Computer Assignment 3

Title:

Sampling with and without replacement the Binomial and Hypergeometric random variables

with

The Poisson random variable as a limiting case of the Binomial random variable.

Introduction

There are occasions when addressing a problem in elementary probability where it is not clear to the beginner whether a problem is an instance of the binomial random variable or the hypergeometric random variable (RV). Some of this can be attributed to learning new and unfamiliar concepts. These are two distinct RV; hence, there should be no confusion. Further investigation will reveal that there is an overlap or common ground between the two probability mass functions (pmf) for the RVs that when understood will help to eliminate uncertainty as to which RV applies in a given problem.

The reader has been exposed to the theory of both the binomial and hypergeometric RV. Also, the reader is aware that the mean and variance of a RV can be obtained by several different approaches: application of the definition of expectation, moment generating functions, and etc. The formulas describing the preceding are well established. What follows then is an attempt to show that the two RV are more closely related than the reader may have supposed.

Arguably this relationship is best seen with the assistance of computer programs and graphics. So, in addition to the mathematical exposition below computer programs are provided. The reader is encouraged to improve upon these programs. Running these programs to visualize the behavior of the random variable's (pmf) will provide the reader with both visual and numerical evidence of how the two RVs behave in similar ways under certain circumstances.

Theory

Suppose that a bin contains N blocks, G of which are green and R of which are red. So $G + R = N$. Consider the random experiment of selecting n blocks. There are two procedures for doing this. First, the blocks can be sampled **with replacement**; this means that after a block is selected, its color is noted, the block is replaced into the bin, and the blocks are mixed. Thus the composition of the bin *remains the same* from selection to selection. The second way to choose them is to sample **without replacement**. This means that blocks are *not* replaced after they are selected.

Let the random variable Y be *the number of green blocks chosen in n drawings.* In sampling *with* replacement the selections are independent from one to next and

$$p = P(\text{green block chosen}) = \frac{G}{N}$$

At any one selection trial. Consequently, Y is binomially distributed in sampling with replacement.

In sampling without replacement successive selections are *not* independent; if a green block is chosen at the first draw, the composition of the bin is altered, so the probability of choosing green at the next draw is changed. Using combinatorial ideas the event $\{X = j\}$ means that among n selections, j are green and $n - j$ are red. Thus j selections must be made from the G green blocks and $n - j$ must be made from the R red blocks. There are ${}_NC_n$ samples of size n from all N blocks. The number of ways to select j green blocks is ${}_GC_j$; the number of ways to choose the remaining $n - j$ blocks from the set of R red ones is ${}_RC_{n-j}$. Consequently, the number of ways to choose j green *and* $n - j$ red blocks is the product

$$ {}_{G}C_{j} \cdot {}_{R}C_{n-j}. $$

The set of values that j can assume is restricted, however. Not only must $0 \leq j \leq n$, but the number of green blocks j cannot exceed the total number G of green bocks; similarly, the number of red blocks $n - j$ cannot exceed the total number R of red blocks. Therefore,

$$ j \leq G \text{ and } n - j \leq R. $$

Probability and Stochastic Processes by Fredrick Solomon pages 98 and 99

Instructions: Do the two exercises below by hand similarly as was asked of the reader in computer assignment 2.

Exercises

A box of 100 ornamental light bulbs contains 40 green and 60 red bulbs. Four are selected at random. Find the probability that three are red, assuming that the sampling is done (a) with replacement and (b) without replacement.

In a Klingon prison there are 40 Vulcans and 60 Humans. The Klingons select 4 at random to be executed. Let the random variable be the number of Vulcans in the group of four to be executed. With the random variable taking on values from zero to four construct two probability tables one using the hypergeometric model and the other using the binomial model. Compare the results of the two approaches.

Instructions: Two programs have been provided. One of the programs outputs the pmf for a binomial RV. The other program outputs the pmf for a hypergeometric RV. Combine these two programs into a single program and include a numerical error. The numerical error is the difference between the valued of the pmf. This error will tend towards zero as the two pmfs become more alike. Run it with at least the values stated. Visually compare the outputs. In which case(s) do the two distributions appear to be equivalent? In these case(s) what is the disparity in their numerical output?

Computer Programs

Binomial vs Hypergeometric

Run your binomial program with 25 and 100 trials. Make bar plots for probabilities of success equal to 0.1, 0.15, 0.25, and 0.5.

Run your hypergeometric program with at least the following suggested amounts in the table.

We are viewing the number of greens as the number of success. The size samples will play a role in making the pmf more or less like that for the binomial pmf.

Green	Red
3	17
5	15
10	30
25	75
50	50

Poisson

A computer program that plots the Poisson probability mass function has been provided. The instructions of what to do with it in this computer assignment are provided in the comments in the file.

Deliverables

The two exercises solved by hand.

A copy of your Python program of the combined binomial and hypergeometric programs. The output from the program showing when the pmfs diverge and when the pmfs merge. The output includes the numerical error between the two pmfs.

The run of the Poisson program with the outputs and the questions answered. (Put the answers in the PDF not in the Python file.)

Rubric

Name and I.D. # Name of Assignment Submission Date	Not Satisfactory: Data requested absent.	Satisfactory: Has all data requested.
Exercise 1.	Not Satisfactory: Missing steps in solution. Answer missing.	Satisfactory: Shows steps in solution towards final answer.
Exercise 2.	Not Satisfactory: Missing steps in solution. Answer missing.	Satisfactory: Shows steps in solution towards final answer.
A block comment at beginning of the combined program summarizing it. Line comments throughout the program explaining its operation.	Not Satisfactory: Conveying incomplete thoughts.	Satisfactory: A brief two or three sentence explanation at the beginning. Sufficient line comments explaining the code.
List references.	Not Satisfactory: No references listed.	Satisfactory: References – assignment handout, internet, students, and etc. listed

In the PDF submitted solutions to exercises 1 & 2, a copy of the .py file of the combined programs that compare the pmfs, the outputs, also, the Poisson outputs with the answers to questions to dropbox and a .py of the combined program to dropbox.	Not Satisfactory: Absent solutions and answers to exercises 1 & 2. Absent outputs. Absent copy of program.	Satisfactory: Solutions to exercises 1 & 2, a copy of the combined Python program file, the output all in the PDF, also, the Poisson output with questions answered and the .py file. These two complete files are submitted to dropbox.

Appendix 1

Random Variables and Distributions

Binomial

A finite number of trials

Independent trials

Two outcomes

The probability of success p is constant (this implies sampling with replacement)

The random variable Y is the number of successes

$P(\{Y = y\}) = {}_nC_y \; p^y q^{(n-y)}$ where $q = 1 - p$ and $\mu = np \;\; \sigma = \sqrt{npq}$

Hypergeometric

A finite number of trials

Two outcomes

The probability of success p is ***not*** constant (this implies sampling without replacement)

Population size is known

The random variable X is the number of successes

$P(\{X = j\}) = \frac{{}_GC_j \;\; {}_RC_{n-j}}{{}_NC_n}$ where G+R = N

Poisson

The occurrences must be random,

The occurrences must be independent of each other,

The occurrences must be uniformly distributed over the interval being used.

$P(\{X = x\}) = e^{-\lambda}\frac{\lambda^x}{x!}$ where $x = 0, 1, 2, \ldots$

Appendix 2

Mean and variance relationship between binomial and hypergeometric distributions.

As the reader is aware a random variable X is said to have a *hypergeometric probability distribution* if and only if

$$p(\{X = j\}) = \frac{\binom{r}{j}\binom{N-r}{n-j}}{\binom{N}{n}}$$

where j is an integer $0, 1, 2, \ldots, n$, subject to the restrictions $j \leq r$ and $n - j \leq N - r$.

Further there is a relationship between the binomial and hypergeometric that may not be familiar to the reader with respect to the mean and standard deviation.

If X is a random variable with a hypergeometric distribution.

$\mu = E(X) = \frac{nr}{N}$ and $\sigma^2 = V(X) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$.

Although the mean and the variance of the hypergeometric random variable seem to be rather complicated, they bear a striking resemblance to mean and variance of a binomial random variable. Indeed, if we define $p = \frac{r}{N}$ and $q = 1 - p = \frac{N-r}{N}$,we can then express the mean the mean and variance of the hypergeometric as $\mu = np$ and

$$\sigma^2 = npq\left(\frac{N-n}{N-1}\right).$$

You can view the factor

$$\frac{N-n}{N-1}$$

in $V(Y)$ as an adjustment that is appropriate when n is large relative to N. For fixed n, as $N \to \infty$,

$$\frac{N-n}{N-1} \to 1.$$

Mathematical Statistics with Applications by Wackerly, Mendenhall, and Scheaffer 5th Ed.