Project 5 Counting and Probability

by

Sri Ram Prabu

ALY6000 : Introduction to Analytics

February 14th, 2025

Problems:

1. Download the data set **ball-dataset.csv** and read it into your script. Each ball in the dataset is represented by a color (red, blue, green, or yellow) and a label (A, B, C, D, or E).

```
ball_data <- read.csv("ball-dataset.csv")
```

2. Create a frequency table as a data.frame or tibble that contains counts for each color of ball (**freq_color**).

```
# A tibble: 4 × 2
  color  counts
  <chr>   <int>
1 blue      307
2 green     192
3 red       408
4 yellow     93
```

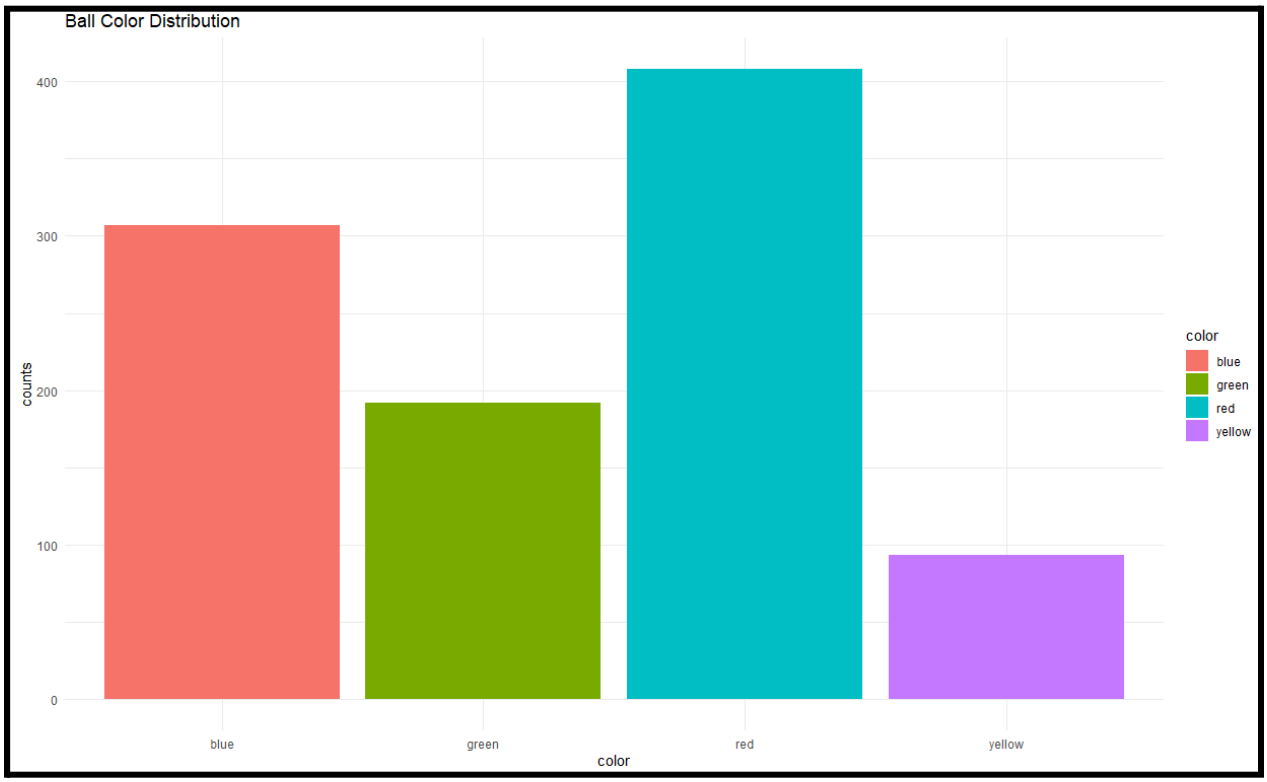
```
freq_color <- as.data.frame(table(ball_data$color))
names(freq_color) <- c("color", "counts")
```

3. Create a frequency table as a data.frame or tibble that contains counts for each label of ball (**freq_label**).

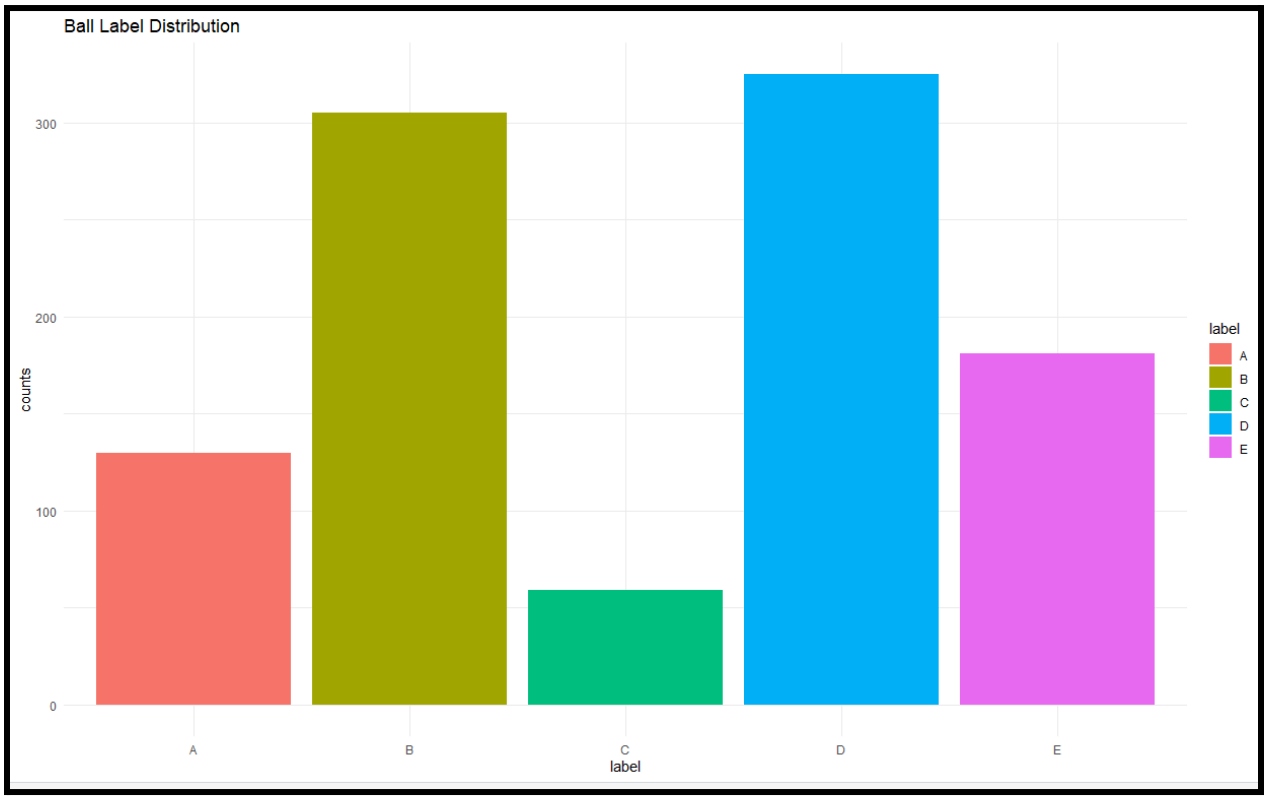
```
# A tibble: 5 × 2
  label counts
  <chr>  <int>
1 A        130
2 B        305
3 C         59
4 D        325
5 E        181
```

```
freq_label <- as.data.frame(table(ball_data$label))
names(freq_label) <- c("label", "counts")
```

4. Create a bar chart of the ball data set representing the counts of the different colors.



5. Create a bar chart of the ball data set representing the counts of the different labels.



6. What is the probability of drawing a green ball (**prob6_result**)?
7. What is the probability of drawing a blue or a red ball (**prob7_result**)?
8. What is the probability of drawing a ball with a label of A or C (**prob8_result**)?
9. What is the probability of drawing a yellow ball with a D (**prob9_result**)?
10. What is the probability of drawing a yellow ball or a ball with a D (**prob10_result**)?
11. What is the probability of drawing a blue ball followed by a red ball without replacement (**prob11_result**)?
12. What is the probability of drawing four green balls in a row without replacement (**prob12_result**)?
13. What is the probability of drawing a red ball followed by a ball with a B without replacement (**prob13_result**)?

```
prob6_result <- sum(ball_data$color == "green") / n_total

prob7_result <- sum(ball_data$color %in% c("blue", "red")) / n_total

prob8_result <- sum(ball_data$label %in% c("A", "C")) / n_total

prob9_result <- sum(ball_data$color == "yellow" & ball_data$label == "D") / n_total

prob10_result <- sum(ball_data$color == "yellow" | ball_data$label == "D") / n_total

n_blue <- sum(ball_data$color == "blue")
n_red <- sum(ball_data$color == "red")
prob11_result <- (n_blue / n_total) * (n_red / (n_total - 1))

n_green <- sum(ball_data$color == "green")
prob12_result <- (n_green / n_total) * ((n_green - 1) / (n_total - 1)) * ((n_green - 2) / (n_total - 2)) * ((n_green - 3) / (n_total - 3))

n_red <- sum(ball_data$color == "red")
n_b <- sum(ball_data$label == "B")
prob13_result <- (n_red / n_total) * (n_b / (n_total - 1))
```

```
> prob6_result
[1] 0.192
> prob7_result
[1] 0.715
> prob8_result
[1] 0.189
> prob9_result
[1] 0.036
> prob10_result
[1] 0.382
> prob11_result
[1] 0.1253814
> prob12_result
[1] 0.001324826
> prob13_result
[1] 0.1245646
```

14. [Challenge] Write a function called **my_factorial** that computes the factorial of a given number.
    - my_factorial (0) = 1
    - my_factorial(3) = 6
    - my_factorial(5) = 120

```
my_factorial <- function(n) {
  if (n == 0) return(1)
  return(prod(1:n))
}
```

```
> my_factorial(3)
[1] 6
> my_factorial(0)
[1] 1
> my_factorial(6)
[1] 720
```

15. Manually create a data.frame or tibble that contains all possible outcomes of flipping the coin four times (**coin_outcomes**).

```
# A tibble: 16 × 4
   first second third fourth
   <chr> <chr>  <chr> <chr>
 1 H     H      H     H
 2 H     H      H     T
 3 H     H      T     H
 4 H     H      T     T
 5 H     T      H     H
 6 H     T      H     T
 7 H     T      T     H
 8 H     T      T     T
 9 T     H      H     H
10 T     H      H     T
11 T     H      T     H
12 T     H      T     T
13 T     T      H     H
14 T     T      H     T
15 T     T      T     H
16 T     T      T     T
```

```
coin_outcomes <- expand.grid(first = c("H", "T"),
                             second = c("H", "T"),
                             third = c("H", "T"),
                             fourth = c("H", "T"))
```

```
> coin_outcomes
   first second third fourth probability
1      H      H     H      H           4
2      T      H     H      H           3
3      H      T     H      H           3
4      T      T     H      H           2
5      H      H     T      H           3
6      T      H     T      H           2
7      H      T     T      H           2
8      T      T     T      H           1
9      H      H     H      T           3
10     T      H     H      T           2
11     H      T     H      T           2
12     T      T     H      T           1
13     H      H     T      T           2
14     T      H     T      T           1
15     H      T     T      T           1
16     T      T     T      T           0
>
```

16. Compute the probability of each row outcome and store it as a column in the data.frame or tibble. You can do this manually or programmatically (**coin_outcomes**).

```
coin_outcomes$probability <- apply(coin_outcomes, 1, function(row) {
  sum(row == "H")
})
```

```
> coin_outcomes$probability
 [1] 4 3 3 2 3 2 2 1 3 2 2 1 2 1 1 0
```

17. There are 5 possible outcomes in our coin dataset if we count the number of heads in each row. For example, the row "H H H H" has 4 heads and the row "H T H T" has 2 heads. Compute the probability of each of the 5 possible outcomes (**num_heads_prob**).

```
num_heads_prob <- as.data.frame(table(coin_outcomes$probability) / nrow(coin_outcomes))
names(num_heads_prob) <- c("num_heads", "probability")
```

```
> num_heads_prob
  num_heads probability
1         0      0.0625
2         1      0.2500
3         2      0.3750
4         3      0.2500
5         4      0.0625
```

18. What is the probability of an outcome of three heads (**prob18_result**)?
19. What is the probability of an outcome of two heads or four heads (**prob19_result**)?
20. What is the probability of an outcome of less than or equal to three heads (**prob20_result**)?

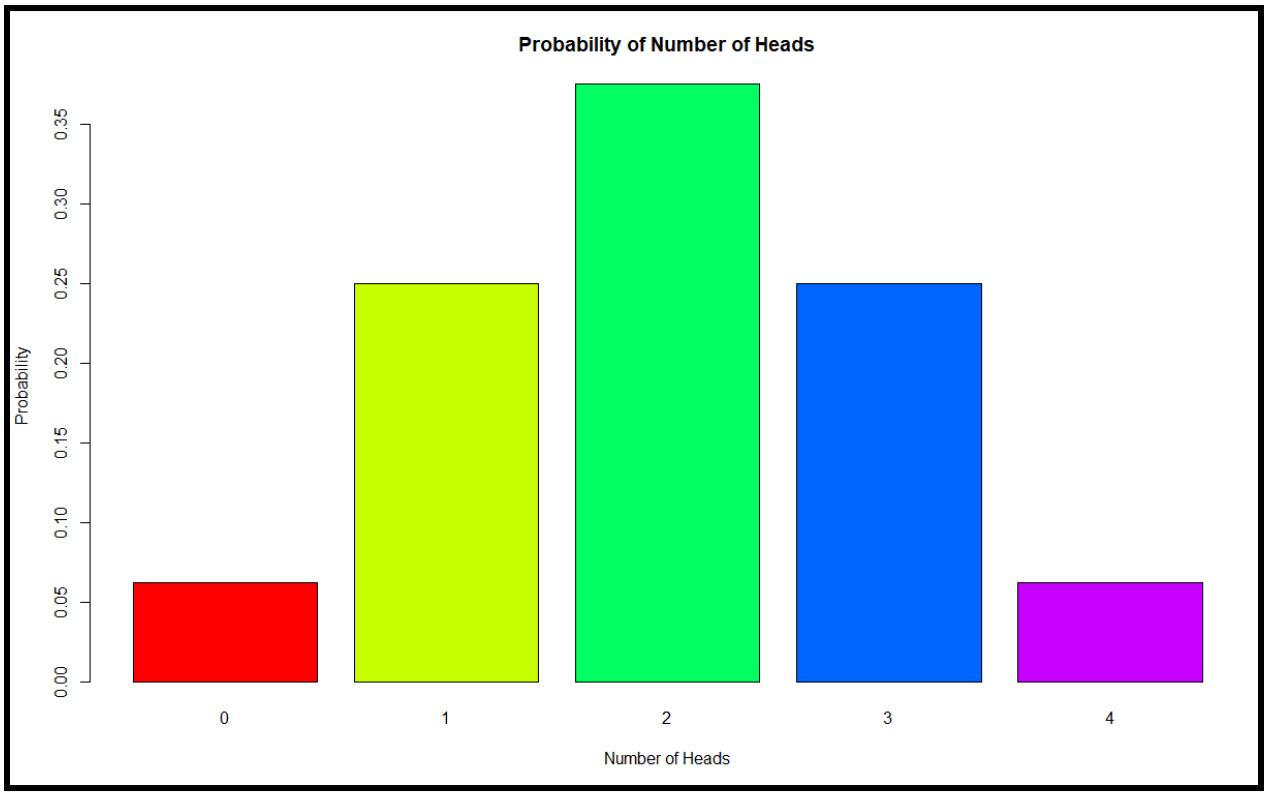
```
prob18_result <- sum(coin_outcomes$probability == 3) / nrow(coin_outcomes)

prob19_result <- sum(coin_outcomes$probability %in% c(2, 4)) / nrow(coin_outcomes)

prob20_result <- sum(coin_outcomes$probability <= 3) / nrow(coin_outcomes)
```

```
> prob18_result
[1] 0.25
> prob19_result
[1] 0.4375
> prob20_result
[1] 0.9375
```

21. Create a bar chart where the *x*-axis is the outcome, and the *y*-axis is the probability.



Soccer Games (Optional Challenge Problems)

22. What is the probability that they will win exactly 10 games (**prob22_result**)?
23. What is the probability that they will win more than one game (**prob23_result**)?
24. How many different ways could you pick five games at random and have three home games and two away games (**prob24_result**)?

```
prob22_result <- (0.75^5) * (0.50^5)

prob23_result <- 1 - ((0.25^5) * (0.50^5) + (5 * (0.75^1) * (0.25^4) * (0.50^5)))

prob24_result <- choose(5, 3)
```

```
> prob22_result
[1] 0.007415771
> prob23_result
[1] 0.9995117
> prob24_result
[1] 10
```

References :

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Ross, S. M. (2019). *Introduction to probability and statistics for engineers and scientists* (6th ed.). Academic Press.

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data.* O'Reilly Media.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the practice of statistics* (9th ed.). W. H. Freeman.

R Documentation. (n.d.). *Base R functions and probability calculations*. Retrieved from https://cran.r-project.org/manuals.html