5 SAMPLING DISTRIBUTIONS

5.06 Sampling distribution proportion

Okay. Imagine you're living in Paris and you know that a proportion of 0.10 of all students identifies as a Hipster. You would like to know what the sampling distribution of this proportion looks like. Note that it doesn't make much sense to compute the population mean here. Your variable of interest is a binary nominal variable: students can choose if they identify as a hipster or not. Means are not relevant with such a binary variable. In this video I'll explain what the **sampling distribution of a population proportion** looks like.

You know that 10 percent of the students in Paris identifies as a Hipster. This means that the population proportion, which is symbolized by Pi, equals 0.10. Now imagine that we draw a sample of n equals 200 from this population. The sample proportion, which is symbolized by p, will be a number close to 0.10. It could be, for instance, 0.09, or 0.12. If you would draw 5 samples, the sample proportions might look something like *this*. A histogram of these sample proportions would look like this. There are 5 values and they all occur once, so they all have a probability of 0.2. Now, if you would draw 25 samples, the distribution would look something like *this*. If you would draw 50 samples something like *this*, and with an infinite number of samples your distribution would look like *this*.

This is the sampling distribution of the sample proportion and the mean of this distribution will be 0.10, which equals the population proportion. To show that we are dealing with the mean of a samp*ling* distribution, the mean is symbolized by Mu-p. We add the p to show that we are dealing with the sampling distribution of the sample proportion in which the scores are not scores of individuals, but sample proportions. As you can see, the exact same logic applies as in the case of the sampling distribution of the sample mean.

In the case of the sampling distribution of the sample *mean*, the distribution is approximately bell-shaped if the population itself is normally distributed or if the sample size is sufficiently large (usually 30 is taken as a minimum). In the case of the sampling distribution of the sample *proportion*, you can only be sure that the distribution is bell-shaped if you have at least 15 positive cases and 15 negative cases, so at least 15 hipsters and 15 non-hipsters. You can express this formally as: $n\pi \geq 15$ and $n(1-\pi) \geq 15$. What does that mean for our example?

First, the product of the sample size and the population proportion should be 15 or more. In our case that's 200 multiplied with 0.10. That equals 20 hipsters and 20 is larger than 15. Second, the product of the population proportion and 1-Pi should be 15 or more. In our case that's 200 times (1-0.10) = 200 times 0.90 = 180, so 180 non-hipsters. We can conclude that the sampling distribution will be bell-shaped because 20 and 180 are both larger than 15.

There is also a pretty straightforward formula to compute the standard deviation of the sampling distribution of the sample proportion. We symbolize the standard deviation by Sigma-p. You know that Sigma stands for the standard deviation. P is added to show that we are talking about the sampling distribution in which the cases are sample proportions, and not individual subjects.

To compute the standard deviation we have to multiply the population proportion with 1 minus the population proportion. Next, we have to divide the outcome by the sample size n and take the square root of the outcome. In our case this boils down to the square root of 0.10 multiplied with 1-0.10 divided by 200. That equals 0.02. The standard deviation of the sampling distribution is 0.02.

So, to conclude. When it comes to binary categorical variables, it doesn't makes sense to compute the population mean or standard deviation. Instead, we compute proportions when we are dealing with categorical variables. So, when it comes to binary variables, we only have the population proportion Pi. A similar logic holds for the sample: we only have the sample proportion p. Yet, when it comes to the samp*ling* distribution of the sample proportion, we *do* have a mean and a standard deviation. These values can easily be computed as long as we know the value of the population proportion.