Sampling distributions and estimation

I. What is a sampling distribution?

A) Review: we have discussed the distribution of a random variable. For instance:

i) We're interested in Pr{three sixes when throwing a single dice 8 times}, => Y has a binomial distribution, or in "official notation", $Y \sim BIN(n,p)$. (The symbol "=>" means "implies")

- why does Y have a binomial distribution? Because I'm interested in:

Pr{three sixes in 8 tosses} which is similar to *Pr{three tails in 8 tosses}*. The only difference here is that the probability of success is 1/6, not ½.

- other binomial examples:

- Pr{3 mutants in a sample of 5}

- Pr{8 people with blood type O in a sample of 20}

- in all cases Y = number of successes, and $Y \sim BIN$.

ii) We're interested in $Pr\{Y \geq 120\}$ where Y is IQ of a randomly chosen person. IQ is normally distributed, and so $Y \sim N(\mu,\sigma)$.

- other normal examples:

- $Pr\{Y \leq y\}$ where y is a specific blood oxygen level. Assuming blood oxygen level is normally distributed (actually, it probably isn't).

- height - a classic example of a normally distributed random variable.

iii) Other distributions:

- Poisson (we discussed this)

- Uniform

- e.g. throwing a die once, the distribution of Y is uniform (any number has the same probability of coming up)

- random number generator, or random number tables are uniform - any number is equally likely.

iv) But we already talked about this. So what's the point here?

- Usually we are not interested in the probability of an individual number, (the examples above). Instead, we are interested in the distribution of the parameters

estimates (or sometimes functions of the parameter estimates).

- What is the distribution of the sample mean?

- What is the probability that the AVERAGE height is greater than *y*? That's much more interesting!

B) So, when we talk about a sampling distribution, we're interested in the distribution of the statistics calculated from our sample, such as the sample mean, sample standard deviation, sample median, etc. If we know this distribution, we can then calculate probabilities.

II. Another way of looking at it (similar to book):

- Let's do an experiment many times.

- Each time we calculate the sample mean.

- What is the distribution of these sample means? [Your book uses the term "meta-experiment"].

- Something else we want to know though:

  - what is the mean of all these sample means?

  - what is the variance and standard deviation of our sample means?

  - also, how do we estimate these?

- We don't need to do *k* number of experiments for this to work. Instead we can use some mathematics and figure it out.

  - the math is similar to that illustrated for the "mathematical" definition of a probability distribution mean (in other words, we use *expected values* to calculate this). We won't go into the details here.

III. So here are the "answers".

- What is the (theoretical) mean for $\bar{Y}$ (the sample mean)?

  - The mean of $\bar{Y} = \mu$

- The (theoretical) standard deviation for $\bar{Y}$:

  - is equal to σ over the square root of *n*

- Here are the equations:

$$\mu_{\bar{Y}} = \mu \qquad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

- See also p. 159 **[159]** *{151}* in your text.

- Are you surprised by the first result? Hopefully not. The second result you'll have to take on faith or do a bunch of math. But think about it intuitively this way:

    - If we do a particular experiment *n* times, you'd have *n* samples each of size *n*. In that case, you might expect to have *n* times *n* in the denominator (after all, you now have $n \times n$ as much data). This is *not* how the math works, but it might give you a feel for what's going on.

- What about the actual distribution of $\bar{Y}$?

    - If *Y* is normal, $\bar{Y}$ is normal (again, hopefully no surprise).

    - If *Y* is not normal, then in almost all cases if *n* is moderately large, then $\bar{Y}$ is still approximately normal!

        - This last result is due to the Central Limit Theorem (CLT). It is one of the reasons that the normal distribution is so important to statistics. In other words, take almost any distribution at all. Take the average. The average will be normally distributed if n is moderately large. We may demonstrate this in recitation.

- What does all this imply? That we can use our normal tables (well, sort of - actually we'll usually wind up using a t-distribution, but more on that next time) to calculate probabilities associated with $\bar{Y}$.

    - What is the Pr{ $\bar{Y} < 6$ feet}? etc.

    - This is much more interesting than just one person being less than 6 feet. Eventually we'll be able to use this approach to ask questions like:

        - is the average height of men and women different? (In probability terms, "what is the probability that the height of men and women is different?")

        - this will lead us directly into confidence intervals and from there into statistical tests and procedures (finally!).

IV. Some comments on your text:

- These concepts are really important, so if you didn't follow everything in class, make sure you go through the text here. It's probably a good idea to read section 5.1 **[5.1]** *{5.1}*. If you have the 2nd or 3rd edition, read through 5.3 **[5.3]**, if you have the 4th edition, read through *{5.2}*.

V. Some examples that illustrate the use of this stuff:

- Example 5.7, p. 160 **[5.9, p. 159]** *{5.2.2, p. 152}*.
- seed weight is normal, $\mu = 500$, $\sigma = 120$ mg.

- sample 4 seeds. What is the mean of $\bar{Y}$ for our four seeds? What is the standard deviation?

$$\mu_{\bar{y}} = \mu = 500\text{mg}$$

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{4}} = 60\text{mg}$$

- Now suppose we want to calculate $\Pr\{\bar{Y} > 550 \text{ mg}\}$.

- You proceed the same way as before, except you need to make sure you use σ for $\bar{Y}$, NOT σ for Y.

- Continuing with the above example in your text.

$$z = \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} = \frac{550 - 500}{60} = 0.83$$

$$so \quad we \quad have$$

$$Pr\{\bar{Y} > 550\} = Pr\{Z > 0.83\} = 1 - 0.7967 = 0.2033$$

- note that we use 60, not 120 in the denominator above.

- If you have the 4th edition, you might also want to check out example *{4.2.4, p. 155}*, which explains the effect of sample size.

VI. Standard Error

(we skipped over the rest of chapter 5 and are now in chapter 6!)

- Remember, we almost never know μ and σ.

- So, we estimate μ with $\bar{y}$, and σ with s. We've discussed this before.

- But, how does all this affect the above if we don't know μ or σ? In particular, what about the standard deviation for $\bar{Y}$??

Well, instead of using σ we use s. In other words, we estimate:

$$\frac{\sigma}{\sqrt{n}} \quad with \quad \frac{s}{\sqrt{n}}$$

This quantity is also known as the Standard error of the mean:

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

So what does the Standard error tell us? Basically how reliable our $\bar{y}$ is as an estimate of the mean. If the SE is small, our $\bar{y}$ is a pretty good estimate; if it's big, our $\bar{y}$ is a lousy estimate. More on this next time.

- The text makes a big fuss about Standard deviation vs. Standard Error. Let's not get bogged down in the details. The basics:

- the standard deviation refers to dispersion associated with *Y*.

- the standard error refers to dispersion associated with $\bar{Y}$ (how good is $\bar{Y}$).

- it is important to know how some of these things are affected by sample size, though.

- Example 6.4, p. 185 **[6.4, p. 182]** *{6.3.2, p. 173 - 174}*:

28 lambs were weighted at birth with the following results:

$\bar{y}$	= 5.17 kg.	
s	= .65 kg.	
SE	= .12 kg.	(you can do the math yourself)

Now what would happen if instead of 28 lambs, we sampled higher numbers of lambs?

- go through fig. 6.3 p. 186 **[6.2, p. 182]** *{6.2.2, p. 174}*