Demo 31: Why *np* > 10 Is a Good Rule of Thumb

Explaining the np > 10 *rule for using the normal approximation in the CI of a proportion*

When you find a confidence interval for a proportion, there is a question of whether the normal approximation is correct. What does that mean? For one thing, the proportions are like a population of ones and zeros. So the distribution of sample proportions will be binomial. The traditional confidence interval formula,

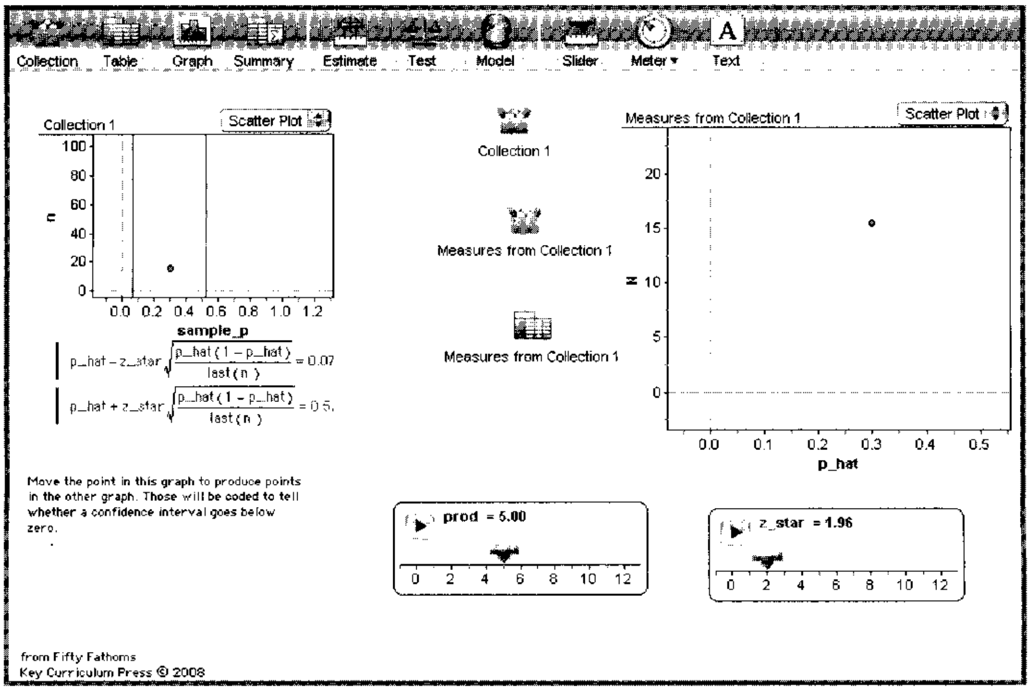
$$CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is based on the idea that the sample proportions will be normally distributed. But didn't we just say that they were binomially distributed? Sure. But the binomial distribution looks pretty normal as long as n is large enough and $\hat{p}$ is far enough from 0 or 1.

But how far is far enough? The rule of thumb is that $np > 10$.[1] But where does that come from? One way to study this question is to ask: What happens if np is small? What can happen if you keep using the normal approximation? For one thing, you can calculate confidence intervals for proportions that extend *below zero.* And that's bad. After all, a confidence interval is the range of proportions that could plausibly give rise to our sample. Are we willing to say that the proportion of "yes" voters might easily be –0.08? I don't think so.

In this demo, you control values for n and $\hat{p}$, and Fathom generates lower bounds for confidence intervals using the normal approximation, plotting them on a separate graph. They're coded to show whether the interval extends below zero or not.

We're looking to figure out what values for n and $\hat{p}$ give "good" confidence intervals. Let's begin!



$p\_hat - z\_star\sqrt{\frac{p\_hat(1-p\_hat)}{last(n)}} = 0.07$

$p\_hat + z\_star\sqrt{\frac{p\_hat(1-p\_hat)}{last(n)}} = 0.5$

Move the point in this graph to produce points in the other graph. Those will be coded to tell whether a confidence interval goes below zero.

prod = 5.00

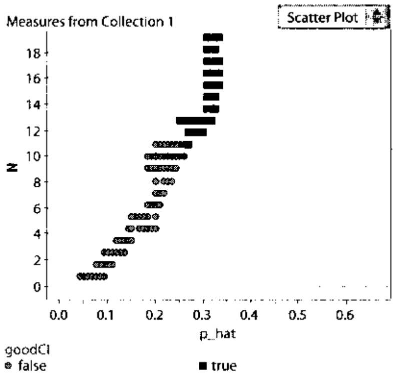
z_star = 1.96

from Fifty Fathoms
Key Curriculum Press © 2008

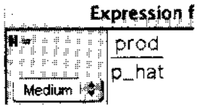
[1] Also, $n(1-p) > 10$, but we'll just focus on the bottom end.

What To Do

- Open **Why np Is Greater Than 10.ftm.** It will look like the illustration.
- Grab the single point in the **Collection 1** graph at left. This point is your controller; by dragging, you change the values for *n* and *p*.[2] The confidence interval extends from the red line to the blue line. Note: The axis bounds for this graph are pretty good; you shouldn't need to change them. Those of the other graph may jump around. Don't let that bother you.
- Carefully move the point down and to the left (toward the origin). Points will appear automatically in the larger graph at right, mirroring your motion in the small graph.



- Drag until you can see two different symbols in the big graph, and a legend appearing for the attribute **GoodCI**, which is either **true** or **false**. In the illustration, the blue squares are good, and the gray circles represent bad confidence intervals—where the interval extends below $p = 0$.
- Continue to drag the point around, in different directions, until you get a good idea of what part of the plane is "good" and what part is not. You should see a region near the axes that's bad. That is, if *n* is small or *p* is small, it may mean that the normal approximation is really bad.
- Let's write a function that approximates the border of this region. Click once on the larger graph—the one with lots of points—to select it.
- Choose **Plot Function** from the **Graph** menu.
- In the formula editor, enter **prod / p_hat**, as shown. Then press **OK** to exit the editor. A curve appears—a rectangular hyperbola.



$\frac{\text{prod}}{\text{p\_hat}}$

- Drag the **prod** slider to change the curve. Make it so that the curve roughly matches the boundary between good confidence intervals and bad ones. You should get a value something like 3.

This value, 3, is *the smallest product of the sample size and the sample proportion* you can possibly imagine using with the normal approximation. If $np < 3$, you're guaranteed a really bad CI—one that extends below zero.

Questions

1 Suppose you did a poll of 50 people and 5 of them said their favorite ice cream was pistachio. Where is that point on the graphs? Is it in the "good" area? What's the (normal approximation) confidence interval for that proportion?

2 We said that $np < 3$ was really bad. And we got that by comparing the points to a function, **N = prod / p_hat.** What's the connection between these two? That is, why is one multiplication and the other division?

Extension

But isn't the rule of thumb $np > 10$? Sure. We have just found the bare minimum here. We know that $np < 3$ is bad; but that doesn't mean $np > 3$ is good. For one thing, we have been identifying "good" as any CI that does not include zero. This is a little strange. Suppose I take a sample of 20 and get 4 successes. The lower bound of the 95% confidence interval (with the normal approximation) is about 0.02. Is that plausible?

Hardly. If I had a process with a 2% chance of success, and I did it 20 times, I'd get 4 or more successes fewer than one time in a thousand.

So, what shall we do to improve our rule of thumb? One way is to remember that we've been looking only

[2] Basically, we have created a two-dimensional slider out of a point. See "Using a Point as a Controller" in Appendix A.

at 95% confidence intervals—ones with $z^* = 1.96$. Control this value with the slider named **z_star.** Bump that up to about 3—that is, look at a CI width of 3 standard errors instead of 2—and see what happens.

Another Question

3 If you take the sample size and multiply it by the sample proportion, don't you get the number of successes? That is, could you say that the rule of thumb "$np > 10$" really means that you have to get at least 10 people to say yes (and at least 10 people to say no) before you can use the normal approximation to the CI? **Sol**

Another Extension

We have really just looked at whether the normal-approximation confidence interval overlaps zero. But another question is whether the overall shape of the distribution matches the binomial. Open **Binomial v Normal.ftm** and explore the difference. It looks like the illustration below.

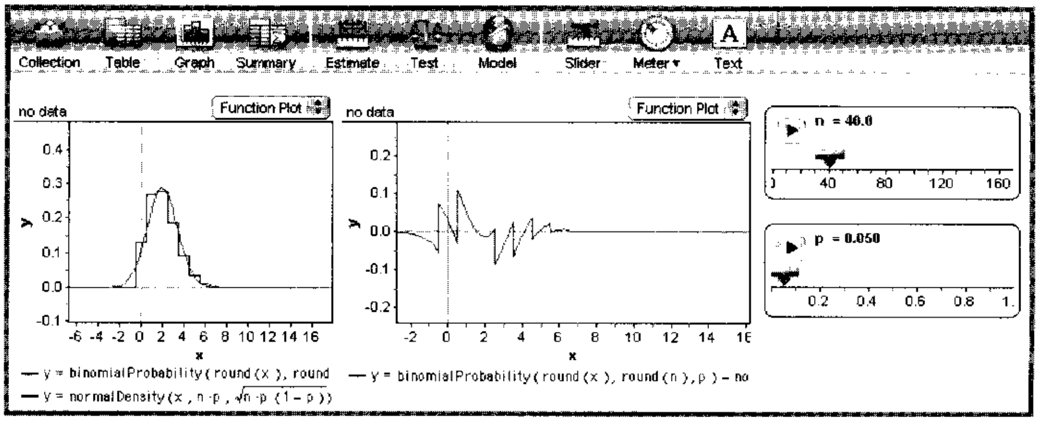
You can see, at left, the normal and binomial density functions plotted together. At right, you see the difference between those functions. The sliders **n** and **p** control the sample size and the probability of success, respectively. If you make **n*p** large, the curves match well, and the difference is small. If **n*p** is small, the difference is all over the place, *whether or not much of the normal function lies below zero.*

Challenges

4 Show that the assertion earlier—"If I had a process with a 2% chance of success, and I did it 20 times, I'd get 4 or more successes fewer than one time in a thousand."—is true.

5 If you have a computer to help you, you don't need the normal approximation—the binomial distribution gives the correct answer. Explain why the binomial is always correct.

6 Perform an experiment to test how much difference it makes to use the normal approximation instead of the binomial. Use a sample size of 10 and a true population proportion of 0.2. Draw repeated samples and construct 95% confidence intervals using both the binomial estimate and the normal approximation. See how many CIs capture the true proportion, and on which side they miss. Vary the sample size and the population proportion. **Sol**



n = 40.0

p = 0.050

— y = binomialProbability (round (x), round

— y = normalDensity (x , n ·p , $\sqrt{n \cdot p\ (1 - p\ )}$)

— y = binomialProbability (round (x), round (n), p) – no