Your Name: ______________________

Section: ______________________

36-201 INTRODUCTION TO STATISTICAL REASONING
Computer Lab Exercise – Lab #10
Relative Frequency, the Central Limit Theorem, and Normal Approximation Calculations

Objectives:

1. To learn how to simulate random data in MINITAB.
2. To investigate the Central Limit Theorem approximation (a.k.a the Normal approximation).
3. To practice probability calculations using the Normal approximation and Z-scores.

Getting Started

For these exercises you will need:

- A good pencil and eraser(s) [ink pens are not recommended!]
- A scientific or business calculator.
- MINITAB.

Copying Files

1. Open the **Student** folder. To do this, double click on the "Cluster HD" icon. Then double click the "documents" folder. You will eventually move the data into this folder.
2. In the following order: From the Apple menu in the upper left hand corner of the titlebar, choose **Chooser**. In the dialog box that appears *click* **AppleShare**. In the bottom left box called "AppleTalk Zones" *select* **BH**. In the top right box where it says "Select a file server" select **HSSHELIOS**. *Click* **Ok**.
3. In the dialog box that appears login using "36201" as your Name and "36201" as your password. Click **Ok**. In the dialog box that appears highlight "Class." Click **Ok**.
4. A "Class" icon will appear to the right. Double click on the icon. Then double click on 36201. You will see many files in a folder.
5. Copy all of the files that you will need to the **Student** folder. Do this by dragging the files into the Student folder. For today, the data file will need is *magazines.dat*.
6. Close all windows.
7. **IMPORTANT:** Drag the *class* icon into the trash. This is necessary to let others in the lab get access to the data.

Background

The *Central Limit Theorem* (CLT) says that if the random variable X is the sum or average of a lot of similar, (nearly) independent random terms, then X will be approximately Normally distributed, with mean μ_X and standard deviation σ_X. The approximation gets better as the number of summands increases.

Therefore to estimate probabilities involving X, we can convert X to the z-score $Z = (X - \mu_X)/\sigma_X$ and look up the corresponding probability for Z in a table of normal probabilities.

In this lab you will explore the validity of the CLT for two very different distributions:

- The Binomial Distribution
- A Bimodal Distribution

and you will practice basic "normal approximation" probability calculations using Z-scores.

I. The Normal Approximation to the Binomial Distribution

> *Suppose you do a survey of n on-campus voters and you ask each voter surveyed whether they would vote for Pat Buchanan in the Presidential Election in 2000. The number of voters X favoring Buchanan in your survey is a Binomial random variable with n trials and success probability p (the probability that each voter would vote for Buchanan). X can be thought of as the sum of a "1" [for Buchanan] or a "0" [for anyone else] for each of the n voters surveyed. Therefore the CLT says that X should be approximately normally distributed, with mean $\mu_X = np$ and standard deviation $\sigma_X = \sqrt{np(1-p)}$. The normal approximation should get better as n grows.*

Suppose the true fraction of persons on campus who favor Buchanan is 10%, so $p = 0.10$. If we repeat the survey 100 times—each time asking just 5 people who they will vote for, and tallying the number who will vote for Buchanan—we will have 100 "Buchanan counts" whose distribution we can compare with the Normal distribution.

> To simulate 100 surveys of $n = 5$ persons each, pull down the **Calc** menu and select **Random Data**. Then select **Binomial** about halfway down the submenu that appears. In the dialog box that follows, type 100 next to "Generate" to generate results from 100 surveys, and type C1 below "Store in" to store the results in. Near the bottom type 5 for "Number of trials" (for 5 persons per survey) and 0.1 for "Probability of Success" (for 10% who favor Buchanan). Then click on **Ok**.

We would like to get a general idea how this data looks, so pull down the **Stat** menu and select **Basic Statistics** and then **Descriptive Statistics** to find basic descriptive statistics for column C1. Enter C1 in the "Variable" block, and select "Graphical Form" instead of "Tabular Form" near the bottom of the dialog box. Click **Ok** and wait for a window summarizing the data to pop up.

Question #1: Compare the theoretical mean and standard deviation for this sample of 100 survey results with the sample values from the graphical summary window. Recall that the theoretical mean is $\mu_X = np$ and the theoretical SD is $\sigma_X = \sqrt{np(1-p)}$.

	Theory	Sample
Mean		
SD		

Question #2: In the graphical summary there should be a normal curve overlaid on the histogram for your 100 surveys. Describe generally how well the histogram follows the normal curve.

Now let us expand the size of the 100 surveys from 5 persons each to 50 persons each. We will again get 100 counts that we can compare to the Normal distribution.

> To simulate 100 surveys of $n = 50$ persons each, follow the same procedure as before, but now type 50 for the "Number of trials" instead of 5 (everything else should be the same as before).

Question #3: Use **Stat**, **Basic Statistics**, **Descriptive Statistics** as you did for Question 1 to get a graphical summary of the data. Compare the theoretical mean and standard deviation for this sample of 100 survey results with the sample values from the graphical summary window. Recall that the theoretical mean is $\mu_X = np$ and the theoretical SD is $\sigma_X = \sqrt{np(1-p)}$.

	Theory	Sample
Mean		
SD		

Question #4: In the graphical summary there should be a normal curve overlaid on the histogram for your 100 surveys. Describe generally how well the histogram follows the normal curve.

♣ **Question #5: Overall, do you think the normal approximation worked better for $n = 5$ or for $n = 50$?**

♣ **Question #6: Use μ_X and σ_X from Question 3 and the Normal approximation to answer the following question:** *In a sample of 50 campus voters, approximately how likely are you to find 8 or fewer who favor Buchanan?*

$$
\begin{aligned}
P[X \leq 8] &= P\left[\frac{X - \boxed{}}{\boxed{}} \leq \frac{8 - \boxed{}}{\boxed{}}\right] \\
&= P\left[Z \leq \boxed{}\right] \\
&\approx \boxed{}
\end{aligned}
$$

To fill in the last box, pull down the **Calc** menu, select **Probability Distributions** and then **Normal**. Click on "Cumulative probability" near the top of the dialog box that appears, and click on "Input constant" near the bottom of the box. Type in the Z-score you calculated in the second line above. MINITAB will print out the value x that you typed in, and also the probability $P[X \leq x]$. Copy this probability into the last box above.

> **Note:** Using MINITAB to look up the exact probability like this is an alternative to using the 68–95–99.7 rule with interpolation as we have done in class. Usually our simpler method gives answers similar to the exact answer.

II. The Normal Approximation to a Bimodal Distribution

Magazine marketing research shows that subscribers to magazines tend to either ignore the magazines they subscribe to, or read them religeously. For a monthly magazine, this leads to a distribution of magazines read like the one in Table 1. Suppose you do a survey of n such magazine readers. The CLT says that the average number of magazines read $\overline{X}$ in your survey should be approximately Normally distributed, with mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$. The normal approximation should get better as n grows.

Magazines Read Per Year	Probability
0	0.13
1	0.11
2	0.09
3	0.07
4	0.05
5	0.04
6	0.02
7	0.04
8	0.05
9	0.07
10	0.09
11	0.11
12	0.13

Table 1: Typical distribution of number of issues read per year, by subscribers to a monthly magazine.

To begin with we input the probability distribution in Table 1 into columns C40 and C41 of MINITAB, using the following three steps:

1. In MINITAB, pull down the **File** menu, and select **Import Text** under the **Other Files** option. You must identify the columns into which the data are to be read, by double-clicking on column names on the left, or by typing column names in at the top. Type C40-c41, since there are two columns of data. Then click on **OK**.
2. A new dialog box will appear with the heading **Minitab** followed by a list of files. Change the folder from **Minitab** to **Desktop**. A new set of folders, as well as the the file *magazines.dat* that you copied from the H&SS server at the beginning of the lab, will appear.
3. Select the the file *magazines.dat* and click on **Open** to import the data into MINITAB. (You may have to tell the Mac to open the file in "MINITAB 10.5 Xtra Power" at this point.) After a brief period, the data will appear in the MINITAB spreadsheet.

Once the probability distribution is imported we will simulate some data from this magazine distribution.

> Pull down **Calc**, select **Random Data** and then **Discrete** near the middle of the submenu. Again type 100 near "Generate" but now under "Store in columns" type C1–C20. Next to "Values in" type C40 and next to "Probabilities in" type C41. Then click **Ok** and wait for the data to be generated.

For this problem we will follow the same format as in the previous problem, but we will do three sizes of survey: $n = 1$, $n = 5$ and $n = 20$.

- We will let C21 represent 100 surveys of 1 person each: Pull down **Calc** and select **Row Statistics**. In the dialog box that appears, click on the "Mean" button, and in the "Input Variables" box type C1. In the "Store results in" box type C21, and then click **Ok**.

- We will let C22 represent 100 surveys of 5 people each: Pull down **Calc** and select **Row Statistics**. In the dialog box that appears, click on the "Mean" button, and in the "Input Variables" box type C1–C5. In the "Store results in" box type C22, and then click **Ok**.

- We will let C23 represent 100 surveys of 20 people each: Pull down **Calc** and select **Row Statistics**. In the dialog box that appears, click on the "Mean" button, and in the "Input Variables" box type C1–C20. In the "Store results in" box type C23, and then click **Ok**.

Question #7: Use **Stat**, **Basic Statistics**, **Descriptive Statistics** to get graphical descriptive statistics of the three samples in C21, C22 and C23. Describe how well each histogram follows the overlaid normal curve.

♣ **Question #8: Describe the evidence that this exercise gives for the CLT, for the magazine data. What happens, as the sample size changes from 1 to 5 to 20 magazine readers per survey?**

♣ **Question #9: Use the value of Mean for $\mu_{\overline{X}}$ and StDev for $\sigma_{\overline{X}}$ in the graphical summary of C23 (samples of size 20) together with the Normal approximation to answer the following question:** *How likely are you to find that the average number of magazines read in a yearly subscription by a random sample of 20 magazine subscribers is less than or equal to 5?*

$$
\begin{aligned}
P[\overline{X} \le 5] &= P\left[\frac{\overline{X} - \boxed{}}{\boxed{}} \le \frac{5 - \boxed{}}{\boxed{}}\right] \\
&= P\left[Z \le \boxed{}\right] \\
&\approx \boxed{}
\end{aligned}
$$

To fill in the last box, pull down the **Calc** menu, select **Probability Distributions** and then **Normal**. Click on "Cumulative probability" near the top of the dialog box that appears, and click on "Input constant" near the bottom of the box. Type in the Z-score you calculated in the second line above. MINITAB will print out the value x that you typed in, and also the probability $P[X \le x]$. Copy this probability into the last box above.

> **Note:** Again, we could have approximated this probability using the 68–95–99.7 rule together with interpolation, but letting MINITAB look up the answer like this is faster and more accurate *(if you are lucky enough to have a computer with* MINITAB *around!)*.

To quit MINITAB *choose the* **Quit** *command in the* **File** *menu. Do not save any changes. Delete files and folders that you might have created. Log out of the Mac if you had to log in.*