Math 130 – Jeff Stratton Name ______Solutions______

The Binomial Model

Goal: To gain experience with the binomial model as well as the sampling distribution of the mean.

Part 1 – The Binomial Model

In this part, we'll derive the binomial model. Remember that a probability distribution is a table, graph, or model giving (1), the possible values of the random variable X, and (2), the probabilities of each value of X.

The binomial model is useful in a very specific set of circumstances. We'll illustrate it using an example.

EXAMPLE:
We are in a computer store that sells Dell computers and Others. Suppose that for a given sale, the probability of selling a Dell computer is 0.80. We observe four sales in a row, and will consider them to be independent. Let X be the number dell computers sold out of the four sales. Fill in the table below:

Outcomes	Value of X	Outcome Probability
OOOO	0	0.2 × 0.2 × 0.2 × 0.2 = 0.0016
OOOD	1	0.2 × 0.2 × 0.2 × 0.8 = 0.0064
OODO	1	0.2 × 0.2 × 0.8 × 0.2 = 0.0064
ODOO	1	0.2 × 0.8 × 0.2 × 0.2 = 0.0064
DOOO	1	0.8 × 0.2 × 0.2 × 0.2 = 0.0064
OODD	2	0.2 × 0.2 × 0.8 × 0.8 = 0.0256
ODOD	2	0.2 × 0.8 × 0.2 × 0.8 = 0.0256
DOOD	2	0.8 × 0.2 × 0.2 × 0.8 = 0.0256
DDOO	2	0.8 × 0.8 × 0.2 × 0.2 = 0.0256
DODO	2	0.8 × 0.2 × 0.8 × 0.2 = 0.0256
ODDO	2	0.2 × 0.8 × 0.8 × 0.2 = 0.0256
DDDO	3	0.8 × 0.8 × 0.8 × 0.2 = 0.1024
DDOD	3	0.8 × 0.8 × 0.2 × 0.8 = 0.1024
DODD	3	0.8 × 0.2 × 0.8 × 0.8 = 0.1024
ODDD	3	0.2 × 0.8 × 0.8 × 0.8 = 0.1024
DDDD	4	0.8 × 0.8 × 0.8 × 0.8 = 0.4096

X	$P(X)$
0	0.0016
1	0.0256
2	0.1536
3	0.4096
4	0.4096

Notice that the probability of each outcome in the table above can be written as the product of the probability of selling a Dell, or one minus that probability.

For example, $P(DDDO) = 0.8 \times 0.8 \times 0.8 \times 0.2 = 0.8^3(1 - 0.8)^1$

This regularity can be exploited to find the probability of *X* using a formula, which is much easier than writing down all the possible outcomes. If we'd observed the next 10 sales, figuring out the outcomes would be a *lot* harder.

We can use the following Binomial Model instead:

$$P(X = x) = {}_nC_x p^x q^{n-x}, \text{ where } {}_nC_x = \frac{n!}{x!(n-x)!}$$

n = number of trials
p = probability of success (and *q* = 1 – *p* = probability of failure)
X = number of successes in *n* trials

Assumes:

- Only 2 possible outcomes for each trial.
- Constant prob. of successs, *p*.
- Independent trials.

Recall that $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

Q1] Use the formula given above to verify each of the probabilities that you calculated above for *X* = 0, 1, 2, 3, 4.

We can use R to compute the number of combinations, though. There's no function native to R that will compute the number of combinations, but we can write one. To do this, we'll open R without R commander.

- Open R
- In the "R Console" window, paste the following onto the line with the ">" symbol and press ENTER.

```
combinations <- function(N,n){
  outnum <- factorial(N)/(factorial(n)*factorial(N-n))
  outnum
  }
```

- Now, to compute the number of combinations of 3 dell computers out of 4, enter

```
> combinations(4,3)
```

$$P(X = 0) = \frac{4!}{0!\,4!}0.8^0 0.2^4 = 0.0016$$

$$P(X = 1) = \frac{4!}{1!\,3!}0.8^1 0.2^3 = (4)0.8^1 0.2^3 = 0.0256$$

$$P(X = 2) = \frac{4!}{2!\,2!}0.8^2 0.2^2 = (6)0.8^2 0.2^2 = 0.1536$$

$$P(X = 3) = \frac{4!}{3!\,1!}0.8^3 0.2^1 = (4)0.8^3 0.2^1 = 0.4096$$

$$P(X = 4) = \frac{4!}{4!\,0!}0.8^4 0.2^0 = 0.4096$$

Q2] Compute the expected value and standard deviation for X.

$$E[X] = \mu = \sum_{i=1}^{n} X_i P(X_i) \text{ and } \sigma = \sqrt{\sum_{i=1}^{n} (X_i - \mu)^2 P(X_i)}$$

$$\begin{aligned}
E[X] = \mu \quad &= \sum_{i=1}^{n} X_i P(X_i) \\
&= 0(0.0016) + 1(0.0256) + 2(0.1536) + 3(0.4096) + 4(0.4096) \\
&= 0 + 0.0256 + 0.3072 + 1.2288 + 1.6384 \\
&= 3.2
\end{aligned}$$

$$\begin{aligned}
\sigma^2 \quad &= \sum_{i=1}^{n} (X_i - \mu)^2 P(X_i) \\
&= (0 - 3.2)^2(0.0016) + (1 - 3.2)^2(0.0256) + (2 - 3.2)^2(0.1536) \\
&\quad +(3 - 3.2)^2(0.4096) + (4 - 3.2)^2(0.4096) \\
&= 0.64 \\
\sigma \quad &= \sqrt{0.64} = 0.8
\end{aligned}$$

Q3] For the binomial model, $E[X] = np$ and $\sigma = \sqrt{npq}$. Do these match the values you computed in Q2?

$$\begin{aligned}
&E[X] = np = 4(0.8) = 3.2 \\
&\sigma = \sqrt{npq} = \sqrt{4 \times 0.8 \times 0.2} = \sqrt{0.64} = 0.8
\end{aligned}$$

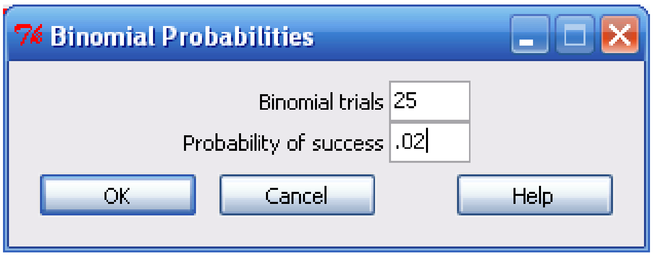
Our calculations match!

Part 2 – Using Rcmdr to compute binomial probabilities

Rcmdr can compute binomial probabilities. As an example, suppose we are sampling 25 light bulbs from an assembly line. The probability that an individual light bulb is defective is 0.02.

Q4] Find the probability that 2 light bulbs are defective.   0.0754

Go to:
Distributions → Discrete distributions → Binomial distribution → Binomial probabilities



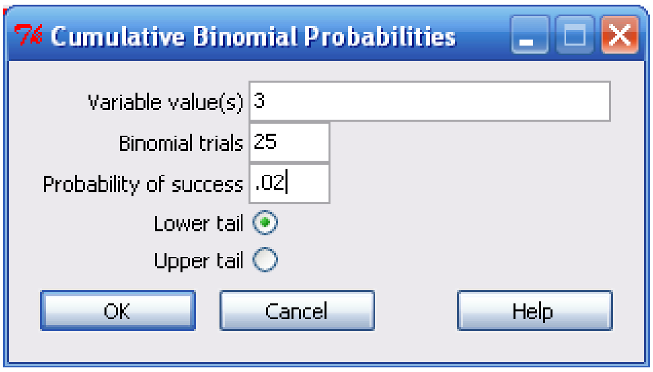
Binomial trials 25
Probability of success .02

This will create a listing of the *X*'s from 0 to 25 along with their probabilities. Next to each value of *X* is its probability

Q5] Find the probability that at most 3 light bulbs are defective.

This probability is trickier. We want $P(X \leq 3)$. We could do what we did above for Question 4 and add up the probabilities, but it is easier to get this done by the computer.

Distributions → Discrete distributions → Binomial distribution → Binomial tail probabilities



Variable value(s) 3
Binomial trials 25
Probability of success .02
Lower tail ●
Upper tail ○

Since we're interested in ≤, we select the "Lower tail". If we were interested in $P(X \geq 3)$, we'd select the "Upper tail."

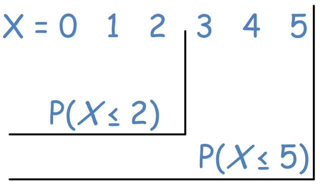
```
> pbinom(c(3), size=25, prob=0.02, lower.tail=TRUE)

[1] 0.9985541
```

Q6] Find the probability that between 3 and 5 (inclusive) bulbs were defective, $P(3 \leq X \leq 5)$. This may take some thought.

There are a couple of approaches here. First, we could individually find the probabilities of 3, 4, and 5 and add them up.

Second, we can think of the possibilities for X:



To get the probability between 3 and 5 inclusive, I can get the probability the X is less than or equal to 5, and subtract off the probability of being less than or equal two. This is 0.9999918 - 0.9867566 = 0.0132.

There's a 1.32% probability that between 3 and 5 light bulbs are defective.

```
> pbinom(c(2,5), size=25, prob=0.02, lower.tail=TRUE)
[1] 0.9867566 0.9999918
```

Part 3 – Plots of the binomial distribution

From the hand calculations you did in part 1, you can see that finding Binomial probabilities by hand would be very tedious if your *n* gets large. However, the binomial distribution can be approximated by another. To see this, we'll make several plots of a binomial distribution.

You may have noticed by now that when you make plots in Rcmdr, they are created in a graphics window found in the R window. Additional graphs overwrite the graphs you've already created. We can split up the graphics window into different panes. In the original R window submit:

```
> par(mfrow=c(2,2))
```

This will split the graphics window into 2 columns and 2 rows.

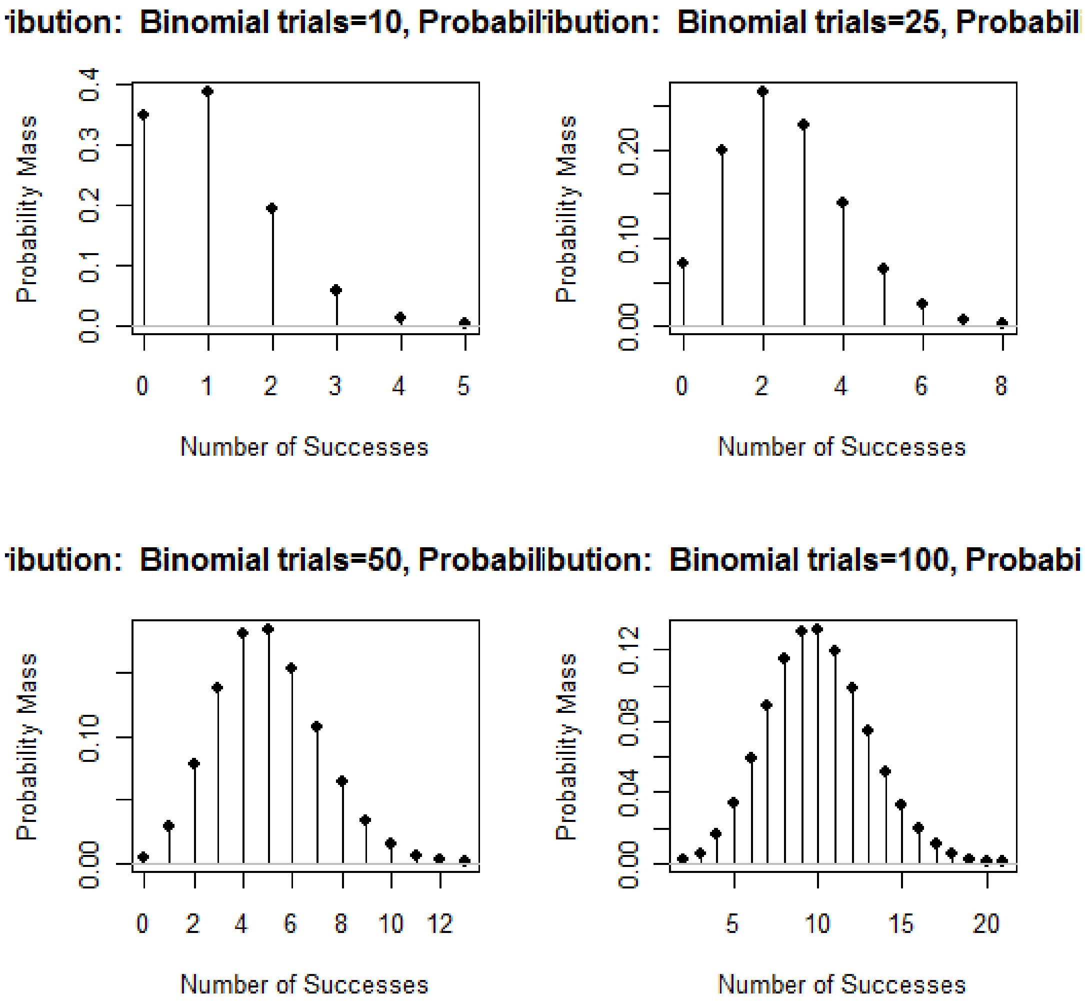
Now, graph a binomial distribution with *n* = 10 and *p* = 0.1. Go to **Distributions → Discrete Distributions → Binomial distribution → Plot binomial distribution.** Put in 10 in the "Binomial trials" window, and 0.1 in the "Probability of success" window. Leave the "Plot probability mass function" radio button selected. Then click OK.

Draw three additional binomial distributions:

- $n = 25$, $p = 0.1$
- $n = 50$, $p = 0.1$
- $n = 100$, $p = 0.1$

As n increases, you should notice a more familiar shape! In the past, people approximated binomial probabilities using normal probabilities when n was large. Now you can see why. Nowadays, computers are powerful enough that the “normal approximation of the binomial” isn’t often needed.

The plots are shown below. We see the distribution looking more and more normal as the sample size n increases.



Part 4 – The Sampling Distribution of the Mean

In this part, we'll investigate the sampling distribution of the mean. Just as a random variable X has a distribution, so does a mean of that variable. For example, let's consider the Class Data and the number of countries visited. Consider the class to be our population of interest. We're going to take samples of size 5 and use that to try to estimate the mean and standard deviation of the number of countries visited.

How does $\bar{X}$ behave? The surprising thing is that it behaves normally with the same center as your population, but a smaller variance.

We'll investigate this using our class data about the number of countries visited. We'll take a bunch (1,000) of random samples of size 5, and calculate the average number of countries visited for each repeated sample. Looking at a plot of these averages, we can see how the sample mean behaves.

We'll use R to investigate this. Rcmdr can't do this for us. Open R, and then go to File → New Script.

This will open a "Script window" in R. You can write R programs there. Then you can highlight sections of code and run it by hitting CTRL-R. Paste the following code into your script window. Then, submit the code in R by going to **Edit → Run all**.

```
###################################
### Script for Lab #5 - MATH 17 ###
###################################

# Read the data
country <- c(1,2,2,3,4,4,4,4,5,5,7,7,8,8,8,9,9,10,12,14)

# Print out the data
country

# Histogram of the original data
Hist(country,main="Number of Countries Visited")

# Get the mean and standard deviation of the class
mean(country)
sd(country)

# Now, consider a sample of 5 students from the class
classamp <- sample(country,5,replace=TRUE)
mean(classamp)
sd(classamp)

# Let's repeat the sampling many times, and look
# at a histogram and the mean/standard deviation of X-bar
sampclass <- matrix(NA,1000,5)
for(i in 1:1000){
  sampclass[i,] <- sample(country,5,replace=TRUE)
  }
```

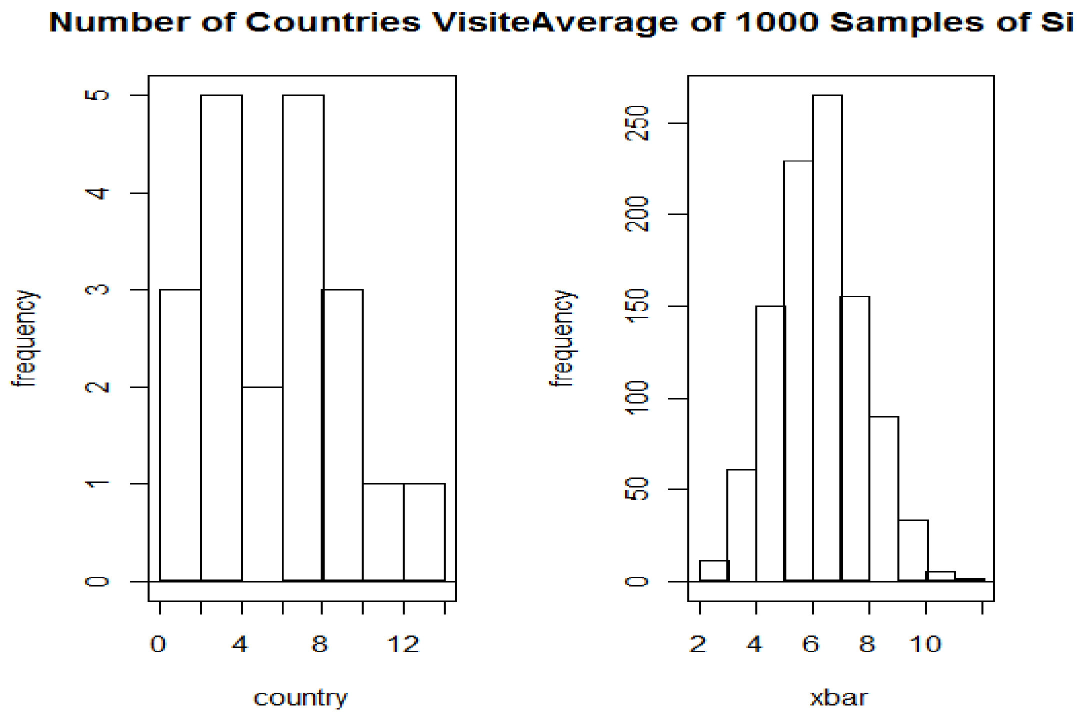
```
# Get a vector of the means of each sample of 5 students
xbar <- apply(sampclass,1,mean)

# Get histogram and mean/sd of the xbars
Hist(xbar,main="Average of 1000 Samples of Size 5")
mean(xbar)
sd(xbar)

# Compare Std. Dev. of xbar to sigma/root(n)
sd(country)/sqrt(5)
```

Q7] Describe the histogram of the original data. What are the mean and standard deviation of the original data?

The original data seem to have two peaks, but the center is at about 5 with a spread from 0 to 13 countries. The mean and standard deviation are 6.3 and 3.28 countries.



Q8] Describe the histogram of the averages of samples of size 5. Find the mean and standard deviation of the sample data. How do they compare to those above? You should find that the standard deviation of your sample xbars is close to the standard deviation of your original data divided by the square root of the sample size.

The histogram of the sample averages looks much more normal. It is centered at about 7 with a spread from 2 to 12. The mean and standard deviation are 6.3006 and 1.52544.

If I take my original standard deviation and divide by the square root of the sample size, I get 1.55665, pretty close to the standard deviation of my sample averages.

Part 4 – More Sampling Distribution of the Sample Mean

For sample means, we will learn about the sampling distribution via an applet (link online). Steer your (Java-enabled) browsers to http://onlinestatbook.com/stat_sim/sampling_dist/index.html. In this applet, when you first hit Begin, a histogram of a normal distribution is displayed at the top of the screen. This is the parent population from which samples are taken (think of it as the bin of balls) except it's showing the distribution. The mean of that distribution is indicated by a small blue line and the median is indicated by a small purple line. Since the mean and median are the same for a normal distribution, the two lines overlap. The red line extends from the mean one standard deviation in each direction.

The second histogram displays the sample data. This histogram is initially blank. The third and fourth histograms show the distribution of statistics computed from the sample data. The option N in those histograms is the sample size you are drawing from the population. We will be exploring the distribution of the sample mean by drawing many samples from the parent distribution and examining the distribution of the sample means we get.

Step 1. Describe the parent population. What distribution is it and what is its mean and standard deviation?

Step 2. You can see the third histogram is already set to “Mean”, with a sample size of N = 5. Click Animated sample once. The animation shows five observations being drawn from the parent distribution. Their mean is computed and dropped down onto the third histogram. For your sample, what was the sample mean?

Step 3. Click Animated sample again. A new set of five observations are drawn, their mean is computed and dropped as the second sample mean onto the third histogram. What did the mean of the sample means (yes, we are interested in the mean of sample means as part of the sampling distribution) change to?

Step 4. Click Animated sample one more time. What did the mean of the sample means update to now?

Step 5. Click 10,000. This takes 10,000 samples at once (no more animation) and will place those 10,000 sample means on the third histogram and update the mean and standard deviation of the sample means. Record the mean and standard deviation of the sample means. What shape does this third histogram have? How do these findings compare to the parent distribution?

Step 6. Hit Clear Lower 3 in the upper right corner. Change N = 5 to N = 25 for the third histogram. Do animated sample at least once (convince yourself it is actually samples of 25 now). Then take 10,000 at once. Record the mean and standard deviation of the sample means. What shape does the third histogram have? How do these findings compare to the parent distribution?

Step 7. Compare the different standard deviations from Steps 5 and 6. What effect does sample size appear to have on standard deviation of the sample means?

Step 8. Hit Clear Lower 3. Change the parent distribution to Skewed. What are the new mean and standard deviation of the parent distribution? Which direction is this distribution skewed?

Step 9. Set N = 5 back for the third histogram. Set “Mean” and N = 25 for the fourth histogram. Hit 10,000 at once. (This will take 10,000 samples of size 5, compute the sample means and put those means in the third histogram, as well as take 10,000 samples of size 25, compute the sample means and put those means in the fourth histogram). What do the distributions look like for the third and fourth histograms? Are they skewed like the parent population? What are the means and standard deviations for each histogram?

Step 10. Hit Clear Lower 3. Change the parent distribution to Custom. Draw in a custom distribution (left click and drag the mouse over the top histogram). Sketch your custom distribution below. What are its mean and standard deviation?

Step 11. Hit 10,000 at once (leave the settings on the third and fourth histograms alone). (You could take animated once to convince yourself it was really drawing from your new distribution). What do the third and fourth histograms look like? Anything like the parent distribution? What are their means and standard deviations?

The Sampling Distribution for the sample mean, $\bar{X}$ can be described as having a mean $\mu_{\bar{X}} = \mu$, the same as that of the population mean. The standard deviation is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
The distribution is exactly normal if the parent population is normal. Finally, the Central Limit Theorem tells us the distribution will be approximately normal with the mean and standard deviation stated above if n is sufficiently large even if the population distribution is not normal.