PREDICT 401 Introduction to Statistical Analysis Data Analysis Assignment #2

Introduction and Objectives:

Contingency tables play an important role in the analysis of discrete variables. This assignment addresses computing probabilities using contingency tables. Also addressed is the use of various discrete and continuous probability functions for problem solving. Since the normal distribution is used to approximate binomial probabilities, the rate of convergence of the binomial distribution to normality will be studied. Different sample size rules will be compared for using the normal distribution to calculate binomial probabilities.

Overview:

There are four parts to the assignment. The first part is a demonstration showing how to construct a contingency table in R, index table elements and calculate probabilities. The second part of the assignment requires calculation of probabilities using a contingency table. The third part requires use of probability functions available with R, both discrete and continuous. This part includes a comparison of exact binomial probabilities versus those calculated using a normal approximation. The fourth part is an investigation of binomial distribution convergence to normality. R is used throughout the assignment. The results are needed to answer quiz questions. Follow instructions carefully. **Before starting be sure to watch the video that pertains to this assignment.**

Supplemental References:

1) Lander, *R for Everyone* pages 185-186.
2) Chihara and Hesterberg, *Mathematical Statistics with Resampling an d R.* pages 87-91.
3) Verzani, Using R for Introductory Statistics, pages 85, 135-136.

Part 1 Table Construction and Probability Calculations

This part of the assignment will use the data in "Hospital.csv". Refer to Black *Business Statistics* page 15 for the data dictionary, and Chapter 4 page 140 problem 2. Use the code in Appendix A of this assignment. Review the problem and execute the code. Comment statements document the program. The table that's generated will be used in Part 2.

Part 2 Probability Calculations

Refer to Black *Business Statistics* Chapter 4 page 140 problem 2, Chapter 5 page 180 problem 2 and Chapter 6 page 220 problem 3. Answer the questions in these problems. The table constructed in Part 1 will be needed. Use library functions dbinom(), dhyper() and pexp(). Lander, *R for Everyone* pages 185-186 lists various library functions. If you have questions, for example with pexp(), type ?pexp() into the console for information. The results of these calculations will be needed for the quiz.

Part 3 Comparison of Probability Calculations

Refer to Black *Business Statistics* Chapter 7 page 254 problem 3. This problem will require the table constructed in Part 1. Assume the binomial distribution can be used without a finite population correction. Complete the following calculations:

1) Determine the probability (using data from Part 1), and calculate the exact result using pbinom(). (Note that pbinom() does not include 225 in the upper tail unless it is started at 224.) Use the function pnorm() with continuity correction to approximate this probability.
2) For the last calculation, subtract 1 to start the pbinom() calculation at the right point for the lower tail. (i.e. start at 39 and request the lower tail). Determine the exact binomial probability and also use the normal approximation with continuity correction to estimate the probability.

Part 4 Study of Distributional Convergence

Refer to Black *Business* Statistics Section 7.3 pages 246-248. When comparing a sample proportion to an assumed proportion p, a common practice is to compute a z value. This number is then compared to the standard normal distribution to determine if the sample proportion differs in a statistically significant way from p. This calculation is justified by the central limit theorem, but it must be remembered using the normal distribution in this manner is an approximation dependent on the sample size and the probability p. Central limit theorem convergence with p close to 0 or close to 1 can require a very large sample size.

Assume random samples are obtained from a binomial distribution with probability p and sample size n. As n becomes larger, the z-value $(\hat{p}-p)/\sqrt{(p(1-p)/n}$, will have a distribution which approaches the normal distribution. For example, with a one-sided test, a z-value of 1.644854 corresponds to a normal distribution tail probability of 5%. Using binomial probabilities based on n and p, it is possible to compute the exact binomial tail probability and compare it to the nominal 5% level for a one-sided test. Doing so allows for an assessment of the adequacy of the sample size n.

There are different sample size rules in use for determining a **minimum** sample which justifies use of the normal approximation. Part 4 of this assignment will investigate three different, commonly used rules mentioned in the literature. They are: 1) both np and n(1-p) >= 5, 2) both np>=9(1-p) and n(1-p)>=9p, and 3) np(1-p) >=15. The first rule is used in the text by Black and the text by Triola. The strengths and limitations of these rules are rarely if ever discussed in detail.

There are a number of reasons for this portion of the assignment. First, alternative sample size selection rules appear in many text books without discussion of their strengths and weaknesses. Second, alternative rules do not perform in the same manner as a function of p. Third, it is useful to be aware of the unusual oscillations that occur with binomial tails probabilities as the sample size changes. In summary, this portion of the assignment aims to inform the student on these issues.

Appendix B provides a program which uses the binomial distribution to calculate tail probabilities as a function of n and p. Computational results are plotted for assessment. The program is designed so that different values of the binomial probability p may be used. For this assignment, execute the program using values of p equal to 0.025, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5. Values for the sample size n will be automatically determined and used. To use the program, substitute each value for p and observe the plot of exact calculated tail probabilities versus sample size. The plot will appear jagged due to the discrete nature of the binomial distribution.

The graph has red horizontal lines above and below the 5% level. The red lines are at 2.5% and 7.5%. For data analysis purposes, **it is highly desirable** to have estimated tail probabilities within the red horizontal lines. This means the sample size is adequate in providing a suitable tail probability. Having calculated tail probabilities within these limits will be considered acceptable for this assignment.

Three vertical lines also appear. The black line is set at a sample size of n=9(1-p)/p. The green line is set at n=5/p. The purple line is set at n=15/p(1-p). These vertical lines represent what each rule projects a minimum sample size should be for the given probability p. The question remains, are the projections adequate? How close do they come to what would be computed using exact binomial probabilities. That is what this assignment is intended to reveal by filling in the table below.

Questions:

- For each p, what minimum sample size is required based on the exact binomial calculation?
- For each p, what minimum sample size is produced by each rule?
- For each p, which rules meet or exceed the minimum sample size required based on the exact binomial calculation?

Rules

1) np >= 5 and n(1-p) >=5
2) np>=9(1-p) and n(1-p)>=9p
3) np(1-p)>=15

p	n	Rule 1	Rule 2	Rule3	Result
0.5					
0.4					
0.3					
0.2					
0.1					
0.05					
0.025					

Procedural steps:

1. Identify for each probability the smallest sample size after which the plotted jagged line of binomial probabilities always stays within the red boundaries. This is the value for n.
2. For each rule enter the sample size n by calculating it directly from the equation. This number is automatically supplied by the assignment program.
3. Compare the sample size results. The preferred rule is the one that produces a minimum sample size which is i) equal to or greater than n (i.e. to what is determined from the exact binomial calculations), and ii) closer to n than the other rules.
4. Be prepared to consider questions such as:
   a) Is one rule preferred for the seven probabilities compared to the other two?
   b) For each value of p which rule is best?
   c) Do the rules generally agree with each other?
   d) If you had to pick one rule, which would it be?
   e) How do the rules compare if p = 4/9?

After completing the assignment be sure to take the quiz. It is possible to exit from the quiz (without submitting) and return after reviewing results in R.

```
# Predict 401 Data Analysis Project #2
# Problems drawn from Analyzing the Databases in Black chpts 4,5,6,7

#----------------------------------------------------------------------
# Part 1
# Appendix A
#----------------------------------------------------------------------

hospital <- read.csv(file.path("c:/RBlack/","Hospital.csv"),sep=",")
str(hospital)

# Page 15 of Black has a hospital data dictionary.
# Chapter 4 page 140 problem 2.

# To generate table with margins, it is necessary to convert the variables to factors.
# In this case, it is equivalent to generating nominal variables for table construction.
control <- factor(hospital$Control)
region <- factor(hospital$Geog..Region)
control_region <- table(control, region)

# Check the structure and print out the table.
str(control_region)
control_region

# Add marginal totals and rename for simplicity.  Print the table.
# The table frequencies can be indexed by row and column.
m_c_r <- addmargins(control_region)
m_c_r
```

```
# Use of labeling with factors.
control <- factor(hospital$Control, labels = c("G_NFed","NG_NP","Profit","F_GOV"))
region <- factor(hospital$Geog..Region, labels = c("So","NE","MW",'SW',"RM","CA","NW"))
control_region <- table(control, region)
addmargins(control_region)

# The following calculations are for problem 2.

# Probability hospital is in Midwest if for-profit?
m_c_r[3,3]/m_c_r[3,8]

# Probability hospital is government federal if in the South?
m_c_r[1,1]/m_c_r[5,1]

# Probability Rocky Mountain or NP Government?
(m_c_r[5,3]+m_c_r[2,8]-m_c_r[2,3])/m_c_r[5,8]

# Probability for-profit in California?
m_c_r[3,6]/m_c_r[5,8]

# Chapter 5 page 180 problem 2-----------------------------------------

# Breakdown of hospitals by service: general hospital=1, psychiatric=2.
# Create a factor out of Service and form a table.
service <- factor(hospital$Service, labels = c("medical", "psychiatric"))
service <- table(service)
addmargins(service)

#----------------------------------------------------------------------
# Part 4
# Appendix B
#----------------------------------------------------------------------
# Evaluation of sample size selection rules.
# Exact probability calculation.

p <- 0.05  #  This is where different probabilities may be substituted.
sample_size <- numeric(0)
tail_prob <- numeric(0)

for (i in 1:80)   # Changes to 80 can lengthen or shorten the x-axis.
{N <- i*5         # Steps of 5 are being used.
 Np <- N*p
  sample_size[i] <- N
  x <- Np+ 1.644854*sqrt((N*p*(1-p)))
  tail_prob[i] <- pbinom(x, size = N, prob = p, lower.tail = FALSE, log.p = FALSE)}
```

```
N_size1 <- 5/p
N_size2 <- 9.0*(1-p)/p
N_size3 <- 15/(p*(1-p))
N_size1
N_size2
N_size3

plot(sample_size, tail_prob, type = "b", col = "blue", ylim = c(0, 0.125),
     main = "Exact")
abline(h = 0.05)
abline(h = c(0.025, 0.075), col = "red")
abline(v = N_size1, col = "green")
abline(v = N_size2, col = "black")
abline(v = N_size3, col = "purple")
legend("topright", legend=c("green is np >= 5","black is np >= 9(1-p)", "purple is np(1-p) >= 15"))
```