POL 345: Asymptotic Confidence Intervals

Tolya Levshin

December 5, 2015

This booklet provides a thorough discussion of how to calculate asymptotic confidence intervals. The mathematics presented here is a tad more advanced than what we've seen so far in this course, but it's not expected that you understand every derivation and every qualification offered herein. For those who are not mathematically-inclined, or are just short on time, feel free to skip to the general summary of the key points on pp. 34-5. You can then browse through those sections of the booklet on which you need more clarification.

Imagine that a coin is tossed six times, and heads come up five times. It may be reasonable to conclude that the coin is biased — indeed, our best guess would be that $\hat{p}_{\text{heads}} = \frac{\text{no. of heads}}{\text{no. of heads+no. of tails}} = \frac{5}{5+1} = \frac{5}{6}$. How accurate is your guess? How likely is it that the coin is actually fair? Or, perhaps, less biased than you deduced? How likely is it that the true probability of heads is even more extreme than you calculated?

Alternatively, suppose that you ask ten students chosen at random from a class of 500, their scores on a recent test. You record their responses and calculate the average — let's say, your calculations yield 84%. How confident can you be that the actual class average, across all 500 students, is only a certain number of points removed from the average that you calculated? Could it be that the students you interviewed aced the test, and so the class-wide average is actually lower, perhaps considerably so? Or, possibly, your interviewees did poorly relative to their peers?

1 Requirements for Uncertainty Estimation

It may be tempting to conclude that, before we can answer these questions, we first need to learn about the data-generating processes that produced the few observations we sampled — for example, their distributional form (is the process given by a

binomial or a normal distribution?) as well as their parameters (what are its mean and standard deviation?). Suprisingly, we require neither. In fact, we can produce arbitrarily accurate answers by making only two assumptions *about our sample* and one assumption about the broader population from which the sample was drawn:

1. the observations in our sample are **statistically independent** — that is to say, the outcome of one observed coin toss does not prejudge future tosses, nor does the test score reported by one student relate to the test score of any other student we interview.

2. the observations in our sample come from the **same population-level distribution**. This assumption excludes the possibility that different observations may be produced by different data-generating processes.

3. the population-level distribution has finite mean and variance. We do not need to know what the distribution is, nor its parameters, but it is important to exclude distributions that lack finite mean and variance. (One prominent example of the latter is the Cauchy distribution, which we do not cover in this course.) The distributions that we work with in this course, like the binomial and the normal, all have defined and finite means and variances, and so you can safely stipulate this assumption as you begin work on the next problem set.

Let X_i denote the value of the i^{th} observation in our sample on some variable of interest, X. For example, X_i could represent whether heads came up on the i^{th} toss of a coin, or the grade of the i^{th} student we interviewed. Let $\mathcal{D}$ denote the (possibly unknown) population-level distribution from which we get to collect only a small sample of observations.

We summarize the first two assumptions in formal notation thusly:

$$X_i \overset{\text{i.i.d}}{\sim} \mathcal{D} \tag{1}$$

Recall that the sign $\sim$ should be read "distributed as", and, in this context, means that, for each observation i, its value on X is a draw from $\mathcal{D}$. `i.i.d.` stands for `independently and identically distributed`: that is to say, each observation is independent from any other (this is the first assumption) and drawn from the same distribution as all the others (this is the second assumption). If our sample consists of such X_i, we then say that it is an i.i.d. sample. You are already familiar with techniques for generating i.i.d. samples from a population of interest: simple random sampling is one important example.

Let's return to the example of interviewing ten students at random from a class of 500. Let X_i denote student i^{th}'s grade, and let $\bar{X}$ denote the average grade of the ten students we interviewed:

$$\bar{X} \;=\; \frac{1}{10}\sum_{i=1}^{10} X_i$$

Now, suppose that the full distribution of the grades in the class is, in fact, normal, with the mean of 80 and standard deviation of 6: $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(80, 36)$. Let $\mu = 80$ denote the class mean, and $\sigma = 6$ the class standard deviation. We will not use this information in the course of our estimation, but it is important to see how to generate an i.i.d. sample from a given population. It is easy to define this class in **R**:

```
> class <- rnorm(n=500, mean=80, sd=6)
```

The function `rnorm(n, mean, sd)` returns `n` random draws from a normal distribution with mean `mean` and standard deviation `sd`.

```
> ## Peek at the grades of the first six students in the class:
> head(class)

[1] 80.90 75.72 81.54 84.55 78.98 96.06

>
> ## Check the average grade in the whole class:
> mean(class)

[1] 79.94

>
> ## Notice that the class average is not quite 80, because
> ## our population in this case is finite and modest in size.
> ## If we were to increase the size of the class considerably,
> ## the class average would become exactly 80.
```

Next, to generate an i.i.d. sample of ten, we simply pick ten students at random from that class, without replacement:

```
> iid.sample <- sample(class, size=10, replace = F)
>
> ## View the grades of students in our sample
> iid.sample

 [1] 75.93 86.79 79.67 79.84 72.81 85.93 84.89 80.56 75.54
[10] 84.65

>
> ## Calculate the mean grade among the students in our sample:
> mean(iid.sample)

[1] 80.66
```

So, $\bar{X}$ =80.66, and the estimation error is $\mu - \bar{X}$ =-0.72. Not bad. In fact, the Weak Law of Large Numbers guarantees that, as the size of our sample grows, the sample mean will converge to the population mean. This means that the mean of an i.i.d. sample is a **consistent** estimator for the population mean. (We will return to this key concept, consistency, in a moment.)

2 Sampling Distributions

Now, in a real estimation problem, you will not know μ and, so, won't be able to evaluate exactly just how accurate $\bar{X}$ is as a guess for the true, but unobserved, value of μ. Instead, you will want to make use of the **sampling distribution** of $\bar{X}$. We formally define the sampling distribution of an estimator $\hat{\theta}$ of some unknown parameter θ as the entire range of values that $\tilde{\theta}$ assumes across (infinitely) many samples, weighted by the frequency of their occurence. This sounds a bit abstract. Let's make this definition more concrete. Suppose that, instead of drawing just one sample from the class, we now draw two more independent samples. Be sure to keep the size of the samples constant throughout:

```
> iid.sample.2 <- sample(class, size=10, replace=F)
>
> iid.sample.3 <- sample(class, size=10, replace=F)
```

The sampling distribution of $\bar{X}$ across the three samples is then given by:

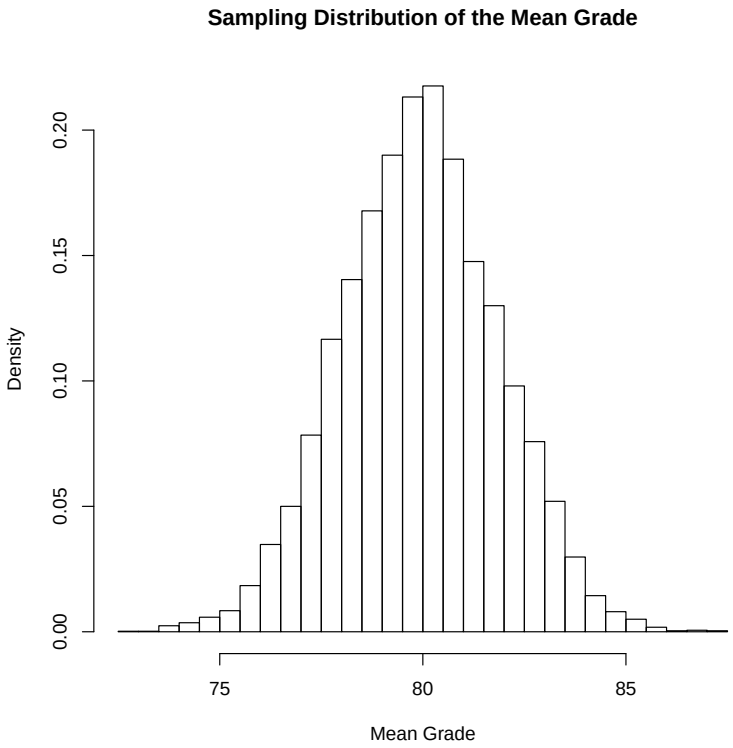
```
> c(mean(iid.sample), mean(iid.sample.2), mean(iid.sample.3))

[1] 80.66 81.35 81.01
```

Now, suppose we draw many more samples from the class, each of the same size as all the others, and record the average grade for each sample. This is easy to do in **R**:

```
> ## Create an empty container vector to record the mean grade
> ## of each sample
> sample.means <- vector()
>
> ## Draw 10,000 samples, each consisting of ten students,
> ## from the class of 500
> for (i in 1:10000) {
+   iid.sample.i <- sample(class, size=10, replace=F)
+   sample.means[i] <- mean(iid.sample.i)
+ }
```

Next, plot the sampling distribution of $\bar{X}$:

```
> hist(sample.means, freq=F, breaks=30,
+      main="Sampling Distribution of the Mean Grade",
+      xlab="Mean Grade")
```



The histogram shows the distribution of $\bar{X}$ calculated 10,000 times across 10,000 random samples. We can see that some samples produced $\bar{X}$ — for example, 74% or 86% — that fell quite far from the true population mean of $\mu = 80\%$. However, as the histogram shows, these wildly inaccurate estimates are relatively infrequent. Indeed, the sampling distribution concentrates most of its mass around the 78-82% range, spiking at $\bar{X} = 80\%$. This tells us that the single most frequent sample mean, across the 10,000 samples we obtained, is exactly 80%.

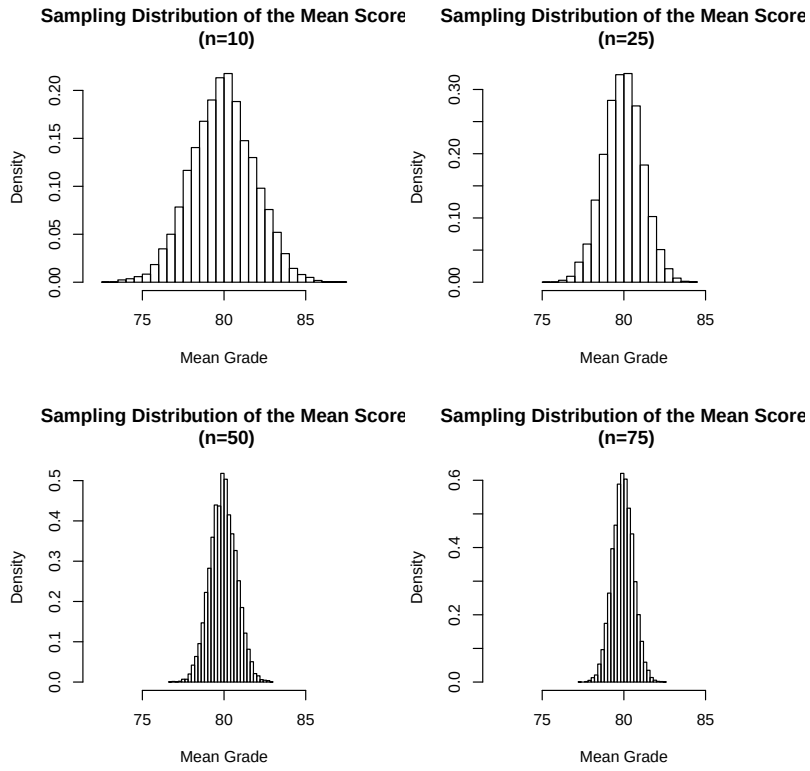
More generally, the sampling distribution of the mean of an i.i.d. sample will *always* display two powerful properties: **asymptotic consistency** and **asymptotic normality**. Consistency refers to the fact that, as the size of our samples begins to grow, the sampling disstribution of the sample mean will begin to converge, ever more tightly, to the population mean. Suppose, for example, that instead of surveying ten students for each of our samples, we now interview 25, 50, and 75:

```
> sample.means.25 <- vector()
> sample.means.50 <- vector()
> sample.means.75 <- vector()
>
```

```
> ## For each iteration, draw samples of 15, 20, and 25
> ## students; then, calculate the means for each of the
> ## three samples:
> for (i in 1:10000) {
+   sample.25 <- sample(class, size=25, replace=F)
+   sample.50 <- sample(class, size=50, replace=F)
+   sample.75 <- sample(class, size=75, replace=F)

+   sample.means.25[i] <- mean(sample.25)
+   sample.means.50[i] <- mean(sample.50)
+   sample.means.75[i] <- mean(sample.75)
+ }
```

Next, plot the histograms for the sampling distributions of $\bar{X}$ for each of the four different sample sizes — 10, 25, 50, and 75 — side-by-side:

```
> par(mfrow=c(2,2))
> hist(sample.means, freq=F, breaks=30,
+      main="Sampling Distribution of the Mean Score\n (n=10)",
+      xlab="Mean Grade", xlim=c(72,88))
> hist(sample.means.25, freq=F, breaks=30,
+      main="Sampling Distribution of the Mean Score\n (n=25)",
+      xlab="Mean Grade", xlim=c(72,88))
> hist(sample.means.50, freq=F, breaks=30,
+      main="Sampling Distribution of the Mean Score\n (n=50)",
+      xlab="Mean Grade", xlim=c(72,88))
> hist(sample.means.75, freq=F, breaks=30,
+      main="Sampling Distribution of the Mean Score\n (n=75)",
+      xlab="Mean Grade", xlim=c(72,88))
```



We can see that $\bar{X}$ is a consistent estimator for $\mu = 80$ (in formal notation, we write: $\bar{X} \xrightarrow{p} \mu$): as the size of our samples — **not** the number of times we repeatedly sample from the population distribution, but the size of each individual sample — increases, the sampling distribution of $\bar{X}$ begins to close in on μ. Indeed, as soon as our samples become sufficiently large, the sampling distribution of $\bar{X}$ will concentrate all of its mass on a single value, $\mu = 80$. The asymptotic consistency of the mean of an i.i.d. sample is **guaranteed** by a famous theorem known as the **Weak Law of Large Numbers**.

Asymptotic normality refers to the fact that the sampling distribution of $\bar{X}$ is approximately normal. Notice that, in calculating the sampling distribution, we did not draw on our knowledge of the population distribution, $\mathcal{N}(80, 36)$. This reflects a more general principle: so long as the population distribution has finite mean and variance, the sampling distribution of the sample mean will *necessarily* be asymptotically normal.[1] This means that we do not need to know anything about the

[1] *Asymptotic* here means simply that the quality of the approximation to the normal distribution depends on the size of our samples; as the size of the samples increases — that is to say, asymptotically — the sampling distribution of the sample mean will approximate a normal distribution

population distribution, except that it has finite mean and variance, to specify the sampling distribution of the sample mean calculated across multiple samples drawn from that distribution. This result is known as the Lindenberg-Lévy **Central Limit Theorem**, and it is formally written thusly:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \quad \xrightarrow{d} \quad \mathcal{N}\Big(\underbrace{\mu}_{=\mathbb{E}(X_i)}, \quad \underbrace{\frac{\sigma^2}{n}}_{=\mathbb{V}(X_i)/n} \Big), \tag{2}$$

where n is the size of our sample, X_i is the value of observation i on X, $\bar{X}$ is the sample mean, $\xrightarrow{d}$ should be read as "convergence in dsitribution", and μ and σ^2 are the mean (expectation) and variance of the population distribution from which our sample is drawn. One alternative, and more popular, way to write the theorem puts $\bar{X}$ on the $z-$scale:

$$\sqrt{n}\Big(\frac{\bar{X} - \mu}{\sigma}\Big) \quad \xrightarrow{d} \quad \mathcal{N}(0, 1) \tag{3}$$

3 Calculating the Mean and Variance of the Sampling Distribution

The Central Limit Theorem reveals a fascinating relationship between the population distribution and the sampling distribution of the sample mean: their means are exactly equal, and their variances differ in scale by the factor of n, the size of the sample. Why must this be? Although it is difficult to prove the asymptoptic normality of the sampling distribution, it is quite straightforward to derive its expectation, $\mathbb{E}(\bar{X})$, and variance, $\mathbb{V}(\bar{X})$. Before we begin, remember that our sample is independently and identically distributed: no two X_i in the sample are statistically dependent, and each $X_i \sim \mathcal{N}(\mu, \sigma^2)$. We specified above that $\mu = 80$ and $\sigma = 6$, but it is not particularly relevant how each X_i is distributed, so long as all are drawn from the same distribution — we may as well write $X_i \sim \mathcal{D}$ for some arbitrary distribution $\mathcal{D}$ — and that distribution has finite mean and variance. Next, substitute $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

more perfectly.

into the expressions for the mean and variance of the sampling distribution:

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n}\mathbb{E}(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n}\mathbb{E}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n}\Big(\mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)\Big)
\end{aligned}$$

At this point, recall that, for any i, $\mathbb{E}(X_i) = \mu$, since all observations are drawn from the same population-level distribution. Thus,

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \frac{1}{n}\Big(\underbrace{\mathbb{E}(X_1)}_{=\mu} + \underbrace{\mathbb{E}(X_2)}_{=\mu} + \cdots + \underbrace{\mathbb{E}(X_n)}_{=\mu}\Big) \\
&= \frac{1}{n}\Big(\underbrace{\mu + \mu + \cdots + \mu}_{\text{a total of } n\ \mu\text{s here}}\Big) \\
&= \frac{1}{n} n\mu = \mu = \mathbb{E}(X_i) \\
\mathbb{E}(\bar{X}) &= \mathbb{E}(X_i)
\end{aligned}$$

The proof for the variance of the sampling distribution proceeds similarly:

$$\begin{aligned}
\mathbb{V}(\bar{X}) &= \mathbb{V}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \\
&= \Big(\frac{1}{n}\Big)^2 \mathbb{V}\Big(\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n^2}\mathbb{V}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}\Big(\mathbb{V}(X_1) + \mathbb{V}(X_2) + \cdots + \mathbb{V}(X_n)\Big)
\end{aligned}$$

Once again, recall that all X_i are drawn from the same distribution, and, so, for every i, $\mathbb{V}(X_i) = \sigma^2$. Plug this into the formula:

$$\begin{aligned}
\mathbb{V}(\bar{X}) &= \frac{1}{n^2}\Big(\underbrace{\mathbb{V}(X_1)}_{=\sigma^2} + \underbrace{\mathbb{V}(X_2)}_{=\sigma^2} + \cdots + \underbrace{\mathbb{V}(X_n)}_{=\sigma^2}\Big) \\
&= \frac{1}{n^2}\Big(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{\text{a total of } n\ \sigma^2\text{s here}}\Big) \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} = \frac{\mathbb{V}(X_i)}{n} \\
\mathbb{V}(\bar{X}) &= \frac{\mathbb{V}(X_i)}{n}
\end{aligned}$$

Thus, even without knowing (almost) anything about the population distribution, we can deduce its mean and variance from the sampling distribution of the sample mean. To confirm that these results do, in fact, obtain, let's return to our example of interviewing students after a test. Recall that we computed the sampling distributions for the mean test score across 10,000 samples of size 10, 25, 50, and 75 students. Compare the means these sampling distributions to the population mean:

```
> ## Compute the class mean
> mean(class)

[1] 79.94

>
> ## Compute the means of the four sampling distributions:
> mean(sample.means)

[1] 79.93

> mean(sample.means.25)

[1] 79.93

> mean(sample.means.50)

[1] 79.95

> mean(sample.means.75)

[1] 79.93
```

Let's further confirm that the variance of $\bar{X}$ is smaller than the class variance by exactly the factor of the size of the sample:

```
> ## Compute the class variance
> var(class)

[1] 37.01

>
> ## Compute the variances of the four sampling distributions,
> ## and multiply each by the size of the respective sample:
> var(sample.means)*10

[1] 36.77

> var(sample.means.25)*25

[1] 35.06

> var(sample.means.50)*50

[1] 32.97

> var(sample.means.75)*75

[1] 30.95
```

One obvious difficulty with applying these results is that, in the real world, we can seldom afford to compute the exact sampling distribution of our estimator. We are often limited to drawing just one sample from a population and calculating a single estimate for that sample. Can we infer anything about the sampling distribution of an estimator, and, by implication, about the population distribution, from just one sample? Perhaps surprisingly, the answer is yes.

4 Confidence Intervals

To begin with, $\bar{X}$ is not just a good guess for the population mean — as we've discussed, the Weak Law of Large Numbers guarantees that $\bar{X}$ is an asymptotically consistent estimate for μ. This does not mean that, for a relatively small sample, its mean will be sufficiently close to μ. But the Central Limit Theorem, which builds

on the Weak Law, does allow us to gauge the proximity of the sample mean, for a given sample, to the population mean. Recall that the variance of the sampling distribution of $\bar{X}$ is given by $\mathbb{V}(X_i)/n = \sigma^2/n$. We can't calculate this quantity exactly, because we don't know what the population variance, σ^2, is. But suppose that we use the sample variance, $\hat{\sigma}^2$, which we do know, as our best guess for the population variance. This guess yields the following estimate for the variance of the sampling distribution of $\bar{X}$:

$$\widehat{\mathbb{V}(\bar{X})} = \frac{\hat{\sigma}^2}{n}$$

The standard deviation of the estimated sampling variance, $\sqrt{\widehat{\mathbb{V}(\bar{X})}} = \hat{\sigma}/\sqrt{n}$, is known as the **standard error** of the estimator, $\bar{X}$. We interpret standard errors as **the estimated average distance from the mean of a sampling distribution**. It is easy to calculate the standard error for $\bar{X}$ from the very first sample of ten students that we created:

```
> se <- sd(iid.sample) / sqrt(length(iid.sample))
>
> se

[1] 1.531
```

We can combine the standard error with the fact that the sampling distribution of $\bar{X}$ is asymptotically normal to gauge how proximate $\bar{X}$ is to the true mean of the sampling distribution, $\mathbb{E}(\bar{X})$, and, by implication, to the true population mean, μ. Suppose, for example, that we should like to construct a confidence interval around our sample mean, calculated from the very first sample of ten students that we interviewed, that contains the true class mean with 95% probability. Or, to formulate the problem in mathematical terms, we want to find the number t, known as the **Margin of Error**, such that

$$p(|\bar{X} - \underbrace{\mathbb{E}(\bar{X})}_{=\mu}| < t) = 0.95$$

It is easy to show that, for any two numbers a and c, $|a| < c \Leftrightarrow -c < a < c$. Therefore,

$$p(|\bar{X} - \mu| < t) = 0.95 \Leftrightarrow p(-t < \bar{X} - \mu < t) = 0.95$$

Why did we include the absolute value operator, instead of simply specifying $\bar{X} - \mu < t$? Because we want to caclulate the probability that the *distance* between the mean of our sample and the true population mean is smaller than the margin of error. That is, we want to calculate an average measure of the proximity of the sample and population means, and that requires us to take account of two possibilities: the sample mean can be either greater than the population mean, or it can smaller than the population mean. In both cases, we want to ensure that the distance between the two is no more than the margin of error.

Remember that Central Limit Theorem stipulates that, as the size of our sample increases, the sampling distribution of $\bar{X}$ becomes normal, and the sampling distribution of the corresponding z-scores, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, becomes standard normal:

$$\sqrt{n}\Big(\frac{\bar{X}-\mu}{\sigma}\Big) \quad \stackrel{d}{\rightarrow} \quad \mathcal{N}(0,1) \tag{4}$$

Thus, the original problem of finding the appropriate t to construct our confidence interval can be simplified by dividing all sides of the inequality inside the probability operator by $\sigma/\sqrt{n}$:

$$\begin{aligned} p(-t < \bar{X} - \mu < t) &= 0.95 \\ p\Big(-\sqrt{n}\frac{t}{\sigma} < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < \sqrt{n}\frac{t}{\sigma}\Big) &= 0.95 \end{aligned}$$

Because we know that $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ is distributed standard normal, it is easy to rewrite the above probability in terms of the standard normal CDF, commonly denoted Φ:

$$\begin{aligned} p\Big(-\sqrt{n}\frac{t}{\sigma} < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < \sqrt{n}\frac{t}{\sigma}\Big) &= 0.95 \\ \Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) - \Phi\Big(-\sqrt{n}\frac{t}{\sigma}\Big) &= 0.95, \end{aligned}$$

where we make use of the fact that, for CDFs of continuous random variables, $p(a < X < b) = F(b) - F(a)$. It can further be shown that, for CDFs of symmetric distributions, such as the normal, F(-a)=1-F(a). Making use of this second rule, the

expression transforms further as follows:

$$\begin{aligned}
\Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) - \Phi\Big(-\sqrt{n}\frac{t}{\sigma}\Big) &= 0.95 \\
\Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) - \left(1 - \Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big)\right) &= 0.95 \\
2 \times \Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) - 1 &= 0.95 \\
2 \times \Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) &= 1.95 \\
\Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) &= 0.975
\end{aligned}$$

To solve for t, we simply need to calculate the quantile of the standard normal distribution that corresponds to the probability of 0.975. This is easy enough to do in **R** using the `qnorm` function:

```
> ## qnorm(p) accepts a probability value p, and
> ## returns the quantile of the standard normal
> ## distribution corresponding to that probability:
> qnorm(0.975)

[1] 1.96
```

Thus,

$$\begin{aligned}
\Phi\Big(\sqrt{n}\frac{t}{\sigma}\Big) = 0.975 \quad \Leftrightarrow \quad \sqrt{n}\frac{t}{\sigma} &= 1.96 \\
t &= 1.96 \times \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

The only remaining difficulty is that we don't know the exact standard deviation of the sampling distribution, $\frac{\sigma}{\sqrt{n}}$. Instead, we estimate it using the standard error for $\bar{X}$, $\frac{\hat{\sigma}}{\sqrt{n}}$, thus yielding the following solution:

$$t = 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}$$

This was a mouthful. Let's return to the original problem of putting together a 95% confidence interval for μ: we wanted to find a number t such that $p(|\bar{X} - \mu| < t) = 0.95$. We solved for the margin of error, t, and found that it is equal to exactly

$1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}$. To put together the confidence interval, we require just a few more algebraic steps:

$$p\left(|\bar{X} - \mu| < 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95$$

$$p\left(-1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} < \bar{X} - \mu < 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95$$

Now, subtract $\bar{X}$ from the inequalities:

$$p\left(-1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} - \bar{X} < \bar{X} - \mu - \bar{X} < 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} - \bar{X}\right) = 0.95$$

$$p\left(-1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} - \bar{X} < -\mu < 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} - \bar{X}\right) = 0.95$$

Finally, multiply the inequalities by -1, keeping in mind that $-a < -x < -c \Leftrightarrow a > x > c$:

$$p\left(1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} + \bar{X} > \mu > -1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} + \bar{X}\right) = 0.95$$

$$p\left(1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} + \bar{X} > \mu > \bar{X} - 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95$$

$$p\left(\bar{X} - 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95$$

This is, then, the general formula for calculating a 95% asymptotic confidence interval for $\bar{X}$:

$$(\bar{X} - 1.96 \times \texttt{standard error}, \bar{X} + 1.96 \times \texttt{standard error})$$

Let's compute the 95% confidence interval for the mean of the sample of ten students that we interviewed:

```
> ## Calculate the margin of error first:
> moe <- qnorm(0.975) * se
>
> ## Or, equivalently:
> moe <- 1.96 * se
>
> ## Calculate the confidence interval:
> c((mean(iid.sample) - moe),
+   (mean(iid.sample) + moe))

[1] 77.66 83.66
```

We can see that the true population mean lies inside the confidence interval.

Interpreting the meaning of the 95% confidence interval is straightforward, at least in frequentist terms: if you were to repeatedly sample from the population of the entire class (infinitely) many times, keeping the size of your samples constant throughout, and recalculate the confidence interval for each of the new samples, then **the true class mean**, which is also the true mean of the sampling distribution of $\bar{X}$, **will be found inside those confidence intervals for nearly 95% of the samples**. (We say "nearly" here rather than "exactly" because we performed two approximations to compute the margin of error: first, we assumed that the size of our sample was sufficiently large to justify invoking the Central Limit Theorem; and, second, we approximated the standard deviation of the sampling distribution of $\bar{X}$ with the standard error calculated from our single sample. The quality of our confidence interval depends directly upon the adequacy of these two approximations.) Because we know what the true value of μ is in this example, we can confirm this property of confidence interval for ourselves:

```
> ## Draw 1,000 samples of 30 students from the class;
> ## construct a 95% confidence interval for each sample,
> ## and record whether the true mean lies inside that
> ## interval using a binary indicator variable
> indicator <- vector()
>
> for (i in 1:1000) {
+   ## Draw a sample of ten students without replacement
+   sample.i <- sample(class, size=30, replace=F)

+   ## Calculate the sample mean:
+   mean.i <- mean(sample.i)
```

```
+   ## Calculate the standard error for the current sample:
+   se.i <- sd(sample.i) / sqrt(length(sample.i))

+   ## Compute the upper and lower bounds of the
+   ## confidence interval for the current sample:
+   lower.i <- mean.i - qnorm(0.975) * se.i
+   upper.i <- mean.i + qnorm(0.975) * se.i

+   ## Record whether the true mean lies inside the
+   ## calculated interval; set the value of the
+   ## indicator variable for the current sample to 1
+   ## if so, to 0 if otherwise
+   indicator[i] <- ifelse(mean(class) >= lower.i &
+                          mean(class) <= upper.i, 1, 0)
+ }
>
> ## Calculate the proportion of samples for which the
> ## 95% confidence interval encompassed the true class mean
> mean(indicator)

[1] 0.951
```

Clearly, one way to improve the accuracy of your confidence intervals is simply to increase the size of your sample. This will reduce the margin of error at the rate of $\frac{1}{\sqrt{n}}$. Conversely, reducing the sample size widens the confidence interval.

Calculating Arbitrary Confidence Intervals

Suppose that you should like to calculate a narrower or wider confidence interval. You would begin by setting the desired **significance level**, α. This specifies the proportion of times that the true population mean will, over (infinitely) repeated sampling, escape your confidence interval. For example, setting $\alpha = 0.05$ or 5% yields the 95% confidence interval: $1 - \alpha = 1 - 0.05 = 0.95$. Recall that the size of the confidence interval entered our calculation of the margin of error at one particular

step, which I reproduce for you below:

$$\begin{aligned}
\Phi\left(\sqrt{n}\frac{t}{\sigma}\right) - \Phi\left(-\sqrt{n}\frac{t}{\sigma}\right) &= 0.95 \\
\Phi\left(\sqrt{n}\frac{t}{\sigma}\right) - \left(1 - \Phi\left(\sqrt{n}\frac{t}{\sigma}\right)\right) &= 0.95 \\
2 \times \Phi\left(\sqrt{n}\frac{t}{\sigma}\right) - 1 &= 0.95 \\
2 \times \Phi\left(\sqrt{n}\frac{t}{\sigma}\right) &= 1.95 \\
\Phi\left(\sqrt{n}\frac{t}{\sigma}\right) &= 0.975 \\
\sqrt{n}\frac{t}{\sigma} &= 1.96 \\
t &= 1.96 \times \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Notice that $0.975 = 1 - \frac{0.05}{2} = 1 - \frac{\alpha}{2}$, for a significance level $\alpha = 0.05$ corresponding to the 95% confidence interval. More generally, it is true that, for an arbitrary confidence interval, the margin of error is given as follows:

$$\begin{aligned}
\Phi\left(\sqrt{n}\frac{t}{\sigma}\right) &= 1 - \frac{\alpha}{2} \\
\sqrt{n}\frac{t}{\sigma} &= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\
t &= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \times \frac{\sigma}{\sqrt{n}},
\end{aligned}$$

where Φ^{-1} is the quantile function of the standard normal distribution, given in **R** by `qnorm`. Therefore, to calculate the confidence interval for some abritrary significance level in **R**, you can use the following code:

```
> ## Specify the significance level; for example, for a 93%
> ## confidence interval, set alpha = 1 - 0.93 = 0.07
> alpha <- 0.07
>
> ## Next, calculate the margin of error:
> moe <- qnorm(1-alpha/2) * se
>
> ## Calculate the lower bound of the confidence interval:
> mean(iid.sample) - moe
```

```
[1] 77.88

>
> ## Calculate the upper bound of the confidence interval:
> mean(iid.sample) + moe

[1] 83.43
```

As you can see, selecting more permissive (higher) values of α narrow the confidence interval, whereas lower significance levels, corresponding to more accurate confidence intervals, produce wider bounds.

5 Applications

We have so far considered calculating point-estimates and confidence intervals for the mean of an unknown population. But suppose that you you should find yourself asked to conduct more complicated estimations. How would you calculate confidence intervals for cross-sectional, before-and-after, or difference-in-differences comparisons? This is not as straightforward a problem as it may at first appear. Remember that the Central Limit Theorem applies to only one type of estimator: the mean of an i.i.d. sample drawn from a population with finite mean and variance. The theorem does not tell us anything about the asymptotic behaviour of more complex estimators, such as, for example, the difference of two sample means.

Fortunately, it is possible to show that, under certain conditions, the asymptotic sampling distributions of each of the three estimators — cross-sectional, before-and-after, and difference-in-differences — are normal. (The proof of this proposition is beyond the scope of this class, and is omitted here.) Specifically, we require that the assignment of treatment be perfectly random, and that individuals in both the treatment and control groups be independent from one another on the variables of interest. In the applications that follow, we shall assume that these two conditions are satisfied. You are allowed to make the same assumption on Problem Set 4.

Throughout, we will use a simple fictitious example to illustrate how to code up these estimations in **R**. Let's return to our example of students taking a test. Suppose that the class is subjected to two difficult tests in close succession. Suppose that the scores for the first test are distributed thusly: $\mathcal{N}(78, 36)$. Suppose the scores for the second test show a noticeable improvement: $\mathcal{N}(83, 25)$. Suppose, finally, that the professor, whether out of scientific cruelty or harmless curiosity, decides to see whether good luck charms can help her students improve their performance. After the

first test, she randomly distributes the charms to exactly half of all students. Now, as a matter of fact, the charms have no effect on performance, whether psychologically as placebos or supernaturally. The professor, not knowing this and too lazy to survey the whole class, collects at random the grades of twenty students whom she gave the charms for both tests, and the grades of twenty students who were left charm-less, again for both tests. Let's code up this scenario:

```
> ## Create the distributions for the two tests
> class.test1 <- rnorm(500, mean=78, sd=6)
> class.test2 <- rnorm(500, mean=83, sd=5)
>
> ## Create a numerical index of all students:
> class.index <- 1:500
>
> ## Randomly select students from the registry
> ## who will receive the charms:
> treatment <- sample(class.index, size=250, replace=F)
>
> ## No further manipulation necessary, as we stipulate
> ## that the treatment cannot cause any difference
> ## in the performance of the students. So, on average,
> ## the performance of both groups, at a point in time,
> ## should be indistinguishable.
>
> ## Draw a sample of 20 students from the treatment group,
> ## and 20 from the control group:
> sample.treated <- sample(treatment, size=20, replace=F)
> sample.control <- sample(class.index[-treatment],
+                          size=20, replace=F)
```

We are now ready to begin.

Cross-Sectional Comparsions

The purpose of cross-sectional comparisons of two groups is to determine whether there is any statistically significant difference between two groups — usually, the treatment and control groups — on some outcome variable of interest. For example, if we select a random sample of patients; if we, further, randomly administer a treatment — for example, a potent experimental drug — to only some of the patients in our sample; and then observe a numerical difference between the means of the

treated and control patients on a certain general index of health, can we conclude that the observed difference is due to our drug? Not necessarily, as it may have been the artefact of random differences between the treated and control patients. To exclude this possibility, we want to put uncertainty bounds on our cross-sectional comparison: that is to say, we should like to calculate a confidence interval for our estimate and examine whether that interval includes the value of 0. If it does not, then we can conclude that, at least at the level of our interval, the observed difference between the two groups is statistically significant.

Recall that the Sample estimate for the Average Treatment Effect (SATE) is given by the difference of means between the treatment and control groups in your sample. Let T_i denote the observed outcome for the i^{th} individual in the *treatment* group, and let C_i denote the observed outcome for the i^{th} individual in the *control* group. Let n_t be the size of the treatment group, and n_c the size of the control group. Then, the familiar formula for the SATE is given by:

$$\text{SATE} \quad = \quad \frac{1}{n_t}\sum_{i=1}^{n_t} T_i - \frac{1}{n_c}\sum_{i=1}^{n_c} C_i,$$

It is possible to show, although we will not discuss the proof of this proposition here, that the SATE is a consistent estimate for the Population-level ATE (PATE): SATE $\xrightarrow{p}$ PATE. So, the mean of SATE's sampling disstribution is exactly the PATE, the quantity we are interested in estimating. It is, therefore, a good point-estimate for the average treatment effect. The same asymptotic behaviour will also hold for the other two estimators we shall consider below, before-and-after and difference-in-differences: their sample estimates will provide consistent approximations to the population-level effects of interest.

It is quite easy to conduct the cross-sectional comparison on our fictitious example of the curious professor randomly distributing good luck charms to students in a (scientific?) effort to see whether the charms would boost their performance. Remember that the SATE is calculated by comparing the means of the treatment and control groups only after the treatment had been administered:

```
> treatment.after <- class.test2[sample.treated]
> control.after <- class.test2[sample.control]
>
> ## Calculate the SATE
> sate <- mean(treatment.after) - mean(control.after)
> sate

[1] -0.6736
```

The difference is slight, but can we be sure that it is due entirely to random variations in the students' performance rather than the effect of the charms? Now, in fact, we know that this difference is caused entirely by random variation, as we've defined the data to ensure that the treatment induces no difference between the two groups. But how could we prove this without recourse to population-level data?

For this, you first need to choose a significance level and then calculate the confidence interval corresponding to that significance level. It would then be easy to check if the value of 0, signifying the null effect of the treatment, is encompassed by that interval. One necessary prerequisite for this analysis is to obtain the standard error of the SATE. We wil derive a general expression for the standard error of this estimator. Begin by calculating the variance of the sampling distribution of the SATE, and be sure to replace the unknown population-level variances with their sample analogues:

$$\mathbb{V}(\text{SATE}) \;=\; \mathbb{V}\Big(\frac{1}{n_t}\sum_{i=1}^{n_t} T_i - \frac{1}{n_c}\sum_{i=1}^{n_c} C_i\Big)$$

To move past this point, you need to consider whether the values on the outcome variable across the treated and control individuals are statistically independent from one another. The answer is affirmative if we assume that the assignment of treatment was perfectly random. (If the treatment was assigned non-randomly, then we should be able to discover certain characteristics of the sampled individuals that allow us to predict who received the treatment, and so introduce the element of statistical dependency between them.) We are given that the treatment was administered randomly. Therefore:

$$\begin{aligned}
\mathbb{V}(\text{SATE}) &= \mathbb{V}\Big(\frac{1}{n_t}\sum_{i=1}^{n_t} T_i - \frac{1}{n_c}\sum_{i=1}^{n_c} C_i\Big) \\
&= \mathbb{V}\Big(\frac{1}{n_t}\sum_{i=1}^{n_t} T_i\Big) + \mathbb{V}\Big(\frac{1}{n_c}\sum_{i=1}^{n_c} C_i\Big) \\
&= \Big(\frac{1}{n_t}\Big)^2 \mathbb{V}\Big(\sum_{i=1}^{n_t} T_i\Big) + \Big(\frac{1}{n_c}\Big)^2 \mathbb{V}\Big(\sum_{i=1}^{n_c} C_i\Big) \\
&= \frac{1}{n_t^2}\mathbb{V}\Big(\sum_{i=1}^{n_t} T_i\Big) + \frac{1}{n_c^2}\mathbb{V}\Big(\sum_{i=1}^{n_c} C_i\Big)
\end{aligned}$$

Before we expand this expression further, you need to consider whether, whithin each of the treatment and control groups, values on the outcome variable across the

sampled individuals are independent. If randomization was properly accomplished, we should have independence not simply across the treatment and control groups, but also within each group. Therefore, we further have:

$$\begin{aligned}
\mathbb{V}(\text{SATE}) &= \frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_t} T_i\big) + \frac{1}{n_c^2}\mathbb{V}\big(\sum_{i=1}^{n_c} C_i\big) \\
&= \frac{1}{n_t^2}\mathbb{V}(T_1 + T_2 + \cdots + T_{n_t}) + \frac{1}{n_c^2}\mathbb{V}(C_1 + C_2 + \cdots + C_{n_c}) \\
&= \frac{1}{n_t^2}\big(\underbrace{\mathbb{V}(T_1)}_{=\sigma_t^2} + \underbrace{\mathbb{V}(T_2)}_{=\sigma_t^2} + \cdots + \underbrace{\mathbb{V}(T_{n_t})}_{=\sigma_t^2}\big) + \frac{1}{n_c^2}\big(\underbrace{\mathbb{V}(C_1)}_{=\sigma_c^2} + \underbrace{\mathbb{V}(C_2)}_{=\sigma_c^2} + \cdots + \underbrace{\mathbb{V}(C_{n_c})}_{=\sigma_c^2}\big),
\end{aligned}$$

where σ_t^2 and σ_c^2 are the population-level variances for the treatment and control groups, respectively. These are the variances we would observe had we randomly partitioned the entire population into treatment and control groups, administered the treatment to the former, and then measured their respective variances on the outcome variable.

$$\begin{aligned}
\mathbb{V}(\text{SATE}) &= \frac{1}{n_t^2}\big(\underbrace{\mathbb{V}(T_1)}_{=\sigma_t^2} + \underbrace{\mathbb{V}(T_2)}_{=\sigma_t^2} + \cdots + \underbrace{\mathbb{V}(T_{n_t})}_{=\sigma_t^2}\big) + \frac{1}{n_c^2}\big(\underbrace{\mathbb{V}(C_1)}_{=\sigma_c^2} + \underbrace{\mathbb{V}(C_2)}_{=\sigma_c^2} + \cdots + \underbrace{\mathbb{V}(C_{n_c})}_{=\sigma_c^2}\big) \\
&= \frac{1}{n_t^2}(\underbrace{\sigma_t^2 + \sigma_t^2 + \cdots + \sigma_t^2}_{\text{a total of } n_t\ \sigma_t^2\text{s here}}) + \frac{1}{n_c^2}(\underbrace{\sigma_c^2 + \sigma_c^2 + \cdots + \sigma_c^2}_{\text{a total of } n_c\ \sigma_c^2\text{s here}}) \\
&= \frac{1}{n_t^2} n_t \sigma_t^2 + \frac{1}{n_c^2} n_c \sigma_c^2 \\
&= \frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c}
\end{aligned}$$

The difficulty with this formula, which gives us the true variance of the sampling distribution of the SATE, is that we never observe the population-level variances, σ_t^2 and σ_c^2. We, therefore, use their sample-level analogues, $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$, which produces the following expression for the standard error of the SATE:

$$\begin{aligned}
\mathbb{V}(\text{SATE}) &= \frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c} \\
\widehat{\mathbb{V}(\text{SATE})} &= \frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_c^2}{n_c} \\
\text{se}_{\text{SATE}} &= \sqrt{\widehat{\mathbb{V}(\text{SATE})}} = \sqrt{\frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_c^2}{n_c}}
\end{aligned}$$

With the standard error in hand, we make use of the asymptotic normality of the sampling distribution of the SATE to construct confidence intervals, just as we did in the previous section. To continue with our example, let's construct a 95% confidence interval for the SATE of good luck charms:

```
> ## Calculate the standard error of the SATE
> se.sate <- sqrt(var(treatment.after)/length(treatment.after) +
+                   var(control.after)/length(control.after))
>
> ## Compute the lower and upper bounds of the
> ## 95% confidence interval for the SATE
> lower.sate <- sate - 1.96 * se.sate
> upper.sate <- sate + 1.96 * se.sate
>
> ## Report the confidence interval
> c(lower.sate, upper.sate)

[1] -2.872  1.525
```

Because the confidence interval includes the value of 0, we can conclude, with 95% confidence — or, more accurately, at the 5% significance level — that the effect of good luck charms on student performance in this particular class is null.

Before-and-After Comparisons

Before-and-after comparisons evaluate longitudinal changes in the outcome variable *for the same units*. Here, we are interested in calculating whether units within a single group — for example, either the treatment or the control group — undergo statistically significant changes through time. Suppose we record the values on the outcome variable for the treatment units before the treatment was administered, and then subsequently record those values immediately after administering the treatment, for the very same units. Let X_i^b denote the value on the outcome variable for observation i before the treatment is administered, and X_i^a denote the value on the outcome variable, for the very same observation, after the treatment is administered. Let n be the size of the group we are examining. It can be either the treatment or the control portions of our sample. We assume there is no attrition in the group, meaning that the number of observations in given group, whether the treated or control, remains the same after as it was before the treatment was administered. Then,

the sample-level Before-and-After (BA) estimate is given by:

$$\widehat{\text{BA}} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\left(X_i^a - X_i^b\right)$$

It is straightforward to calculate the sample BA estimate for our fictitious example:

```
> ## Record the test scores of the students in the
> ## treatment and control groups, before the
> ## charms were distributed:
> treatment.before <- class.test1[sample.treated]
> control.before <- class.test1[sample.control]
>
> ## Calculate the sample BA estimate separately for
> ## the treatment and control groups:
> sba.treated <- mean(treatment.after) - mean(treatment.before)
> sba.treated

[1] 4.902

>
> sba.control <- mean(control.after) - mean(control.before)
> sba.control

[1] 8.441
```

These are non-zero estimates. Do our sample-level before-and-after estimates reflect underlying population dynamics, or are they, rather, artefacts of random variation within the treatment and control group? Now, we know, from the way that we set up the data, that student performance did improve from the first to the second test. However, to answer this question without recourse to population-level data, we first stipulate, once again without proof, that the asymptotic sampling distribution of the BA estimator is normal, and that it is a consistnet estimator for the population-level BA effect.

How can we calculate its standard error? As in the case of the SATE, we begin

by calculating the variance of the sampling distribution of $\widehat{\text{BA}}$:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \mathbb{V}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(X_i^a - X_i^b\big)\Big) \\
&= \Big(\frac{1}{n}\Big)^2\mathbb{V}\Big(\sum_{i=1}^{n}\big(X_i^a - X_i^b\big)\Big) \\
&= \frac{1}{n^2}\mathbb{V}\Big(\sum_{i=1}^{n}\big(X_i^a - X_i^b\big)\Big) \\
&= \frac{1}{n^2}\mathbb{V}\Big(\big(X_1^a - X_1^b\big) + \big(X_2^a - X_2^b\big) + \cdots + \big(X_n^a - X_n^b\big)\Big)
\end{aligned}$$

Before we can proceed further, it is important to consider whether any dependencies exist between the values of the outcome variable across units within a single group of your sample. For example, do the units in the treatment group exhibit any dependencies, at a point in time? As discussed above, randomization ensures that no such dependencies occur. Therefore, we can proceed as usual:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \frac{1}{n^2}\mathbb{V}\Big(\big(X_1^a - X_1^b\big) + \big(X_2^a - X_2^b\big) + \cdots + \big(X_n^a - X_n^b\big)\Big) \\
&= \frac{1}{n^2}\Big(\mathbb{V}\big(X_1^a - X_1^b\big) + \mathbb{V}\big(X_2^a - X_2^b\big) + \cdots + \mathbb{V}\big(X_n^a - X_n^b\big)\Big)
\end{aligned}$$

What about the same unit considered at two different points in time? Unless we argue that the assignment of treatment wipes the slate clean for every unit, it is highly implausible to claim that the value of a unit on the outcome variable at one time is statistically independent from that value, for the very same unit, at some future time. Because of this dependency, it is necessary to incoporate a covariance component in the calculation of within-unit variances.

Let's begin with the general case. Suppose that you have two dependent random variables, X and Y. Then,

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{Cov}(X,Y) \tag{5}$$

$$\mathbb{V}(X-Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\text{Cov}(X,Y), \tag{6}$$

where $\text{Cov}(X,Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big]$, but can also be calculated thusly: $\text{Cov}(X,Y) = r(X,Y) \times \sigma_X\sigma_Y$, where $r(X,Y)$ is the correlation coefficient between X and Y.

Applying these formulas to our expression, we get:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \frac{1}{n^2}\Big(\mathbb{V}\big(X_1^a - X_1^b\big) + \mathbb{V}\big(X_2^a - X_2^b\big) + \cdots + \mathbb{V}\big(X_n^a - X_n^b\big)\Big) \\
&= \frac{1}{n^2}\big(\underbrace{\mathbb{V}(X_1^a)}_{\sigma_a^2} + \underbrace{\mathbb{V}(X_1^b)}_{\sigma_b^2} - 2\text{Cov}(X_1^a, X_1^b) + \cdots + \underbrace{\mathbb{V}(X_n^a)}_{\sigma_a^2} + \underbrace{\mathbb{V}(X_n^b)}_{\sigma_b^2} - 2\text{Cov}(X_n^a, X_n^b)\big),
\end{aligned}$$

where σ_a^2 and σ_b^2 are the population-level variances of the "after" and "before" groups; that is to say, had we randomly administered the treatment to the entire population, and then compared the treated units in the whole population after and before the treatment was administered, their variances would be given by σ_a^2 and σ_b^2, respectively. Notice that the population-level variance for all observations at a given time is constant. This is because randomization ensures that there are no differences between observations, either across the treatment and control groups, or within the groups, with respect to the administration of treatment.

The formula then further simplifies as follows:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \frac{1}{n^2}\big(\underbrace{\mathbb{V}(X_1^a)}_{\sigma_a^2} + \underbrace{\mathbb{V}(X_1^b)}_{\sigma_b^2} - 2\text{Cov}(X_1^a, X_1^b) + \cdots + \underbrace{\mathbb{V}(X_n^a)}_{\sigma_a^2} + \underbrace{\mathbb{V}(X_n^b)}_{\sigma_b^2} - 2\text{Cov}(X_n^a, X_n^b)\big) \\
&= \frac{1}{n^2}\big(n\sigma_a^2 + n\sigma_b^2 - 2\text{Cov}(X_1^a, X_1^b) - 2\text{Cov}(X_2^a, X_2^b) - \cdots - 2\text{Cov}(X_n^a, X_n^b)\big)
\end{aligned}$$

We must now pause once again to consider the covariance structure of our observations. We've already stipulated that there are no dependencies between observations within a group, at a point in time. Only temporal dependencies — that is, for the same observation considered at different points in time — are allowed. But how should we model these latter dependencies? One simple solution is to assume that the temporal dependency is identical across all the observations — that it to say, that the relationship between an observation's "before" and "after" values on the outcome variable remains constant across observations. Our formula then simplifies

as follows:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \frac{1}{n^2}\big(n\sigma_a^2 + n\sigma_b^2 \underbrace{-2\text{Cov}(X_1^a, X_1^b) - 2\text{Cov}(X_2^a, X_2^b) - \cdots - 2\text{Cov}(X_n^a, X_n^b)}_{\text{a total of n terms here}}\big) \\
&= \frac{1}{n^2}\big(n\sigma_a^2 + n\sigma_b^2 - n \times 2\text{Cov}(X_i^a, X_i^b)\big) \\
&= \frac{1}{n^2}\big(n\sigma_a^2 + n\sigma_b^2 - n \times 2\text{Cov}(X_i^a, X_i^b)\big) \\
&= \frac{1}{n^2}n\sigma_a^2 + \frac{1}{n^2}n\sigma_b^2 - \frac{1}{n^2}2n\text{Cov}(X_i^a, X_i^b) \\
&= \frac{\sigma_a^2}{n} + \frac{\sigma_b^2}{n} - \frac{2\text{Cov}(X_i^a, X_i^b)}{n},
\end{aligned}$$

which gives us the expression for the *exact* variance of the asymptotic sampling distribution of $\widehat{\text{BA}}$. Once more, we can't calculate this quantity precisely because we do not know the population-level σ_a^2, σ_b^2, or $\text{Cov}(X_i^a, X_i^b)$. Instead, we replace these three quantities with their sample analogues, $\hat{\sigma}_a^2, \hat{\sigma}_b^2$, and $\widehat{\text{Cov}(X_i^a, X_i^b)}$, and plug them into the formula to obtain the standard error of $\widehat{\text{BA}}$:

$$\begin{aligned}
\mathbb{V}(\widehat{\text{BA}}) &= \frac{\sigma_a^2}{n} + \frac{\sigma_b^2}{n} - \frac{2\text{Cov}(X_i^a, X_i^b)}{n} \\
\widehat{\mathbb{V}(\widehat{\text{BA}})} &= \frac{\hat{\sigma}_a^2}{n} + \frac{\hat{\sigma}_b^2}{n} - \frac{2\widehat{\text{Cov}(X_i^a, X_i^b)}}{n} \\
\text{se}_{\widehat{\text{BA}}} &= \sqrt{\widehat{\mathbb{V}(\widehat{\text{BA}})}} = \sqrt{\frac{\hat{\sigma}_a^2}{n} + \frac{\hat{\sigma}_b^2}{n} - \frac{2\widehat{\text{Cov}(X_i^a, X_i^b)}}{n}}
\end{aligned}$$

With the standard error in hand, we make use of the asymptotic normality of the sampling distribution of the SATE to construct confidence intervals, just as we did in the previous section. Let's perform these calculations for our fictitious example:

```
> ## Specify the size of the treated sample:
> n.t <- length(treatment.after)
>
> ## Specify the size of the control sample:
> n.c <- length(control.after)
>
> ## Calculate the standard errors for the BA
> ## estimate of the treatment group
> se.sba.treated <- sqrt(var(treatment.after)/n.t +
```

```
+                          var(treatment.before)/n.t -
+                          2 / n.t * cov(treatment.after,
+                                        treatment.before))
>
> ## Calculate the lower and upper bounds of the
> ## 95% confidence interval for the SBA of the
> ## treatment group:
> lower.sba.treated <- sba.treated - 1.96 * se.sba.treated
> upper.sba.treated <- sba.treated + 1.96 * se.sba.treated
>
> ## Calculate the standard errors for the BA
> ## estimate of the control group
> se.sba.control <- sqrt(var(control.after)/n.c +
+                          var(control.before)/n.c -
+                          2 / n.c * cov(control.after,
+                                        control.before))
>
> ## Calculate the lower and upper bounds of the
> ## 95% confidence interval for the SBA of the
> ## treatment group:
> lower.sba.control <- sba.control - 1.96 * se.sba.control
> upper.sba.control <- sba.control + 1.96 * se.sba.control
>
> ## Report the confidence intervals for the treatment
> ## and control groups:
> c(lower.sba.treated, upper.sba.treated)

[1] 1.374 8.429

> c(lower.sba.control, upper.sba.control)

[1]  5.996 10.885
```

Neither interval contains the value of 0, which tells us that we can conclude, at the 5% significance level, that the improvement in scores from the first to the second test, across both the treatment and control groups, analyzed separately, is not due to random variation but underlying population dynamics — for example, the entire class studying harder for the second, as opposed to the first, test.

Difference-in-Differences

The asymptotic behaviour of the familiar Difference-in-Differences (DiD) estimator should present few difficulties after the foregoing discussions of the cross-sectional and before-and-after estimators. Combining the notation from the previous two subsections, let T_i^b denote the value on the outcome variable for the i^{th} observation in the treatment group, before the treatment is administered; T_i^a denote the value of the very same observation, but after the treatment is administered; C_i^b denote the value on the outcome variable for the i^{th} observation in the control group, before the treatment is administered; and C_i^a the value of the very same observation, but after the treatment is administered. Let n_t be the number of treated observations, and n_c the number of control observations. As before, we assume there is no attrition. Then, for our sample, the DiD estimate is given by:

$$\widehat{\text{DiD}} \quad = \quad \frac{1}{n_t}\sum_{i=1}^{n_t}\left(T_i^a - T_i^b\right) - \frac{1}{n_c}\sum_{i=1}^{n_c}\left(C_i^a - C_i^b\right)$$

It is easy to calculate the DiD estimate for our fictitious example:

```
> DiD <- mean(treatment.after) - mean(treatment.before) -
+   (mean(control.after) - mean(control.before))
> DiD

[1] -3.539
```

We state it without proof that $\widehat{\text{DiD}}$ is a consistent estimate for its population-level analogue, and that its asymptotic sampling distribution is normal. The tricky part is, once more, deriving its standard error. As before, we begin by solving for the variance of its sampling distribution:

$$\mathbb{V}\left(\widehat{\text{DiD}}\right) \quad = \quad \mathbb{V}\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\left(T_i^a - T_i^b\right) - \frac{1}{n_c}\sum_{i=1}^{n_c}\left(C_i^a - C_i^b\right)\right)$$

We pause to consider whether the treatment and control groups are statistically independent. As discussed above, this is guaranteed by the randomization assumption.

Therefore,

$$\begin{aligned}
\mathbb{V}\big(\widehat{\text{DiD}}\big) &= \mathbb{V}\Big(\frac{1}{n_t}\sum_{i=1}^{n_t}\big(T_i^a - T_i^b\big) - \frac{1}{n_c}\sum_{i=1}^{n_c}\big(C_i^a - C_i^b\big)\Big) \\
&= \mathbb{V}\Big(\frac{1}{n_t}\sum_{i=1}^{n_t}\big(T_i^a - T_i^b\big)\Big) + \mathbb{V}\Big(\frac{1}{n_c}\sum_{i=1}^{n_c}\big(C_i^a - C_i^b\big)\Big) \\
&= \Big(\frac{1}{n_t}\Big)^2\mathbb{V}\big(\sum_{i=1}^{n_t}(T_i^a - T_i^b)\big) + \Big(\frac{1}{n_t}\Big)^2\mathbb{V}\big(\sum_{i=1}^{n_c}(C_i^a - C_i^b)\big) \\
&= \underbrace{\frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_t}(T_i^a - T_i^b)\big)}_{=\mathbb{V}(\widehat{\text{BA}}_t)} + \underbrace{\frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_c}(C_i^a - C_i^b)\big)}_{=\mathbb{V}(\widehat{\text{BA}}_c)}
\end{aligned}$$

This fork in the road should look familiar. We've already computed the exact variance of the sampling distribution of the BA estimator, and can now simply plug in that solution, separately, for the treatment and control groups:

$$\frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_t}(T_i^a - T_i^b)\big) = \frac{\sigma_{t,a}^2}{n_t} + \frac{\sigma_{b,t}^2}{n_t} - \frac{2\text{Cov}(T_i^a, T_i^b)}{n_t},$$

where, once again, $\sigma_{t,a}^2, \sigma_{t,b}^2$, and $\text{Cov}(T_i^a, T_i^b)$ are the population-level variances of the treatment group after and before the administration of the treatment, and the covariance of the treatment group with itself at those two points in time. To repeat, these are the quantities we would be able to compute if we were to parse the whole population into treatment and control groups, record the values of the treatment group before the treatment was administered as well as afterward, and calculate the resulting covariance. The expression for the control group is similar:

$$\frac{1}{n_c^2}\mathbb{V}\big(\sum_{i=1}^{n_t}(C_i^a - C_i^b)\big) = \frac{\sigma_{c,a}^2}{n_c} + \frac{\sigma_{b,c}^2}{n_c} - \frac{2\text{Cov}(C_i^a, C_i^b)}{n_c}$$

Plugging these into our solution for the variance of the sampling distribution of $\widehat{\text{DiD}}$, we get:

$$\begin{aligned}
\mathbb{V}\big(\widehat{\text{DiD}}\big) &= \frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_t}(T_i^a - T_i^b)\big) + \frac{1}{n_t^2}\mathbb{V}\big(\sum_{i=1}^{n_c}(C_i^a - C_i^b)\big) \\
&= \frac{\sigma_{t,a}^2}{n_t} + \frac{\sigma_{b,t}^2}{n_t} - \frac{2\text{Cov}(T_i^a, T_i^b)}{n_t} + \frac{\sigma_{c,a}^2}{n_c} + \frac{\sigma_{b,c}^2}{n_c} - \frac{2\text{Cov}(C_i^a, C_i^b)}{n_c}
\end{aligned}$$

This expression gives us the exact variance of the sampling distribution of the DiD estimator. Because we don't know the population-level variance and covariance terms in the expression, we replace them with their sample-level analogues and get, instead, the expression for the standard error of $\widehat{\text{DiD}}$:

$$
\begin{aligned}
\mathbb{V}\big(\widehat{\text{DiD}}\big) &= \frac{\sigma^2_{t,a}}{n_t} + \frac{\sigma^2_{b,t}}{n_t} - \frac{2\text{Cov}(T^a_i, T^b_i)}{n_t} + \frac{\sigma^2_{c,a}}{n_c} + \frac{\sigma^2_{b,c}}{n_c} - \frac{2\text{Cov}(C^a_i, C^b_i)}{n_c} \\
\widehat{\mathbb{V}\big(\widehat{\text{DiD}}\big)} &= \frac{\hat{\sigma}^2_{t,a}}{n_t} + \frac{\hat{\sigma}^2_{b,t}}{n_t} - \frac{2\widehat{\text{Cov}(T^a_i, T^b_i)}}{n_t} + \frac{\hat{\sigma}^2_{c,a}}{n_c} + \frac{\hat{\sigma}^2_{b,c}}{n_c} - \frac{2\widehat{\text{Cov}(C^a_i, C^b_i)}}{n_c} \\
\text{se}_{\widehat{\text{DiD}}} &= \sqrt{\widehat{\mathbb{V}\big(\widehat{\text{DiD}}\big)}} \\
&= \sqrt{\frac{\hat{\sigma}^2_{t,a}}{n_t} + \frac{\hat{\sigma}^2_{b,t}}{n_t} - \frac{2\widehat{\text{Cov}(T^a_i, T^b_i)}}{n_t} + \frac{\hat{\sigma}^2_{c,a}}{n_c} + \frac{\hat{\sigma}^2_{b,c}}{n_c} - \frac{2\widehat{\text{Cov}(C^a_i, C^b_i)}}{n_c}}
\end{aligned}
$$

The final expression for the standard error may look imposing, but it is actually quite easy to calculate if you have already calculated the BA standard errors for the treatment and control groups. Indeed, note that

$$
\begin{aligned}
\widehat{\mathbb{V}\big(\widehat{\text{DiD}}\big)} &= \underbrace{\frac{\hat{\sigma}^2_{t,a}}{n_t} + \frac{\hat{\sigma}^2_{b,t}}{n_t} - \frac{2\widehat{\text{Cov}(T^a_i, T^b_i)}}{n_t}}_{=\widehat{\mathbb{V}(\text{BA}_t)}} + \underbrace{\frac{\hat{\sigma}^2_{c,a}}{n_c} + \frac{\hat{\sigma}^2_{b,c}}{n_c} - \frac{2\widehat{\text{Cov}(C^a_i, C^b_i)}}{n_c}}_{=\widehat{\mathbb{V}(\text{BA}_c)}} \\
&= \widehat{\mathbb{V}(\text{BA}_t)} + \widehat{\mathbb{V}(\text{BA}_c)} \\
&= \text{se}^2_{\widehat{\text{BA}_t}} + \text{se}^2_{\widehat{\text{BA}_c}} \\
\text{se}_{\widehat{\text{DiD}}} &= \sqrt{\widehat{\mathbb{V}\big(\widehat{\text{DiD}}\big)}} = \sqrt{\text{se}^2_{\widehat{\text{BA}_t}} + \text{se}^2_{\widehat{\text{BA}_c}}}
\end{aligned}
$$

To illustrate this final point, we calculate the DiD estimate and the corresponding 95% confidence interval using our prior estimates of the standard errors for the BA comparisons:

```
> se.did <- sqrt(se.sba.treated^2 + se.sba.control^2)
>
> ## Calculate the lower and upper bounds of the 95%
> ## confidence interval:
> lower.did <- DiD - 1.96 * se.did
> upper.did <- DiD + 1.96 * se.did
```

```
>
> ## Report the confidence interval
> c(lower.did, upper.did)

[1] -7.8307  0.7528
```

Combining the results of the cross-sectional, before-and-after, and difference-in-differences, we can conclude, at the 5% significance level: that the causal effect of good luck charms on student performance in our fictitious class is altogether null (this was demonstrated by the cross-sectional analysis); that student performance improved from the first to the second test (this was demonstrated by the before-and-after analysis); and that this improvement was not due to the charms (this was demonstrated by the difference-in-differences analysis).

6 General Instructions

I need to calculate asymptotic confidence intervals for an estimate. How should I proceed?

1. **Calculate your point-estimate**: apply the estimator you're given, whether it be the mean, SATE, BA, or DiD, to your sample. The resulting number is your best guess for what the true effect looks like at the level of the population from which your sample was drawn.

    - This is subject to the three assumptions specified on pp. 1-2 of this booklet.

2. **Calculate the standard error of your estimator**: this step requires identifying the sampling distribution of your estimator and calculating its variance. After obtaining an expression for the variance, replace all population-level terms with their sample-level analogues. This will give you the standard error. The formulas for the standard errors of four popular estimators are given below:

    - **Sample mean**: $\frac{\hat{\sigma}}{\sqrt{n}}$
    - **Cross-sectional**: $\sqrt{\frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_c^2}{n_c}}$
    - **Before-and-after**: $\sqrt{\frac{\hat{\sigma}_a^2}{n} + \frac{\hat{\sigma}_b^2}{n} - \frac{2\widehat{\text{Cov}(X_i^a, X_i^b)}}{n}}$
    - **Difference-in-differences**: $\sqrt{\widehat{\text{se}^2_{\text{BA}_t}} + \widehat{\text{se}^2_{\text{BA}_c}}}$

3. **Select the desired significance level**, α: for a confidence interval of $n\%$, set $\alpha = (100 - n)/100$. For example, for the 95% confidence interval, set $\alpha = (100 - 95)/100 = 5/100 = 0.05$.

4. **Calculate the margin of error**: for asymptotic confidence intervals — that is, confidence intervals computed by approximating the sampling distribution of your estimator to the standard normal distribution — the margin of error is always given by the following formula: $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \times$ `standard error`

    - Use `qnorm` to calculate Φ^{-1} in **R**. For example, for a 95% confidence interval, $\alpha = 0.05$, and the critical value can be found by inputting `qnorm(1-0.05/2)` or, equivalently, `qnorm(0.975)`

6. **Calculate the confidence interval**: the lower bound of the confidence interval is given by `Point Estimate - Margin of Error`, and the upper bound by `Point Estimate` + `Margin of Error`.

7. **Intepret the confidence interval**: across (infinitely) many samples drawn from a population of interest, if we were to recalculate the confidence interval for each one of those samples, the true parameter being estimated will be encompassed by nearly $(1-\alpha)\times 100\%$ of those confidence intervals.

8. **Draw conclusions regarding statistical significance, if desired**: if the problem asks you to estimate a causal effect, note whether the confidence interval contains the value of 0 (null effect). If it does not, conclude that, at the significance level of α, the causal effect is due to underlying population dynamics and not random variation in your particular sample. If it does, conclude that, at the significance level of α, the causal effect is statistically indistinguishable from random variation that we would expect to see by chance.