5 SAMPLING DISTRIBUTIONS

5.05 Three distributions

Many social, political and religious groups have their own sacred texts. Hipsters also have an almost holy canonical text. It is a book called "On the Road". We are interested in the question how much time Hipsters in New York have spent reading "On the Road". Assume we know that in the entire population the mean time Hipsters have spent reading the book is 943 minutes. You also know that the population standard deviation equals 212 minutes. You have drawn a simple random sample of 200 respondents from this population. The mean reading time in this sample equals 867 minutes. The standard deviation in the sample is 188 minutes. In this video I will talk about three distributions that are of importance to our research project: the population distribution, the data or sample distribution and the samp*ling* distribution. I will also show you how you can compute the probabilities of selecting individuals with particular scores and samples with particular sample means from this population.

As said, three distributions are of importance here. The first one is the population distribution. It looks like *this*; it's approximately bell-shaped. I've already told you what the mean is: 943. The subjects in the population are *all* New York Hipsters. So, the mean means that if we add up the "On the Road - reading times" of all New York hipsters and divide that by the total number of New York Hipsters, we get a value of 943 minutes. The population mean is symbolized by Mu. The standard deviation in the population, symbolized by Sigma, measures the variability of all the individual reading times in the population around the mean reading time. It equals 212 minutes.

The second distribution is the data or sample distribution. It is the distribution of the sample data. *This* is what it looks like. It is, just like the population distribution, approximately bell-shaped. The cases here are the 200 sampled respondents. The mean of this distribution is symbolized by X-bar. It equals 867 minutes. As you can see, this mean is not very far removed from the population mean of 943 minutes. The standard deviation in the sample is 188 minutes. The mean and standard deviation are symbolized by Roman characters because they are sample statistics. The mean and standard deviation in the population are symbolized by Greek symbols because they are population parameters.

The third distribution is the samp*ling* distribution of the sample mean. And *this* is what it looks like. Because of the Central Limit Theorem it is normally distributed. The cases in this distribution are not individuals, but an indefinite number of samples of 200 respondents from our population of New York Hipsters. The mean of the sampling distribution of the sample mean is the mean of these infinite sample means. The value of the mean of the sampling distribution equals the mean of the population distribution which was 943 minutes. To show that we are talking about the mean of the samp*ling* distribution, we add X-bar to indicate that it is a mean of sample means, and not a mean of individual scores. The standard deviation of the sampling distribution equals the standard deviation of the population divided by the square root of N. That makes 212 / SQRT(200) = 15.

You need to remember that this third distribution, the sampling distribution, is a *theoretical* distribution. We don't actually collect an infinite number of samples. That's not even possible. It's not necessary either, because we now what the sampling distribution looks like, as long as we know the values of the mean and standard deviation in the population.

The nice thing about normal distributions is that we can find probabilities by changing original scores into z-scores and by employing the z-table. Now, if we would like to know what the probabilities are

of selecting random samples or subjects from a population, we can apply this logic to sampling distributions, and, as long as they are normally distributed, to population distributions.

Now imagine you select a random Hipster from the population. What is the probability that this Hipster has a reading time of 1000 minutes or more? Well, first we want to know how many standard deviations a person with a reading time of 1000 is removed from the mean. So, we have to compute this person's z-score in the population. *This* is the formula, so the z-score is 1000 – 943 / 212 = 0.27. We're interested in the area to the right of this value. If we look it up in our z-table we find that the chance of selecting a person with a reading time of 1000 minutes or more is 39 percent.

Now imagine we draw a simple random sample of n = 200 from the population. What is the probability that the sample mean is 1000 minutes or higher? Now, pay close attention, this is a completely different question. We're not talking about selecting *a specific person* from the population. Instead, we're talking about a statistic, based on *a specific sample* from the population. Therefore, we don't make use of the *population* distribution, but of the *sampling* distribution of the sample mean.

In general, the procedure is the same, but now we make use of other means and standard deviations. So *this* is how we compute the z-score now. We subtract the mean of the sampling distribution of the sample mean from the sample mean we're interested in. That makes 1000 – 943. We divide by the standard deviation of the sampling distribution. That's Sigma divided by the square root of n. That is 212 / SQRT(200). That's 15. So 1000 – 943 / 15. That gives a z-score of 3.8. If we look it up in our z-table we find that the chance of drawing a sample with a mean reading time of 1000 minutes or more is 0.01 percent.

So, you need to be very careful when deciding which distribution to use. If you're interested in selecting individual subjects, you should use the population distribution. But if you're interested in selecting samples you should use the sampling distribution. In the actual research practice confusion between a population and a sampling distribution will almost never occur because you wouldn't know what your population looks like. The only thing you know is what your sample looks like. Later on you'll see how we can make use of the sampling distribution *without* information about the population distribution!