Report for Module 4 R Practice

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

1. Introduction

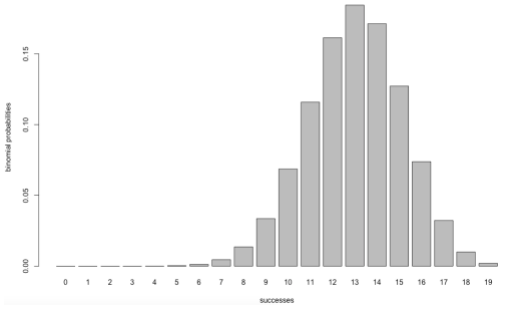
The goal of this project was to analyze the behavior of binomial distributions with different sample sizes and determine when it can be approximated by a normal distribution with minimal error. Using RStudio, we ran simulations for binomial distributions with varying numbers of trials to better understand the conditions for normal approximation. We examined cases with few, moderate, and many trials and used visualizations to illustrate these differences. The focus was to determine how the sample size influences the distribution symmetry and the feasibility of applying a normal approximation.

2. Exploratory Data Analysis

Several charts were generated using RStudio to visualize binomial distributions and evaluate the conditions for normal approximation. Below are the major charts and their findings:

- **Barplot for 20 Trials**: We set the number of trials to 20 with a success probability of 0.65. The condition for normal approximation (size * prob > 5, size * (1 - prob) > 5) was satisfied, but the distribution appeared skewed to the left. The barplot for this distribution was generated using the following R code:

```
# set trials <-c(0:20) and calculate dbinom(q,size=20,prob=0.65)
par(mfrow=c(1,1))
trials <-c(0:20)
trials
size = 20
prob=0.65
# check conditions for normal approximation
# size*prob > 5, size*(1-prob)>5
probabilities1 <-dbinom(trials,size=20,prob=0.65)
probabilities1
successes1 <-trials[1:20]
successes1
binomial_probabilities1 <-probabilities1[1:20]
successes1 <-factor(successes1)
successes1
barplot(binomial_probabilities1, names.arg = successes1, xlab = "successes", ylab = "binomial probabilities")
```



  This plot demonstrated that although the requirements for normal approximation were technically met, the distribution still exhibited left skewness, showing that it was at the threshold of approximation feasibility.
- **Pushed-to-Left Barplot for 20 Trials**: By excluding the first few bars of the initial plot (successes 0-6), we obtained a more symmetric histogram, making it easier to approximate the distribution using a normal curve. The corresponding code used was:

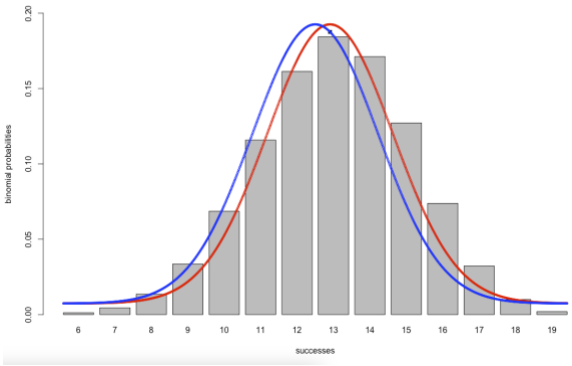
```
successes1 <-trials[7:20]
successes1
binomial_probabilities1 <-probabilities1[7:20]
successes1 <-factor(successes1)
successes1
barplot(binomial_probabilities1, names.arg = successes1,ylim=c(0,0.20), xlab = "successes", ylab = "binomial probabilities")
```

- **Normal Approximation Curves**: We plotted normal approximation curves over the barplot to evaluate the quality of approximation with and without continuity correction. The code snippet used for the normal curves was:

```r
# since we need to create now a normal curve, we need to create a vector for the continuous variable
# Z - think about the z-scores (from -3.49 to 3.49) and we will calculate for each of those the corresponding X
# using the formula Z = X-mean/sd , solve for X = Z*sd+mean
# this is the vector of the normal distributed variable X with mean = 13 and sd = 2.13
x <- seq(-4,4,length=1000)*sd + mean

# now we will create the density function for each of those values of x
# Note that this integral does not exist in a simple closed formula. It is computed numerically.
# Let's see an example: What does the following mean?
# dnorm(13,mean,sd) we will see in a moment

hx.red <- dnorm(x,mean+0.5,sd)  # Normal Curve with correction
hx.blue <- dnorm(x,mean,sd)     # Normal Curve without correction is shifted to the left and not a good approximation
par(new=TRUE)  # this allows us to graph on top of the previous graph
plot(x,hx.red, col = "red",axes = FALSE, xlab = "", ylab = "", cex=0.45)
points(13+0.5,dnorm(13,mean+0.5,sd),pch=15)
par(new=TRUE)
plot(x,hx.blue, col = "blue",axes = FALSE, xlab = "", ylab = "", cex = 0.45)
```



These curves demonstrated how the normal approximation is influenced by the correction factor. The red curve (with correction) was more aligned with the barplot, indicating a better approximation.

3. Data Analysis Methodologies

We employed binomial and normal distribution calculations to analyze the data. Various sample sizes were used to determine how closely the binomial distribution could be approximated by a normal curve. For smaller trials, such as 5 and 10, the distribution was highly skewed and did not resemble a normal distribution. For larger trials, such as 200, the distribution became much more symmetric, fulfilling the requirements for normal approximation.

The following comparisons were made using RStudio:

- **Small Sample Size (5 Trials)**: The graph was not symmetric, and the conditions for normal approximation were not satisfied (size * prob < 5). Thus, a normal approximation would be highly inaccurate.
- **Moderate Sample Size (10 Trials)**: The distribution was less skewed but still did not achieve symmetry. The conditions were borderline, suggesting that a normal approximation could be applied, but with significant error.
- **Large Sample Size (200 Trials)**: The distribution was symmetric and satisfied the conditions for normal approximation. Both the blue and red normal curves closely followed the barplot.

4. Conclusion

From this analysis, we can conclude that the feasibility of using a normal approximation for binomial distributions depends significantly on the number of trials and the probability of success. When the number of trials is small, the binomial distribution is skewed, making the normal approximation unsuitable. As the sample size increases, the distribution becomes more symmetric, meeting the conditions for applying a normal approximation effectively. For 200 trials, the approximation error is minimized, suggesting that larger sample sizes are ideal for applying normal distribution assumptions.