Distributions, functions, transformations

In this lecture we will start thinking about variables in terms of distributions. We will see how we can perform simple arithmetic operations, such as addition and multiplication on entire variables to perform linear transformations. We'll discuss one transformation in particular, the *z*-transformation, and see how it's used to standardise the values of a variable. Finally, we will talk about how we can use simple maths to compare groups on a measured variable of interest.

AUTHOR

Milan Valášek

AFFILIATION

University of Sussex

PUBLISHED

Sept. 2, 2021

Contents

Introduction

In the previous lecture, we introduced the basic *descriptive statistics*, the measures of central tendency and spread, and talked about the concepts of the sampling distribution and the standard error. In this lecture, we start playing with the mean and standard deviation and see how we can manipulate them to create comparable distributions.

Before we do that, we need to get used to thinking about things—variables—in terms of their distributions.

The shape of things

For the purpose of this lecture, we will only be talking about *continuous* variables. If you need a refresher on what these are, check out Lecture 4.

As we saw in previous lectures, we can plot an observed variable on a *histrogram* to visualise the distribution of its values. For example, if we measure the height of 500 women and plot the values, we get something like Figure 1.

Plot R code

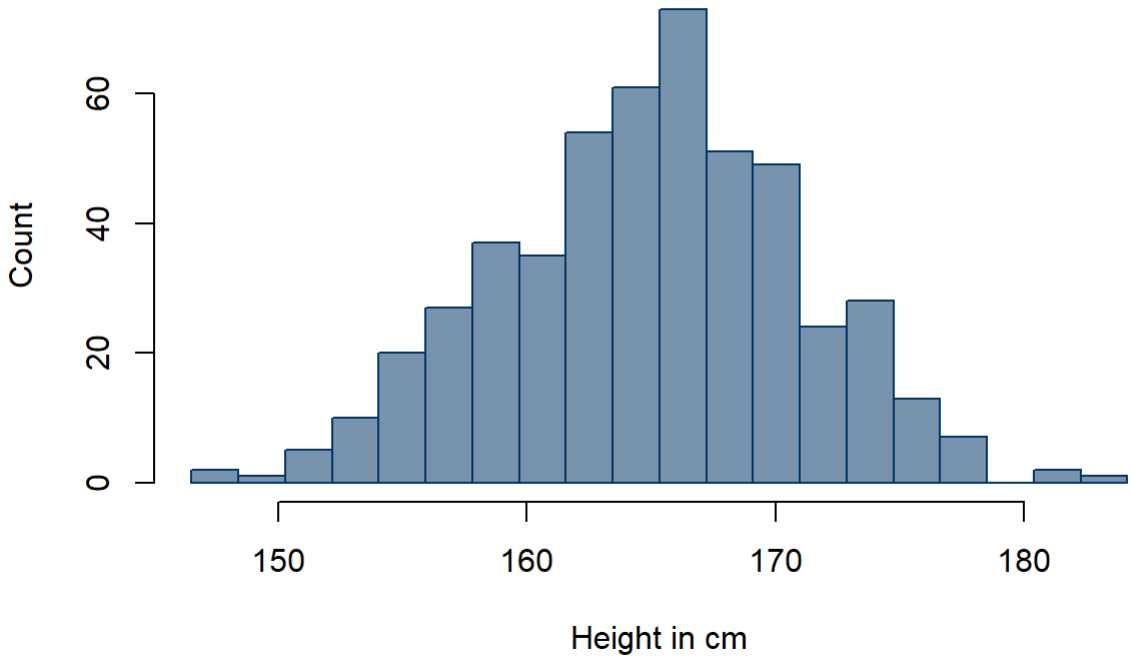


Figure 1: Distribution of height on a sample of 500 women. This is not real data.

As you can see, the vast majority of the measured heights were roughly in the 155-175 centimetre range, with only a few people in the sample being shorter than 155 cm or taller than 175 cm. You can also see that the distribution is roughly symmetrical around its mean and has the shape of a bell characteristic of a normal distribution. This is not a coincidence: height is a normally-distributed variable as you found out in Lecture 6. Of course, the shape isn't as smooth as the normal curve you saw in the same lecture. This is because 500 observations is too few to smooth out any statistical fluctuations due to sampling. So, just by chance alone, we can end up with a few more 173-cm-tall people and a few fewer 172-cm-tall people than we would expect based on what we know about the normal distribution.

Because height is a continuous variable and no two people are the exact same height, to plot the variable on a histogram, we have to assort the values into bins. Each bar on the histogram in Figure 1 represents the number of people whose height falls within a 1 cm range (150-151 cm, 151-152 cm, *etc.*). How wide we make the bins is an arbitrary decision but it's a good idea not to make them so wide that the distribution gets unrecognisable o so narrow as to have too many gaps in the histogram (see Figure 2).

Plot R code

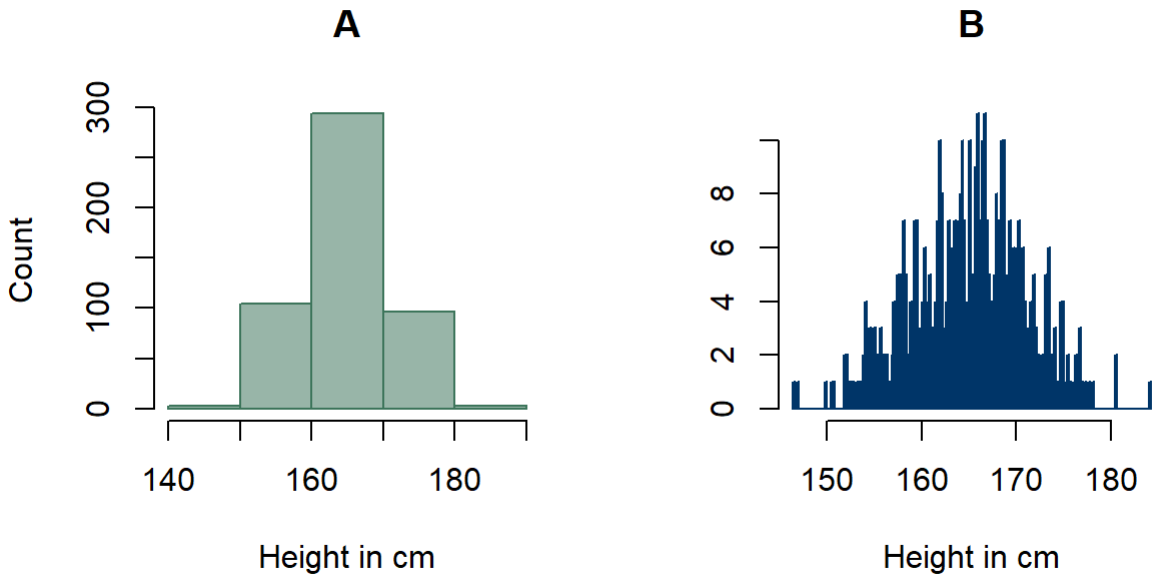


Figure 2: Histograms with (A) too few bars to see the distribution in enough detail and (B) too many bars.

Now, if we could collect an infinite number of observations, we could make the bins as narrow as we wanted, even *infinitely* narrow. This would give us an idealised shape of the normal distribution: the normal curve. The process of making the bars narrower and narrower until each only represents a single point on a line is visualised in Figure 3

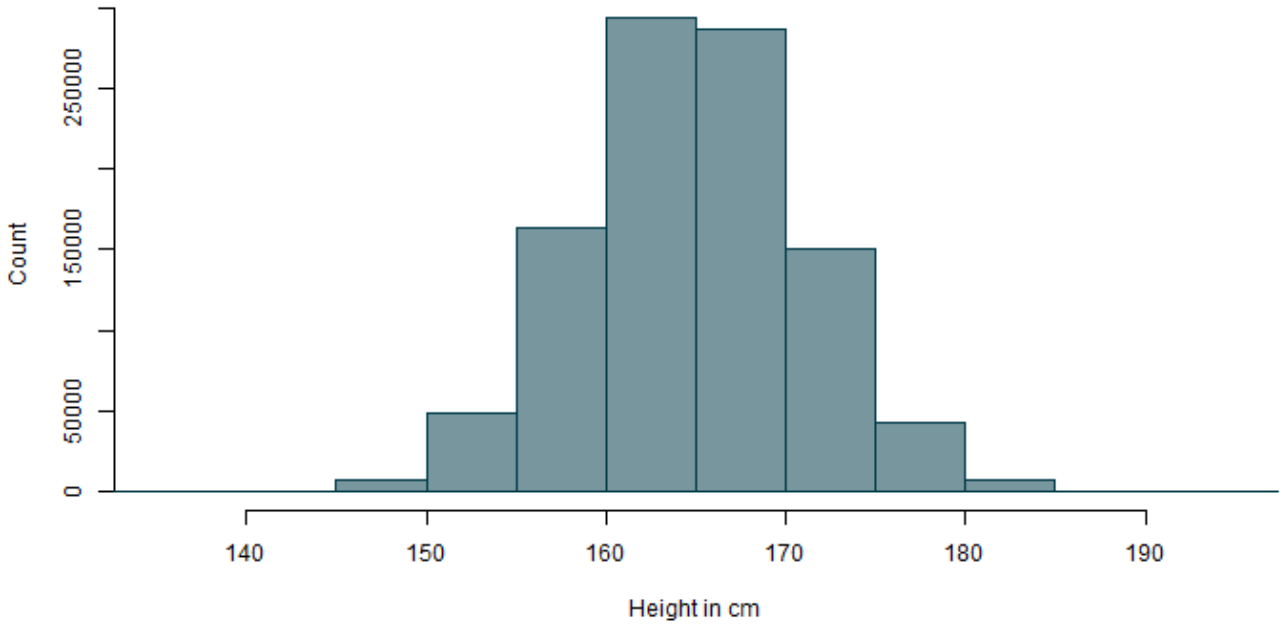


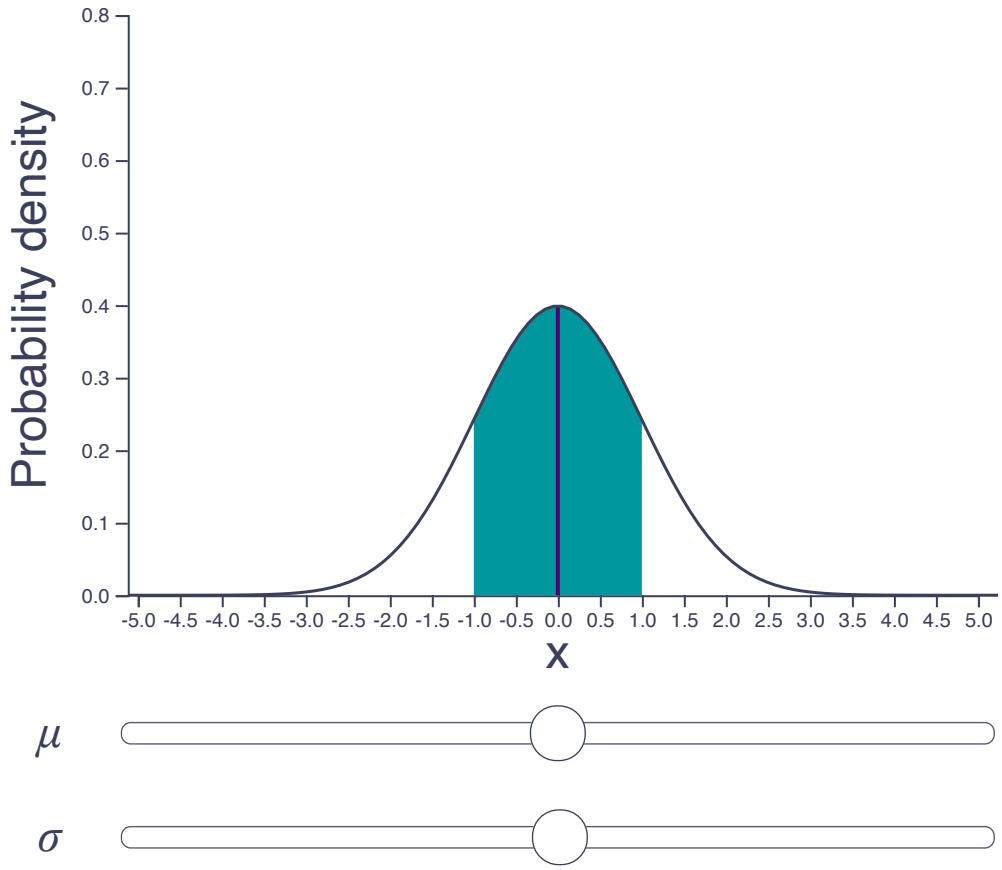
Figure 3: From a histogram of a variable to an ideal normal curve

Because we will mostly be talking about continuous normal variables, we can visualise them as this kind of curve. This allows us to abstract from the individual values and talk about variables in terms of their means and standard deviations. Everything that applies in general to this abstraction of a variable in the form of the curve, applies to the actual values. The curve is just a tool that allows us to talk about variables without having to worry about details that are *not relevant* for the topic at hand.

OK, to reiterate what we said in the previous lecture, we can describe key properties of a variable using measures of *central tendency* and *spread*. The mean tells us where the centre of a variable is and the standard deviation tells us the average distance between the mean and each value in the variable.

In a normally distributed variable, the majority – about 68% – of all the values are concentrated within ±1 standard deviation to either side of the mean.

The larger the standard deviation, the more spread out the variable is. The interactive visualisation below shows a normal distribution with a $\mu = 0$ and $\sigma = 1$. This is called the standard normal distribution. Have a little play around with the it using the sliders to make sure the concepts of standard deviation and mean feel intuitive.



The standard normal distribution

The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$. Use the top slider to shift the mean of the distribution along the x-axis and the bottom slider to change the standard deviation (possible range is between 0.5 and 3).

The purple line shows the mean of the distribution and the green shading shows the distance of 1 standard deviation to either side of the mean. That means that *the shaded area is 2 standard deviations wide*.

(You *really* don't need to worry about what "probability density" means. It's a fairly complicated concept and it's on the y-axis label only so that the plot is technically correct)

There are two important things here to notice.

First, the mean and the standard deviation are independent of one another: You can change one and the other will stay the same.

Second, neither shifting the mean, not changing the standard deviation of a distribution doesn't change its *fundamental shape*. Yes, changing the value of the standard deviation does make the distribution appear flatter or pointier but the relative position of the individual points on the line with respect to each other does not change! In other words, the distance between any two points may get larger or smaller but it gets so equally for any two points. For that reason, the shape of the curve remains the same: It is still true that about 68% of all the values are within ±1 standard deviation from the mean (in the shaded bit), no matter how large or small the value of the standard deviation is.

To illustrate this, let's take our example of people's heights (Figure 1). In this example, we measured height in centimetres. It should be pretty obvious that your height doesn't change depending on the units you measure it in: you're the same height whether you get measured in centimetres, metres, millimetres, or feet and inches. It's just the value that changes.

So, if we measured the height of the sample of women in that example in metres, the shape of the data would remain the same. Figure 4 below shows exactly this. The distribution in panel A has a standard deviation of 5.99, while the one in panel B has a standard deviation of 0.06. But as you can see, *it's the very same distribution*: it is just displayed on different scales (centimetres vs metres, respectively)

Plot R code

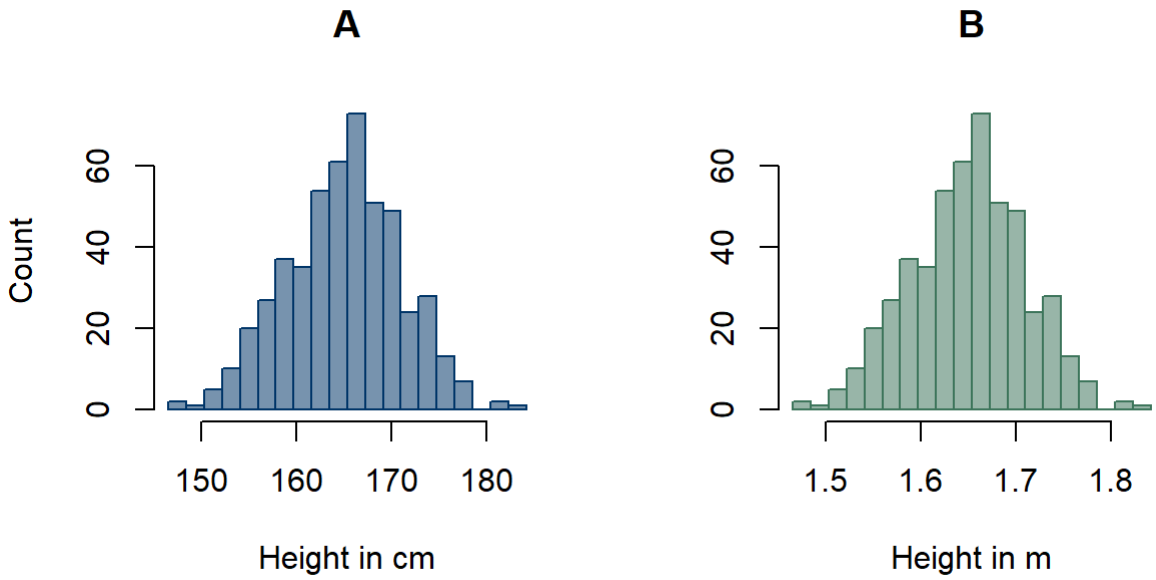


Figure 4: Histograms of participants heights measured in (A) centimetres and (B) metres.

This is why the standard deviation is sometimes referred to as the scale parameter of a distribution, while the mean is the location parameter.

Changing the mean shifts the location of a distribution to the left or right along the number line, while changing the standard deviation changes the scale of the distribution is displayed.

Of course, we haven't yet talked about *how* we can change the mean or the standard deviation of a variable but it's important that you understand what effects these changes have.

Transformations

So how *do* we change things like the mean and the standard deviation of a variable?

Since we're talking about observed variables, let's switch from talking about means and standard deviations in general to talking about the sample mean ($\bar{x}$) and sample standard deviation (SD). The only reason for doing this is so that we can use the symbols instead of having to write it all out every time.

We know from the previous lecture that both $\bar{x}$ and SD are sensitive to outliers and that we can change their values to be arbitrarily small or large by moving a few points around. However, only changing the values of a selection of observations will alter the shape of the distribution. That is not good. We can decide to switch our measurement unit of height from centimetres to feet and inches but we have to do it consistently for all observations. We can't just take a few points and convert them. That would mess up *the relationships between individual observations*!

The same principle applies to any data transformations we may want to perform.

Functions

A transformation is just a mathematical function that takes an input and returns an output. Pretty much in the same way as functions in R do it. Take, for example the *second power*: 2^2=4, 3^2=9, 4^2=16 and so on.

We can think of this operation as a function that takes an input, x and returns the output x^2. We would define the function as $f(x)$ (pronounced *f of x* or *function of x*):

$$f(x) = x^2$$

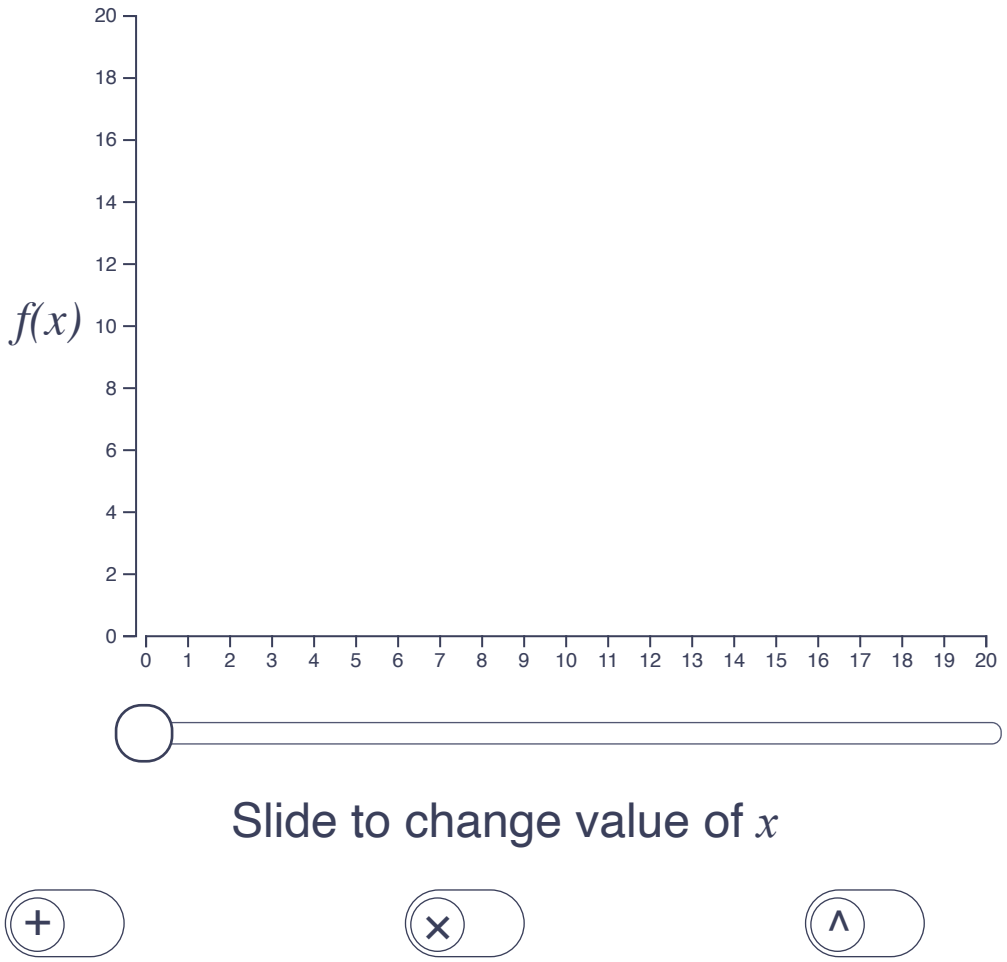
The above is a general formulation of a function that takes any number and returns it's square.

We can create a graph of the function by plotting x on the, well, x-axis and $f(x)$ on the y-axis. Have a look at the applet below. It shows this kind of a graph. You can use the switches to show three functions:

- one that *adds* some number a to x: $f(x) = x + a$,
- one that *multiplies* x by some number a: $f(x) = x \times a$,
- and one that *raises* x *to the power* of some number a: $f(x) = x^a$

The formula appears next to the switch when you flip it on. The default setting is $a = 0$ for addition and $a = 1$ for multiplication and *exponentiation* You can drag the number in the formula left and right to change the value of a that gets added to, multiplied by, or raises x. Finally, you can move the slider to change the value of x and see how the graph of the function looks.

Do spend a few minutes playing with the visualisation to develop intuition about these three functions.



Processing math: 100%

Once again, notice a few interesting and important things:

1. For the default settings, the functions are identical, because for any x, $x + 0 = x \times 1 = x^1$
2. Subtraction is just *addition of a negative number*: $x - a = x + (-a)$
3. Division is just *multiplication by a number between 0 and 1*: $x \div a = x \times \frac{1}{a}$)
4. Taking a^{th} root of a number is just *raising the number to the power of* 1/a: $\sqrt[a]{x} = x^{\frac{1}{a}}$
5. While the graphs of the addition and multiplication functions are straight lines for all values of a, the graph of the exponentiation function is a curve for all but two values of a, 0 and 1.
6. All points along the individual lines are evenly spaced in the case of addition and multiplication but not exponentiation (except for a equal to 0 or 1).

The last two points are *really* important!

Addition and multiplication do not change the relative distances between individual inputs. Their graphs are straight lines, which is why they are referred to as linear transformations.

Some other functions, such as exponentiation, do change these relative distances and thus are non-linear: their graphs are not straight lines.

Notice also what addition and multiplication do. Addition shifts the values of x up and down along the y-axis, while keeping the distances between points unchanged. Multiplication, by contrast, spreads or "squishes" the values of x along the y-axis. Perhaps things are now slowly clicking into place: We have two linear transformations, addition and multiplication, the former of which changes the location of a bunch of numbers and the latter of which changes their spread.

When addition and multiplication are applied to variables, they are referred to as centring and scaling, respectively.

Centring

Centring is the subtraction of a fixed value from each observation of a variable. (Remember, subtraction is just *addition of a negative number*.) This has the effect of shifting the entire variable to an *arbitrary location along the x-axis*.

You can technically centre a variable by subtracting *any* value from it but the most frequently used method is mean-centring:

$$f(x) = x - \bar{x}$$

Applying this transformation results in shifting the variable so that it's mean is at the zero point and the individual values of the mean-centred variable tell us how far that observation is from the mean of the entire variable.

Figure 5 shows our height variable in its original form next to its mean-centred form.

Plot R code

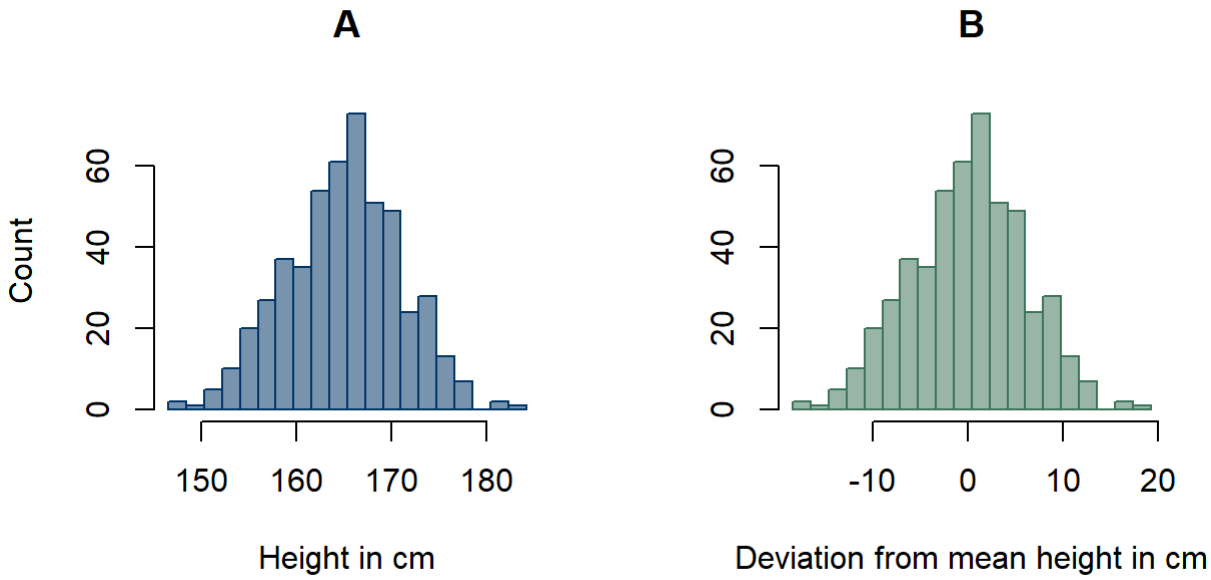


Figure 5: Histograms of participants heights: (A) raw data (B) mean-centred.

It's crucially important to understand that mean-centring does not alter the shape of the variable, nor does it change the scale at which the variable is measured. It only changes the interpretation of the values from the raw scores to differences from the mean.

The variable in panel B in Figure 5 is still measured in centimetres and still has the same standard deviation as the original height variable (panel A).

Scaling

Scaling is the division of each observation of a variable by a fixed value. (Remember, division by x is just *multiplication by the inverse 1/x*.) This has the effect of stretching or squishing the entire variable *in the direction of the x-axis*.

Just like with centring, you can technically scale a variable by dividing it by *any* value. For example, in Figure 4, panel B we scaled the height variable by 100 to transform it from height in centimetres to height in metres. The most frequent method of scaling variables, however, is by their standard deviation:

$$f(x) = \frac{x}{SD(x)}$$

Figure 6 shows our height variable in its original form next to its form scaled by the standard deviation.

Plot R code

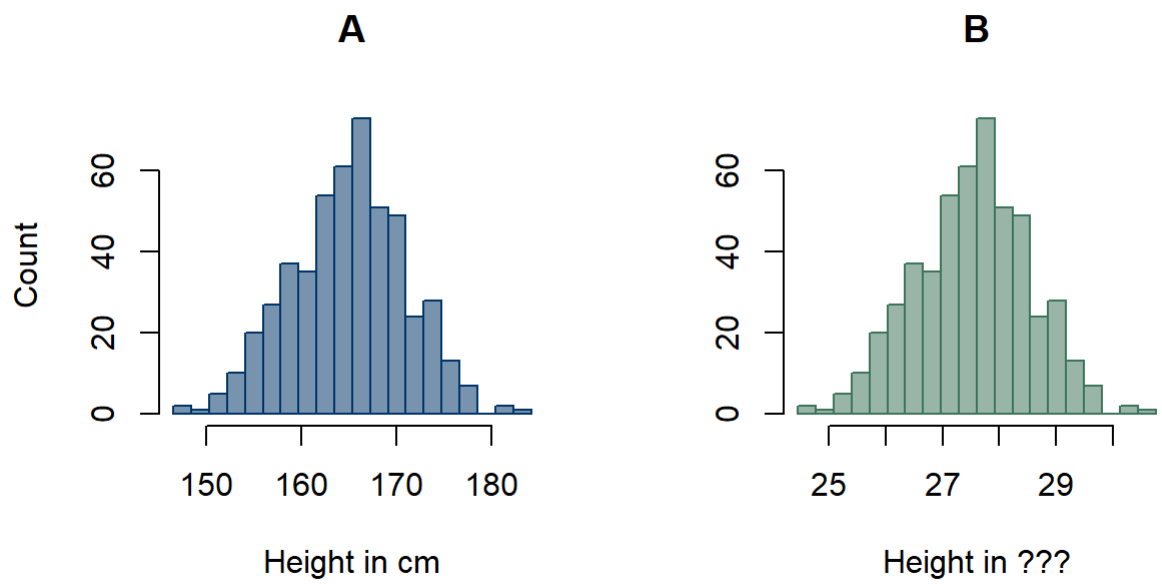


Figure 6: Histograms of participants heights: (A) raw data (B) scaled by *SD*.

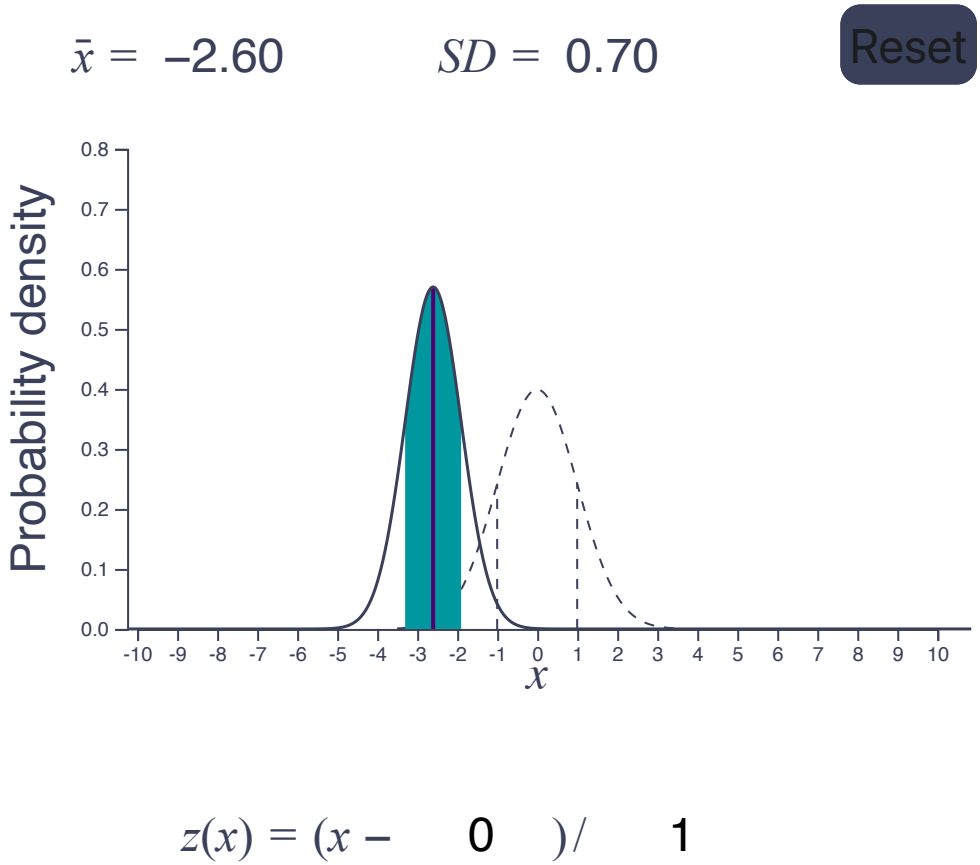
Just like with centring, the fundamental shape of the variable's distribution did not change as a result of scaling.

Unlike centring, however, scaling does change the scale on which the variable is measured. After all, that's why it's called scaling...

Unfortunately, the scale on which the variable in panel B of Figure 6 is shown is not very interpretable: the individual numbers don't mean much.

To make scaling by standard deviation more useful, we need to make sure the variable is mean-centred first!

To understand why, have a little play around with the visualisation below. For the time being, leave the subtraction control (purple circle) alone. Instead, drag the division control (green circle) left and right to see what happens.



Processing math: 100%

Transforming variables

Drag the controls (purple and green circles) left and right to change the value by which the variable plotted gets transformed. Pay attention to the order of operations: once you move the division control, you won't be able to change the subtraction control.

Double-click on the control to reset it.

Click Reset to get a new example.

Dashed curve represents the standard normal distribution.

As you can see, when you scale a variable that has not been mean-centred, the mean of the variable is also affected by the scaling factor. The mean of the resulting variable will thus be the original mean / the scaling factor. This is because the entire stretching/squishing operation happens with respect to the 0 point: everything moves further away/ closer to 0. This was quite useful when we were transforming height from the centimetre scale to the metre scale. We wanted the mean to change proportionally! If the mean height in the sample is 165 cm, we want the value to scale when converting to metres: it should be 1.65 m, not 165 m.

But, if we're scaling the variable by it's standard deviation, the resulting mean of $\bar{x}/SD$ is not tremendously interpretable. This is what happened in panel B of Figure 6

Go back to the visualisation and this time, mean-centre the variable before you start changing its scale.

Do it!

In this scenario, it is still true that the scaling happens with respect to the 0 point on the x-axis. But now, the 0 is the centre of our variable and so, its mean does not get affected by scaling. That's because for a mean of 0, $\bar{x}/SD = 0$ irrespective of the value of SD.

Remember to mean-centre a variable before you scale it by its *SD*!

The *z*-transform

The combination of first mean-centring and then scaling a variable by its *SD* is known as the *z*-transform or standardisation. The formula for this function, let's call it z, is:

$$z(x) = \frac{x - \bar{x}}{SD(x)}$$

Figure 7 shows what a *z*-transformed height variable looks like compared to its original form.

Plot R code

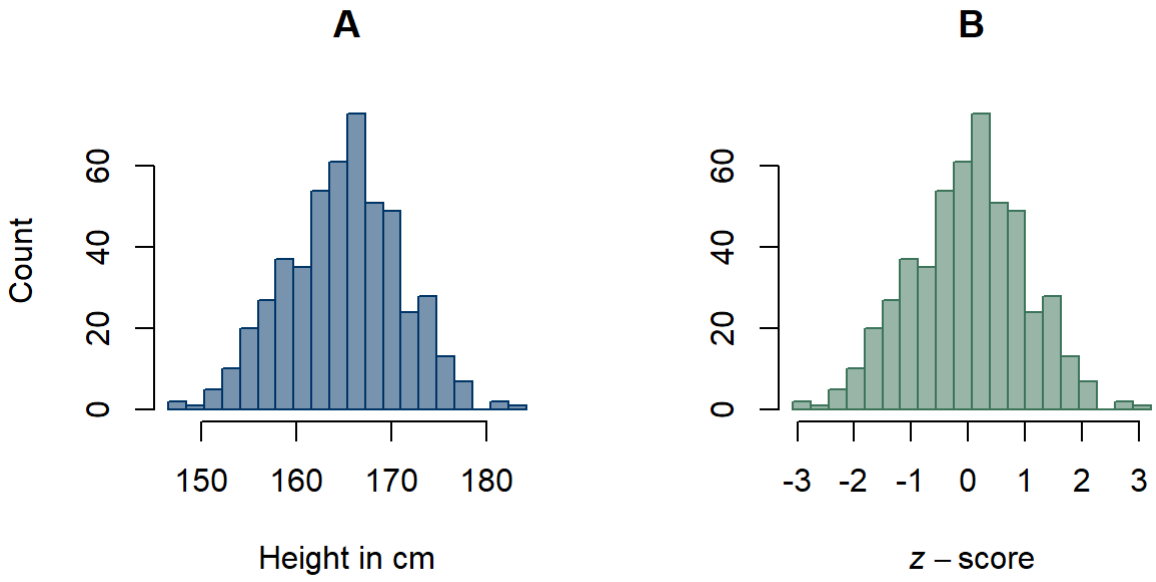


Figure 7: Histograms of participants heights: (A) raw data (B) *z*-transformed data.

Again, the shape of the variable remains intact and the relative differences between any two values in the variable are preserved. That's because standardisation is a linear transformation, just like addition and multiplication.

This time, however, the scores are interpretable. Standardisation (or *z*-transformation) converts values of a variable to numbers that can be interpreted in terms of the number of standard deviations from the mean.

Return to the visualisation once again and try *z*-transforming the variable. It should perfectly fit the dashed curve representing the *standard normal* distribution (a normal distribution with $\mu = 0$ and $SD = 1$).

Try it now!

z-scores

Values of a standardised/*z*-transformed variables are called *z*-scores. To repeat, they are interpreted as the distance from the mean in units of standard deviation.

This interpretation is independent of the actual value of *SD* in the original variable!

So, for example, let's say that, in the sample of 500 women whose height we've been plotting over and over again, the mean height was 164.94 cm with a *SD* of 5.99 cm. Let's say we standardised the variable. A person with a *z*-score of 1 will be *one* SD *taller than average*: 164.94 + (1 × 5.99) = 170.93 cm. Someone with a *z*-score of &mius;0.8 will be 0.8 *SD* shorter than the average person in the sample: 164.94 + (−0.8 × 5.99) = 160.15 cm.

R know-how

The interpretation of *z*-scores is the same no matter what variable we are working with! This is very useful for comparing scores on different variables or across different groups.

But before we can discuss this topic, we need to talk about comparing means.

Making comparisons

Apples and apples: Comparing groups

When we talk about comparing groups on some variable in the context of quantitative research, we are talking about looking at the average difference in that variable between groups. In other words, we are asking, how different are the groups *on average*.

Let's extend our, at this stage well-milked, height example and imagine that, in addition to the 500 women, we also measured the heights of 500 non-binary people. Suppose, we find that the mean height in the two groups is $\bar{x}_{\text{w}}$ = 164.94 and $\bar{x}_{\text{nb}}$ = 170.73, respectively. Figure 8 shows a grouped histogram of these data.

Plot R code

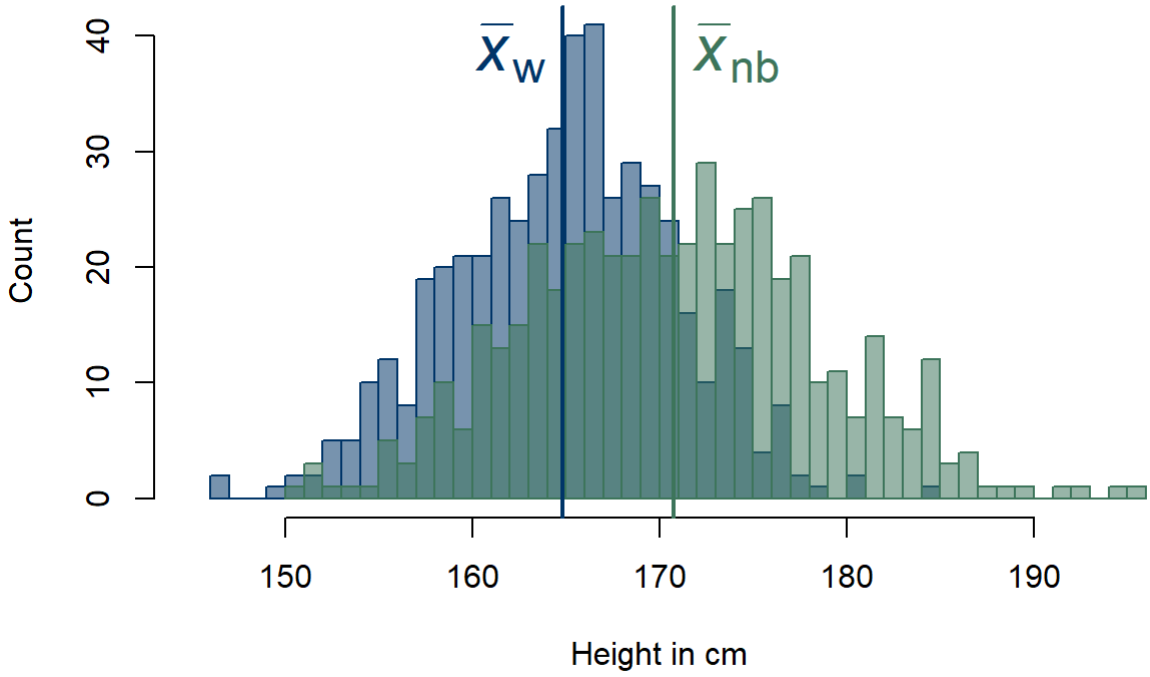


Figure 8: Distribution of height on samples of 500 women (blue) and 500 non-binary people (green) with group means indicated by vertical lines. This is not real data.

In our simulated example data, there is a substantial overlap between the two distributions. However, you can also clearly see that there is a difference in *average* heights of women and non-binary people. To quantify this difference, all we need to do is subtract the mean of one group from the mean of the other:

$$\begin{aligned} diff_{\text{height}} &= \bar{x}_{\text{w}} - \bar{x}_{\text{nb}} \\ &= 164.94 - 170.73 \\ &= -5.79 \end{aligned}$$

The sign indicates the direction of the difference. If the number is positive, that means that the first group's mean is larger than that of the second group. If the number is negative, the opposite is true. Of course, it is completely arbitrary which group is *first* and which is *second*.

Differences between groups vs differences within groups

People sometimes argue against the existence of meaningful differences between all kinds of groups of people by claiming that any differences observed between said groups are smaller than differences that exist *within* each of the groups. Therefore, the argument goes, these between-group differences are irrelevant or actually non-existent.

Regardless of how sympathetic we may or may not be towards the motivation behind these arguments, they are not statistically literate. Take another look at Figure 8: Height of the people in our non-binary sample ranges from 150 cm to 195 cm. And that's just in our sample. The range in the population is bound to be wider than 45 cm.

If we are to follow the argument above, the difference between the height of women and non-binary people would have to be about half a metre (that's about 20 in) to recognise that there's actually a difference. That's really not a reasonable criterion for what magnitude of differences should be considered *"real"*.

There are better ways of tackling arguments of those who aim to divide people into fixed categories so let's chuck this one out!

Remember: To get the mean difference between two groups *on the same variable*, subtract the mean of one of the groups from the mean of the other.

Apples and oranges: Comparing across groups/variables

Nyari is a 172 cm tall woman. Karim is a 179 cm tall non-binary person. Comparing their respective heights is trivial: we can easily tell that Karim is taller than Nyari. But what if we wanted to know how their heights compare relative to their groups/populations? Is Nyari taller as a woman than Karim is as enby?

To answer this question, we can use... you guessed it, *z*-scores. By standardising the height variable of each group, we get variables that are on the same scale: the scale of units of *SD*.

Remember the formula for the *z*-transform: $z(x) = \frac{x - \bar{x}}{SD(x)}$.

Let's use this formula to calculate Nyari's and Karim's height *z*-scores. To do that, all we need is to know the means and standard deviations of their respective groups. Here are the statistics in our pretend samples shown in Figure 1:

	$\bar{x}$	*SD*
Women	164.94	5.99
Non-binary	170.73	7.74

This is all the info we need to calculate the *z*-scores for our two guys. Why not do it in R?

```
# brackets are important because of order of operations!
z_nyari  <-      (      172      -      164.94  )      /      5.99
z_karim  <-      (      179      -      170.73  )      /      7.74
z_nyari
```

```
[1] 1.178631
```

```
z_karim
```

```
[1] 1.068475
```

So now we know that Nyari's *z*-score of height is about 1.18, while Karim's is 1.07. Nyari is 1.18 standard deviations *taller than the average woman* and Karim is 1.07 *taller than the average non-binary person*. That means, that Nyari, despite being shorter in absolute terms, is actually *relatively taller* (relative to her population).

We could use the same principle to compare values of variables measured on any scale. Imagine Nyari earns £38,400 per year here in the UK. She just got a job offer in Germany with an agreed salary of EUR 4,270 per month. Is she going to be relatively better off if she takes the job?

Again, we can use *z*-scores to see how the two salaries fare with respect to what people in the UK and Germany make. Say, the average *annual* wage in the UK is £37,428 (*SD* = 4,266), while the average *monthly* wage in Germany is EUR 3,880 (*SD* = 351.6). Again, we can use R to calculate the *z*-scores:

```
mean_annual_UK <-      37428
sd_annual_UK <-      4266
mean_monthly_GER <-      3880
sd_monthly_GER <-      351.6

z_nyari_UK <-      (      38400    -      mean_annual_UK)      /
sd_annual_UK
z_nyari_GER <-      (      4270     -      mean_monthly_GER)
/      sd_monthly_GER
z_nyari_UK
```

```
[1] 0.2278481
```

```
z_nyari_GER
```

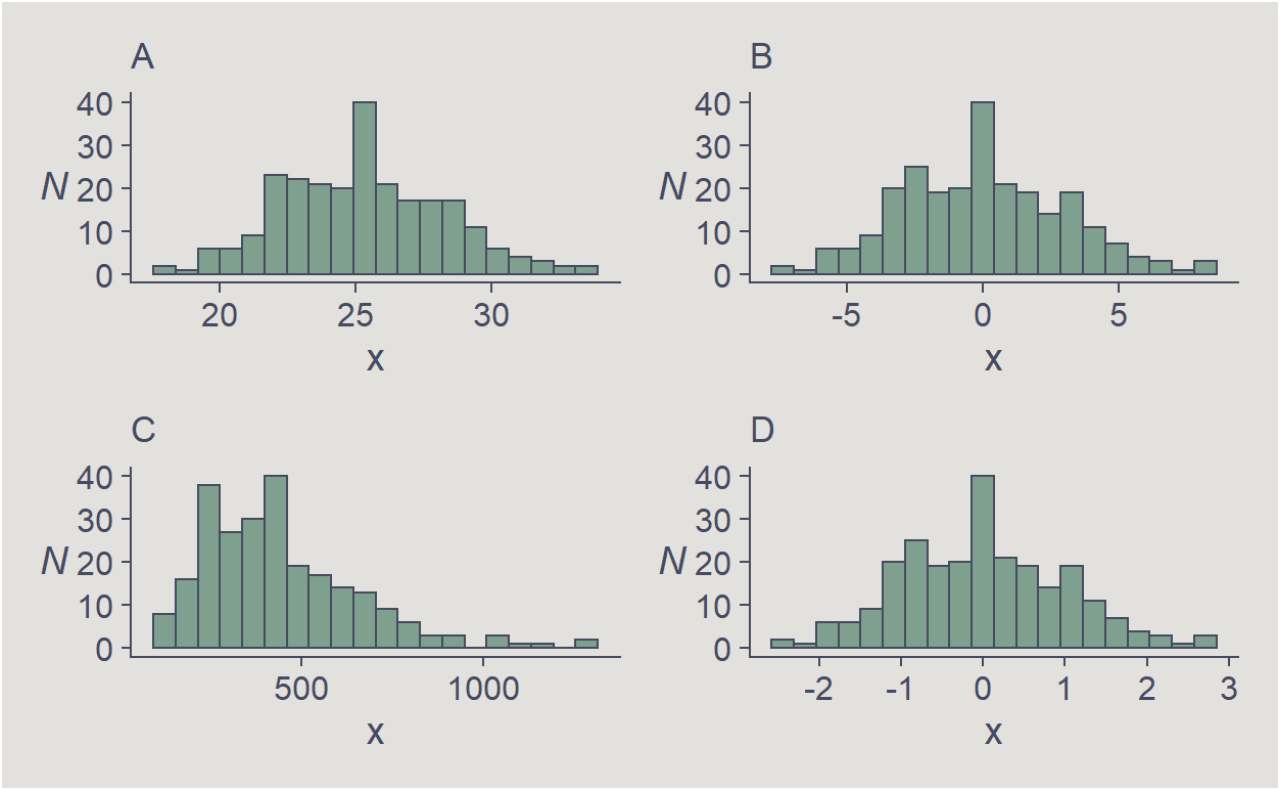
```
[1] 1.109215
```

As you can see from the two *z*-scores, Nyari would make substantially more relative to what people in the country make if she took the job in Germany.

> *z*-scores are very useful for comparing values of variables measured on different scales or in different units!

Check your understanding

Here is a set of plots of the same random normal variable after various transformations:



QUESTION 1

Which of the plots shows a standardised (*z*-transformed) variable?

A

B

C

D

QUESTION 2

Which of the plots shows a non-linear transformeation?

A

B

C

D

QUESTION 3

Assuming the variable in question is age and the sample is drawn from a population of university

D

QUESTION 4

What is the *approximate* standard deviation of the untransformed distribution?

1

3

5

7

Here are the data on the heights of two groups of people:

	Group 1	Group 2
1	165.0	173.5
2	169.1	175.4
3	168.1	166.7
4	168.8	160.0
5	174.4	166.2
6	172.1	167.5
7	165.0	166.8
8	159.0	168.7
9	160.9	166.3
10	180.6	172.5

QUESTION 5

What is the mean of Group 1?

Give answer to 2 decimal places.

Submit

QUESTION 6

What is the standard deviation of Group 1?

QUESTION 8

What is the *z*-score of the data point in row 9 with respect to Group 2?

Give answer to 2 decimal places.

Submit

QUESTION 9

Without calculating anything, which row in Group 1 has the largest *z*-score?

Submit

QUESTION 10

Without calculating, what number do we nead to *subtract* from the scores in Group 2 in order for the data to have a mean of 0?

404.06
168.36
-168.36
-323.25

Recap

In this lecture, we talked about how we can think about the distributions of variables in terms of the normal curve and how the mean and *SD* reflect the position and spread of this curve.

We also introduced the concept of one kind of mathematical functions called transformations. We saw, how some transformations, such as centring or scaling don't change the relative distances between individual values of a variable. These are linear transformations. Others, such as

converts the values of any variable into units of *how far the value is from the mean of the whole variable in terms of numbers of standard deviations*.

Finally, we learned how we can compare group averages by *subtracting the means of the groups* and how we can use *z*-scores to compare values of variables measured on different scales or in different units.