NORMAL RANDOM VARIABLES AND THE CENTRAL LIMIT THEOREM

In the birthday simulation in class, we saw that the distribution of the averages of three birth dates is narrower than the distribution of the birth dates themselves. If we had more people, we would also have seen that, whereas the original birth dates (Y) came from a close to uniform distribution, the distribution of the averages ($\bar{Y}_3$) was more mound-shaped. This is an illustration of the idea behind the *Central Limit Theorem.* In fact, there are several versions of Central Limit Theorems. A very broad summary is:

Central Limit Theorem, General Version: The distribution of the sum (or mean) of a large enough number of independent random variables is approximately normal.

(Actually, we need some hypotheses saying that the random variables involved aren't too horribly badly behaved – like being so "heavy tailed" that they have infinite variance.)

Exercise: Is the CLT for a sum an entirely different theorem from the CLT for a mean? (What is the relationship between the distribution of the sum and the distribution of the mean of a collection of random variables?)

Natural Question: How large is "large enough"?

Web Simulations:

Pictures: http://www2.austin.cc.tx.us/mparker/1342/cltdemos.htm
(Original distribution in gray; distribution of mean dark line; distribution of normal with same mean and standard deviation as mean dotted)

Demo: http://www.ruf.rice.edu/%7Elane/stat_sim/sampling_dist/

Applications:

1. A person's height is the sum of the heights of lots of parts: ankle, lower leg, upper leg, pelvis, lots of vertebrae, and head. If these are independent, the CLT predicts an approximately normal distribution of heights (at least for adults of the same sex). Empirical data are consistent with normality.

2. Errors (e.g., measurement errors, errors in manufacturing parts) may be the result of many smaller errors adding up, so it is reasonable to expect that in many cases (if the smaller errors are independent) that errors are normally distributed. Indeed, Gauss studied normal distributions because of an interest in studying errors; normal distributions are often called Gaussian distributions.

3. The central limit theorem tells us that many sampling distributions (e.g., of means, but also others) are approximately normally distributed. (This allows us to form confidence intervals and do hypothesis tests based on the normal distribution.)

4. Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with the same probability of success p. Then the CLT says that for n large enough, the distribution of $X_1 + \ldots + X_n$ is approximately normal. But $X_1 + \ldots + X_n$ is just the binomial random variable with n trials, each with probability of success p. Hence, *binomial random variables (for large enough n) are approximately norma*l.
Note: The Central Limit Theorem also says that the approximating normal distribution has the same mean and variance as the binomial random variable that it is approximating.

5. A Galton Board (or Galton box, or quincunx, or bean machine) uses a variation of the idea in #4. This consists of a vertical or inclined board with staggered pins arranged in a triangle or rectangle. Balls are dropped through the board. Each time a ball hits a pin, it has probability ½ of going in each direction. So if X_i is the distance traveled horizontally when the ball hits the ith pin, X_i is like a Bernoulli random variable, except with values -1 and +1 instead of 0 and 1. The total distance the ball goes left or right as it falls is the sum $X_1 + \ldots + X_n$, where n is the number of rows in the arrangement of pins. By the CLT, this distance is approximately normal for n large. There are simulations at
http://www.rand.org/statistics/applets/clt.html (rectangular)
http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html (rectangular + info)
http://www.teacherlink.org/content/math/interactive/flash/quincunx/quincunx.html
(triangular).

There are variations with probabilities other than ½, but you need a larger n to get a distribution close to normal. There's a simulation that allows you to vary both n and p at http://www.jcu.edu/math/isep/Quincunx/Quincunx.html.

Working with normal distributions

The pdf of the normal distribution is not integrable in closed form. That is, *there is no formula for its antiderivative*. So tables or technology are needed to evaluate areas under a normal curve. But the following reasoning shows that we really just need tables for the standard normal ($\mu = 0$ and $\sigma = 1$). The standard normal is traditionally called Z.

Recall: In Problem 10 of the handout Cumulative Distribution Functions, we showed that if $Y = aZ + b$, then Y is normal with parameters $\mu = b$ and $\sigma = a$. Thus
$(Y - \mu)/\sigma = (Y - b)/a = Z$ is standard normal

So, for example, if Y is normal with parameters μ and σ, then

$$\begin{aligned} P(a < Y < b) &= P((a - \mu)/\sigma < (Y - \mu)/\sigma < (b - \mu)/\sigma) \\ &= P((a - \mu)/\sigma < Z < (b - \mu)/\sigma) \\ &= P(Z< (b - \mu)/\sigma) - P(Z < (a - \mu)/\sigma\ ) \quad \text{(Draw a picture!)} \end{aligned}$$

Tables of normal values in particular give us the *empirical rule*:

The area under a normal curve between $\mu - \sigma$ and $\mu + \sigma$ is about 0.68.

The area under a normal curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 0.95.

The area under a normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 0.997.

(Draw a picture!)

Thus "standard deviations" are a natural unit for talking about normal random variables. (*This will be used in Problems 2 and 4 below.*)

Problems:

1. *The purpose of this problem is to do your own simulation illustrating the Central Limit Theorem, to get a better feel for what is going on.*

A. a. Calculate (from the theory we have already developed) the mean and the variance of a binomial random variable with parameters p = 0.5 and n = 10. Also calculate the *standard deviation* (defined as the square root of the variance) of this random variable. (Note from item 4 above that the approximating normal distribution has the same mean and standard deviation as the binomial random variable it is approximating.)

b. (*See additional handout for tips on how to do this*.) Now simulate a sample from the binomial random variable with p = 0.5 and n = 10, using the following steps:

i. Sample 10,000 values (or 1000, if your hardware or software complains) from each of ten independent Bernoulli random variables (each with parameter p = 0.5)

ii. Sum the samples in (i) to obtain a sample from a binomial random variable with p = 0.5 and n = 10

iii. (*See pointers below before doing this*.) Draw a graph showing, *superimposed in the same graph*, both a *density* histogram (See the handout Random Variables if you don't remember what a density histogram is) of the 10,000 values of your sum, *and* a normal curve *with the mean and standard deviation you calculated in part (a).*

iv. Also use the software to calculate the *sample* mean and *sample* standard deviation of the sample you obtained in (ii). Compare with what you calculated in part a. (What you calculated in part (a) are the *population* mean and standard deviation for both the binomial random variable and the normal random variable; what you calculated using software are estimates of these.)

<u>Hand in</u>:

I. Part a.

II. The first ten samples from each of the ten binary variables plus their sum (from i and ii of part b).

III. Your graph (density histogram plus superimposed normal curve).

IV. Your code or an outline of the steps you used (unless you used the instructions given for doing the simulation on Minitab).

V. Part (iv) of b.

Things to think about as you do this problem:

- If your software doesn't give you an option to make a density histogram, you will need to think about how to scale the frequency histogram to make a density histogram. (See the handout Random Variables for ideas.)

(Problem cont'd on next page)

- Think about what bins (intervals) are appropriate for your histogram. The default of your software might *not* be a good choice.

B. Do a second related simulation or group of simulations that you think is interesting. *Everyone in the class should choose a different simulation*. You may choose from one of the following types of simulation, or something similar. You will need to figure out how to alter the instructions for part A. *Hand in the same items as for part (A), plus a description of exactly what you are simulating (e.g.,your choice of p and n if you are doing Choice(i).)*
Possible choices:

i. Simulate a binomial random variable with p not equal to 1/2 as a sum of Bernoulli random variables. (The farther p is from 0.5, the larger you will need n to be; or you might try this for a couple of different values of n and compare.)

ii. Do a simulation of a normal random variable by adding randomly drawn values from two or more *different* random variables. (For example, you might choose ten or more Bernoulli random variables, each with a different value of p; or ten with = 0.5 and ten with p = 0.75; or ten or more different uniform random variables; or ten Bernoulli and ten uniform, etc.)

The remaining problems are intended to give you some practice using normal models to investigate questions of real-world interest. For these problems, you will need to use a graphing calculator, tables of normal values, an online calculator, or software such as Excel to get values for the standard normal distribution.

2. *Background*: Bone densitometry (measurement of bone density) is often used as a screening test for possible future hip fracture. A 1991 paper in the British Medical Journal examined studies comparing femoral neck bone density of women with hip fractures and of control groups of women in the same age range without hip fractures. Three of the studies had problems with their methodology. Since the remaining five studies used different units of measurement, the authors converted all the measurements to units of standard deviations for the particular study. *(See comments right before the problems.)* In these five studies, the difference in average bone density between the two groups (women with hip fractures and the controls) ranged between 0.4 and 0.75 standard deviations. Since the studies had different numbers of subjects, the authors used a weighted average of these differences as their estimate of the difference in average bone density between women with hip fractures and those without. The weighted average was one half standard deviation (with women with hip fractures having the lower average bone density). The authors then used a normal model to examine detection rates and false positive rates for bone density as a test for risk of femoral neck fracture. In this problem, you will carry out the analysis that they did.

a. Draw, *in the same picture*, the pdf's for the normal models of bone density for the two groups (women with hip fractures, controls). Remember that you are using "one standard deviation" as your unit. So the means of the two normal pdf curves should be ½ unit apart. (To check the quality of your picture, remember from an early problem that the point of inflection of the normal curve is one standard deviation from the mean.)

b. Suppose a cut-off of one standard deviation below the mean of the controls is used as a cut-off for deciding if a woman is at risk for a hip fracture. Calculate the TPP and FPP (see the handout "Plots Summarizing Medical Tests") for this test.

c. Repeat with cut-off 1.5 standard deviations below the mean of the controls.

d. Repeat with cut-off 2 standard deviations below the mean of the controls. (*cont'd*)

e. Now make a plot of TPP vs FPP and mark each of the three tests (parts b, c, and d).

f. Based on your results in parts (a) – (e), discuss how good bone mineral density testing is as a screening test for future hip fracture.

3. (*You will need to use software or a programmable graphing calculator for this problem.*) Climate models reviewed in the 2007 report of the Intergovernmental Panel on Climate Change (IPCC) project that global surface temperatures are likely to increase by 1.1 to 6.4 °C (2.0 to 11.5 °F) between 1990 and 2100. Since global warming may cause more variability in local temperatures, the standard deviation of the distribution of temperatures could also increase. This problem uses a normal model to get a feel for what changes in mean and standard deviation could mean for extreme temperatures in Austin.

a. August high temperatures in Austin are approximately normally distributed with mean 96.5 and standard deviation 4.5. Using a normal model for August high temperatures, investigate the effect of different increases in mean (within the range projected by the IPCC) and standard deviation on the number of August days in Austin with high temperature 100 or more.

One way to do this is with Excel. Check to be sure the function NORMDIST is in your version of Excel. If not, if you can add the Data Analysis package, that should give the function. NORMDIST(x,mean,standard_dev,TRUE) gives the cdf of the normal distribution with mean and standard deviation as given, as a function of x. In other words, NORMDIST(x,mean,standard deviation,TRUE) calculates the probability (as a decimal) that a randomly chosen value from the normal distribution with given mean and standard deviation has value ≤ x. *Note*: If you use the Open Office spreadsheet instead of Excel, use NORMDIST(x; mean: standard_dev; 1) instead of NORMDIST(x,mean,standard_dev,TRUE).

Be sure to include each of the following in report of your investigation:

i. Both the percentage and typical number of August days when the high temperature in Austin is 100 or higher for several combinations of means and standard deviations within the range of interest. (Consider at least standard deviations 5.5 and 6.5 as well as 4.5.)

ii. A graph that shows how the different possible future means and standard deviations affect the typical number of Austin August days with highs 100 or more.

Also include anything else of interest!

b. There is some evidence that the actual distribution of high temperatures may be slightly skewed to the left (that is, have a larger tail to the left) rather than normal. If we used this model instead of a normal model, would your answers in (a) be smaller or larger?

4. Standardized tests are usually constructed to have a normal distribution of scores. Lily’s class took two such standardized tests, one in English and one in math. Her scores and the mean and standard deviation for each test are shown in the table below:

Test	Lily’s score	Mean	Standard deviation
English	85	75	7.7
Math	65	55	6

a. Which test did she do better on? [Hint: Measure in standard deviations.]

b. After the math test, the students discussed it and realized that everyone in the class had missed one particular question, worth four points. The class hadn’t covered the material needed in that question.

i. Andy scored 52 on the test. What would be the change in his percentile rank if his score were 4 points higher? Draw a picture illustrating this.

ii. Brenda scored 68 on the test. What would be the change in her percentile rank if her score were 4 points higher? Draw a picture illustrating this.

iii. If 100,000 students took the test, about how many would Andy pass by in percentile if his score were 4 points higher?

iv. Same as (iii) but for Brenda.

v. Compare your answers for Andy and Brenda. Do they surprise you? Might they surprise your principal?