Lab 3: Probability and a Computer Simulation of Drawing Tickets from Boxes

OBJECTIVES: In the first half of the lab, we will be drawing candies out of a bag to look at various probability schemes. In the second half of the lab, we will perform a computer simulation of an experiment concerned with drawing tickets with replacement from a box.

DATA: The data for the first part of the lab will be constructed by the students. The data for the second half will be simulated (constructed) by Systat.

DIRECTIONS: Divide into groups (the lab instructor will determine how many groups). Read the instructions that follow and answer the questions below. Your answers should be written neatly and turned in to the lab instrutor.

1. Suppose you have 30 chocolate kisses, **17** of which are **silver** and **13** of which are **red**. Consider the (theoretical) situation where you draw one kiss from the bag, and then you draw another one. The important ideas in this situation are that: (i) you do not replace the first kiss before you draw the second one, and (ii) you draw each time from the bag in such a way that every kiss in the bag is equally likely to be selected (the idea of a random sample). In questions a-c and e-g, you will be asked to calculate probabilities associated with outcomes of this theoretical situation, and in questions d and h you will be asked to verify these probabilities experimentally.

   a. What is the probability that the wrappers on the two kisses are the same color?
   b. Different colors?
   c. What is the probability of choosing at least one red kiss?
   d. Now count out 17 silver kisses and 13 red kisses and put them in the bag. Draw a kiss at random, record its color. Without replacing the first kiss, draw a second kiss at random. Record its color. Repeat this procedure 30 times. Are your results close to the theoretical answers you found in a-c above? Note: as an example, if I count that there were 15 times where I drew 2 kisses of the same color, then my empirical estimate for the probability in question 1 would be $15/30 = 0.50$.
   e. Now suppose there are **24 silver** and **6 red** kisses in the bag. What is the probability of obtaining a match?
   f. What is the probability that the kisses do not match?
   g. What is the probability of obtaining at most one silver kiss?
   h. Now count out 24 silver kisses and 6 red kisses and put them in the bag. Repeat the procedure of drawing kisses as before. Again, do it 30 times in all. Are your results close to the theoretical answers you found in e-g above?
   i. Note how the probability of obtaining a match changed as you increased the imbalance between the number of kisses that were silver and the number that were red. How would you *minimize* the probability of obtaining a match (of course, assume that the total number you start with is even)?

2. For the second part of this lab, we are going to use Systat to simulate the results of the following experiment:

   Experiment: Suppose that a box contains 4 tickets. These 4 tickets have the numbers 1, 3, 3, and 9 written on them. Our experiment is to draw a ticket from the box, record the result, and then replace the ticket (we assume that each of the 4 tickets is equally likely to be drawn). Then we draw another ticket from the box, record the result, replace that ticket, and so on. We continue until we have done this 100 times. Then we let Y be the sum of the 100 numbers which we recorded. We want to ask what is $P(370 < Y < 430)$?

   Simulation: Observe that the probability that we draw any of the 4 tickets from the box on any given draw is 1/4, since we have assumed that each of the tickets is equally likely to be drawn. We will generate 100 U(0,1) random numbers. We will say that we have drawn the ticket labelled "1" for all the random numbers which are less than or equal to 0.25. We will say that we have drawn a "3" for all the random numbers which are greater than 0.25 and less than or equal to0.5. We will also say that we have drawn a "3" for all the random numbers which are greater than 0.5 and less than or equal to 0.75. And lastly, we will say that we have drawn a "9" for all the random numbers which are greater than 0.75 and less than or equal to 1. Then we will add up these simulated draws to get one simulated value of Y.

   We will repeat this many times so that when we are through, we will have generated many simulated values of Y. Once we are done generating these simulated values of Y, we will find the proportion of the number of Y's which fall in the interval $(370, 430)$. This will be our simulated estimate for $P(370 < Y < 430)$. You will see in class that the true probability is approximately 0.6826.

   We now proceed to the simulation. Please follow the instructions carefully.

   a. Before doing anything, look around to see if a Systat command window is already open. If it is, go on to the next step. If not, go to WINDOW/Command in order to open up a command window. This is very important.

   b. Turn on CAPS LOCK.

   c. Open a new Systat file. Title the first column UNIF1, and use DATA/FILL WORKSHEET to fill the worksheet to 100 rows. Under DATA/MATH, let UNIF1=URN. This fills the UNIF1 column with 100 random U(0,1)'s. Now go under FILE/SAVE AS and save the file under your first name.

   d. Find the "SYSTAT Command" window. In this window type DATA, and then hit <return>. Type USE and a space, and then copy the file name which is listed directly after ESAVE just above where you are in the same window (to copy use <Apple C>, to paste use <Apple V>). You will need to include the quatation marks and everything inside them. Now hit <return> and type in the following series of commands, hitting <return> at the end of each line:

      SAVE NEWDATA

      DIM Y(50)

```
FOR I=1 TO 50
LET UNIF1=URN
IF UNIF1<=0.25 THEN LET Y(I)=1
IF UNIF1>0.25 AND UNIF1<=0.75 THEN LET Y(I)=3
IF UNIF1>0.75 THEN LET Y(I)=9
NEXT
RUN
```

It is possible that you will receive some error messages along the way or get confused. This is the most important step, so make sure to ask your lab instructor if you have any questions at all.

The computer will take a few seconds to process what you have just done.

What Systat has just done for you is to simulate 50 different experiments of drawing 100 times from our box of tickets! Explain in your own words what each of the IF..THEN statements in the above program are doing (i.e. to what part of the **Simulation** description are these statements related)? Why don't we need to have separate statements for the 2 cases 0.25<UNIF1<=0.5 and 0.5<UNIF1<=0.75?

e. Now go to FILE/OPEN and double click on the "SystatWork" folder (note, the first area you are prompted with is the UserWork Area, but you need to go to Desktop/Macintosh HD/ Systat/Systat Work). Ask your lab instructor if you can't find it. When you get to the right place, you should see NEWDATA listed as one of the files from which to choose. We want to actually edit this data file which we have just created, so click the "Edit" button once. You should now be looking at this huge data file which you have just created.

f. We are no longer interested in UNIF1, so double click on UNIF1, and go to EDIT/CLEAR VARIABLE. Note that if you want to delete columns (or rows) in the future, all you have to do is double click on that column (or row) and then type <Apple B>. At this point, you might want to take a few seconds to surf around the editor window and try to figure out what this data set has to do with the **Experiment** that we talked about above. Do the experiments correspond to the columns or the rows (hint: how many columns are there?, how many rows?)?

g. Now go to STATS/Stats/Statistics. We are only interested in the sums of the 50 columns, so turn off all the statistics except for SUM. Also, choose "Save statistic" in this window. This option will allow us to save the 50 sums to their own data file, so that we can look at them by themselves. When you think you've got it set up correctly, click on OK. If you do not select any variable, Systat will assume that you want the sums of each individual column, and it will return them all. It should ask if you want to save the file as "NEWDATA Stats". Click on "Save." It may ask you to replace the existing "NEWDATA Stats" file; that's fine, go ahead and replace it.

h. Now for our last fancy manipulation. The 50 sums have been saved to the data file "NEWDATA Stats", but Systat chooses to put one sum in each column (so there are 50 columns, each of which has only 1 element). We want a data file that has all 50 sums (i.e. Y's) in only one column.

Go to the command window and type "DATA." This will put us back in programming mode. (If Systat prompts you with something like "Save current file", type OK, and then click the "Save" button. Again, go ahead and "Replace" the file that is there if it asks you. This merely resaves the huge data file from which we just deleted the UNIF1 column.)

Now type "USE" and then a space and then copy down the name of the file which is directly above your cursor in the command window (it should have "NEWDATA Stats" at the end of it). Hit <return>. If you have problems here ask your lab instructor for help.

Now type the following series of commands, hitting <return> at the end of each line:

SAVE NEWDATA2

TRANSPOSE

RUN

(Note that it is fine to overwrite whatever files it asks you to) These commands save the data which was in the file "NEWDATA Stats" into the file "NEWDATA2", but all of our data will now be in one column.

i. Now go to FILE/OPEN (again you'll be prompted with the UserWork Area, but go to Desktop/Macintosh HD/Systat/Systat Work). We want to pull the NEWDATA2 file into our data editor. Make sure to click on the "Edit" button, not the "Use" button.

j. We need to count how many of these simulated Y's are less than or equal to 370 or greater than or equal to 430. This will be easy to do if we just sort the list. Go to DATA/SORT, select COL(1), and click OK. This will save the sorted values into (yet another) new data file. Finally go to FILE/OPEN and choose this new sorted file. Now figure out the proportion of Y values which fall (strictly) between 370 and 430. Is this proportion close to the theoretical probability of 0.6826?