Assignment 1

Introduction to Bayesian Data Analysis 2025

Ana Sofia Acevedo (03805441)

Preamble

- **Points**: Assignment 1 comprises of 6 tasks, 2 points each (12 in total). 2 points are obtained for complete and correct answers. 1 point is obtained for a proper approach or if only part of the task is solved.
- **Submission**: Hand in the assignment as a PDF `Markdown` report. The report should show the results, the code that produced the results, and additional text or comment. The report should appear clean and be uploaded on Moodle until Wednesday, June 04, 9:45 am.
- **Collaboration**: Reports can be handed in as team work (max. 2 people). When working in teams, declare this on page 2. However, each collaborator needs to hand in a report via Moodle, stating their name, student number (p. 1), and their machine specification (p. 2).
- **Permitted and Prohibited**: You may use materials from this class (e.g., slides, code on GitHub) as well as online forums such as Stack Overflow to write your code. However, you are not allowed to post questions from the assignment online or prompt them (including paraphrases) to LLMs/chatbots. All use of LLMs/chatbots is generally not allowed. Solutions may not be shared with other students from the class (except 1 potential collaborator).

Authorship Information

1. Declaration of Collaboration

☐ Yes (Collaborator name)
☒ No

2. Declaration of Authorship

☒ I certify that this assignment represents my own work. I have not used any unauthorized or unacknowledged aids as stated in the preamble, including free or commercial systems or services offered on the internet or text generating systems embedded into software. I did not copy code from someone else nor did I share my code with someone else.

3. System Information

☒ I confirm that I generated the submitted PDF report myself using R version 4.5.0 (2025-04-11 ucrt) and `Quarto/RMarkdown`.

Machine stamp: x86_64-w64-mingw32/x64

Timestamp: 2025-06-02 10:56:36 CEST

```r
# load packages here
library(tidyverse)
library(ggplot2)
library(gtools)
```

Task Set 1

For tasks 1.1-1.3, assume you throw 3 dice—normal dice with 6 sides each ;-)

Task 1.1

Create a data frame with all possible combinations (outcomes) that can result from throwing all three dice. Each row should represent one possible outcome. Print the first/last 10 rows and the total number of possible outcomes to the output.

Answer

The number of possible outcomes can be computed as a permutation, where repeated numbers are allowed. The library *gtools* has a function that does exactly this.

```r
perm <- as.data.frame(permutations(6,3, repeats.allowed = TRUE))
print(head(perm, 10))
```

```
   V1 V2 V3
1   1  1  1
2   1  1  2
3   1  1  3
4   1  1  4
5   1  1  5
6   1  1  6
7   1  2  1
8   1  2  2
9   1  2  3
10  1  2  4
```

```r
print(tail(perm, 10))
```

```
    V1 V2 V3
207  6  5  3
208  6  5  4
209  6  5  5
210  6  5  6
211  6  6  1
212  6  6  2
213  6  6  3
214  6  6  4
215  6  6  5
216  6  6  6
```

```r
print(nrow(perm))
```

```
[1] 216
```

Task 1.2

Create a data frame showing all possible sums that can result from throwing the three dice (e.g., if all three dice show a `1`, the sum is `3`) along with their probabilities. Print the data frame to the output and visualize the distribution in a plot.

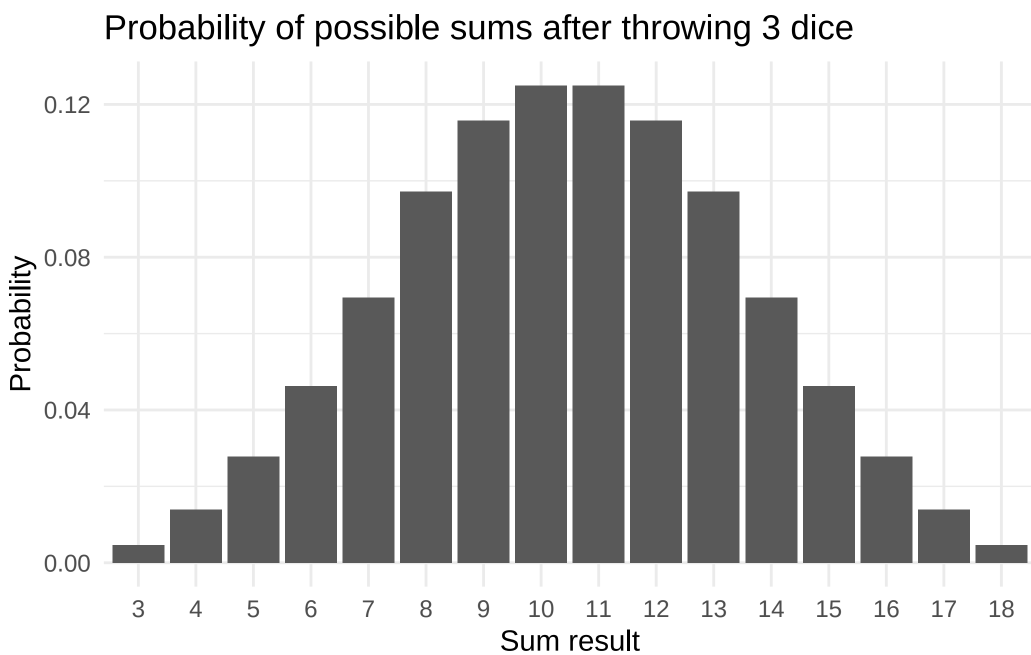
Answer

```r
data <- as.data.frame(table(rowSums(perm)))
colnames(data) <- c("sum", "prob")
data$prob <- data$prob/sum(data$prob)
print(data)
```

```
   sum       prob
1    3 0.00462963
2    4 0.01388889
3    5 0.02777778
4    6 0.04629630
5    7 0.06944444
6    8 0.09722222
7    9 0.11574074
8   10 0.12500000
9   11 0.12500000
10  12 0.11574074
```

```
11   13 0.09722222
12   14 0.06944444
13   15 0.04629630
14   16 0.02777778
15   17 0.01388889
16   18 0.00462963
```

```r
ggplot(data, aes(x=sum, y=prob)) +
  geom_bar(stat = "identity") +
  labs(title = "Probability of possible sums after throwing 3 dice",
       x = "Sum result",
       y = "Probability") +
  theme_minimal()
```



Task 1.3

Which of the following events is more likely?

☒ A. The sum is $\geq$ `12`, *given* that at least one of the dice shows a `4`.
☐ B. The sum is $\leq$ `10`, *given* that at least one of the dice shows a `5`.

Answer

From Bayes' theorem, the probabilities for each sum given the respective conditions can be calculated. In this sense,

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

The probability for the first scenario is

```
perm$total <- rowSums(perm)
num_1 <- nrow(perm[(perm$total>=12) & (perm$V1==4 | perm$V2==4 | perm$V3==4),])/nrow(perm)
den_1 <- nrow(perm[(perm$V1==4 | perm$V2==4 | perm$V3==4),])/nrow(perm)
num_1/den_1
```

```
[1] 0.4065934
```

The probability for the second scenario is

```
num_2 <- nrow(perm[(perm$total<=10) & (perm$V1==5 | perm$V2==5 | perm$V3==5),])/nrow(perm)
den_2 <- nrow(perm[(perm$V1==5 | perm$V2==5 | perm$V3==5),])/nrow(perm)
num_2/den_2
```

```
[1] 0.3296703
```

Task Set 2

For Task 2.1-2.3, assume you toss a globe `10` times, leading to either land or water.

Task 2.1

Compute the probability of all possible numbers of occurrence of land, given the candidate proportion of `.3`. Report the results in a summary table and a plot and indicate whether the plot shows a probability distribution or a likelihood function.

Answer

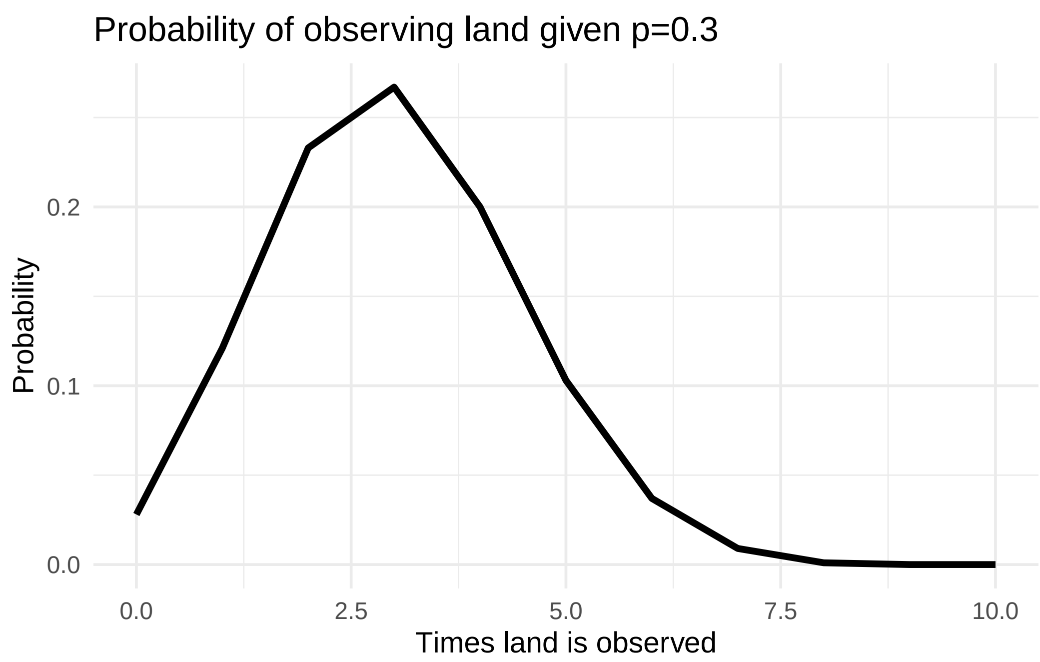
Since the candidate proportion of land (that is, the parameter) is fixed, the plot shows a probability distribution of the number of times land is observed given the proportion.

```
# i is the number of times that land is observed,
# so 10-i is the number of times water is observed
df <- data.frame(land = 0:10, probs = rep(0,11))
for (i in 0:10){
  df[i+1, 2] <- dbinom(i, size = 10, prob = 0.3)
}
df$probs <- round(df$probs, 3)
print(df)
```

```
   land probs
1     0 0.028
2     1 0.121
3     2 0.233
4     3 0.267
5     4 0.200
6     5 0.103
7     6 0.037
8     7 0.009
9     8 0.001
10    9 0.000
11   10 0.000
```

```
ggplot(df, aes(x = land, y = probs)) +
  geom_line(linewidth = 1.2) +
  labs(title = "Probability of observing land given p=0.3",
       x = "Times land is observed",
```

```
      y = "Probability") +
  theme_minimal()
```



Task 2.2

Assume you observe 8 water. Take the candidate proportions of land `cp =` 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. For each of these candidates, compute the probability of this observation. Report the results in a summary table and a plot and indicate whether the plot shows a probability distribution or a likelihood function.

Answer

Since in this case the observation is fixed and the parameter is variable, the plot shows a likelihood function that is normalized. To solve this task, I compared the results I go using the code I wrote during class (first part), and then compared it to what I got by using the binomial distribution function.

```
# Coded during class
counter_prob_1 <- function(data, cp){
  L <- sum(data == "L")
  W <- sum(data == "W")
```

```r
  probs <- (cp)^L * ((1 - cp))^W
  probs <- probs/sum(probs)
  data.frame(cp, probs)
}

data <- c(rep("W", 8), "L","L")
cp_seq <- seq(0, 1, 0.1)
results <- counter_prob_1(data, cp = cp_seq)
results$probs <- round(results$probs, 3)
print(results)
```

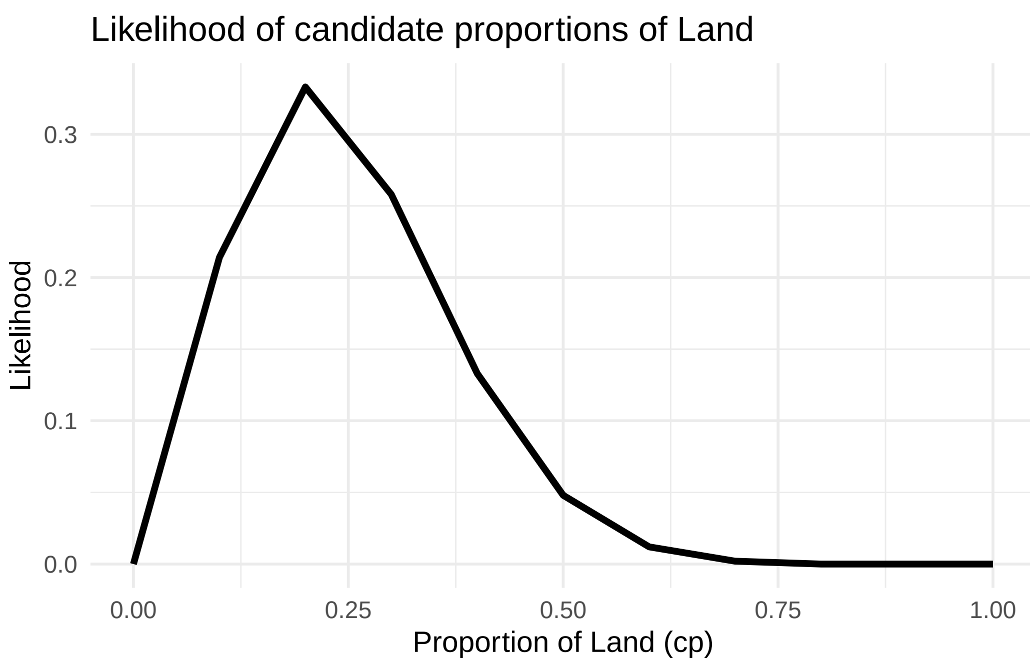
```
    cp probs
1  0.0 0.000
2  0.1 0.214
3  0.2 0.333
4  0.3 0.258
5  0.4 0.133
6  0.5 0.048
7  0.6 0.012
8  0.7 0.002
9  0.8 0.000
10 0.9 0.000
11 1.0 0.000
```

```r
# Code using the binomial distribution function
df_1 <- data.frame(cp = seq(0, 1, 0.1), probs = rep(0,11))
count <- 1
for (i in seq(0, 1, 0.1)){
  df_1[count,2] <- dbinom(2, size = 10, prob = i)
  count <- count+1
}
df_1$probs <- df_1$probs / sum(df_1$probs)
df_1$probs <- round(df_1$probs, 3)
print(df_1)
```

```
    cp probs
1  0.0 0.000
2  0.1 0.214
3  0.2 0.333
4  0.3 0.258
5  0.4 0.133
```

```
6  0.5 0.048
7  0.6 0.012
8  0.7 0.002
9  0.8 0.000
10 0.9 0.000
11 1.0 0.000
```

```r
# Plot with the results
ggplot(results, aes(x = cp, y = probs)) +
  geom_line(linewidth = 1.2) +
  labs(title = "Likelihood of candidate proportions of Land",
       x = "Proportion of Land (cp)",
       y = "Likelihood") +
  theme_minimal()
```



Task 2.3

For each candidate proportion of land, compute the probability of all possible number of occurrences of land. Report the results in a summary table, showing the probability distributions as columns and the likelihood functions as rows.

Answer

Since the rows represent the likelihood function, they do not need to necessarily be normalized to add up to 1, while the columns represent a probability distribution function and thus have to add up to 1. Therefore, the binomial distribution function was used.

```r
cp_seq <- seq(0, 1, 0.1)
land <- c(0:10)
df_2 <- data.frame(matrix(0, nrow=length(land), ncol=length(cp_seq)+1))
colnames(df_2) <- c("land", as.character(cp_seq))
df_2$land <- land

for (i in land){
  for (j in cp_seq){
    df_2[i+1,as.character(j)] <- dbinom(i, size = 10, prob = j)
    df_2[i+1,as.character(j)] <- round(df_2[i+1,as.character(j)], 3)
  }
}

print(df_2)
```

```
   land 0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9 1
1     0 1 0.349 0.107 0.028 0.006 0.001 0.000 0.000 0.000 0.000 0
2     1 0 0.387 0.268 0.121 0.040 0.010 0.002 0.000 0.000 0.000 0
3     2 0 0.194 0.302 0.233 0.121 0.044 0.011 0.001 0.000 0.000 0
4     3 0 0.057 0.201 0.267 0.215 0.117 0.042 0.009 0.001 0.000 0
5     4 0 0.011 0.088 0.200 0.251 0.205 0.111 0.037 0.006 0.000 0
6     5 0 0.001 0.026 0.103 0.201 0.246 0.201 0.103 0.026 0.001 0
7     6 0 0.000 0.006 0.037 0.111 0.205 0.251 0.200 0.088 0.011 0
8     7 0 0.000 0.001 0.009 0.042 0.117 0.215 0.267 0.201 0.057 0
9     8 0 0.000 0.000 0.001 0.011 0.044 0.121 0.233 0.302 0.194 0
10    9 0 0.000 0.000 0.000 0.002 0.010 0.040 0.121 0.268 0.387 0
11   10 0 0.000 0.000 0.000 0.000 0.001 0.006 0.028 0.107 0.349 1
```