Lecture 6: Samples, Populations, and Distributions

Dr Lincoln Colling
04 November 2021

UNIVERSITY OF SUSSEX

Plan for today

Part I: Samples and populations

- The relationship between samples and populations

Part II: Distributions

- The binomial distribution
- The normal distribution
  - Describing the normal distribution, and processes that produce normal distributions
  - Processes that don't produce normal distributions, and describing deviations from the normal distribution

Part III: Distributions and samples

Samples and populations

A key use of statistics is to make inferences (or claims) about populations from *the information* we get from samples

Example

You're interested in the average height of

people in the UK

How can you go about collecting some data that will allow you to make claims about the average height of people in the UK?

Samples and populations

To you have (at least) two options:

Option 1

- Measure the height of *all the people in the UK* and then work out the average
- But that's over 66 million people, so it'll take you a very long time and maybe some people don't want to be measured

Option 2

- Measure a *subset* of *all the people in the UK* and use the average of this subset to figure out plausible values for the average height of people in the UK

In this example, the subset of people is the sample and

all the people in the UK

is the population

The relationship between samples and populations

After we've taken a sample we'll want to use information from this sample to figure out something about the population

But what's the relationship between the population and the sample?

The sample should hopefully resemble the population in some way

- For example, the average of the sample should resemble the average of the population

- But we don't know the average of the population (if we did, then we wouldn't need the sample), so how would be *know* whether our sample resembles the population?

We can do a *thought experiment* to try and figure out some factors that will influence whether the sample *resembles* the population

Relationship of sample to population

Let's think back to our question about *the average height of people in the UK*

Factor 1: Variation in the population

If all members of the population are identical then the height of one person would be the same as the average height of two people, or 100 people, or the entire population, because people only come in one height

When there is no variation in the population then the sample average will be identical to the population average

Relationship of sample to population

Factor 2: Size of the sample

If our sample is large enough so that it *includes all members of the population* the sample and the population are the same thing

When the sample includes the entire population then the sample average will be identical to the population average

These are extreme cases but they suggest that population variation and sample size will influence the relationship between samples and populations

Relationship of sample to population

So if we have a big sample and/or small population variation then will our sample resemble the population?

For a particular sample there is no way of knowing whether it resembles the population or not, because we don't know what the population looks like!

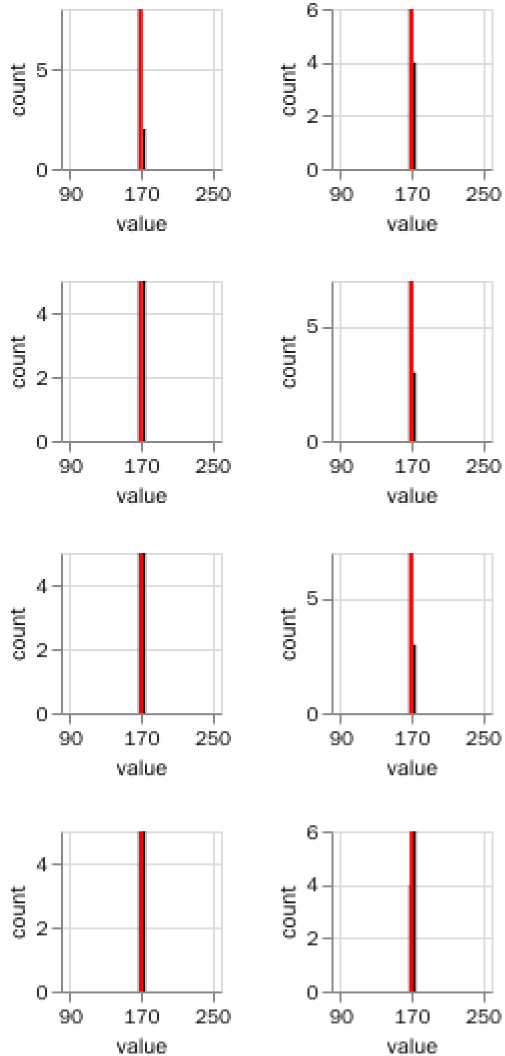
A better way to think about this is in terms of repeated sampling

- If we take lots of samples from the same population then will those samples on average be closer to the population?
- These two factors (sample size, and population variation) will influence whether the samples resemble the population on average

If our sample size is big enough then samples will on average resemble the population...

...but what counts as big enough will depend on the population variation

Repeated sampling from the same population



How big is the sample?

Small | Medium | Large

How similar are people in the population?

Very Similar | Somewhat Similar | Very Different

Start | Stop

Distributions

Before we start talking about distributions let's think about what they are and where they come from

We'll do another thought experiment

- We'll take a coin, and we'll flip it.

- Two outcomes are possible

    1. The coin lands showing *heads*

    2. The coin lands showing *tails*

Of the two possible events

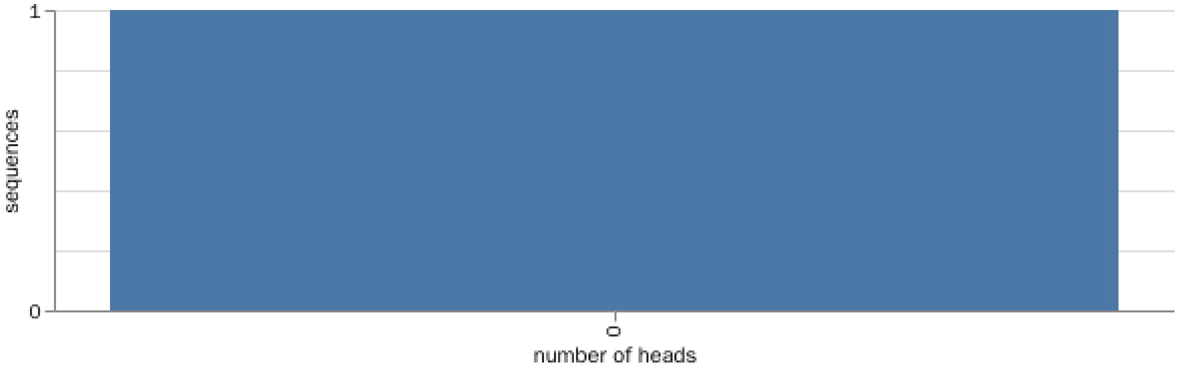
1. One produces 0 heads
2. One produces 1 head

Now let's add more coins. As we do, we'll count up the number of sequences that produces 0 heads, 1 head, 2 heads, 3 heads etc

The binomial distribution

coins  0

When there are **0** coins there are **1** possible sequences.



Plotting the frequency of outcomes

We'll treat the *number of heads in a sequence* as our *outcome*

As we add more and more coins we can plot the frequency of each *possible outcome*.

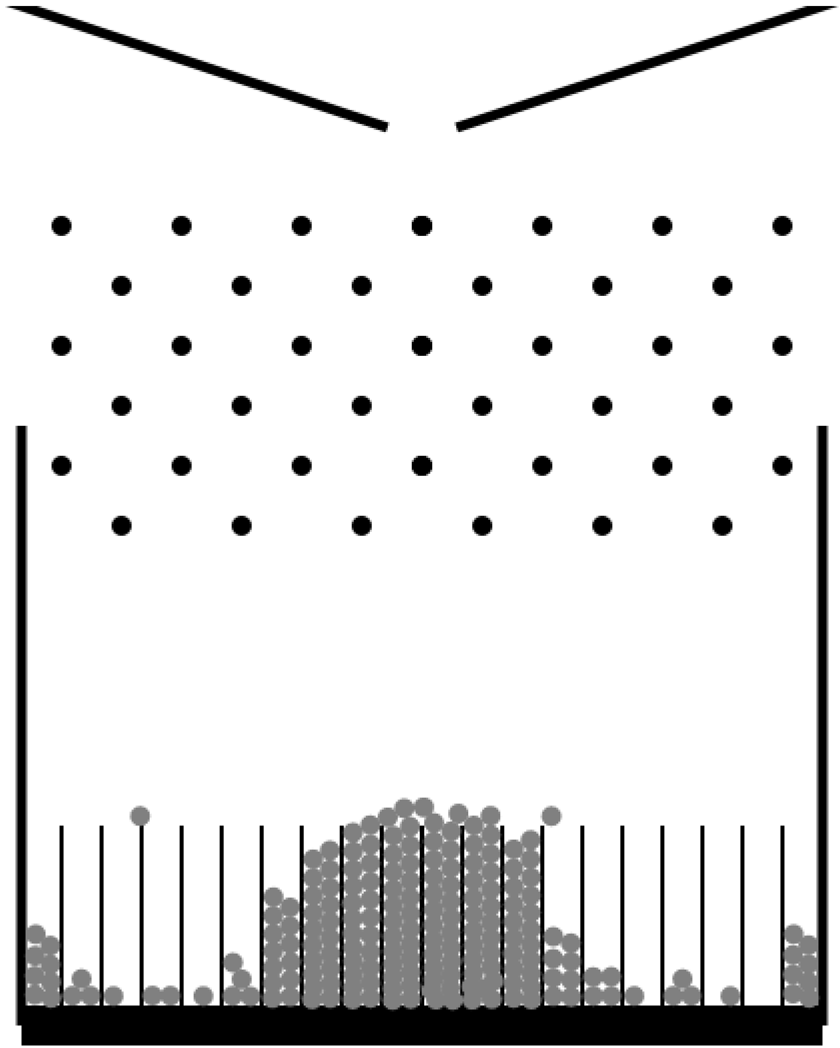
- This frequency plot starts to take on a characteristic shape
- This shape can be described mathematically using the binomial distribution

The binomial distribution describes the frequency of outcomes in our coin flipping example[1]

[1] In our thought experiment we assume that every *possible* sequence of Heads and Tails occurs, and that it occurs only once.

Natural processes that produce binomial distribution

Balls falling through a *bean machine* approximate a binomial distribution



Drop balls

The normal distribution

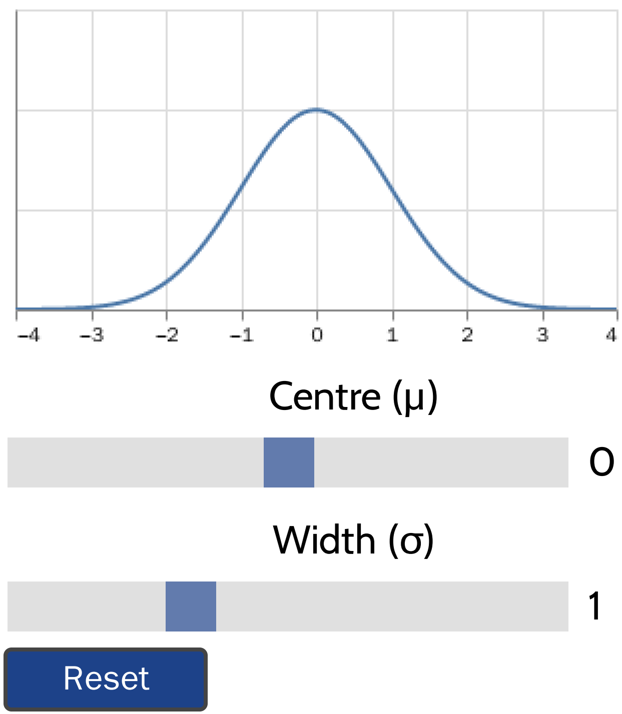
The shape seen in the binomial distribution is also seen in another distribution called the normal distribution.

Differences between the binomial distribution and the normal distribution:

- The binomial distribution is bounded and the normal distribution is not
  - The binomial distribution ranges from 0 to n (where n is the number of coins you've flipped)
  - The normal distribution ranges from $-\infty$ to $+\infty$
- The binomial distribution is discrete and the normal distribution is continuous
  - You can only have 0 heads, 1 head, etc., and not 1.5 heads
  - *Normal distribution* represents all outcomes between $-\infty$ and $+\infty$

The normal distribution as a model

- The normal distribution is a mathematical abstraction (nothing in real life perfectly follows a normal distribution)
- But we can use it as a model of real-life frequency distributions



The normal distribution can be described by two parameters:

- The μ parameter controls where it is centred
- and the σ parameter controls how wide it is.

Processes giving rise to normal distribution

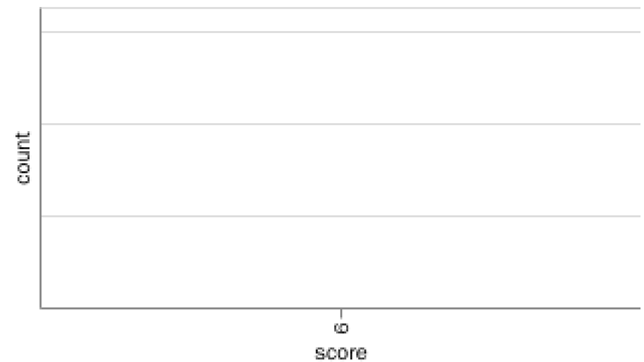
To see how a natural process can give rise to the normal distribution let's consider a simple *dice game*

The rules of the game

1. A group of x players roll a dice n times
2. A player's score is calculated by adding all the values of the dice rolls
    - For example, if they rolled the dice three times ($n = 3$) and the dice showed 1, 4, and 4 then their score would be 9 ($1 + 4 + 4 = 9$)

If you have enough dice rolls then the players' scores will be normally distributed

A dice game simulation



Dice rolls? 30 Players? 100 Roll!

Add Multiply

As you increase the number of dice rolls the frequency distribution of *players scores* will start to look like a normal distribution

But you also need enough players to clearly see shape

Natural processes are analogous to the dice game

There are many natural processes that are analogous to the dice game

We can imagine other processes that work like the dice game

- For example, a developmental process might work similarly.
    - At each point in time some value can be added on to the person's current height just like players scores can increase by some amount on each dice roll.

The numbers you add aren't important... it's the *adding* that's important

Natural processes are analogous to the dice game

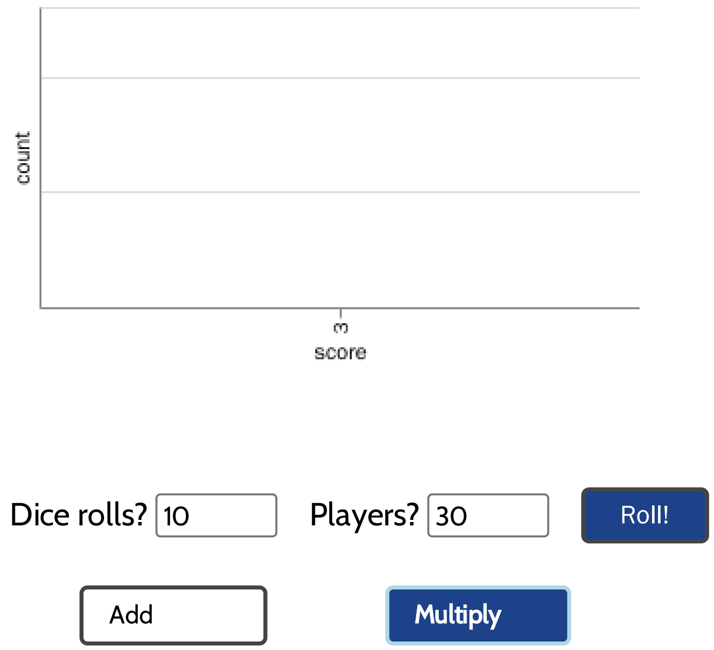
A person’s score can increase by either 1, 2, 3, 4, 5, or 6 after each roll, and with a balanced dice an increase of 1 will be no more common than an increase of 6, or 5 etc

But even if the dice were unbalanced then a normal distribution would still appear.

The numbers that you add isn't the important thing... the adding is what's important

If instead the numbers were multiplied then we wouldn't see a normal distribution

A different dice game simulation



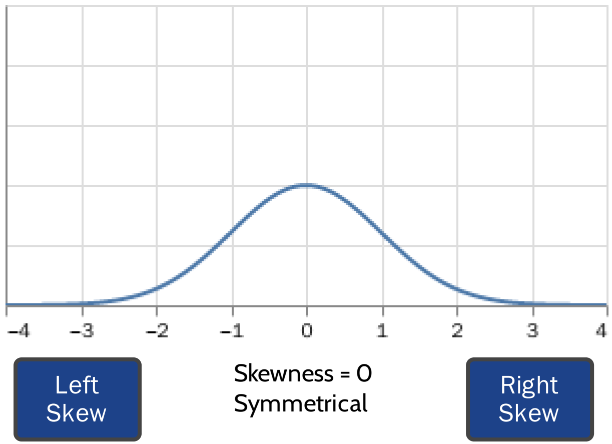
We can change the rules of the dice game so that we multiply instead of add

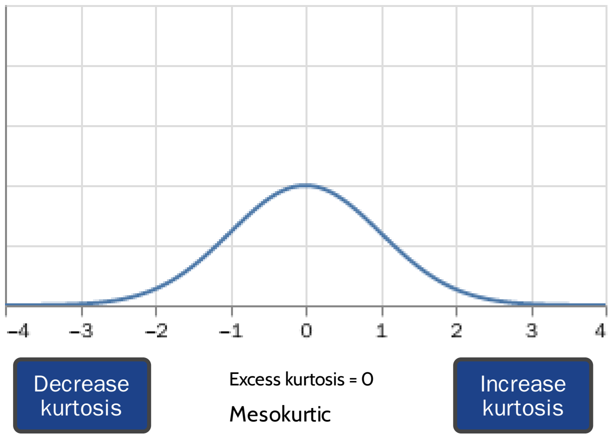
- This won't produce a normal distribution.
- The distribution will be skew

Describing deviations from the normal distribution

When distributions deviate from the normal distribution this can happen in two ways

- The distribution can be asymmetrical
- The size of the tails can change





- Asymmetry is quantified by skew
- The thickness of the tails is quantified by kurtosis (or excess kurtosis if given *relative* to the normal distribution)

Distributions and samples

Now that we know a little about distributions we can return to samples

So far, with the dice game we've just been concerned with whether the scores of the players within a game are normally distributed

But can we say anything about distribution of scores across games?

Let's return to the problem of sampling people and measuring their height

- After we've got our sample and we've taken our measurements let's add up all the measurements, and calculate the total height of our sample.
- Now let's take another sample and add all the measurements again

What can we say about how these sums will be distributed?

Because we're taking sums, we know that the total heights will be normally distributed!

Distributions, samples, and averages

What if instead of taking a regular sum we can first *divide each value in the sample by the sample size and then calculate the sum?

This turns the *sum* into an average, but because we're still dealing with sums, the averages will be normally distributed too!

That is, we can expect that if we take lots of samples (select a group of people and measure their height), and then we work out the average of each sample, then these sample averages will be normally distributed

This will happen irrespective of how the population is distributed

- That is, even if height isn't normally distributed then the average height from lots of samples will still be normally distributed[1]

[1]Remember, whether a dice shows 1, 2, 3, etc isn't normally distributed. Each value is equally likely

Preview of the sampling distribution

The fact that sample averages are normally distributed underlies the concept of the sampling distribution

The sampling distribution will underlie many of the statistical procedures you'll learn about, and it'll be covered in more detail in the next lecture!