Figuring Out Standard Errors

Here's another way to look at standard errors. First look closely at the formulas and where the numbers come from. The formulas are a little different than how things are presented in the book or in class so that I can show how similar they are underneath it all:

	In the box		**In the sample**		
Start with:	**Center**	**Spread**	**Value of interest**	**How calculated**	**Estimate of sampling error**
Sum	Average of box	Standard deviation of box	Total sum from draws	# of draws * mean of box	$\sqrt{\#\ of\ draws} * SD\ of\ box$
Count	# correct/# of outcomes	$\sqrt{\frac{\#\,correct*\#\,incorrect}{\#\,outcomes*\#\,outcomes}}$	Total count correct	# of draws*P(correct outcome)	$\sqrt{\#\ of\ draws} * SD\ of\ box$
Percent	# correct/# of outcomes* 100%	$\sqrt{\frac{\#\,correct*\#\,incorrect}{\#\,outcomes*\#\,outcomes}}$	Percent correct	# of draws*P(correct outcome) * 100%	$\frac{\sqrt{\#\ of\ draws} * SD\ of\ box}{\#\ of\ draws} * 100\%$
Average	Average of box	Standard deviation of box	Average of the draws	$\frac{\#\ of\ draws * mean\ of\ box}{\#\,of\ draws}$	$\frac{\sqrt{\#\ of\ draws} * SD\ of\ box}{\#\ of\ draws}$

So one thing you should notice is that all four of these start out with either the average or mean of the box or the fraction of outcomes that meet your criterion of interest (classifying and counting) All four calculate a SD of the box, though there are two ways in which this is done. All four standard errors multiple the center of the box by the number of draws, though two do one additional step with the outcome of interest. And all four calculate the SE by multiplying the SD of the box by the squareroot of the number of draws. Two of the four then take it one additional step.

Now, how would you use this?

Let's start first from a box: Imagine a town that has 30,000 homes. You are going to select a simple random sample of 100 homes to examine. How many people live in those 100 hundred homes. Can you guess before you draw the sample? Do you feel comfortable with your guess?

Well if I tell you the following: On average in this town there are 2.5 people per home, SD 1.5. Then you could guess:

100 homes sampled * 2.5 per home = 250 people

But you might be off a little bit. How much? Well going back to our original information, just by chance you would estimate that you might be off:

$$\sqrt{100\ homes\ sampled} * 1.5 = 15\ people$$

That is, in this town of 30,000 homes if you were to select 100 homes at random, you would expect to find 250 ±15 people (about 2/3 of the time--where does this come from? The normal curve…you've gone out 1 SE to either side of your estimate) living in those 100 homes.

How would you use this information in the real world? Well, what if you had to plan for emergency services for this town? What if you had to predict before something happened how many shelters you would have to create in case people needed a place to sleep in if some percentage of the town (like 100 out of 30000 homes were completely destroyed)? Your best bet is 250, but just to be on the safe side you might add a few more slots. How many, how about 30 (2 SE from the estimate)? That ought to cover about 95% of the possibilities (or chance error in sampling 100 homes). See how much better off you are with a reasoned guess. If not you might have figured 250 beds for people plus a 100 estimated by a gut feeling just to be safe…far, far more costly and unnecessary.

Let's ask a different question. How many of these 100 homes include at least one child of school age? Let's make an informed guess…

Can you do it without any information? Pretty hard. It could be 100 homes; it could be none (maybe this town is a retirement village…). What if I tell you that 1/2 of homes in this town contain at least one child of school age.

Well then here's the prediction: You expect to see 100 * 0.5 = 50 of the homes with at least one child of school age.
50 exactly???? Well, maybe not. But close to 50. How close?????

$$\sqrt{100\ \textit{homes sampled}} * \sqrt{0.5 * 0.5} = 5\ \textit{homes}$$

So the prediction would be that out of 100 randomly selected homes you would expect 50 ± 5 homes to contain at least one child of school age.

Again, this might be very useful in city planning.

Now, let's ask questions differently. What if you didn't know the population at all? What if you only knew the sample? How would you use Standard Errors?

Well, imagine that you work for the public health service. You want to know how many children in California are immunized for measles. What percent? Give it a guess? Feeling lucky? No idea? Well, you're not alone in that sensation.

Ok, now let's say you take a simple random sample of 100 children in California and find out that 82% are immunized for measles.

Now, what's your guess about all Californian children (the box…).

Are you certain that it is exactly 82%? Well, no. It should be close to 82%, but how close?

Well, the SE for the percent is:

$$\frac{\sqrt{100\textit{children}} * \sqrt{0.82 * 0.18}}{100\ \textit{chilren}} * 100\% = 3.8\%$$

So a 95% CI would be 82% ± 2 * 3.8% or about 74% to 90%

That is 95% of time this interval that you have just created (using the observed percent and the estimated SE) will include the value for the parameter (the true percent of Californian children who are immunized). It's a more educated and reasoned guess than either a number off the top of your head, or the 82% you observed in your sample in that it gives a range of possible and reasonable values. But note two things about it: 1> you're going to be wrong 5% of the time when you make an interval like this, and you can't possibly know whether or not this is one of those times, and 2> the true percent of California children who are immunized is fixed, real, and unmeasurable. It is either in this interval or it is not. The 95% chance only refers to our chances of being correct in creating an interval that includes the population value. It doesn't mean there is a 95% chance that the population percent is in the interval. It does mean you have a 95% of creating an accurate interval.

Let's take another situation. You are planning to go to grad school and need to figure out how much it is going to cost you so that you can begin to hit up family members for money. In the absence of any information, how do you make your prediction? Well, it's a three year program, you figure it'll cost you $25,000 a year just going from what goes on with you, so off the top of your head you estimate $75,000. You mention it to your parents and they start coughing loudly (choking actually).

Well, that's one estimate. But there are many things you may not know. For example, there may be fellowships, grants, awards. There may be paid-internships or all sorts of paid learning opportunities like ta'ing (not uncommon in graduate training unlike undergrad years) that cut the personal financial cost of education.

You take a poll of 100 recent graduates of the program you are planning to enter (assume it is a simple random sample of new graduates from the population) and find that they average $20,000 out of pocket for their graduate education (SD = $10,000). How much do recent graduates spend out of pocket on average for going to grad school?

Well, now you can get a much closer, more accurate estimate. Because the sample mean is an element in the theoretical sampling distribution of means, and because the mean of the sampling distribution of the means approaches the mean of the population over repeated drawing of samples (this is a part of the central limit theorem…central means mean, limit means approaches or becomes), then your best bet for the mean of the sampling distribution of the means is a known value in that sample (which your observed mean is). That would be like, I have a distribution of numbers, give me a bet as to it's mean. Don't know? Ok. I 'll tell you 1 and only 1 number in the distribution of numbers--100,979. Now, tell me the mean of the distribution. Well, you still don't know, but my guess is you will guess the number I just gave you. Why, because intuitively you know that a mean is contained within the range of the numbers in the distribution (somewhere between the lowest and highest value), so knowing at least one number gives at least some chance of being right, given that your other choice is to choose one number from an infinite number of possibilities.

So you might bet that the average amount of out-of pocket cost would be $20,000. (look back at the formula in the table to track how you got this so that you can match your intuition to the formula).

$20,000 exactly? Well, close. How close? Remember, you need to plan.

The SE or estimate of chance or sampling error in predicting this mean is:

$$\frac{\sqrt{100\ \textit{grads polled}} * \$10{,}000}{100\ \textit{grads polled}} = \$1000$$

So a 95% CI would be $20,000 ± 2 * $1,000 or about $18.000 to $22,000.

What does this mean? Your bet for the average amount of out-of-pocket expense by new grads in your chosen field is that it lies somewhere between $18k and $22k. You are simply betting on what you think the average amount. Not the range (for some it costs nothing, for some $100k), not the average deviation among the population of new grads (that clearly would have to be far greater than the interval you've created…how do you know? Well in your sample alone 2/3 of those your poll report approximately being out-of-pocket between $10k and $30k, 95% $0-$40k, so it is hard to imagine that it would vary only $4k in the population…if you can't see this, go back and figure it out…it comes from the mean, the SD, and the normal curve. And thinking about the population, it's pretty hard to believe there is such homogeneity in cost).

Again, your 95% confidence interval refers to your chances that you have created an interval that includes the true population mean. The true population mean is either 100% inside the interval or not. It doesn't move around. It simply is, exactly where it is. The chances are in your work, your bet, your estimation. Most of the time you are right, rarely you are wrong, but this time you are either right or wrong and there's no way of knowing which it is. Just like flipping a coin. Over the long haul it'll come up heads 50% of the time, but every time you flip it, you don't what will happen. You'll either guess right or guess wrong and you don't know which is going to be.

Notice how tight your prediction is now. Far better than the guess you made just thinking about your own life ($75k). It now takes into account the actual experiences of 100 of those going through the program ($20k) and allows for some slop due to chance in selecting those 100 people ($18k-$22k).

So why do we use SE's? Because they help us to make, more often than not, more accurate estimations about either what will happen if we start from a certain place (like a town of 30,000 homes) and try to predict what will happen or if we start from what we observe (like a sample of 100 children) and try to estimate what is true for everyone. In the former, we are predicting what will occur. In the latter, we are estimating what is true in something we cannot measure.