Sampling distributions and the standard error

Sampling distributions.

In the last chapter we briefly discussed distributions other than the normal and binomial. We also (very briefly) discussed expected values. This idea of *expected* values will now let us move from calculating probabilities for a single variable (Y) to a group of variables represented by ($\bar{Y}$).

In statistics we are usually not that interested in the probability that Y is less than some value. Rather, we are interested in the probability that $\bar{Y}$ is less than some value. For example, we're not very interested in the probability that an individuals blood pressure dropped by 10 points (systolic). The probability that the *average* blood pressure (represented by $\bar{Y}$) dropped by 10 points is much more interesting.

To do this for a single observation we needed to know (1) the distribution of Y and (2) the parameters of Y (the parameters of the distribution for Y). So if we're interested in $\bar{Y}$, we need to know the same two things. What is the distribution of $\bar{Y}$, and what parameters do we need to know?

As you might guess, the distribution of $\bar{Y}$ depends to some degree on the distribution of Y. Suppose we know that $Y \sim N$. What does that imply for $\bar{Y}$? In other words, $\bar{Y} \sim ?$.

We can't show a mathematical proof of the following in an introductory class, but let's look at the answers:

If $Y \sim N \implies \bar{Y} \sim N$.

If $Y \nsim N \implies \bar{Y} \sim N$ if n large enough (but see below).

(Remember they symbol "$\implies$" means *implies*).

The first result is probably somewhat intuitive. The second requires some advanced calculus (as mentioned, we can't prove these results here).

The second result is due to something called the ***C**entral **L**imit **T**heorem* or CLT. It is one of the reasons that the normal distribution is so important in statistics. If we are really interested in $\bar{Y}$, then we can use the normal distribution to figure out probabilities *regardless* of what the original distribution looks like.

There are, however, two important considerations. One practical, one theoretical:

1. How big does n have to be? Well that depends how badly not normal the distribution of Y is. We'll learn how to evaluate this later.

2. In theory, the CLT only works if our original distribution has a mean (i.e., if the distribution of Y has a mean, μ). This generally not an issue in practice, but it is something you should be aware of (an example of a distribution without a mean is the Cauchy distribution - look it up on Wikipedia if you're interested).

Okay, so now we know a little about the possible distribution of $\bar{Y}$. Let's assume for now that $\bar{Y}$ does have a normal distribution, so we know we need to know μ and σ for $\bar{Y}$ to figure out probabilities associated with $\bar{Y}$.

In other words, we need to know $\mu_{\bar{Y}}$ and $\sigma_{\bar{Y}}$. To do this, we remember the section on expected values. The equations for expected values (and variances) can be used to calculate not just the mean of Y, but also the mean of $\bar{Y}$ (similarly for the variance). All we need to do is substitute $\bar{y}$ for y in the first part of the equation and solve. Since this is way too complicated for an introductory class, we'll just look at the answers:

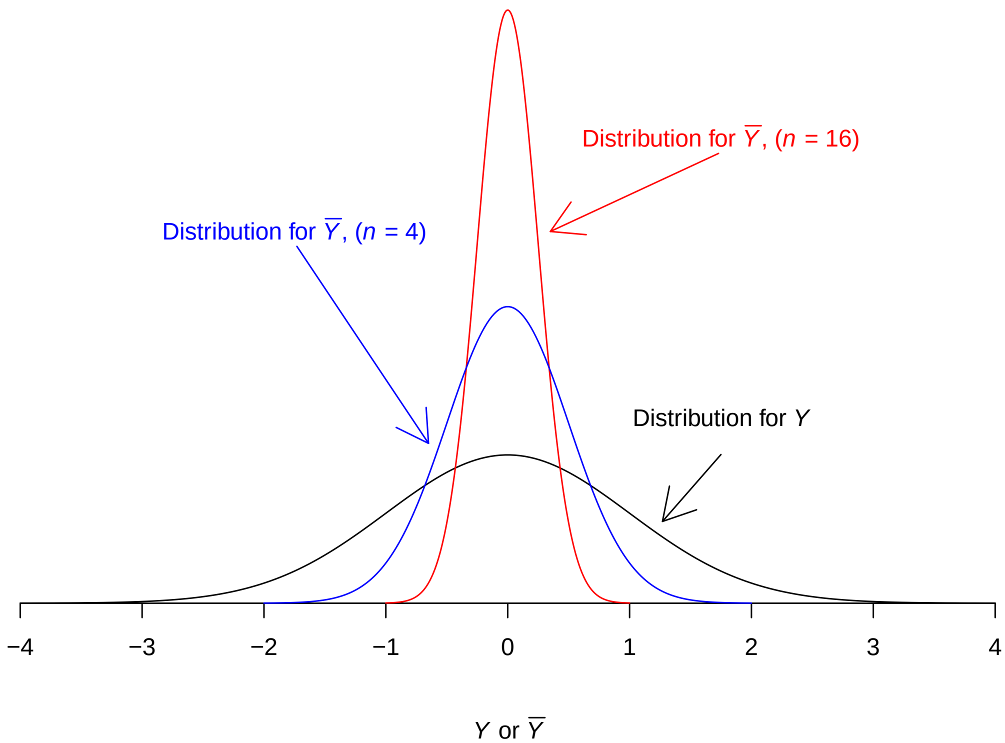
$$\mu_{\bar{Y}} = \mu_Y = \mu.$$

$$\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n} = \sigma/\sqrt{n}.$$

Are you surprised by the first result? Hopefully not. If $\bar{y}$ estimates μ, then $\bar{\bar{y}}$ should estimate μ as well (using $\bar{\bar{y}}$ as the sample mean of our $\bar{Y}$'s).

The second result is a bit more difficult to understand as it's not quite as intuitive. Probably the best way to think about it is that if our sample size increase, $\bar{Y}$ should become less variable. For example, if you have $n = 1$, then $\bar{y} = y$ and the variability of $\bar{Y}$ is the same as the variability of Y. But if our sample size increases, $\bar{Y}$ becomes less variable.

Or, to put it another way, the $\bar{y}$'s are closer to each other than the y's are to each other. Here is a graphical representation of what's going on:



So, now we've figured out everything we need to start calculating probabilities using $\bar{Y}$ instead of Y. We know the distribution of $\bar{Y}$ and we know the parameters of $\bar{Y}$. Let's do an example.

Suppose we have secret knowledge and know that for women the true average height (μ) is 64 inches and the true standard deviation (σ) is 3.5 inches. In other words we have:

$\mu = 64$ inches, and

$\sigma = 3.5$ inches.

Let's calculate the probability that a single woman is more than 6 feet (= 72 inches) tall:

$$z = \frac{y - \mu}{\sigma} = \frac{72 - 64}{3.5} = 2.29$$
$$\Longrightarrow$$
$$Pr\{Y > 72\} = Pr\{Z > 2.29\} = 0.0110$$

So the probability of a woman being more than 6 feet tall is 1.1%, a small but real number.

But let's suppose we now get the height of 5 women. What is the probability that the *average* height of our 5 women is more than 6 feet? In other words we want $Pr\{\bar{Y} > 72\}$.

Here's how we do things. First we note that we have:

$\mu_{\bar{Y}} = \mu = 64$ inches, and

$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{5}} = 1.5652$ inches.

We can calculate z just as before, but now we need to be a little careful with our denominator:

$$z = \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{72 - 64}{\frac{3.5}{\sqrt{5}}} = 5.11$$
$$\Longrightarrow$$
$$Pr\{\bar{Y} > 72\} = Pr\{Z > 5.11\} = 1.61 \times 10^{-7}$$

Notice the absurdly small value for the probability (we had to use R to get it!). This is because it is *extremely* unlikely that any group of 5 women (sampled at random) will have an average height greater than 6 feet.

What you should remember from all this is that once we start looking at the probabilities of $\bar{Y}$, the probabilities we get can be drastically different from those for Y.

Introducing the standard error.

We just learned how to calculate probabilities for $\bar{Y}$. Unfortunately, most of what we learned isn't terribly useful. The reason for this is that we hardly ever know μ or σ. Notice that this is true not just for probabilities associated with $\bar{Y}$ but also for probabilities associated with Y. We don't have magical abilities to figure out μ and σ so we need to start thinking about how we can calculate probabilities without knowing the parameters of our distribution.

As it turns out, not knowing μ is not usually a problem. The reason for this is that we make guesses about μ, so not knowing μ is okay. But to make guesses about μ we still need to know σ, and as mentioned, we don't know σ.

This is where the standard error comes in. Since we don't know σ, we have to estimate σ with s. In other words, when we try to calculate probabilities, we can't use $\frac{\sigma}{\sqrt{n}}$ in our denominator since we don't know σ. Instead we will use $\frac{s}{\sqrt{n}}$. This latter quantity is very important and is called the ***S**tandard **E**rror*, or SE of $\bar{Y}$. We have:

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

So what does the Standard error (SE) tell us? It tells us how reliable (how good) our $\bar{y}$ is. If our SE is small, our $\bar{y}$ is probably doing a good job estimating μ. If our SE is large, our $\bar{y}$ isn't doing so good.

What does the SE not tell us? It does *not* measure the variability of Y. The sample standard deviation s does that. It is easy to get confused about this. Incidentally, the standard deviation of $\bar{y}$ is the same as $SE_{\bar{y}}$, although we usually don't refer to this as the standard deviation of $\bar{y}$.

$$s_{\bar{y}} = SE_{\bar{y}} = \frac{s}{\sqrt{n}} \quad (\neq s \text{ or } s_y)$$

It is also important to realize that the SE can vary depending on what we are interested in. Right now we want to know the variability of $\bar{Y}$. Later we will be interested in such quantities as the variability of $\bar{Y}_1 - \bar{Y}_2$. This quantity will have a different SE (we will designate it $SE_{\bar{y}_1 - \bar{y}_2}$). More about this will be explained as needed.

So now that we know what the $SE_{\bar{y}}$ is, how do we use it? When we calculate probabilities we can now use the SE instead of $\sigma_{\bar{y}}$ which is an unknown quantity. However, we do run in to a problem. If we use

$$\frac{\bar{y} - \mu_{\bar{y}}}{s_{\bar{y}}} = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}}$$

to calculate z, we discover that z no longer has a normal distribution (unless n is large). So how do we calculate probabilities? We will have to use the T distribution. More on the T distribution and how to deal with this will be given in the next chapter on confidence intervals.

Finally, we want to discuss the relation ship of $\bar{y}$ and s to μ, σ and the $SE_{\bar{y}}$ more closely. In particular, we want to fund out what happens to these quantities as our sample size, n, increases:

What happens to $\bar{y}$ as sample size increases?

Suppose we take a sample of $n = 25$ heights in men and calculate $\bar{y}$. Now suppose we do this again with $n = 100$. Does $\bar{y}$ change? Of course it does, but it doesn't change much. Our $\bar{y}$ for $n = 25$ will be close to the $\bar{y}$ for $n = 100$. Suppose we now use $n = 100,000$. Again, $\bar{y}$ won't change much. But notice that as $n \to \infty$, $\bar{y} \to \mu$.

What happens to s as the sample size increases?

The sample standard deviation, s, behaves in a way similar to $\bar{y}$. As our sample size increase, s may vary a bit, but as $n \to \infty$, $s \to \sigma$.

What happens to the $SE_{\bar{y}}$ as sample size increases?

Remember that $SE_{\bar{y}} = s/\sqrt{n}$, so as the sample size, n, increases, $SE_{\bar{y}}$ gets smaller and smaller. In other words, the bigger your sample size, the less variability in your $\bar{y}$. This makes sense. If, for example, $\bar{y} \to \mu$, then $\bar{y}$ is becoming less and less variable. After all, μ is a constant. If $\bar{y}$ actually reaches μ then there is *no* variability left in $\bar{y}$.

What happens to μ as sample size increases?

Nothing. It's important to realize that μ is not a random variable. We may not know what it is, but it is not random. Sample size has no effect on μ.

What happens to σ as sample size increases?

Nothing. Same as for μ, it's important to realize that σ is not a random variable. Sample size has no effect on σ.