V S N Sai Krishna Mohan Kocherlakota

Exploring Probability Distributions

28th October 2023

INTRODUCTION

In this project, I utilized a build different data set and used the complete knowledge gained from the subject. Tried to plot Q-Q plot and different visualizations. It was a very good start for my data analysis and learning the basics has given me a lot of confidence and knowledge.

KEY-FINDINGS

Part-1: - Baseball Data

1. `prob1_result <- dbinom(5, 7, win)`
   - This line uses the `dbinom` function to calculate the probability of having 5 wins out of 7 games with a win probability of 0.65. The result is stored in `prob1_result`.

3. `prob3_result <- 1 - sum(prob2_result$probability[6:8])`
   - This line calculates the probability of not winning 6, 7, or 8 games out of 7. It subtracts the sum of these probabilities from 1.

8. This sets the random seed to ensure reproducibility and then generates 1000 random samples from a binomial distribution with 7 trials (games) and a success probability of 0.65 (winning).

```
> #8
> set.seed(10)
> samples <- rbinom(1000, 7, win)
```

Part-2: - Call Center

11. `prob11_result <- dpois(x = 6, lambda)`

   - This line calculates the probability of receiving exactly 6 calls in a time with an average arrival rate of 7 calls per hour (lambda).

15. `prob15_result <- qpois(0.9, lambda*hours)`

   - This calculates the number of calls that corresponds to the 90th percentile of the Poisson distribution with an average arrival rate of 7 calls per hour (lambda) over 8 hours.

Part-3: - Life span of light bulbs

19. This code calculates the probability that a randomly selected light bulb's lifespan falls between 1800 and 2200 hours, assuming a normal distribution with a mean of 2000 and a standard deviation of 100 hours.

```
> #19
> z_upper <- (2200 - mean_life_span)/sd_life_span
> z_lower <- (1800 - mean_life_span)/sd_life_span
```

20. `prob20_result <- 1 - pnorm((2500 - mean_life_span)/sd_life_span)`

   - This code calculates the probability that a randomly selected light bulb's lifespan exceeds 2500 hours.

25. This code performs a bootstrap resampling simulation. It generates 1000 samples of 100 light bulbs each, calculates the sample means for each sample, and stores these means in `prob25_result`.

```
> #25
> set.seed(1)
>
> num_samples <- 1000
> sample_size <- 100
>
```

```
> prob25_result <- c()
>
> for (i in 1:num_samples) {
+
+   sample_data <- rnorm(sample_size, mean_life_span, sd_life_span)
+   prob25_result[i] <- mean(sample_data)
+ }
```

26.This code creates a histogram of the sample means to visualize the distribution of sample means obtained through the bootstrap resampling.

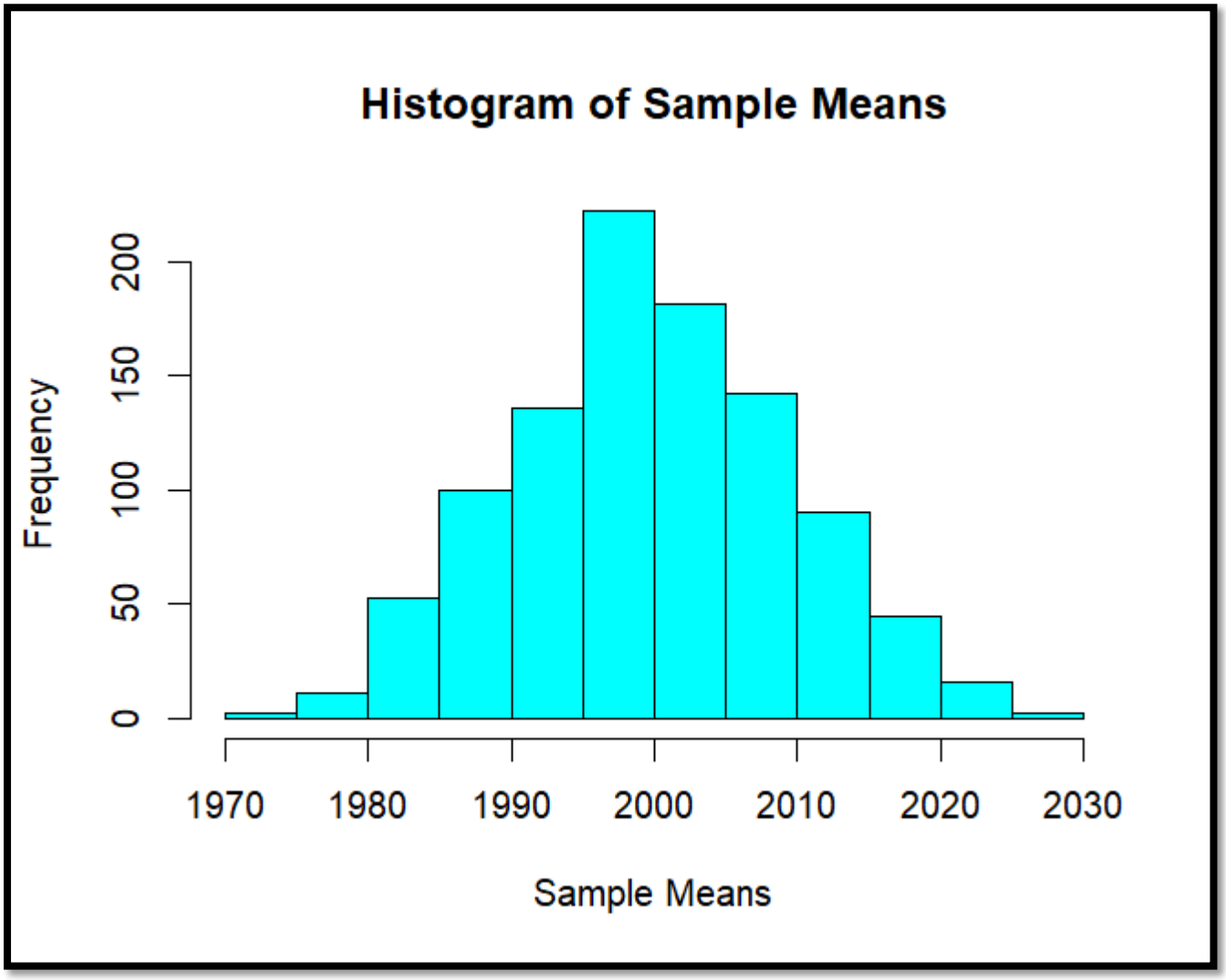

Figure1: - Histogram of Sample means

Part-4: - Flipper length of penguins

The R code uses the `palmerpenguins` dataset to perform various data visualization tasks using the `ggplot2` library. Here's a breakdown of what each part of the code does:

1. The code loads the `palmerpenguins` library and displays the first few rows of the penguins dataset using `head(penguins)`.

2. It filters the dataset to create a subset containing only Adelie penguins and assigns it to the variable `adelie`.

3. It then provides a summary of the `adelie` dataset using `summary(adelie)`.

4. The code generates a quantile-quantile (QQ) plot for the flipper length of Adelie penguins using `qqnorm` and adds a red line using `qqline` to assess the distribution's normality.

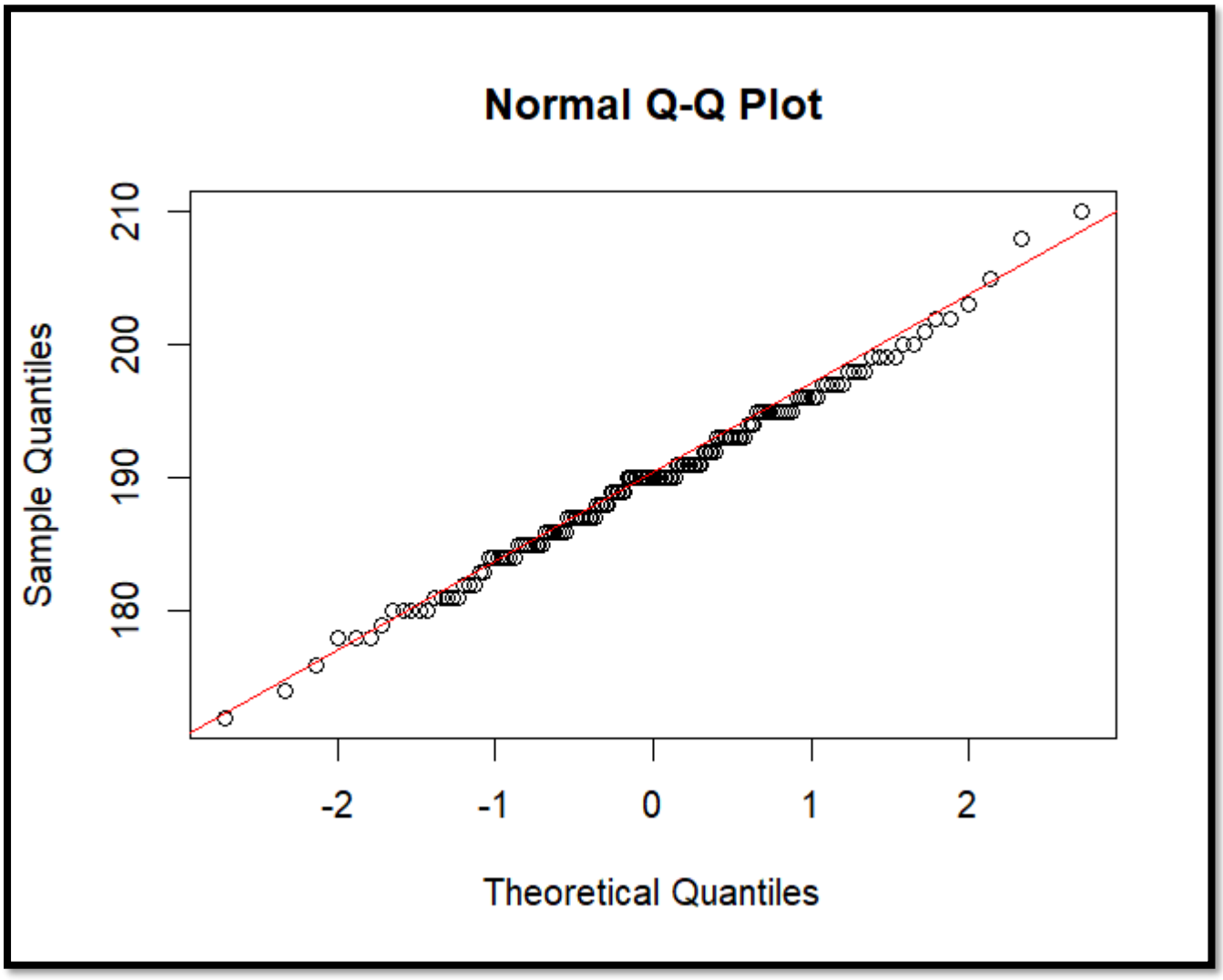

Figure2: - Q-Q Plot

5. It creates a histogram of the flipper length of Adelie penguins using `ggplot2`, specifying the binwidth, fill color, and labels. However, some lines are commented out that could be used to scale the y-axis.

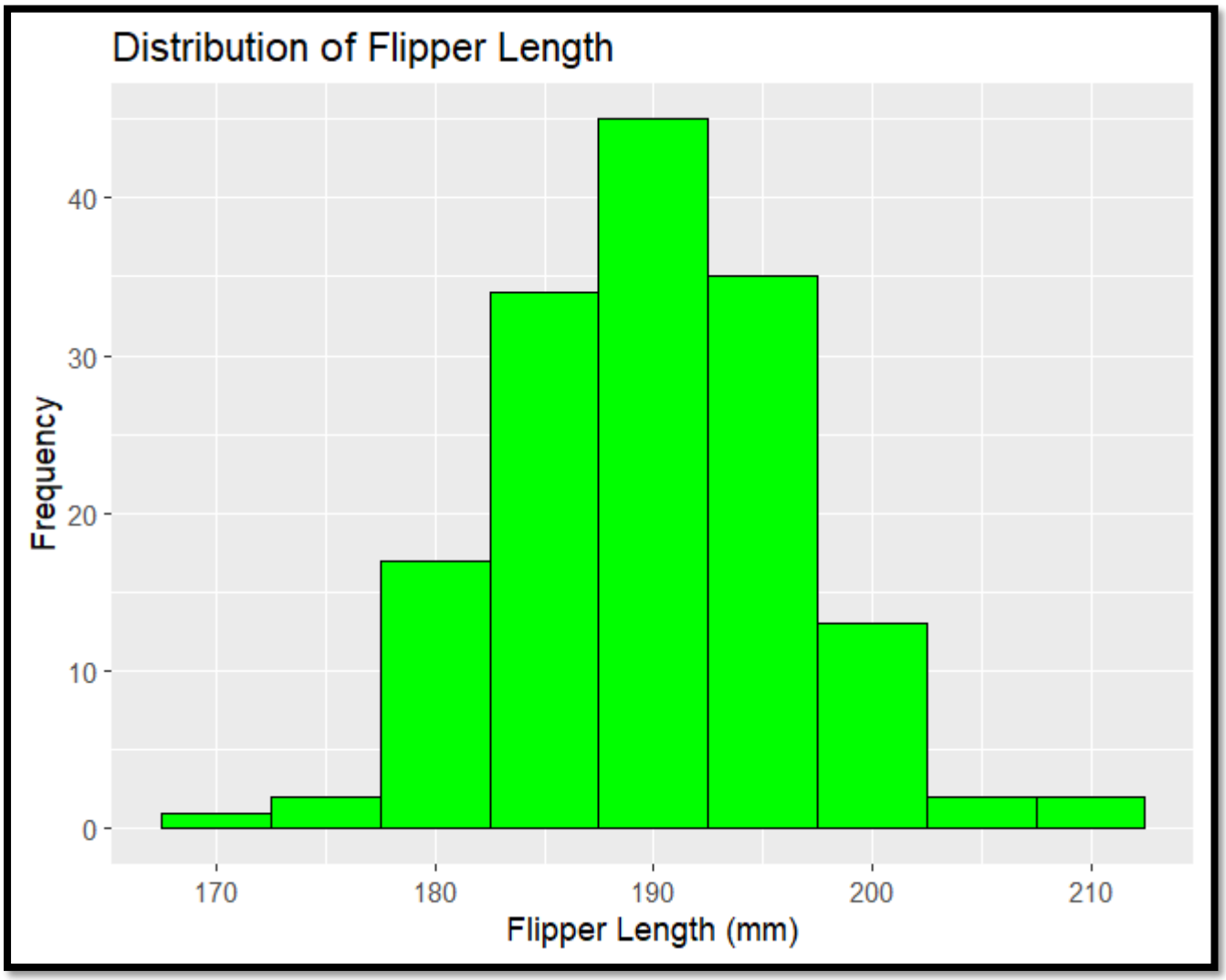

Figure3: - Distribution of Flipper length

6. The code filters the dataset to create a subset containing only Gentoo penguins and assigns it to the variable `gentoo`.

7. It creates a scatter plot between flipper length and bill depth for Gentoo penguins using `ggplot2`. It uses different colors for data points and adds a linear regression line (method = "lm") without a shaded area (se = FALSE). The plot is also given titles and labels.

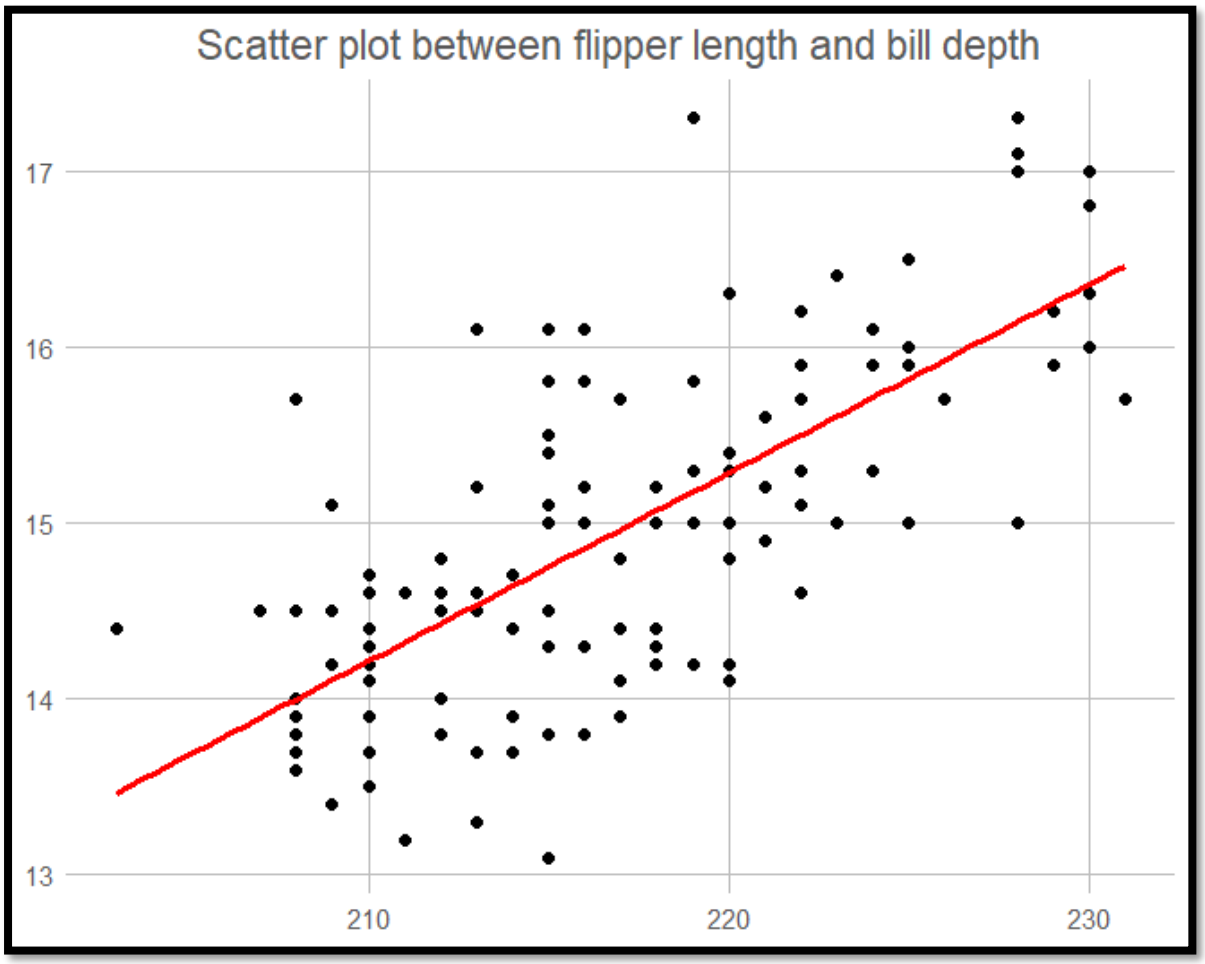

Figure4: Scatter plot

CONCLUSION

In summary, the provided code encompasses a comprehensive analysis of various scenarios, including baseball data, call center, lifespan of light bulbs and soccer games. Here are the key takeaways:

1. Baseball Data Analysis:
This code is a practical example of how to analyse and work with binomial probability distributions in the context of baseball games. It provides insights into the likelihood of different outcomes and the expected values and variability associated with these outcomes.

2. Call Center:
This code is useful for analysing and modelling call center activity, including assessing probabilities of different call volumes, percentiles, and statistical characteristics. The random sampling and calculations help gain insights into expected call volumes and their variability, which can be valuable for capacity planning and resource allocation in a call center setting.

3. Lifespan of light bulb:
This code is valuable for assessing various aspects of light bulb lifespans, including probabilities, percentiles, and estimates of central tendency and variability. The use of bootstrap resampling helps account for uncertainty in the estimates by repeatedly resampling from the data, providing insights into the distribution of sample means. The histogram visualization aids in understanding the distribution of sample means.

4. Flipper length of penguins:
In conclusion, the code demonstrates various data visualization techniques using ggplot2 and basic data exploration. It focuses on the characteristics of Adelie and Gentoo penguins, allowing for a visual examination of their flipper length, bill depth, and their distribution characteristics. These visualizations are essential for understanding and summarizing the data, making it easier to draw insights and make informed decisions based on the penguins dataset.

```
> test_file("project6_tests.R")
[ FAIL 0 | WARN 0 | SKIP 0 | PASS 24 ]
```

CITATIONS

1. **Module-6 Overview: -**
   Module 6 | Resources: ALY6000 70917 Introduction to Analytics SEC 19 Fall 2023 CPS [BOS-A-HY] (instructure.com)