Week 10: Continuous Probability Distributions

Data Analysis for Psychology in R 1

Alex Doumas

Department of Psychology
The University of Edinburgh

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

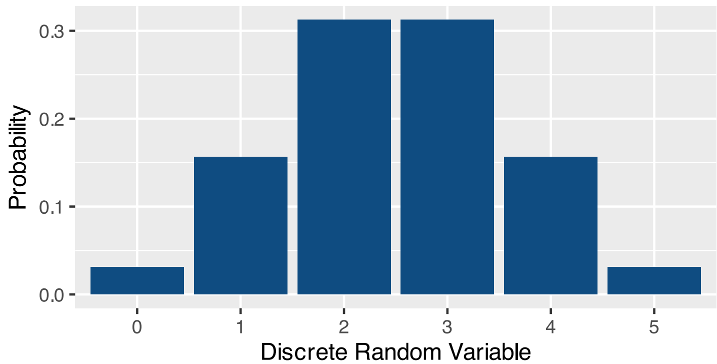
Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

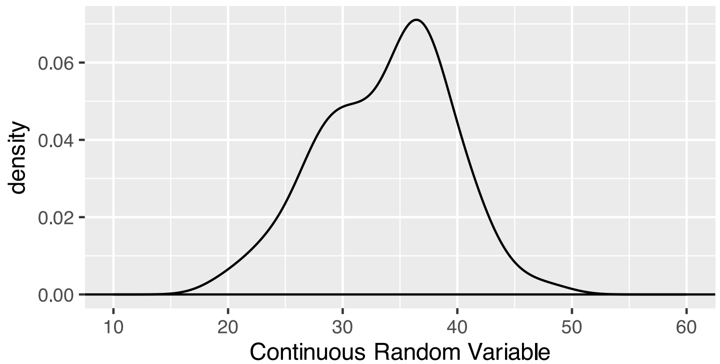
This Week's Learning Objectives

1. Understand the key difference between discrete and continuous probability distributions
2. Apply understanding of continuous probability distributions to the example of a normal distribution
3. Understand how to use a range from a continuous probability distribution
4. Introduce other continuous probability distributions

Discrete vs. continuous

- Recall that a *discrete probability distribution* describes a random variable that produces a discrete set of outcomes
- By contrast, a *continuous probability distribution* describes a random variable that produces a continuous set of outcomes
    - Temperature
    - Height
    - Reaction Time
- If you have arbitrary precision of measurement, you have a continuous random variable
- While a discrete probability distribution is jagged, a continuous probability distribution is smooth





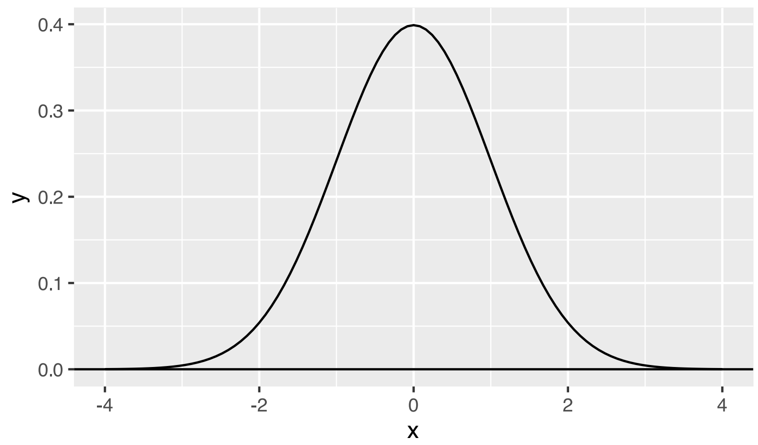
Discrete vs. continuous

- Continuous probability distributions differ from discrete in two other important ways
  - $P(X = x) = 0$
  - Continuous probability distributions are described using the **probability density function (PDF)**, rather than the **probability mass function**
- Now, let's take a look at perhaps the most widely used continuous probability distribution...

Normal distribution

- A **normal distribution** (AKA the Gaussian distribution) is a continuous distribution

- It is uni-modal (one peak) and symmetrical



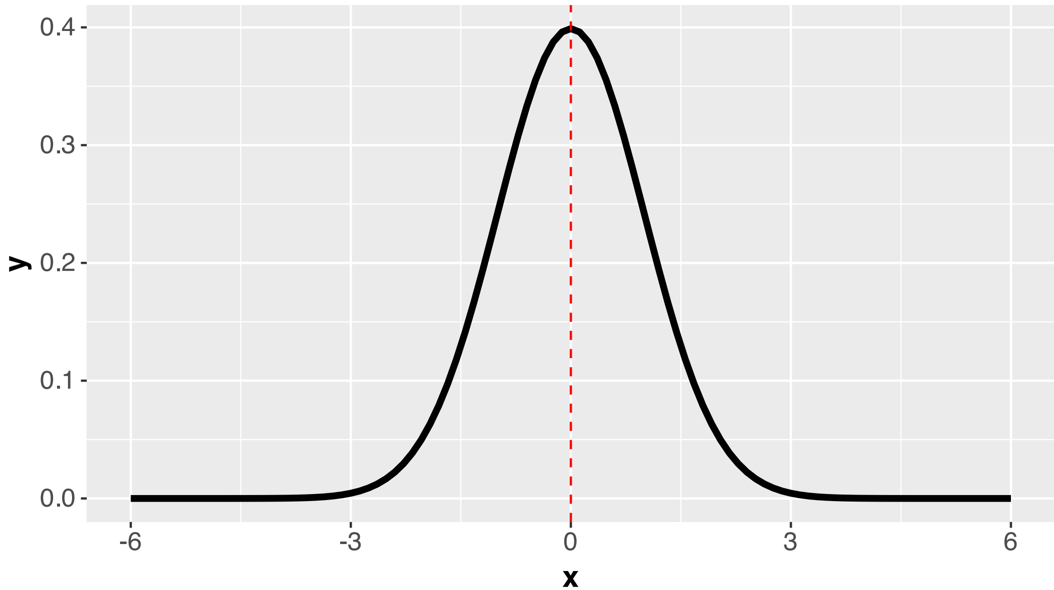
Normal: PDF

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- A bit scary!
- But the basic points are:
    - It is a function of data *x*
    - And *two* parameters μ and σ (mean and SD)
- There is not one single normal distribution
- We have a family of different distributions that are defined by their mean, μ, and standard deviation, σ

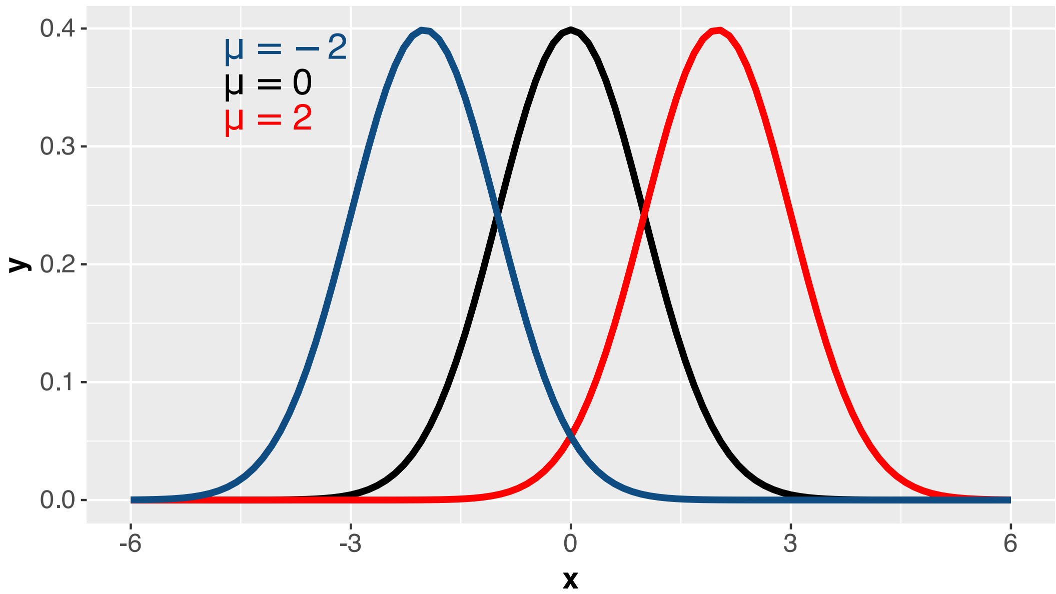
The Standard Normal Distribution

- The **standard normal distribution** is a normal distribution where $\mu = 0$ and $\sigma = 1$



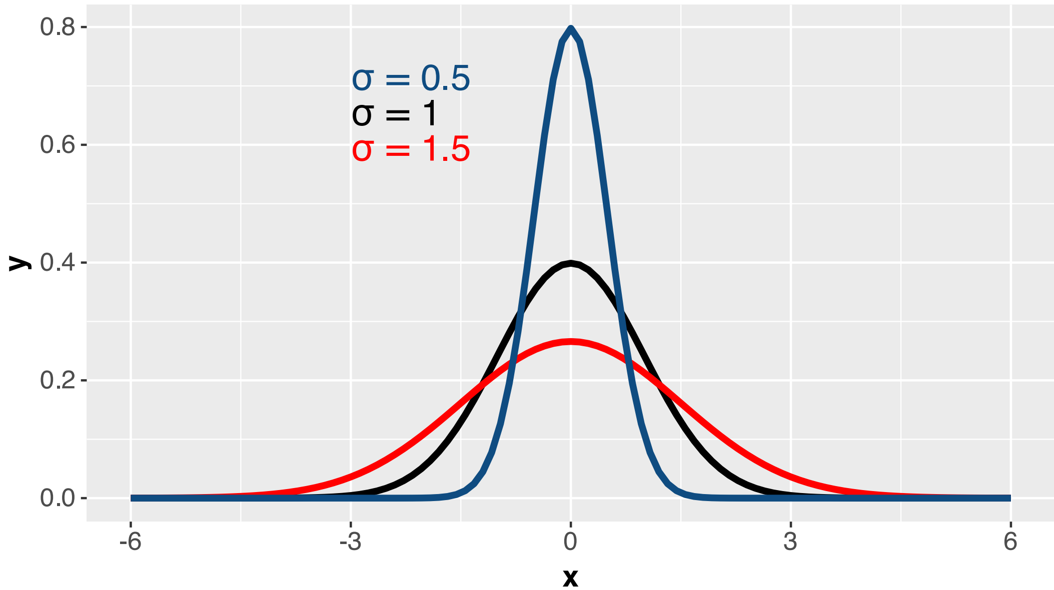
Different Normal Distributions - Adjusting μ

- Adjusting μ changes where the curve is centered on the x-axis



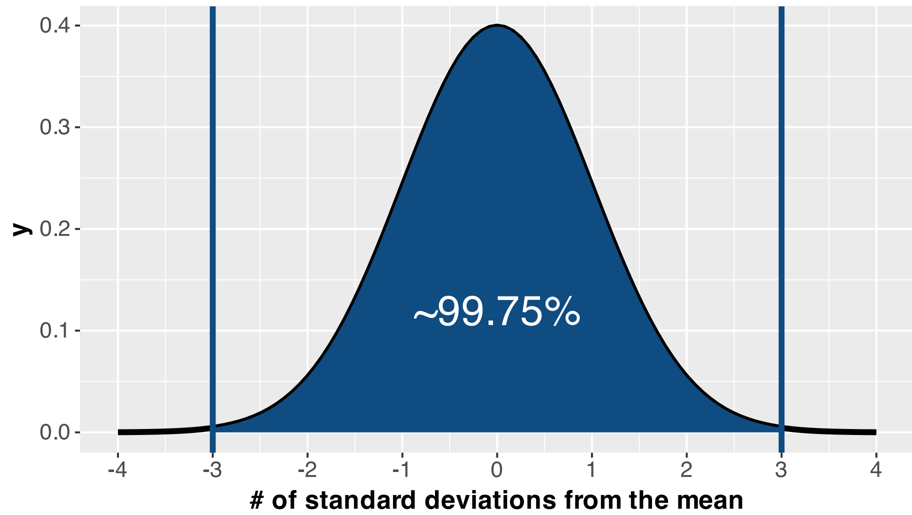
Different Normal Distributions - Adjusting σ

- Adjusting σ changes the shape of the curve



Properties of Normal Distributions

- Properties of any normal distribution:
  - $\approx$ 68% of area falls under 1 SD on either side of mean
  - $\approx$ 95% of area falls under 2 SD on either side of mean
    - *Exactly* 95% falls under +/- **1.96 SD**
  - $\approx$ 99.75% of area falls under 3 SD on either side of mean



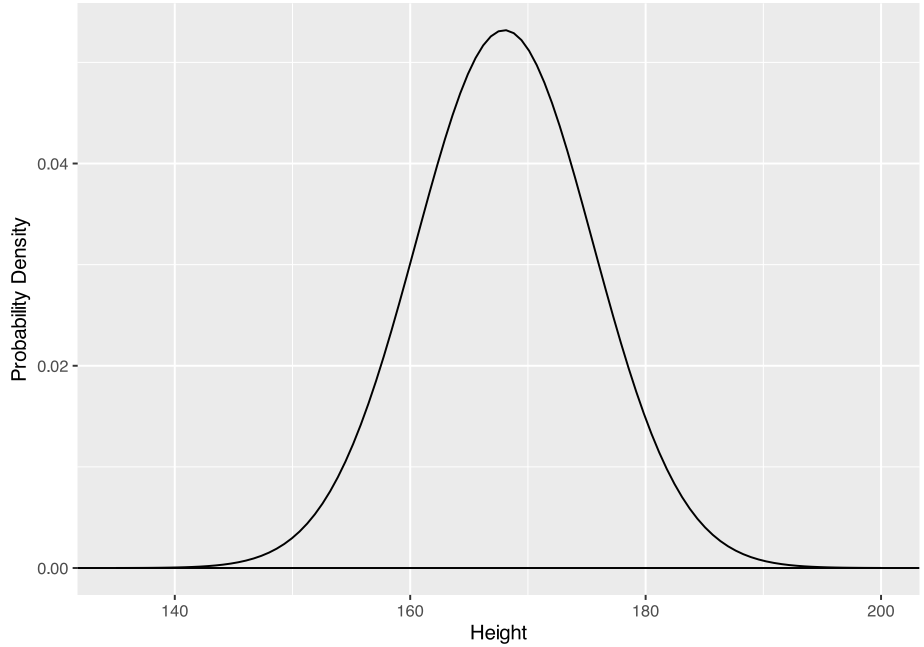
Questions?

Using the PDF of the normal distribution

- Let's use the normal distribution to illustrate how continuous probability distributions work
- With a discrete random variable it makes sense to ask: 'What's the probability associated with a specific value of the random variable?'.
    - e.g. what the probability of getting heads on a fair coin?
- With a continuous random variable it makes sense to ask about ranges of scores
    - e.g. what's the probability of sampling someone between 1.75 and 1.8 meters tall if we sample students from a university?
    - Remember that the probability of any single value (e.g. exactly 1.764736525678943655 meters) is 0
    - The total probability (1) is divided between an infinite number of possible values that the variable could take, as the variable is continuous

Using the PDF of the normal distribution

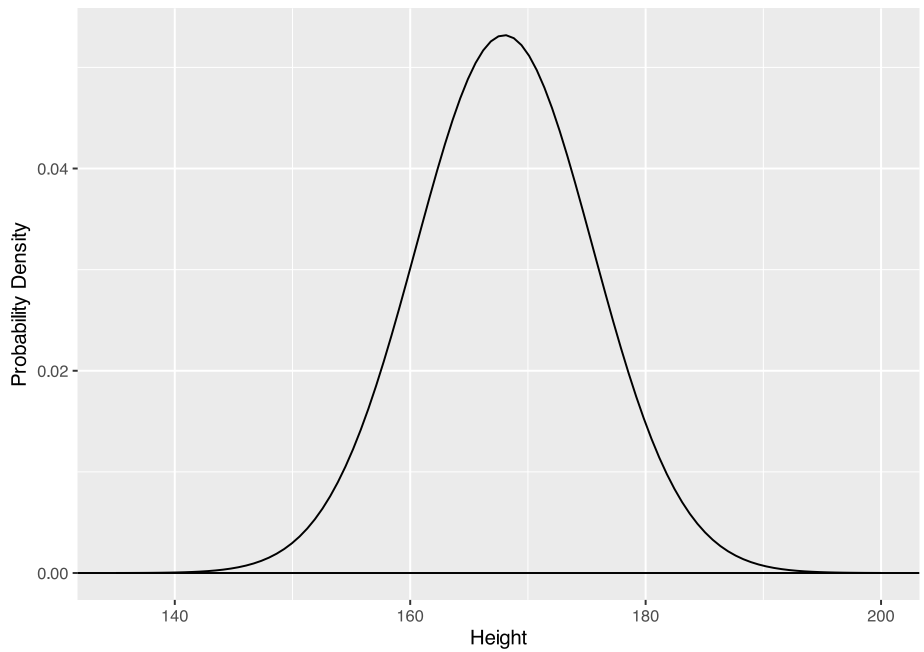
- Let's imagine that in some course, student height is normally distributed
  - $\mu = 168$ cm
  - $\sigma = 7.5$ cm
- We can ask what is the probability of sampling someone between 175 and 180 cm?
  - This question translates to: $P(175 \leq x \leq 180) =?$
  - Let's unpack this...



Using the PDF of the normal distribution

$$P(175 \leq x \leq 180) = ?$$

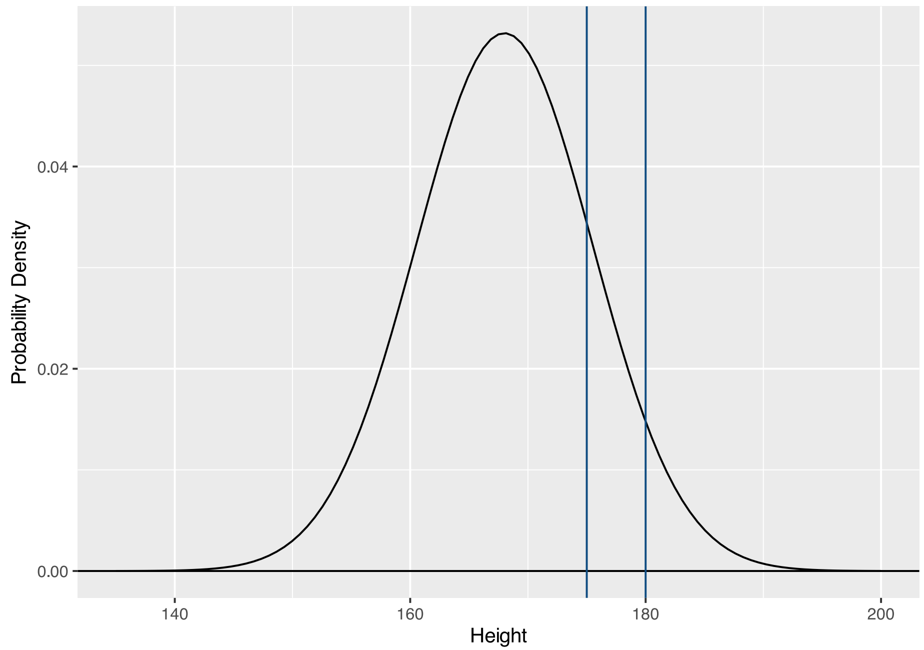
- Let's draw these boundaries on our plot



Using the PDF of the normal distribution

$$P(175 \leq x \leq 180) = ?$$

- Let's draw these boundaries on our plot
- What is the value of the area under the curve between these two lines?

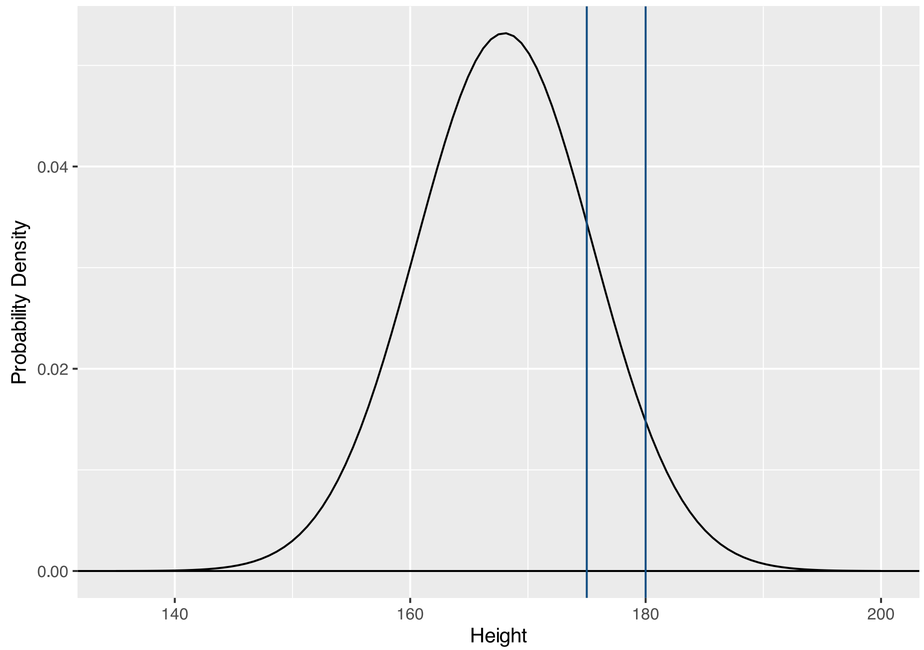


Using the PDF of the normal distribution

- We get the area under a curve by calculating an integral

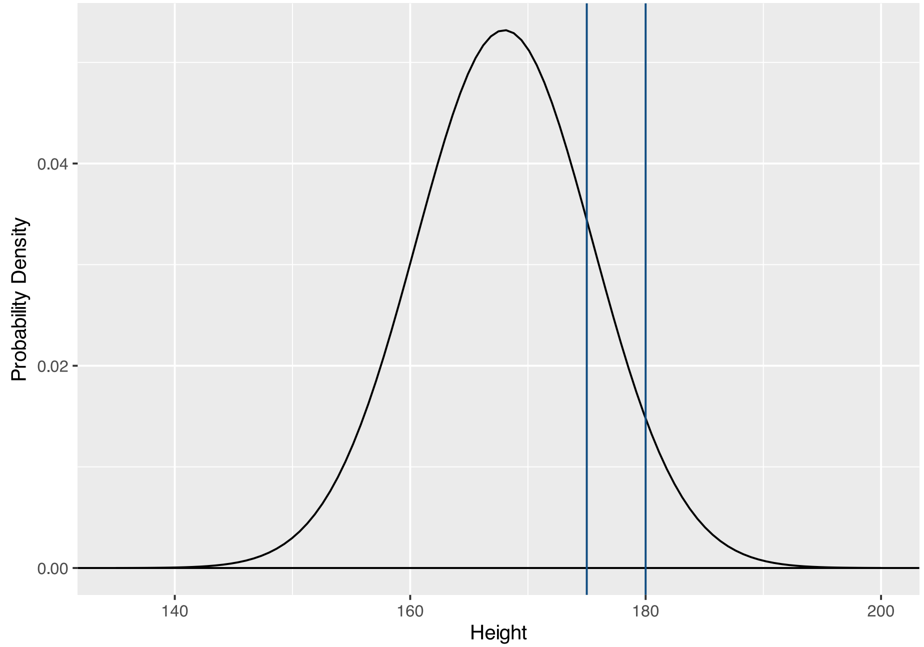
$$\int_a^b f(a)\,dx$$

  - Don't worry, you don't need to know the details of integrals, but you may encounter the equation above
  - This equation can be read as: The integral of values falling between vertical lines a and b on the function a of variable x
  - We can calculate this value using the probability density function



Using the PDF of the normal distribution

- `pnorm(x, mean, sd)`
    - *x* is the upper threshold; the function will output the probability of all values less than this
    - *mean* and *sd* give the parameters of the function
    - Returns the area under the normal distribution below x
    - Remember, the normal curve changes based on the values of μ and σ, so it makes sense that this PDF requires these parameters



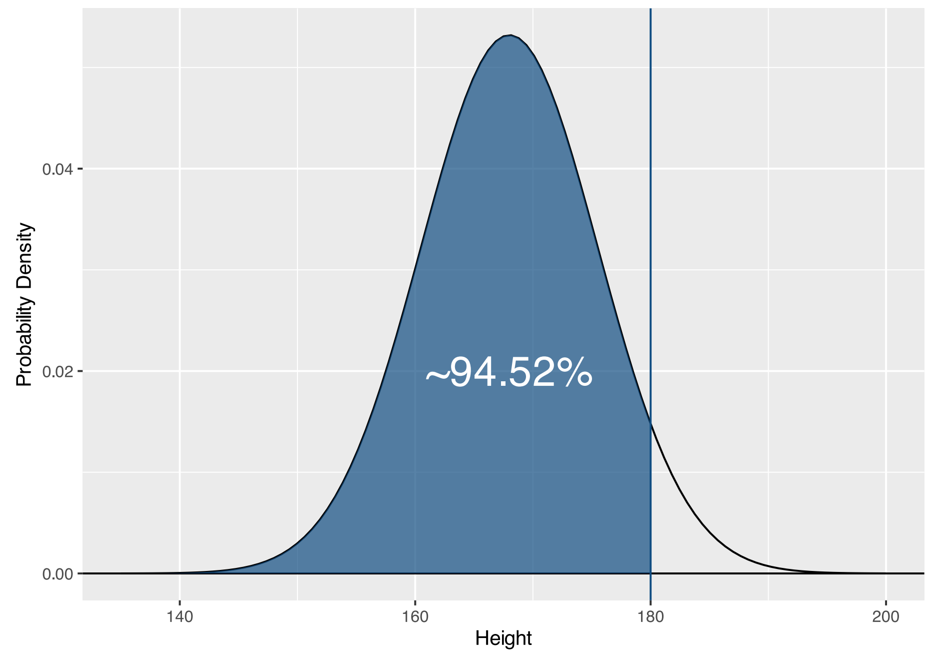
Using the PDF of the normal distribution

- `pnorm(x, mean, sd)`
    - *x* is the upper threshold; the function will output the probability of all values less than this
    - *mean* and *sd* give the parameters of the function
    - Returns the area under the normal distribution below x
    - Remember, the normal curve changes based on the values of μ and σ, so it makes sense that this PDF requires these parameters

```
pnorm(180, mean=168, sd=7.5)
```

```
## [1] 0.9452007
```

> **Test Your Understanding:** How do you interpret this output?

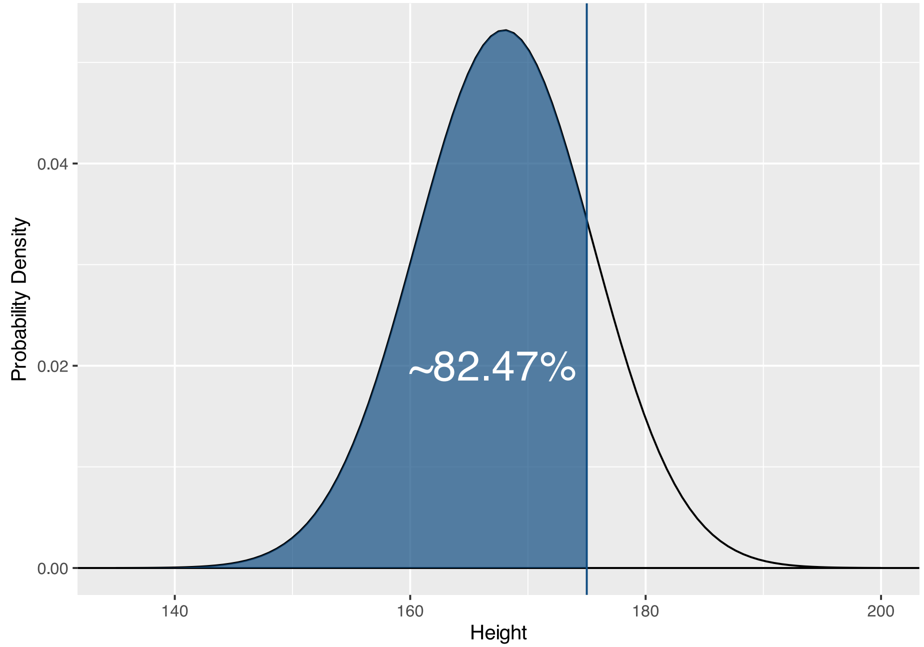


Using the PDF of the normal distribution

- We can also calculate the area under the curve below 175:

```
pnorm(175, mean=168, sd=7.5)
```

```
## [1] 0.8246761
```



> **Test Your Understanding:** Now you know that 94.52% of student heights fall below 180 cm, and 82.47% of student heights fall below 175 cm. How do you calculate the probability of selecting a student whose height falls between 175-180 cm?

Using the PDF of the normal distribution

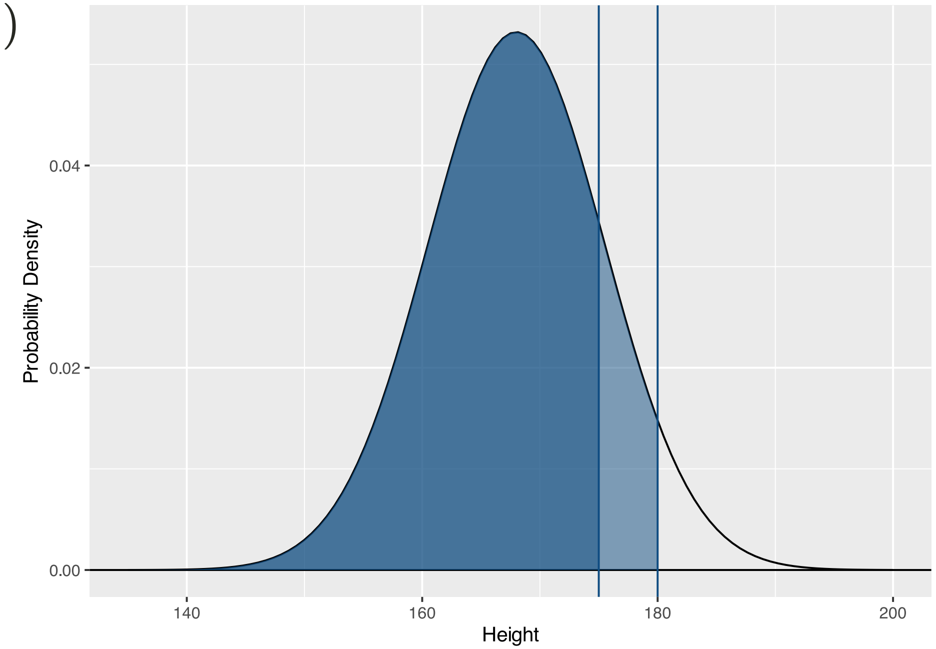
- $P(175 \leq x \leq 180) = P(X < 180) - P(X < 175)$

```
p180 <- pnorm(180, mean=168, sd=7.5)
p175 <-  pnorm(175, mean=168, sd=7.5)

p180-p175
```

```
## [1] 0.1205247
```

- So, the probability of randomly selecting a student with a height between 175 and 180 is 0.12



Using the PDF of the normal distribution

- We can also ask about the probability of a sampled element having a value from one of 2+ ranges
- What is the probability that a person will have a height below 151 or greater than 185? $P(x \leq 151 \text{ } or \text{ } x \geq 185)$

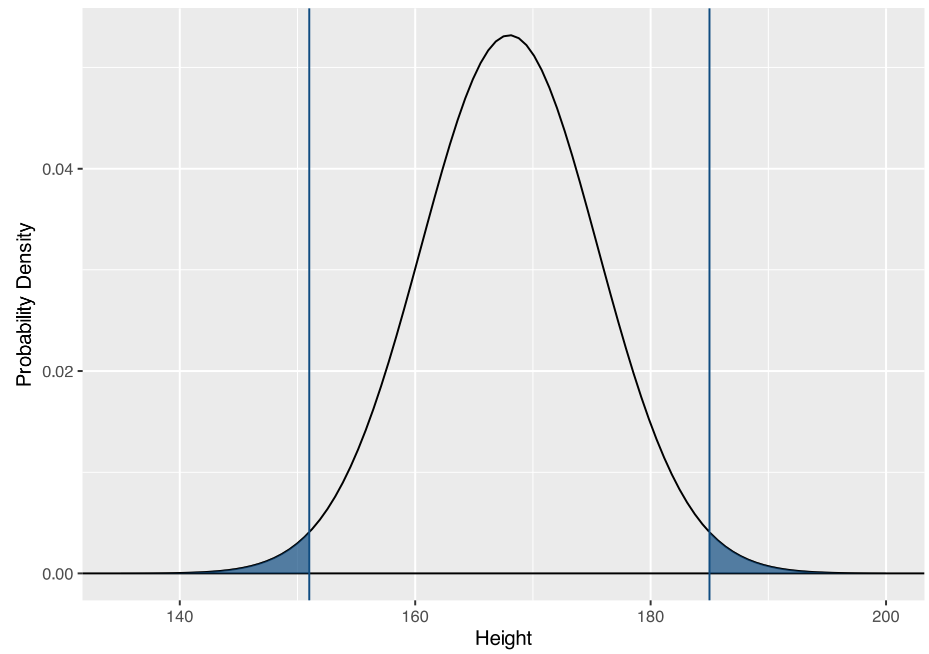
```
pnorm(151, mean=168, sd=7.5)
```

```
## [1] 0.0117053
```

```
1 - pnorm(185, mean=168, sd=7.5)
```

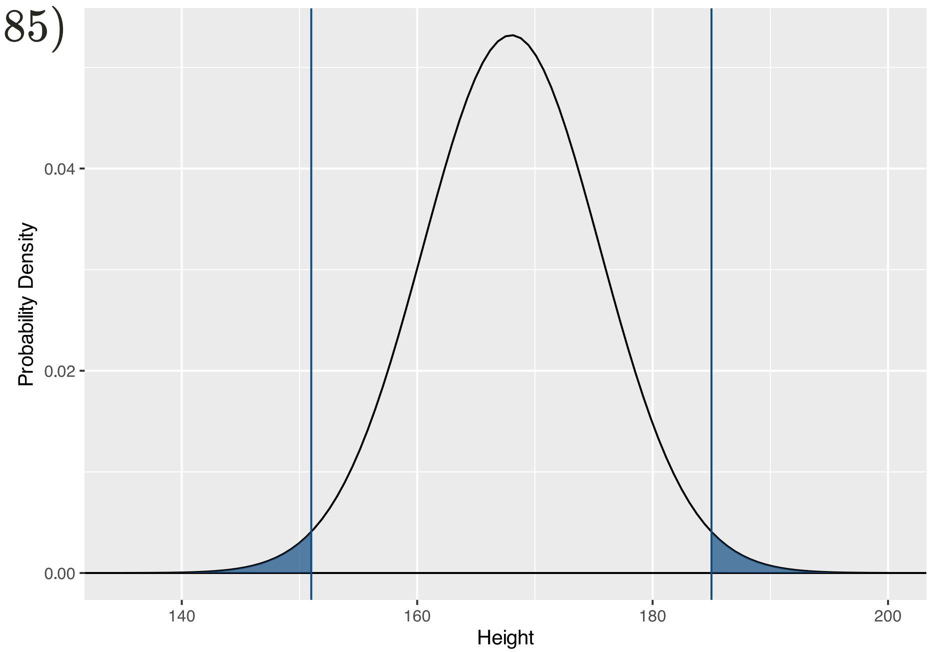
```
## [1] 0.0117053
```

- **Test your understanding:** Why are we subtracting a value from 1 here?



Using the PDF of the normal distribution

- $P(x \leq 151 \cup x \geq 185) = P(x \leq 151) + P(x \geq 185)$
- $0.01 + 0.01 = 0.02$



Using the PDF of the normal distribution

- What if I wanted to know where the 5% of the most extreme values (i.e., smallest and largest) in this distribution fall?
    - The normal distribution is symmetric, which means that there are the same number of extreme values at the bottom and top end
    - This means the most extreme 5% will be the 2.5% at the bottom of the distribution and the 2.5% at the top
    - So our question is: What is the height below which there are only 2.5% of students, and what is the height above which there are only 2.5% of students?

Using the PDF of the normal distribution

- To get these values, you can use `qnorm(p, mean, sd)` - this is the inverse function of `pnorm()`
- For a normally distributed range of heights with a mean of 168 cm and a SD of 7.5 cm:
- The height below which 2.5% of students fall:

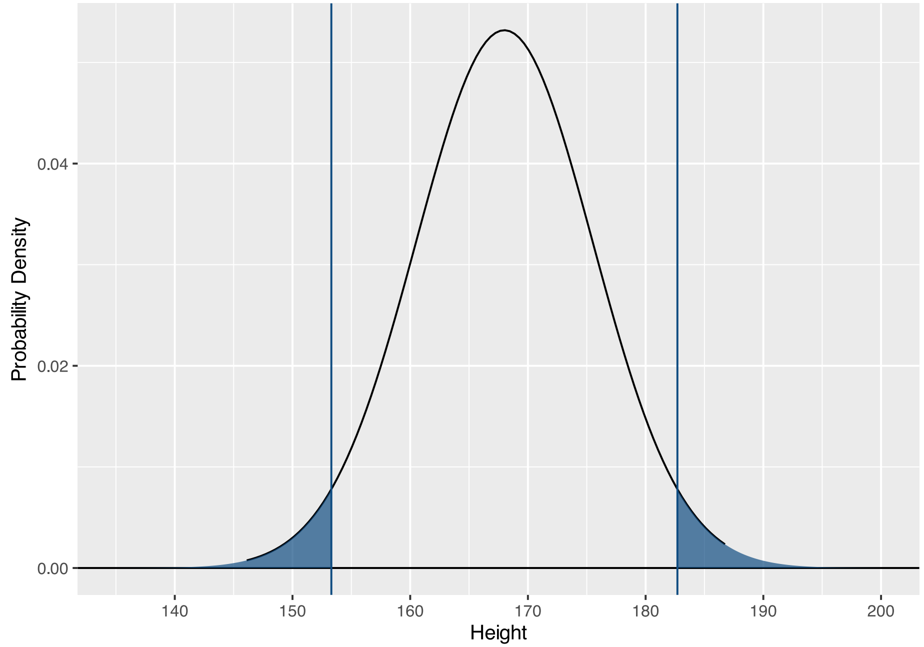
```
qnorm(.025, mean=168, sd=7.5)
```

```
## [1] 153.3003
```

- The height above which 2.5% of students fall:

```
qnorm(.975, mean=168, sd=7.5)
```

```
## [1] 182.6997
```



Take this knowledge forward

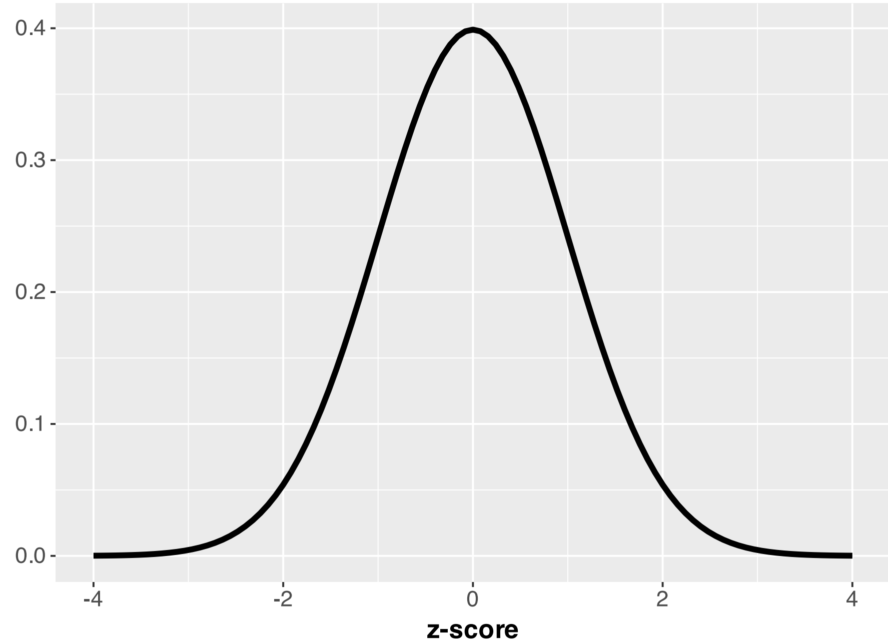
- These examples might seem a bit far-fetched (when will you ever need to calculate extreme heights?), but this will be incredibly relevant when you discuss:
    - 1- and 2-tailed distributions
    - p-values
    - Distributions of test statistics
- You may find it helpful to come back and review these slides when you get to these topics later in the course

Questions?

Remember z-scores

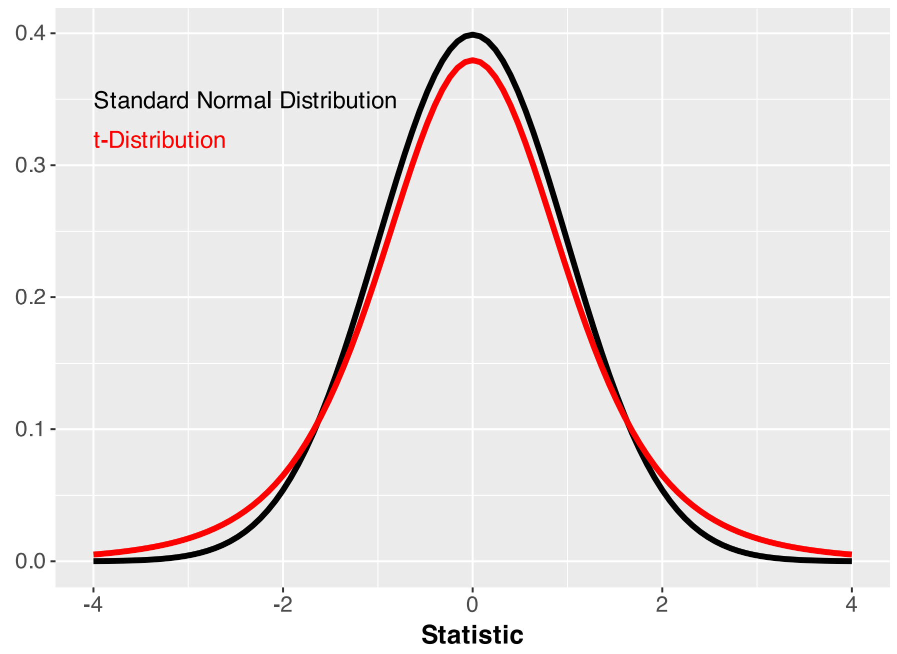
$$z = \frac{x - \mu}{\sigma}$$

- It is quite typical to present a normal distribution in terms of **z-scores**.
- z-scores standardise values of x
  - The numerator: converts x to deviations from the mean
  - The denominator: scales these deviation values based on the spread of the variable (SD)
- The result is the **standard normal distribution**, also known as the z-distribution



Standard normal vs. t distribution

- There are other continuous probability distributions you'll be working with next semester, such as the t-distribution
- The t distribution is a bit like the z-distribution, but the shape differs slightly
    - When calculating t, we replace the population SD (σ) with the sample SD (s)
    - As a result, the tails of the t-distribution are slightly higher to account for extra variability, or uncertainty from using an estimate (s) rather than the actual population value (σ)



Summary of today

- Continuous probability distributions
- The normal distribution
- Using the normal distribution to make estimates about the probability of events
- Using the normal distribution to find values at the extremes of the distribution
- The normal distribution and the t-distribution

- Tomorrow, I'll present a live R session focused on continuous probability distributions

- Next week, we will talk about samples and populations

This week

Tasks

- Attend both lectures
- Attend your lab and work together on the lab tasks
- Complete the weekly quiz
    - Opened Monday at 9am
    - Closes Sunday at 5pm

Support

- **Office hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)
- **Piazza**: help each other on this peer-to-peer discussion forum
- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)