Lecture 18 - Polling

Friday, February 12, 2021 10:53 AM

Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

1. Define probability of a “bad event”
2. Apply CLT
3. Convert to a standard normal
4. Solve for n

This goal is a statement of confidence in our answer (commonly called a confidence interval). The steps above walk through each step of the work to solve for the required value of n.

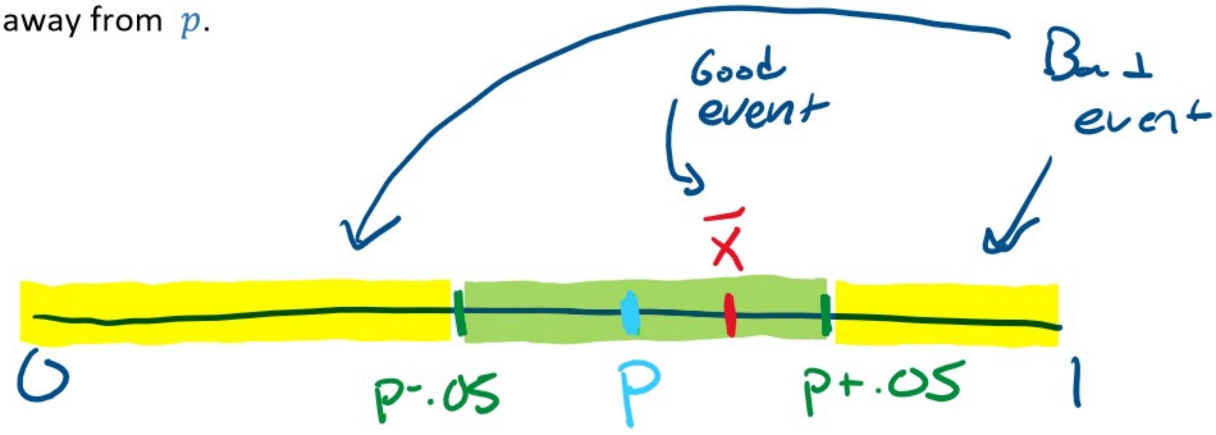
1) Define probability of "bad event"

Our estimate is bad if it's far away from the true value p. It's possible for $\bar{X}$ to be far from p (e.g., if we get a unlucky sample), so we want to make sure our estimate is close with high probability.

Therefore, the "bad event" is if our estimate is far off (defined as more than 5% in our goal) and we want to make sure that event is unlikely (occurs less than 2% of the time).

$$(*) \quad \Pr(|\bar{X} - p| > 0.05) \leq 0.02$$

Visually, we can imagine this as a number line, where the bad even is when $\bar{X}$ is far away from p.



So the equation $P(|\bar{X} - p| > 0.05) \leq 0.02$ is trying say that with high probability (at least 98%), there is at most an error of 5%. Usually the error term makes sense, but it's a little less clear why this is a statement of probability.

Remember that $\bar{X}$ itself is a random variable. This means it is possible for it to take on a value outside of this range. We want to make sure n is large enough that this "bad event" is unlikely.

In other words, our goal is to find which value of n does our equation (*) hold.

To do this, we need to discuss the distribution of $\bar{X}$, which requires uing the Central Limit Theorem (CLT).

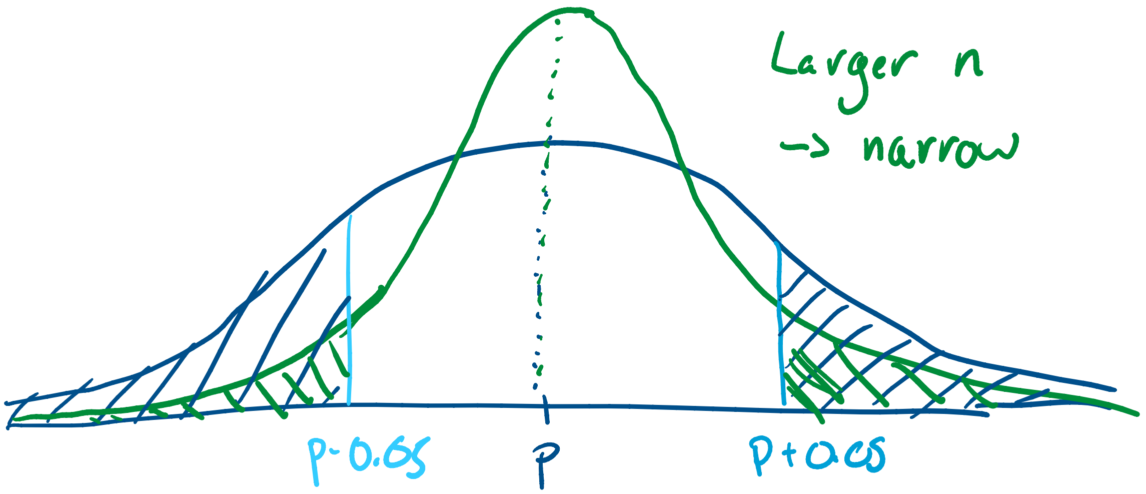
2) CLT

Since the $\bar{X}$ is the sum of n i.i.d Bernoulli random variables, we can consider applying the CLT, since we are going to be talking about the case where n gets large.

Applying the CLT directly to the definition of $\bar{X}$ yields:

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

We can visualize the distribution of $\bar{X}$ as the following. Note that since the variance has n in the denominator, as we increase n, we expect the distribution to become more narrow, and closer to it's expectation.



So as n gets larger, the curve gets more and more probability near the mean, we can say that the probability of the tails (e.g., being far away from the mean) goes down as n increases. So then our question comes to, how do we choose n such that at most 2% of the time, the sample mean is outside of this range (e.g., in the tails, far away).

To do this, we need to normalize to a standard normal so we can do computations on it.

3) Convert to Standard Normal

We will eventually want to be able to look up into a Φ-table to find the appropriate value of n. In order to do so, we have to normalize $\bar{X} - p$ by its standard deviation (square root of the variance).

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1)$$

$$\Pr(|\bar{X}-p| > 0.05)$$

$$= \Pr\left(\left|\underbrace{\frac{\bar{X}-p}{\sqrt{p(1-p)/n}}}_{Z\sim N(0,1)}\right| > \underbrace{\frac{0.05}{\sqrt{p(1-p)/n}}}_{a}\right)$$

We simplified the side on the right by just calling that value a. Unfortunately, this is not going to be the most useful since

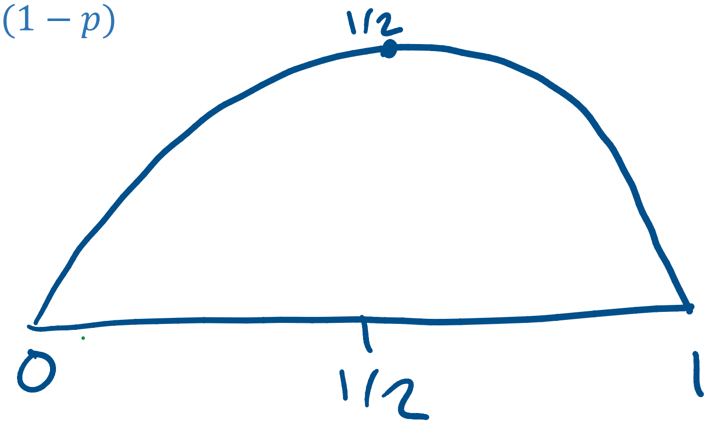
$$a = \frac{0.05}{\sqrt{p(1-p)/n}}$$

Depends on knowing the true ratio p! This means we are a bit stuck since we can't do any computations involving a number we don't know.

It's very common to face a road block like this when working in math, so a common trick is to see if you can "bound" the answer rather than find it exactly. Remember, our goal was to find an upper-bound on the probability of the bad event in the first place (≤ 0.02). Well this probability depends on which area we look at in the tails in the image above, so we can pick the smallest possible value of a and use that as a conservative estimate of the error probability.

It turns out that the denominator of a, is a bit simpler to analyze. It turns out that $\sqrt{p(1-p)} \leq \frac{1}{2}$ for all values of $0 \leq p \leq 1$. We can see this visually by plotting out $\sqrt{p(1-p)}$

$$\sqrt{p(1-p)} \leq \frac{1}{2}$$

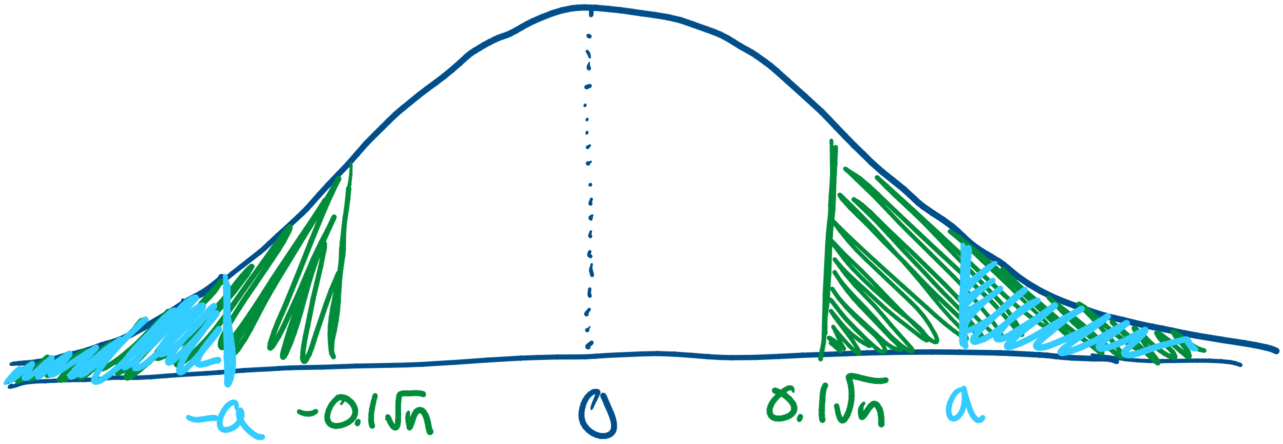


One of the trickiest parts about this example is thinking about the bound we just argued. We argued that the denominator term $\sqrt{p(1-p)}$ is **at most** $1/2$. In other words $\sqrt{p(1-p)}$ can be less than $1/2$, but will never be greater. This then forms a **lower bound** on the value of a since we made an **upper bound** on one of the terms in the denominator. Remember that with fractions, a larger denominator leads to a smaller number. So the $\sqrt{p(1-p)}$ can be at most $1/2$ which will result in the smallest a possible, but if $\sqrt{p(1-p)}$ were smaller, a would be larger.

Mathematically, we can then say

$$a = \frac{0.05}{\sqrt{p(1-p)/n}} \underset{\substack{\uparrow \\ \text{by algebra}}}{=} \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}} \underset{\substack{\uparrow \\ \text{by bound above}}}{\geq} \frac{0.05\sqrt{n}}{1/2} = 0.1\sqrt{n}$$

So now we can say that a will be at least $0.1\sqrt{n}$! The reason why we lower-bound a, can also be seen visually in the graph below. a is some fixed number (that depends on p) and we are trying to find the probability of the "bad event" which is out in the shaded tails. So since we don't know a, we find the smallest possible value it can take on ($0.1\sqrt{n}$). This is definitely find an upper bound on the probability of the bad event, since it's a conservative estimate (treating a to be as close to the center as it possibly could).



The blue area (at least a away) is the actual area of interest, but since we don't know p, we can find an upper bound on this probability (the green area) but finding the lowest possible value a can take on, which is $0.1\sqrt{n}$.

So now we are getting very close to the result. We now just need to figure out which value of n is necessary such that

$$\Pr(|Z| > 0.1\sqrt{n}) \leq 0.02$$

So now we are getting very close to the result. We now just need to figure out which value of n is necessary such that this holds.

4) Solve for n

We are getting to almost be done. Most of the work now is just changing the left-hand side of the equation above to be one that we can look up in a Φ-table.

$$\begin{aligned}
&\Pr(|Z| > 0.1\sqrt{n}) \\
&= \Pr(Z < -0.1\sqrt{n}) + \Pr(Z > 0.1\sqrt{n}) \\
&= 2\Pr(Z > 0.1\sqrt{n}) \\
&= 2(1 - \Pr(Z \leq 0.1\sqrt{n})) \\
&= 2(1 - \Phi(0.1\sqrt{n}))
\end{aligned}$$

Therefore, we can now do algebra to simplify to just $\Phi(0.1\sqrt{n})$

$$2(1-\Phi(0.1\sqrt{n})) \leq 0.02$$

$$1-\Phi(0.1\sqrt{n}) \leq 0.01$$

$$0.99 \leq \Phi(0.1\sqrt{n})$$

So now we just need to figure out what cell in the Φ-table corresponds to a probability of < 0.99. The first such value in the table we have is 2.33, which means

$$2.33 \leq 0.1\sqrt{n}$$

$$n \geq \left(\frac{2.33}{0.1}\right)^2$$

$$\geq 542.89$$

$$\boxed{\geq 543}$$

And now finally, we got to the answer! We must have at least $n \geq 543$ in order for our confidence intervalu of at most 5% away from true value, 98% of the time to hold!

The most important part of this example is the high-level idea of applying the CLT to a process and using standardization to turn it to a standard normal so that we can use a Phi-table to look up the solution. Most of the other details in this example are just that, details for this particular example. The most important take-away is the high-level process and seeing that the CLT actually can be used to estimate values!