

# Lecture 6: Samples, Populations, and Distributions

In this lecture we'll learn about samples and populations and how samples and populations can be described with distributions.

---

AUTHOR

Dr Lincoln Colling

AFFILIATION

University of Sussex

PUBLISHED

Nov. 4, 2021

---

## Contents

### Samples, populations, and distributions

- Samples and populations

  - The relationship between samples and populations

- Distributions

  - The binomial distribution

  - The normal distribution

    - Describing the normal distribution

    - Processes giving rise to normal distributions

    - Processes that don't produce normal distributions

    - Describing deviations from the normal distribution

- Distributions and samples

## Samples, populations, and distributions

This lecture is made up of two parts. **Part one** will introduce the concepts of **populations** and **samples** and how they are related to one another. In **Part Two**, we'll learn about **distributions** and about how natural processes can give rise to particular kinds of distributions. In particular, we'll learn about how natural phenomena can give rise to the **normal distribution**, and we will lay some of the groundwork for understanding the **sampling distribution**, which will be introduced in a later lecture. Finally, at the end of **Part Two**, we'll try to put everything together, and we'll look at the relationship between **samples**, **populations**, and **distributions**.

---

# Samples and populations

---

One of the key things that we want to do with *statistics* is to make **inferences** about **populations** from *the information* we get from **samples**. That is, we often want to make a judgement, or draw a conclusion, about an aspect of the population when all we have access to is a sample.

We'll get to more formal definitions of *populations* and *samples* shortly, but first, let's make things more concrete by introducing an example.

Let's say you're interested in the **average height** of **people in the UK**. The "easy" way to find an answer to this question is to measure **all the people in the UK** and then work out the **average height**. Doing this will give you the exact answer to your question. But if you can't measure everyone in the UK, then what do you do?

One option is to select a smaller group, or subset, of people from the UK. You can then measure the height of people in this group, and then try to use this information to figure out plausible values for the average height of people in the UK.

In this example, the group (or groups) you're making claims about is the population. You want to claims about **the average height** of **people in the UK**. And the **sample** is a subset of this population—the smaller group of people that you were eventually able to measure.

It's important to note that there isn't a **single** population. What counts as the population will depend on the claim you're making. For example, let's say I'm interested in testing the claim, "Do **people in East Sussex** show an interference effect on the Stroop task?". Here the **population** would be **people in East Sussex**. If, however, I want to make claims about **people in general**, then the **population** might be **all living humans**. The **sample** is always going to be a subset of the **population**.

## WEIRD samples

+

## The relationship between samples and populations

Let's assume that we have explicitly defined our **population** (for example, as *all people in the UK*) and we've collected a **sample** by taking measurements from a **subset** of this population. What is the relationship between this sample and the population from which it was drawn?

The *sample* should **resemble** the *population* in some way. Most often we're interest in making **inferences** about **averages**—for example, **average** performance or **average** score on a measure or a test. In the example I introduced earlier, we were interested in **average height**. But we might also be interested in things a difference between two averages—for example, whether there is a difference in **average depression levels** before and after some intervention, or whether **average response times** are different between the two conditions of a Stroop task. Ideally then, the **average height** of our **sample** should **resemble** the **average** height of our **population**, or the *average response time difference* in our **experimental sample** should *resemble* the *average response time difference* in our population. But if we don't know the **average** of our **population**, then how will we know whether our

**sample** *resembles* it?

In short, **we can't know for sure**. But we can think of a couple of things that will **influence** the relationship between our **sample** and the **population**. To figure out what these are, let's do a thought experiment and think of some **extreme cases**.

First, consider the case where **all the members** of a *population* are **identical**. If this were the case, then our **sample** will have an **identical** average to the population. The height of one person would be the same as the average height of two people, which would be the same as the average height of 100 people, which would be same as the average height of the population because people only come in one height. But if the **members** of the **population** are all **different** from one another, then there is no guarantee that the **sample's average** will **resemble** the **population's average**.

The second extreme scenario is if our sample is **very large**. Let's say that it is so large that it includes **all the members of the population**. If this were the case, then, by definition, our **sample average** would be **identical** to the **population average**. However, if our sample is smaller than the entire population, then once again, there is no guarantee that the **sample's average** will **resemble** the **population's average**.

Based on this reasoning, we can say that two things will influence whether your *sample* resembles your *population*. These are 1) the amount of **variation** in our population, and 2) the **size** of our sample.

Importantly, however, and barring the extreme cases above, for **any particular sample** we won't know whether it **resembles** the population or not, because, remember, we don't know the average of the population. Instead, we should think about these two factors as influencing **how likely** it is for samples to resemble the population. But what does this mean?

One way to think about this is in terms of **repeatedly** taking samples from the same population. For example, if we take a large sample from the population—large, but not so large as to include the entire population—then we can't say that our **particular** sample will resemble the population. But if we take many samples (of that size), then we can say that **on average** those samples will be closer to the population than would be the case for a collection of smaller samples.

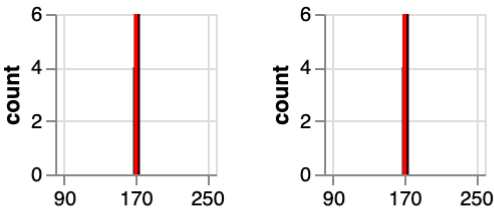
The same reasoning applies to **variation in the population**. If there is **less variation** in the **population**, then the samples drawn from that population will tend to be closer to each other and closer to the population average. But again, we won't be able to say whether **a particular sample** has an **average** that is close to the population average.

Of course, sample size and population variation exert their influence together. If we want our sample averages to be close to the population average, then we need samples that are **big enough**, but what counts as **big enough** will depend on the **population variation**. Therefore, knowing whether our sample is big enough depends on knowing the population variation. Unfortunately, we don't know this; however, there is a way to estimate it. But that's a topic for another lecture.

The example in Box 1 allows you to explore the relationship between samples and the population. You can make adjustments see to how the relationship is influenced by sample size and **variation in the population**.

In this example, the population average is **170**. On the left side of Box 1, there are **frequency plots** or **histograms** of nine different samples that have been drawn from the population. The **average** of **each sample** is shown with a red vertical line. You can make two adjustments using the options on the right. You can change the **size of samples**, and you can change how **similar the members of the population are** to each other. Click **Start** to start drawing samples from the population. Make adjustments to the **sample size** and the **population variation** and look at the **averages of the samples**. Some settings produce samples that look very different from each other, and some settings produce samples that look very similar to each other.

In the lecture next week, you'll learn more about how to calculate the average and the spread of a sample.

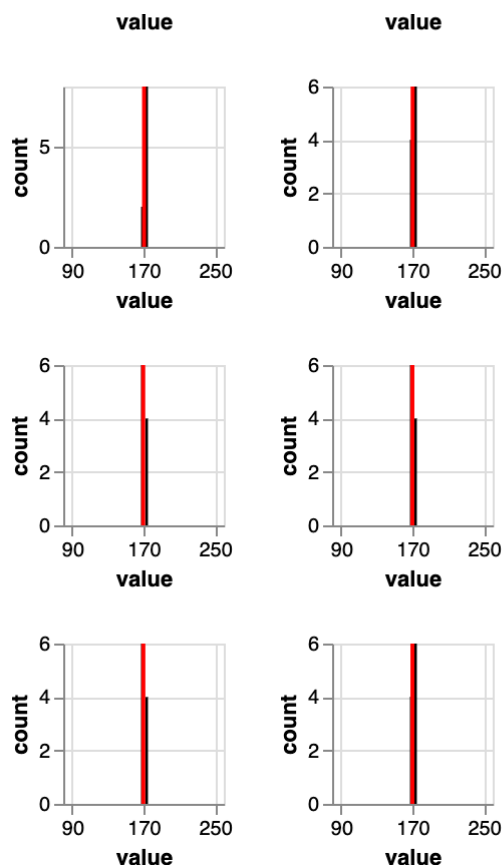


How big is the sample?

Small

Medium

Large



How similar are people in the population?

Very Similar

Somewhat Similar

Very Different

Start

Stop

## Explanation

When you draw samples from the population there are two things that will influence how similar the samples are to each other.

1. Whether the samples are large or small
2. Whether the individuals in the population are similar to each or not

Click the button to generate samples. When are the same averages close to each other and when are they very different?

When each sample is small then the samples can be very different from each. Larger samples tend to be more similar to each other.

When the individuals in the population are similar to each other then the samples are similar to each other. But when they are very different from each other then the samples can be very different.

**Box 1** The relationship between the population and samples.

## Distributions

The second half of this lecture is about distributions. However, before we start talking about distributions, let's do another simple thought experiment to try and understand how distributions come about.

### The binomial distribution

In our thought experiment, we'll take a coin, and we'll flip it. When we flip a coin, one of two outcomes

is possible. Either the coin will land showing heads, or it will land showing tails. We can say that there are two possible events or two possible sequences of events (*sequences* will make more sense when we add more coins) than can happen when we flip a coin.

Now let's think of each of the sequences and count up how many heads are showing in each. In the first sequence, where the coin lands showing tails, no heads occur. In the second sequence, where the coin lands showing heads, there is one head. Therefore, we have one sequence that produces 0 heads, and one sequence that produces one head. And those are all the possible sequences.

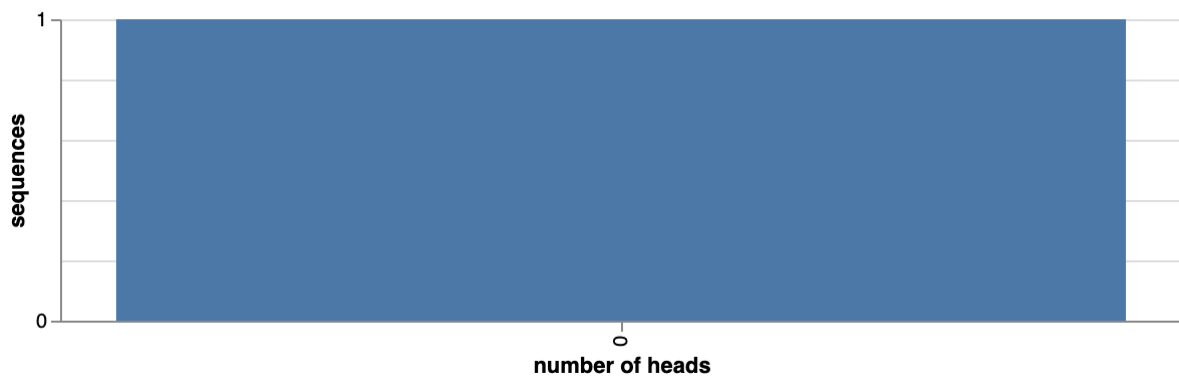
But now let's make it more complicated. Let's flip two coins. Now there's a greater number of possible sequences. We can list them:

1. The first coin shows heads, and so does the second (HH),
2. The first coin shows heads and the second shows tails (HT)
3. The first coin shows tails and the second shows heads (TH)
4. and the first coins shows tails and the second shows tails (TT)

Therefore, there are four possible sequences. Let's count up the **number of sequences** that lead to 0 heads, one head, two heads, etc. If we do this, we'll see that one sequence leads to 0 heads (TT). Two sequences lead to 1 head (HT, and TH). And one sequence leads to 2 heads (HH).

Let's now add more coins. Things will get trickier from here because the number of sequences rapidly goes up. With three coins, there would be eight possible sequences, and with four coins, there would be 16 possible sequences. Figuring out the number sequences, and the nature of the sequences (whether they produce 0 heads, one head, two heads, etc.) quickly becomes difficult. To make things easier, we'll draw a plot. First, we'll draw a plot to trace out the sequences. We'll use different coloured dots to indicate heads and tails. We can do this in the form of the branching tree diagram shown in Box 2. Once we've visualised the sequences, it's easy to count up how many sequences result in 0 heads, one head, two heads etc. We can put our counts on another plot. For this, we'll make a frequency plot or histogram. On the x-axis, we'll have the number of heads. And on the y-axis, we'll have the count of how many sequences result in that number of heads. This frequency plot is also shown in Box 2.

When there are 0 coins there are 1 possible sequences.



**Box 2** Sequences of coin flips (top) and the number of sequences producing a particular number of heads.

You can adjust the slider in Box 2 to change the number of coins you want to flip. Increasing the number of coins increases the number of possible sequences, and it changes the number of ways of getting one head, two heads, three heads and so on changes (however, there's always only one way to get 0 heads and one way to get all heads). Notice that as you adjust the slider and add more and more coins, the frequency plot takes on a characteristic shape. You can mathematically model the shape of this plot using a **binomial distribution**.

In our coin flipping example, we created this shape by counting up the number of sequences that produced various quantities of heads. But if we look around at **natural processes**, we'll see that this shape occurs often.

One natural process that gives rise to this shape is the "bean machine"<sup>1</sup>. In a bean machine, small steel balls fall from the top of the device to the bottom of the device. On their way down, they bump into pegs. When one of the balls hits a peg, it has a roughly equal chance of bouncing off to the left or the right, not unlike a coin which has a roughly equal chance of landing heads up or tails up. At the bottom of the device are equally-spaced bins for collecting the balls. If enough balls are dropped into the device, then the **distribution** of balls across the bins will start to take on the shape of the **binomial distribution**. Very few balls will be at the very edges, because this would require the balls to bounce left or right every time.

Similarly, very few sequences of coin flips result in large numbers of heads or large numbers of tails. The greatest number of balls are seen in the bins near the middle. The balls that land here have bounced left and right a roughly equal number of times. Again a similar pattern can be seen with the coin flips.

In Box 3, I've included a computer simulation of a bean machine. Press the Start button to display the bean machine and watch the balls drop. Press Replay to drop more balls.







Start

Replay

Stop

**Box 3** The bean machine showing beans falling and collecting the a bell shape.

## The normal distribution

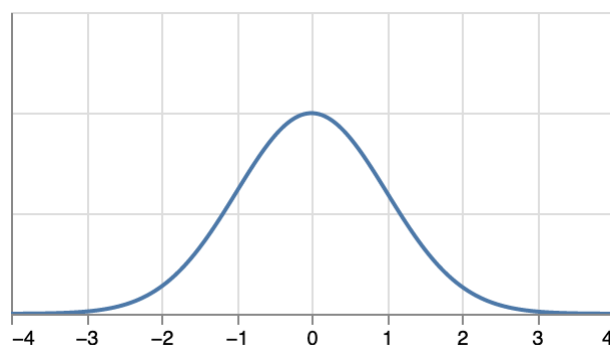
The shape seen in the **binomial distribution** is also seen in another distribution called **the normal distribution**. There are two key differences between the normal distribution and the binomial distribution.

The **binomial** distribution is **bounded**. That means that one end represents 0 heads and the other end represents all heads. That is, the distribution can only range from 0 to  $n$  (where  $n$  is the number of coins that have been flipped)—it is bounded at 0 and  $n$ . The **normal distribution**, however, ranges from negative infinity to positive infinity. Additionally, for the **binomial distribution**, the steps along the x-axis are **discrete**. That is, you can have 0 heads, one head, two heads and so on, but you can have anything in between—for example, it's not possible to have sequences of coin flips that results in 1.5 heads. In contrast, the normal distribution is **continuous**.

The **normal distribution** is a mathematical abstraction, but we can use it as a **model** of real-life frequency distributions. That is, we can use it as a model of **populations** that are produced by certain kinds of natural processes. Because normal distributions are unbounded and continuous, nothing, in reality, is normally distributed. For example, it's impossible to have infinity or negative infinity of anything. This is what is meant by an **abstraction**. But natural processes can give rise to frequency distributions that look a lot like normal distributions, which means that normal distributions can be used as a model of these processes.

### DESCRIBING THE NORMAL DISTRIBUTION

Before we see how natural processes can give rise to the normal distribution, let us take a look at one. Box 4 shows an example of a **normal distribution**. Two parameters can be used to describe the normal distribution. The location parameter denoted  $\mu$  describes where the distribution is centred. The scale parameter denoted  $\sigma$  describes the width of the distribution. When you learn how to represent populations with distributions, then these two parameters will correspond to the population average and population variation, respectively.



Centre ( $\mu$ )

0

Width ( $\sigma$ )

1

Reset

**Box 4** An example normal distribution. It is described by two parameters. The  $\mu$  parameter controls where it is centred and the  $\sigma$  parameter controls how wide it is.

## PROCESSES GIVING RISE TO NORMAL DISTRIBUTIONS

To see how natural processes can give rise to the normal distribution, let us take a look at the simple simulation in Box 5. The simulation in Box 5 is a model of a dice game. In this dice game, each player rolls a dice a certain number of times and then **adds** up the values of all the rolls. For example, if a player rolled the dice three times and they got 1, 4, and 3, then their score would be 8 ( $1 + 4 + 3 = 8$ ). We can plot a frequency distribution of how many players got each of the possible scores. If you have enough players and enough dice rolls, then this frequency distribution will begin to look a lot like normal distribution (See Box 4).

Click **Roll!** to see it in action. Try adjusting the number of dice rolls and the number of players, and you'll see how the distribution looks more and more like a normal distribution. (For now, we'll just ignore the boxes that say Add and Multiply, but we will discuss those later).

count

**Box 5** A dice game gives rise to a normal distribution when a player's score is calculated by *adding* the values from all their rolls. *Multiplying* the values does not give rise to the normal distribution

Although this is a **dice game**, we can imagine other processes that might work similarly. For example, a **developmental process** might work similarly. Let's say that we have a developmental process like height. At each **point in time** some **value** can be **added** on to the person's current height—that is, at each time point a person can **grow** by some amount just like player's scores can increase by some amount on each dice roll. If you have enough people, and time points, then the **distribution of height** will start to look like a normal distribution. This applies for any process that develops or grows by **adding** bits to its current value.

In the dice game, a person's score can increase by either 1, 2, 3, 4, 5, or 6 after each roll. And if the dice is balanced an increase of 1 will be no more common than an increase of 6, or 5, or any other number in-between. But even if we had unbalanced dice, so people were more likely to roll a 6 than a 1, then a normal distribution would still appear. A normal distribution appearing isn't dependent on the numbers that are added at each round. Rather, the important thing is that numbers are *added* (or *subtracted*, because subtraction is just the addition of a negative number). If instead of *adding* we *multiplied* the numbers, then a normal distribution would not appear.

## PROCESSES THAT DON'T PRODUCE NORMAL DISTRIBUTIONS

We won't cover other distribution shapes in much detail, but let us take a look at an example of a process that doesn't produce a normal distribution. To do this, we'll just modify the dice game in [Box 5](#).

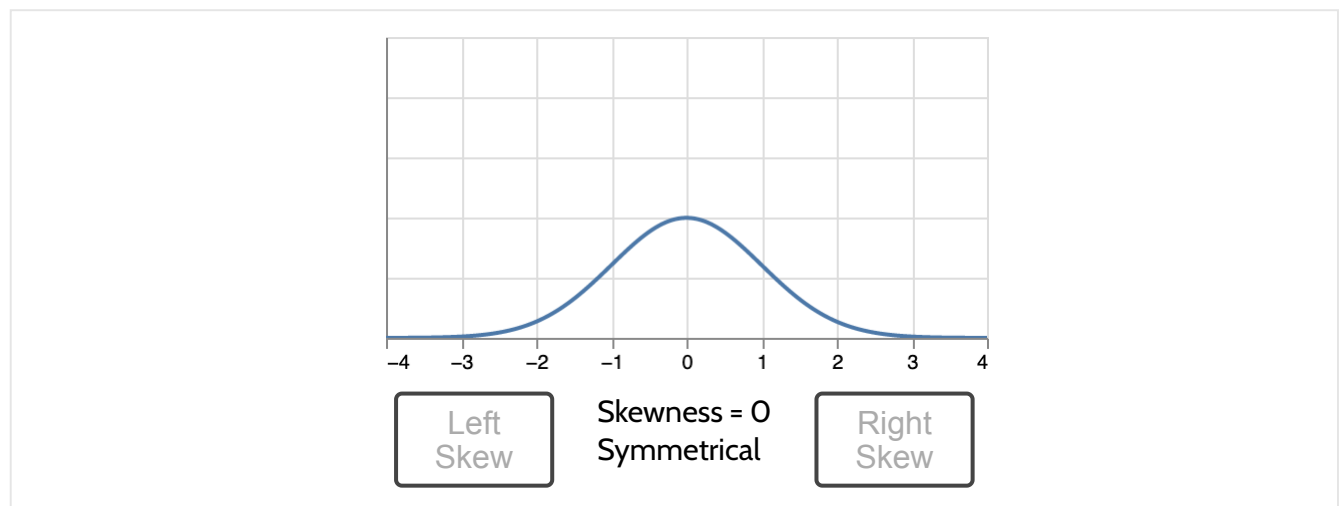
In [Box 5](#), click the option that says **Multiply**. Doing so changes the rules of the dice game so that a player's score is determined by *multiplying* together the values of their dice rolls. For example, under the new rules, if a player rolled 1, 4, 3 then their score would be 12 ( $1 \times 4 \times 3 = 12$ ). Now try clicking Roll! to see the shape of the distribution. This new distribution has an extreme **skew**. The vast majority of players have fairly low scores, but a tiny minority of players have extremely high scores. When you have a process that grows by multiplying, then you'll get a distribution that looks like this.

In psychology, we won't study many processes that grow like this, but some processes that grow like

this will be very familiar to you. Think about something like wealth. Wealth tends to grow by multiplying. For example, earning interest or a return on investment of 10% means that the new value of the investment is 1.10 times the original value. This feature of wealth growth explains why, for example, in the UK, the top 10% of people control more wealth than the bottom 50%.

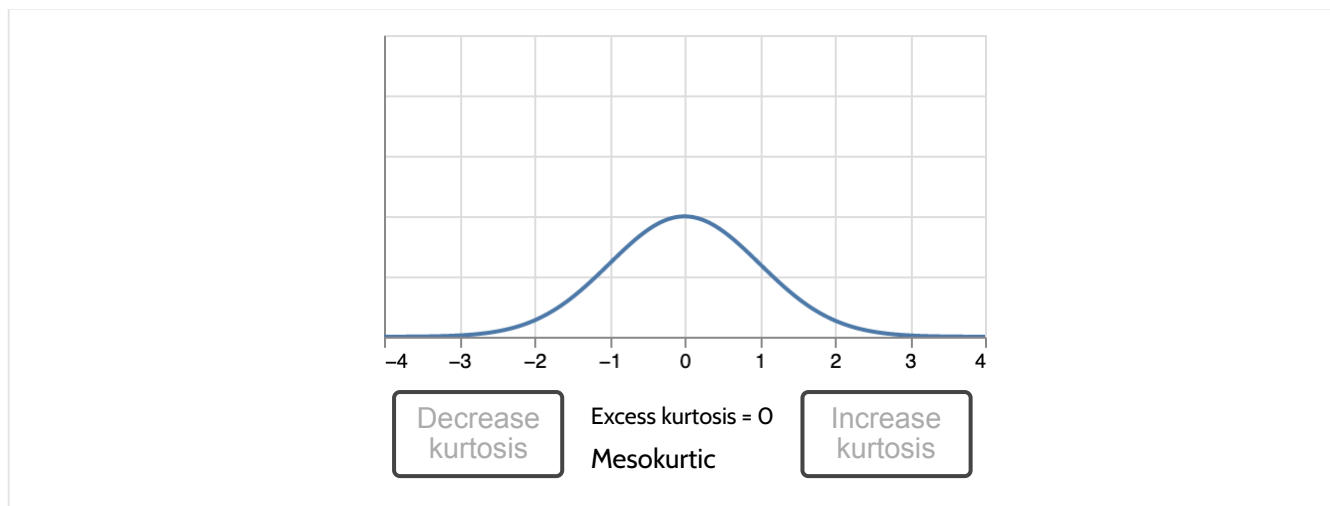
## DESCRIBING DEVIATIONS FROM THE NORMAL DISTRIBUTION

When you clicked the *Multiply* button, the dice game produced a distribution that was **skew**. **Skew** is a technical term that describes one way in which a distribution can *deviate* from a *normal distribution*. The *normal distribution* is **symmetrical**, but a **skew** distribution is not. A left-skewed distribution has a longer *left tail*, and a **right-skewed** distribution has a longer *right tail*. Use Box 6 to explore **skewness**.



**Box 6** An example of a skew distribution. Negative, or left skew, means the left tail is longer. Positive, or right skew, means the right tail is longer.

Apart from **skew**, deviations from the **normal** distribution can occur when a distribution either has fatter or skinnier **tails** than the normal distribution. The *tailedness* of a distribution is given by its **kurtosis**. The kurtosis of a distribution is often specified with reference to the **normal distribution**. In this case, what is being reported is **excess kurtosis**. A distribution with **positive excess kurtosis** has a higher kurtosis value than the normal distribution, and a distribution with **negative excess kurtosis** has a lower kurtosis value than the normal distribution. In your research methods courses, you probably won't come across many distributions that have negative excess kurtosis. However, the distribution that describes dice rolls is one such distribution, and this will be discussed briefly later in this course. You will encounter distributions with positive excess kurtosis more often. In particular, the *t*-distributions, a distribution with positive excess kurtosis, will be used in several of the statistical procedures that you will learn. In Box 7, you can explore excess kurtosis. When excess kurtosis is set to 0, then the figure displays a normal distribution. Distributions with no excess kurtosis are called *mesokurtotic*. When excess kurtosis is negative, the figure displays a thin-tailed or *platykurtotic* distribution. And when excess kurtosis is positive, the figure displays a fat-tailed or *leptokurtotic* distribution.



**Box 7** An example of excess kurtosis. Positive excess kurtosis results in fat tails (extreme values are more common) and negative excess kurtosis results in skinny tails (extreme values are less common)

## Distributions and samples

Now that we've covered distributions, we can return to samples and see how the two are related to one another. We saw in the *dice game* (see Box 5) that when we produce values by **adding**, then those values will start to look like a normal distribution as long as we add up enough values.

Importantly, there's nothing normally distributed about the numbers we're. For example, a balanced dice will show the values 1, 2, 3, 4, 5, and 6 with roughly equal frequency. But when you produce values but adding up the result of lots of dice rolls, then those values will be normally distributed.

Now let's try to relate this to the idea of samples. Let's say that we have some unknown population, and we can measure members of that population. That is, we can take a sample. And let's further say that when we've measured all the members of our sample, we'll just add up all those measurements to arrive at some final value. Now let's repeat that process. We'll take another sample, sampling the same number of members as before, and add up all the values to arrive at some final value. We can repeat this process over and over until we have lots of sums (created by adding up all the values in the sample). How will these values be distributed?

This situation is exactly analogous to the dice game. When we measure a member of the population, then that is analogous to rolling the dice. Adding the values in the sample to arrive at the final value is analogous to adding up all the dice rolls to get to the player's score. And the player's score is equivalent to the sum of all the measurements in our sample. Since the players' scores are normally distributed (provided we have enough dice rolls) the sample sums will also be normally distributed, provided we use a large enough sample size.

We can now add another mathematical twist to this story. Let's say that before we add up all the values, we divide each value by the number of values (so if we have a sample size of 10, we'll divide each value by 10). Now our sum is an average (a mean to be exact).

Now think about what this means. It means that when we take samples from a population, no matter

NOW THINK ABOUT WHAT THIS MEANS. It means that when we take samples from a population, no matter how the population is distributed if our samples are big enough, then our sample averages will be **normally distributed**. This insight will underlie the concept of the **sampling distribution**, which you'll learn about in a later lecture, and it is one of the most fundamental ideas in statistics.

---

## Footnotes

1. The "[bean machine](#)" [[↗](#)].

## References

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioural and Brain Science* 33 (2–3): 61–135. <https://doi.org/10.1017/S0140525X0999152>.