SOC6707 Intermediate Data Analysis

Monica Alexander

Week 2: Probability! Chance! Randomness!

Announcements

- Assignment 1 is up
- tidyverse, some plotting, some probabilities
- Also added a 'template' (helpful? maybe not?)
- Also due is research proposal
- Datasets you need are on Quercus
- Discussion boards are up

What's the point of today

- Population –> Sample –> Population
- We are introducing randomness, but trying to make meaningful inferences despite this
- Need to know basic probability concepts
- This helps us to talk about distributions of the statistical quantities we are interested in
  - e.g. population means, but also regression coefficients

Random variables

From first week: A **random variable** is a variable whose values depend on the outcomes of a random process.

Examples

- Flipping a coin four times and recording the number of heads
- Randomly sampling six people and recording their height
- A toddler randomly selecting a Lego car

Probability essentials

Probability

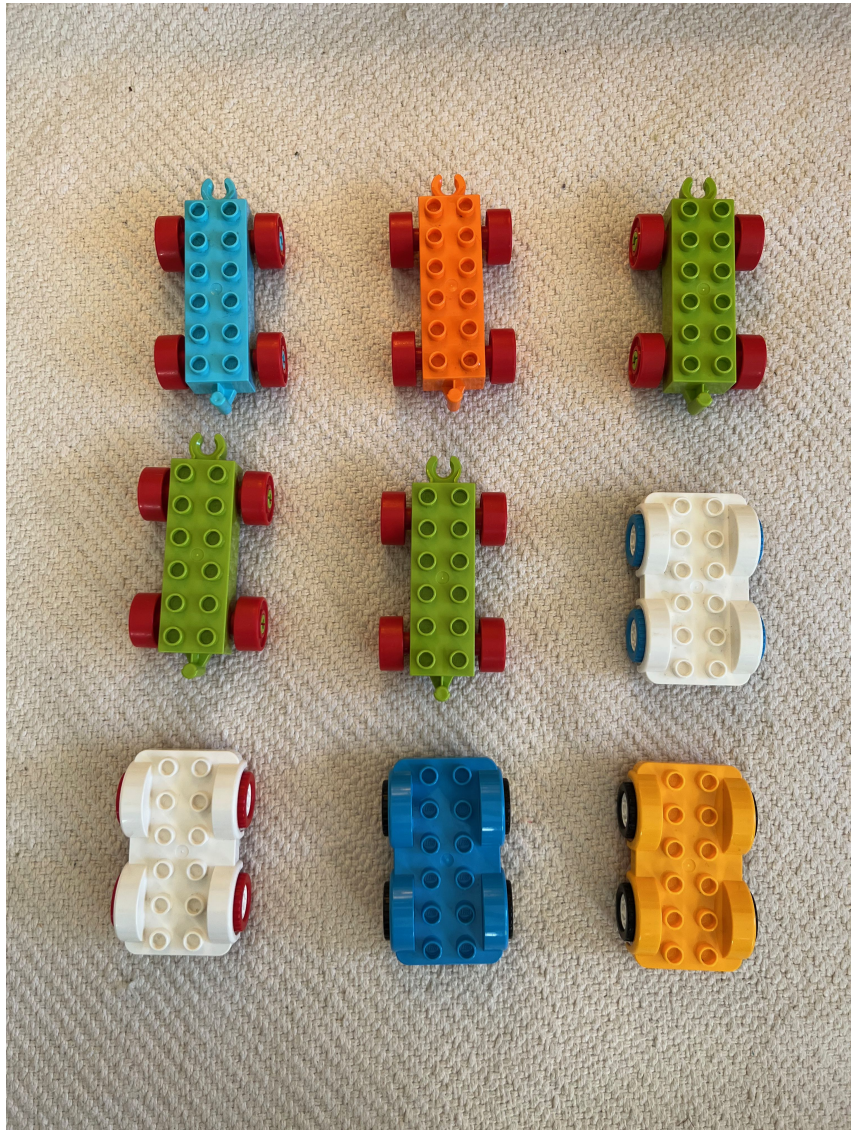
- Based on our sample or other random process (as in the coin flipping or a toddler choosing Lego), we would like to make valid statements about the underlying population or quantity of interest
- Probability is one tool that will help us do that
- Probability is all about talking about the chance of something (an event happening or observing a particular thing)
- There is uncertainty associated with the event or observation, and probability helps us to quantify this

Definitions

- **Events**: things that can happen
  - what's an example of an event when flipping a coin once? Four times?
  - what's an example of an event of sampling six people's heights?
- **Probability function**: a rule that assigns a value $P(A)$ to each event A. We know
  - Probability is positive
  - Probability is at most 1
  - The sum of probablities of all possible events is 1

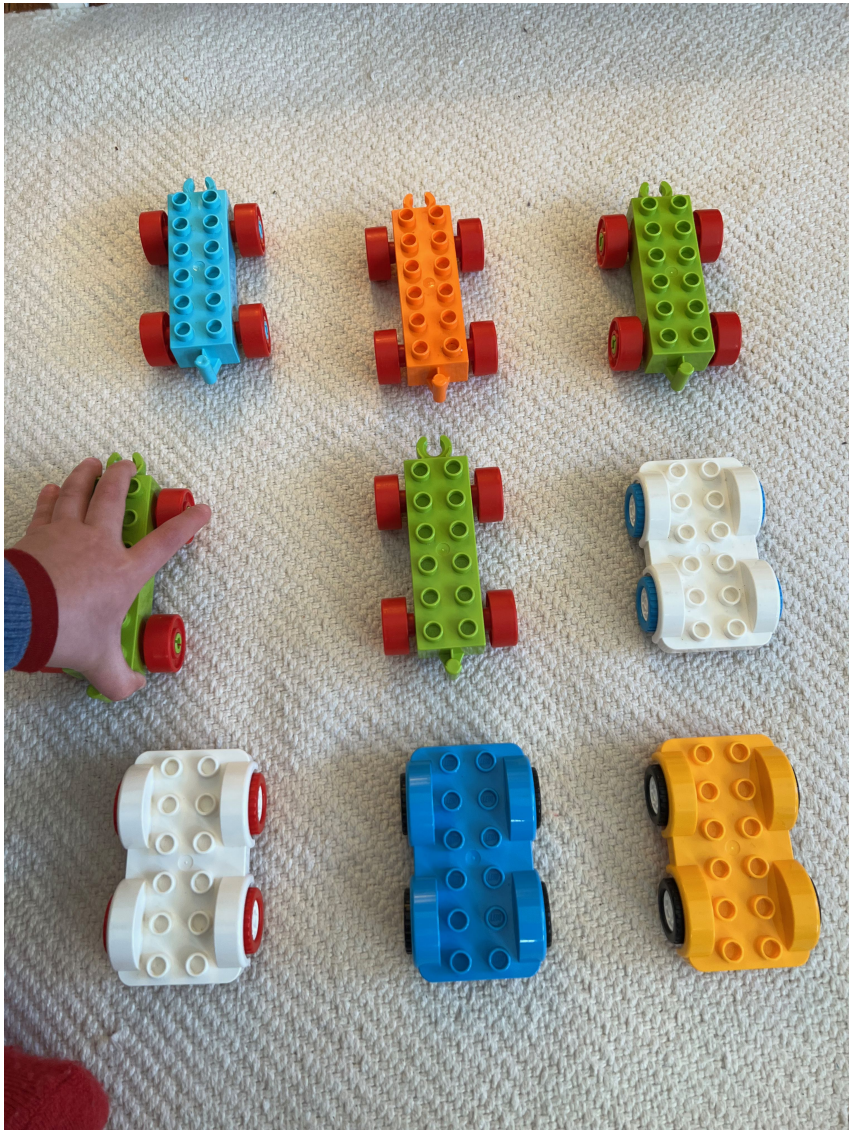
Lego example

We have the following lego trains and cars:



Lego example

My son randomly draws out one vehicle



Lego example

Let's define some events:

- A = "Choose a train"
- B = "Choose a vehicle that is blue"

What is $P(A)$? What is $P(B)$?

Probability is just counting!

Probability is just counting!

Conditional probability

- The probability of something happening given we know something else
- $P(B|A)$ is conditional probability i.e. the probability of B given that A is true
- Lego examples
  - what is $P(B|A)$?
  - what is the probability that the vehicle is a train given it has red wheels?
  - what is the probability that the vehicle is white given it is a car?

Conditional probability

Conditional probability is important for us

- What's the probability that someone work's remotely given they work in finance (vs hospitality?)
- What's the probability that someone graduates college given their parent's did?

Probability distributions

Back to coin flipping example

- The process of tossing a coin four times qualifies as an experiment
- We can observe the outcome of each toss, and the outcome is uncertain.
- Our random variable of interest was the number of heads

First, let's look at possibilities. On the first toss, we could observe an outcome of heads (H) or tails (T). On each of the remaining three tosses, we could observe an H or a T. Thus, the possibilities for four tosses can be enumerated as follows:

- HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, and TTTT.

We can see that there are 16 different possible outcomes when listed as simple events.

Flipping a coin

We can enumerate these possible outcomes in a table with the associated probability and observed number of heads

event	probability	number of heads
HHHH	0.0625	4
HHHT	0.0625	3
HHTH	0.0625	3
HHTT	0.0625	2
HTHH	0.0625	3
HTHT	0.0625	2
HTTH	0.0625	2
HTTT	0.0625	1
THHH	0.0625	3
THHT	0.0625	2
THTH	0.0625	2
THTT	0.0625	1
TTHH	0.0625	2
TTHT	0.0625	1
TTTH	0.0625	1
TTTT	0.0625	0

Flipping a coin

Using this we can work out different probabilities. e.g. probability of 3 heads

$P(X = 3) = P(HHHT \text{ or } HHTH \text{ or } HTHH \text{ or } THHH) = 4/16$

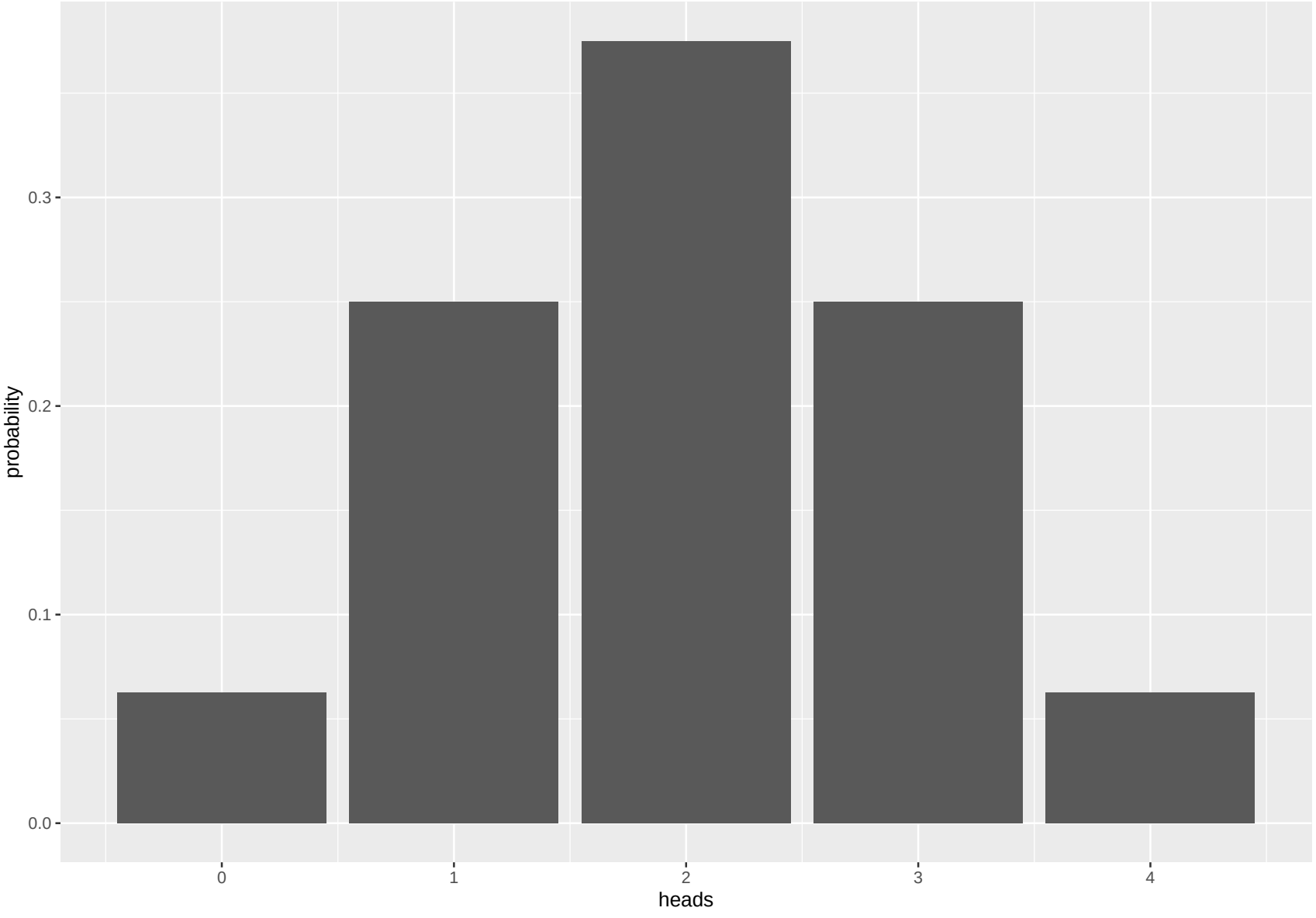
Probability distribution for the number of heads

Given our RV of interest is the number of heads and that all events are mutually exclusive, we can summarize the table as

Number of heads (X)	P(X)
4	1/16
3	4/16
2	6/16
1	4/16
0	1/16

We have a **probability distribution** for the number of heads. That is, a rule or function that associates the probability of observing that particular value with each value of a random variable. The probability distribution for a **discrete** RV (like # heads) is called a **probability mass function**

Probabilities as areas



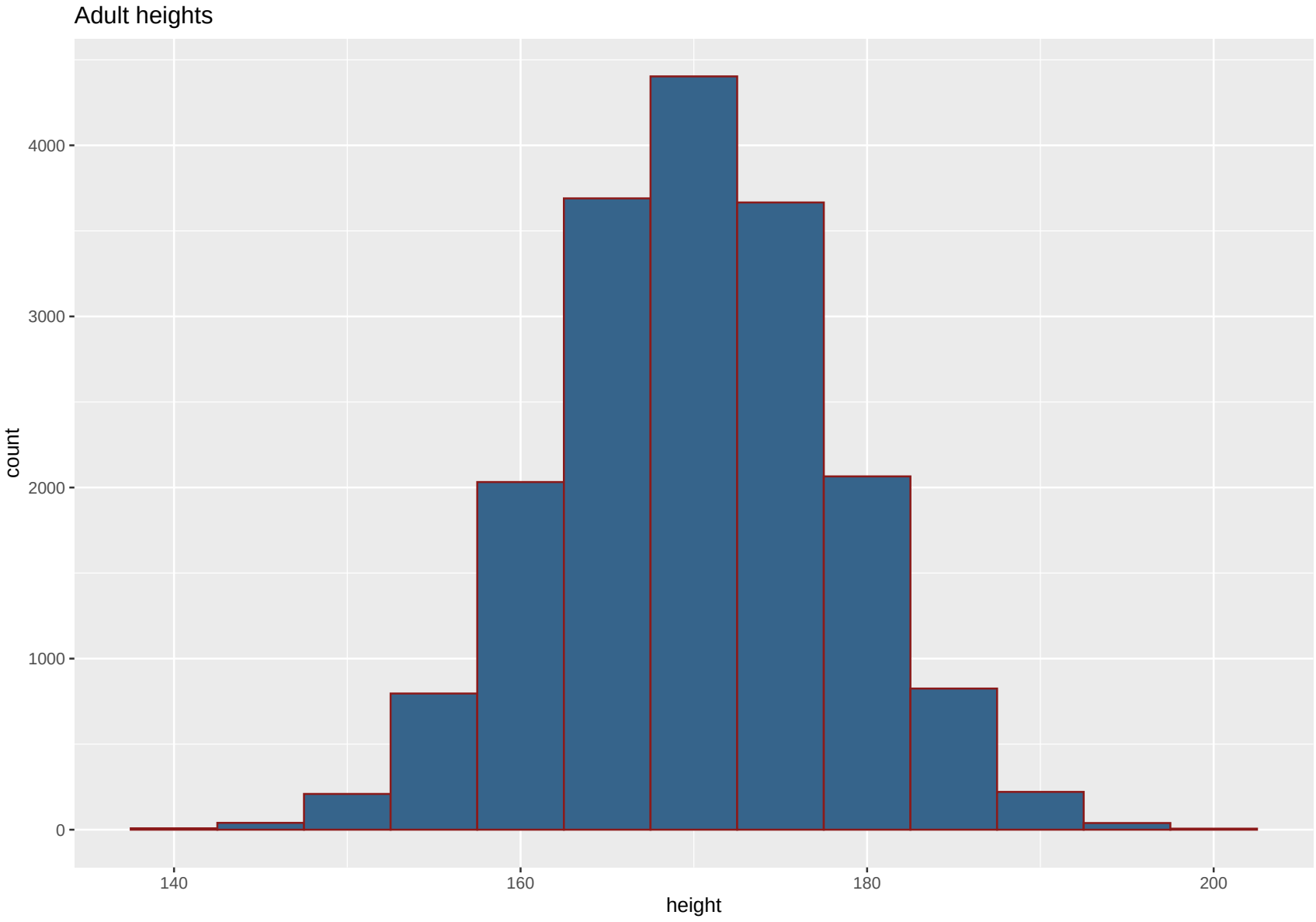
Probabilities as areas

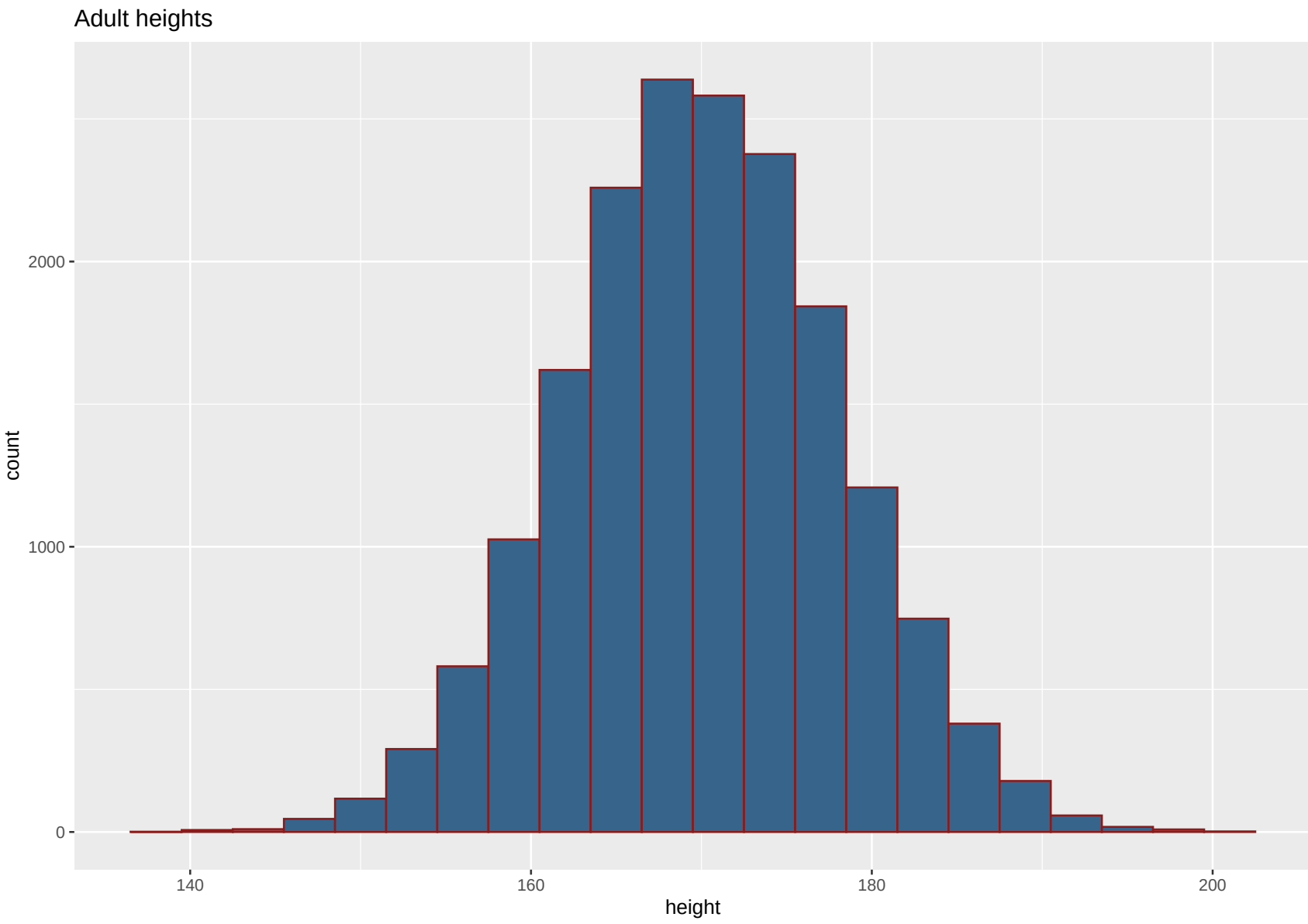
- To calculate probabilities, can sum up the area of the rectangles
- E.g. $P(X \geq 3)$ would be the sum of the right two rectangles
- What is $P(1 \leq X \leq 3)$?
- What is $P(1 \leq X < 3)$?

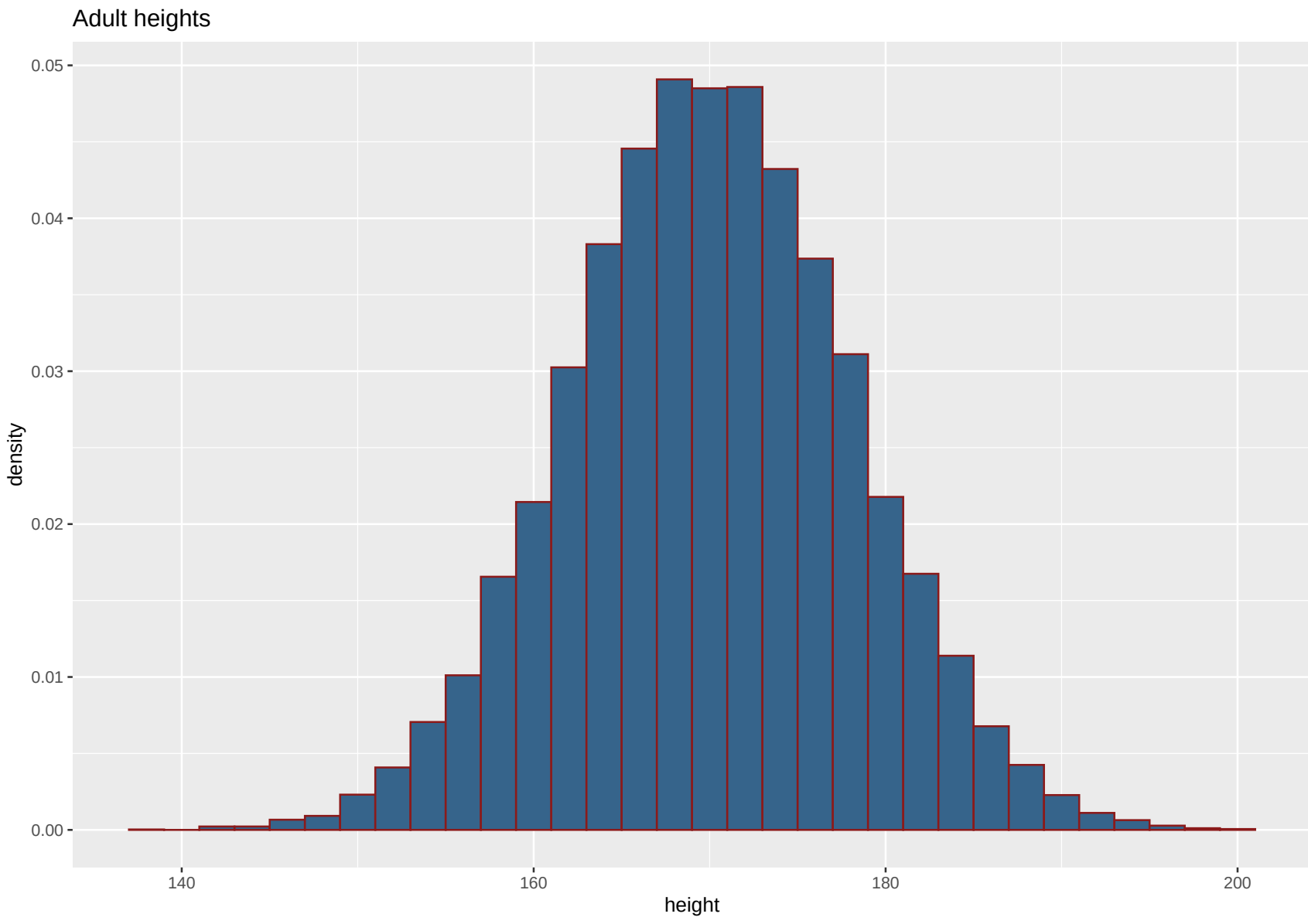
Continuous random variables and probability distributions

- So far we have talked about a discrete RV
- But a lot of our RVs of interest are continuous (e.g. height)
- Can think about in the same way (defining probability distributions, expected values, etc)
- Instead of having a table of values making up the probability distribution (or pmf), we have a mathematically defined function
- A probability distribution for a continuous RV is called a **probability density function**

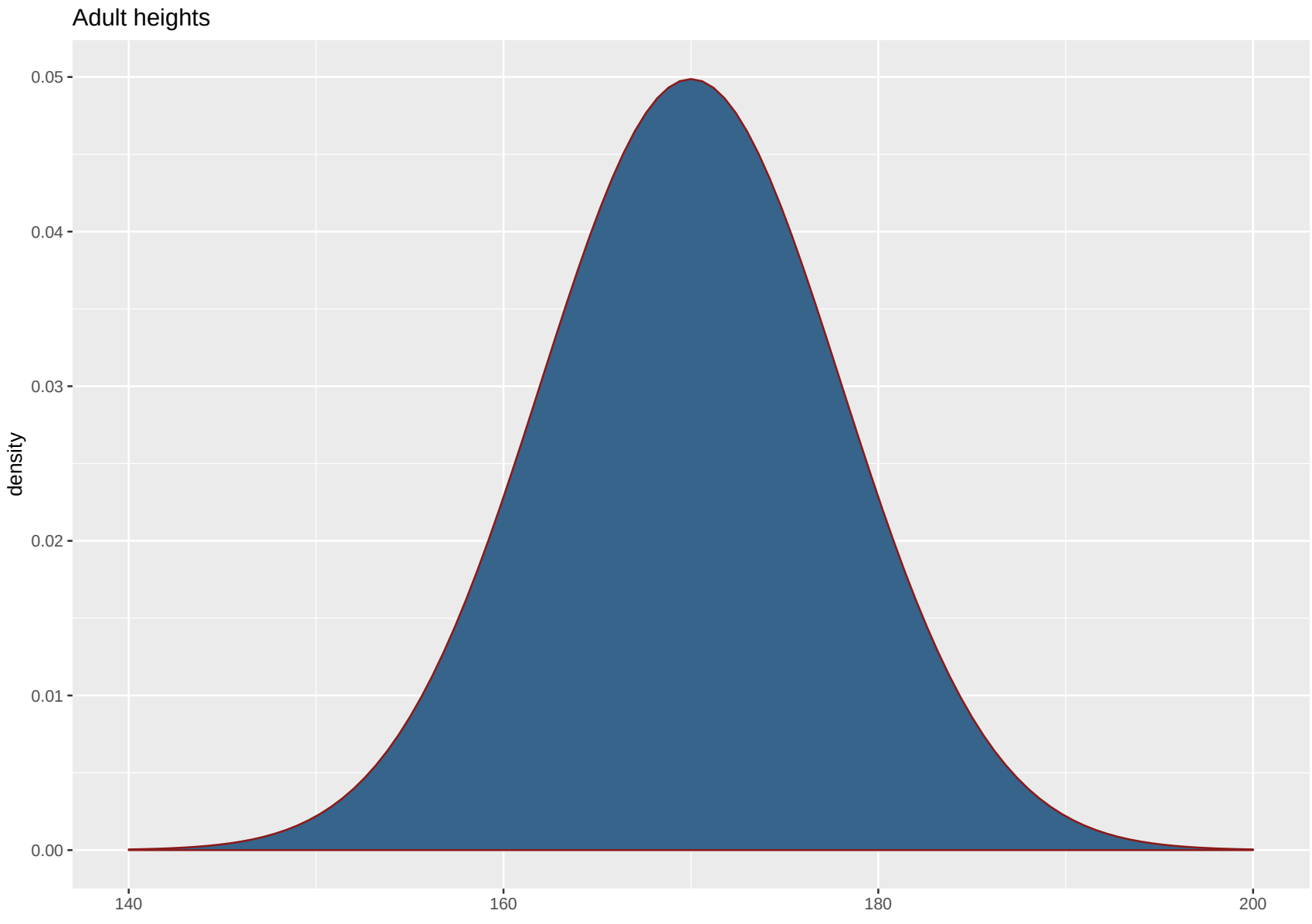
A continuous probability distribution is just a histogram with infinitely small bins



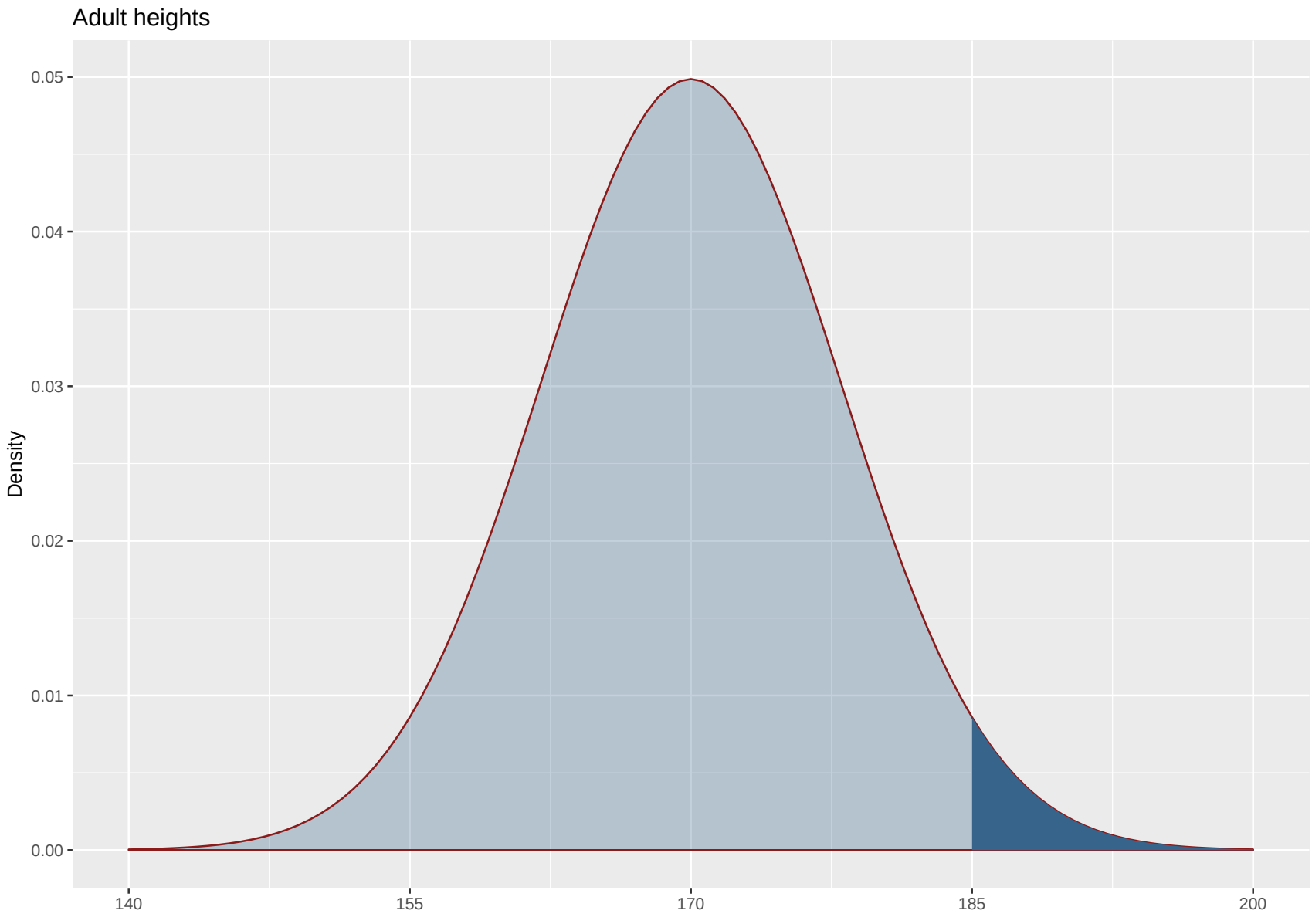

Adult heights
count
2000
1000
0
140
160
180
200
height


Adult heights
density
0.05
0.04
0.03
0.02
0.01
0.00
140
160
180
200
height

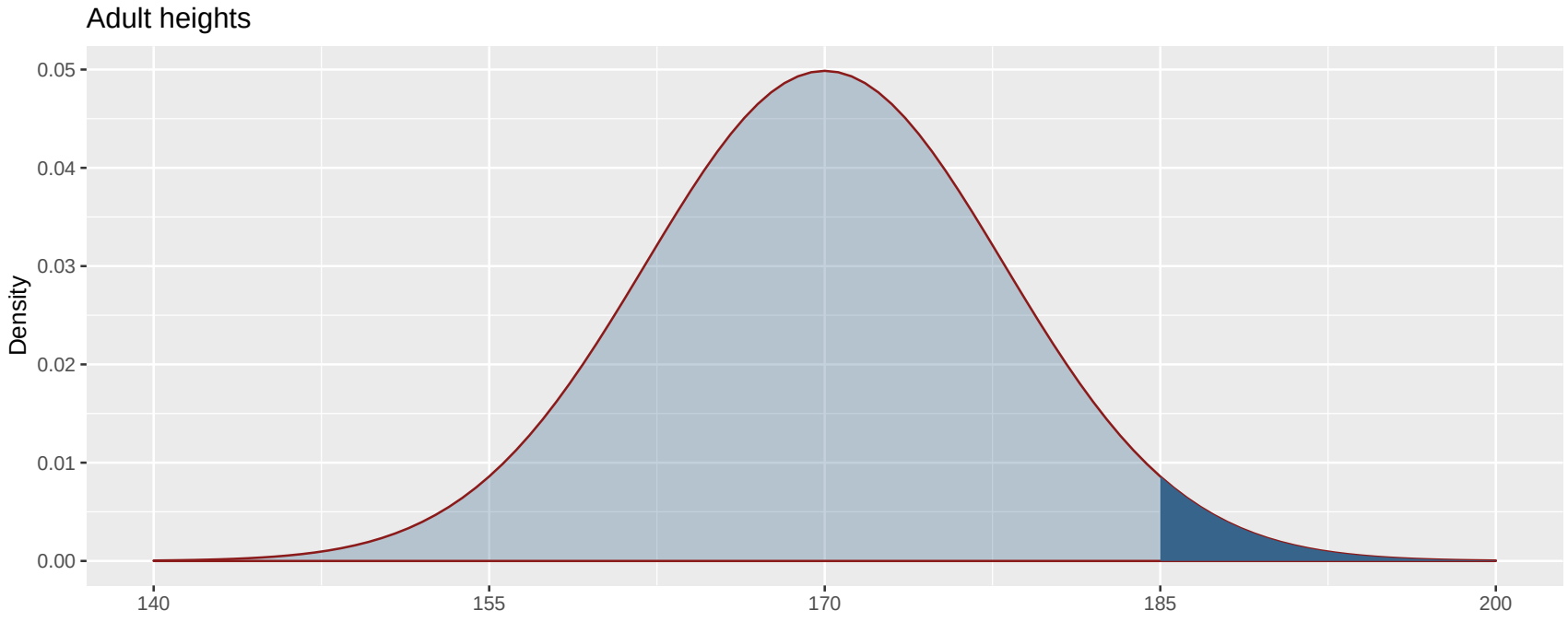
Probability density function



Probabilities as areas



Probabilities as areas

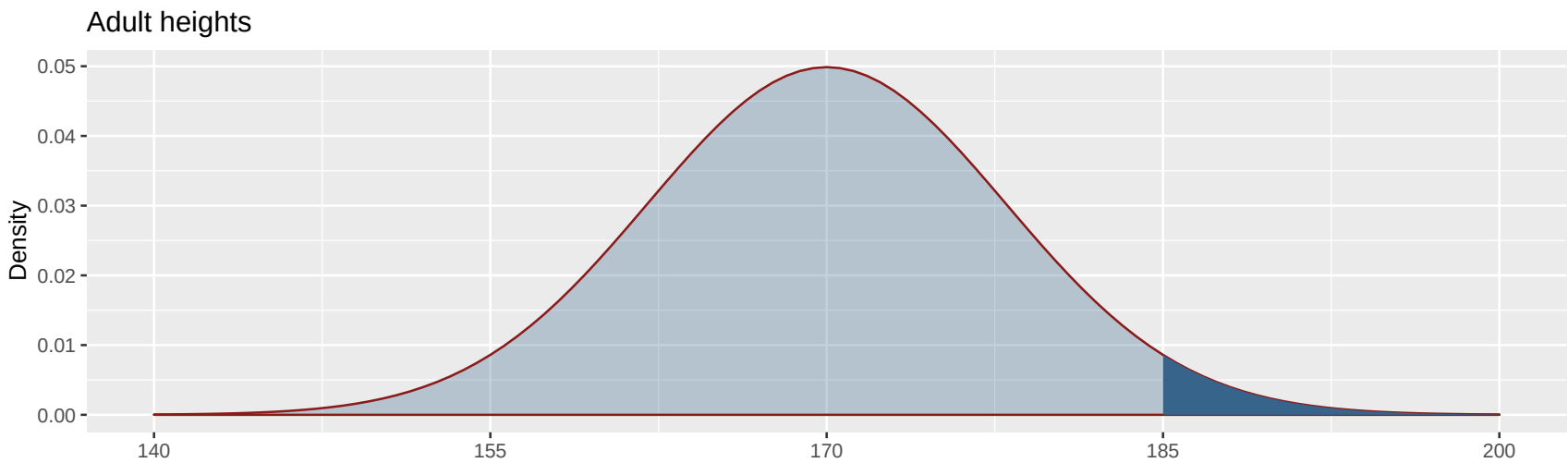


- The probability that height is greater than 185cm i.e. $P(X > 185)$
- Like summing up very tiny histogram bins above a certain point

Probability as areas

Important notes

- The sum of the whole area under the curve is equal to 1 (because we know all probabilities have to sum to one)
- A value is either greater than or less than/equal to a number
- So can express probabilities as the complement
  e.g. $P(X > 185) = 1 - P(X \leq 185)$



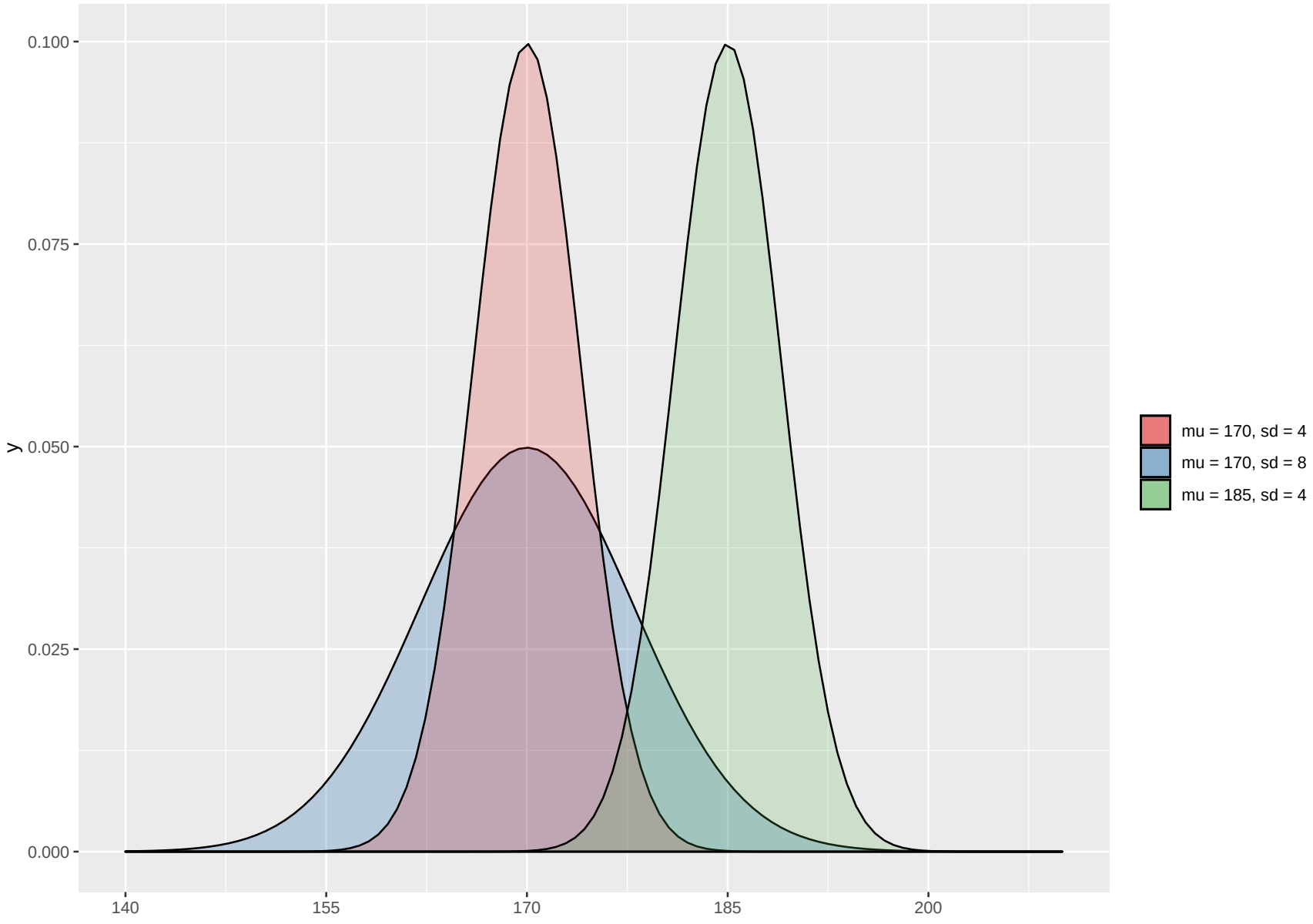
The normal distribution

The normal distribution

- One of the most important continuous probability distributions
- Is described by the formula

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

- The shape is determined by two **parameters**, μ and σ
- If we were to plot $f(x)$ as a function of x, we would obtain a normal distribution that would be centered at whatever value of μ we specified, and it would have a standard deviation equal to σ.

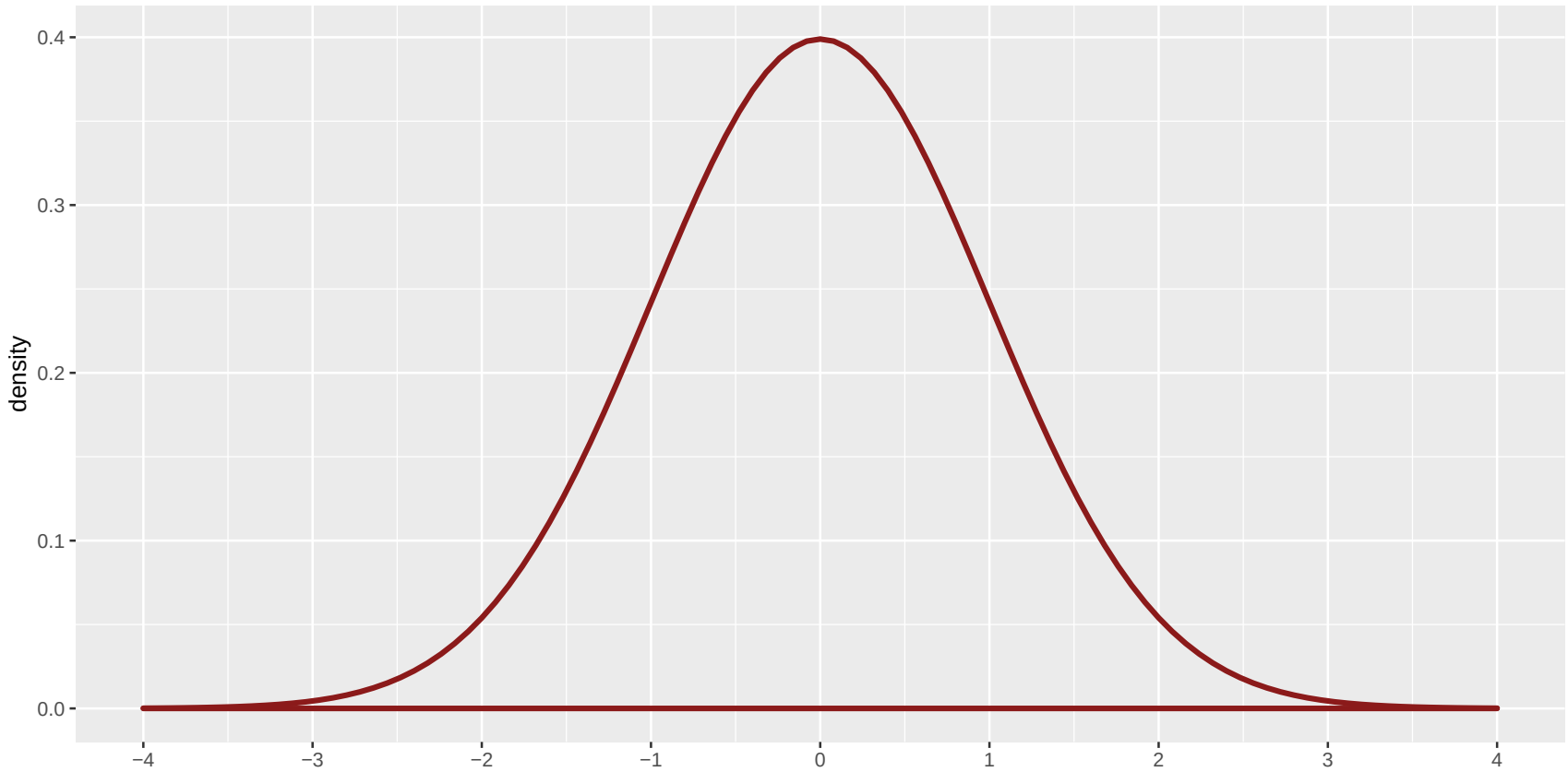

0.100
0.075
0.050
0.025
0.000
y
140
155
170
185
200
mu = 170, sd = 4
mu = 170, sd = 8
mu = 185, sd = 4

The normal distribution

- Many variables naturally resemble the normal distribution (or can be transformed to be so)
- Height, weight, intelligence...
- Strong relationships with other distributions
- Many sample statistics are normally distributed (more later)

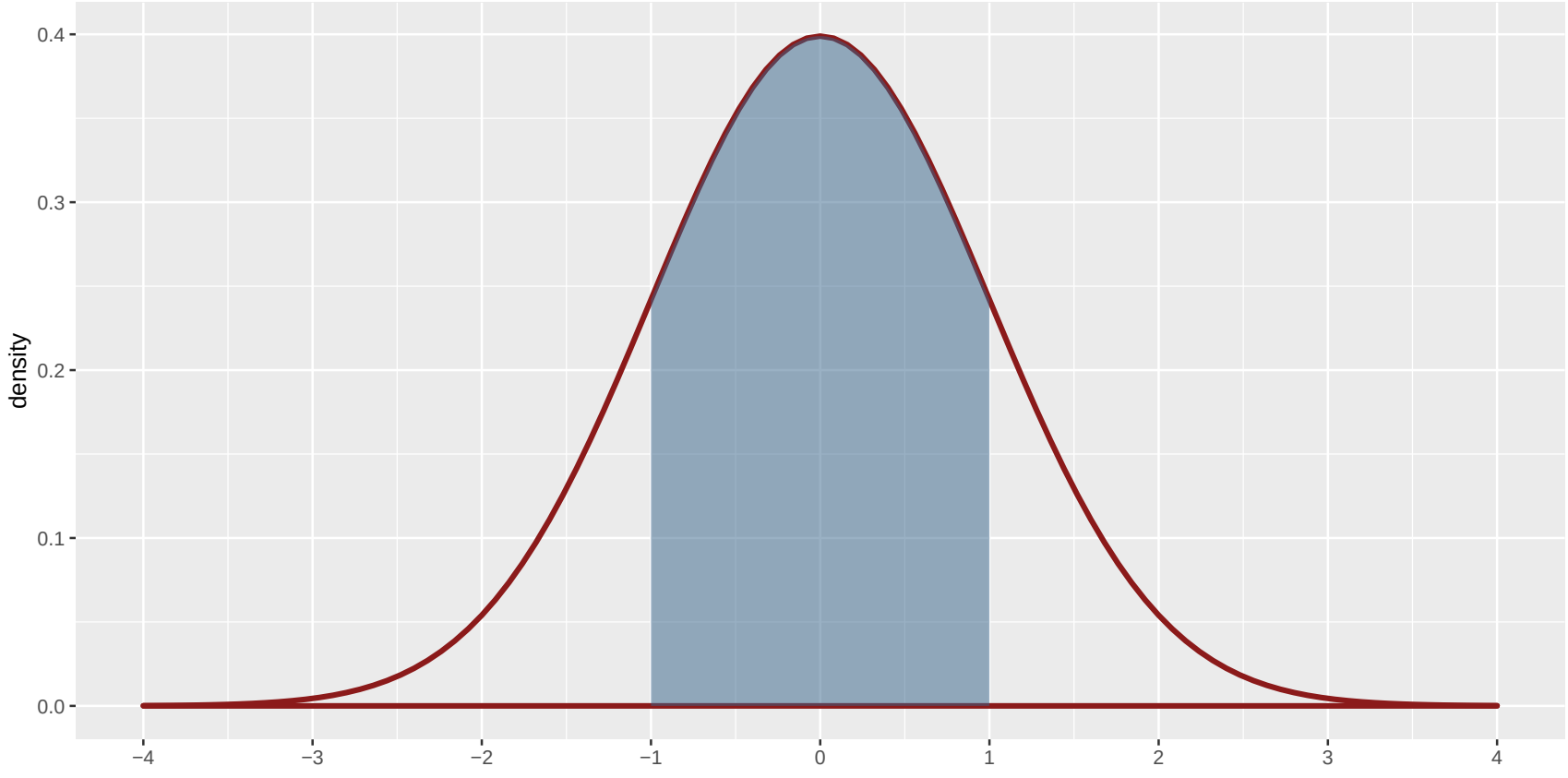
The standard normal distribution

A special case of the normal distribution with $\mu = 0$ and $\sigma = 1$.



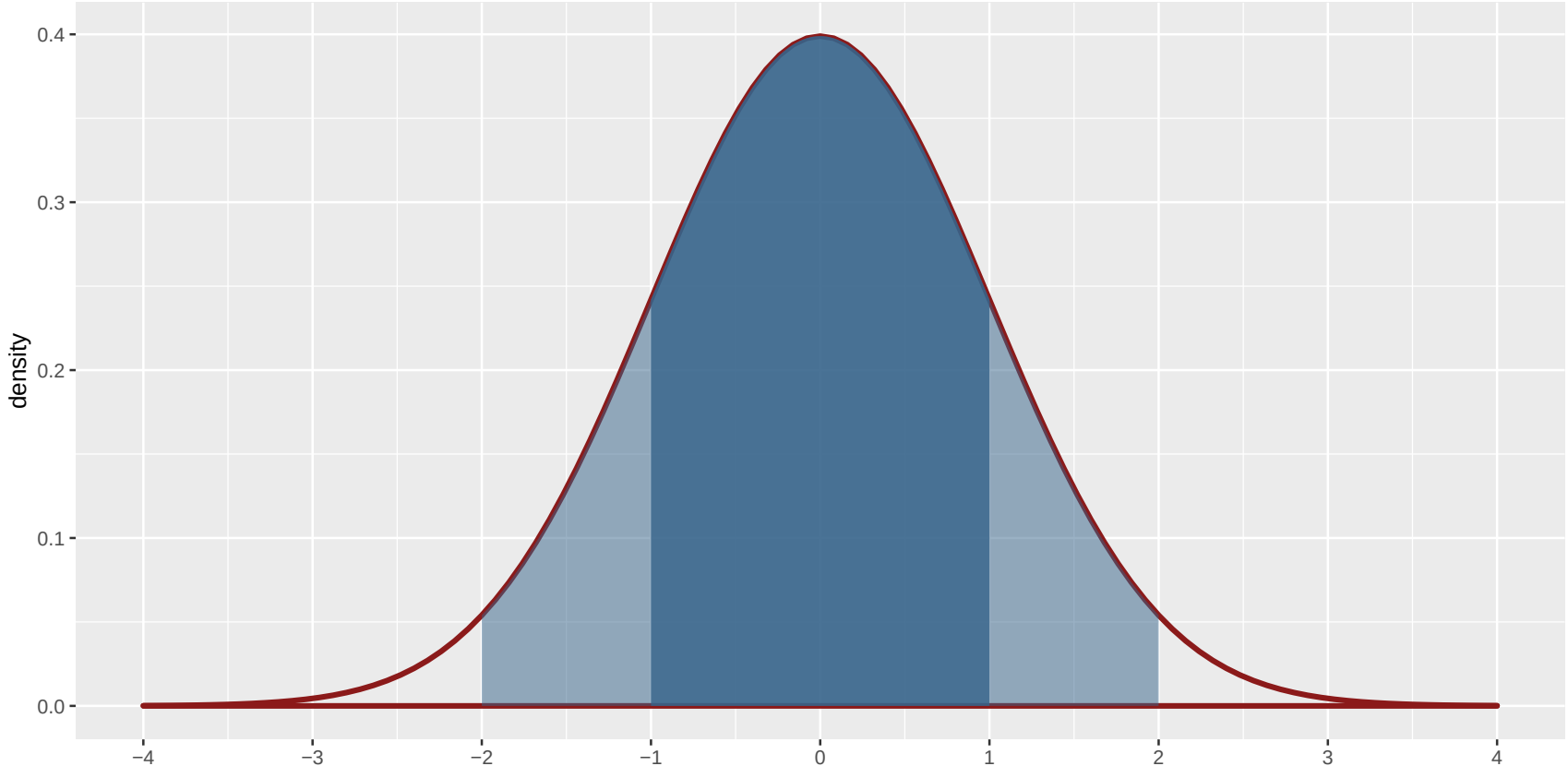
The standard normal distribution

- ~68% of area within 1 standard deviation



The standard normal distribution

- ~95% of the area within 2 standard deviations



Any normal distribution can be transformed into the standard normal

Say X is normally distributed with mean μ and variance σ^2. We can write this as

$$X \sim N(\mu, \sigma^2)$$

We can transform X using the z-**transformation**

$$\frac{X - \mu}{\sigma}$$

Call this transformed version Z i.e. $Z = \frac{X-\mu}{\sigma}$. Then

$$Z \sim N(0, 1)$$

we can refer to the transformed version as **Z-scores**.

Z-scores

- Z-scores tell you the number of standard deviations by which the value of a raw score is above or below the mean value.
- In the heights example, the mean $\mu = 170$ and standard deviation $\sigma = 8$.

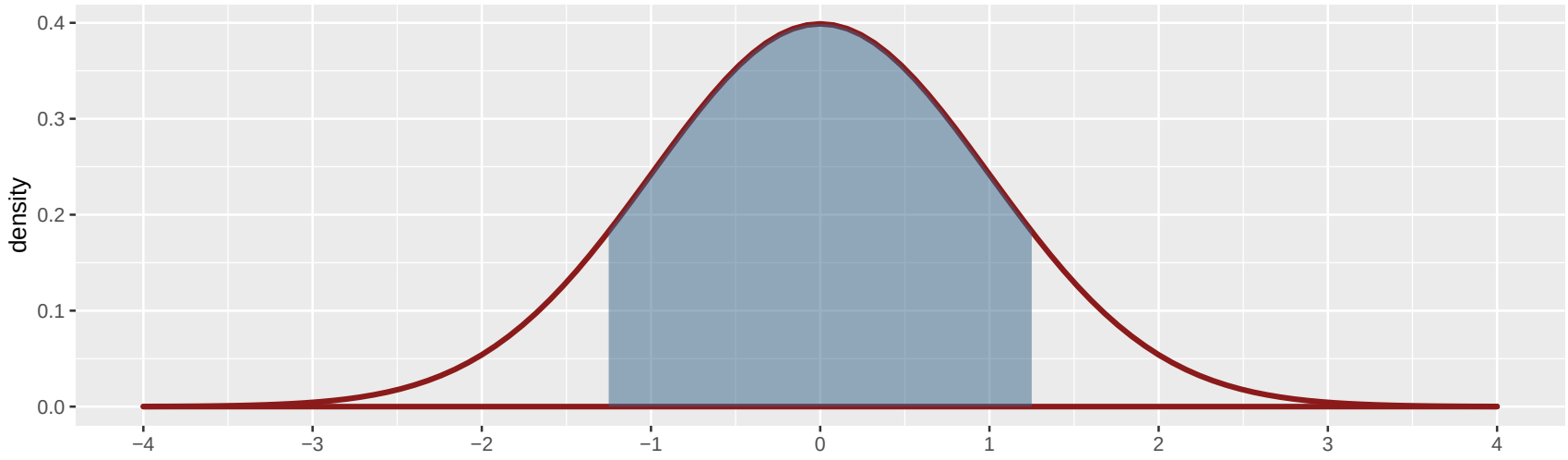
Rohan is 180cm. What is his Z-score?

$$Z = \frac{180 - 170}{8} = 1.25$$

So Rohan is 1.25 standard deviations above the mean height.

- Monica is 168cm, so her Z-score is -0.25. So she is 0.25 standard deviations below the mean height.

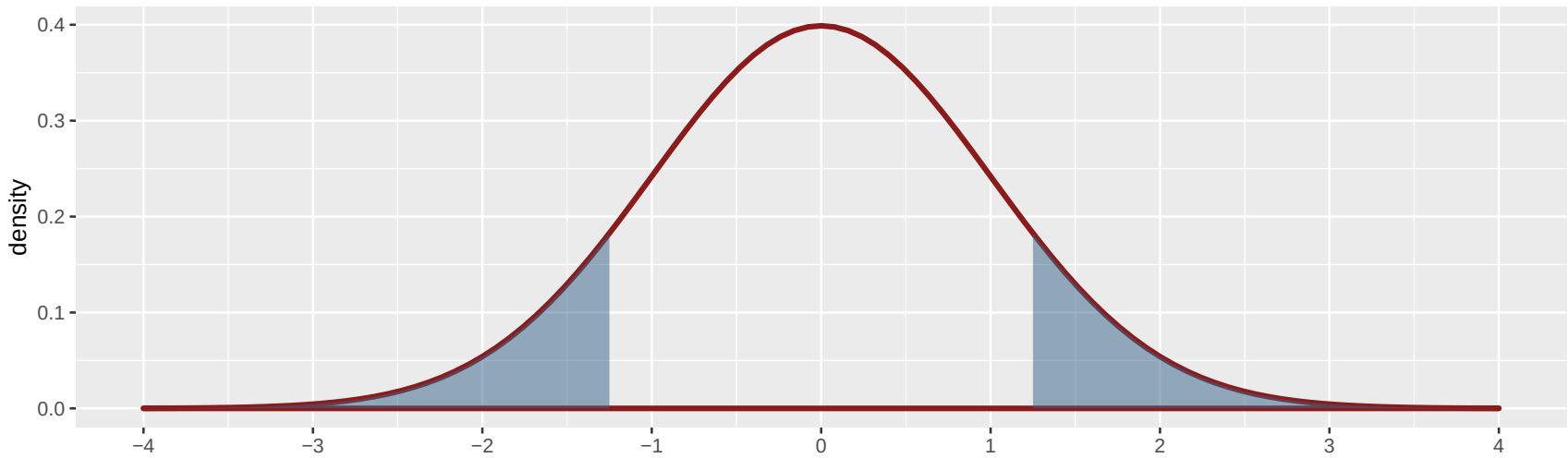
Z-scores



- Recall that ~95% of the area was within 2 standard deviations.
- We can flip this and ask what proportion of the area of a standard normal is within 1.25 standard deviations?
- The answer is ~79%

Z-scores

- ~79% of the area is within 1.25 standard deviations
- Alternative interpretation: If we randomly draw a value from a standard normal distribution, the value will be between [-1.25,1.25] 79% of the time.
- Conversely, 21% of values will fall outside that range



Finding areas under the curve using R

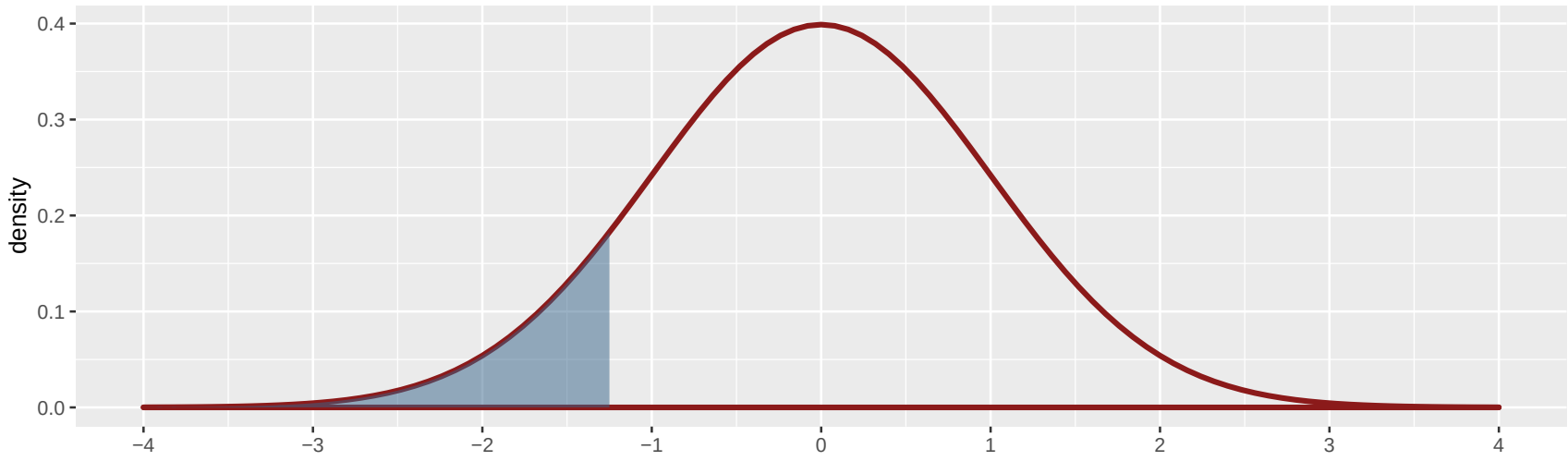
You can find the area under a normal curve in R using the `pnorm` function, which returns the cumulative probability from $-\infty$ to the value supplied to the `pnorm` function.

For example, find the area below z = -1.25:

```
pnorm(-1.25)
```

```
## [1] 0.1056498
```

This corresponds to



Finding areas under the curve using R

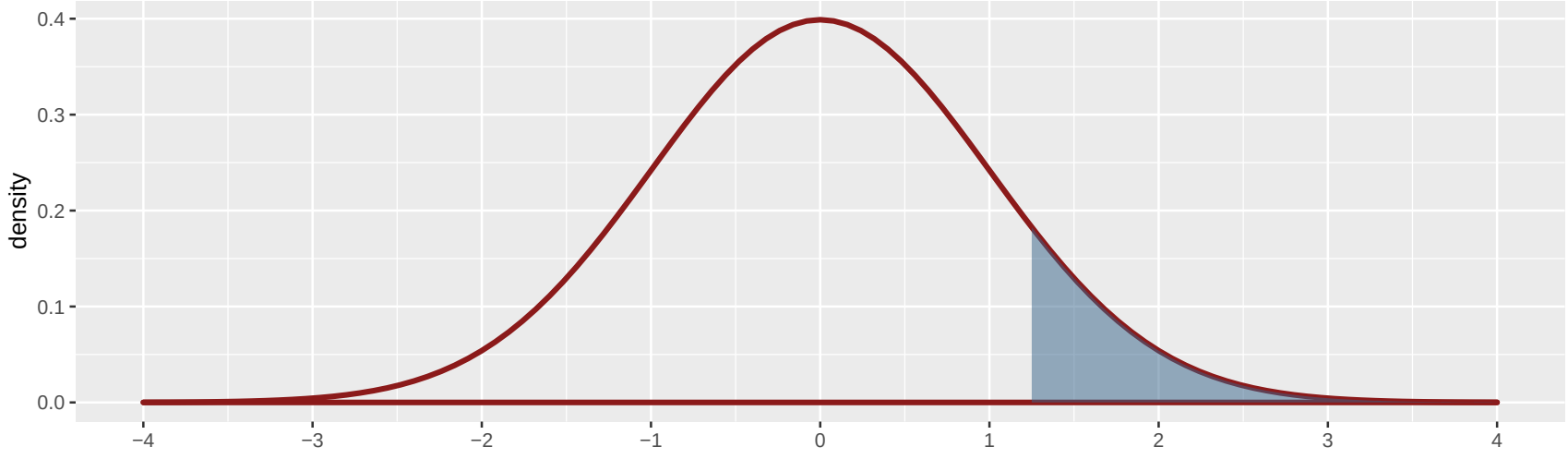
To find the area above 1.25:

```
1-pnorm(1.25)
```

```
## [1] 0.1056498
```

```
# or
# pnorm(1.25, lower.tail = FALSE)
```

This corresponds to



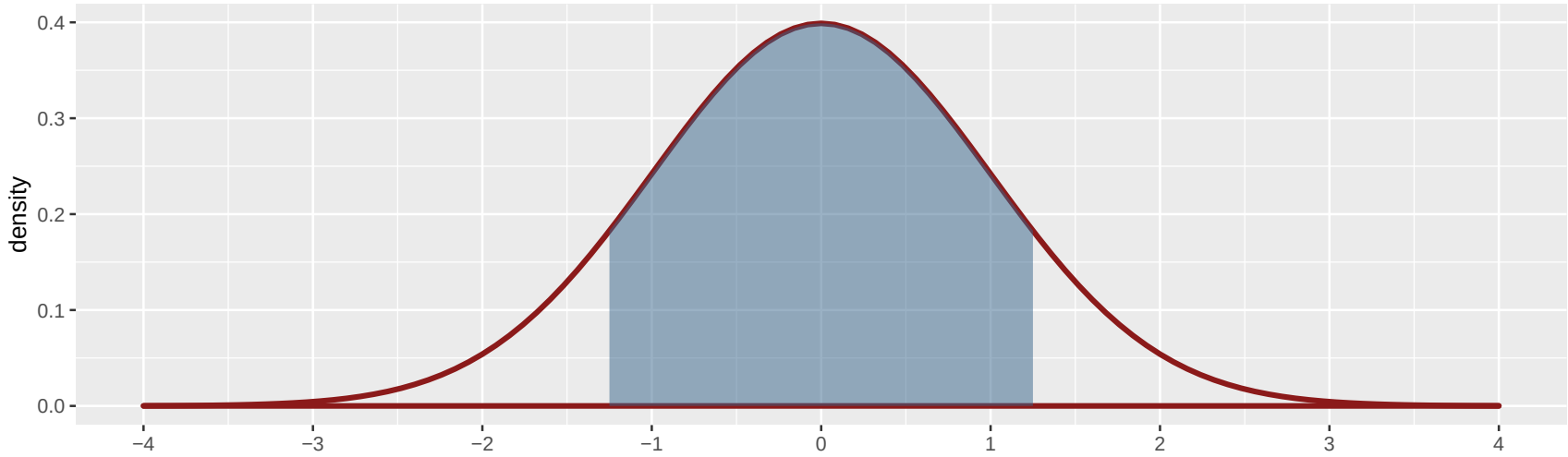
Finding areas under the curve using R

To find the area between -1.25 and 1.25

```
pnorm(1.25) - pnorm(-1.25)
```

```
## [1] 0.7887005
```

This corresponds to



Sampling distributions and the central limit theorem

Why do we care about the Normal distribution so much

- It's just one distribution
- It tends to over-simplify the real distribution of outcomes/variables

BUT it turns out that the distribution of summary statistics that we're interested in (e.g. means) tend towards being Normal

- this is important for regression and inference

Sampling distributions

A **sampling distribution** is a probability distribution for a statistic based on repeated samples.

Say we are interested in taking a random sample of people's heights, X and calculating the mean height for that sample. So our statistic of interest is the mean height, $\bar{X}$.

Randomly sampling heights

We first take a random sample of 12 people and get the following heights

```
##  [1] 158.1 174.9 171.9 154.1 157.4 174.4 168.7 166.2 165.4 167.5 177.9 154.5
```

The observed mean height of this sample is 165.9.

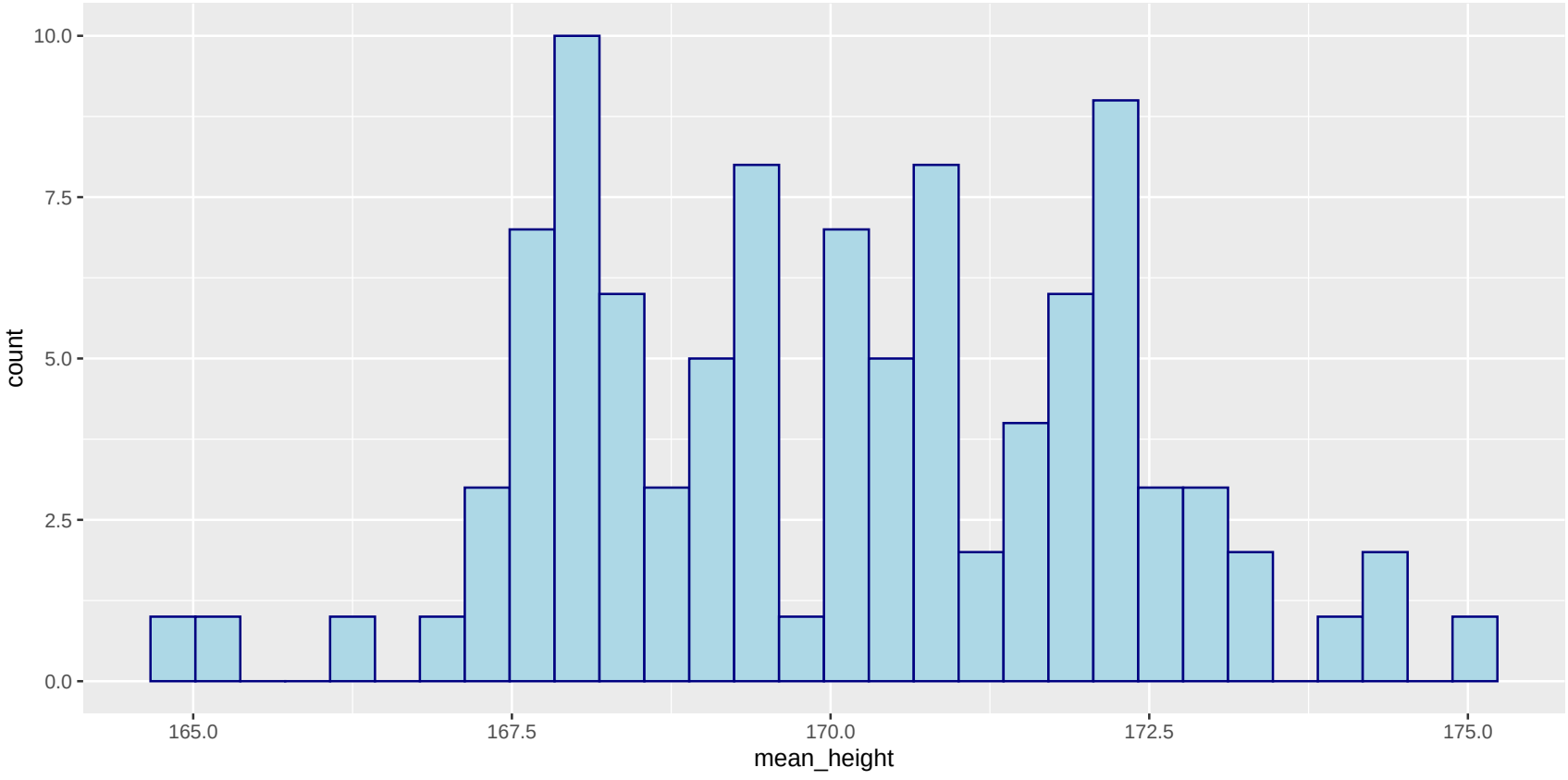
We take a random sample of another 12 people and get the following heights:

```
##  [1] 170.5 158.7 166.6 171.4 169.7 165.2 168.7 166.2 174.7 171.7 168.5 169.2
```

The observed mean height of this sample is 168.4.

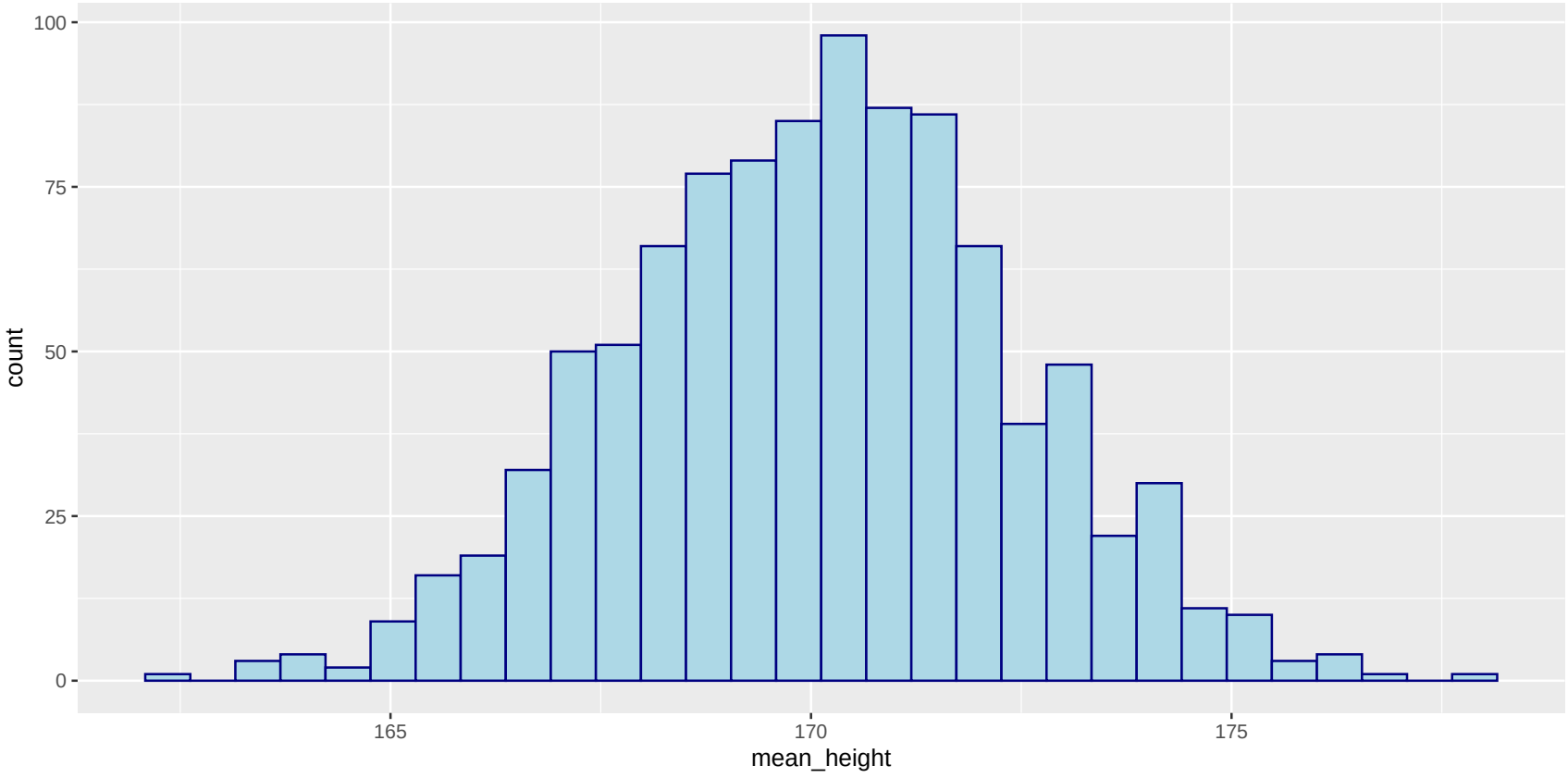
Randomly sampling heights

Say that I keep doing this process again and again and again, and end up with 100 observations of mean height. I can plot a histogram of these means:



Randomly sampling heights

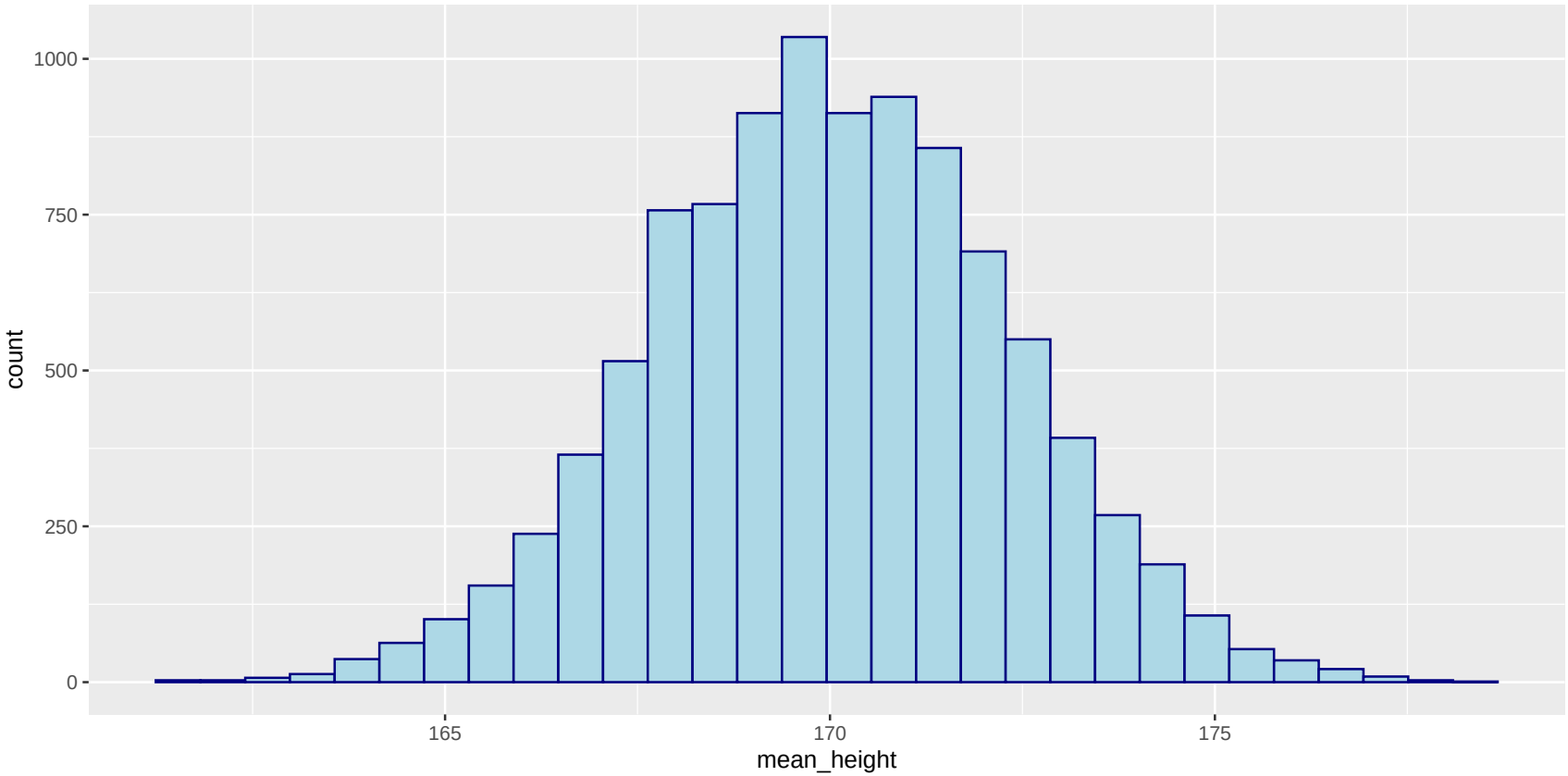
Okay, what if I take 1000 samples. I plot a histogram of the means again



What do you notice?

Randomly sampling heights

What about 10,000 samples?



The central limit theorem

The distribution of the sum (or mean) of a set of independent random variables will **tend towards** a normal distribution.

- "tend towards" means as the number of observations of the sum or mean gets larger, the distribution will become more normal
- The central limit theorem holds even if the original variables themselves are not normally distributed.

The central limit theorem

For a random variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$, the central limit theorem results in the following distribution for the mean $\bar{X}$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

What does this mean?

- The mean $\bar{X}$ will be centered at the same value as X
- The variance of $\bar{X}$ depends on the variance of the original random variable X and also the number of samples of the mean we have, n.

The quantity $\frac{\sigma}{\sqrt{n}}$ is also called the **standard error of the mean**.

Sampling heights

Before, we were repeatedly taking a random sample of 12 heights.

When we did this 100 times, we had 100 observations of $\bar{X}$. The mean was 170 and the standard error of the mean was 0.078.

When we did this 10000 times, we had 100 observations of $\bar{X}$. The mean was 170 and the standard error of the mean was 8×10^{-4}.

Summary

- Probability concepts
- Probability distributions
- Probabilities as areas
- The normal distribution
- The standard normal distribution and Z scores
- Sampling distributions
- The central limit theorem
- Standard error of the mean

Lab

- Finish off tidyverse bit
- Calculating random variables and probabilities
- Intro to ggplot