SOC6302 Statistics for Sociologists

Monica Alexander

Week 4: Probability! Chance! Randomness!

Announcements

- A1 due ‘today’

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

Coin toss example

Imagine tossing a coin 4 times. Say we are interested in the number of heads that turns up. The observed outcomes are:

```
## [1] "T" "H" "T" "H"
```

So the number of heads is 2. But we can toss it another 4 times. The second set of observed outcomes are

```
## [1] "T" "T" "H" "T"
```

So the number of heads is 1.

The number of heads is a **random variable** that depends on the random process of flipping a coin.

Heights example

Say we are interested in heights of people in Canada. We take a random sample of 6 people. Their heights are (in cm)

```
## [1] 177.16 169.96 181.95 188.82 175.19 177.94
```

We sample another 6 people. Their heights are

```
## [1] 164.53 180.91 175.36 166.61 188.82 165.57
```

So height is a random variable that depends on the random process of sampling the population

Notation

- Call our random variable of interest X
  - in coin example $X =$ number of heads
  - in heights example $X =$ height
- After we observe values we denote these with lower case x
  - coin example $x = 2$ and $x = 1$
  - heights example $\{x_1 = 177.16, x_2 = 169.96, x_3 = 181.95, x_4 = 188.82, x_5 = 175.19, x_6 = 177.94\}$ etc

Probability essentials

Probability

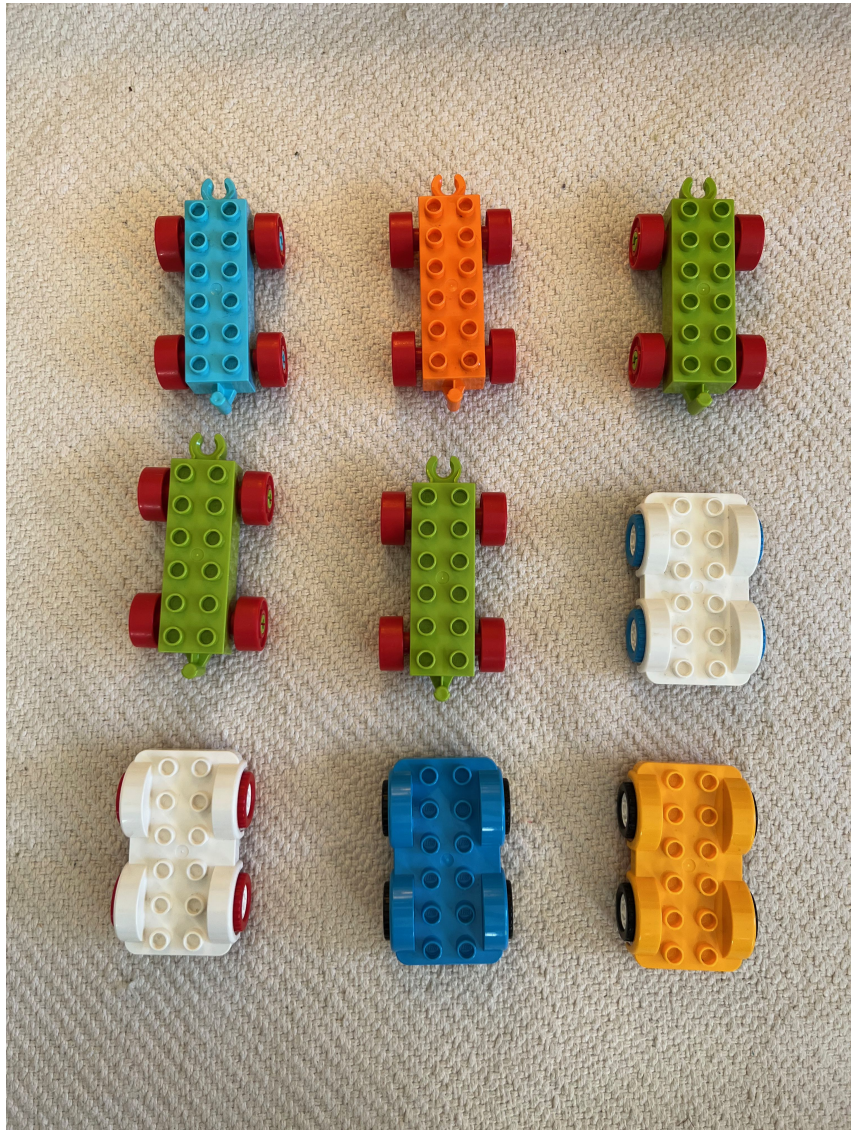
- Based on our sample or other random process (as in the coin flipping or a toddler choosing Lego), we would like to make valid statements about the underlying population or quantity of interest
- Probability is one tool that will help us do that
- Probability is all about talking about the chance of 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

- **Experiment**: An experiment can be any process, in a laboratory or otherwise, where we can observe the result of a process and the result of that process is uncertain.
- **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

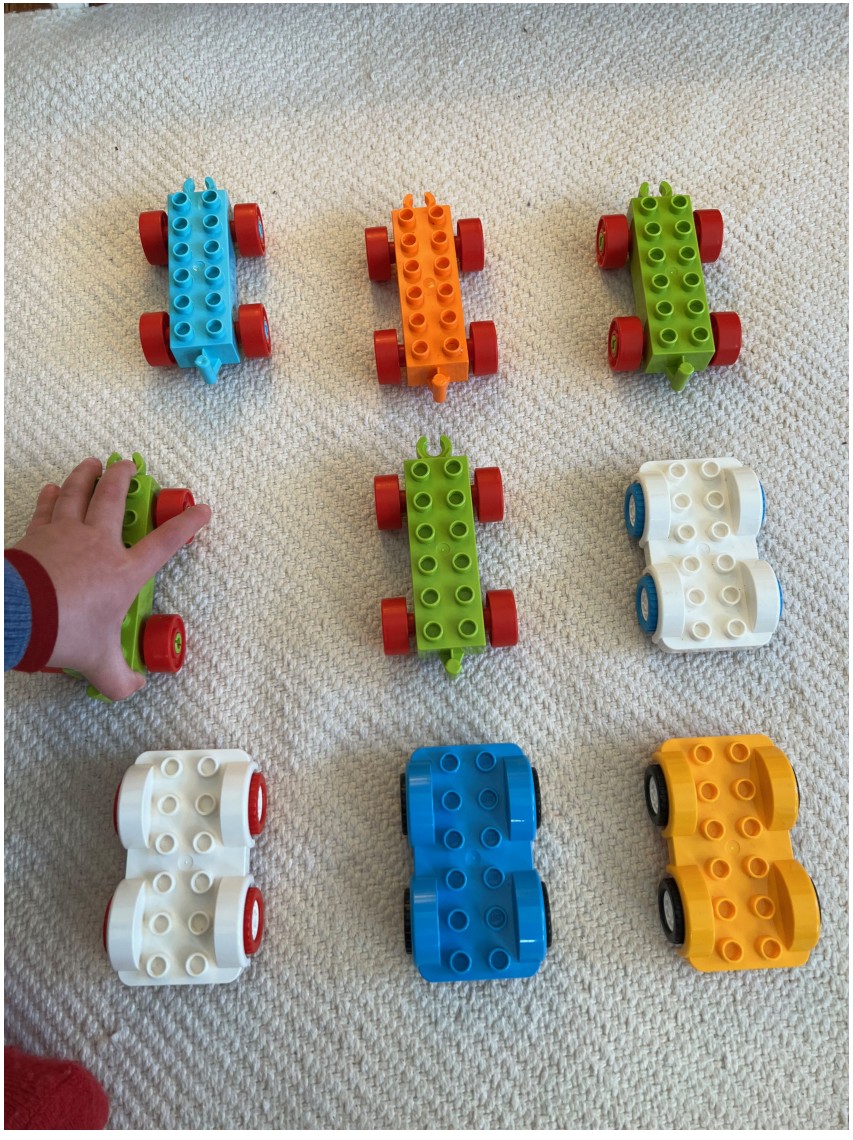
Lego example

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!

Additive / Union rule

What is $P(A \text{ or } B)$? That is the probability that the vehicle is a train or is the color blue?

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Note that if A and B are **mutually exclusive** then they can't happen together so $P(A \text{ or } B) = P(A) + P(B)$.

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?

Multiplicative / Intersection rule

What is $P(A \text{ and } B)$? That is the probability that the vehicle is a train and is the color blue?

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

Independence

If two events A and B are independent, then $P(A)$ is not affected by the condition B, and vice versa, so we can say that $P(A|B) = P(A)$ and likewise, $P(B|A) = P(B)$, so the multiplicative rule becomes

$$P(A \text{ and } B) = P(A) \times P(B)$$

Complements

the complement of any event A is the event [not A], i.e. the event that A does not occur. It is denoted A^c.

Lego practice

Interpret and calculate the following

- $P(B|A)$
- $P(A|B)$
- $P(A^c)$
- $P(A|B^c)$

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**

The expected value of a random variable

- In the first week, we discussed summary measures of data (measures of central tendency, spread, range)
- These types of measures are also useful to summarize sampling distributions of random variables.

Intuitively, on average, we would expect to see about two heads and two tails in any sequence of four trials. Imagine that we were to replicate the four tosses many, many times and take note of the number of heads in each of the series of four tosses. After many, many series, we could find the mean number of heads across the large number of series; it would seem reasonable to expect that the mean should be quite close to two. This is the **expected value**.

The expected value of a random variable

For a discrete random variable, X, with a known probability distribution $P(X_i)$ and where X_i is the ith outcome in the set of k simple events:

$$E(X) = X_1 \times P(X_1) + X_2 \times P(X_2) + \ldots + X_k \times P(X_k) = \sum_{i=1}^{k} X_i \times P(X_i) = \mu$$

The expected value is a weighted mean of all the possible values of the RV, weighted by their probabilities. It is given the symbol μ.

Calculate the expected value for the number of heads in four coin flips.

Expected value

The variance of a random variable

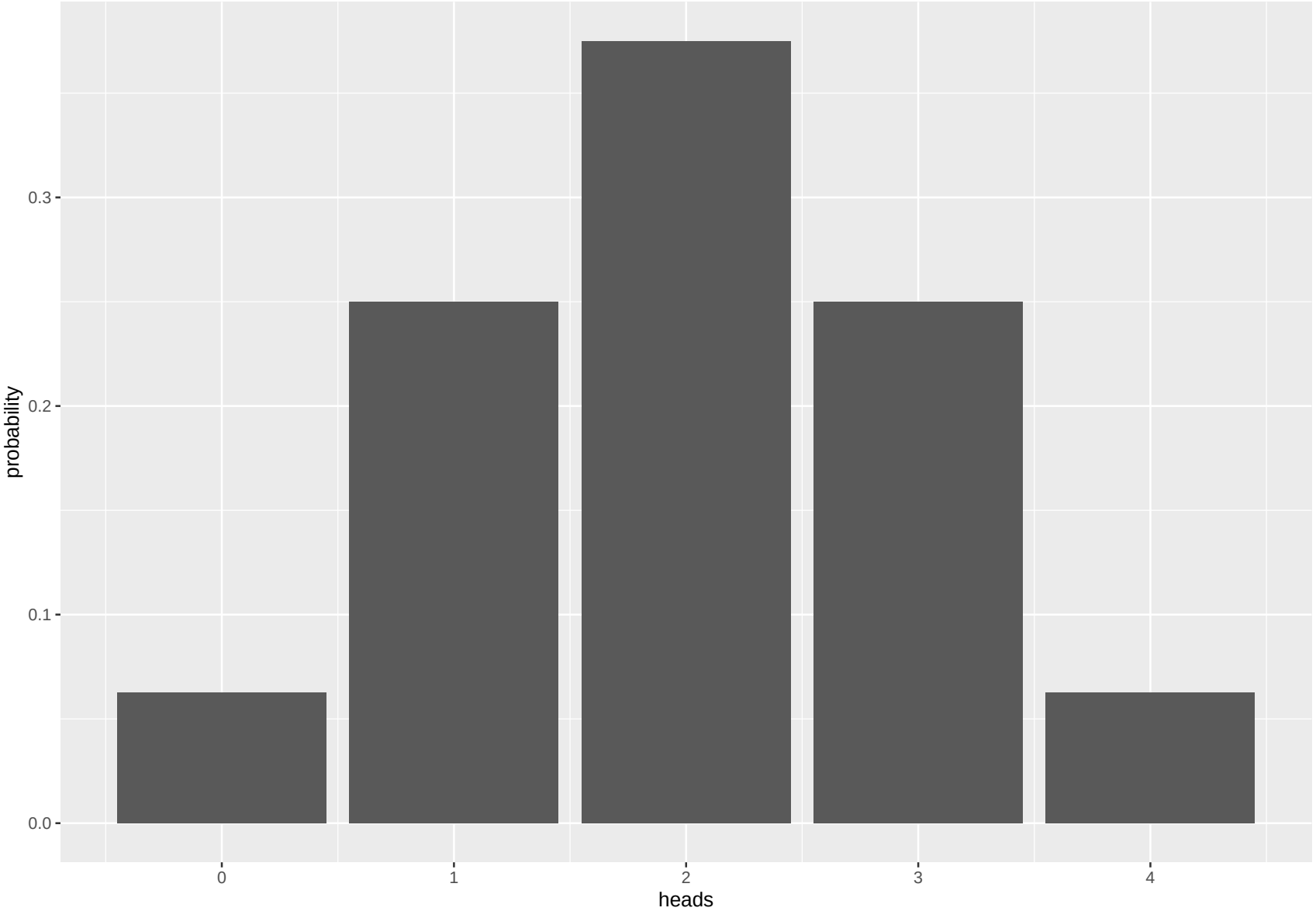
In week 1 we defined the variance is the average of the squared deviations from the mean. We can use the definition of expected value to derive the variance, given the probability distribution

$$\begin{aligned}\sigma^2 &= E\left[(X-\mu)^2\right] \\ &= [X_1 - E(X)]^2 \times P(X_1) + \ldots + [X_k - E(X)]^2 \times P(X_k) \\ &= \sum_{i=1}^{k} [X_i - E(X)]^2 \times P(X_i)\end{aligned}$$

Calculate the variance for the number of heads in four coin flips.

Variance

Probabilities as areas



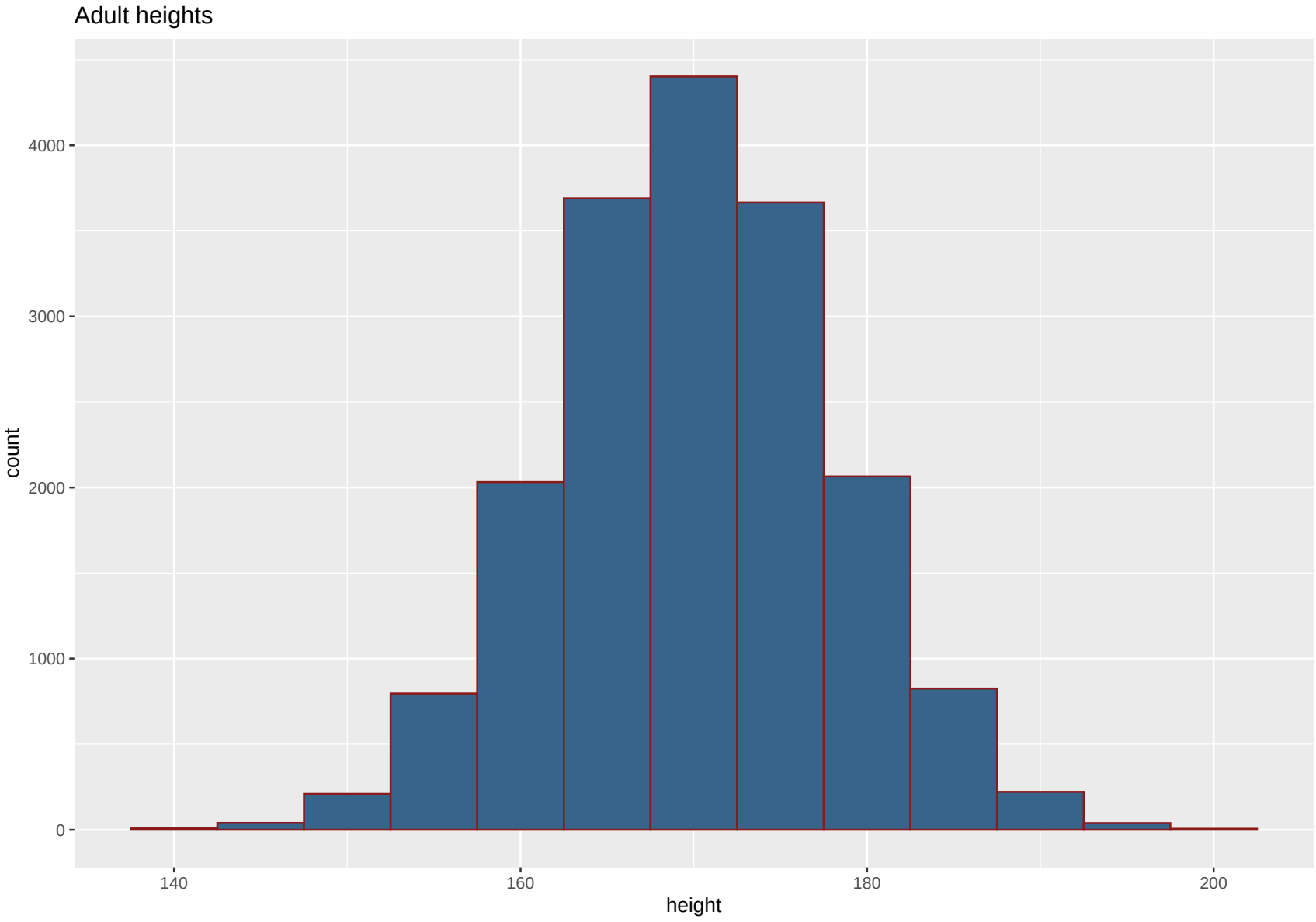
Probabilities as areas

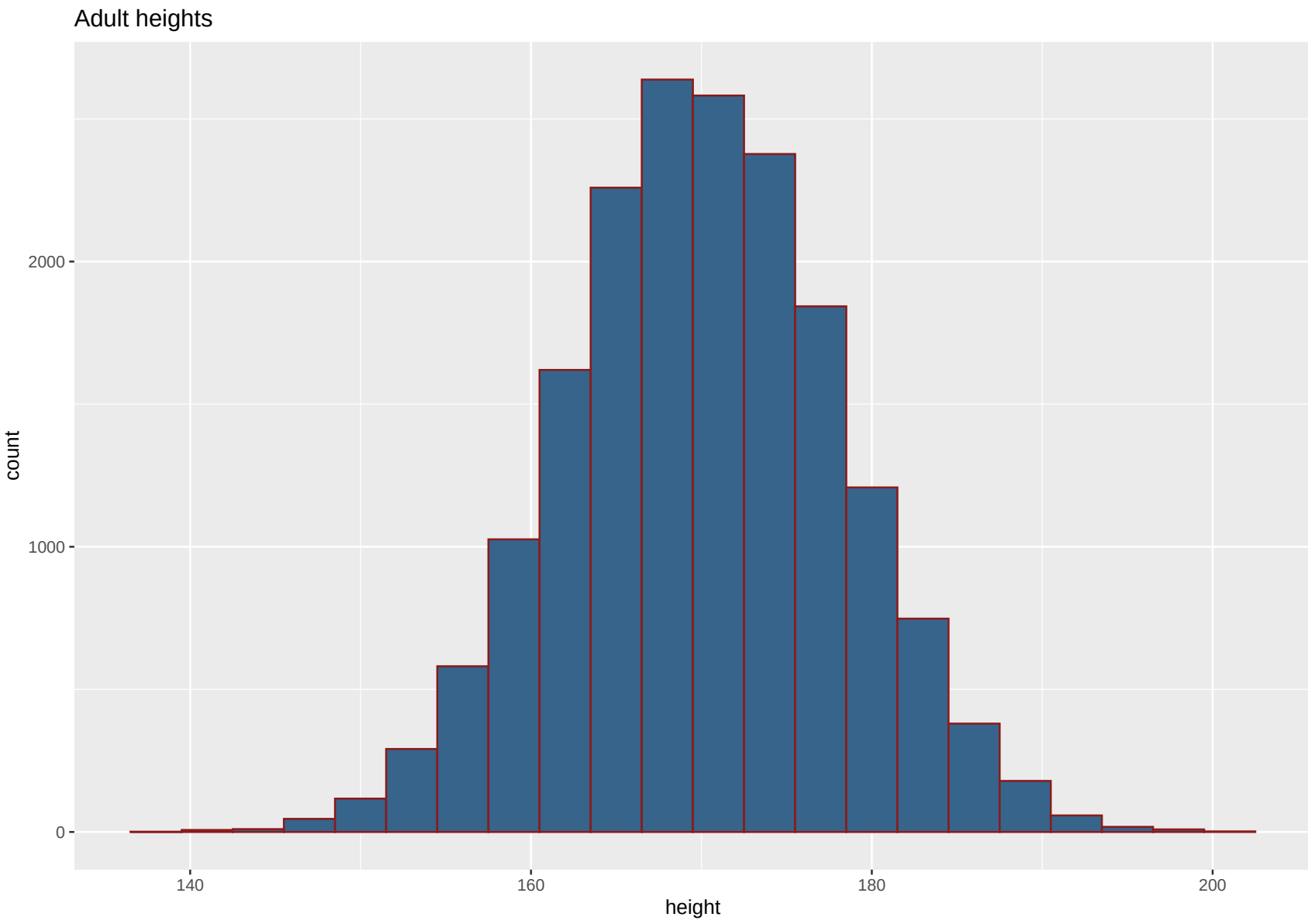
- To calculate probabilities, can sum up the area of the rectangles
- E.g. $P(X \geq 3)$ would be the sum of the right two rectangles
- What is $P(1 \leq X \leq 3)$?
- What is $P(1 \leq X < 3)$?

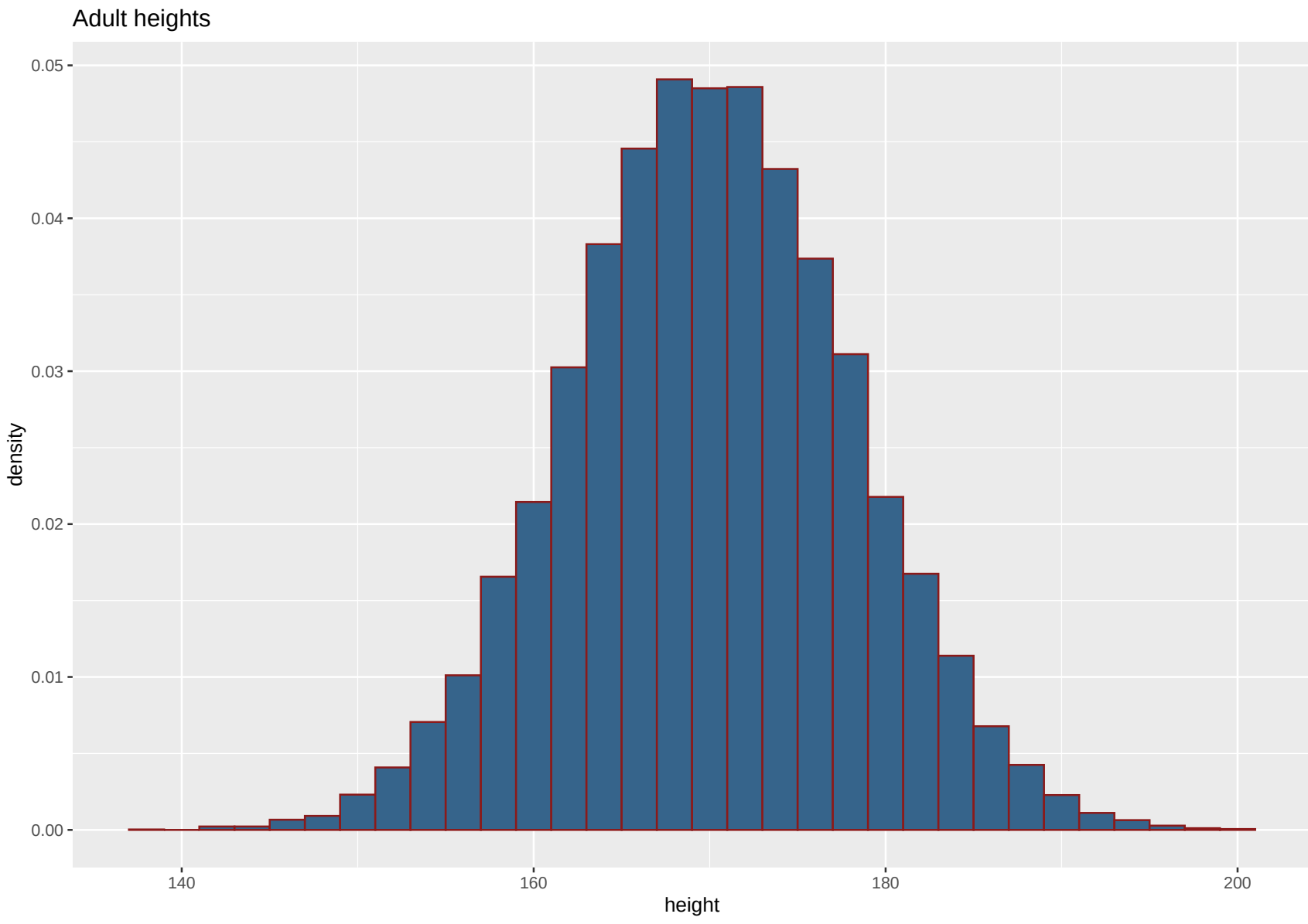
Continuous random variables and probability distributions

- 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**

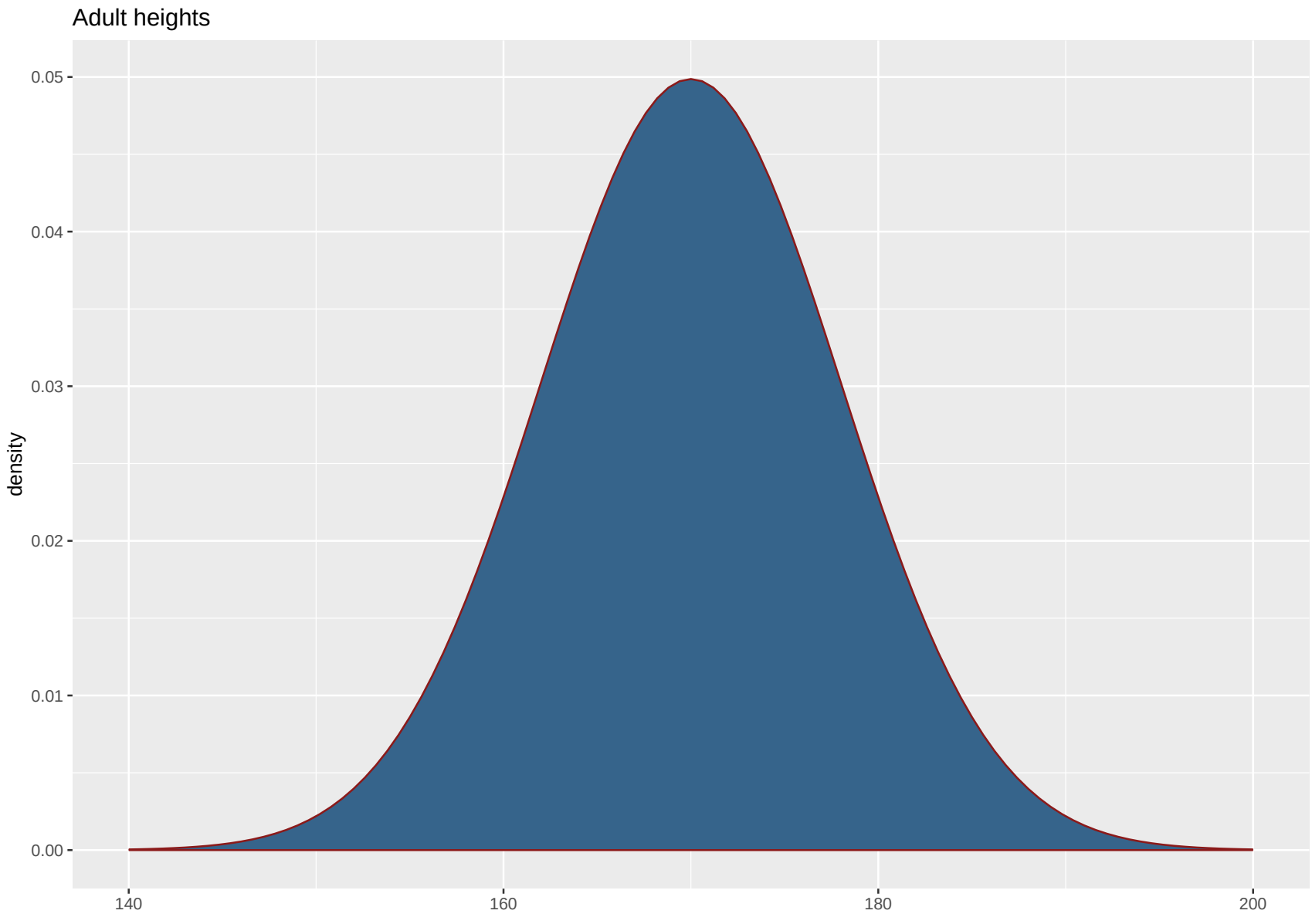
A continuous probability distribution is just a histogram with infinitely small bins



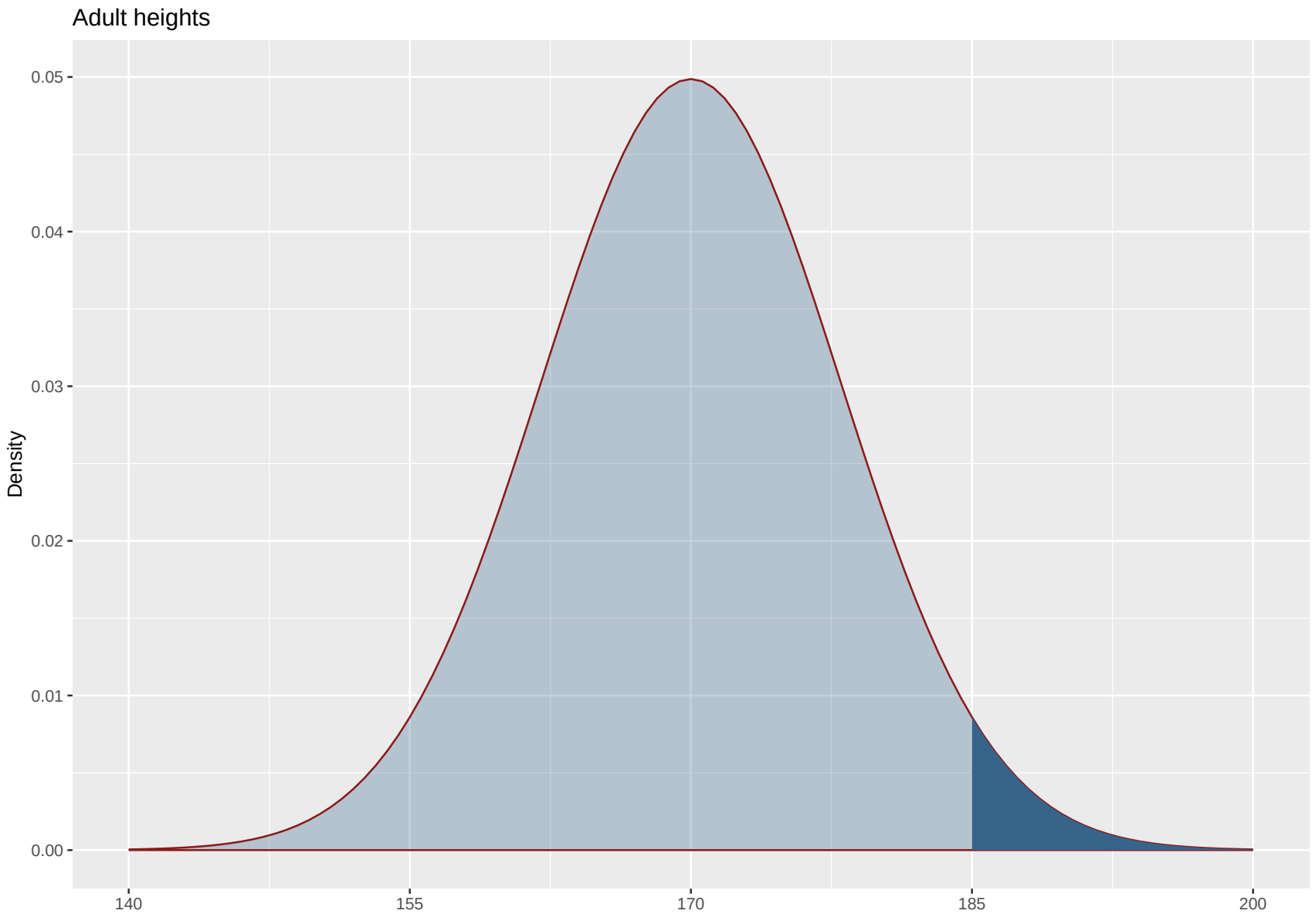

Adult heights
count
2000
1000
0
140
160
180
200
height


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

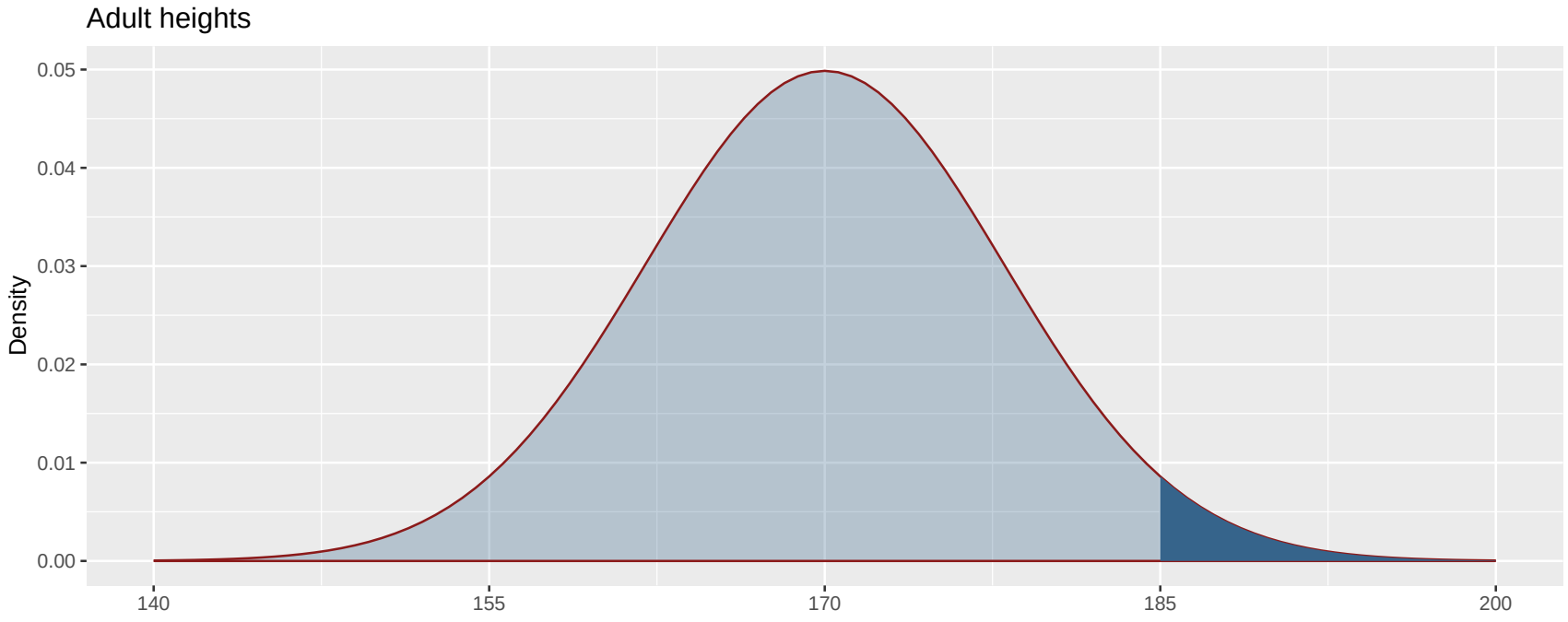
Probability density function



Probabilities as areas



Probabilities as areas

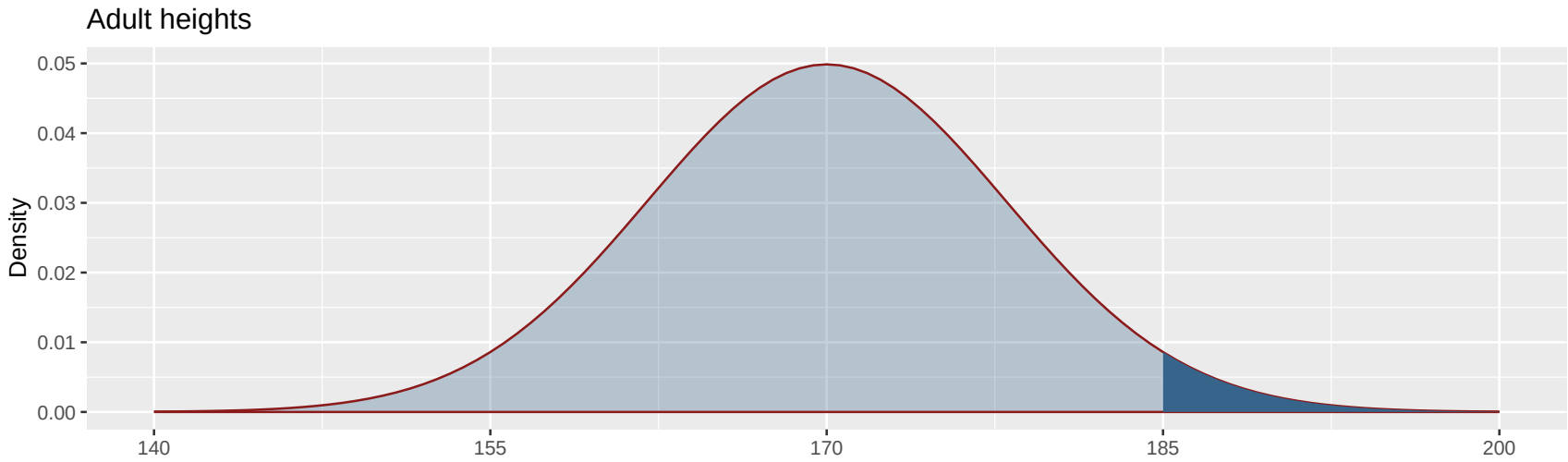


- The probability that height is greater than 185cm i.e. $P(X > 185)$
- Like summing up very tiny histogram bins above a certain point

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. $P(X > 185) = 1 - P(X \leq 185)$



Summary

- Probability concepts
  - Additive rule, mutually exclusive events, multiplicative rule, independence, complements
- Probability distributions
  - Discrete RV = probability mass function
  - Continuous RV = probability density function
- Probabilities as areas

Observing data and sampling distributions (intro)

Simulating outcomes in R

We can **simulate** coin flips in R using the `sample` function

```
possible_events <- c("H", "T")
coin_flips <- sample(possible_events, size = 4, replace = TRUE)
coin_flips
```

```
## [1] "H" "T" "H" "T"
```

Simulating outcomes in R

We can do this simulation experiment more than once and count the number of observed heads each time. For example, the output below show the results of 100 experiments (of four coin flips) and the number of times each value of heads was observed.

heads	n
0	7
1	23
2	45
3	17
4	8

The sampling distribution

heads	n	probability
0	7	0.07
1	23	0.23
2	45	0.45
3	17	0.17
4	8	0.08

A **sampling distribution** is a probability distribution for a statistic based on repeated samples. Here, our random-sample-based statistic is the number of heads in four coin flips.