CIS 2033, Spring 2017

Instructor: David Dobor

April 23, 2017

We consider a very practical application of the weak law of large numbers and of the central limit theorem. The application has to do with polling.

The polling problem

A certain referendum is about to take place. We're close enough to the day of the referendum so that the voters have made up their minds. A fraction p of the population of these voters are going to vote "yes" on the day of the referendum, but the referendum has not yet taken place, and you would like to predict what p actually is.

p is the fraction of the population of voters that will vote "yes" in a referendum.

So you go ahead and select, at random, a number of people out of the population. You then interview the people you have selected and for each person you record their answer, whether they intend to vote "yes", or whether they intend to vote "no". You then let

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th selected person says "yes",} \\ 0 & \text{if the } i\text{-th selected person says "no".} \end{cases}$$

When we say that the people are randomly selected, what we mean is that we choose them uniformly from the population. And since there's a fraction p that will vote yes, the random variable X_i will be 1 with probability p, and therefore its expected value will be equal to p: $\mathbf{E}[X_i] = p$.

Since X_i is a Bernoulli random variable with success probability p, its mean is

$$\mathbf{E}[X_i] = p,$$

and its variance is

$$\sigma^2_{X_i} = p\,(1-p).$$

In addition, we assume that we select these people independently. Note that if we select people independently, there's always a chance that the first person polled will be the same as the second person polled, something that we do not really want to happen. However, if we assume that the population is very large, or even idealize the situation by assuming that the population is infinite, then this is never going to happen, and this will not be a concern.

So how do we proceed? We look at the results that we get from the people that we polled. We count how many said "yes", divide that count by n, and this gives us the fraction of "yes"es *in the sample* of people that we have polled:

M_n is the fraction of the people in the sample (the n polled voters) who will vote "yes".

$$M_n = \frac{X_1 + \ldots + X_n}{n}.$$

Now M_n is a pretty reasonable estimate for the unknown true fraction p, the fraction of "yes"es in the overall population.

Suppose your boss has asked you to find out the exact value of p. What should your response be? Well, there is *no way* to calculate p exactly on the basis of the limited number of people polled.

Therefore, there is going to be some error in our estimation of p.

Your boss reflects on this for a little while but soon comes back and says: "OK, then try to give me an estimate of p that is very accurate. I would like you to come up with an estimate which is correct within one percentage point. Can you do this for me?"

Your boss would like a "small error", e.g., within one percentage point: $\mid M_n - p \mid < 0.01$.

Your answer might be: "OK, let me try polling 10,000 people, and see if I can guarantee for you such a small error."

But after you think about the situation a little more, you realize that there is no way of guaranteeing such a small error with certainty. What if you get really unlucky, and the people that you poll happen to be not representative of the true voting population?

There is no way of guaranteeing *with certainty* that the error $\mid M_n - p \mid$ will be withing one percentage point (i.e., will be < 0.01).

So you come back to your boss and say: "I cannot guarantee with certainty that the error is going to be small, but I may be able to guarantee that the error that I get is small with high probability." Or, alternatively, "I'm going to guarantee for you that the probability that we get an error that's bigger than 0.01 is small:

$$\mathbf{P}\big(\mid M_{10,000} - p \mid \geq 0.01\big) \leq \text{"something small"} \qquad (1)$$

However, you can give a guarantee in the form of equation (1), which says that "the chances that the fraction of 'yes'es in your sample, M_n, deviates from the true fraction of the population of voters that will vote "yes", p, can be made "small".

So how small is it going to be? Let's try to derive a bound on this probability of an error larger than one percentage point.

Using the weak law of large numbers to find the "something small" in equation (1)

First, recall what exactly were the calculations that we carried out when we derived the weak law of large numbers. We saw that the weak law of large numbers follows directly from Chebyshev's inequality that says that for any random variable X with finite mean μ and variance σ^2, it holds that

$$\mathbf{P}\big(\mid X - \mu \mid \geq \epsilon\big) \leq \frac{\sigma^2}{\epsilon^2}, \quad \text{for any } \epsilon > 0. \qquad (2)$$

The statement in (2) is the familiar Chebyshev inequality.

Now, let our M_n play the role of X in Chebyshev's inequality (2). To rewrite (2) for our particular purpose, we need to know the variance of M_n and we need to know its mean. We computed these quantities earlier and found that $\mathbf{E}[M_n] = p$, and $\sigma^2_{M_n} = \frac{p\,(1-p)}{n}$.

So, using the calculations that we carried out when we derived the weak law of large numbers, we know that

$$\mathbf{P}(\mid M_{10,000} - p \mid \geq 0.01) \leq \frac{\sigma^2}{10000 \cdot (0.01)^2} = \frac{p(1-p)}{10^4 \cdot 10^{-4}} = p(1-p).$$

OK, but this expression depends on p, and we do not know what p is. (Of course if we knew p, there would be no need to conduct the poll in the first place.)

However, if you take the expression $p(1-p)$ and plot it as a function of p, what you obtain is a plot that you see in Figure 1. The maximum value of $p(1-p)$ is attained when p is equal to $1/2$, in which case we get a value of $1/4$.

That is, the variance of the Bernoulli is, at most, $1/4$. Therefore, rewriting the prvious equation, we obtain the following bound:

$$\mathbf{P}(\mid M_{10,000} - p \mid \geq 0.01) \leq p(1-p) \leq \frac{1}{4} \tag{3}$$

So you tell your boss: "*If I sample 10,000 people, then the probability of an error more than the one percentage point is going to be less than 25%.*"

At which point, your boss might reply: "Well, a probability of a large error of 25% is just too big. This is unacceptable. I would like you to have a probability of error that's less than 5%."

So suppose now that we want to reduce the error, and make it only 5% or less. How are we going to proceed? We go back and take another look at Chebyshev's inequality, which we wrote down for our particular exmple – and which is inequality (3) – and rewrite it again for any n, not just for $n = 10000$:

$$\mathbf{P}(\mid M_n - p \mid \geq 0.01) \leq \frac{\sigma^2}{n \cdot (0.01)^2} = \frac{p(1-p)}{n \cdot 10^{-4}} \leq \frac{1}{4 \cdot n \cdot 10^{-4}}.$$

Now if we want to make

$$\frac{1}{4 \cdot n \cdot 10^{-4}} \leq 0.05 = \frac{5}{10^2}, \tag{4}$$

then we should solve (3) for n to obtain that

$$n \geq \frac{10^6}{20} = 50,000. \tag{5}$$

So at this point, you go back to your boss and tell them that one way of guaranteeing that the probability of a large error is less than or

Since M_n is the sum on n independent Bernoulli(p) random variables divided by n, the mean of M_n is

$$\mathbf{E}[M_n] = \frac{np}{n} = p,$$

and its variance is

$$\sigma^2_{M_n} = \frac{np\,(1-p)}{n^2} = \frac{p\,(1-p)}{n}.$$

In particular, since we take $n = 10,000$, we have that

$$\sigma^2_{M_{10,000}} = \frac{p\,(1-p)}{10^4}.$$

Figure 1: A plot of $\sigma = p(1-p)$. The maximum occurs when $p = 1/2$, at which point $\sigma = 1/4$.

equal to 5% is to take n equal to 50,000. This number of people to poll, 50, 000, will suffice to achieve the desired specs:

$$\mathbf{P}\big(\mid M_{50,000} - p \mid \geq 0.01\big) \leq 0.05$$

Notice that the desired specs have two parameters. One is the accuracy, 0.01, that you want, and the other reflects the confidence, 0.05, with which the accuracy is going to be achieved.

So you can say that you are at least 95% confident that the sample mean is within 1 percentage point of the true population mean, p, when $n = 50,000$ people are sampled.

Now 50,000 is a pretty large number. If you look at the results of polls, as they are presented in newspapers or on websites, they usually tell you that there's an accuracy of plus or minus *three percentage points*, not one percentage point. That helps things a little: it means that you can do with a somewhat smaller sample size.

And then there's another effect. Our calculations were based on the Chebyshev inequality. But Chebyshev's inequality is not all that accurate. It turns out that if we use more accurate estimates of the probability $\mathbf{P}(\mid M_n - p \mid \geq \epsilon)$, we will find that actually much smaller values of n will be enough for our purposes.

We'll do that next.

Applying the Central Limit Theorem to Polling

We will now use the central limit theorem to approximate the quantity $\mathbf{P}(\mid M_n - p \mid \geq 0.01)$.

Recall that the central limit theorem involves the standardized version of the random variable S_n, where S_n stands for the sum of the X_is:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}, \tag{6}$$

Let $S_n = X_1 + \ldots X_n$. Then, by the central limit theorem, we know that

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

is approximately standard normal for large enough n.

and we know that the random variable Z_n is approximately standard normal for sufficiently large n.

What we want to do now is to take the event

$$\mid M_n - p \mid \geq 0.01 \tag{7}$$

and rewrite it in an equivalent way, but in a way that involves the random variable Z_n.

μ in (6) is p

σ in (6) is $\sqrt{p(1-p)}$

First, we note that in (6) we have a μ and a σ, so we should keep in mind what these are. For a Bernoulli random variable, the mean is p and the standard deviation σ is $\sqrt{p(1-p)}$.

Next, let's look at the event in (7). M_n is S_n/n and so we can say that

$$\mid M_n - p \mid \geq 0.01 \quad \text{is the same as} \quad \left| \frac{S_n}{n} - \frac{np}{n} \right| \geq 0.01,$$

which is the same as

$$\left| \frac{S_n - np}{n} \right| \geq 0.01. \tag{8}$$

So (7) and (8) are the same event, and (8) is beginning to look like the expression in (6): p is the same as μ, but there is a little bit of a difference in the denominator terms. So let's see what we can do. Let's take the event in (8) and multiply both sides of the inequality by $\sqrt{n}$. This causes the denominator term on the left hand side to become just $\sqrt{n}$ and we get a $\sqrt{n}$ term in the numerator on the other side:

$$\left| \frac{S_n - np}{n} \right| \geq 0.01 \quad \text{is the same as} \quad \left| \frac{S_n - np}{\sqrt{n}} \right| \geq 0.01\sqrt{n}.$$

Finally, we multiply the denominators on both sides by σ, and we obtain this equivalent representation:

$$\left| \frac{S_n - np}{\sqrt{n}} \right| \geq 0.01\sqrt{n} \quad \text{is the same as} \quad \left| \frac{S_n - np}{\sqrt{n}\sigma} \right| \geq \frac{0.01\sqrt{n}}{\sigma}. \tag{9}$$

But now we notice that in (9) we do have the random variable Z_n that we wanted, and so we managed to express the event $| M_n - p | \geq 0.01$ in terms of the random variable Z_n. In short, what we have is that

$$\mathbf{P}(| M_n - p | \geq 0.01) = \mathbf{P}\Big(| Z_n | \geq \frac{0.01\sqrt{n}}{\sigma} \Big),$$

and using the central limit theorem to approximate this last probability we write:

$$\boxed{\mathbf{P}(| M_n - p | \geq 0.01) \approx \mathbf{P}\Big(| Z | \geq \frac{0.01\sqrt{n}}{\sigma} \Big).} \tag{10}$$

where Z stands for the standard normal random variable with mean equal to 0 and variance equal to 1.

Z stands for the standard normal with mean 0 and variance 1:

$$Z \sim \mathcal{N}(0,1)$$

,

If somebody gives us the value of n, we would like to be able to calculate the probability $\mathbf{P}(| M_n - p | \geq 0.01)$ using the approximation in (10). But there's a slight difficulty because σ is a function that depends on p, and p is not known.

However, as we discussed in the previous section, we do know that σ is always less than or equal to $1/2$ (see figure 1 on page 3). This suggests that we could use the worst-case value of the standard deviation – replace σ by $1/2$ – and instead look at this probability:

$$\mathbf{P}\ \big(| Z | \geq 0.02\sqrt{n}\ \big). \tag{11}$$

Figure 2: The shaded area under this standard normal curve corresponds to the probability in expression (10).

Now, which one of the following two probabilities is larger,

$$\mathbf{P}\Big(| Z | \geq \frac{0.01\sqrt{n}}{\sigma} \Big) \quad \text{or} \quad \mathbf{P}\ \big(| Z | \geq 0.02\sqrt{n}\ \big)? \tag{12}$$

Note that since $\sigma \leq 1/2$, we have that $1/\sigma \geq 2$, and therefore

$$\frac{0.01\sqrt{n}}{\sigma} \geq 0.02\sqrt{n}.$$

It follows then that

$$\mathbf{P}\Big(\mid Z \mid \geq \frac{0.01\sqrt{n}}{\sigma} \Big) \leq \mathbf{P}\left(\mid Z \mid \geq 0.02\sqrt{n} \right). \tag{13}$$

This is shown in figure 3. The blue shaded area corresponds to the probability on left-hand-side of (13), and the red shaded area corresponds to the probability on the right-hand-side of that inequality.

Figure 3: Under this standard normal curve, the area shaded in blue is smaller than the area shaded in red. The blue area is the probability on the left-hand-side of (13). The red area is the probability on the right-hand-side of (13).

So if somebody gives us a value of n, we should be able to calculate the probability in (11). How do we calculate it?

The probability that the *absolute value* of Z is above $0.02\sqrt{n}$, is equal to the sum of the probabilities of the two tails that are shaded in red in figure 2 or figure 3. But because of the symmetry of the normal distribution, this is twice the probability of each one of the tails:

$$\mathbf{P}\left(\mid Z \mid \geq 0.02\sqrt{n} \right) = 2 \cdot (1 - \Phi(0.02\sqrt{n})).$$

Putting all of this together, we finally have a bound for the desired probability, which is expressed in terms of the standard normal CDF:

$$\mathbf{P}(\mid M_n - p \mid \geq \ 0.01) \leq 2 \cdot (1 - \Phi(0.02\sqrt{n})) \tag{14}$$

Let us put (14) to use and try $n = 10,000$:

$$\begin{aligned} \mathbf{P}(\mid M_{10,000} - p \mid \geq \ 0.01) &\leq 2 \cdot (1 - \Phi(0.02\sqrt{10,000})) \\ &= 2 \cdot (1 - \Phi(2)) \\ &= 2 \cdot (1 - 0.9772) \\ &= 0.046. \end{aligned}$$

So if we use 10, 000 people in our sample, then we will get an accuracy of one percentage point with very high confidence. In other words, the probability that we do not meet the specification, so that the accuracy is worse than one percentage point, is quite small. That probability is 0.046 – about 4%. And that's pretty good.

0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	.9032	.9049	.9066	.9082	.9099	.9115
1.4	.9192	.9207	.9222	.9236	.9251	.9265
1.5	.9332	.9345	.9357	.9370	.9382	.9394
1.6	.9452	.9463	.9474	.9484	.9495	.9505
1.7	.9554	.9564	.9573	.9582	.9591	.9599
1.8	.9641	.9649	.9656	.9664	.9671	.9678
1.9	.9713	.9719	.9726	.9732	.9738	.9744
2.0	.9772	.9778	.9783	.9788	.9793	.9798

Figure 4: A fragment of the standard normal table. We are looking for the value of the CDF at point 2.

Suppose that your boss now tells you: "I really only want the probability of not meeting the specs to be 5%."

- Specs: $\mathbf{P}\left(\mid M_n - p \mid \geq 0.01 \right) \leq 0.05$.

You look at your results and say: "with 10,000, I achieved a probability of a large error that's less than 5%. This means that I probably have some leeway in reducing the size of my sample."

What could the size of the sample be and still meet those specs? What we're trying to do here is that we have (13), the approximation for the probability of interest, and we want to set this probability to a value of 0.05:

$$\mathbf{P}(\mid M_n - p \mid \geq 0.01) \leq 2 \cdot (1 - \Phi(0.02\sqrt{n})) = 0.05 \qquad (15)$$

Then we want to ask, what is the value of n that will result in the probability of not meeting the specs, which is 0.05?

So we do the algebra and we find that (15) corresponds to requiring that $\Phi(0.02\sqrt{n}) = 0.975$. What's the interpretation of this? We want to choose n so that the probability of the two green tails in figure 5 is 5%. This means that we want the probability of one of the tails shown in figure 4 to be 2.5%.

Figure 5: The total area shaded in green corresponds to the probability of 0.05. Therefore, because of symmetry, each tail has the probability of 0.025. This implies that $\Phi(0.02\sqrt{n})$ must be 0.975.

So we peek into the standard normal table (or rush to our favorite software package) in hopes of finding the value for which the CDF is equal to 0.975. We look around and we find that 0.975 corresponds to 1.96.

0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	

Figure 6: A fragment of the standard normal table. We are looking for the value of the standard normal at which its CDF is equal to 0.975. This happens at point 1.96.

This tells us that $0.02\sqrt{n} = 1.96$, which we solve for n to find that $n = 9,604$. This is indeed some reduction from the 10,000 that we had originally.

How does this relate to the real world? When you read newspapers or check websites about the polls, you never see sample sizes that are about 10,000. You usually see sample sizes of the order of 1,000, sometimes even smaller.

How can they do that? Well, they can do that because the specs that they impose are not as tight as the specs that we have here. Usually, they tell you that the results are accurate within three percentage points, let's say, instead of one percentage point. By moving from 0.01 to 0.03, if you repeat the calculations, you will find that the sample size of about 1,000 will actually do.