-n$ marbles ($n \leq k$), so the two players have k marbles between them. The game proceeds through a series of rounds. In each round A has probability p of winning the round, and B has probability $1-p$ of winning the round, where $p \in (0, 1/2)$ is a fixed number. If A wins a round then B must give A one of his marbles. If B wins the round then A must give B one of his marbles. All of the rounds are independent from one another. A player wins the entire game of marbles when they have all k marbles, and the other player has none.

Can you give a formula for the probability that player A wins the game in terms of n and k and p?

If you have completed all the questions on this assignment email `henry.reeve@bristol.ac.uk`.