Probability

Mathematical background for statistics

Download the section 8 .Rmd handout to `STAT240/lecture/sect08-probability-basics`.

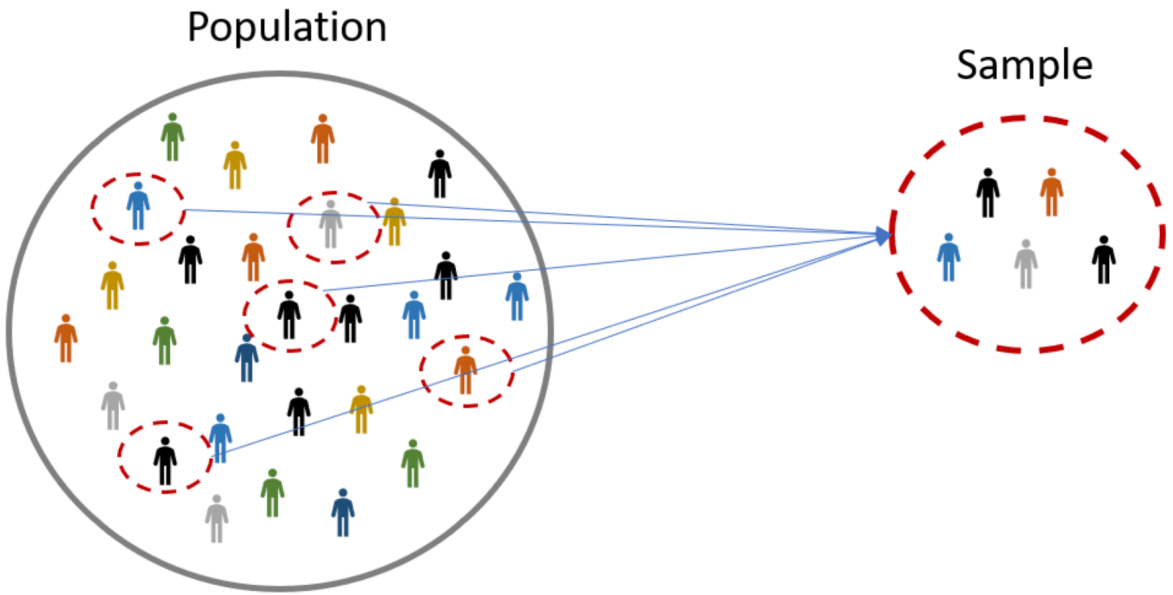
Material in this section is covered by Chapters 9 and 10 on the notes website.

In statistics, we work with a sample of data.

Freeze duration:
118, 151, 121, 96, 110, 117, 132, 104, 125, ...

Heights:
65, 68, 70, 72, 64, 63, ...

We assume random samples are drawn from a **population**.



Goal: learn about population

- What is the true average freeze duration?
- What is the true mean height of US males?

We take a random sample, which we *can* measure.

This introduces uncertainty. Statistics is based on **probability**, the mathematics of randomness.

Probability is used informally all the time.

- I might miss my bus if I don't hurry up.
- I will need to have a good day on the final to get an A.

- On another day, we would've won that game.
- I didn't even leave late or anything, I just hit every red light.

We represent data collection as a **random process**.

- Any "experiment" with a random outcome
- Flip a coin twice
- Measuring a random item from a population

Repeating this process and recording the results is how we generate data.

Most analyses we do will focus on **numeric** data.

- Height
- Number of siblings
- GPA

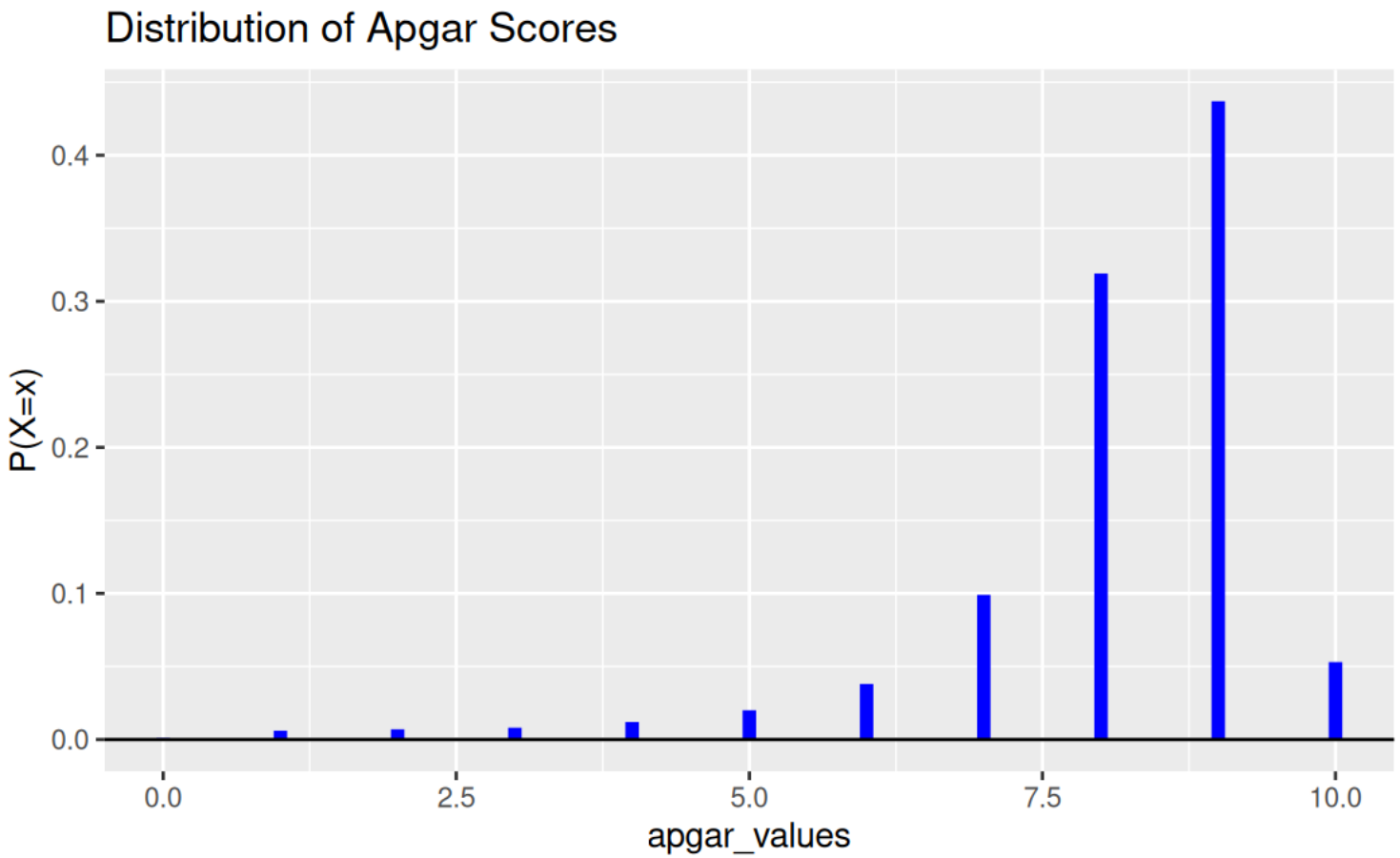
Numeric data is sampled from a population of numbers. This is represented with a **random variable** (RV).

The Apgar score for newborns is on a scale from 0-10 based on condition immediately after birth.

x	0	1	2	3	4	5
$P(X = x)$	0.001	0.006	0.007	0.008	0.012	0.020

x	6	7	8	9	10
$P(X = x)$	0.038	0.099	0.319	0.437	0.053

X is a RV, x is a possible value, and $P(X = x)$ is the **probability**.


Distribution of Apgar Scores
P(X=x)
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
apgar_values

The **probability** of an outcome is a number between 0 and 1 giving:

- The probability of seeing the outcome for the next sampled item
- The proportion of items in the population with this outcome

If we ran the random process infinite times, it is the percentage of time we would get the given outcome.

Some probabilities are intuitive.

- $P(\text{Heads})$ from the flip of a fair coin
- $P(3)$ from rolling a fair die

For RVs, these are described as $P(X = x)$.

These can get complicated quickly. Roll two dice, and let S be their sum. What is $P(S = 10)$?

A **probability distribution** gives all possible RV values and their corresponding probabilities. For example, the Apgar score.

x	0	1	2	3	4	5
$P(X = x)$	0.001	0.006	0.007	0.008	0.012	0.020
x	6	7	8	9	10	
$P(X = x)$	0.038	0.099	0.319	0.437	0.053	

Properties of probability distributions:

- Probabilities must be between 0 and 1.
- Probabilities must all sum to 1.

If there are a few finite outcomes, we can list the probabilities in a table.

To recap:

- A **random process** or **experiment** is the process that randomly generates data
- A **random variable** is the quantity of interest corresponding to each outcome
- A **probability distribution** assigns a probability to each possible value of the RV

These are theoretical tools representing a population of interest.

Consider flipping a coin three times.

- What are the possible outcomes?
- What is the probability of each outcome?

Let RV X count the number of heads out of three coin flips.

- What are the possible values of X?
- What is the probability distribution of X?

A **discrete** random variable can only take on specific values.

- Number of people in a household
- Number of goals scored in a soccer match

A **continuous** random variable can take on any value in a range.

- Time to walk to class
- Inches of rainfall

The Apgar score and the "count of heads" RVs are both discrete.

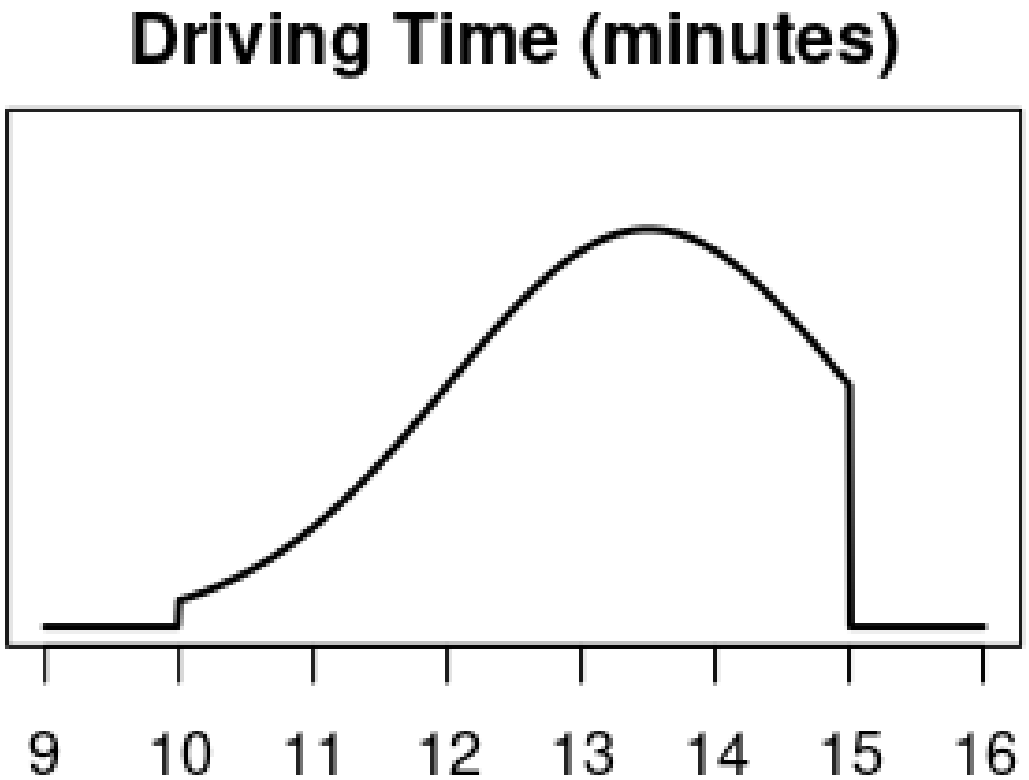
- Values can be enumerated in a table
- Probabilities sum to 1
- The numbers in between are impossible

Just because 0 and 1 are possible values, it doesn't mean (0.1, 0.535, 0.9999) also have probability.

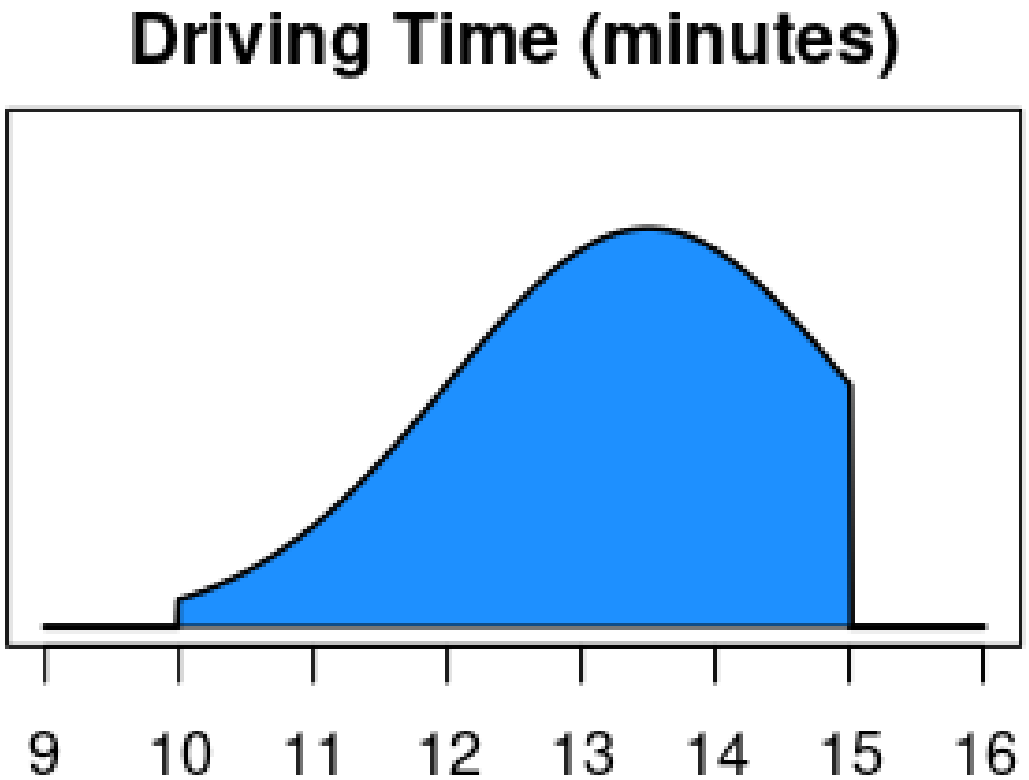
A **continuous** random variable can take on any value in a specific range. The shape of a continuous RV is a density function, similar to `geom_density`.

We draw the pdf as a smooth curve.

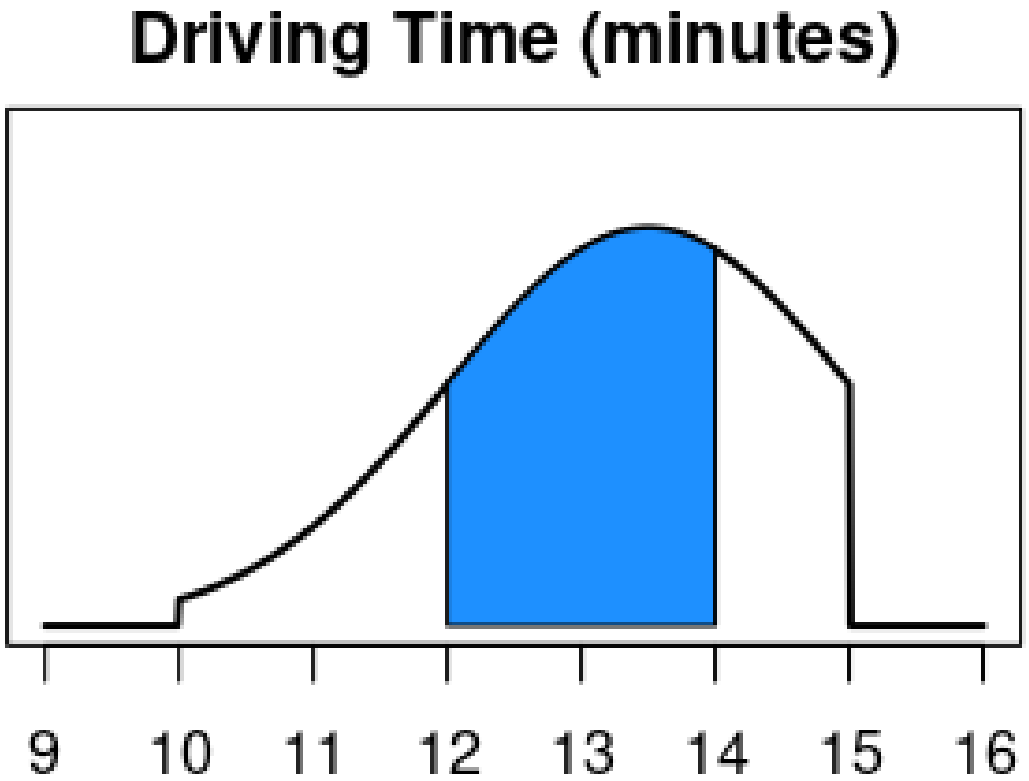
It takes 10 to 15 minutes to drive to campus, depending on traffic. The curve of the pdf represents all possible driving times.



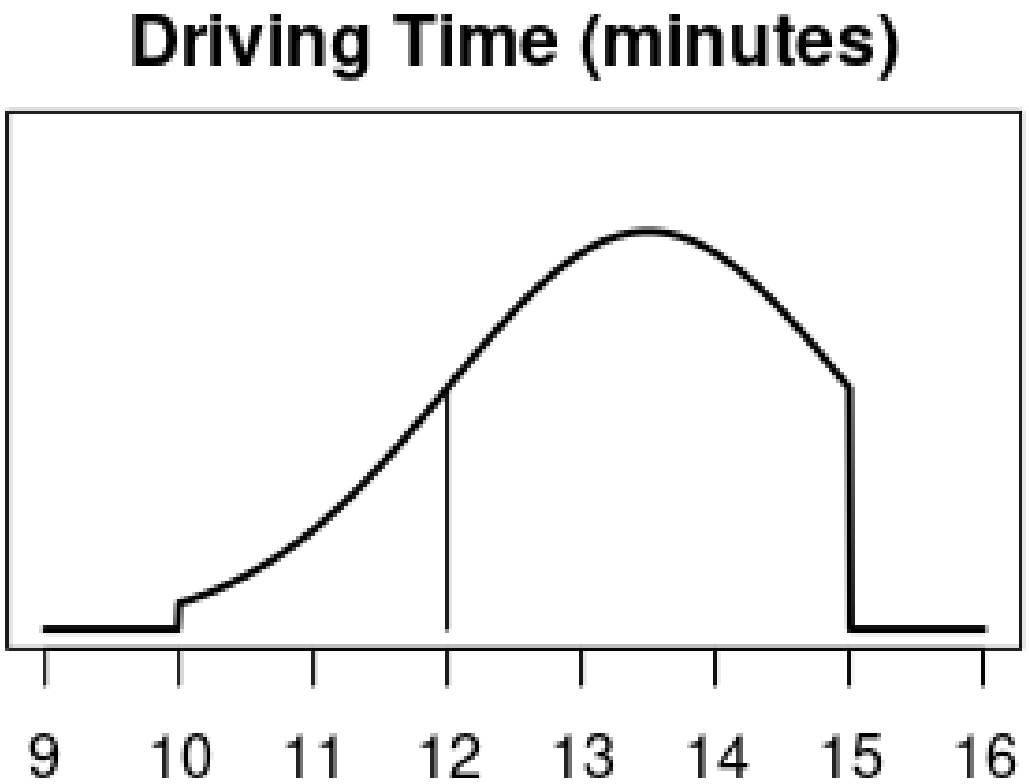
Probabilities are given by the *area* under the pdf.



The area under the entire pdf is 1.

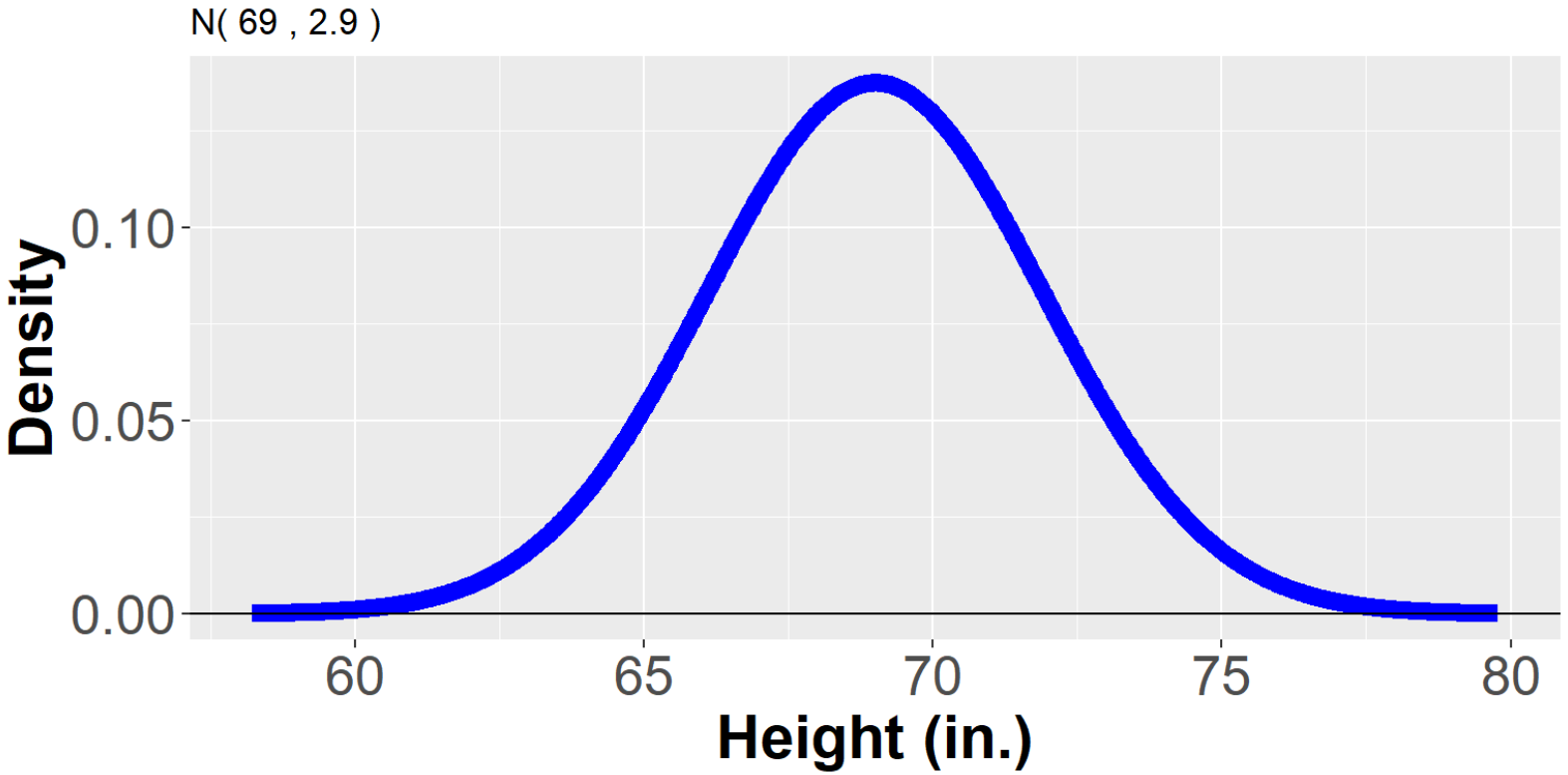


$P(12 \leq \text{driving time} \leq 14)$.

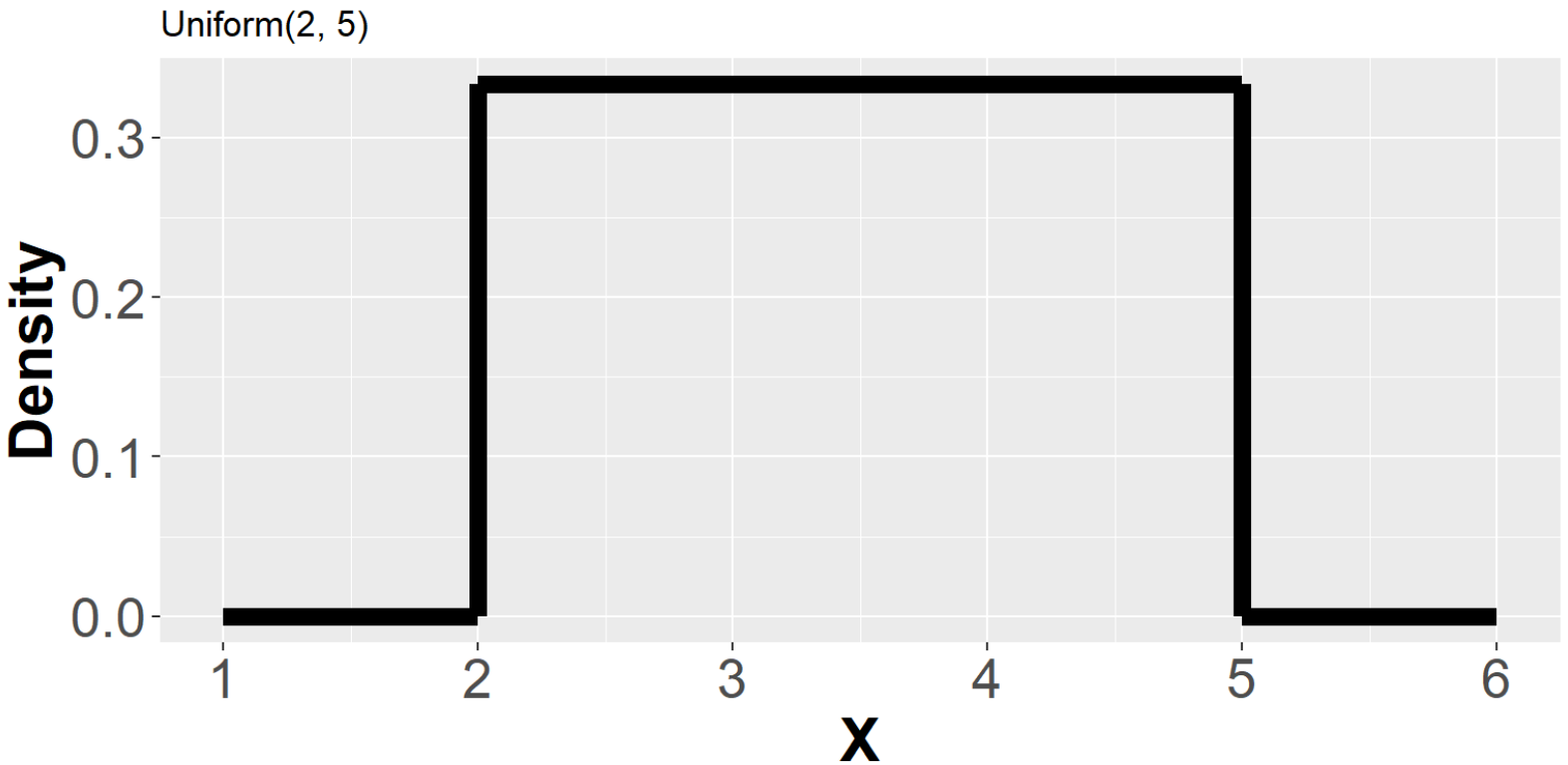


The probability of a single value (like $P(\text{driving time} = 12)$ is 0.

Height of US males:



Random number between 2 and 5:



RVs are numerical, which makes them a useful framework for studying populations.

Using the values of the RV, we can calculate summary measures about a population.

- Expected value: population mean (center)
- Variance: population variance (spread)

Suppose we have a population where 90% of the values are 0, and 10% of the values are 1.

$$P(X = 0) = 0.9, \qquad P(X = 1) = 0.1$$

Intuitively, it doesn't make sense to say the mean is $(1 + 0)/2 = 0.5$. The mean is a weighted average:

$$(0)0.9 + (1)0.1 = 0.1 = E(X)$$

This RV mean is called **expectation**.

In general, the expectation of a discrete RV is:

$$E(X) = \mu = \sum_x x \cdot P(X = x)$$

which is a weighted average of the values of X.

If we were able to take a sample of infinite x_i's and take their mean, $\bar{x}$ would be equal to μ.

The **population variance** is represented by how spread out the RV is.

Specifically, the **variance** of a RV is the average squared distance from the RV to its mean.

We start by taking $[x - E(X)]^2$.

Just like with expectation, we take a *weighted* average based on the probabilities.

The variance of a discrete RV is

$$V(X) = \sigma^2 = \sum_x [x - E(X)]^2 \cdot P(X = x)$$

The **standard deviation** of X is the square root of the variance.

$$sd(X) = \sqrt{V(X)} \quad \text{or} \quad \sigma = \sqrt{\sigma^2}$$

Expectation and variance have the same interpretation for continuous RVs.

- $E(X)$: population mean
- $V(X)$: population variance

To calculate the mean and variance of a continuous RV, we use calculus, which is not a part of 240.

Let's work with a small discrete example. Recall:

- Probabilities must sum to 1
- $$E(X) = \sum_x x \cdot P(X = x)$$
- $$V(X) = \sum_x [x - E(X)]^2 \cdot P(X = x)$$