Agenda: #57F AP Stats Ch18A

Title: Confidence Intervals for Proportions

Objectives: Set observed proportions as population proportions, then find the confidence interval.

1) W.Up : (Worksheet 15min)

2) Different Dice Questions

(Elbow Partners 15min)

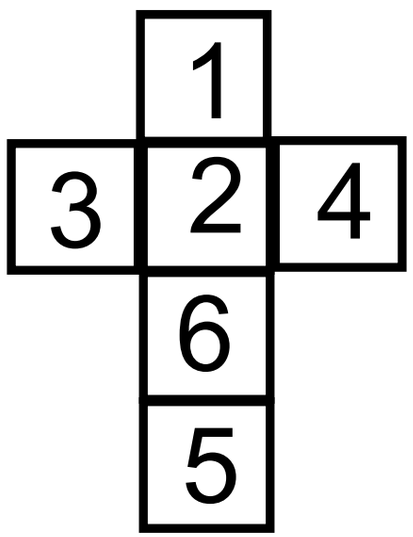
3) #52 Quiz A&B Review (Independent 15min)

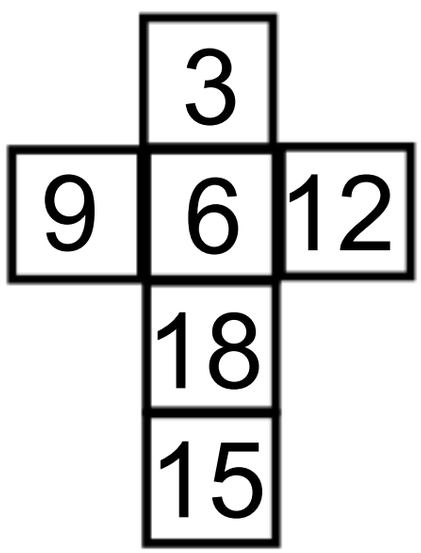
4) Chapter Intro Reading (Group 15min)

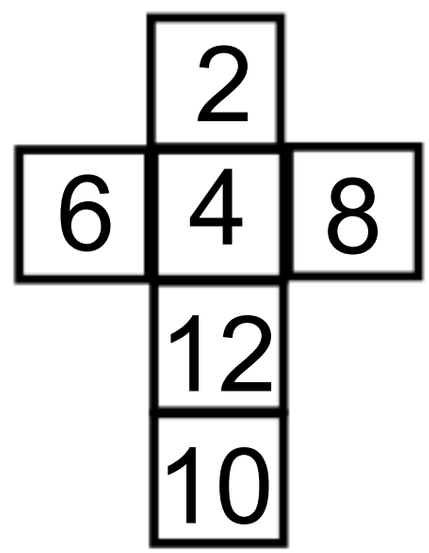
5) Coral Reef Questions (Elbow Partner + Stamp Check 15min.)

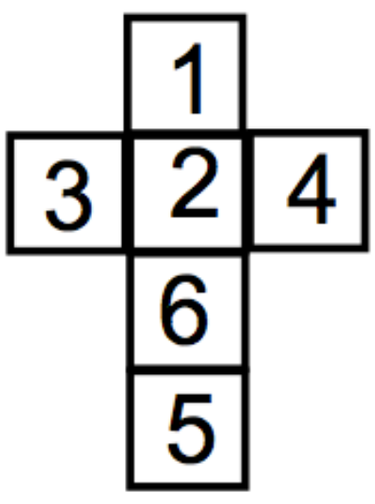
(Reading 15min).

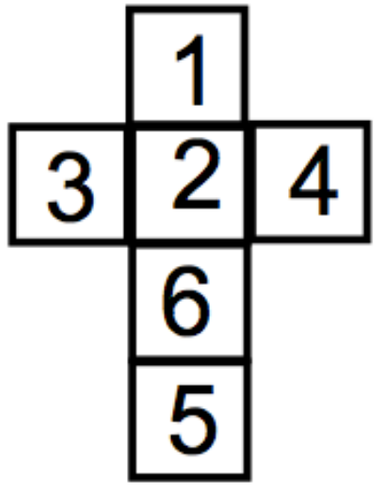
6) Global Warming Questions & Closing Activities (15min)

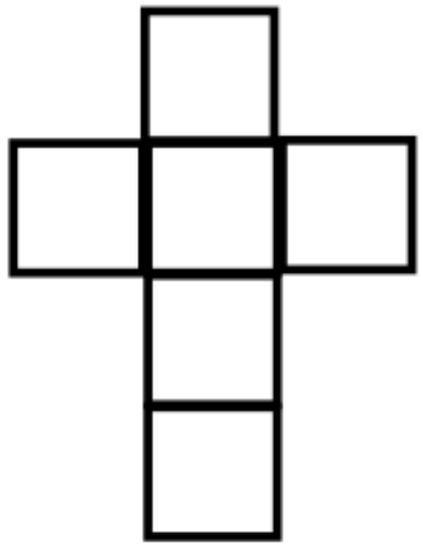

1
3
2
4
6
5


3
9
6
12
18
15


2
6
4
8
12
10

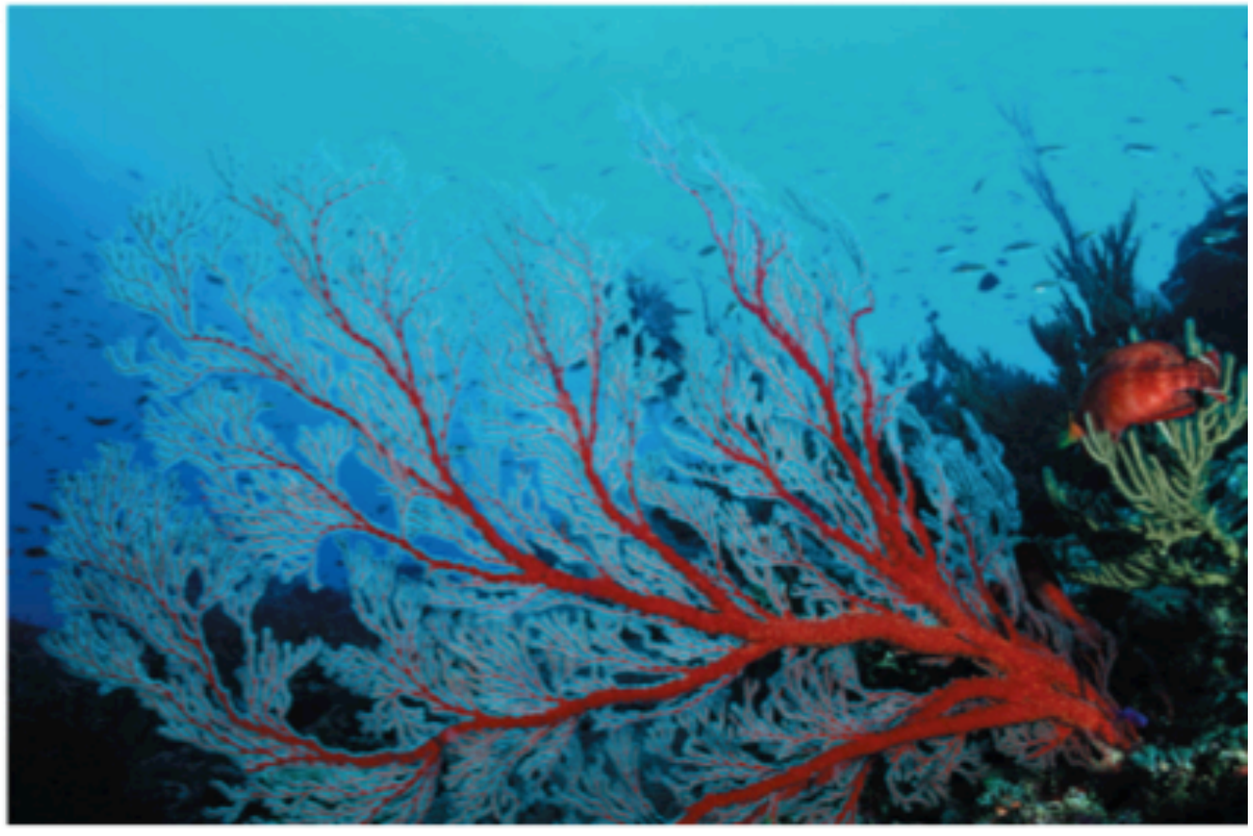

1
3
2
4
6
5


1
3
2
4
6
5



chapter

18 Confidence Intervals for Proportions



Coral reef communities are home to one quarter of all marine plants and animals worldwide. These reefs support large fisheries by providing breeding grounds and safe havens for young fish of many species. Coral reefs are seawalls that protect shorelines against tides, storm surges, and hurricanes, and are sand "factories" that produce the limestone and sand of which beaches are made. Beyond the beach, these reefs are major tourist attractions for snorkelers and divers, driving a tourist industry worth tens of billions of dollars.

But marine scientists say that 10% of the world's reef systems have been destroyed in recent times. At current rates of loss, 70% of the reefs could be gone in 40 years. Pollution, global warming, outright destruction of reefs, and increasing acidification of the oceans are all likely factors in this loss.

Dr. Drew Harvell's lab studies corals and the diseases that affect them. They sampled sea fans[1] at 19 randomly selected reefs along the Yucatan peninsula and diagnosed whether the animals were affected by the disease *aspergillosis*.[2] In specimens collected at a depth of 40 feet at the Las Redes Reef in Akumal, Mexico, these scientists found that 54 of 104 sea fans sampled were infected with that disease.

Of course, we care about much more than these particular 104 sea fans. We care about the health of coral reef communities throughout the Caribbean. What can this study tell us about the prevalence of the disease among sea fans?

We have a sample proportion, which we write as $\hat{p}$, of 54/104, or 51.9%. Our first guess might be that this observed proportion is close to the population proportion, p. But we also know that because of natural sampling variability, if the researchers had drawn

a second sample of 104 sea fans at roughly the same time, the proportion infected from that sample probably wouldn't have been exactly 51.9%.

What *can* we say about the population proportion, p? To start to answer this question, think about how different the sample proportion might have been if we'd taken another random sample from the same population. But wait. Remember—we aren't actually going to take more samples. We just want to *imagine* how the sample proportions might vary from sample to sample. In other words, we want to know about the *sampling distribution* of the sample proportion of infected sea fans.

Confidence Intervals for Proportions

Coral reef communities

sample sea fans

Depth 40 feet at the Las Redes Reef in Akumal, Mexico

54 of 104 sea fans sample were infected.

disease "aspergillosis"

$\hat{p}$ .... $\frac{54}{104}$ or 51.9%

We guess this observed proportion is close to the population proportion p.

But we can't be exactly ~~much~~ matched "51.9%", if we had drawn the second sample.

What can we say about the population "p"?

Are we going to take more samples? - ... No

So just imagine how sample proportions might vary from sample to sample.

⇓

Sampling distribution of the sample proportion

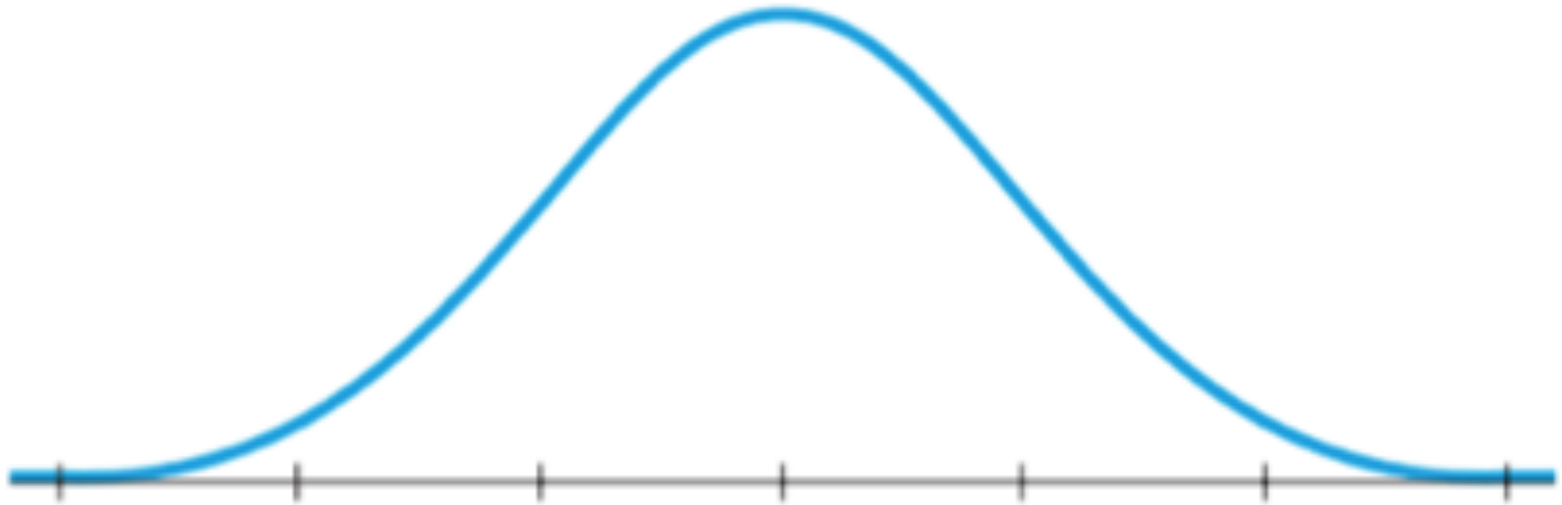
A Confidence Interval

Let's look at our model for the sampling distribution. What do we know about it? We know it's approximately Normal (under certain assumptions, which we must be careful to check) and that its mean is the proportion of all infected sea fans on the Las Redes Reef. Is the infected proportion of *all* sea fans 51.9%? No, that's just $\hat{p}$, our estimate. We don't know the proportion, p, of all the infected sea fans; that's what we're trying to find out. We do know, though, that the sampling distribution model of $\hat{p}$ is centered at p, and we know that the standard deviation of the sampling distribution is $\sqrt{\frac{pq}{n}}$.

Now we have a problem: Since we don't know p, we can't find the true standard deviation of the sampling distribution model. We do know the observed proportion, $\hat{p}$, so, of course we just use what we know, and we estimate. That may not seem like a big deal, but it gets a special name. Whenever we estimate the standard deviation of a sampling distribution, we call it a **standard error.**[3] For a sample proportion, $\hat{p}$, the standard error is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

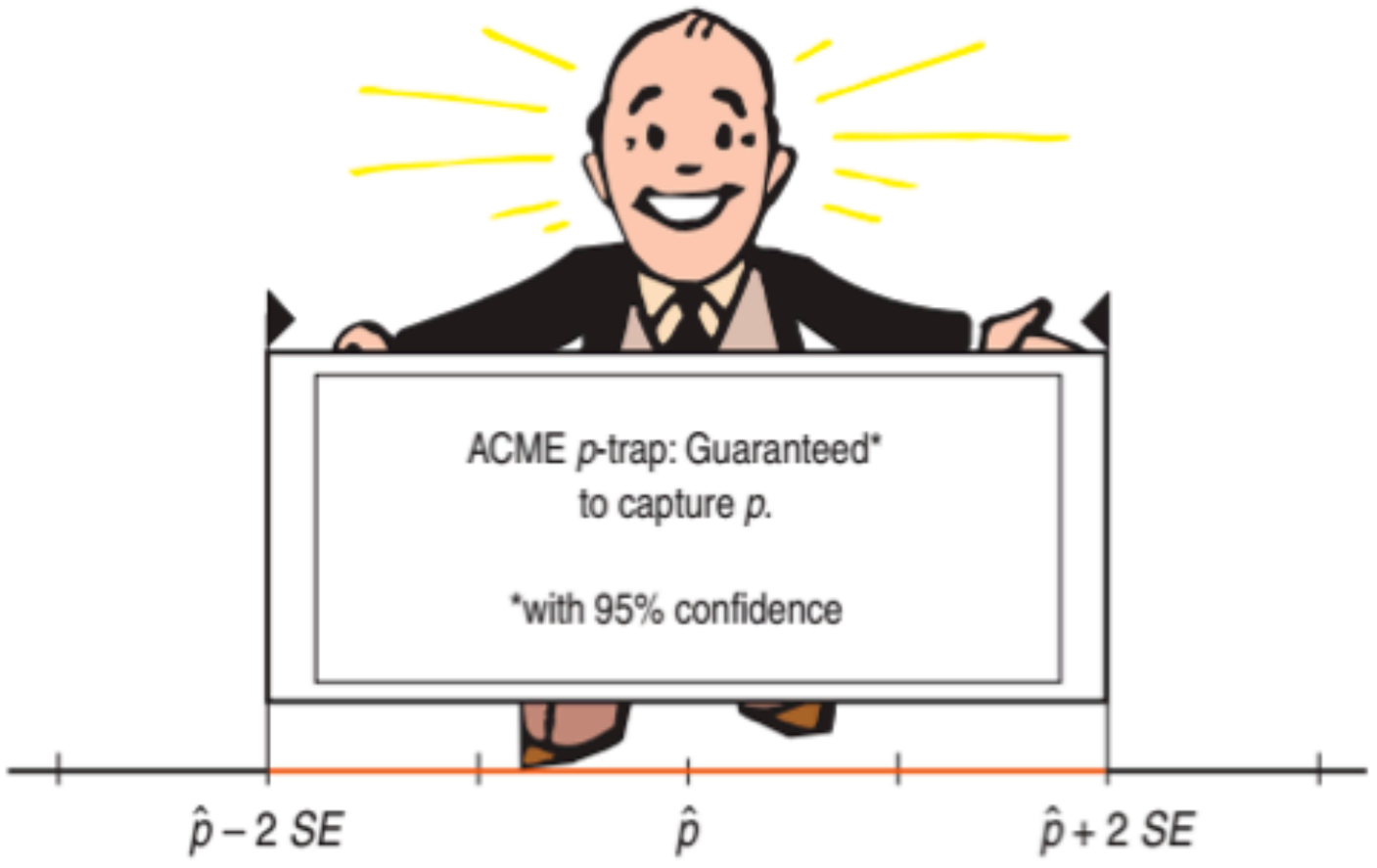
Now we know that the sampling model for $\hat{p}$ should look like this:



Great. What does that tell us? Well, because it's Normal, it says that about 68% of all samples of 104 sea fans will have $\hat{p}$'s within 1 *SE*, 0.049, of p. And about 95% of all these samples will be within $p \pm 2$ *SE*s. But where is *our* sample proportion in this picture? And what value does p have? We still don't know!

We do know that for 95% of random samples, $\hat{p}$ will be no more than 2 *SE*s away from p. So let's look at this from $\hat{p}$'s point of view. If I'm $\hat{p}$, there's a 95% chance that p is no more than 2 *SE*s away from me. If I reach out 2 *SE*s, or 2×0.049, away from me on both sides, I'm 95% sure that p will be within my grasp. Now I've got him! Probably.

Of course, even if my interval does catch p, I still don't know its true value. The best I can do is to produce an interval, and even then I can't be positive it contains p.



So what can we really say about p? Here's a list of things we'd like to be able to say, in order of strongest to weakest and the reasons we can't say most of them:

1. **"51.9% of *all* sea fans on the Las Redes Reef are infected."** It would be nice to be able to make absolute statements about population values with certainty, but we just don't have enough information to do that. There's no way to be sure that the population proportion is the same as the sample proportion; in fact, it almost certainly isn't. Observations vary. Another sample would almost certainly yield a different sample proportion.
2. **"It is *probably* true that 51.9% of all sea fans on the Las Redes Reef are infected."** No. In fact, we can be pretty sure that whatever the true proportion is, it's not exactly 51.900%. So the statement is not true.
3. **"We don't know exactly what proportion of sea fans on the Las Redes Reef is infected, but we *know* that it's within the interval $51.9\% \pm 2 \times 4.9\%$. That is, it's between 42.1% and 61.7%."** This is getting closer, but we still can't be certain. We can't know *for sure* that the true proportion is in this interval—or in any particular interval.
4. **"We don't know exactly what proportion of sea fans on the Las Redes Reef is infected, but the interval from 42.1% to 61.7% *probably* contains the true proportion."** We've now fudged twice—first by giving an interval and second by admitting that we only think the interval "probably" contains the true value. And this statement is true.

That last statement may be true, but it's a bit wishy-washy. We can tighten it up a bit by quantifying what we mean by "probably." We saw that 95% of the time when we reach out 2 *SE*s from $\hat{p}$ we capture *p, so we can be 95% confident that this is one of those times*. After putting a number on the probability that this interval covers the true proportion, we've given our best guess of where the parameter is and how certain we are that it's within some range.

5. **"We are 95% confident that between 42.1% and 61.7% of Las Redes sea fans are infected."** Statements like these describe **confidence intervals.** They're the best we can do.

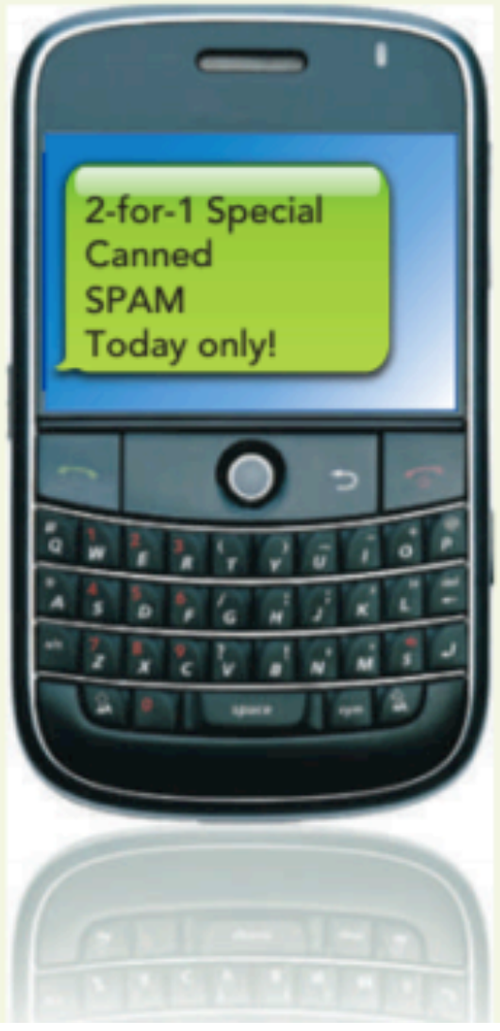
Each confidence interval discussed in the book has a name. You'll see many different kinds of confidence intervals in the following chapters. Some will be about more than *one* sample, some will be about statistics other than *proportions,* and some will use models other than the Normal. The interval calculated and interpreted here is sometimes called a **one-proportion *z*-interval.**[4]

Just Checking

A Pew Research study regarding cell phones asked questions about cell phone experience. One growing concern is unsolicited advertising in the form of text messages. Pew asked cell phone owners, "Have you ever received unsolicited text messages on your cell phone from advertisers?" and 17% reported that they had. Pew estimates a 95% confidence interval to be 0.17 ± 0.04, or between 13% and 21%.

Are the following statements about people who have cell phones correct? Explain.

1. In Pew's sample, somewhere between 13% and 21% of respondents reported that they had received unsolicited advertising text messages.
2. We can be 95% confident that 17% of U.S. cell phone owners have received unsolicited advertising text messages.
3. We are 95% confident that between 13% and 21% of all U.S. cell phone owners have received unsolicited advertising text messages.
4. We know that between 13% and 21% of all U.S. cell phone owners have received unsolicited advertising text messages.
5. 95% of all U.S. cell phone owners have received unsolicited advertising text messages.



What Does "95% Confidence" Really Mean?

What do we mean when we say we have 95% confidence that our interval contains the true proportion? Formally, what we mean is that "95% of samples of this size will produce confidence intervals that capture the true proportion." This is correct, but a little long winded, so we sometimes say, "we are 95% confident that the true proportion lies in our interval." Our uncertainty is about whether the particular sample we have at hand is one of the successful ones or one of the 5% that fail to produce an interval that captures the true value.

Back in Chapter 17 we saw that proportions vary from sample to sample. If other researchers select their own samples of sea fans, they'll also find some infected by the disease, but each person's sample proportion will almost certainly differ from ours. When they each try to estimate the true rate of infection in the entire population, they'll center *their* confidence intervals at the proportions they observed in their own samples. Each of us will end up with a different interval.

Our interval guessed the true proportion of infected sea fans to be between about 42% and 62%. Another researcher whose sample contained more infected fans than ours did might guess between 46% and 66%. Still another who happened to collect fewer infected fans might estimate the true proportion to be between 23% and 43%. And so on. Every possible sample would produce yet another confidence interval. Although wide intervals like these can't pin down the actual rate of infection very precisely, we expect that most of them should be winners, capturing the true value. Nonetheless, some will be duds, missing the population proportion entirely.

On the next page you'll see confidence intervals produced by simulating 20 different random samples. The red dots are the proportions of infected fans in each sample, and the blue segments show the confidence intervals found for each. The green line represents the true rate of infection in the population, so you can see that most of the intervals caught it—but a few missed. (And notice again that it is the *intervals* that vary from sample to sample; the green line doesn't move.)

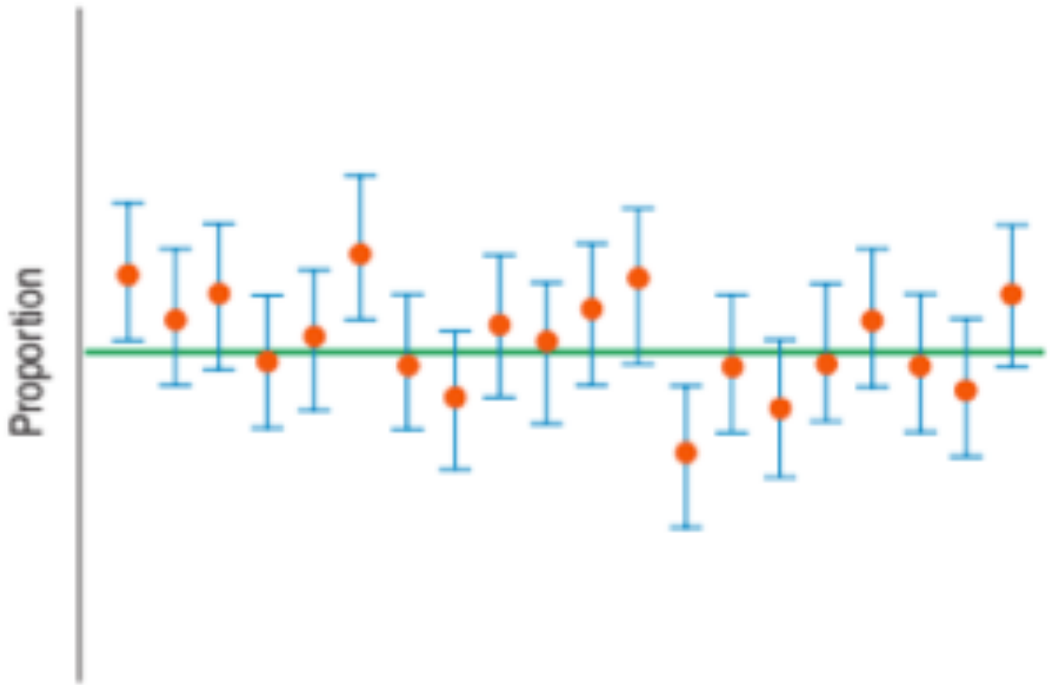


Figure 18.3

The horizontal green line shows the true percentage of all sea fans that are infected. Most of the 20 simulated samples produced confidence intervals that captured the true value, but a few missed.

Of course, there's a huge number of possible samples that *could* be drawn, each with its own sample proportion. These are just some of them. Each sample proportion can be used to make a confidence interval. That's a large pile of possible confidence intervals, and ours is just one of those in the pile. Did *our* confidence interval "work"? We can never be sure, because we'll never know the true proportion of all the sea fans that are infected. However, the Central Limit Theorem assures us that 95% of the intervals in the pile are winners, covering the true value, and only 5% are duds. *That's* why we're 95% confident that our interval is a winner!

So, What *Can* I Say?

Technically, we should say, "I am 95% confident that the interval from 42.1% and 61.7% captures the true proportion of sea fans on the Las Redes Reef that are infected." That formal phrasing emphasizes that *our confidence (and our uncertainty) is about the interval, not the true proportion.* But you may choose a more casual phrasing like "I am 95% confident that between 42.1% and 61.7% of Las Redes sea fans are infected." Because you've made it clear that the uncertainty is yours and you didn't suggest that the randomness is in the true proportion, this is OK. Keep in mind that it's the interval that's random and is the focus of both our confidence and doubt.

A Confidence Interval

The word "probably" ⇒ quantifying this

Sampling distribution $\hat{p}$ is centered at p.

and

Standard deviation of the sampling distribution is $\sqrt{\frac{pq}{n}}$.

But we don't know P, and when we estimate the standard deviation of sampling distribution, we call it standard error.

SE

$$SE_{(\hat{p})} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

For the SCA fans.

$\hat{p} = 0.519$

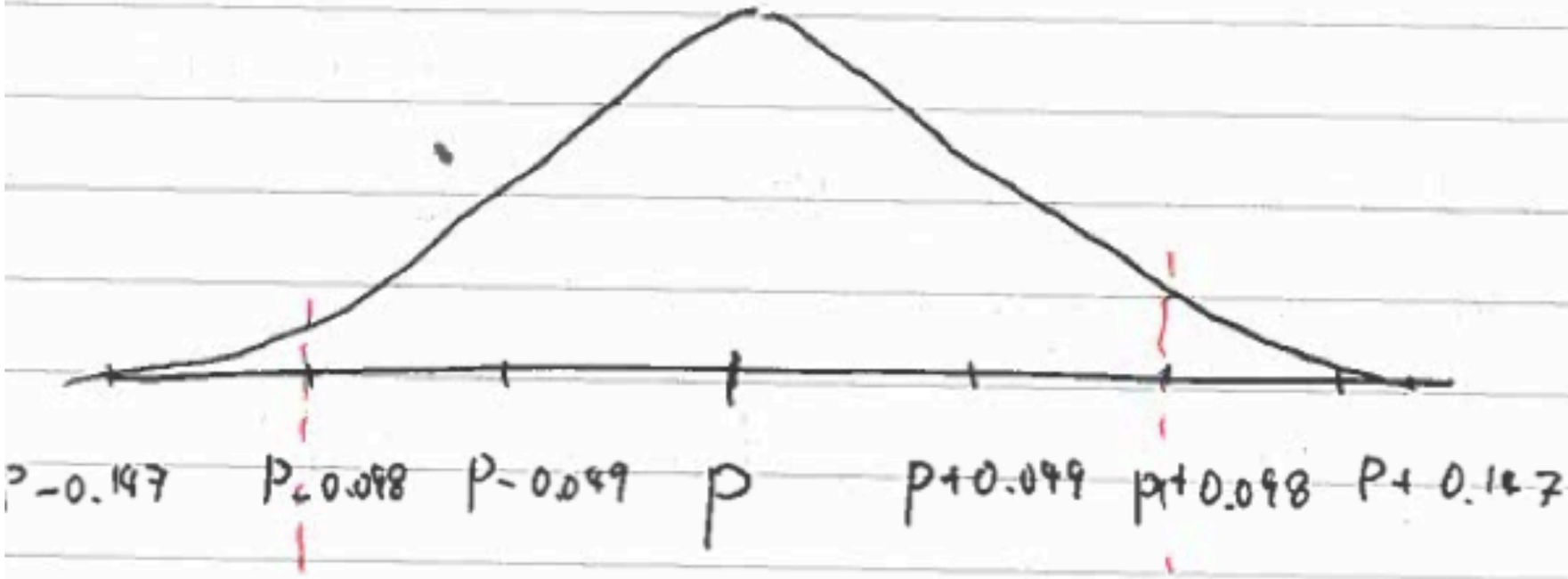
$\hat{q} = 1 - 0.519$

$= 0.481$

$SE(\hat{p}) = \sqrt{\frac{(0.519)(0.48}{104}}$

$= 0.049$

$= 4.9\%$



~~with $\hat{p} \pm 2$ SEs~~

68% of all sample of 104 sea fans will have $\hat{p}$ within 1 SEs.

95% of all sample of 104 sea fans will have $\hat{p}$ within 2 SEs.

No.

Date

Page 441

$\hat{p}$ view

from $\hat{P}$.

There is a 68% chance that p is no more than ± 1 SEs away ✓

There is a 95% " that P is no more than ± 1 SEs away

from $\hat{P}$.

----- what can we say about p?

1. "51.9% of all sea fans on the Las Redes Reef are infected"

No way to be sure

2. "It is probably true that 51.9% of all sea fans on the Las Redes Reef are infected."

No. we can pretty sure that it's NOT EXACTLY 51.9%.

3. We don't know exactly what proportion of sea fans on the Las Redes Reef is infected, but we know that it's within the interval $51.9\% \pm 2 \times 4.9\%$ (42.1% — 61.7%)

We still can be certain.

4. "We don't know exactly what proportion of sea fans on the Las Redos Reef is infected, but the interval from 42.1% to 61.7% probably contains the true proportion."

5. We are 95% confident that between 42.1% and 61.7% of Las Redes sea fans are infected.

There are many different kinds of confidence intervals.

The interval we just did in this chapter is sometimes called

a one-proportion z-intervals.

For Example POLLS AND MARGIN OF ERROR

In April and May 2011, the Yale Project on Climate Change Communication and the George Mason University Center for Climate Change Communication interviewed 1010 U.S. adults about American's global warming beliefs and attitudes.[5]

QUESTION: It is standard among pollsters to use a 95% confidence level unless otherwise stated. Given that, what do these researchers mean by their confidence interval in this context?

r Example POLLS AND MARGIN OF ERROR

In April and May 2011, the Yale Project on Climate Change Communication and the George Mason University Center for Climate Change Communication interviewed 1010 U.S. adults about American's global warming beliefs and attitudes.[5]

QUESTION: It is standard among pollsters to use a 95% confidence level unless otherwise stated. Given that, what do these researchers mean by their confidence interval in this context?

ANSWER: *If this polling were done repeatedly, 95% of all random samples would yield confidence intervals that contain the true proportion of all U.S. adults who believe that there's a lot of disagreement among scientists about global warming.*

---

[5]Among their questions, they asked what respondents thought were the views of scientists. Among respondents, 40% agreed with the alternative "There is a lot of disagreement among scientists about whether or not global warming is happening." The investigators provide a confidence interval from 37% to 43%.