1 Confidence intervals – practice and examples

Much like questions about the normal distribution, there are only really two things we can do with confidence intervals: construct an interval corresponding to a given level of confidence, or determine various parameters which define a supplied confidence interval (the level of confidence fixing other model parameters, or the sample size fixing the desired confidence, etc.). The first example here is worked through pretty thoroughly to help build understanding; the second is meant to be more of a direct approach (with far less discussion).

For more practice, the book has some decent questions in section 4.1 (as well as some which seem to be fairly irrelevant for the course). You can also Google some example problems, but I haven't been able to find any excellent question repositories as yet; and keep in mind that questions on the Internet are often dubious.

Question 1: a TA wants to move his office hour from 3:00pm Wednesday to 9:30am Wednesday. He emails the class to see how well the change would work for the students. 5 students reply; 3 say the change is better for them, and 2 say it is worse (due to a conflict with Econ 11).

(a) *What is a "reasonable" statistical setup of this question?*

When we are concerned with the central limit theorem and its applications (such as confidence intervals), we don't need to make assumptions about the statistical form of our problem, or the distributions of any of the random variables it contains. This is because the central limit theorem states that

$$\sqrt{N}\left(\frac{\overline{X}_N - E(X)}{\sigma}\right) \sim_{\text{appx}} N(0,1)$$

for N sufficiently large, *regardless of the distribution of* X.

However, in the name of understanding the question its useful to think about what is being asked. In this case, we appear to have a set of $N = 5$ observations of a Bernoulli random experiment; we'll use the convention that the outcome of the experiment is 1 if the time change is good, and 0 if it is bad. This tells us that when we talk about a confidence interval for the mean, we are talking about a range of values which may contain the population mean with some probability; here, the population mean is the probability that the change is good for a randomly-drawn student.

(b) *Is it acceptable to use a normal approximation for this question?*

The short answer is, "no." We give two justifications:

- As in class, we claim that the normal approximation leading to a confidence interval is valid so long as $N \geq 30$. Here, $N = 5$, so the approximation is not good.
- Consider what we discussed about approximating discrete distributions by continuous distributions: we said that the approximation was only justified if $Np \geq 5$ and $N(1-p) \geq 5$. If we assume that the population mean is close to the sample mean – which is not necessarily a good assumption – we can see that these conditions do not hold.

Regardless, the normal approximation in this question is not mathematically or statistically robust. However, we are going to apply it anyway (the point here is to learn the standard methods for confidence analysis, rather than fall victim to errors in the particular model; as a tip on pedagogy, it would have been good to present an example with better properties, but respecting reality was paramount).

(c) *Assume the population variance is* $\sigma^2 = 0.24$*; using a normal approximation, construct a* 95% *confidence interval for* μ*, the population mean.*

We will attack this problem in two ways. First, we will take the direct approach; we look for an interval which is symmetric about the sample mean such that μ lies in this interval with the specified

probability (95%). That is,

$$\Pr\left(\mu \in [\overline{X} - E, \overline{X} + E]\right) = 95\%$$

Notice that this can be restated and rearranged algebraically as

$$\begin{aligned}
\Pr\left(\mu \in [\overline{X} - E, \overline{X} + E]\right) &= \Pr\left(\overline{X} - E \leq \mu \leq \overline{X} + E\right) \\
&= \Pr\left(-E \leq \mu - \overline{X} \leq E\right) \\
&= \Pr\left(\left|\mu - \overline{X}\right| \leq E\right) \\
&= \Pr\left(\left|\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma}\right| \leq \frac{\sqrt{N}E}{\sigma}\right)
\end{aligned}$$

The central limit theorem tell us that $|\frac{\sqrt{N}(\mu-\overline{X})}{\sigma}| \sim_{\text{appx}} N(0,1)$. So how can we determine E? Following statistical convention, we let α be such that $0.95 = 1-\alpha$ (less obliquely, $\alpha = 0.05$ in this case). Letting $Z = \frac{\sqrt{N}(\mu-\overline{X})}{\sigma}$, we know by the rule of complementation that

$$\Pr\left(|Z| \leq \frac{\sqrt{N}E}{\sigma}\right) = 1 - \left(\Pr\left(Z < -\frac{\sqrt{N}E}{\sigma}\right) + \Pr\left(Z > \frac{\sqrt{N}E}{\sigma}\right)\right)$$

But since the distribution of Z is symmetric about 0, we also know

$$\Pr\left(Z < -\frac{\sqrt{N}E}{\sigma}\right) = \Pr\left(Z > \frac{\sqrt{N}E}{\sigma}\right)$$

We can put these facts together to see

$$\Pr\left(|Z| \leq \frac{\sqrt{N}E}{\sigma}\right) = 1 - 2\Pr\left(Z > \frac{\sqrt{N}E}{\sigma}\right)$$

Since we know that we're looking for a $100(1-\alpha)\%$ confidence interval (think about why we threw the 100 multiple in front of that expression), we want

$$\begin{aligned}
& & 1 - \alpha &= 1 - 2\Pr\left(Z > \frac{\sqrt{N}E}{\sigma}\right) \\
\iff & & \alpha &= 2\Pr\left(Z > \frac{\sqrt{N}E}{\sigma}\right) \\
\iff & & \frac{\alpha}{2} &= \Pr\left(Z < -\frac{\sqrt{N}E}{\sigma}\right) \\
\iff & & \frac{\alpha}{2} &= \Phi\left(-\frac{\sqrt{N}E}{\sigma}\right) \qquad (1) \\
\iff & & -\Phi^{-1}\left(\frac{\alpha}{2}\right) &= \frac{\sqrt{N}E}{\sigma} \qquad (2)
\end{aligned}$$

Now we can begin actually solving. Since $1-\alpha = 0.95$, $\Phi^{-1}(\frac{\alpha}{2}) = 1.96$ – this can be obtained by table lookup. We have enough algebraic power now to solve

$$1.96 = \frac{\sqrt{N}E}{\sigma}$$

Since $N = 5$ and $\sigma^2 = 0.24$, we plug and chug and get

$$E = 0.4294$$

Returning to our claim that $[\overline{X} - E, \overline{X} + E]$ was a 95% confidence interval for μ, we now only need to determine $\overline{X}$. By the implicit setup in this experiment, we saw 3 "successes" out of 5 runs of the experiment, so our sample mean is $\overline{X} = 0.6$. Then we find that

$$[0.6 - 0.4294, 0.6 + 0.4294] = [0.1706, 1.0294]$$

is a 95% confidence interval for μ.

Having solved this problem the "mathier" way from first principles, we can discuss solving this using the shortcut method described in lecture. From the formulas in the lecture slides, we know that a $100(1-\alpha)\%$ confidence interval for μ can be constructed as

$$\left[\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}}, \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right]$$

From our discussion of the long way to solve the question, it's apparent that we're making a direct connection here to $E = \frac{z_{\alpha/2}\sigma}{\sqrt{N}}$. That is, if $z_{\alpha/2}$ describes the number of standard deviations away from the mean our confidence interval must contain, the error term will take this form. The only remaining question regards what value we should substitute in for $z_{\alpha/2}$.

The notation $z_{\alpha/2}$ is intentionally suggestive: the z regards the fact that we're concerned with a standard normal, and the $\frac{\alpha}{2}$ connects with our earlier discussion culminating in equation (2). That is, since $1 - \alpha$ gives us the probability contained within the central portion of the distribution – defining our confidence interval – $\frac{\alpha}{2}$ gives us the probability of being in either one of the tails. Then we can look up $z_{\alpha/2}$ in a z-table, if it exists, or we can use (2) to see

$$z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$$

With this, we again have sufficient information to construct our confidence interval. Plugging in for all of our known values, we find

$$\left[0.6 - \frac{1.96\left(\sqrt{0.24}\right)}{\sqrt{5}}, 0.6 + \frac{1.96\left(\sqrt{0.24}\right)}{\sqrt{5}}\right] = [0.1706, 1.0294]$$

is a 95% confidence interval for μ.

The real advantage of this shortcut method is, frankly, that it's much shorter. We were able to use the general formula and knowledge about where $z_{\alpha/2}$ comes from to build a confidence interval quickly, with the exact same result. On a test, using the quicker method is definitely the right way to go.

(d) *Does the confidence interval we came up with make sense?*

This is a brief aside. We know that μ, the population mean, is a parameter measuring the probability that changing the section time is good for any particular student. With the upper bound of our confidence interval being 1.0294, we've included a set of values in our confidence interval *which μ can never equal*. This is a little counterintuitive.

What's going on here is that, as mentioned in part (b), the normal approximation isn't all that great in this case, but we used it anyway. If you look at the equations, you can see that by raising N we will tighten the bounds of our confidence interval; so if we'd had a more reasonable sample size we wouldn't run into this problem in this particular question.

However, we could counter the effects of a larger sample size by increasing the desired confidence. Since the normal distribution contains positive probability mass out to infinity (on both sides), we can always choose a confidence level large enough so that we'd have "funky" values in our confidence interval. This is a function of the exact same trouble as before: the normal approximation (and the

central limit theorem) are not precise until $N \to \infty$. With any finite sample, we're going to have issues with accuracy. We just claim that you can generally ignore them with $N \geq 30$.

To completely beat the point into the ground, think about what happens in this question if $N = 30$ (and 18 students think the change is good while 12 think it is bad) and we increase the desired confidence level to 99.99993%. Of course, you can't look this value up in the book – although your calculator might be able to do it – so you can assume that $z_{\frac{1-0.9999993}{2}} = 4.8250$. What's key in using this as a counterexample is that we had to cook up an objectively ridiculous confidence level to break things; with the usual 95% or 99% confidence levels we'd still be okay.

(e) *With what confidence can we say $\mu \in [0.5, 0.7]$?*

The trick here is to notice that we can invert our usual application of the confidence interval formula,

$$\mu \in \left[\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}}, \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}\right]$$

In this case, since we have $\overline{X} = 0.1$, we know

$$[0.5, 0.7] = [0.6 - 0.1, 0.6 + 0.1]$$

Appealing to the the standard form of a confidence interval, we know then that

$$\frac{z_{\alpha/2}\sigma}{\sqrt{N}} = 0.1$$

With $\sigma^2 = 0.24$ and $N = 5$, this tells us $z_{\alpha/2} = 0.4564$. From equation (1), and the fact that the standard form of the error term is $E = \frac{z_{\alpha/2}\sigma}{\sqrt{N}}$ we see

$$\frac{\alpha}{2} = \Phi\left(-\frac{\sqrt{N}E}{\sigma}\right) = \Phi\left(-\frac{\sqrt{N}\frac{z_{\alpha/2}\sigma}{\sqrt{N}}}{\sigma}\right) = \Phi\left(-z_{\alpha/2}\right)$$

Then $\frac{\alpha}{2} = 0.3240$. Since this construction yields a $100(1-\alpha)\%$ confidence interval for μ, we see that

$$1 - \alpha = 0.3519$$

and so $[0.5, 0.7]$ is a 35.19% confidence interval for μ.

Notice that we went directly through this question using the known equation for confidence intervals. We could have done this more indirectly by moving $\Pr(\cdot)$ around and playing with it that way, but it would have taken significantly longer to obtain the same result. Using the known form is a real time-saver.

(f) *With what confidence can we say $\mu \geq 0.5$?*

This answer follows pretty directly from part (e). We know that

$$\Pr\left(\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma} < -0.4564\right) + \Pr\left(\left|\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma}\right| \leq 0.4564\right) + \Pr\left(\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma} > 0.4564\right)$$

where we use ± 0.4564 from our obtained $z_{\alpha/2}$ value from part (e). By symmetry about the mean, we know

$$\Pr\left(\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma} < -0.4564\right) = \Pr\left(\frac{\sqrt{N}\left(\mu - \overline{X}\right)}{\sigma} > 0.4564\right)$$

By algebra and the solution to (e) we then have

$$\Pr\left(\left|\frac{\sqrt{N}\left(\mu-\overline{X}\right)}{\sigma}\right|\leq 0.4564\right)+\Pr\left(\frac{\sqrt{N}\left(\mu-\overline{X}\right)}{\sigma}>0.4564\right)=0.6760$$

We can rearrange the terms in our probabilities to see

$$\begin{aligned}
&\Pr\left(\left|\frac{\sqrt{N}\left(\mu-\overline{X}\right)}{\sigma}\right|\leq 0.4564\right)+\Pr\left(\frac{\sqrt{N}\left(\mu-\overline{X}\right)}{\sigma}>0.4564\right)\\
&=\Pr\left(\left|\mu-\overline{X}\right|\leq(0.4564)\frac{\sigma}{\sqrt{N}}\right)+\Pr\left(\mu-\overline{X}>(0.4564)\frac{\sigma}{\sqrt{N}}\right)\\
&=\Pr\left(-(0.4564)\frac{\sigma}{\sqrt{N}}\leq\mu-\overline{X}\leq(0.4564)\frac{\sigma}{\sqrt{N}}\right)+\Pr\left(\mu-\overline{X}>(0.4564)\frac{\sigma}{\sqrt{N}}\right)\\
&=\Pr\left(-(0.4564)\frac{\sigma}{\sqrt{N}}\leq\mu-\overline{X}\right)\\
&=\Pr\left(-(0.4564)\frac{\sigma}{\sqrt{N}}+\overline{X}\leq\mu\right)\\
&=0.6760
\end{aligned}$$

From part (e), we know that $-(0.4564)\frac{\sigma}{\sqrt{N}}+\overline{X}=0.5$. Then we see directly that $[0.5,+\infty)$ is a 67.6% confidence interval for μ.

We could also have computed this directly from part (e): if $[0.5,0.7]$ is a 35.19% confidence interval for μ which is symmetric about the sample mean, we know that either tail is such that there is $\frac{1-0.3519}{2}=0.3241$ probability that the population mean lies within that particular tail. Then the confidence of the interval $[0.5,+\infty]$ is simply $0.3519+0.3241=0.6760$, implying a 67.6% confidence interval.

The confidences reflect, in the guise of the question, the probability that switching office hour times is a good idea (that is, the switch is good for any random student with probability 0.5 or above). This generalizes our notion of confidence intervals slightly: where we were previously discussing only confidence intervals which were symmetric about the sample mean – two-sided confidence intervals – we see now that we can drop one bound to obtain a one-sided confidence interval for the population mean, giving us the confidence that the true value of the population mean is greater than (or less than, had we flipped the question around) a particular value.

A more generic form for this one-sided confidence interval is

$$\mu\in\left[\overline{X}-\frac{z_\alpha\sigma}{\sqrt{N}},+\infty\right]$$

Notice that where we usually have $z_{\alpha/2}$ in the definition, we now have z_α. This is because where we previously cared about "outside" probability entering into both tails – giving us half of the "outside" probability in each tail – we now only have one tail outside of our interval. We can then put full "outside" mass into that tail, leading to z_α being the term of interest. In this case, if we backed into the question and said that we wanted to build a 67.6% one-sided confidence interval for μ (to $+\infty$), we could look up in the table to find $z_\alpha=\Phi^{-1}(0.6760)$, or $z_\alpha=0.4564$. That we'd obtain the same result is obvious.

As a side note, we again run into some intuitional issues regarding why we would care if $\mu>1$, since we know that the true value of the population parameter is such that $0\leq\mu\leq 1$. This is because we are using a continuous approximation for the true distribution of the sample mean; there is also 0 probability that $\overline{X}>1$, but our approximation will assign some nonzero probability to this occurring.

The best intuition for getting around this is possibly that if our continuous distribution would assign $\overline{X} > 1$, the "most reasonable" real-world assignment is $\overline{X} = 1$. So we lump all of the tail probability into the one possible value – $\overline{X} = 1$ – and account for this in the continuous distribution by including cases where $\mu > 1$.

Notice that we could apply similar appeals to symmetry to determine the confidence of any arbitrary interval, for example $[0.5, 0.8]$ *even though it is not symmetric about the sample mean.* This is only slightly more difficult, but isn't all that useful in this context.

(g) *Suppose the population variance is unknown. What is a* 95% *confidence interval for* μ*?*

In this question, we change the problem by removing knowledge of the population variance. When this is unknown, we use the t distribution and the sample variance to compute confidence intervals. To begin, we compute the sample variance:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 \\ &= \frac{1}{4}\left(3(1-0.6)^2 + 2(0-0.6)^2\right) \\ &= \frac{1}{4}\left(0.48 + 0.72\right) \\ S^2 &= 0.3 \end{aligned}$$

The generic form of the confidence interval is the same,

$$\mu \in \left[\overline{X} - E, \overline{X} + E\right]$$

In fact, the explicit form doesn't change that much either,

$$\mu \in \left[\overline{X} - \frac{t_{\alpha/2}(N-1)S}{\sqrt{N}}, \overline{X} + \frac{t_{\alpha/2}(N-1)S}{\sqrt{N}}\right]$$

The two differences between this form and that for the normal (i.e., the case with known variance) is that we use the sample standard deviation S rather than the population standard deviation σ, and rather than using the z-value $z_{\alpha/2}$ we use the t-value $t_{\alpha/2}(N-1)$. The meaning behind the t-value is essentially identical to that behind the z-value (for more on this, read part (c) of this question), except that it is drawn from a different – but still symmetric about the mean – distribution.

It's worth noting that the t-value is dependent on the number of observations. In some sense, this allows the distribution to "know" how accurate our sample variance is. As the number of observations grows large, the t distribution approaches the normal.

Now, since we have calculated S, all that remains is to find the proper $t_{\alpha/2}(N-1)$ value. Since a 95% confidence interval implies $\alpha = 0.05$, we look for $t_{0.025}(4)$. The book tells us this is $t_{0.025}(4) = 2.776$. We now have all of the pieces we need to compute the desired confidence interval. We see

$$E = \frac{t_{0.025}(4)S}{\sqrt{N}} = 0.6801$$

Then a 95% confidence interval for μ is

$$\mu \in [-0.0801, 1.2801]$$

There are two points of note in this:

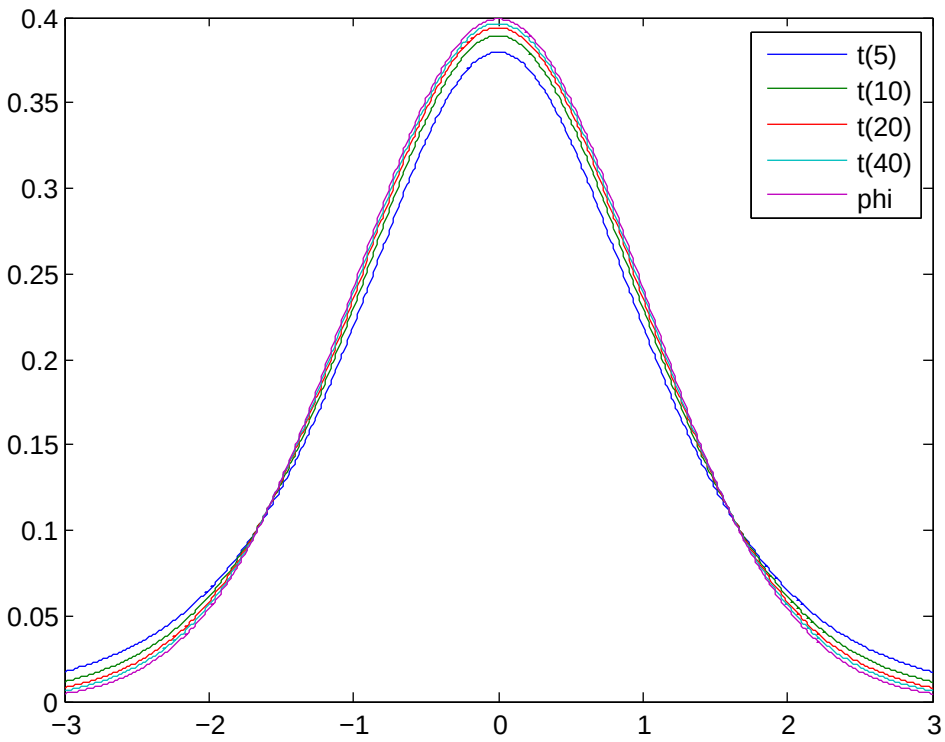

Figure 1: t PDFs with various degrees of freedom; as the degree of freedom grows, the t PDF approaches ϕ, the normal PDF

(i) this confidence interval is even more "ridiculous" than that in part (c). That is, our confidence interval now contains all possible values of μ! Again, this is a function of the fact that using a continuous approximation of this discrete problem was not a good choice. Still, in general (in Econ 41) this will be the proper choice to make, and it's worth working through here.

(ii) this confidence interval is larger than that derived with known population variance (using the z-statistic). Intuitively, this is because there is more uncertainty in the system: since we don't know the actual variance involved, we have to adjust our estimators (here, the confidence interval) to account for the fact that it could be larger, it could be smaller. In general, to reach the same level of confidence with more uncertainty we will need to broaden our interval estimators; we could also take the opposite approach, keeping our confidence interval fixed while reducing the confidence we have in it.

It may be worthwhile to repeat the previous exercises (particularly (e) and (f)) assuming that the population variance is unknown. The trick is that nothing really changes, other than that we approximate with the sample variance and the t-statistic, rather than the population variance and the z-statistic. Even though not much is different, practice makes perfect!

Question 2: in 2014, I will eat my 30^{th} Thanksgiving dinner. Assume that (as of 2014) I rest for – on average – 1 hour after eating, with a standard deviation of 5 minutes, and then I nap for – on average – 3 hours, with a standard deviation of 15 minutes. Assume both standard deviations are known and not sampled, and that the time spent resting and the time spent napping are independent.

(a) *Construct a 95% confidence interval for the amount of time I rest after eating Thanksgiving dinner (not including napping).*

As is usual, we have

$$\mu_r \in \left[\overline{X}_r - \frac{z_{\alpha/2}\sigma_r}{\sqrt{N}}, \overline{X}_r + \frac{z_{\alpha/2}\sigma_r}{\sqrt{N}}\right]$$

Here, $\sigma = 5$, $\overline{X}_r = 60$, and $N = 30$. From the table in the back of the book, we know $z_{\alpha/2} = 1.96$ when we are interested in 95% confidence intervals (more directly, $z_{0.025} = 1.96$).

With these numbers in hand, we can see directly that

$$\mu_r \in [58.2108, 61.7892]$$

(b) *With what confidence can we say that the duration of post-rest nap is within* 5 *minutes of* 3 *hours?*

We know that we would like

$$\mu_n \in [175, 185]$$

Since $\overline{X}_n = 180$, this means that our margin of error is $E = 5$ (we can also see this directly from the question statement). Then we want

$$\frac{z_{\alpha/2}\sigma_n}{\sqrt{N}} = 5$$

With $\sigma_n = 15$ and $N = 30$, this tells us $z_{\alpha/2} = 1.8257$. Looking up this z-value in the tables in the back of the book, we see that $\mu_n \in [175, 185]$ represents a 93.21% confidence interval for μ_n.

(c) *If the experiment continues, in what year will my total post-dinner rest-and-nap time be known to within* ± 5 *minutes with* 95% *confidence?*

To begin, we consider distributional facts about total post-dinner rest-and-nap time. Since the time spent resting X_r and the time spent napping X_n are independent, we know that the variance of their sum is the sum of their variances,

$$\text{Var}\left(X_r + X_n\right) = \text{Var}(X_r) + \text{Var}(X_n)$$

Here, $X_r + X_n$ is the total time spent resting and napping (rest-and-nap time); for ease of notation, let $T = X_r + X_n$. Then we know

$$\sigma_T = \sqrt{\sigma_r^2 + \sigma_n^2}$$

Suppose we'd like to construct a 95% confidence interval with margin of error $E = 5$; our only method of "getting there" by repeating the experiment is to raise the size of our sample. We see

$$\frac{z_{\alpha/2}\sigma_T}{\sqrt{N}} = E \quad \Longrightarrow \quad N = \frac{z_{\alpha/2}^2\left(\sigma_r^2 + \sigma_n^2\right)}{E^2}$$

Again, $z_{0.025} = 1.96$, so we see

$$N = \frac{1.96^2\left(5^2 + 15^2\right)}{5^2} = 38.4160$$

So we need to run at least 38.4160 experiments to achieve this margin of error with 95% confidence. Since experiments are discrete – we can run only 1 or 0 experiments at a time, not 0.4160 of an experiment – we'll need to take $\hat{N} = 39$ samples.

If our 30^{th} sample is taken in 2014, the 39^{th} sample will be taken in 2023. Then in 2023, we will know the mean of the total post-eating rest-and-nap time to within ± 5 minutes with 95% confidence.

Question 3: parking structure 7 – across from Pauley Pavilion – contains 6 bike lockers which are reservable by phone. Let X be a random variable denoting the number of bike lockers which are registered as occupied even though they contain no bicycle. Over 50 observations, we witness the following data:

X	#
0	0
1	0
2	1
3	3
4	10
5	20
6	16

(a) *What are the sample mean and variance?*

These are fairly direct calculations,

$$\begin{aligned}\overline{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ &= \left(\frac{1}{50}\right)2 + \left(\frac{3}{50}\right)3 + \left(\frac{10}{50}\right)4 + \left(\frac{20}{50}\right)6 + \left(\frac{16}{50}\right)6 \\ &= 4.9400\end{aligned}$$

$$\begin{aligned}s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \\ &= \left(\frac{1}{49}\right)(2-4.94)^2 + \left(\frac{3}{49}\right)(3-4.94)^2 + \left(\frac{10}{49}\right)(4-4.94)^2 + \left(\frac{20}{49}\right)(5-4.94)^2 + \left(\frac{16}{49}\right)(6-4.94)^2 \\ &= 0.9555\end{aligned}$$

(b) *What is a 99% confidence interval for the mean number of falsely-occupied bike lockers?*

Since we do not know the population variance, we must use the sample variance to compute this confidence interval. When we use the sample variance, we construct confidence intervals using the t-distribution rather than the standard normal distribution. The expression of a confidence interval looks very similar, the only difference being that we employ a t-value rather than a z-value. Our confidence interval will then take the form

$$\mu \in \left[\overline{X} - \frac{t_{\alpha/2}(N-1)s}{\sqrt{N}}, \overline{X} + \frac{t_{\alpha/2}(N-1)s}{\sqrt{N}}\right]$$

We can't quite look up $t_{0.005}(49)$ in the back of the book – 49 is out of range – so take it on faith that $t_{0.005}(49) = 2.68$ (this isn't too far from what we *can* find in the back of the book, $t_{0.005}(30) = 2.75$, so it jibes with intuition). From here, we've got all the values we need to explicitly compute the confidence interval, since we calculated the sample mean and variance in part (a) above. Then we have

$$\mu \in [4.5695, 5.3105]$$

is a 95% confidence interval for μ.

(c) *How many more observations are needed to determine the mean number of falsely-occupied bike lockers to within* ± 0.01 *with* 99% *confidence? Assume that the sample variance does not change.*

First, note that the assumption needed to attack this question is not a good assumption. We are not likely to see a constant sample variance as we increase the size of our sample: data is random! Still, without making this assumption there is nowhere we can get reasonably quickly.

There is a further issue which will make this problem practically unsolvable (without numerical tools such as MATLAB): the 99% confidence interval is determined by the t-value $t_{0.005}(N-1)$, which is dependent on the sample size! This is not something we can easily invert on paper, so we're going to have to make some simplifying assumptions and hand-wave a little bit.

We know that a confidence interval should look like

$$\mu \in \left[\overline{X} - E, \overline{X} + E\right]$$

Since we want a 99% confidence interval with margin of error $E = 0.01$, and we know $E = \frac{t_{\alpha/2}(N-1)s}{\sqrt{N}}$, we can compute

$$\frac{\sqrt{N}}{t_{0.005}(N-1)} = 100s = 97.75$$

Here's where the hand-waving comes in: as $N \to \infty$, the t distribution approaches the normal distribution; from the z-tables, we know $z_{0.005} = 2.5758$. Since the t-value lies above the z-value everywhere, this will be a good lower bound for $t_{0.005} \geq 2.5758$. We can plug in to find

$$\frac{\sqrt{N}}{t_{0.005}(N-1)} \leq \frac{\sqrt{N}}{z_{0.005}} \quad \Longrightarrow \quad \frac{\sqrt{N}}{z_{0.005}} \geq 97.75 \quad \Longrightarrow \quad \sqrt{N} \geq 97.75(2.5758) \quad \Longrightarrow \quad N \geq 63397$$

So we need at least sixty thousand samples, roughly. A calculator will tell us that $t_{0.005}(63396) = 2.5759$, so our approximation of the t-value was not that far off (actually solving this explicitly with a computer yields $N = 63400$).

To finish hand-waving, an acceptable answer would be that we need at between 60000 and 65000 samples to obtain a confidence interval with $E = 0.01$ with 99% confidence. That is one sample per day for almost 200 years; it goes to show that arbitrary confidence is not always the best objective: we probably could have done just as well in whatever purpose we were using these measurements for with $E = 0.1$, requiring roughly 650 observations at 99% confidence (I'll leave that math to you), and probably just as well with still a lower level of confidence. However, these decisions are subject to the nature of the experiment and the model we are trying to fit.