Chapter Eight: Estimation

Simona Simona

4/12/2020

In the previous chapter, we indicated that the two main areas of inferential statistics are parameter estimation and hypothesis testing. We are now coming back to deal with these two and we do so in this chapter and in the next. This chapter deals with parameter estimation, which we said entails taking the sample statistic such as the sample mean and using it to say something about the population parameter such as the population mean. The question we are trying to answer in estimation is "what is the population mean?". As we have seen in the previous chapter, we don't usually know population parameters and so, we use sample values to estimate them. Based on the sampling distribution, we know that there are many possible samples that could have been collected from the same population. Our aim is to find the best approach to estimate population parameters on the basis of the sample values.

A specific sample value that we use to estimate the population value is called an **estimate**. There are two types of estimates namely **point estimate** and **confidence interval estimate**. We will discuss point estimates first and then confidence intervals.

Point Estimates

For means, the mean that we calculate from the sample is the point estimate of the population mean. That is to say that when we obtain a sample randomly from the population and we want to compute the population mean for a particular variable, we calculate the mean of the concerned variable from the sample and use it to represent the mean of the variable in the population. The sample mean $\bar{x}$ is the point estimate of the population mean μ. There are many other estimators such as the sample standard deviation s which is the point estimate of the population standard deviation σ. The sample variance s^2 is the point estimate of the population variance σ^2.

For an estimate to properly represent the population value, it needs to have two properties: **unbiasedness** and **efficiency**. A sample statistic is an unbiased estimator of the population parameter if the expected values of the sampling distribution of the statistic is equal to the population mean. In other words, the mean of the repeated samples from the population should be equal to the population mean. We know that this is in fact true because we spent the previous proving this to be the case using the properties of the sampling distribution and the central limit theorem. Efficiency is the extent to which the sampling distribution clusters around the mean. Clustering is a function of the standard deviation of the sampling distribution, which is the standard error. In other words the smaller the standard error, the greater the clustering and the higher the efficinecy. Smaller standard error mean that the sample statistics will be closer to the popultion mean compared to those with larger standard errors. Statistics has proven that for the mean, sample estimators are unbiased and efficient. We just need to remember that larger sample sizes and more likely to have greater efficiency and unbiased.

Point estimates have the disadvantage of not being very reliable. This is because they enlist only one value to represent the population prameter. For example, imagine that you asked 10% of student population for the yearly income and concluded that their average income is £12000 based on the income of the sample. There is greater uncertanty in this value being the average income for the student population of sampling variability. As we know, our sample is but one of the many possible samples that can be drawn from the population. Therefore £12000 would probably be very close but it may not be exactly the value of the population. To deal with the uncertanties in point estimates, we are much safer if we gave a range of values in which your estimated average in likely to fall. That is what confidence intervals bring.

Confidence intervals

Confidence intervals (CI) or interval estimates are a range of values likely to include the true value of the population parameter being estimated, with some degree of confidence. Confidence intervals are more difficult to compute but are a better representation of reality compared to point estimates. They allow us to estimate parameters withing a confidence interval of values between the **lower bound** and the **upper bound**. For example, instead of saying the yearly student income is £12000, we would look smater if we said, we expected the yearly student income to be between £11000 and £13000. Notice that confidence intervals put the point estimate in between. But there is still some uncertainty in this calculation. That is, we are not really 100% sure that our population means lies between those two bounds. We would be even more accurate if we supplied our level of confidence in the the range of expected yearly student income. For example we would say we are 95% confident that the yearly student income falls between £11000 and £13000.

The degree of confidence (95% in our example above) is called **confidence level**. Sometimes we don't talk in terms of confidence levels, but rather in terms of the amount of risk we are willing to tolerate of being wrong. Being wrong in this case means our interval estimate does not include the population mean. If we are 95% confident that the yearly student income is between £11000 and £13000, it means there is a 5% chance that this prediction is incorrect, that the true student yearly income average falls outside this range, either it is smaller than £11000 or bigger than 13000. This risk of being wrong is called the **alpha level** and it is represented by a greek letter alpha (α). See the figure below

We have a few options about confidence levels. We can set them at 90%, 95% or 99%. We must however, be aware that there us a tradeoff. If we have a higher confidence level of capturing the population mean, then we have less risk of missing it. Meaning our apha level will be small. A lower confidence level results in greater risk of wrong prediction. For isntance, 99% confidence, will result in only 1% chance of wrong estimation while 90% confidence gives us 10% risk.

Deciding the confidence level or alpha level is usually the first step in computing confidence intervals. In the social sciences, we most often use 95% confidence. After deciding the confidence level, the next step is to determine the boundaries of the confidence interval. In doing so, we will use the standard normal curve. Our upper bound will be on the right hand side of the mean and the lower bound will be on the left. We will need several things. First is the population mean itself for us to calculate its boundaries. But of course we don't have it, so we approximate it using our sample mean with some margin of error. Secondly we need the standard error, which is the standard deviation of the sampling distribution. We also don't have it but statisticians have told us we can approximate

that using the standard deviation of the sample and the sample size

$$\frac{s}{\sqrt{n}} \approx \sigma_{\bar{x}}$$

where s is the standard deviation, n is the sample size, and $\sigma_{\bar{x}}$ is the standard error of the mean. Thirdly we need the z-sore values within which 95% of cases fall. We know this already. Remember in the last chapter we came accross one of the properties of the normal curve which said approximately 95% of the area under the curve falls between -2 and +2 standard deviation from the mean. Bingo! so, the z-score we are looking for is 2. But we need the exact value not an approximated one. We will use the Z-distribution table to find the exact values between which 95% of the area under the curve is bounded. You can do Google search of "standard normal distribution". Most tables will report values for the proportion of the area of the curve to the left of the Z-scores. In that case, the proportion you are looking is up to the upper bound and you get that by sabracting 0.025[1] from 1 and you get 0.975. Now carefully scan the table to find 0.975 and record the z-scores associated with it beginning with the row and then column. If you have done it well you should get 1.9 on the x-axis and 0.6 on the y-axis. Combining them gives us 1.96.

We can also do this very quickly in R by using the function `qnorm()`. It gives the z-score associated with the given probability density function (PDF) which is given by supplying 0.975 as an argument in the `qnorm()` function as follows:

```
qnorm(0.975)
```

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

We are done with our ingrediants for calculating the confidence interval. The equation below calculates the the confidence interval:

$$\mu = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

The only symbol we haven't yet met in the above equation is z^* which represents the value of Z associated with the confidence level we desire for our estimate. We know this value to be 1.96. Before we look at examples to ground our understanding, let's briefly explain the equation above. We start with the sample mean which approximates the population mean, and then proceed with two separate equations in one. For the lower bound, we sabtract the product the z-score and approximated standard error from the sample mean, and the upper bound is given by using the addition operator in the same equation. Hopefully, we have not forgotted that there is still a chance, albeit small (0.05 or 5%) of not capturing the true value of the mean. Since the normal curve is symetrical, the probability of the true mean being outsise the lower bound is 0.025 and it is by the same probability that the true mean is above the upper bound.

Examples Lets start with the question we have been working with so far. Supposed you wanted to find the amount of the yearly earnings of the University of Glasgow student population as part of your term assignment. You were only able to interview a random sample of 300 students and when you calculated the mean from the earnings variable, you got £12000 with a standard deviation of £4600. How do you calculate the 95% confidence interval for population mean?

[1] Because our alpha level is 5% and the the normal curve is symetrical. We divide the apha by 2 to get the remaing area to the right of our normal curve. Which would be 0.025

Answer

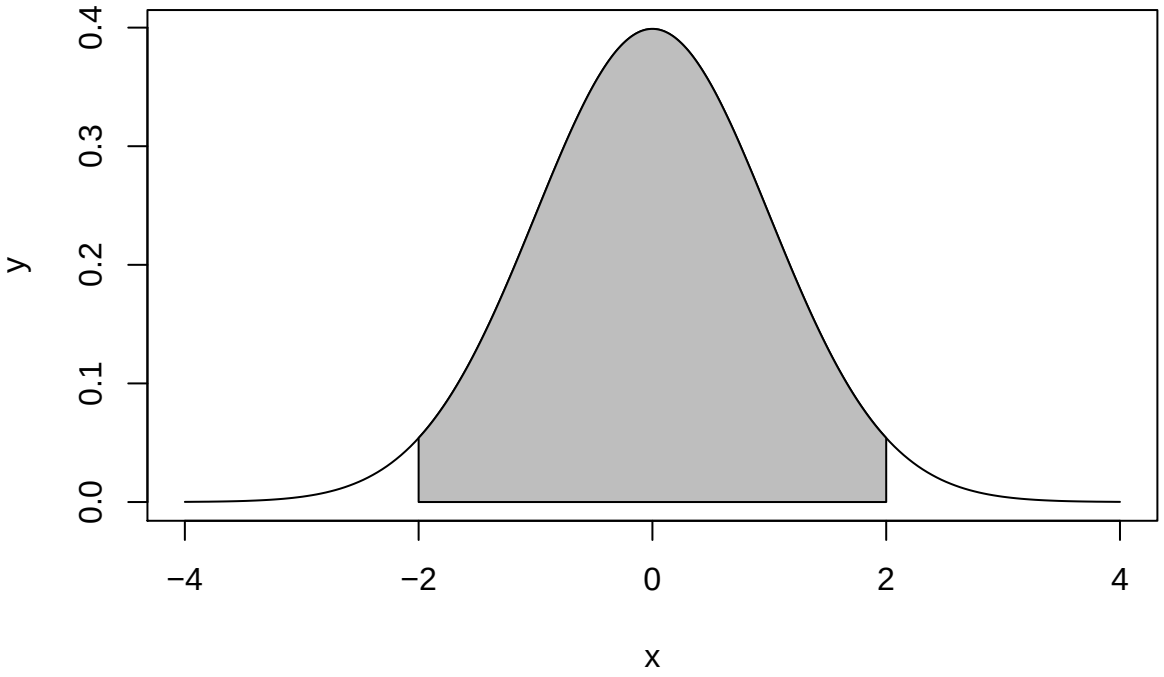
The first thing you need to do is summerising the information we have been given in the problem and relating it to the equation you will use to solve the problem. In this case you have:

$n = 300$
$\bar{x} = £12000$
$s = £4600$

As we indicated above, the first step is for you to select the level of confidence. You have already been given this in the question. So you are now going to the second step of working out the boundaries of the confidence intervals using our equation:

$$\mu = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

Step three is to find the values of z^* which puts 95% cases in the middle of the standard normal curve. This can mathematically be written as $\text{Prob}(-z^* < z < z^*) = 95\%$. The figure below shows the boundaries we are looking. We have already calculated this z-score value to be 1.96.



You will now need to input the values in the equation as follows:

$$\mu = 12000 \pm 1.96 \frac{4600}{\sqrt{300}}$$

$$= 12000 \pm 520.54$$

$$\mu = 11479.46, 12520.54$$

Therefore, we are 95% confident that the yearly income among the student population is between £11,479.46 and £12,520.54. You can also say that the error associated with using the sample as an estimate of the population is 520.54. You can see that the range of between £11000 and £12000 we

have been working with was not very accurate. It turns out the confidence intervals are not as big.

Example 2

You are a member of a research group studying the prevalence of fake news posts on social media. You devise a research strategy that interviews 670 members of the public asking them how many fake posts they have seen over the past one month. Upon collecting and analysing the data, you discover from your sample that the average number of suspected fake news materials that a person sees in a month is 159 with a standard deviation of 63. Calculate the 90% and 99% confidence interval of the population mean.

Answer

We have two questions in one. The first one requires us to calculate confidence intervals using the 90% confidence level while we need to use the 99% in the second part of the question. We will address the questions one by one beginning with the 90% confidence level. As usual we begin by stating what we have:

$n = 670$
$\bar{x} = 159$
$s = 63$

The answer to this question follows the steps that we used in the first example. We are given the confidence levels we need to use and therefore we begin by finding the z^* associated with each of the confidence levels. Since we already how to calculate z-scores manually, we use the `qnorm()` function in R to work them out fast. Remember that what R gives us is the PDF associated with the given z-score.

```
qnorm(0.95)
```

```
## [1] 1.644854
```

```
# The density curve showing the PDF comes here!
```

Now you can apply the equation to get the equation we are now familer with

$$\mu = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

Big sample size versus small sample size

Confidence intervals in R