5 SAMPLING DISTRIBUTIONS

5.01 Sample and population

Almost all statistical studies are based on samples. Imagine you want to know to what extent students in London identify themselves as Hipsters. It is almost impossible to ask all students, so you decide to draw a sample of, say, 200 respondents, and you assess to what extent they see themselves as Hipsters. The great thing about statistics is that it can help you to draw conclusions about *all* students in London (which is the **population**), based on an analysis of *only these 200* respondents (which is the **sample**). In this video I'll explain the difference between sample and population in a little more detail.

If you have selected a sample of 200 respondents from a population of about 300.000 students (which is the total number of students in London), you basically focus on a subset of your population. If you measure a couple of variables, like gender, age, attended university, etcetera, you can do all kinds of computations. You can do univariate analyses and compute modes, means, and standard deviations. You could also do bivariate analyses and compute Pearson's r correlation coefficients or do regression analyses. All numerical summaries resulting from these computations are fully based on your sample and they are called **statistics**. In general, the methods for summarizing sample data are called **descriptive statistics**.

However, in the actual research practice, we are often not so much interested in summaries of a specific *sample* (in our case the 200 selected students), but our real goal is to make statements about the entire underlying population (so, in our case all 300.000 students in London). If we are employing the data obtained from a sample to draw inferences about a population, we are using methods of **inferential statistics**. We use the computed statistics to draw conclusions about the corresponding population **parameters**. Statistics are displayed by Roman letters. For instance, $\bar{x}$ is the mean and s is the standard deviation in a sample. Parameters, however, are displayed by Greek letters. μ stands for the mean in a population and σ for the standard deviation in a population.

Imagine you ask 200 respondents to what extent they see themselves as Hipsters. They could indicate their self-perceived 'Hipsterness' on a scale from 0 to 10. 0 means that someone doesn't see him- or herself as a Hipster at all, and 10 indicates that a person fully identifies him- or herself as a Hipster. This is an ordinal variable, measured on a scale of *more* than 10 categories, so we will treat this variable as if it were a quantitative one. Now imagine that the mean 'Hipsterness' score in the sample is 3.12.

The central question now is what the mean 'Hipsterness score' in the wider population is. You know the relevant statistic in your sample ($\bar{x}$ equals 3.12), but what you *actually* want to know is what the mean 'Hipsterness' score in the wider population is. You want to know, in other words, the value of population parameter μ. Methods of inferential statistics can help us to answer such questions.
So, if you want to know more about that, or about Hipsters, watch the videos in this module carefully!