Confidence Intervals: Three Views
Class 23, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to find z, t and χ^2 confidence intervals using the corresponding standardized statistics.

2. Be able to use a hypothesis test to find a confidence interval for an unknown parameter.

3. Refuse to answer questions that ask, in essence, 'given a confidence interval what is the probability or odds that it contains the true value of the unknown parameter?'

2 Introduction

Our approach to confidence intervals in the previous reading was a combination of standardized statistics and hypothesis testing. Today we will consider each of these perspectives separately, as well as introduce a third formal viewpoint. Each provides its own insight.

1. **Standardized statistic.** Most confidence intervals are based on standardized statistics with known distributions like z, t or χ^2. This provides a straightforward way to construct and interpret confidence intervals as a point estimate plus or minus some error.

2. **Hypothesis testing.** Confidence intervals may also be constructed from hypothesis tests. In cases where we don't have a standardized statistic this method will still work. It agrees with the standardized statistic approach in cases where they both apply.

This view connects the notions of significance level α for hypothesis testing and confidence level $1 - \alpha$ for confidence intervals; we will see that in both cases α is the probability of making a 'type 1' error. This gives some insight into the use of the word confidence. This view also helps to emphasize the frequentist nature of confidence intervals.

3. **Formal.** The formal definition of confidence intervals is perfectly precise and general. In a mathematical sense it gives insight into the inner workings of confidence intervals. However, because it is so general it sometimes leads to confidence intervals without useful properties. We will not dwell on this approach. We offer it mainly for those who are interested.

3 Confidence intervals via standardized statistics

The strategy here is essentially the same as in the previous reading. Assuming normal data we have what we called standardized statistics like the standardized mean, Studentized mean, and standardized variance. These statistics have well known distributions which depend on hypothesized values of μ and σ. We then use algebra to produce confidence intervals for μ or σ.

Don't let the algebraic details distract you from the essentially simple idea underlying confidence intervals: we start with a standardized statistic (e.g., z, t or χ^2) and use some algebra to get an interval that depends only on the data and known parameters.

3.1 z-confidence intervals for the mean: normal data with known standard deviation

z-confidence intervals for the mean of normal data are based on the standardized mean, i.e. the z-statistic. We start with n independent normal samples

$$x_1, x_2, \dots, x_n \sim \mathrm{N}(\mu, \sigma^2).$$

We assume that μ is the unknown parameter of interest and σ is known. Notationally, let's write the (unknown) true value of μ as μ_0

We know that the standardized mean is standard normal:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

For the standard normal critical value $z_{\alpha/2}$ we have: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Thus,

$$P\left(-z_{\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \mid \mu = \mu_0\right) = 1 - \alpha$$

A little bit of algebra puts this in the form of an interval around μ:

$$P\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

We can emphasize that the interval depends only on the statistic $\overline{x}$ and the known value σ by writing this as

$$P\left(\left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \ \text{ contains } \ \mu \mid \mu = \mu_0\right) = 1 - \alpha.$$

This is the $(1 - \alpha)$ z-confidence interval for μ. We often write it using the shorthand

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error.

Make sure you notice that the probabilities are conditioned on $\mu = \mu_0$. As with all frequentist statistics, we have to fix hypothesized values of the parameters in order to compute probabilities.

3.2 t-confidence intervals for the mean: normal data with unknown mean and standard deviation

t-confidence intervals for the mean of normal data are based on the Studentized mean, i.e. the t-statistic.

Again we have $x_1, x_2, \dots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, but now we assume both μ and σ are unknown. As we did above, let's write the (unknown) true value of μ as μ_0. We know that the Studentized mean follows a Student t distribution with $n-1$ degrees of freedom. That is,

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1),$$

where s^2 is the sample variance.

Now all we have to do is replace the standardized mean by the Studentized mean and the same logic we used for z gives us the t-confidence interval: start with

$$P\left(-t_{\alpha/2} < \frac{\overline{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2} \mid \mu = \mu_0\right) = 1 - \alpha.$$

A little bit of algebra isolates μ in the middle of an interval:

$$P\left(\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

We can emphasize that the interval depends only on the statistics $\overline{x}$ and s by writing this as

$$P\left(\left[\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \ \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \ \text{ contains } \ \mu \mid \mu = \mu_0\right) = 1 - \alpha.$$

This is the $(1-\alpha)$ t-confidence interval for μ. We often write it using the shorthand

$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error.

3.3 Chi-square confidence intervals for variance: normal data with unknown mean and standard deviation

You guessed it: χ^2-confidence intervals for the variance of normal data are based on the standardized variance, i.e. the χ^2-statistic.

We follow the same logic as above to get a χ^2-confidence interval for σ^2. Because this is the third time through it we'll move a little more quickly.

We assume we have n independent normal samples: $x_1, x_2, \dots, x_n \sim \mathrm{N}(\mu, \sigma^2)$. We assume that μ and σ are both unknown and write the (unknown) true value of σ as σ_0. The standardized variance is

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1).$$

We know that the X^2 statistic follows a χ^2 distribution with $n-1$ degrees of freedom.

For Z and t we used, without comment, the symmetry of the distributions to replace $z_{1-\alpha/2}$ by $-z_{\alpha/2}$ and $t_{1-\alpha/2}$ by $-t_{\alpha/2}$. Because the χ^2 distribution is not symmetric we need to be explicit about the critical values on both the left and the right. That is,

$$P(c_{1-\alpha/2} < X^2 < c_{\alpha/2}) = 1 - \alpha,$$

where $c_{\alpha/2}$ and $c_{1-\alpha/2}$ are right tail critical values. Thus,

$$P\left(c_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < c_{\alpha/2} \mid \sigma = \sigma_0\right) = 1-\alpha$$

A little bit of algebra puts this in the form of an interval around σ^2:

$$P\left(\frac{(n-1)s^2}{c_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{c_{1-\alpha/2}} \mid \sigma = \sigma_0\right) = 1-\alpha$$

We can emphasize that the interval depends only on the statistic s^2 by writing this as

$$P\left(\left[\frac{(n-1)s^2}{c_{\alpha/2}}, \ \frac{(n-1)s^2}{c_{1-\alpha/2}}\right] \ \text{ contains } \ \sigma^2 \mid \sigma = \sigma_0\right) = 1-\alpha.$$

This is the $(1-\alpha)$ χ^2-confidence interval for σ^2.

4 Confidence intervals via hypothesis testing

Suppose we have data drawn from a distribution with a parameter θ whose value is unknown. A significance test for the value θ has the following short description.

1. Set the null hypothesis $H_0 : \theta = \theta_0$ for some special value θ_0, e.g. we often have $H_0 : \theta = 0$.

2. Use the data to compute the value of a test statistic, call it x.

3. If x is far enough into the tail of the null distribution (the distribution assuming the null hypothesis) then we reject H_0.

In the case where there is no special value to test we may still want to estimate θ. This is the reverse of significance testing; rather than seeing if we should reject a specific value of θ because it doesn't fit the data we want to find the range of values of θ that do, in some sense, fit the data. This gives us the following definitions.

Definition. Given a value x of the test statistic, the $(1-\alpha)$ confidence interval contains all values θ_0 which are not rejected (at significance level α) when they are the null hypothesis.

Definition. A type 1 CI error occurs when the confidence interval does not contain the true value of θ.

For a $(1-\alpha)$ confidence interval the type 1 CI error rate is α.

Example 1. Here is an example relating confidence intervals and hypothesis tests. Suppose data x is drawn from a binomial(12, θ) distribution with θ unknown. Let $\alpha = 0.1$ and create the $(1-\alpha) = 90\%$ confidence interval for each possible value of x.

Solution: Our strategy is to look at one possible value of θ at a time and choose rejection regions for a significance test with $\alpha = 0.1$. Once this is done, we will know, for each value of x, which values of θ are not rejected, i.e. the confidence interval associated with x.

To start we set up a likelihood table for binomial(12, θ) in Table 1. Each row shows the probabilities $p(x|\theta)$ for one value of θ. To keep the size manageable we only show θ in increments of 0.1.

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1. Likelihood table for Binomial(12, θ)

Tables 2-4 below show the rejection region (in orange) and non-rejection region (in blue) for the various values of θ. To emphasize the row-by-row nature of the process the Table 2 just shows these regions for $\theta = 1.0$, then Table 3 adds in regions for $\theta = 0.9$ and Table 4 shows them for all the values of θ.

Immediately following the tables we give a detailed explanation of how the rejection/non-rejection regions were chosen.

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 2. Likelihood table for binomial(12, θ) with rejection (orange)/non-rejection (blue) regions for $\theta = 1.0$

$\theta\backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Table 3. Likelihood table: rejection (orange)/non-rejection (blue) regions for $\theta = 1.0$ and 0.9

$\theta \backslash x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	**0.13**	0.24	0.28	0.21	0.07	0.073
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	**0.23**	0.24	0.17	0.07	0.01	0.052
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	**0.21**	0.14	0.06	0.02	0.00	0.077
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	**0.12**	0.05	0.02	0.00	0.00	0.092
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	0.077
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.052
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.073
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.026
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000

Table 4. Likelihood table: rejection (orange)/non-rejection (blue) regions for $\theta = 0.0$ to 1.0

Choosing the rejection and non-rejection regions in the tables

The first problem we confront is how exactly to choose the rejection region. We used two rules:

1. The total probabilitiy of the rejection region, i.e. the significance, should be less than or equal to 0.1. (Since we have a discrete distribution it is impossible to make the significance exactly 0.1.)

2. We build the rejection region by choosing values of x one at a time, always picking the unused value with the smallest probability. We stop when the next value would make the significance more that 0.1.

There are other ways to choose the rejection region which would result in slight differences. Our method is one reasonable way.

Table 2 shows the rejection (orange) and non-rejection (blue) regions for $\theta = 1.0$. This is a special case because most of the probabilities in this row are 0.0. We'll move right on to the next table and step through the process for that.

In Table 3, let's walk through the steps used to find these regions for $\theta = 0.9$.

- The smallest probability is when $x = 0$, so $x = 0$ is in the rejection region.
- The next smallest is when $x = 1$, so $x = 1$ is in the rejection region.
- We continue with $x = 2, \dots, 8$. At this point the total probability in the rejection region is 0.026.
- The next smallest probability is when $x = 9$. Adding this probability (0.09) to 0.026 would put the total probability over 0.1. So we leave $x = 9$ out of the rejection region and stop the process.

Note three things for the $\theta = 0.9$ row:

1. None of the probabilities in this row are truly zero, though some are small enough that they equal 0 to 2 decimal places.

2. We show the significance for this value of θ in the right hand margin. More precisely, we show the significance level of the NHST with null hypothesis $\theta = 0.9$ and the given rejection region.

3. The rejection region consists of values of x. When we say the rejection region is shown in orange we really mean the rejection region contains the values of x corresponding to the probabilities highlighted in orange.

Think: Look back at the $\theta = 1.0$ row and make sure you understand why the rejection region is $x = 0, \dots, 11$ and the significance is 0.000.

Example 2. Using Table 4 determine the 0.90 confidence interval when $x = 8$.

Solution: The 90% confidence interval consists of all those θ that would not be rejected by an $\alpha = 0.1$ hypothesis test when $x = 8$. Looking at the table, the blue (non-rejected) entries in the column $x = 8$ correspond to $0.5 \le \theta \le 0.8$: the confidence interval is $[0.5, 0.8]$.

Remark: The point of this example is to show how confidence intervals and hypothesis tests are related. Since Table 4 has only finitely many values of θ, our answer is close but not exact. Using a computer we could look at many more values of θ. For this problem we used R to find that, correct to 2 decimal places, the confidence interval is $[0.42, 0.85]$.

Example 3. Explain why the expected type one CI error rate will be at most 0.092, provided that the true value of θ is in the table.

Solution: The short answer is that this is the maximum significance for any θ in Table 4. Expanding on that slightly: we make a type one CI error if the confidence interval does not contain the true value of θ, call it θ_{true}. This happens exactly when the data x is in the rejection region for θ_{true}. The probability of this happening is the significance for θ_{true} and this is at most 0.092.

Remark: The point of this example is to show how confidence level, type one CI error rate and significance for each hypothesis are related. As in the previous example, we can use R to compute the significance for many more values of θ. When we do this we find that the maximum significance for any θ is 0.1 ocurring when $\theta \approx 0.0452$.

Summary notes:
1. We start with a test statistic x. The confidence interval is random because it depends on x.

2. For each hypothesized value of θ we make a significance test with significance level α by choosing rejection regions.

3. For a specific value of x the associated confidence interval for θ consists of all θ that aren't rejected for that value, i.e. all θ that have x in their non-rejection regions.

4. Because the distribution is discrete we can't always achieve the exact significance level, so our confidence interval is really an 'at least 90% confidence interval'.

Example 4. Open the applet https://mathlets.org/mathlets/confidence-intervals/. We want you to play with the applet to understand the random nature of confidence intervals and the meaning of confidence as (1 - type I CI error rate).

(a) Read the help. It is short and will help orient you in the applet. Play with different settings of the parameters to see how they affect the size of the confidence intervals.

(b) Set the number of trials to $N = 1$. Click the 'Run N trials' button repeatedly and see that each time data is generated the confidence intervals jump around.

(c) Now set the confidence level to $c = 0.5$. As you click the 'Run N trials' button you

should see that about 50% of the confidence intervals include the true value of μ. The 'Z correct' and 't correct' values should change accordingly.

(d) Now set the number of trials to $N = 100$. With $c = 0.8$. The 'Run N trials' button will now run 100 trials at a time. Only the last confidence interval will be shown in the graph, but the trials all run and the 'percent correct' statistics will be updated based on all 100 trials.

Click the run trials button repeatedly. Watch the correct rates start to converge to the confidence level. To converge even faster, set $N = 1000$.

5 Formal view of confidence intervals

Recall: An interval statistic is an interval I_x computed from data x. An interval is determined by its lower and upper bounds, and these are random because x is random.

We suppose that x is drawn from a distribution with pdf $f(x|\theta)$ where the parameter θ is unknown.

Definition: A $(1-\alpha)$ confidence interval for θ is an interval statistic I_x such that

$$P(I_x \text{ contains } \theta_0 \mid \theta = \theta_0) \;=\; 1-\alpha$$

for all possible values of θ_0.

We wish this was simpler, but a definition is a definition and this definition is one way to weigh the evidence provided by the data x. Let's unpack it a bit.

The confidence level of an interval statistic is a probability concerning a random interval and a hypothesized value θ_0 for the unknown parameter. Precisely, it is the probability that the random interval I_x (computed from random data x) contains the value θ_0, given that the model parameter truly is θ_0. Since the true value of θ is unknown, the frequentist statistician defines 95% confidence intervals so that the 0.95 probability is valid no matter which hypothesized value of the parameter is actually true.

6 Comparison with Bayesian probability intervals

Confidence intervals are a frequentist notion, and as we've repeated many times, frequentists don't assign probabilities to hypotheses, e.g., to the value of an unknown parameter. Rather they compute likelihoods; that is, probabilities about data or associated statistics given a hypothesis (note the condition $\theta = \theta_0$ in the formal view of confidence intervals). Note that the construction of confidence intervals proceeds entirely from the full likelihood table.

In contrast Bayesian posterior probability intervals are truly the probability that the value of the unknown parameter lies in the reported range. We add the usual caveat that this depends on the specific choice of a (possibly subjective) Bayesian prior.

This distinction between the two is subtle because Bayesian posterior probability intervals and frequentist confidence intervals share the following properties:

1. They start from a model $f(x|\theta)$ for observed data x with unknown parameter θ.

2. Given data x, they give an interval $I(x)$ specifying a range of values for θ.

3. They come with a number (say 0.95) that is the probability of something.

In practice, many people misinterpret confidence intervals as Bayesian probability intervals, forgetting that frequentists **never** place probabilities on hypotheses (this is analogous to mistaking the p-value in NHST for the probability that H_0 is false). The next section explores this mistake in some detail. The harm of this misinterpretation is somewhat mitigated by that fact that, given enough data and a reasonable prior, Bayesian and frequentist intervals often work out to be quite similar.

For an amusing example illustrating how they can be quite different, see the first answer in the link just below (involving chocolate chip cookies!). This example uses the formal definitions and is really about confidence sets instead of confidence intervals.

https://stats.stackexchange.com/questions/2272/whats-the-difference-between-a-confidence-interval-and-a-credible-interval

7 Misinterpreting confidence intervals

It is very tempting to think that given a 95% confidence interval for, say, the mean, the probability that the true mean is in the confidence interval is 95%.

We know this can't be true because the value of the mean can only be hypothesized and Frequentists don't assign probabilities to hypotheses. To be more concrete, if the mean is θ and the confidence interval is $[45, 55]$ then the statement $45 \le \theta \le 55$ is a hypothesis, so asking for the probability that $45 \le \theta \le 55$, is asking for the probability of a hypothesis.

The mistake is subtle and hard to wrap your mind around. It boils down to a question of what is being randomly sampled. Here is an attempt to explain the issue.

First, consider a test for a disease. Assume a person is given the test. Let T^+ be a positive test and D^+ be that they have the disease. Assume the test is 95% accurate, i.e. $P(T^+|D^+) = 0.95$. We know (base rate fallacy) that this does not imply that $P(D^+|T^+) = 0.95$.

Let's look at this from a different angle: Implicit in $P(T^+|D^+) = 0.95$ is the following experiment: Draw a random person from the set of all people with the disease and give them the test. Then 95% will test positive. That is, the population sampled is all people with the disease and the event considered is that the chosen person in that population tests positive.

For $P(D^+|T^+)$, the experiment is to draw a random person from the set of all people who tested positive. The probability is the fraction who have the disease. That is, the population sampled is all people who test positive and the event considered is that the chosen person in that population has the disease.

We can't expect $P(T^+|D^+)$ and $P(D^+|T^+)$ to be the same, since we're sampling from different populations and looking at different events. The probability $P(D^+|T^+)$ can be computed using Bayes' theorem from the (prior) probability $P(D^+)$ and the likelihoods $P(T^+|D+)$, $P(T^+|D^-)$.

Confidence intervals are a little more abstract, but the analysis is similar. Just as in testing for a disease, the populations sampled will be different. One source of difficulty is that the

events are essentially the same in both cases.

Let's assume we have a distribution with unknown mean θ_0. We generate some data and compute a 95% confidence interval for the mean. The value of 95% comes from the following implied experiment: imagine having run many trials and created a confidence interval for each one. Then, 95% of confidence intervals contain the true mean. In notation,

$$P(\text{random interval contains } \theta_0) = 0.95.$$

That is, the random sample is drawn from the set of all confidence intervals generated by our trials. The event in question is that the chosen interval contains the true mean.

What if we run one experiment and generate the 95% confidence interval, call it I. To a Frequentist, the true mean is not random and we have a fixed interval. So, to the Frequentist, it makes no sense to ask about the probability θ_0 is in I.

To a Bayesian, it is fine to consider θ_0 as randomly drawn from a probability distribution –they often interpret it as a description of the uncertainty of our knowledge. So, they can ask for the probability

$$P(\text{random } \theta_0 \text{ is in a given } I).$$

So, here, the random sample is drawn from the set of possible means and the event considered is that the chosen mean is contained in the given interval.

As in the disease testing example, what population is being randomly sampled is different in the two cases, i.e in the first we have a random interval, in the second we have a random value for the true mean. As noted above, in both cases the event is that the true mean is in the interval.

We finish by noting that $P(\text{random } \theta_0 \text{ is in } I)$ can be computed using Bayes' theorem and depends on the prior distribution for the true mean and the likelihoods that each mean will generate the given confidence interval. The formula is a little unwieldy. Here it is.

- Call the interval I and the true mean θ_0.
- Call the data I. This is a shorthand for the data that the interval I is based on.
- Let $p(\theta)$ be the prior probability that $\theta_0 = \theta$.
- The likelihood $f(I|\theta)$ is the probability (or density) that the experiment would produce the interval I given $\theta_0 = \theta$.
- Let $p(\theta|I)$ be the posterior probability that $\theta_0 = \theta$. This is the updated probability found using the Bayes' theorem.

Bayes' theorem gives us

$$p(\theta|I) = \frac{f(I|\theta)p(\theta)}{f(I)}, \text{ where } f(I) = \sum_{\theta} f(I|\theta)p(\theta).$$

So we have, $P(\theta_0 \in I|I) = \sum_{\theta \text{ in } I} p(\theta|I)$. (As usual, if θ has a continuous range of values, then the sums will be replaced by integrals.)

18.05 Introduction to Probability and Statistics
Spring 2022