Chapter 9, part 2: Confidence intervals

S520

These notes are written to accompany Trosset chapter 9.5.

Some things the CLT says are approximately Normal

To recap, suppose we have a large IID sample $X_1, \ldots, X_n$, where each X has finite expected value μ and finite variance σ^2. By the Central Limit Theorem:

1. The sample mean $\bar{X}$ is an approximately Normal random variable with expected value μ, variance σ^2/n, and standard deviation $\sigma/\sqrt{n}$.
2. The sum $S = \sum_{i=1}^n X_i$ is an approximately Normal random variable with expected value $n\mu$, variance $n\sigma^2$, and standard deviation $\sigma\sqrt{n}$.
3. Let the **error** of the sample mean be defined as sample mean minus population mean: $E = \bar{X} - \mu$. Then E is an approximately Normal random variable with expected value 0, variance σ^2/n, and standard deviation $\sigma/\sqrt{n}$.

Confidence intervals

The approximate normality of the error turns out to be the key property. We know that, for example, 68% of the time, a Normal random variable is between ± 1 standard deviation of its expected value. Well, we also know that E has expected value 0 and standard deviation $\sigma/\sqrt{n}$. So about 68% of the time, E will be between $\pm\sigma/\sqrt{n}$. That is, about 68% of the time, the difference between the sample mean and the true population mean μ is between $\pm\sigma/\sqrt{n}$.

This lets us hedge our bets. If someone asks us us "what is the population mean μ?", we could give a single number (**point estimate**) as our answer: our best guess is it's the observed sample mean, $\bar{x}$. But we probably wouldn't be exactly right – there's almost always some error.

Instead, we can give an **interval estimate**: "It's $\bar{x}$, plus or minus $\sigma/\sqrt{n}$." If we had the mean of a large sample and we knew what σ was, then the interval we'd get would contain the true μ about 68% of the time.

There's one catch: It's pretty rare that we know what the population standard deviation σ is. But if n is really large, then either the plug-in estimator $\hat{\sigma}$ or the sample standard deviation s should be close enough to σ to use for this purpose. If you really do have large n, then it shouldn't matter which of these two you use, but we'll use s simply because that's what the `sd()` function in R gives us. So for very large n, we can construct the interval estimate as

$$\bar{X} \pm \frac{s}{\sqrt{n}}$$

and about 68% of the time, we'll end up with an interval that contains the true value of μ. We call this a **68% confidence interval for the population mean.**

Well, 68% isn't that great – we generally want to be right more often than that. Let's say we want such a probability $1 - \alpha$ of getting an interval that contains the true μ. We'll call $1 - \alpha$ our **level of confidence**. The 68% above came about because going up and down one SD captures 68% of the Normal distribution:

```
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

To generalize, we need to find a q such that `pnorm(q) - pnorm(-q)` $= 1 - \alpha$. From the definition of quantiles and the symmetry of the Normal, this is just

$q =$ `qnorm(1 - alpha/2)`

Why? Then `pnorm(q)` is $1 - \alpha/2$ by definition, `pnorm(-q)` is $\alpha/2$ by symmetry, and `pnorm(q) - pnorm(-q)` is $1 - \alpha$ as required.

Example. What does q have to be to get a 95% chance of including the population mean?

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Double-check this works:

```
q = qnorm(0.975)
pnorm(q)
```

```
## [1] 0.975
```

```
pnorm(-q)
```

```
## [1] 0.025
```

```
pnorm(q) - pnorm(-q)
```

```
## [1] 0.95
```

Round this to `qnorm(0.975) = 1.96` and memorize it.

More generally, a $(1 - \alpha)$ Central Limit Theorem confidence interval for μ is

$$\bar{X} \pm q \frac{s}{\sqrt{n}}$$

where q is the $(1 - \alpha/2)$ quantile of the standard Normal distribution, i.e. `qnorm(1 - alpha/2)`.

Example: Faculty salaries

The file `faculty100.txt` contains a simple random sample of 100 IU Bloomington faculty salaries.

```
salaries = scan("faculty100.txt")
mean(salaries)
```

```
## [1] 91108.09
```

```
sd(salaries)
```

```
## [1] 48289.88
```

We now know how to construct a 95% confidence interval for the mean of *all* IU Bloomington faculty salaries. Take the sample mean and add and subtract 1.96 estimated SDx of the error.

```
mean(salaries) - qnorm(0.975) * sd(salaries) / sqrt(length(salaries))
```

```
## [1] 81643.45
```

```
mean(salaries) + qnorm(0.975) * sd(salaries) / sqrt(length(salaries))
```

```
## [1] 100572.7
```

We get the interval estimate \$81,643 to \$100,573. We don't care about a few dollars (and couldn't estimate to that accuracy even if we did), so round to the nearest thousand: \$82,000 to \$101,000.

A "confidence" trick

Once you've got the data, it's no longer straightforward to talk about probability. You can talk about probability before you take the sample, or if you're talking about a theoretical sample. But once you get a particular sample, that'll determine a particular interval, and it'll either contain the true value or it won't. So it's not strictly speaking true to say there's a 95% probability that a particular confidence interval you have in front of you contains the mean you're trying to estimate. It does or it doesn't. (Subjective probability doesn't save us here, as there is no obligation to have the subjective belief that the interval in front contains the population mean with 95% probability just because it's a 95% confidence interval – in fact, sometimes this would be a pretty laughable belief.)

Instead, we fudge and call it an (approximate) 95% confidence interval for μ. A 95% confidence interval is an interval estimate constructed in such a way that 95% of the time, you'll get an interval that contains the true value (where "95% of the time" implies a large number of repetitions.) That means 5% of the time you'll be wrong, but that's statistics.

We should note that just about everyone gets this interpretation of confidence intervals wrong – even people who have been doing statistics for years. Fortunately it's rarely a costly mistake, so only the most pedantic instructors would put in on your midterm and make it a decent fraction of your grade.

Confidence intervals for a proportion

In fact, we don't often use the method given above to find confidence intervals for a population mean. The reason is that s isn't quite the same as σ, so we usually prefer methods that specifically account for this extra uncertainty.

There's one exception where estimating σ is really easy. Suppose each X is a Bernoulli trial: it's either 0 or 1, and it's 1 with probability p. Back in chapter 4, we found that:

$$EX = p$$

and

$$\sigma = \sqrt{p(1-p)}$$

As long as we have a good-sized IID sample (and we don't have an extreme value of p like 0.02 or 0.98), then not only can we estimate p accurately, we can use this value of p to estimate σ accurately as well.

Let $\bar{X}$ be a random variable representing the average of n IID Bernoulli(p) trials. That is, it's just the proportion of the sample that are 1 (or say "yes" or have whatever characteristic you're studying.) Then by the CLT, for large n, $\bar{X}$ will have a Normal distribution. (Of course the *sum* will be Binomial, but here we're interested in the average.) The expected value of this Normal distribution will be $EX = p$, so the "sample proportion" $\bar{X}$ will be an unbiased estimator of p. The standard deviation of this Normal distribution will be $\sigma/\sqrt{n} = \sqrt{p(1-p)/n}$.

A $(1-\alpha)$ CLT confidence interval for the population proportion p is

$$\bar{X} \pm q\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

where $q =$ `qnorm(1 - alpha/2)`.

Note that in this Bernoulli case, the notation $\hat{p}$ ("p hat") is sometimes used instead of $\bar{X}$, to emphasize that our estimate is a sample proportion. We'll try to avoid this notation because we'll have a lot of p's and P's floating around throughout this chapter.

Example: Sample surveys

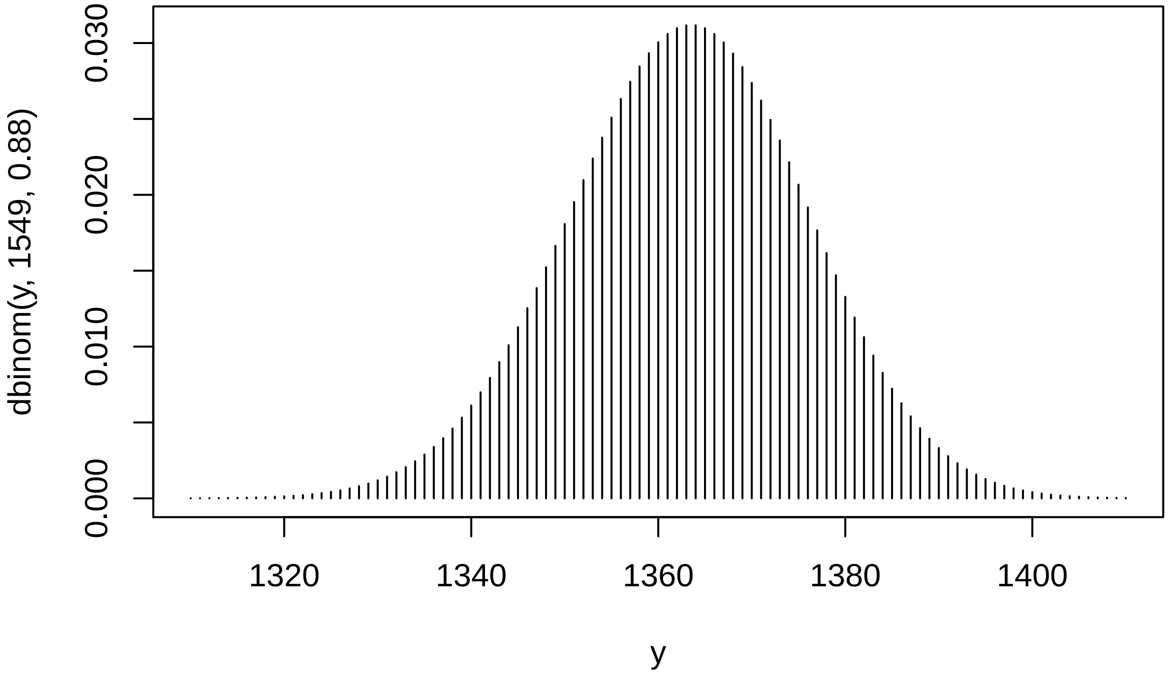
A 2016 Pew Research survey asked a sample of 1,549 U.S. adults their opinions on a number of health issues. Out of the sample, 88% agreed that "the benefits of childhood vaccines for measles, mumps and rubella outweigh the risks."

Let p be the proportion of all U.S. adults who would agree that "the benefits of childhood vaccines for measles, mumps and rubella outweigh the risks." What's a 95% confidence interval for p?

Before we jump to the formula, let's review the assumptions we'll be making.

- We assume we have an IID random sample from the population of U.S. adults. This is not literally true. Firstly, Pew almost certainly took their sample without replacement, but that hardly matters because the population of U.S. adults is so large that it doesn't matter if you replace or not. Secondly and importantly, some people never respond to surveys, so Pew almost certainly had to adjust the data to avoid bias due to nonresponse. This is completely legitimate, but it means the method we'll use to calculate confidence intervals won't be completely accurate. Nevertheless, we don't know exactly how they made their adjustments, so we'll stick with our method.
- We assume the sample size is large enough that the Central Limit Theorem kicks in and the distribution of $\bar{X}$ is approximately Normal. This should be fine: we should be a bit careful if we're dealing with an extreme probability, but 88% is not quite extreme. If you're not convinced, let's take a quick look at what a Binomial$(1549, 0.88)$ distribution looks like:

```
y = 1310:1410
plot(y, dbinom(y, 1549, 0.88), type="h")
```



Looks Normal. (You could take repeated samples, calculate $\bar{x}$, and draw a normal QQ plot, but that would be overkill.)

Let's proceed with the interval.

```
n = 1549
x_bar = 0.88
sd_error = sqrt(x_bar * (1 - x_bar) / n)
# Lower bound
x_bar - qnorm(0.975) * sd_error
```

```
## [1] 0.8638172
```

```
x_bar + qnorm(0.975) * sd_error
```

```
## [1] 0.8961828
```

Our 95% confidence interval goes from 86.4% to 89.6%.

Another way of writing this is 88% ± 1.6%. Here, 1.6 percentage points is the **margin of error at 95% confidence**. When a survey gives a margin of error, then unless otherwise stated it's almost always at 95% confidence.

But wait: if you Google this survey, you'll find that Pew states the margin of error for percentages based on this sample is 4%. That isn't especially close to our number. What happened?

There are two issues. One is that Pew's number doesn't just apply to this question, it's intended to apply to *all* questions based on this sample. The sample percentages for these questions aren't all 88%, they're all over the place. To be on the safe side, Pew gives the *maximum* margin of error for any such question. The true standard deviation of the error in a proportion is:

$$SD(error) = \sqrt{\frac{p(1-p)}{n}}$$

From calculus or by drawing a picture, we see this is maximized at $p = 0.5$. So $p = 0.5$ gives the maximum margin of error. Let's see if using this number gets our calculation any closer to Pew's.

```
max_MoE = qnorm(0.975) * sqrt(0.5 * 0.5 / 1549)
max_MoE
```

```
## [1] 0.0248996
```

Our calculation gives a maximum margin of error of 2.5 percentage points. We're still off.

The second and more important reason is what we alluded to above: Pew is doing some fairly complex corrections for nonresponse and other biases. Our method of finding confidence intervals and margins of error assumes no bias. But if there really is bias, our method will underestimate the margin of error. That means that if there's bias, our confidence intervals might not be right 95% (or 90% or 80%) of the time.

The morals:

- Our method for Central Limit Theorem confidence intervals, like just about all the methods you learn in introductory statistics, works best with large, unbiased samples.
- In many real-life situations, you don't have large, unbiased samples. That's not necessarily the end of the world – there are often methods designed for these situations, and they're often just wrinkles on things we learn in this course. You just have to learn them.
- If you have no choice but to use methods that assume large unbiased samples when your sample is small or biased, the performance may be poor or appalling. A method that's supposed to have 95% confidence may have a *dramatically* lower success rate.

Planning your study

Let's suppose you're planning a study, and you think you can take a random sample with no or negligible bias. One thing you *don't* want to happen is to end up with a confidence interval with a total length of 10 percentage points when you really wanted one with a total length of 4 percentage points. (By "total length" we mean the upper bound minus the lower bound.)

One way to proceed is to pick a length L for your confidence interval (and a level of confidence), then work out what value of n you need to get a confidence interval of that length.

For our CLT intervals for a population mean, if σ were known, the total length would be

$$\left(\bar{X} + q\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - q\frac{\sigma}{\sqrt{n}}\right)$$

which is just

$$2q\frac{\sigma}{\sqrt{n}}$$

So the idea is to set

$$L = 2q\frac{\sigma}{\sqrt{n}}$$

and solve for n. We'll leave it to you to do the algebra (or look up the answer on Trosset p. 225.)

L you decide for yourself (or are given) and q you find from `qnorm`. What about σ? If you don't know, just guess (or do a pilot study.) It's better to guess a bit high rather than a bit low, because you really don't want to spend a year doing a study and then find out your confidence interval is too wide because you underestimated σ.

The same idea holds for confidence intervals for p. Set

$$L = 2q\sqrt{\frac{p(1-p)}{n}}$$

The safest guess for p is the one that maximizes the margin of error i.e. $p = 0.5$.

Example. I want to take a simple random sample of U.S. to estimate the proportion that like ice cream. Assuming I can take an unbiased sample, how large a sample do I need to guarantee a 95% confidence interval of length at most 6 percentage points?

We write:

$$0.06 = 2 \times 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}}$$

and solve this for n; this gives $n = 1067.1$. We can't have 0.1 of a person, so we round; to be extra-cautious, we round up to $n = 1068$.

One-sided confidence intervals

These exist but I've never used them, since in practice I don't know why you would want (for example) just a lower bound rather than both a lower and an upper bound. See Trosset 9.5.2 should you ever want to make one of these.

Extra for nerds: Do confidence intervals work?

Let's do a simulation. We'll repeatedly take simple random samples of size 100 from the population of IU Bloomington faculty salaries (in the file `faculty.txt`, scanned into R as `salaries.all`.) We'll create a (nominal) 95% confidence interval from each sample. Then we'll see what percentage of the confidence intervals we get contain the true population mean, which we can find exactly using `mean(salaries.all)`. Since we're taking a truly unbiased sample, we should get reasonably close to 95%. (The full R syntax here is beyond the scope of the course, but copy-paste if you feel like it.)

```
salaries.all = scan("faculty.txt")
faculty.interval = function(n){
  salaries.sample = sample(salaries.all, size=n)
  x.bar = mean(salaries.sample)
  s = sd(salaries.sample)
  lower = x.bar - qnorm(0.975) * s / sqrt(n)
  upper = x.bar + qnorm(0.975) * s / sqrt(n)
  return(c(lower, upper))
}
simulations = replicate(1000, faculty.interval(100))
lower.list = simulations[1,]
upper.list = simulations[2,]
too.low = sum(upper.list < mean(salaries.all))
too.high = sum(lower.list > mean(salaries.all))
just.right = 1000 - too.low - too.high
print(c(too.low, too.high, just.right))
```

```
## [1]  40  14 946
```

It's a bit lower than 95%, but close. As hinted above, in chapter 10 we'll learn a method that has closer to the right level of coverage.