6 CONFIDENCE INTERVALS

6.06 Choosing the sample size

Okay, I'm going to investigate how much sleeping hours parents in Amsterdam lose in the first weeks after the birth of their first child. Because it's impossible to ask this question to *all* new parents in Amsterdam, I'm going to draw a simple random sample. One crucial question then is how large this sample should be. Is a sample of 50 respondents sufficient? Or do I need at least 300 or maybe even 1000 respondents? This video is about choosing the sample size of a study. I'll talk about situations in which we are interested in means and about situations in which we are interested in proportions.

Let me start with the sample size when it comes to means. The choice for your sample size depends on three main factors. The first one is how precise you would like to be. Remember that a confidence interval is computed by adding to and subtracting from a point estimate a certain margin of error. How large do you allow this margin of error to be? The smaller you want the margin of error to be, the larger your sample size should be. Second, your sample size also depends on the confidence level you want to use. With a larger confidence level you also need a larger sample size. Finally, the choice of you sample size depends on the variability in your data. The larger the standard deviation of your variable is, the larger your sample size should be.

This is displayed by the following formula. The size of your sample for which a confidence interval for population mean Mu has a margin of error M equals the standard deviation in your population squared, multiplied with the z-score corresponding to your chosen significance level squared, divided by the margin of error you allow, also squared. Of course, you don't know the population standard deviation. And because you don't know your sample standard deviation either (you have, after all, not drawn your sample yet), you need to estimate this value by means of an educated guess.

Let me show you how you could do this in practice. I am going to investigate how much sleeping hours new parents in Amsterdam lose after the birth of their child. Now imagine that I want the confidence level to be 95 percent. The z-score corresponding to this confidence level is 1.96. I also don't want the margin of error to be wider than 0.3 hours. M thus equals 0.3. We can complete this part of the formula and this part of the formula. Now we have to come up with an educated guess of Sigma. If there is an existing study in which the variable 'lost sleeping hours' is included, and you know the standard deviation of this variable, you can simply use this value to estimate Sigma. However, if such a study doesn't exist, we have to come up with an educated guess. I assume that some parents won't sleep less at all, and that some of them will sleep about 5 hours less. Yet, I don't expect many parents to sleep *more* than they did before they got their first child, and I also don't expect many of them to lose more than 5 hours of sleep. So, if we assume that the variable is normally distributed and that 95% of the new parents sleep between 0 and 5 hours less, the distribution in the population would look like *this*. The mean is 2.5 hours and the standard deviation is 1.25 hours. After all, 95% falls within about 2 standard deviations of the mean. 2 standard deviations equals 2.5, so one standard deviation equals 1.25. If we now complete the formula, this is the result: 1.25 squared times 1.96 squared, divided by 0.3 squared. That's about 66.69. We round up, and conclude that we need at least 67 respondents.

We can also do this if we are interested in a proportion instead of a mean. Suppose I want to know the proportion of babies that like to poo while their diaper is being changed. I therefore ask a simple random sample of new parents if their babies like to defecate during the diaper-changing process. How large should my sample size be if I want to work with the 99% confidence level and I want to allow a margin of error of 0.10? Well, the formula looks very much like the previous formula. *Here* it

is. The sample size for which a confidence interval for a population proportion π has a margin of error M equals: p times one minus p times the z-score corresponding to your chosen significance level squared, divided by the margin of error you allow, squared as well. We know the values of M and z. They are 0.10 and 2.58 respectively. (Note that the value 2.58 comes from the z-table; it is the z-score corresponding to the 99% confidence level.) What we don't know is the value of p. Again, if you can make a guess based on previous research with the same variable you can use the value of p coming from this previous study. If not, you should make an educated guess as I did before. Or… you could go for the so-called 'safe approach'. This is how you do that. In the formula you can see that the sample size depends on the value of p multiplied with one minus p. The largest possible value this multiplication could take is 0.25, and that only happens if p (and therefore also 1 minus p) equals 0.5. Just try it. We can now complete the formula: 0.5 times 0.5 times 2.58 squared, divided by 0.10 squared. That equals 166.41, which makes 167 respondents.

Such computations can be very important. In an ideal world, you would just go for a super large sample of at least 1000 respondents or so. However, in the real world, we have to deal with limited time and often we don't have enough money to draw very large samples. Computing how large our sample should be in the way I just did, can help us keep the costs to a minimum. On the other hand, it also avoids conducting an experiment or survey where it would be clear from the onset that for a given sample size and variability, you would never be able to estimate a statistic with the desired margin of error.