Report for Module 3 R Practice

Group 10: Jiancong(Kevin) Zhu, Yate Zhang, Chris Wang, Yuchen Bi

1. Introduction

The purpose of this project is to analyze a dataset named "Hospital.csv," which contains information on 200 hospitals and various attributes, including geographical region, type of control, service provided, bed count, admissions, and other operational data. This project aims to calculate and interpret different probabilities using classical probability methods, such as marginal, joint, and conditional probabilities, in addition to Bayesian inference.

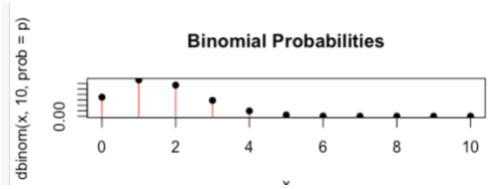
2. Exploratory Data Analysis (EDA)

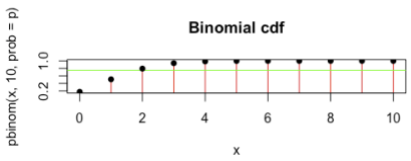
In the binomial distribution plot, the random variable was set to 10 trials with a success probability of 0.16 (calculated from the proportion of psychiatric hospitals in the dataset). The red bars represent the probabilities, and a cumulative distribution plot was also generated to examine the behavior of the distribution. Similarly, I also generated hypergeometric distribution plots, which considered the difference between sampling with replacement and without replacement. A final set of plots illustrated the Poisson distribution with a calculated average value (lambda) of 1.6.

```
x <- seq(0,10)
par(mfrow = c(2,1))
plot(x,dbinom(x, 10 , prob = p), type = "h", main = "Binomial Probabilities",
col = "red")
points(x,dbinom(x, 10 , prob = p), pch = 16, cex = 1)
```

And

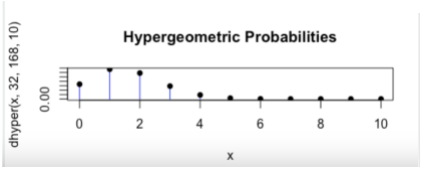
```
par(mfrow = c(2,1))
plot(x,pbinom(x, 10 , prob = p), type = "h", main = "Binomial cdf", col =
"red")
abline(h = 0.75, col = "green")
points(x,pbinom(x, 10 , prob = p), pch = 16, cex = 1)
```

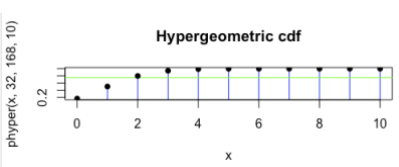




The red bars represent the probabilities, and the cumulative distribution plot examines the behavior of the distribution. Similarly, we generated hypergeometric distribution plots, which considered the difference between sampling with replacement and without replacement:

```
plot(x,dhyper(x, 32 , 168, 10), type = "h",main = "Hypergeometric
Probabilities", col = "blue")
points(x,dhyper(x, 32 , 168, 10), pch = 16, cex = 1)
par(mfrow = c(1,1))
plot(x,phyper(x, 32 , 168, 10), type = "h", main = "Hypergeometric cdf" , col
= "blue")
abline(h = 0.75, col = "green")
points(x,phyper(x, 32 , 168, 10), pch = 16, cex = 1)
par(mfrow = c(1,1))
```

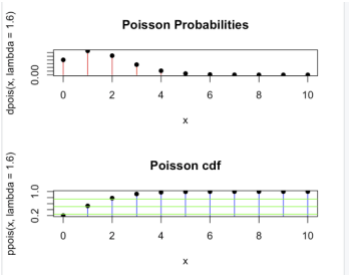




The final set of plots illustrated the Poisson distribution with a calculated average value (lambda) of 1.6:

```
# Using n = 10 and p = 0.16 from before we can choose lambda = 1.6 and
compare.

par(mfrow = c(2,1))
plot(x,dpois(x, lambda = 1.6), type = "h", main = "Poisson Probabilities",
col = "red")
points(x,dpois(x, lambda = 1.6), pch = 16, cex = 1)
plot(x,ppois(x, lambda = 1.6), type = "h", main = "Poisson cdf" , col =
"blue")
points(x,ppois(x, lambda = 1.6), pch = 16, cex = 1)
abline(h = c(0.25, 0.5, 0.75), col = "green")
par(mfrow = c(1,1))
```



These charts provided insights into the relationships between the hospital attributes and allowed for comparisons among different distributions to understand how each method describes the dataset. For example, the comparison between binomial and hypergeometric distributions highlights the subtle differences in probability when considering whether sampling is done with or without replacement.

3.Data Analysis Methodologies

The dataset analysis incorporated a variety of classical probability methods, as described below:

1. **Marginal, Joint, and Conditional Probabilities**: These calculations aimed to uncover basic insights about the hospital dataset, including marginal and joint probabilities based on hospital type, region, and other characteristics. For instance, we computed the marginal probability that a hospital is classified as "for-profit" (22.5%) using the following code:

```
> mcr[3,8]/mcr[5,8]
[1] 0.225
```

Conditional probabilities were also calculated, such as the probability that a hospital in the Midwest is a for-profit facility:

```
> (mcr[5,5]+mcr[2,8]-mcr[2,5])/mcr[5,8]
[1] 0.485
```

Conditional probabilities were also calculated, such as the probability that a hospital in the Midwest is a for-profit facility:

```
> mcr[3,3]/mcr[3,8]
[1] 0.2444444
```

2. **Bayes' Theorem**: Bayes' Theorem was used to derive posterior probabilities, such as the probability that a hospital is a for-profit hospital given that it is in the Midwest. These calculations provide insight into the relationship between region and type of hospital control. The code below calculates the intersection and marginal probabilities to apply Bayes' Theorem:

```
> numerator <- (mcr[3,8]/mcr[5,8])*(mcr[3,3]/mcr[3,8])
>
> # The numerator is the intersection probability of MW and for-profit.
>
> marginal <- mcr[,8]/mcr[5,8]
> conditional <- mcr[,3]/mcr[,8]
> intersection <- marginal*conditional
> numerator/sum(intersection[1:4])
[1] 0.1833333
```

4. Conclusion

The analysis provided significant insights into hospital characteristics, including the probability distribution of different types of hospitals, and the impact of regional factors. The calculated probabilities helped reveal the relationships between different attributes of the dataset, such as control type and region. For instance, the probability that a hospital is for-profit given its region is an important insight that indicates regional disparities in hospital control.

The distribution analysis, using binomial, hypergeometric, and Poisson distributions, highlighted the differences between sampling methodologies and how they influence the interpretation of data. These distributions are critical in assessing rare events and understanding the differences between independent trials and dependent events in sampling.