

Name: _____
Class #: _____
Instructor: Ralf Becker

Class: _____
Section #: _____
Assignment: Confidence Intervals

Question 1: (0 points)

Introduction to Point and Interval Estimation

In the section on sampling distribution we established the link between population parameters and their sample estimators, in particular the link between the population mean and its sample average. In the Section on Hypothesis Testing we then established a particular technique (hypothesis testing) we can use when we want to learn from a (known) sample something about the unknown population parameter. Of course, different samples will yield different estimates of the population parameter: this is the idea of **sampling variability**. It is because of this sampling variability that hypothesis testing could only make probabilistic statements.

In this section we will learn about an alternative (yet linked) way to express the uncertainty arising from sampling variability. Confidence intervals are a popular tool to communicate this uncertainty, one which, in some respects, may deliver a more intuitive way of presenting the uncertainty we are facing.

Intro to Confidence Intervals



Question 2: (1 point)

Calculating Confidence Intervals

Sampling Variability

Let's recall the link between population parameter (here population mean μ) and sample statistic (here sample mean \bar{X}). Consider the simplest case of sampling from a normal distribution, $X \sim N(\mu, \sigma^2)$ with the intention of estimating μ . The obvious estimator is the sample mean, \bar{X} , with sampling distribution $\bar{X} \sim N(\mu, \sigma^2/n)$. Here, μ is unknown but for now we assume that σ^2 is also known.

The variance of the distribution of \bar{X} measures the dispersion of this distribution, and thus gives some idea of the range of values of \bar{X} that might be obtained in drawing different samples of size n . This measure, $\sigma_{\bar{X}}^2 = \sigma^2/n$, is therefore a measure of sampling variability and can be combined with the actual sample estimate, \bar{x} , in the following way:

$$\bar{x} \pm k\sigma_{\bar{X}},$$

with the value k chosen suitably (we will get to that later). What we get from this is an interval that is centred around the sample mean \bar{x} . The actual (unknown) population mean will or will not be contained in this interval, we just don't know.

So how could such an interval be useful? Before we get to this, let's calculate an interval we were able to calculate following our understanding of sampling distributions.

Assume that $X \sim N(7, 9)$ and that you draw a sample of size $n = 20$. Then we can calculate the probability that $\Pr(5.685 \leq \bar{X} \leq 8.314)$. To calculate this we use our knowledge that $\bar{X} \sim N(7, 9/20)$ which then allows us to calculate the value of that probability to be

$$\begin{aligned}\Pr(5.685 \leq \bar{X} \leq 8.314) &= \Pr\left(\frac{5.685 - 7}{\sqrt{9/20}} \leq Z \leq \frac{8.314 - 7}{\sqrt{9/20}}\right) \\ &= \Pr\left(\frac{5.685 - 7}{\sqrt{9/20}} \leq Z \leq \frac{8.314 - 7}{\sqrt{9/20}}\right) \\ &= \Pr(-1.96 \leq Z \leq 1.96) = 0.95\end{aligned}$$

where you can get the probability from the normal probability table. So, if we knew that the population mean was 7, then we would expect that with a 95% probability we would get a sample mean between 5.685 and 8.314. Where did these two values actually come from? They were calculated from

$$\mu \pm k\sigma_{\bar{X}}$$

where we used $\sigma_{\bar{X}}^2 = 9/20$ and $k = 1.96$ value as that is the value in the standard normal distribution that cuts off 2.5% in each tail. This is the interval into which, with a 95% probability we should expect the sample mean to fall, if we knew the population mean ($\mu = 7$).

Of course we do not know the population mean in most interesting applications and all we have is a sample mean \bar{x} . Can we just replace the unknown population mean μ with \bar{x} ? Yes we can, that is how we get to

$$\bar{x} \pm k\sigma_{\bar{X}},$$

If we again chose the $k = 1.96$ we call this the 95% confidence interval for the population mean. Continuing with the example we now learn that we obtained a sample mean of 7.931. This now allows us to calculate a 95% confidence interval

$$7.931 \pm 1.96 \cdot \sqrt{9/20} \Rightarrow [6.2031, 9.6589]$$

What interpretation can we give this confidence interval with lower limit/bound $c_L = 6.2031$ and upper limit/bound $c_U = 9.6589$? An intuitive interpretation would be "There is a 95% probability that the population mean is inside this interval." Unfortunately that is an incorrect interpretation. This interpretation treats the unknown population mean as if it was a random variable, but it is not. It is assumed to be fixed but unknown. What is random is the outcome of the sample mean \bar{x} and therefore the confidence interval (as the lower and upper limits, c_L and c_U , of that interval are centred around the particular outcome $\bar{x} = 7.931$ of the random variable \bar{X} these limits are also random variables, C_L and C_U). Recall, capital letters represent the random variable (like C_L) and small letters represent sample draws of that random variable (e.g. c_L).

Therefore, the correct interpretation is the following. If you did take 100 samples, then 95% of the associated confidence intervals should contain the population mean, μ .

Suppose that a random sample of size 50 is drawn from the distribution of household incomes, where the latter is supposed to be $N(\mu, 4)$, and that the mean of the sample is $\bar{x} = 22$. If we choose $k=1.96$, the 95% confidence interval is

$$\underline{\quad} \pm 1.96 \cdot \underline{\quad} \Rightarrow [\underline{\quad}, \underline{\quad}]$$

If you were to calculate the 99% confidence interval, the interval would

- (a) stay unchanged
- (b) become wider
- (c) become narrower

If you increased the sample size, the interval would

- (a) stay unchanged
 - (b) become wider
 - (c) become narrower
-

Question 3: (1 point)

Interpreting Confidence Intervals

This interval is defined by the sample values of C_L and C_U : these are the lower and upper confidence bounds and they have sample realisations:

$$c_L = \bar{x} - 1.96SE(\bar{X}), \quad c_U = \bar{x} + 1.96SE(\bar{X}).$$

The interval estimate

$$\bar{x} \pm 1.96SE(\bar{X})$$

contains the measure of sampling variability ($SE(\bar{X}) = \sigma_{\bar{X}}$).

The choice of k as $k = 1.96$ is now determined by the desired confidence level. Why choose the latter to be 0.95 or 95%? This is really a matter of convention and you will often also see 90% or 99% confidence intervals.

It is a common abuse of language to call the interval

$$\bar{x} \pm 1.96SE(\bar{X})$$

the confidence interval - indeed it is so common that this abuse will be allowed. Strictly, this is **an** interval estimate, which is now seen to be a combination of a point estimate, \bar{x} , of μ , and a measure of sampling variability determined by the constant k , which sets the confidence coefficient or level, in this case, 0.95.

There is a common misinterpretation of a confidence interval, based on this abuse of language, which says that **the confidence interval**

$$\bar{x} \pm 1.96SE(\bar{X})$$

contains μ with 95% confidence} Why is this a misinterpretation? For the following reasons:

- μ is unknown
- so this "confidence interval" does or does not contain μ ;
- since μ is unknown, we will \textbf{never} know which is true;

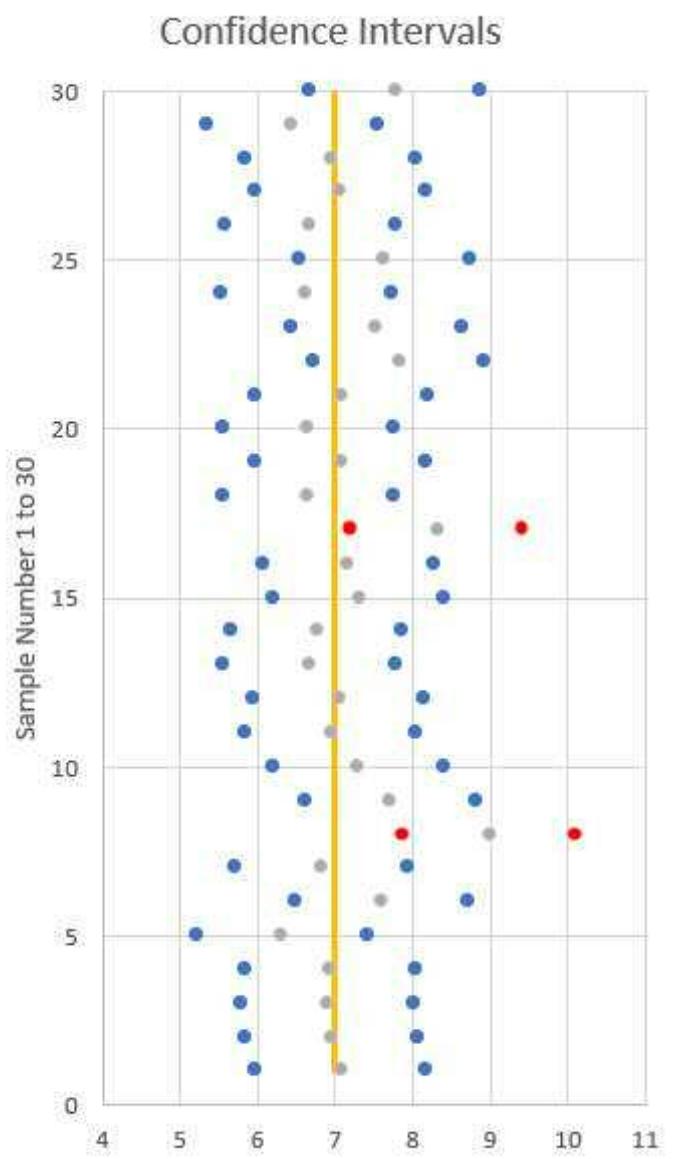
A better interpretation is based on the relative frequency interpretation of probability. If samples are repeatedly drawn from the population, say $X \sim N(\mu, \sigma^2)$, and the interval estimates ("confidence interval") for a 95% confidence level calculated for each sample, about 95% of them will contain μ .

This video (<https://youtu.be/Uby0i8GcJww>) illustrates this interpretation using a simple Excel simulation.

Confidence Interval Simulation in Excel



The following image illustrates thirty 90% confidence intervals from a population with mean 7 (yellow line indicates the population mean). The confidence intervals are centred around the point estimate \bar{x} represented by grey dots. c_L and c_U are in blue. The two red confidence intervals are the ones which do not contain the population mean.



However, this doesn't help when only a single sample is drawn. You just don't know whether that particular confidence interval does, or does not contain the true population mean. The way you can interpret a particular confidence interval is you have a point estimate (\bar{x}) and a measure of sampling variability.

Suppose that a random sample of size 50 is drawn from the distribution of household incomes, where the latter is supposed to be $N(\mu, 5)$. Notice that σ^2 here is supposed to be **known** to equal 5. Suppose that the mean of the sample is $\bar{x} = 18$. Then, the 95% confidence interval for μ is (allowing the abuse of language)

$$\begin{aligned}\bar{x} \pm 1.96SE(\bar{X}) &= 18 \pm 1.96\sqrt{\frac{5}{50}} \\ &= 18 \pm 0.62 \\ &= [17.38, 18.62].\end{aligned}$$

Here the measure of sampling variability around $\bar{x} = 18$ is ± 0.62 . One might reasonably conclude that since this measure of sampling variability is small compared to \bar{x} , $\bar{x} = 18$ is a relatively precise estimate of μ . To refer to precision here is fair, since we are utilising the variance of the sampling distribution of \bar{X} .

A random sample of size 20 is drawn from the distribution of household savings (measured in 1000s of UKP, negative numbers mean that a household is in debt), where the latter is supposed to be $N(\mu, 7)$. Notice that σ^2 here is supposed to be **known** to equal 7. Your sample mean is $\bar{x} = -23$. Then, the 95% confidence interval for μ is

$$\underline{\hspace{2cm}} \pm 1.96 \cdot \underline{\hspace{2cm}} \Rightarrow [\underline{\hspace{2cm}}, \underline{\hspace{2cm}}]$$

How big should your sample be such that the measure of sampling variability ($1.96 SE(\bar{X})$) is smaller than 1?

$n \geq \underline{\hspace{2cm}}$ (give the smallest possible natural number)

This video walks you through the calculations required for a problem similar to the above exercise.

Confidence Interval calculation



Question 4: (1 point)

Other Confidence Levels

Instead of choosing k so that

$$\Pr \left(-k \leq \frac{\bar{X} - \mu}{SE(\bar{X})} \leq k \right) = 0.95,$$

we can choose k such as to deliver any desired probability, usually expressed as $1 - \alpha$:

$$\Pr \left(-k \leq \frac{\bar{X} - \mu}{SE(\bar{X})} \leq k \right) = 1 - \alpha.$$

Since

$$Z = \frac{\bar{X} - \mu}{SE(\bar{X})} \sim N(0, 1),$$

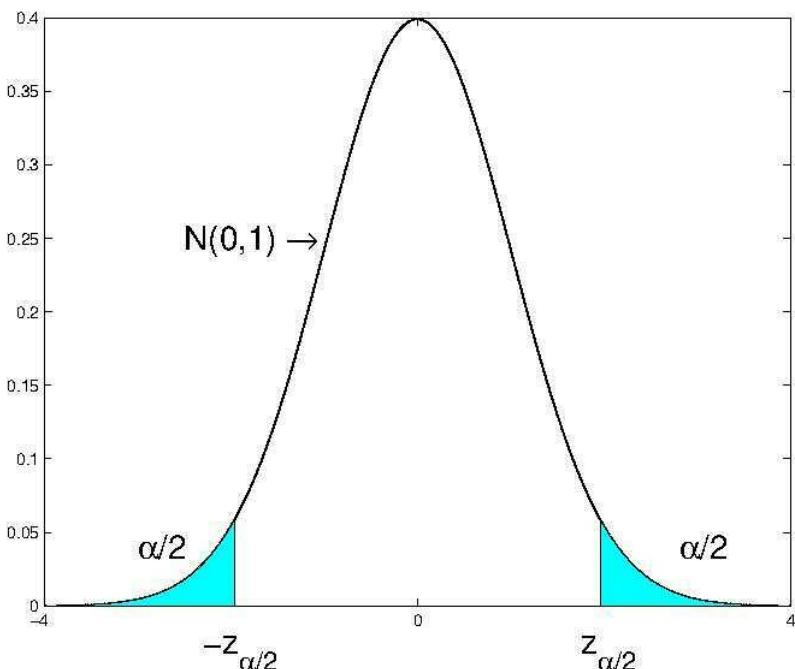
we can find from the table of the standard normal distribution the value (percentage point) $z_{\alpha/2}$ such that

$$\Pr(Z > z_{\alpha/2}) = \frac{\alpha}{2}.$$

This implies that

$$\Pr \left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2} \right) = 1 - \alpha.$$

This is clear from the familiar picture in the Figure below.



This delivers the following, commonly used values, to calculate confidence intervals assuming that the sample mean follows a normal distribution (as it does when the population is normal and the population variance is known):

α	$k = z_{\alpha/2}$
0.20	1.28
0.10	1.65
0.05	1.96
0.01	2.58

If we were to chose $\alpha = 0.01$ the confidence interval's lower and upper confidence bounds for the earlier example (population normally distributed with $\sigma^2 = 5$, sample size $n = 50$ and sample mean $\bar{x} = 18$) would be calculated as:

$$\begin{aligned}[c_L, c_U] &= \bar{x} \pm z_{\alpha/2} SE(\bar{X}) \\ &= 18 \pm 2.58 \sqrt{\frac{5}{50}} \\ &= 18 \pm 0.82 \\ &= [17.18, 18.82].\end{aligned}$$

Notice that the measure of sampling variability has increased from 0.62 for a 95% confidence interval to 0.82 for a 99% confidence interval. This illustrates the general proposition that the confidence interval gets wider as the confidence coefficient ($1 - \alpha$) is increased. There has been no change in the precision of estimation here.

Why the use of $1 - \alpha$ in the probability statement underlying the confidence interval? The random variables C_L and C_U are designed here to make

$$\Pr(\mu \in [C_L, C_U]) = 1 - \alpha$$

and therefore

$$\Pr(\mu \notin [C_L, C_U]) = \alpha.$$

Recall that in this probability statement it is the lower and upper bounds which are random variables, and not the population mean μ .

A random sample of size 16 is drawn from the distribution of annual household incomes (measured in thousands of pounds), where the latter is supposed to be $N(\mu, 12)$. Notice that σ^2 here is supposed to be known to equal to 12. Your sample mean is $\bar{x} = 33$. Then, the 80% confidence interval for μ is

$$\underline{\hspace{2cm}} \pm \underline{\hspace{2cm}} \cdot (\underline{\hspace{2cm}})^{1/2} \Rightarrow [\underline{\hspace{2cm}}, \underline{\hspace{2cm}}]$$

How big should your sample be such that the measure of sampling variability ($z_{\alpha/2} SE(\bar{X})$) is smaller than 1?

$$n \geq \underline{\hspace{2cm}} \text{ (provide the smalles possible natural number)}$$

Question 5: (1 point)

Which distribution to use?

We started the argument with a calculation we were familiar with from the discussion of sampling distributions. We asked a question like "What is $\Pr(c_L \leq \bar{X} \leq c_U)$?" and then based our calculation on what we knew about the sampling distribution of \bar{X} . In the case we discussed $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ as the population was assumed to be normally distributed with known $\sigma^2 = 5$.

The assumption, amongst these, which seemed the most problematic was that of a known population variance (σ^2). In a situation where we wish to estimate an unknown population mean (μ) it is unlikely that we know the population variance σ^2 . In the section on hypothesis testing we discussed how the sample mean \bar{X} is distributed when σ^2 has to be estimated by the sample variance s^2 . The following table represents what we know about the sampling distribution of the standardised sample mean \bar{X} , $t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$ in this situation:

	Sample Size (n)	
Dist of \bar{X}	small	large
Normal	$t \sim t_{n-1}$	$t \sim N(0, 1)$
not Normal	$t \sim ?$	$t \sim N(0, 1)$

This will lead us to adjust the way in which we calculate confidence intervals. In the situation where we could assume that \bar{X} was normally distributed the lower and upper confidence bounds were calculated as follows:

$$[c_L, c_U] = \bar{x} \pm z_{\alpha/2} SE(\bar{X}) = \bar{x} \pm z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$$

Now, we shall adjust this, acknowledging that we usually have to estimate the population variance from the sample, using s^2 :

$$[c_L, c_U] = \bar{x} \pm t_{n-1, \alpha/2} SE(\bar{X}) = \bar{x} \pm t_{n-1, \alpha/2} \sqrt{\frac{s^2}{n}}$$

The interpretation stays exactly the same as earlier and you should note that the only difference is that we use the sample estimate s^2 and we use the t-distribution to obtain our estimate of the sampling variation $t_{n-1, \alpha/2} \sqrt{\frac{s^2}{n}}$. You understand that, for large sample sizes the t-distribution tends towards the standard normal distribution.

This following video walks you through the different cases.

Confidence Intervals - Which distribution to use



This is the same as the previous example, but now assuming that the population variance σ^2 is unknown. Household income in UKP '000 is $X \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. A random sample of size $n = 5$ (previously 50) yields $\bar{x} = 18$ (as before) and $s^2 = 4.5$. Here,

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_4.$$

For a 95% confidence interval for μ , we need the percentage point $t_{4,0.025}$ such that

$$\Pr(T \leq t_{4,0.025}) = 0.975.$$

From the t-distribution this is found to be

$$t_{4,0.025} = 2.776.$$

The confidence interval for μ is then

$$\begin{aligned}\bar{x} \pm t_{n-1,\alpha/2} \sqrt{\frac{s^2}{n}} &= 18 \pm (2.776) \sqrt{\frac{4.5}{5}} \\ &= 18 \pm 2.634 \\ &= [15.366, 20.634].\end{aligned}$$

For comparison with the original example, if we had used $\sigma^2 = 5$ with a sample of size 5 the resulting normal-based confidence interval would be

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} &= 18 \pm (1.96) \sqrt{\frac{5}{5}} \\ &= 18 \pm 1.96 \\ &= [16.04, 19.96].\end{aligned}$$

This normal-based confidence interval is narrower than the t based one: this is the consequence of the extra dispersion of the t distribution compared with the $N(0, 1)$ distribution. The underlying reason for the increased dispersion is, of course, the fact that we do not know the value of σ^2 . Although we have a sample estimate s^2 we need to acknowledge that there is sampling variation with respect to this estimate on top of the sampling variation we have for \bar{x} .

A random sample of size 120 is drawn from a distribution, where the latter is known to be $N(\mu, \sigma^2)$. Notice that σ^2 has to be estimated with the sample variance, here $s^2 = 12$. Your sample mean is $\bar{x} = 40$. Then, the 99% confidence interval for μ is

$$\text{_____} \pm \text{_____} \cdot \text{_____} = [\text{_____}, \text{_____}]$$

Question 6: (1 point)

From the earlier table you will see that the t-distribution is only the right distribution to use if you know that the population distribution is actually a normal distribution. But as when we discussed hypothesis testing, the central limit theorem will ensure that a sample mean is normally distributed if the sample size is sufficiently large. In that case we will construct a $100(1 - \alpha)$ confidence interval from

$$\bar{x} \pm z_{\alpha/2} \sqrt{\frac{s^2}{n}}.$$

Consider the earlier example (using $n = 50$), but now not assuming that sampling takes place from a normal distribution. We assume that the sample information is $n = 50$, $\bar{x} = 18$ and $s^2 = 4.5$. Then, an approximate 95% confidence interval is (assuming that the sample size is sufficiently large for the CLT to do its magic)

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \sqrt{\frac{s^2}{n}} &= 18 \pm 1.96 \sqrt{\frac{4.5}{50}} \\ &= 18 \pm 0.588 \\ &= [17.412, 18.588].\end{aligned}$$

In general, how this compares with the exact confidence interval based on knowledge of σ^2 depends on how good s^2 is as an estimate of σ^2 . Nothing can be said about this usually.

A random sample of size 100 is drawn from a distribution of incomes. You recognise that you should not assume that these incomes are normally distributed. Notice that σ^2 has to be estimated with the sample variance, here $s^2 = 8$. Your sample mean is $\bar{x} = 39$. Then, the 95% confidence interval for μ is

$$\underline{\hspace{2cm}} \pm \underline{\hspace{2cm}} \cdot \underline{\hspace{2cm}} \Rightarrow [\underline{\hspace{2cm}}, \underline{\hspace{2cm}}]$$

Additional resources

- Salman Khan on such a confidence interval (<https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/small-sample-size-confidence-intervals>)
-

Question 7: (1 point)

Confidence intervals for population proportions

As for hypothesis testing you will learn that we can perform inference on many unknown parameters. Here we will merely introduce one extra such confidence interval, a confidence interval for a population proportion π . We assumed that a random sample is obtained from the distribution of a Bernoulli random variable, a random variable X taking on values 0 and 1 with probabilities $1 - \pi$ and π respectively. The sample mean \bar{X} here is the random variable representing the sample proportion of 1's, and so is usually denoted P or \hat{P} , "the" sample proportion. (Note, here we use \hat{P} to denote the sample estimator of the proportion and \hat{p} as the actual sample proportion. In some books and websites you will see P and p only representing these two objects. As the population proportion is usually labelled π this would be sufficient.)

When you learned about descriptive statistics you learnt that proportions can be seen as means of a variable which only takes values of 0 and 1. It is therefore not a surprise that the CLT works on sample proportions (\hat{P}) as well and makes them, for sufficiently large samples, normally distributed random variables.

Since

$$E[\hat{P}] = \pi, \quad \text{var}[\hat{P}] = \frac{\pi(1-\pi)}{n},$$

$$\frac{\hat{P} - \pi}{\sqrt{\text{var}[\hat{P}]}} \sim N(0, 1) \quad \text{approximately.}$$

So, one could hope to use this to provide an approximate confidence interval for π . There is a minor complication here in that $\text{var}[\hat{P}]$ depends on the unknown parameter π , but, there is an obvious estimator (\hat{P}) available which could be used to provide an estimator of $\text{var}[\hat{P}]$.

Following the previous reasoning, we argue that

$$\frac{\hat{P} - \pi}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \sim N(0, 1) \quad \text{approximately.}$$

This knowledge can then be used to construct a sample confidence interval from

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Note that this differed from the approach taken when we required the variance of \hat{P} when hypothesis testing. In that case we did not use the sample proportion \hat{p} but rather the hypothesised (under H_0) population proportion to calculate $\text{Var}[\hat{P}]$.

A random sample of 300 households is obtained, with 28% of the sample owning a tablet computer. An approximate 95% confidence interval for the population proportion of households owning a tablet computer is

$$\begin{aligned}
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.28 \pm (1.96) \sqrt{\frac{0.28(1-0.28)}{300}} \\
&= 0.28 \pm 0.0508 \\
&= [0.229, 0.331].
\end{aligned}$$

For such an apparently large sample size, this is quite a wide confidence interval. Better precision of estimation would require a larger sample size.

A random sample of 1021 randomly selected students is asked whether they intend to vote at the next election (in whichever country they are registered to vote). In the sample 714 students indicate that they intend to vote at the next general election in their country. Calculate a 80 % confidence interval for the population proportion of students intending to vote.

_____ $\pm 1.28 \cdot$ _____ $\Rightarrow [$ _____ , _____]

Additional resource

- This is the Khan Academy example of a confidence interval for a proportion (<https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/confidence-interval-example>) . The standard error of \hat{P} is calculated in a different way but with the same result.

Question 8: (1 point)

Opinion polling reality

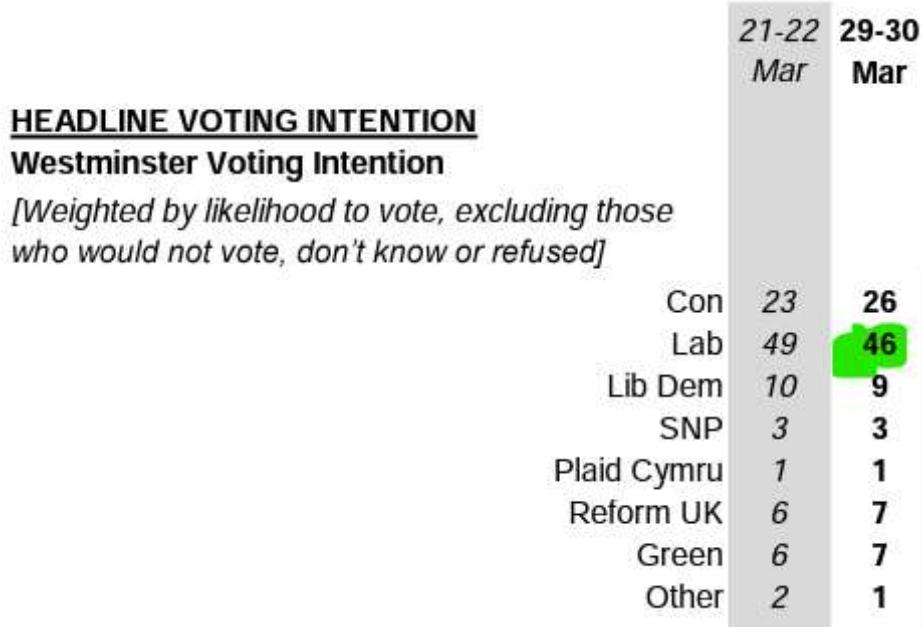
Estimating proportions is a very common task. Market research companies will often do this on behalf of clients. One of the companies which is busy in this sphere is YouGov (<https://yougov.co.uk/>). Visit their website to see the wide variety of research they undertake. One of the regular pieces of research undertaken is an investigation of voting intentions. Here is a link to the article (<https://yougov.co.uk/topics/politics/articles-reports/2023/03/31/voting-intention-con-26-labour-46-29-30-march-2023>) to the survey on voting intentions from 31 March 2023. You will remember from earlier discussions that, when sampling, we would normally assume that we have taken a random sample from the population of interest. One interesting aspect of such work is that organisations like YouGov will often reweigh the observations. This means that not all observations will get the same weights. We dealt with this in the descriptive statistics section. The reason why this is done is to ensure that the investigated sample has demographic properties that the population of interest is assumed to have.

In this document (https://docs.cdn.yougov.com/8iqsf1dct/TheTimes_VI_230330_W.pdf) we can get survey details for the 31 March 2023 YouGov voting intention survey. In there you can see (on the bottom of page 4) that YouGov targets (amongst others) the following characteristics

Brexit Vote	Target percentage	Actual obs	Actual percentage	Weighted obs
Remain	35.2%	801	40.0%	705
Leave	37.2%	768	38.4%	745
Did not vote / Did not know	27.5%	433	21.6%	551

There were 2002 responses and for instance there were 801 respondents that voted remain in the Brexit vote. That is an actual percentage of 40% which is more than the 35.2% targeted. Therefore, the weights of these observations are down-weighted such that they effectively represent only 705 observations (approximately the target percentage). Equally, there were too few respondents who did not vote and therefore these observations receive a larger weight.

Below you can see a snippet from the resulting voting intention.



As you can see YouGov estimates that 46% of votes will go to the Labour party. When performing a hypothesis test or calculating a confidence interval you will need the standard error for the sample proportion \hat{P} . We used, in the previous example

$$\sigma_{\hat{P}}^2 = \frac{\hat{p}(1 - \hat{p})}{n}$$

Unfortunately, when using weighted observations as in the YouGov example, this formula is not correct, as it does not take into account that some observations had a higher weight than others. How to calculate the standard error for a weighted sample proportion is beyond the scope of this. For the example below use $\sigma_{\hat{P}}^2$ as introduced above, it will continue to be in the right ballpark.

Calculate the 95 % confidence interval for the proportion of Labour votes.

$$\underline{\hspace{2cm}} \pm 1.96 \cdot \underline{\hspace{2cm}} \rightarrow [\underline{\hspace{2cm}}, \underline{\hspace{2cm}}]$$

Question 9: (0 points)

Summary

In this section we discussed confidence intervals. They are a very powerful and popular way to communicate that any estimate for a population quantity (say the mean or proportion) is merely an estimate and that really we are not certain. As the p-value in hypothesis testing, the confidence interval has a slightly "stilted" interpretation. However, confidence intervals are a very intuitive way to communicate the uncertainty or margin of error.

Having said that, often the communication opinion poll results (say in TV, TikTok or newspapers) does not acknowledge such uncertainty. Here is an example of typical reporting on opinion polls. An institution (here Quinnipac University, well known for their polling operation) conducted some opinion polls (<https://poll.qu.edu/poll-release?releaseid=3863>) and presents a write-up of the results. At the bottom of that report you do get a clear statement about the margins of error (typically calculated as 95% confidence intervals). Then some other organisation picks up on these poll results and does not mention any margin of error (here the respectable Brookings Institute (<https://www.brookings.edu/blog/fixgov/2023/01/06/polls-show-americans-are-divided-on-the-significance-of-january-6/>)). This is a very typical pattern which you can confirm by checking how another Quinnipac poll on the conservative US Supreme Court judge Thomas (<https://poll.qu.edu/poll-release?releaseid=3842>) was reported on by CNN (<https://edition.cnn.com/2022/04/06/politics/justice-clarence-thomas-wife-ginni-poll-recusal/index.html>).

In the UK, the BBC actually does have guidance on how opinion polls are to be reported (<https://www.bbc.com/editorialguidelines/guidance/surveys#reportingopinionpolls>):

Reporting Opinion Polls

This guidance applies whether we are reporting on polls the BBC has itself commissioned or on polls commissioned by other organisations, especially if they are members of the British Polling Council. We should always make clear who has carried out the poll and who has commissioned it (as well as giving information about the size and nature of the sample, the margin of error and the dates of the fieldwork).

But that does not ensure that this guidance is always adhered to as reading through this BBC news story confirms (<https://www.bbc.com/news/world-us-canada-44324545>).

What should you take away from this? Reporting on statistical uncertainty is not easily done correctly (recall the awkward interpretations of p-values and confidence intervals). It is therefore perhaps not surprising that many news organisations shy away from such communication. It is therefore some of your skill, after having studies on this course, to properly understand such uncertainty and contribute to communicating it properly.

Here are two additional pieces of interesting reading in this area.

- This is an interesting report by the UK Parliament on the way Media report on polling results. (<https://publications.parliament.uk/pa/ld201719/ldselect/lppdm/106/10606.htm>)
- This is an interesting investigation by YouGov on the probabilities the public associate with terms like "very likely" or "almost impossible". (<https://yougov.co.uk/topics/politics/articles-reports/2020/10/01/how-likely-likely>)