Probability

Mathematical background for statistics

Download the section 7 .Rmd handout to `STAT240/lecture/sect07-probability`.

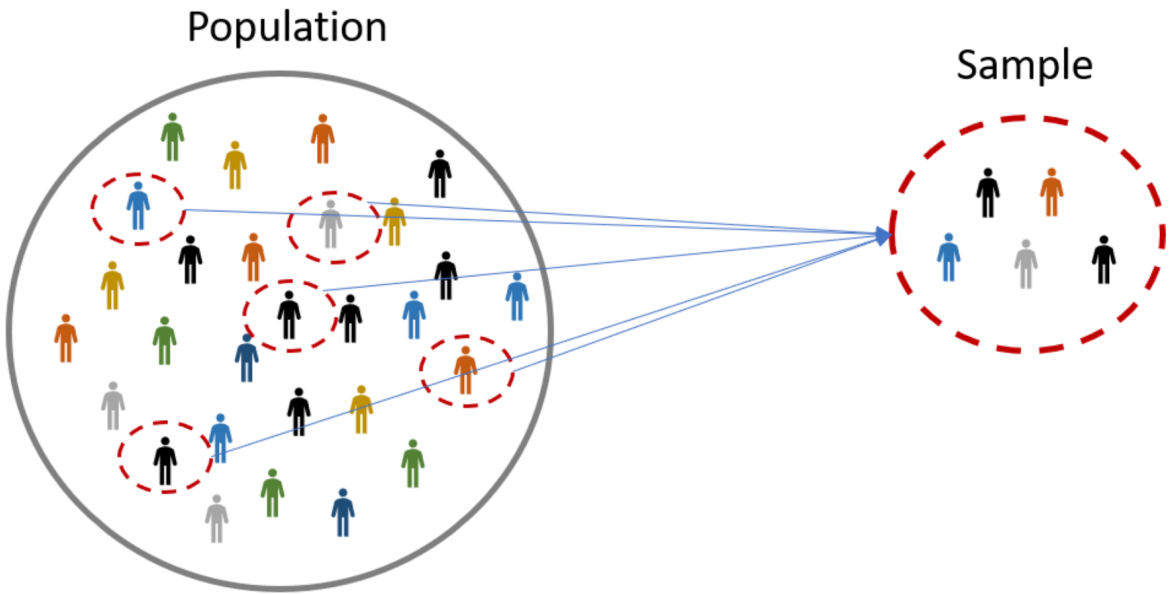
Material in this section is covered by Chapters 9 and 10 on the notes website.

In statistics, we work with a sample of data.

Penguin flipper lengths:
181, 196, 195, 193, 190, 181, ...

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 flipper length?
- What is the true average height of US males?

Briefly discuss: how can we perform a sample that *represents* the population of interest?

Bad news: sampling introduces uncertainty. What does a sample tell us about a population?

The fix:

- Build a theoretical model for the population
- Use **probability** to connect population and sample

Probability is used 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.

Represent sampling as a **random process**.

- Any "experiment" with a random outcome
- Flip a coin twice
- Measuring the flipper length of a random penguin

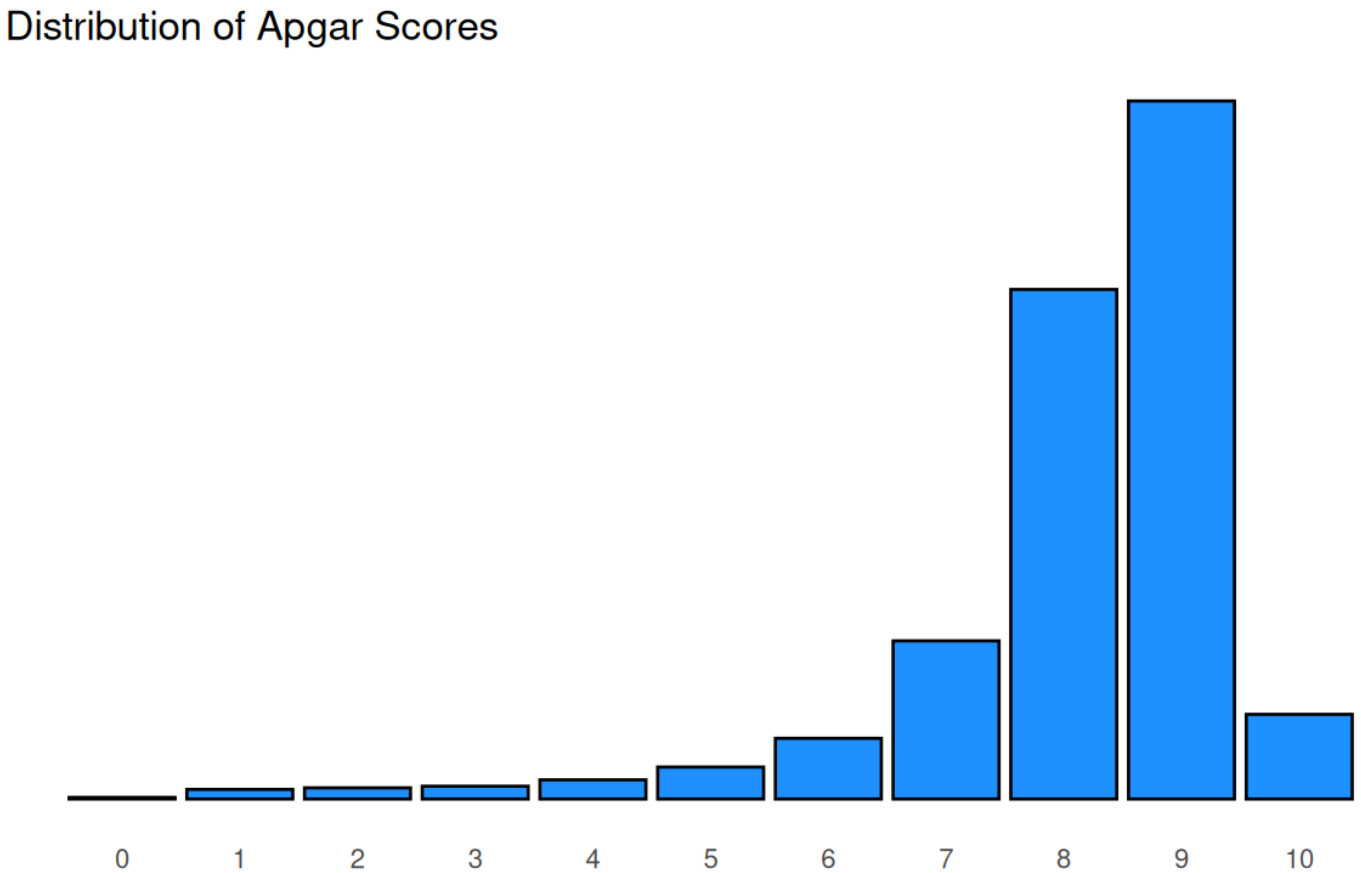
We can think of a population as a process for generating data.

Wil will focus on **numeric** data.

- Height
- Number of siblings
- GPA

A random, numeric process is represented by a **random variable** (RV).

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



An RV X has a **support** of possible values.

- X = process, x = specific outcome

x	0	1	2	3	4	5	6	7	8	9	10

But how common is each value?

x	0	1	2	3	4	5
	0.001	0.006	0.007	0.008	0.012	0.020

x	6	7	8	9	10
	0.038	0.099	0.319	0.437	0.053

- Probability of a random draw
- Breakdown of population values

A **probability distribution** assigns a number between 0 and 1 to each possible value of an RV.

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, write $P(X = x)$.

- Remember X and x are different!

The **probability distribution** of Apgar scores:

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** generates data
- A **random variable** is the numeric representation of the process
- A **probability distribution** assigns a probability to each value of the RV's **support**

These tools represent a population.

Flip 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 is the support of X?
- What is the probability distribution of X?
- Bonus: graph X in R

A **discrete** RV can only take on specific values.

- Number of people in a family
- Number of goals scored in a match

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

- Time to walk to class
- Inches of rainfall

The Apgar score is discrete.

- Support can be enumerated in a table
- Probabilities sum to 1
- Values inbetween are impossible

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

The distribution of a discrete RV is called a **probability mass function** (pmf).

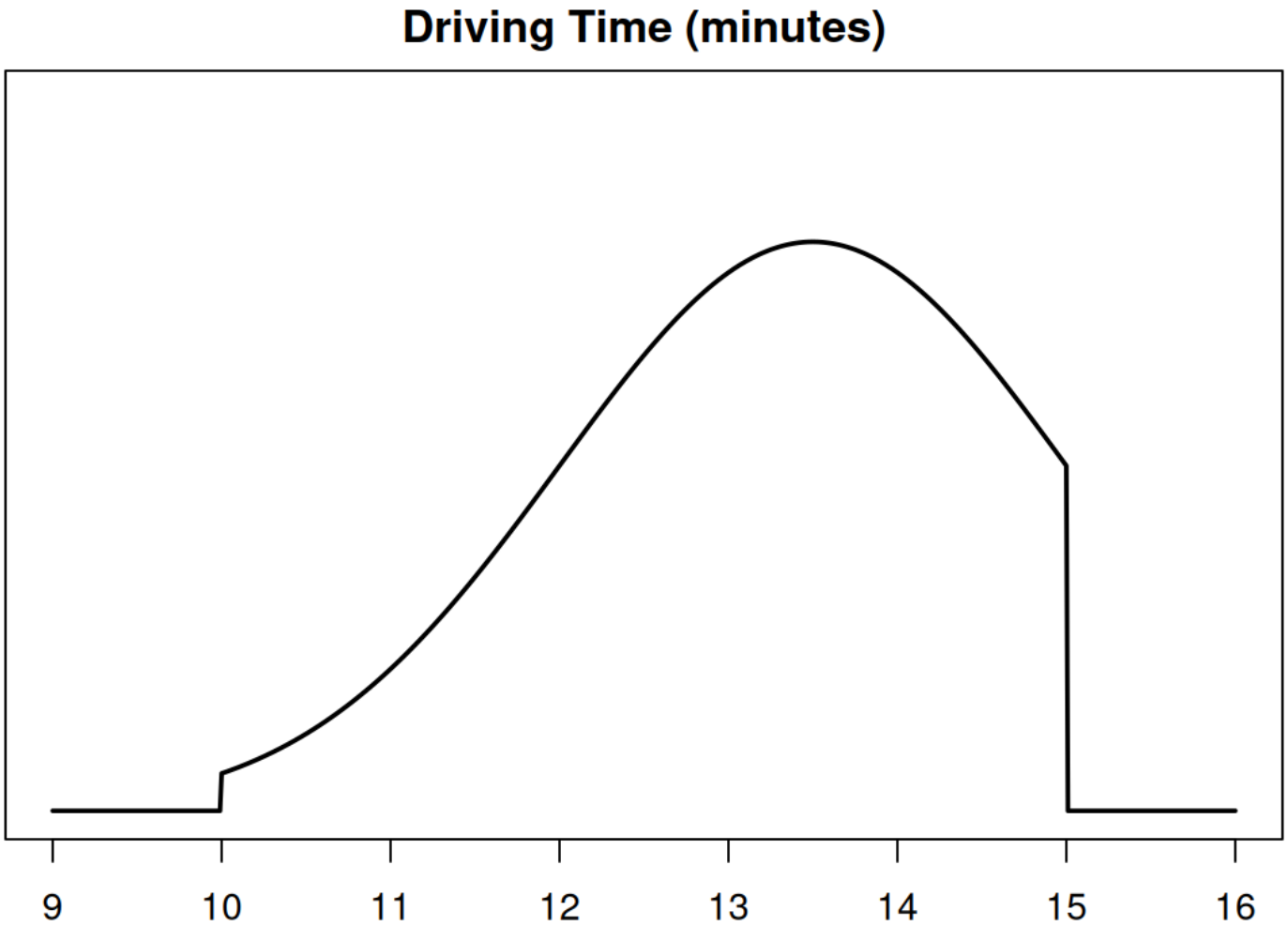
x	0	1	2	3	4	5
$P(X = x)$	0.001	0.006	0.007	0.008	0.012	0.020

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

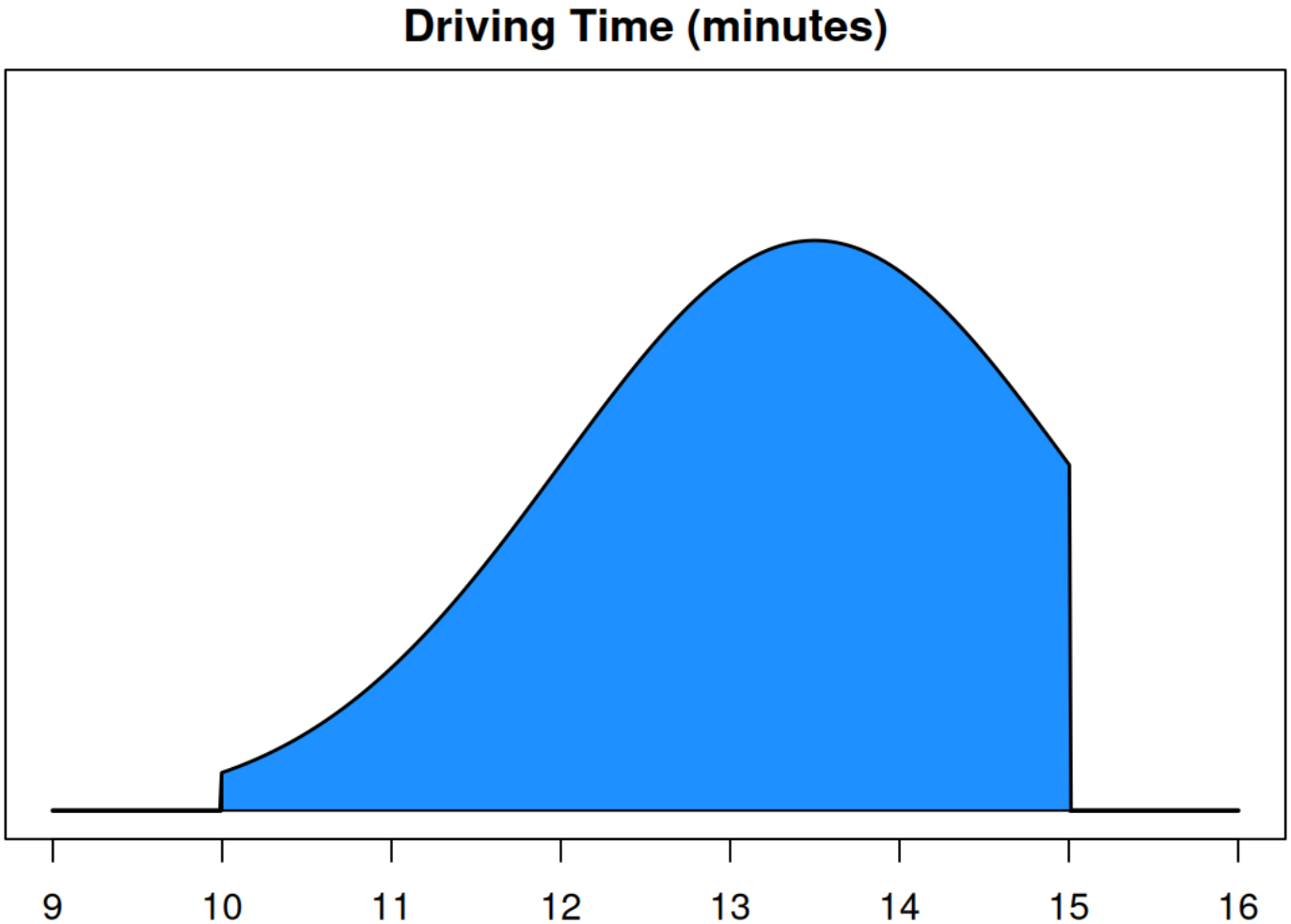
A **continuous** random variable's support is a range.

We draw the probability as a smooth curve, similar to `geom_density()`. It is a **probability density function** (pdf).

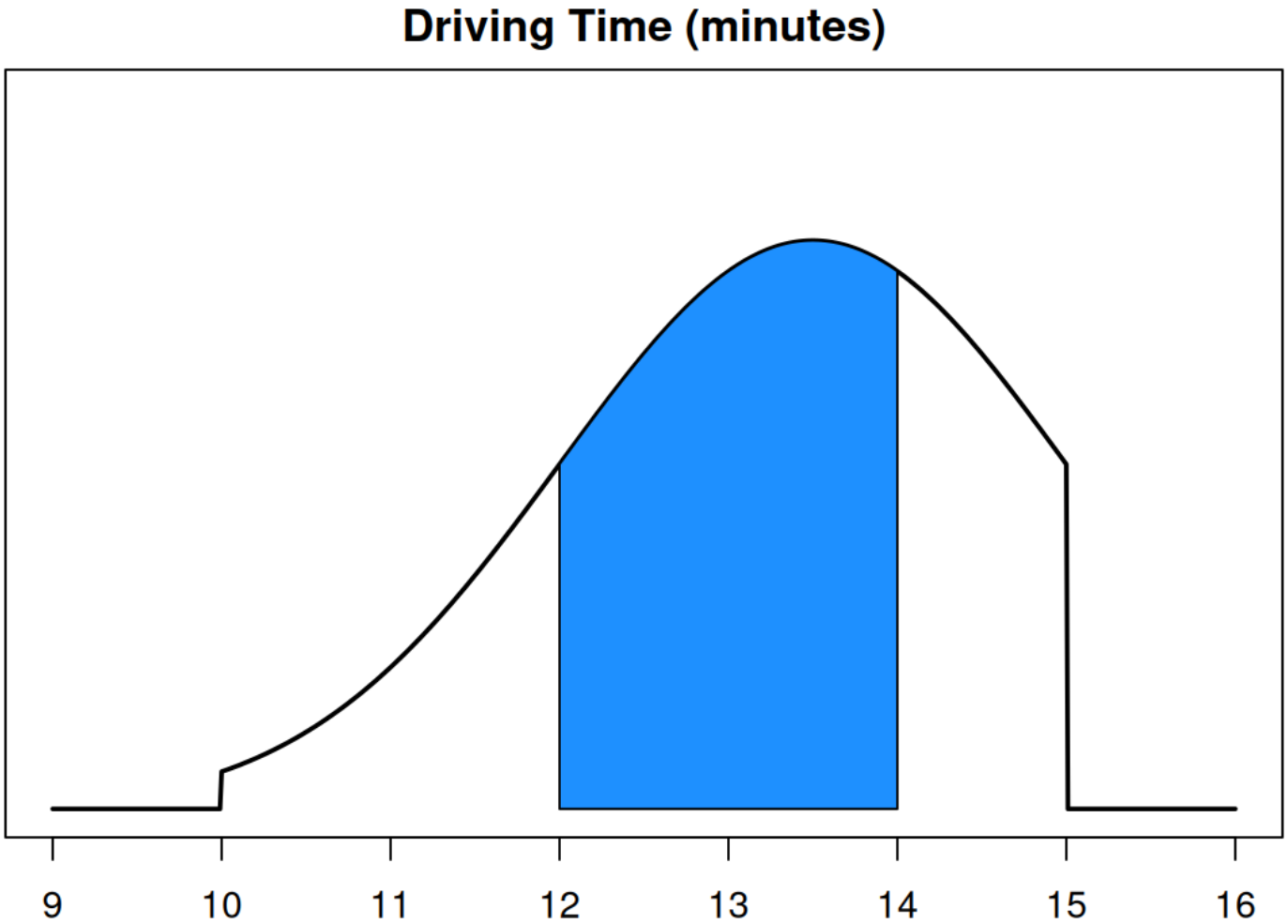
For example, it takes 10 to 15 minutes to drive to campus, depending on traffic.



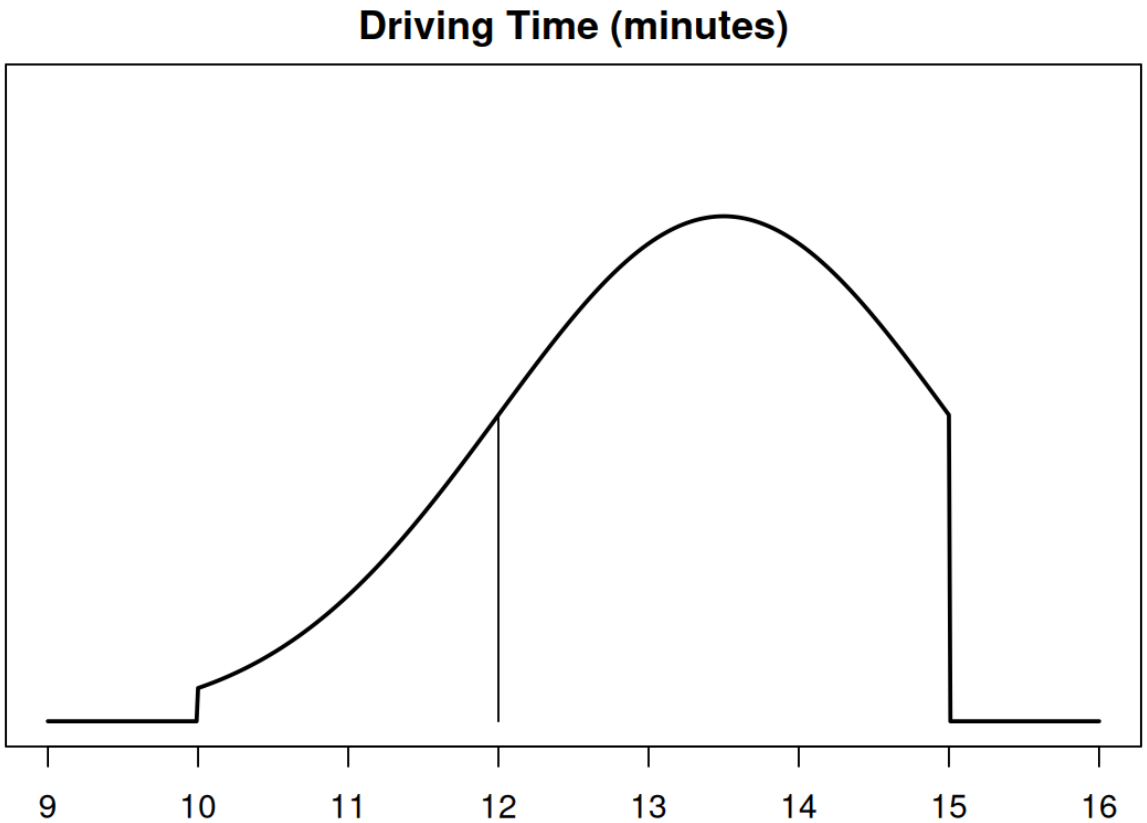
Probabilities are given by the *area* under the pdf.



The total area is 1.

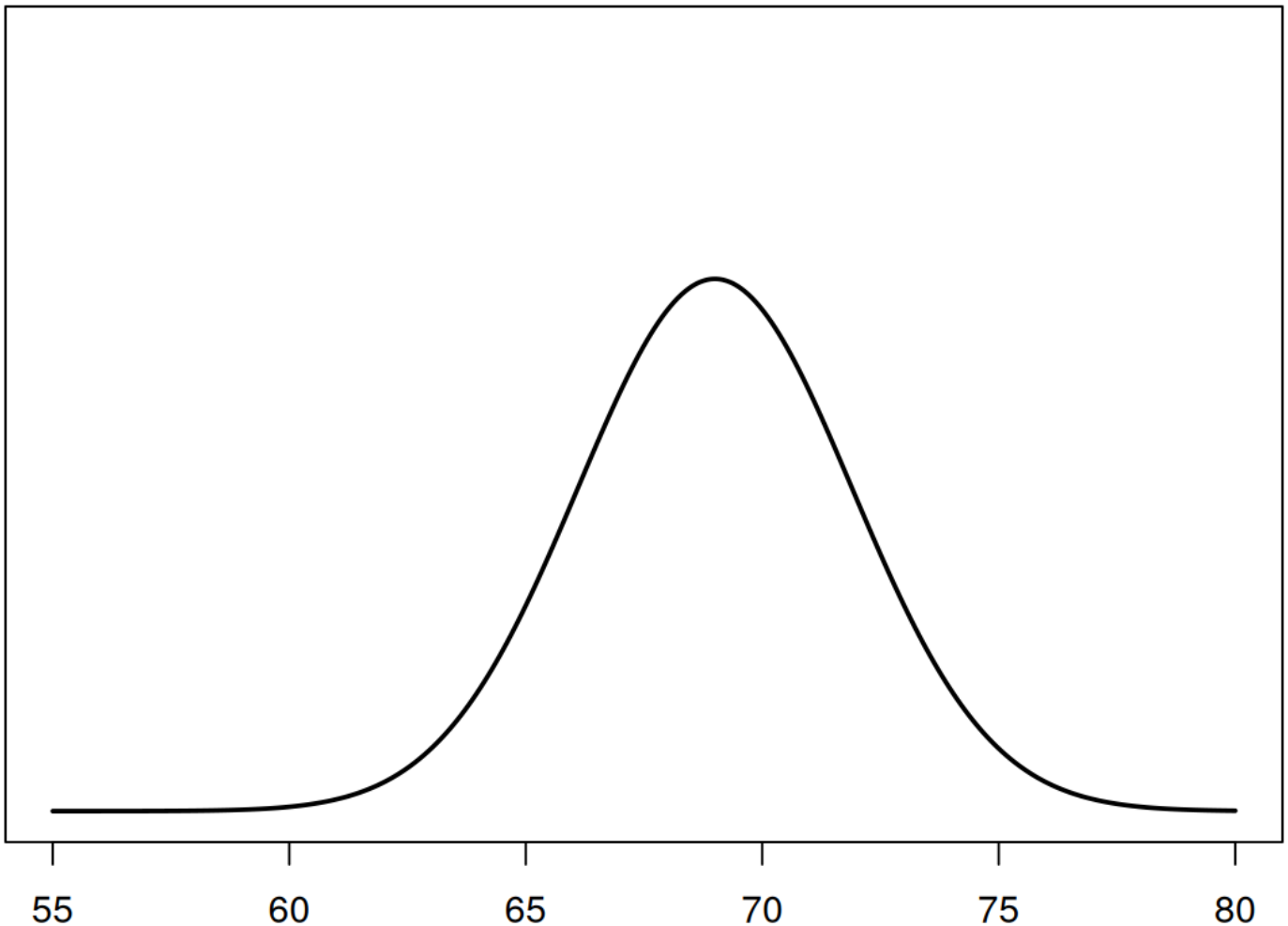


$P(12 \leq \text{driving time} \leq 14)$.

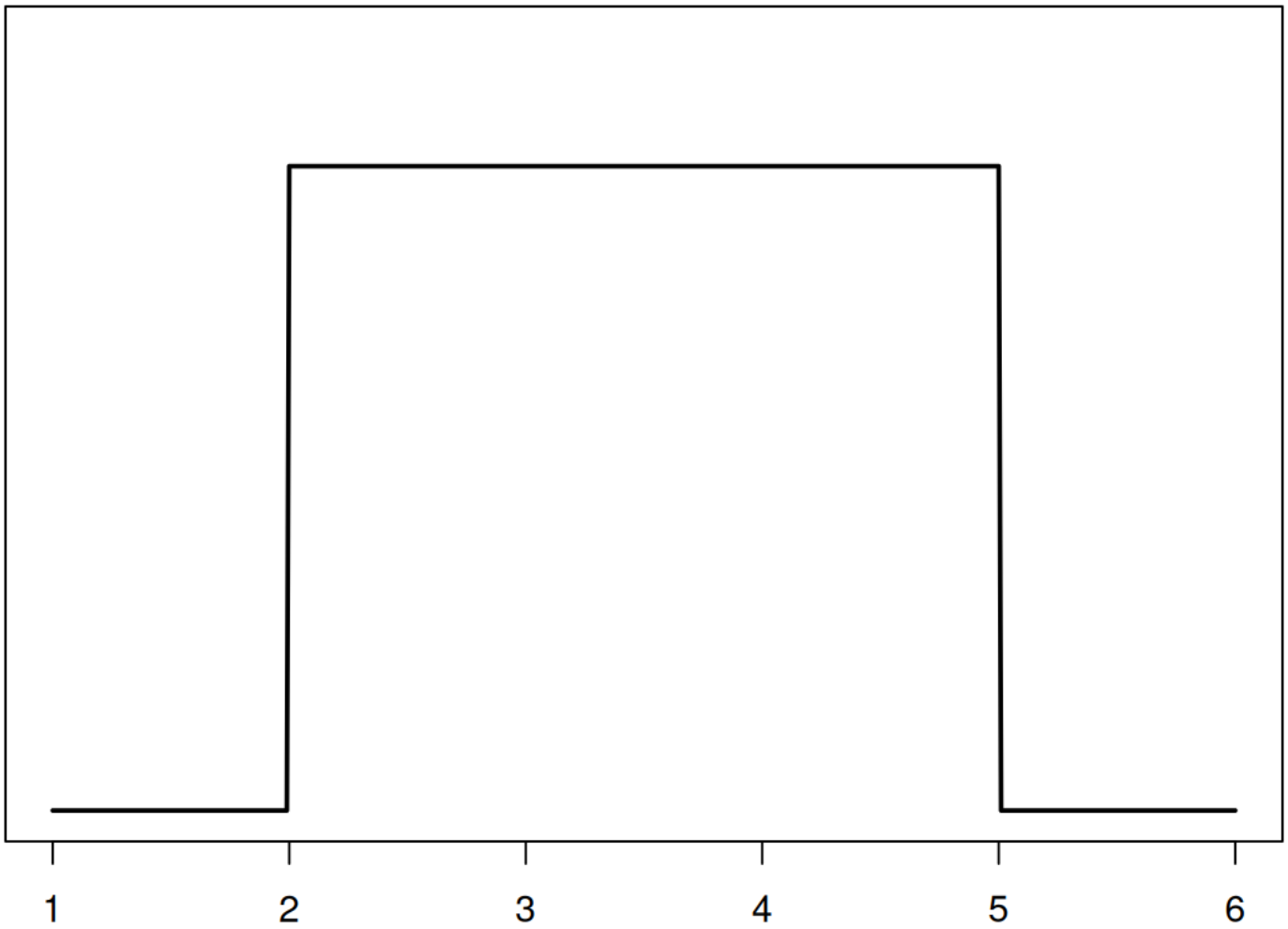


The probability of a single value like $P(\text{driving time} = 12)$ is 0.

Height of US males:



Random number between 2 and 5:



We can use RVs to calculate summary measures.

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

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

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

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