Stat 13, Tue 5/8/12.

1. Collect HW 4.
2. Central limit theorem.
3. CLT for 0-1 events.
4. Examples.
5. σ versus $\sigma/\sqrt{n}$.
6. Assumptions.

Read ch. 5 and 6. Ignore the normal approximation to the binomial.
Hw5 is due Tue, 5/15. Midterm 2 is Thur, 5/17.

2. Central Limit Theorem (CLT).

Suppose we're interested in the avg income in the Los Angeles population. We think it's $36000.

We sample 100 people and calculate their mean income, $\bar{x}$. Say $\bar{x}$ = $35000.

That sample mean $\bar{x}$ will always be a bit different from the population mean, μ. There is a sense in which μ could be $36000, and yet $\bar{x}$ could be $35000, just by chance.

The question is: what is the chance of this happening? How much different should be expect $\bar{x}$ and μ to be?

To get at this question, we have to think about sampling over and over, repeatedly, each time taking 100 people at random. For each sample, we'll get a different $\bar{x}$.

Imagine looking at this list of all these different $\bar{x}$s.

It turns out that, provided n is large (here n = 100), and provided each sample is a SRS (simple random sample), then:

this list of $\bar{x}$s is NORMALLY DISTRIBUTED with mean μ, and the std deviation of the $\bar{x}$s is $\frac{\sigma}{\sqrt{n}}$

where σ is the std deviation of the population.

That's called the CLT (central limit theorem) or the NORMAL approximation.

What does this mean?

It means that the sample mean is typically about μ, but off by around $\sigma/\sqrt{n}$. It means that 68% of the times we sample, the sample mean is within $\sigma/\sqrt{n}$ of μ.

Note that this doesn't depend on what the population looks like. Even if the population is not normally distributed, even if it's all 0s and 1s, we can use the normal table to see what the probability is that the sample mean $\bar{x}$ is in some range.

3. CLT for proportions.

What about if you have a population of 0s and 1s, and you sample from this population and are interested in the total number of 1s in your sample. Your book calls this total X. e.g. the number of people in your sample with type A blood. Usually we're not interested in X: we're interested in the PERCENTAGE, $\hat{p}$ with A blood in the sample, and we want to use this to estimate the PERCENTAGE, p, with type A blood in the population.
The formulas for X in the book are unnecessary and are also potentially misleading. They don't matter once you realize $\hat{p}$ is $\overline{x}$ and p is μ for such problems. If we view each value as 1 or 0, then the population percentage (p) is the same as the population mean μ. And the sample percentage ($\hat{p}$) is the same as the sample mean $\overline{x}$ of all the 0s and 1s in the sample. So we don't ever need separate formulas for $\hat{p}$ and p. They are special cases of the formulas for $\overline{x}$ and μ, where the data happen to be 0s and 1s.

Again, the CLT says that $\overline{x}$ is normally distributed with mean μ and the std deviation of $\overline{x}$ is $\frac{\sigma}{\sqrt{n}}$. This $\frac{\sigma}{\sqrt{n}}$ is called the standard error for $\overline{x}$.

Remember though that σ is the std dev of the population. If all the values in the population are 1's and 0's, then σ is the std dev of those numbers. In such situations, $\sigma = \sqrt{pq}$ where p = percentage of ones in population, and q = 1 − p = percentage of zeros.

Forget the book's equations for $\hat{p}$ and especially the equations for x which might confuse you. They have a crazy formula for the SE for a proportion involving n+4 instead of n. I have no idea where they got that from. Ignore it and use n.

4. Examples.
The average cow produces 20 pounds of wet manure per day, with a sd of 6 pounds.

Suppose we take a SRS (simple random sample) of 144 cows.
What is the probability that our sample mean will be greater than 21 pounds?
This is a SRS, and n is large, so $\bar{x}$ is N(20, 6/√n) = N(20, .5). So we're asking what the chance is that **a draw from a normal distribution with mean 20 and std dev 0.5 will be at least 21**.
You know how to do this (subtract the mean and divide by the sd): this is the chance that a draw from a N(0,1) will be at least (21−20) ÷ 0.5 = 2.00, which from the tables is 1 − 0.9772 = 0.0228, or 2.28%.

What is the probability that the amount for one individual cow will be greater than 21?
Can't answer that question. Not enough info.
But suppose we also knew that the population of cow excretions was normally distributed. Then we could do it.
The answer would be: the chance that a draw from a N(20, 6) is at least 21, which is the chance that a N(0, 1) is at least (21-20) ÷ 6 = 0.17, which from the tables is 1 − .5675 = 0.4325, or 43.25%.

Another example.
Suppose that 90% of the cows are dairy cows. In a SRS of 144 cows, what's the probability that at least 80% of the cows in our sample are dairy cows.
Answer: this is P($\hat{p}$ > 0.80) = P($\bar{x}$ > 0.80),
which is the probability that a draw from a N(0.9, $\frac{\sigma}{\sqrt{n}}$) is > 0.80.
What's $\frac{\sigma}{\sqrt{n}}$?
n = 144.
$\sigma = \sqrt{pq} = \sqrt{(0.9)(0.1)} = \sqrt{0.09} = 0.3$.
So, $\frac{\sigma}{\sqrt{n}}$ = 0.3/12 = 0.025.

We want the probability that a draw from a N(0.9, 0.025) is at least 0.80.
This is the probability that a N(0, 1) draw is at least (0.80−0.9) ÷ 0.025 = −4.00.
-4.00 is off the charts. For the closest number on the chart, the probability of being greater than that number is 0.9998 or 99.98%, so all we can say is the probability is at least 99.98%.

5. σ versus $\frac{\sigma}{\sqrt{n}}$.

The issue here is whether you are interested in one draw or a sample of draws.

σ tells you how much ONE DRAW typically differs from the mean.

$\frac{\sigma}{\sqrt{n}}$ is how much $\bar{x}$, the mean of SEVERAL DRAWS, typically differs from the mean.

6. Assumptions.
The assumptions are important here.
The CLT says that if we have a SRS, n is large, and σ is known, then $\bar{x}$ ~ N(μ, $\frac{\sigma}{\sqrt{n}}$).

Don't need SRS if the obs. are iid (independent and identically distributed) with mean μ.
Don't need n large if population is normal.
But, if n is small and pop. is not known to be normal, we're stuck. Need more info.

Large n?
For non-0-1 problems, n ≥ 25 is often large enough. Let's use this rule of thumb.
For 0-1 problems, the general rule of thumb is that $n\hat{p}$ and $n\hat{q}$ must both be ≥ 10.
In other words, there must be at least 10 of each type in the sample.

If n is large and you don't know σ, you can plug in s, the sd of your sample.
If n is small, pop. is normal, and you use s for σ, then the distribution is t, not N.

Next: Confidence Intervals (CIs).