Discussion Sheet 13
The Central Limit Theorem

We now study the most important Theorem that let's us make inferences about populations from samples based on the normal distribution and distributions related to it.

**Random Sample**

If a sample of \( n \) things is selected from a population in such a way that every set of \( n \) things in the population has an equal probability of being selected, then that sample is called a random sample.

1. Suppose that all of a doctor's patient records are stored in alphabetical order by patient's last name. If a sample of 200 by is selected by taking first 200 records, would that be likely to be a random sample with respect to whether the patient had arthritis?

   Comment:
   This is likely to be a random sample since names are for the most part unrelated to type of illness — here arthritis. The one exception would be if arthritis had a strong hereditary component and then names in alphabetical order might contain several from the same family and thus a greater likelihood that others had arthritis if one did. However, without a strong hereditary component for arthritis this would be a reasonably random sample.

2. Suppose instead that the doctor's records above were stored by age and then alphabetically by patient last name. If a sample of 200 by taking first 200 records, would that be likely to be a random sample with respect to whether the patient had arthritis?

   Comment:
   This would clearly not be a random sample with respect to whether a patient had arthritis. There is a strong relationship between age and having arthritis. If the first records were those of younger patients then the probability of drawing one with arthritis would be less than would be true for the whole population of patients. If the first records were those of older patients then the probability of drawing one with arthritis would be greater than would be true for the whole population of patients. In either case, this procedure would be unlikely to yield a random sample with respect to having arthritis.

**Sampling Distribution**

Suppose that a large number of random samples in a certain situation is taken and if a particular statistic (e.g., the mean or the variance) is calculated for each sample taken. The set of statistics calculated in this way would have its own distribution, not necessarily that of the original population. That distribution would be called the sampling distribution of that statistic.

3. Suppose that a sample of 100 values is drawn from a population with a uniform distribution and the mean of that sample calculated. Suppose we did this for 5000 different samples of 100 values from that population. Would the sampling distribution are represented by those 5000 different means have a uniform distribution? What can we know about this sampling distribution?

   Comment:
   At this point all we can say is that it seems unlikely that the distribution of the 5000 different means would have a uniform distribution. In a mean, higher and lower values tend to be erased to find a value more in the center. This means that there would likely be more means in the center of the distribution and fewer high and low means. Statisticians call this kind of distribution "mound shaped". At this point we can't say much beyond that based simply on logic.
If we have the sampling distribution of a particular statistic, the mean of that sampling distribution is simply called the mean of the sampling distribution. The standard deviation of that sampling distribution is traditionally called the standard error of the sampling distribution.

4. When polls or surveys are reported on television they are usually labeled as accurate "plus or minus n%." That percentage is the standard error for the result reported which is usually a proportion of the voters or a certain group that believe or will do a certain thing. Can we tell anything about this standard error if we know the mean and standard deviation of the original population (not of any sample drawn from it) but don't know what kind of distribution it has?

Comment:
Not really, at least without further information. If the original population had a large standard deviation, then it seems likely that the standard error would be larger but we do not know that for sure.

5. Look Figure 7.9 on pages 304 and 305 of your textbook. This shows the relative frequencies of the mean of 1000 samples of size 5, 15, 25, 50, and 100 drawn from a population with a uniform distribution from 0 to 1 (see Figure 7.8 a at the top of page 304). Relative frequency is an approximation to the sampling distribution of a statistic, in this case the mean. What can you say about the mean of the sampling distribution for the samples of different sizes (5, 15, etc.)? What can you say about the standard error of the sampling distribution for the samples of different sizes? What can you say about shape of the sampling distribution for the samples of different sizes? How do these three things appear to change as the size of a sample gets larger?

Comment:
Notice that the vertical scales on the five figures change slightly and bear that in mind when making comparisons. It appears that the "top of the hill" and therefor the mean is somewhere around .5 in all of the figures. It doesn't change much. On the other hand as the sample size increases from figure to figure, the spread of the distribution appears to be smaller implying a smaller standard error. Further, even for a small sample size, the distribution appears to be mound shaped even though the original distribution was uniform. It seems that the distributions for larger sample sizes are even more sharply mound shaped.

6. For the uniform distribution shown at the top of page 304, what is its mean? What is its standard deviation? How do these relate to the means and standard errors of the sampling distributions approximated by the relative frequency histograms shown for samples of different sizes?

Comment:
We know that for a random variable, \( y \), with a uniform distribution and pdf

\[
p(y) = \begin{cases} \frac{1}{b-a} & a < y < b \\ 0 & \text{elsewhere} \end{cases}
\]

then the mean is \( \mu = \frac{a+b}{2} \) and the variance is \( \sigma^2 = \frac{(b-a)^2}{12} \). Here we have

\[
p(y) = \begin{cases} \frac{1}{1-0} & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}
\]

so the mean is \( \mu = \frac{a+b}{2} = \frac{0+1}{2} = .5 \) and the variance is \( \sigma^2 = \frac{(b-a)^2}{12} = \frac{(1-0)^2}{12} = \frac{1}{12} \) so the standard deviation is \( \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{12}} = .289 \). The means of the sampling distributions pictured are about the same as the mean of the uniform distribution, .5. The standard errors change and may well
be smaller than the standard deviation of the original uniform distribution. In fact, from Figure 7.9 (all parts except (a)), the entire spread of the distribution is between \(0.5 - 0.289 = 0.211\) and 
\(0.5 + 0.289 = 0.789\) so the standard error is certainly smaller than 0.289 (since the spread is several standard errors wide.

7. Look Figure 7.11 on pages 306 to 308 of your textbook. This shows the relative frequencies of the mean of 1000 samples of size 5, 15, 25, 50, and 100 drawn from a population with a standard normal distribution (see Figure 7.10a at the top of page 306). What can you say about the mean of the sampling distribution for the samples of different sizes (5, 15, etc.)? What can you say about the standard error of the sampling distribution for the samples of different sizes? What can you say about shape of the sampling distribution for the samples of different sizes? How do these three things appear to change as the size of a sample gets larger?

Comment:
Again, it appears that all of the means of the sampling distributions are about 0. The standard errors appear to decrease as the sample size increases since the spread of the histograms is smaller. The distributions are mound shaped and more sharply so for larger sample sizes.

8. For the normal distribution shown in Figure 7.10a, what is its mean? What is its standard deviation? How do these relate to the means and standard errors of the sampling distributions approximated by the relative frequency histograms shown for samples of different sizes?

Comment:
We are given that mean is \(\mu = 0\) and the standard deviation is \(\sigma = 1\) (that is, it is a standard normal distribution). The mean of the sampling distributions appear to be about 0, the same as that of the normal distribution, regardless of how large the sample size is. The standard error appears to be getting smaller as the sample size gets larger and is certainly not \(\sigma = 1\) as in the original population. We have \(\mu - 1\sigma = 0 - 1 = -1\) and \(\mu + 1\sigma = 0 + 1 = 1\) but in all parts of Figure 7.11 except (a) the entire spread of the distribution is less than from -1 to 1 yet is likely to be several standard errors wide. So the standard errors are likely to be smaller than \(\sigma = 1\).

9. Look at Figure 7.12 on pages 308 to 310 of your textbook. This shows the relative frequencies of the mean of 1000 samples of size 5, 15, 25, 50, and 100 drawn from a population with an exponential distribution (see Figure 7.10b at the top of page 306). What can you say about the mean of the sampling distribution for the samples of different sizes (5, 15, etc.)? What can you say about the standard error of the sampling distribution for the samples of different sizes? What can you say about shape of the sampling distribution for the samples of different sizes? How do these three things appear to change as the size of a sample gets larger?

Comment:
The mean of the sampling distribution appears to be about 1 regardless of the sample size. The standard error of the sampling distribution again appears to be smaller for larger sample sizes. The distributions are mound shaped and fairly symmetric and even more so as the sample size is larger.

10. Based on your answers to Questions #5 through #9, what would you say is true in general about the means of a sampling distribution of means regardless of the probability distribution of the original population? Of the standard error of the sampling distribution? Of the shape or type of the sampling distribution? How does this change in general as the samples involved in calculating the statistic get larger?

Comment:
The means of sampling distributions appear to be about the same as the means of the underlying population regardless of sample size. The standard errors appear to be less than the standard deviation of the underlying population and to be smaller for larger sample sizes. The distributions of
the sampling distributions are mound shaped (something like a normal distribution) and get more sharply mound shaped and symmetrical as the sample size increases. This is true whether we start with an underlying population that is uniform, normal, or strongly skewed likely the exponential distribution and draw samples from that population.

### Central Limit Theorem

If a random sample of \( n \) observations, \( y_1, y_2, \ldots, y_n \) is drawn from a population for a random variable \( Y \) with a finite mean \( \mu \) and variance \( \sigma^2 \) then \( \bar{Y} \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 / n \) for sufficiently large \( n \) regardless of the distribution of the original random variable \( Y \).

Note: "Sufficiently large" turns out to be about \( n \geq 30 \).

11. Based on the Central Limit Theorem, how would you answer the basic questions posed in Question 10?

   For all sample sizes, the sampling distributions should have a mean, \( \mu \), equal to that of the original underlying population. They should have standard errors that are smaller for larger sample sizes since the standard error for the sampling distribution of samples of size \( n \) is \( \sigma^2 / n = \sigma / \sqrt{n} \). The sampling distributions are not only mound shaped but approximately normal, with a better approximation for larger sample sizes.

### Worksheet 13 (§7.1-§7.2, §7.4, §7.7)
The Central Limit Theorem and Distributions Related to the Normal

This worksheet is a continuation of Discussion Sheet 13. Please complete that Discussion Sheet if you haven’t already. Discussion Sheet 13 and this worksheet combined to cover §7.1-§7.2, §7.4, and §7.7 of the textbook. Be sure to read those sections.

1. Suppose \( y \) is a random variable representing the true weight of a 16 oz can of coffee. Suppose that the mean of \( y \) is 16.1 oz. and that the standard deviation is 0.2. What is the mean or expected value of \( \bar{y} \), that is, what is \( E(\bar{y}) \) in a sample of 10,000 cans of this coffee?

   Comment:
   We know \( \mu_y = 16.1 \) and \( \sigma_y = 0.2 \). From the Central Limit Theorem we know that the expected value of \( \bar{y} \), that is \( E(\bar{y}) \), for a sample of 10,000 cans should be
   \[
   \mu_{\bar{y}} = E(\bar{y}) = \mu_y = 16.1
   \]

2. What is the standard deviation of the mean, \( \sigma_{\bar{y}} \), for 10,000 cans of this coffee?

   Comment:
   Since \( \sigma_y = 0.2 \) and by the Central Limit Theorem for a sample of size \( n \) is \( \sigma_{\bar{y}} = \sigma_y / \sqrt{n} \), we should have \( \sigma_{\bar{y}} = \sigma_y / \sqrt{n} = 0.2 / \sqrt{10000} = 0.2 / 100 = 0.002 \).
3. What kind of distribution is the distribution of the means for all the possible lots of 10,000 cans of this coffee?

Comment:
Since we can safely assume that 10,000 \((n \geq 30)\) is "sufficiently large", then the distribution of the sampling distribution for samples of size 10,000 should be normal or a close approximation of normal.

**Sampling Distributions related to the Normal Distribution**
There are not an infinite number of different distribution types you will need to learn besides the binomial distribution and the normal distribution. We use basically four others: the c2 distribution, the t distribution, and the F distribution (see pages 322-323). We will use tables for each of these and learn more about them as we need them.