Lecture 8: Distributions, functions, transformations

A little maths goes a long way

Dr Milan Valášek
16 November 2020

UNIVERSITY OF SUSSEX

Overview

The shape of things

- Histograms
- The normal curve

Transformations

- Functions
- The z-transform

Comparing things with maths

- Comparing groups
- Comparing scores across groups
- Comparing scores across variables

The shape of things

For the purpose of this lecture, we will only be talking about *continuous* variables!

- The vast majority of the measured heights are roughly in the 155-175 centimetre range
- The distribution is roughly symmetrical around its mean and has the shape of a bell characteristic of a *normal distribution*
    - The shape isn't perfectly smooth in a finite sample

Plot | R code

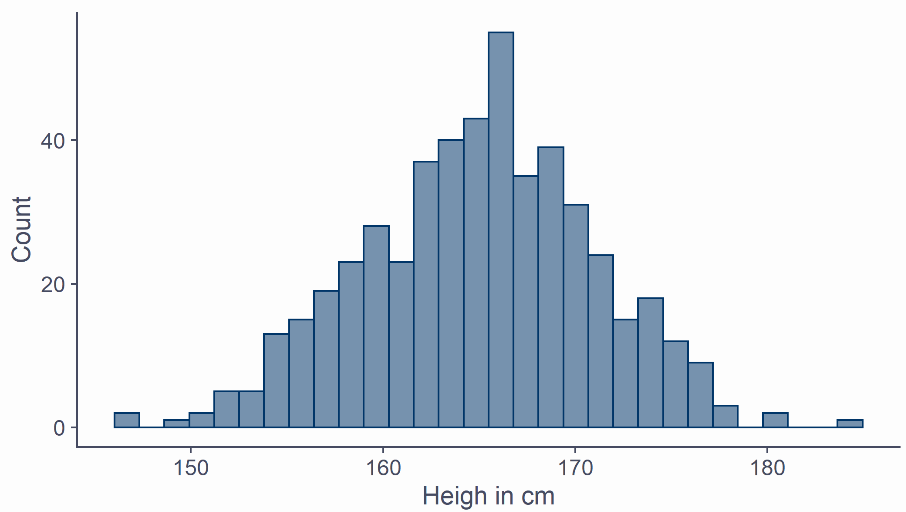


Fig 1 Distribution of height on a sample of 500 women. This is not real data.

Histograms

- Height is a continuous variable so 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 represents the number of people whose height falls within a given range

Plot | R code

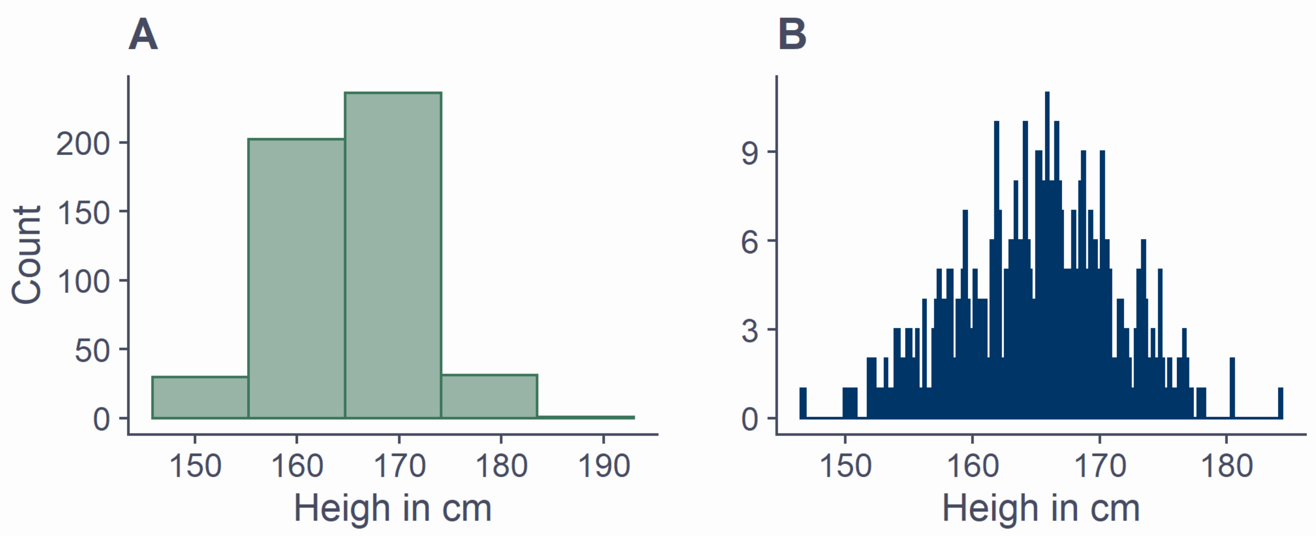


Fig 2 Histograms with (A) too few bars to see the distribution in enough detail and (B) too many bars.

Ideal curves

- If we could collect an infinite number of observations, we could make the bins *infinitely* narrow
- This would give us an idealised shape of the normal distribution: the normal curve.
- Because we will mostly be talking about continuous normal variables, we can visualise them as this kind of curve

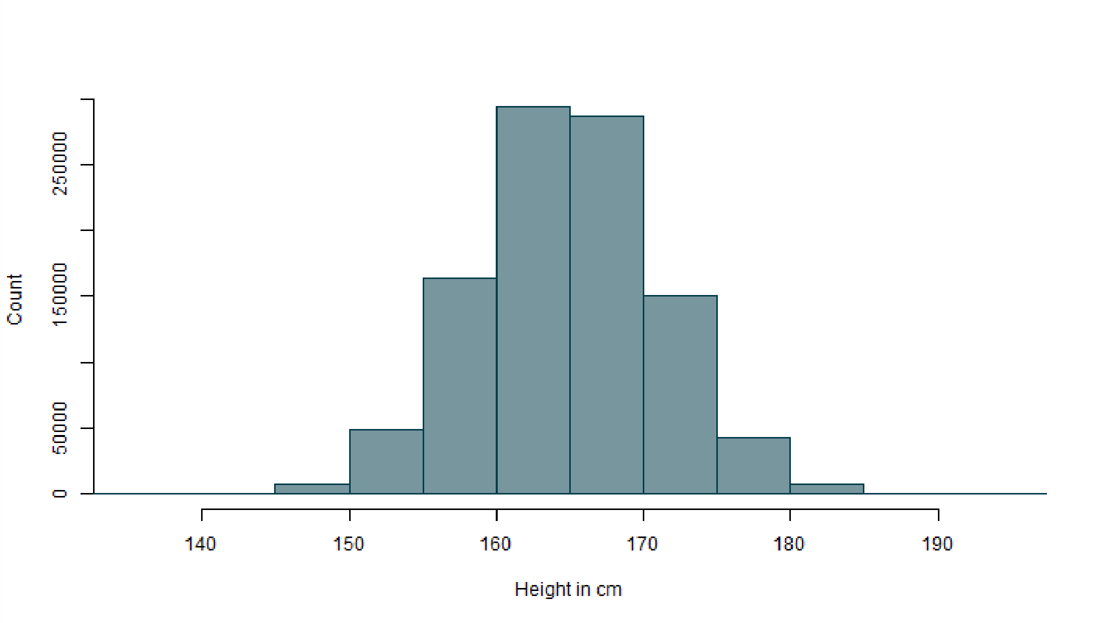
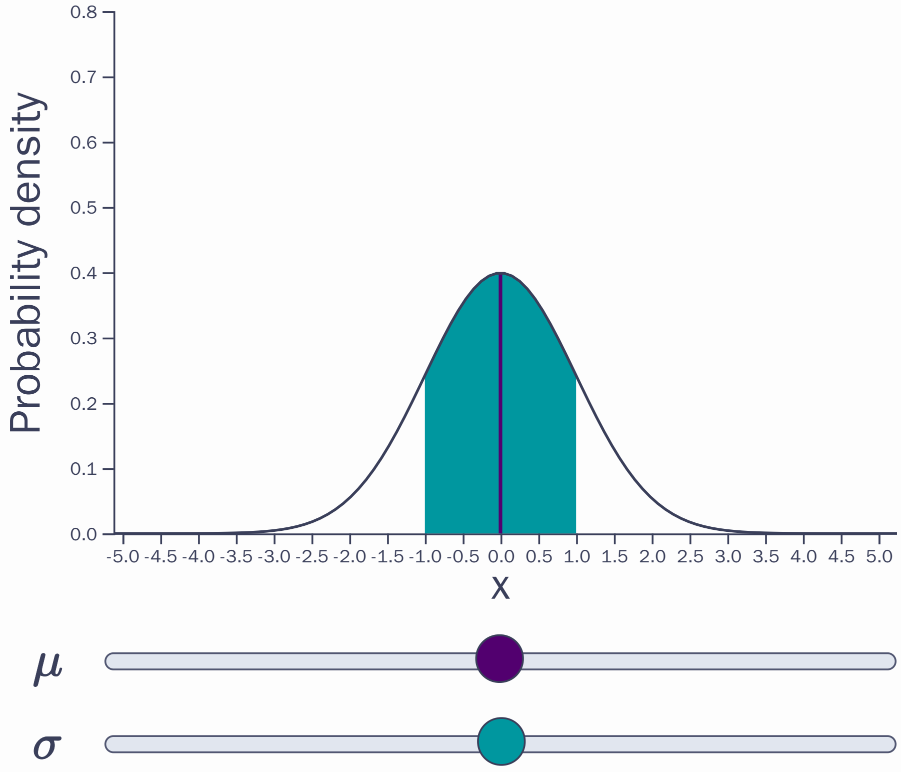


Fig 3 From histogram to an idealised shape.

The normal distribution

- We can describe key properties of a variable using measures of *central tendency* and *spread*
- 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 normal distribution



The normal distribution

- Mean and standard deviation are independent of one another
- Neither shifting the mean, not changing the standard deviation of a distribution doesn't change its *fundamental shape*
    - Relative position of the individual points on the line with respect to each other does not change!
    - It is still true that about 68% of values are within ±1 standard deviation from the mean

Same shape, different scale

Plot | R code

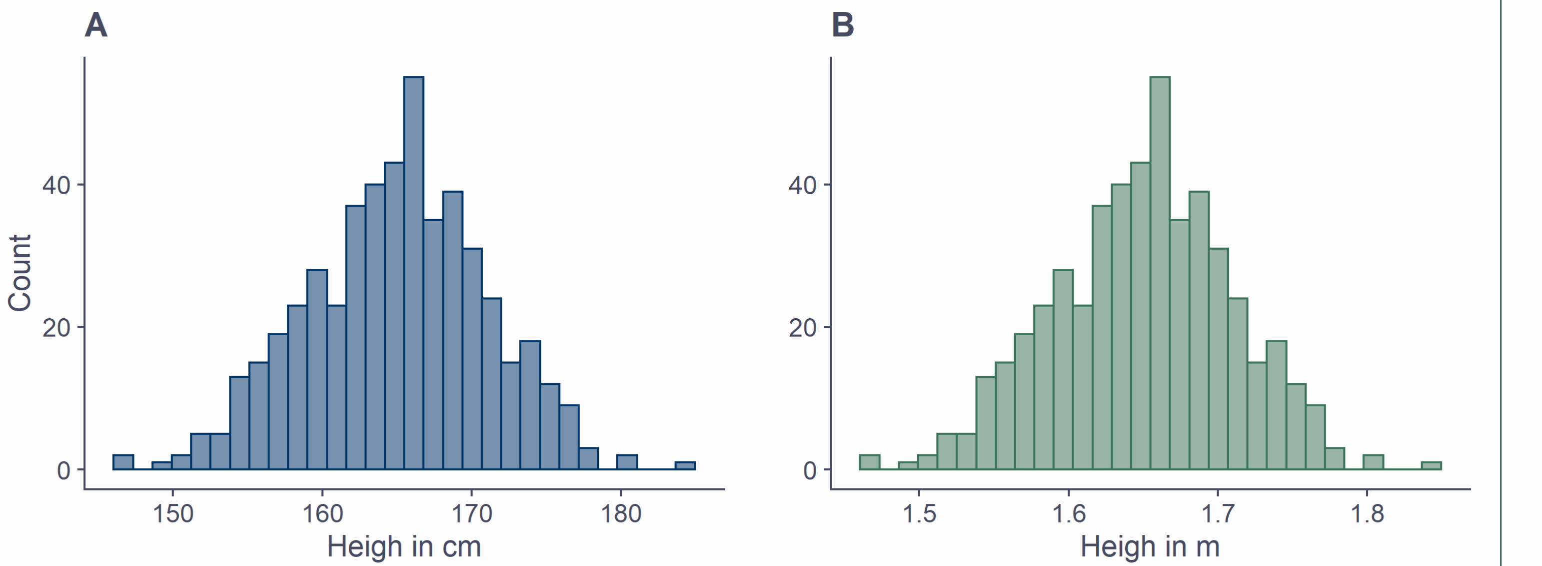


Fig 4 Histograms of participants heights measured in (A) centimetres and (B) metres.

Transformations

(From now on we'll be talking about sample mean, $\bar{x}$, and sample standard deviation, SD)

- How do we change $\bar{x}$ and SD without changing the shape of the variable?
  - Only changing the values of a selection of observations will alter the shape of the distribution - *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
- This *preserves the relationships between individual observations*!

Functions

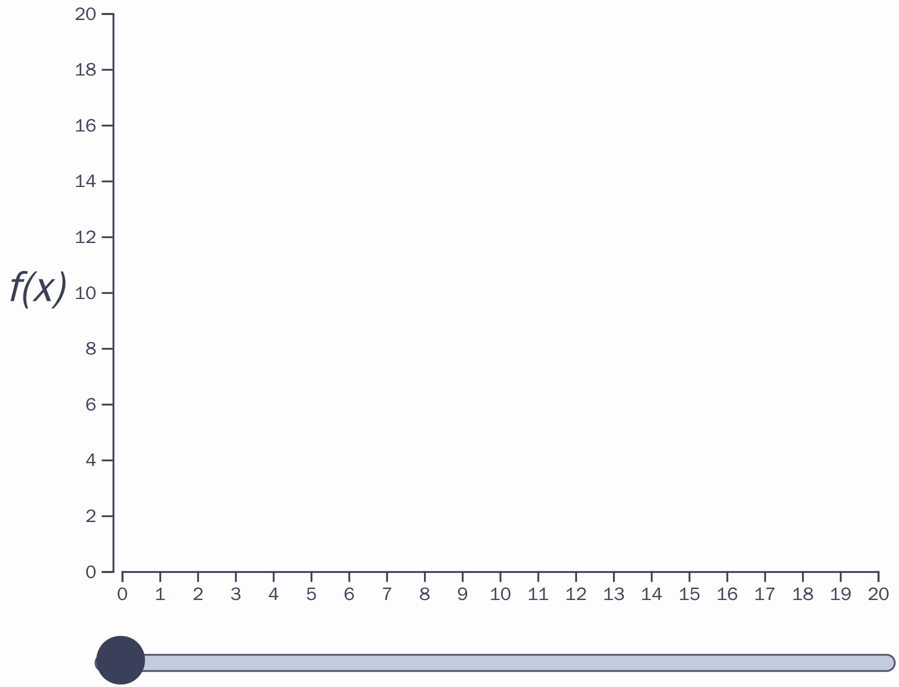
Let's play a game!

Functions

- CONGRATS! You have just discovered the *identity* function: $f(x) = x$
- A transformation is just a mathematical function that takes an input and returns an output (just like a function in R)
- 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.

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

Graph of *f(x)*



Slide to change value of x

Centring and scaling

- Addition shifts the values of x up and down along the y-axis, while keeping the distances between points unchanged
- Multiplication, spreads or "squishes" the values of x along the y-axis
- 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
- You can technically centre a variable by subtracting *any* value from it but the most frequently used method is mean-centring:

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

- Mean-centring does not alter the shape of the variable, nor does it change the scale at which the variable is measured

Plot | R code

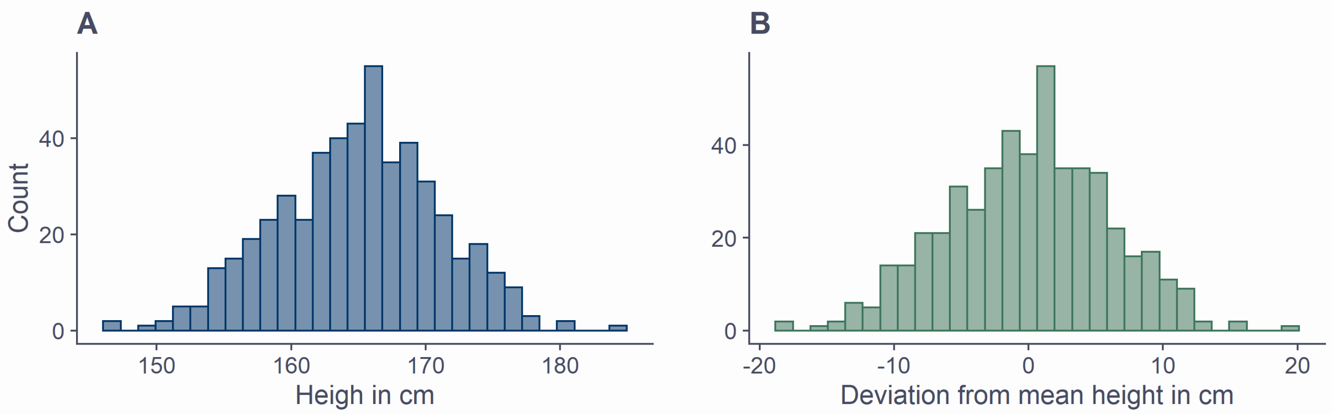


Fig 5 Histograms of participants heights: (A) raw data (B) mean-centred.

Scaling

- Scaling is the division of each observation of a variable by a fixed value
- This has the effect of stretching or squishing the entire variable *in the direction of the x-axis*
- The most frequent method of scaling variables is by their standard deviation:

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

Plot | R code

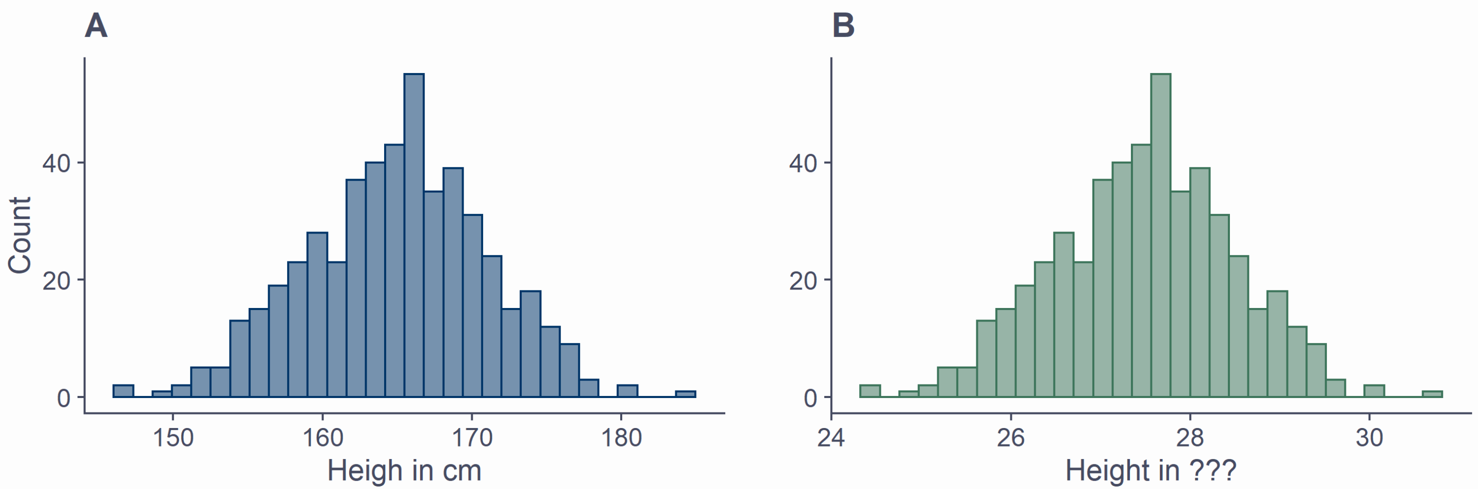
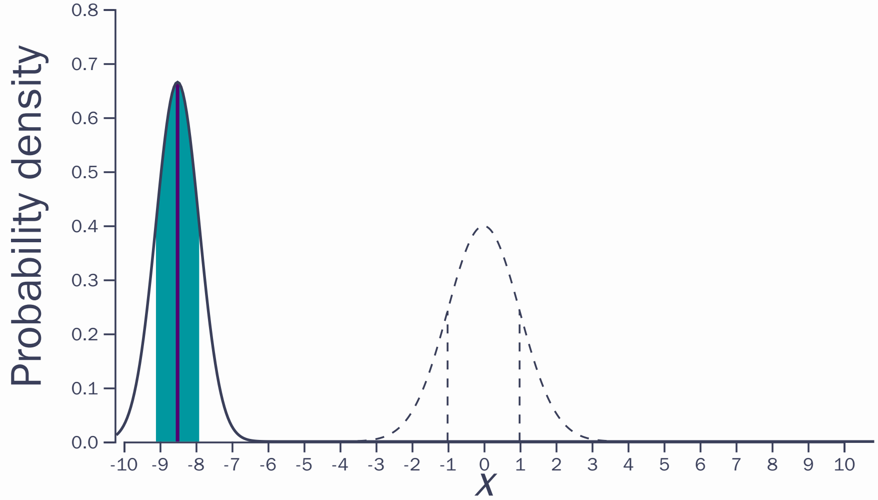


Fig 6 Histograms of participants heights: (A) raw data (B) scaled by *SD*.

The *z*-transform

$\bar{x} = -8.50$ $SD = 0.60$ Reset



$z(x) = (x - 0)/1$

The *z*-transform

- First mean-centring and then scaling a variable by its *SD*
- AKA, standardisation.

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

- Shape of the variable remains intact and the relative differences between any two values in the variable are preserved
    - Standardisation is a linear transformation (like addition and multiplication)

Plot | R code

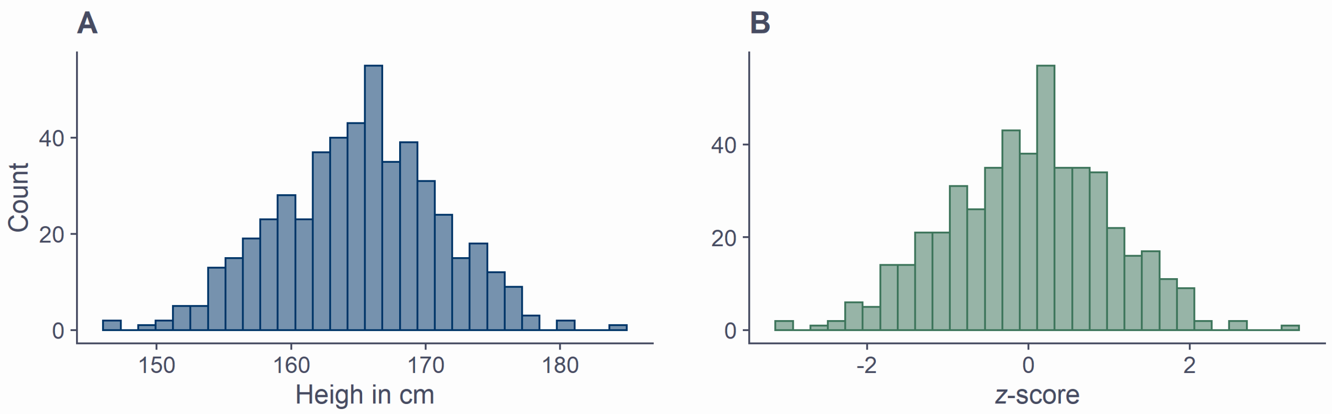


Fig 7 Histograms of participants heights: (A) raw data (B) *z*-transformed data.

z-scores

- Values of a standardised/*z*-transformed variables
- Distance from the mean in units of standard deviation.
- This interpretation is independent of the actual value of *SD* in the original variable!
- A person with a *z*-score of 1 will be *one* SD *taller than average*: $164.98 + (1 \times 6) = 170.98$ cm.
- Someone with a *z*-score of -0.8 will be 0.8 *SD* shorter than the average person in the sample: $164.98 + (-0.8 \times 6) = 160.17$ cm.

Comparing groups

We can compare groups by asking how different are the groups *on average*.

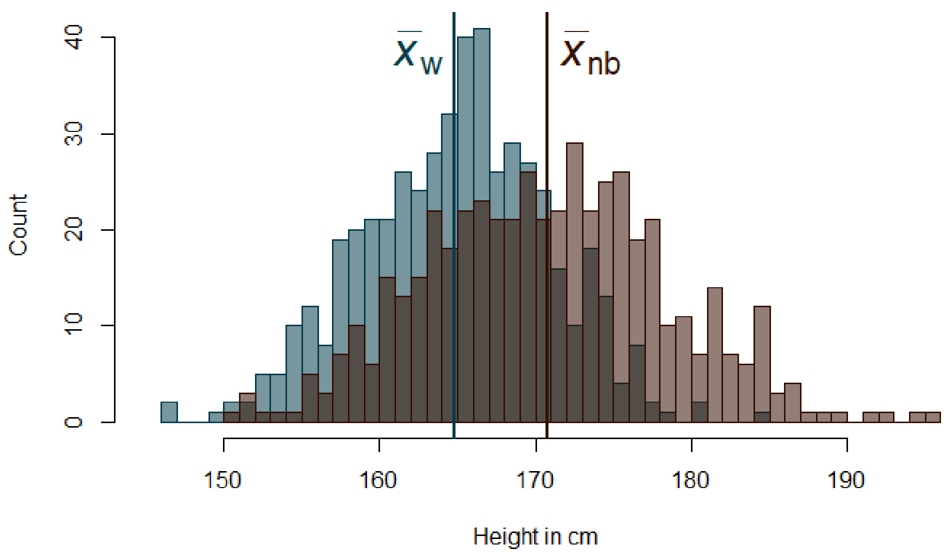


Fig 8 Comparing distributions of heights in a sample of women (green) and non-binary people (brown). Data are not real.

$$
\begin{aligned}
diff_{\text{height}} &= \bar{x}_{\text{w}} - \bar{x}_{\text{nb}} \\
&= 164.98 - 170.74 \\
&= -5.77
\end{aligned}
$$

Comparing across groups

Nyari is a 172 cm tall woman; Karim is a 179 cm tall non-binary person

What if we wanted to know how their heights compare relative to their groups/populations?

We can use *z*-scores: $z(x) = \frac{x - \bar{x}}{SD(x)}$

	$\bar{x}$	SD
Women	164.98	6.00
Non-binary	170.74	7.74

```r
(172 - 164.98) / 6 # Nyari
```

```
## [1] 1.17
```

```r
(179 - 170.74) / 7.74 # Karim
```

```
## [1] 1.067183
```

Comparing across variables

- We could use the same principle to compare values on of variables measured on any scale
- 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?
- Average *annual* wage in the UK is £37,428 (*SD* = 4,266)
- Average *monthly* wage in Germany is EUR 3,880 (*SD* = 351.6)

```
(38400 - 37428) / 4266 # Nyari's UK salary z-score
```

```
## [1] 0.2278481
```

```
(4270 - 3880) / 351.6  # Nyari's German salary z-score
```

```
## [1] 1.109215
```

Recap

- We often think about the distributions of variables in terms of the normal curve
- Mean and *SD* reflect the position and spread of this curve
- Transformations are mathematical functions we can use to manipulate variables
- 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 *exponentiation* (*e.g.*, x^2) do change the proportions of the transformed variables
  - These are non-linear transformations

Recap

- The z-transform, AKA standardisation, is a two step transformation consisting of *first* mean-centring the variable and then scaling it by its *SD*
- It 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*
- We can compare group averages by *subtracting the means of the groups*
- We can use z-scores to compare values of variables measured on different scales or in different units


That's all Folks!