I. Introduction

Previously we have looked at the law of averages and see that over many repetitions of a random event, the outcome is close to what we would call "an expected value". Chapter 18's function is to explain the logic behind section 17.3 and it begins with a discussion of how many ways (patterns) can be constructed for a simple coin toss.

What you see is that very quickly, as the sample size increases, that is, say from flipping a coin 5 times, to 100, the number of possible patterns becomes very large. This was done to prove that the entire pattern of resulting outcomes is close to the normal curve – particularly when the number of tosses (or sample size or number of draws) is large.

II. The Probability Histogram (18.2-18.3)

This pattern of all possible outcomes resulting from many repetitions of a random process is called a "probability histogram" (it's a graph of <u>all possible</u> flips of size 5 or size 20 or any size for example). You could also call this graph of all possible resulting outcomes a "sampling distribution" (it's the pattern of all possible samples, other texts use this name).

The "expected value" is the center or "mean" of the probability histogram.

The "standard error" is the spread of the probability histogram.

Until now, you have used histograms to represent data. You will now use histograms to represent chance (or probability). **A histogram that represents chance by area is called a *probability histogram (or sampling distribution).***

III. The Normal Approximation (18.4)

In a sentence: the normal approximation means replacing the probability histogram by the normal curve in Table A 105 before computing areas under the curve.

You will use probability histograms for the sum of draws from your box models. As the number of draws increases, the probability histogram for the sum of draws starts to resemble (or approximates) the normal curve. You can then use the normal curve to figure chances. This is often much easier than calculating the chances directly (this is what approximating does for you). For example, there is about a 68% chance that the sum of the draws will be within one SE when the number of draws is large. You know this from the normal curve.

In this section (p. 317-318), Freedman teaches you how to apply a "continuity correction" to the calculation of a Z score. It is useful if your probability histogram does not look very "smooth". You do not need to know how calculate Z scores this way in this class.

IV. On applying the Normal Approximation (18.5)

Section 5 in your textbook makes an important point: lopsided boxes need more draws before you can use the normal approximation. Furthermore, the normal curve approximation is for sums (and ultimately averages and percentages), but not other functions of the draws, such as products.

For example, take a lottery and graph it's possible outcome and also construct a box, you have a very lopsided graph and a lop-sided box model (in terms of fractions of winners and losers). Not drawn to scale, it looks like this:

To use the normal approximation for a probability histogram that does approximately follow the normal curve, you need only the expected value and standard error. To get these from a box model, you need the number of draws, the average of the box, and the SD of the box.

V. The Central Limit Theorem

Back to this fact: With enough draws, the probability histogram of a sum is approximated by the normal curve.

This leads us to the Central Limit Theorem: (p. 325 of your text)

"When drawing at **random** with **replacement** from a box, the **probability histogram for the sum will follow a normal curve**, even if its contents do not. The histogram must be put into standard units (Z) and the number of draws must be reasonably large.

This results in our ability to use the normal curve to make statements of chance or probability or percentages and to think of the areas under curve in Table A 105 as probabilities/changes instead of areas.

Large depends on what is in the box. The less normal it is, the more draws you need.

The central limit theorem explains why many distributions (e.g. height, eyesight, IQ, hearing, blood pressure, blood sugar) tend to be close to the normal distribution. The key ingredient is that the random variable being observed should be the sum or mean of many independent (not biased in any way) and identical (the chance is the same, draw after draw) random processes.

Note: Keep in mind the two different types of *convergence* discussed in chapter 18. With more and more repetitions of the draws (e.g. different tosses of size 10), the histogram for the data converges to the probability histogram (not all probability histograms are normal). With more and more draws from the box with each repetition (bigger and bigger sample sizes), the probability histogram *for the sum* gets smoother and smoother and approximates the normal curve better.