

# What are the chances?

INTRODUCTION TO STATISTICS



George Boorman

Curriculum Manager, DataCamp

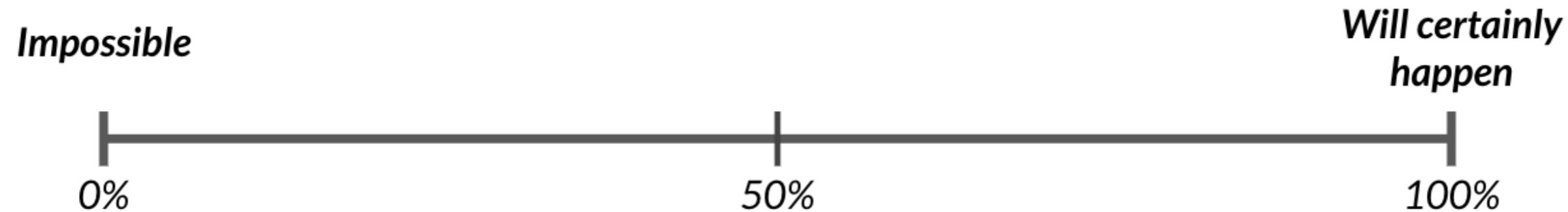
# Measuring chance

*What's the probability of an event?*

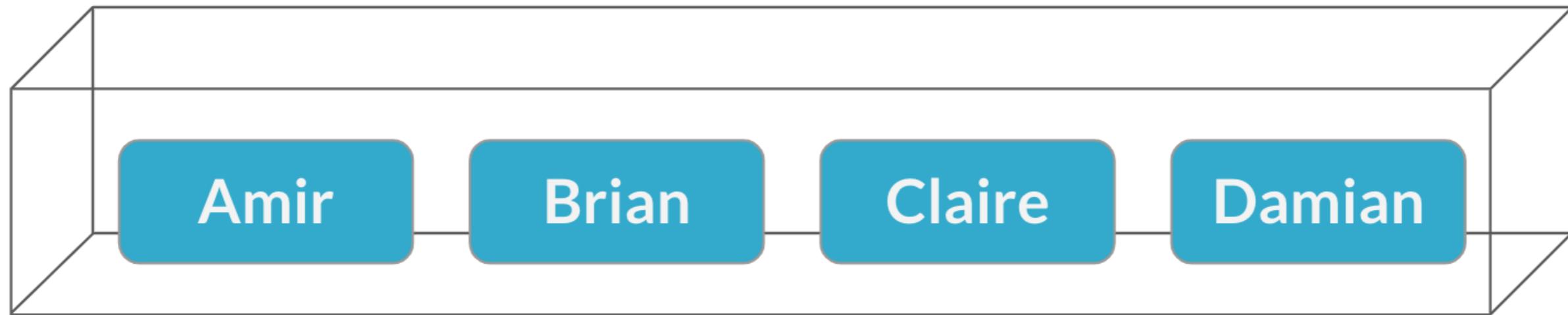
$$P(\text{event}) = \frac{\# \text{ ways event can happen}}{\text{total } \# \text{ of possible outcomes}}$$

*Example: a coin flip*

$$P(\text{heads}) = \frac{1 \text{ way to get heads}}{2 \text{ possible outcomes}} = \frac{1}{2} = 50\%$$

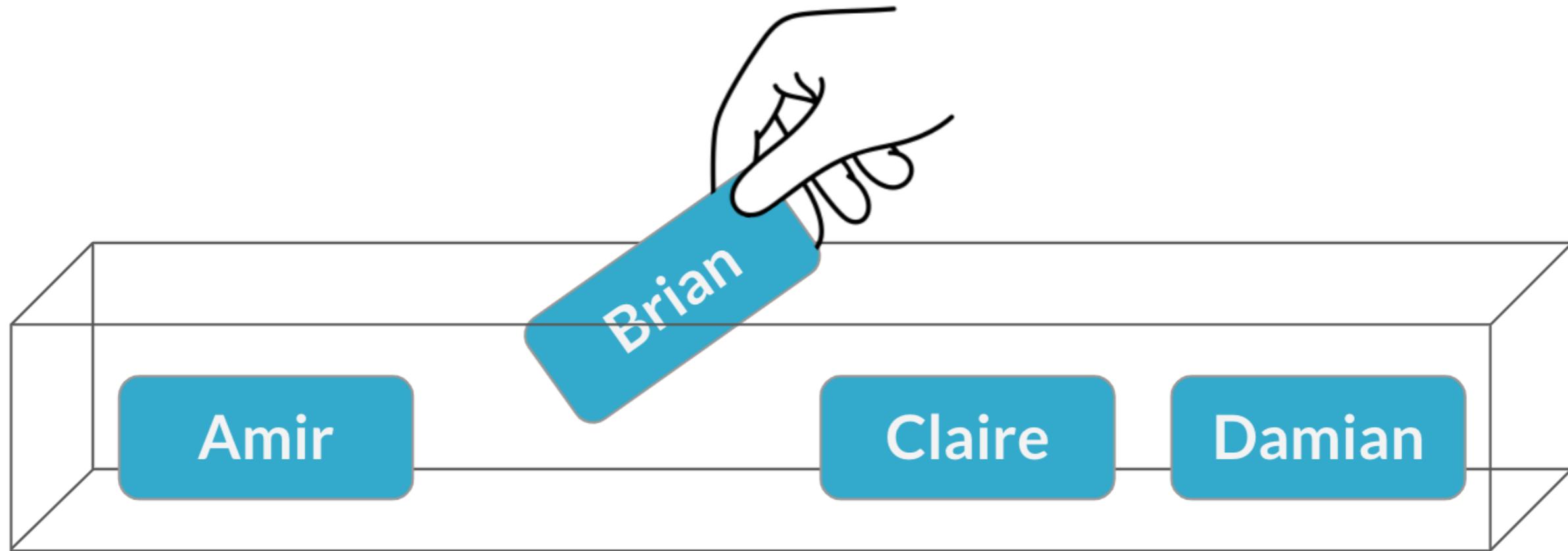


# Assigning salespeople



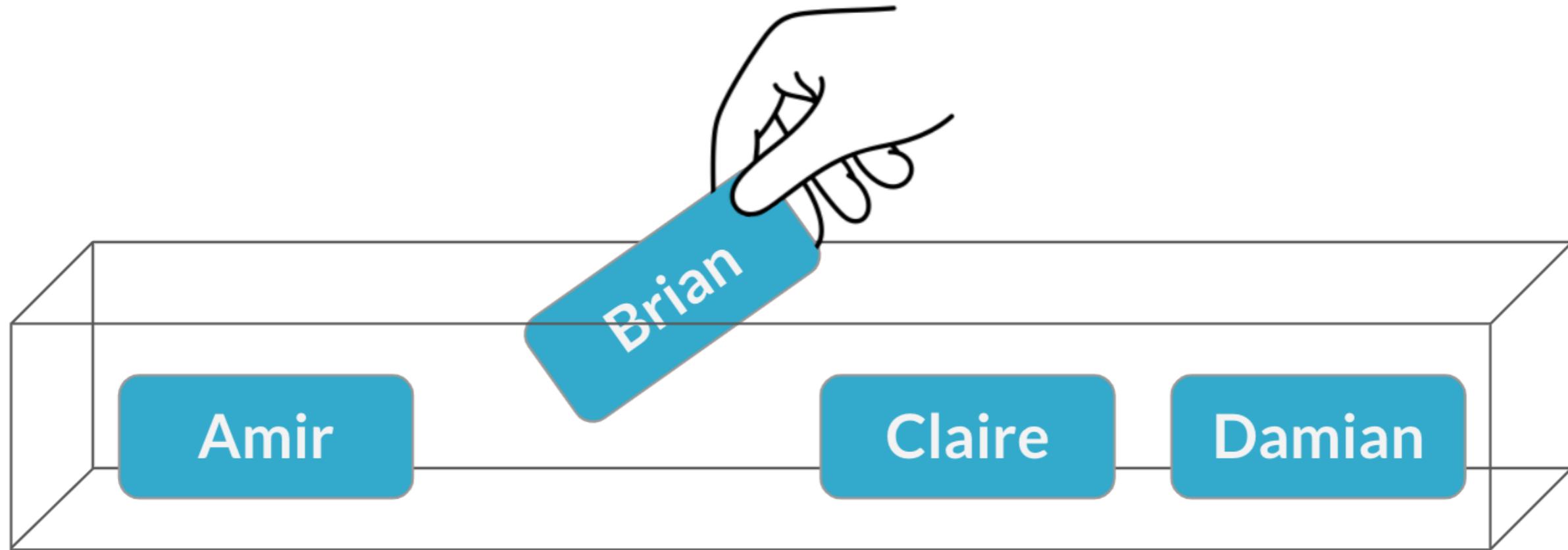
- Sampling

# Assigning salespeople

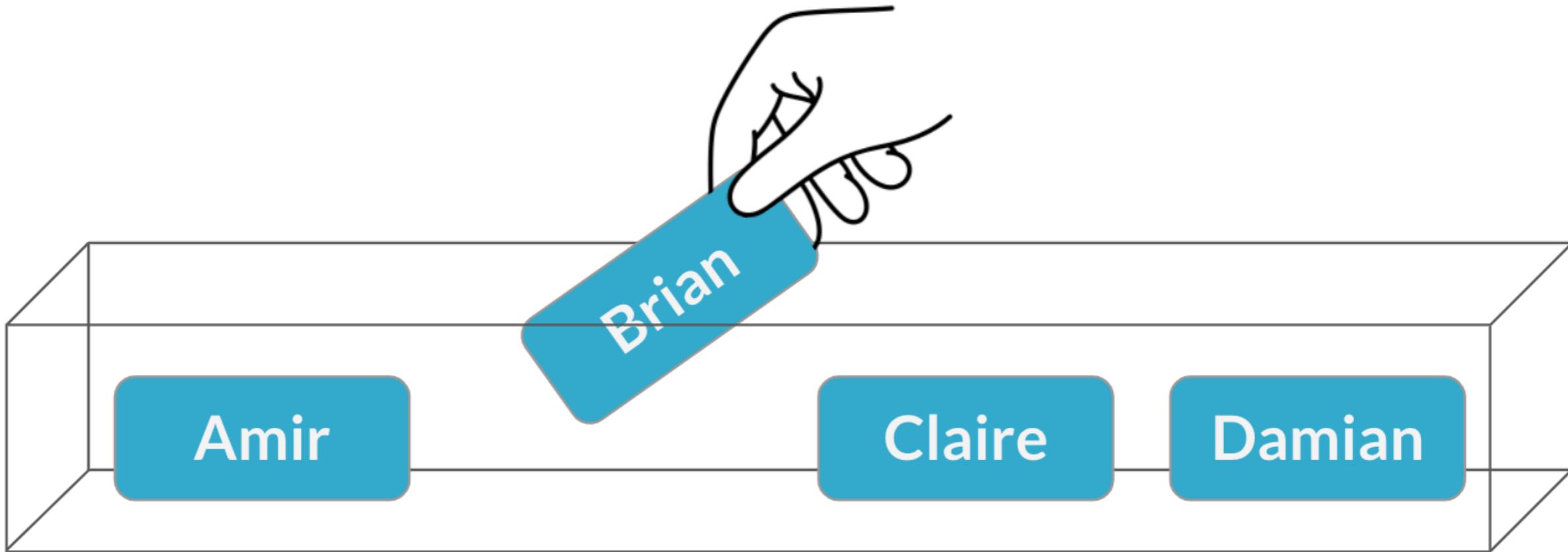


$$P(\text{Brian}) = \frac{1}{4} = 25\%$$

# Morning meeting



# Afternoon meeting



$$P(\text{Brian}) = \frac{1}{4} = 25\%$$

- Sampling with replacement

# Independent probability

*Two events are **independent** if the probability of the second event **does not** change based on the outcome of the first event.*

# Online retail sales

Order Number	Product Type	Net Quantity	Gross Sales	Discounts	Returns	Net Sales
200	Basket	13	3744.0	-316.80	0.00	3427.20
201	Basket	12	3825.0	-201.60	-288.0	3335.40
202	Basket	17	3035.0	-63.25	0.00	2971.75
203	Art & Sculpture	47	2696.8	-44.16	0.00	2652.64
204	Basket	17	2695.0	-52.50	-110.00	2532.50



<sup>1</sup> Image credit: <https://unsplash.com/@rodriguezedm>

# Probability of an order for a jewelry product

Product Type	Order Count
Basket	551
Art & Sculpture	337
Jewelry	210
Kitchen	161
Home Decor	131
...	...
<b>Total</b>	<b>1767</b>

# Probability of an order for a jewelry product

$$P(\text{Jewelry}) = \frac{\text{Order Count}(\text{Jewelry})}{\text{Sum}(\text{Total Order Count})}$$

$$P(\text{Jewelry}) = \frac{210}{1767}$$

$$P(\text{Jewelry}) = 11.88\%$$

# Probabilities for all product types

Product Type	Order Count	Probability
Basket	551	31.18%
Art & Sculpture	337	19.07%
Jewelry	210	11.88%
Kitchen	161	9.11%
Home Decor	131	7.41%
...	...	...
<b>Total</b>	<b>1767</b>	<b>100%</b>

# **Let's practice!**

## **INTRODUCTION TO STATISTICS**

# Conditional probability

INTRODUCTION TO STATISTICS



George Boorman

Curriculum Manager, DataCamp

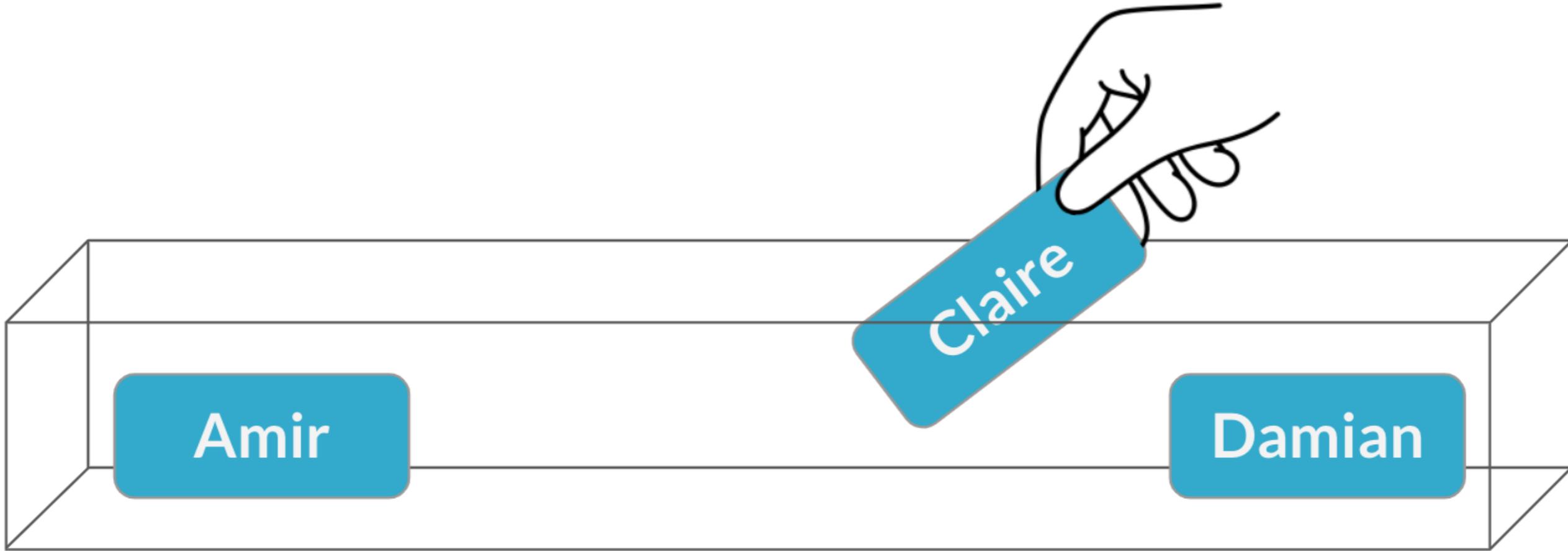
# Multiple meetings

*Sampling without replacement*



# Multiple meetings

*Sampling without replacement*



$$P(\text{Claire}) = \frac{1}{3} = 33\%$$

# Dependent events

*Probability of the second event is affected by the outcome of the first event*

**Sampling without Replacement**

*First pick*

*Second pick*

Amir

Brian

Damian

Claire

# Dependent events

*Probability of the second event is affected by the outcome of the first event*

## ***Sampling without Replacement***

*First pick*

Amir

*Second pick*

Brian

Damian

Claire

Claire

0%

# Dependent events

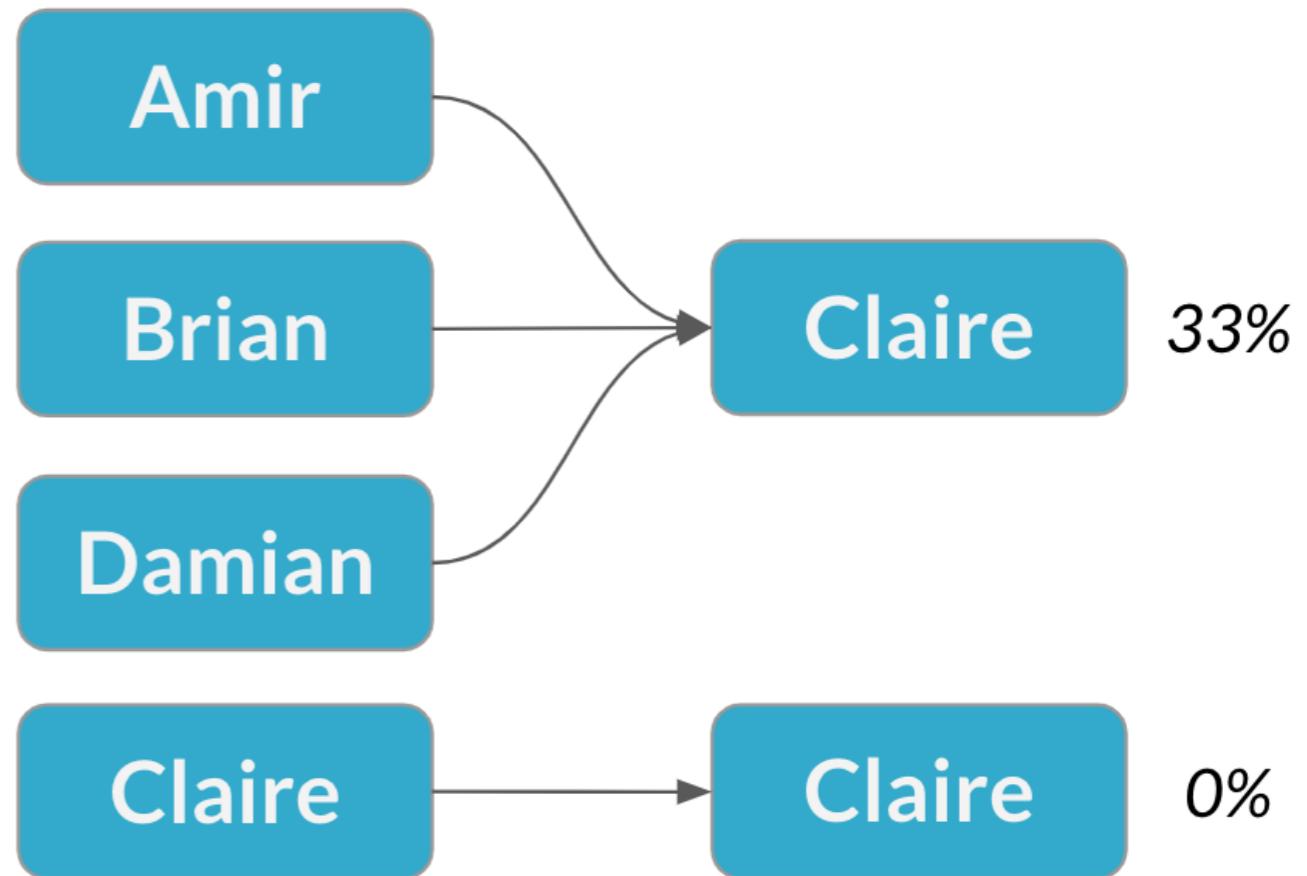
*Probability of the second event is affected by the outcome of the first event*

Sampling without replacement = each pick is dependent

## **Sampling without Replacement**

First pick

Second pick



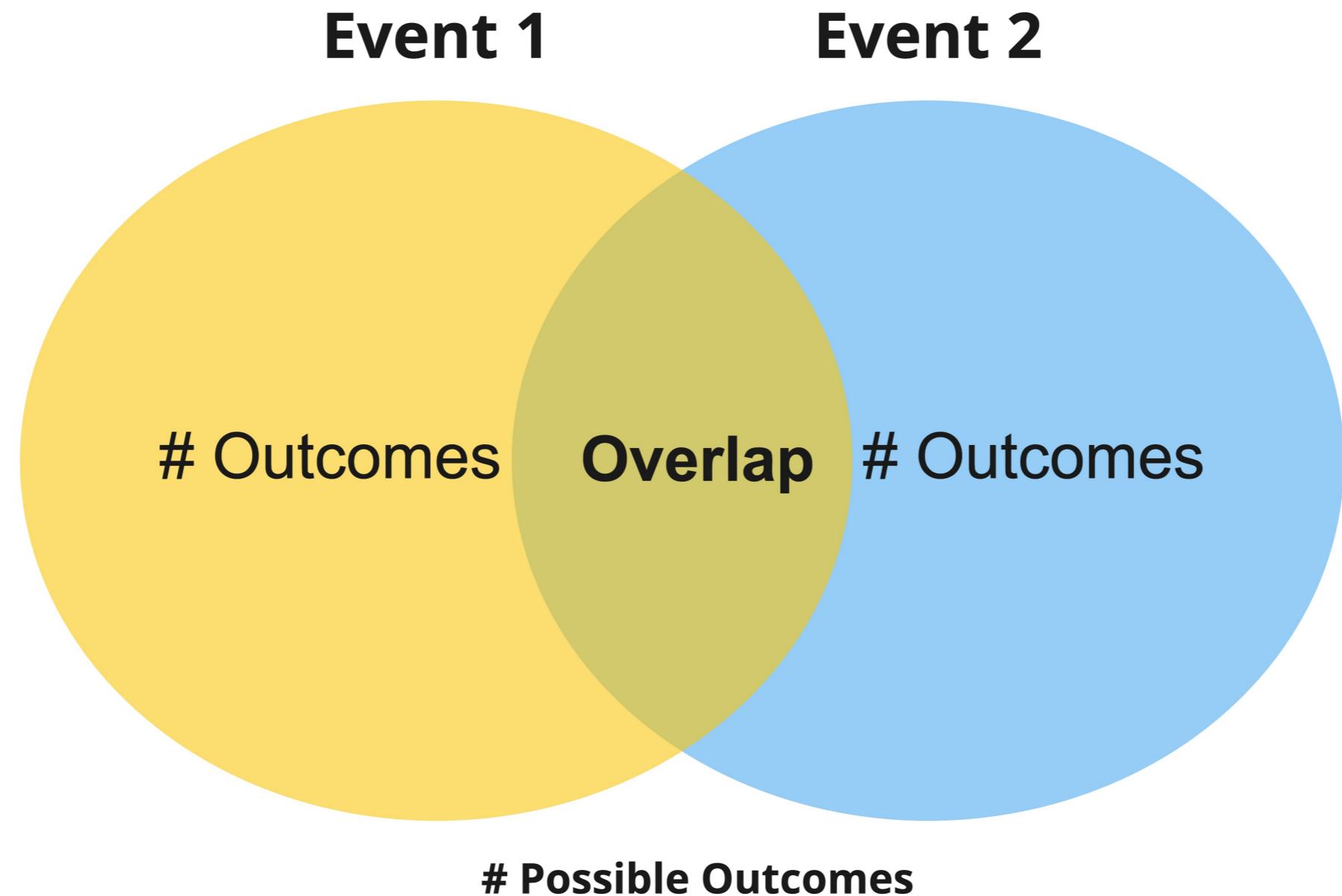
# Conditional probability

- **Conditional probability** is used to calculate the probability of dependent events
  - The probability of one event is **conditional** on the outcome of another

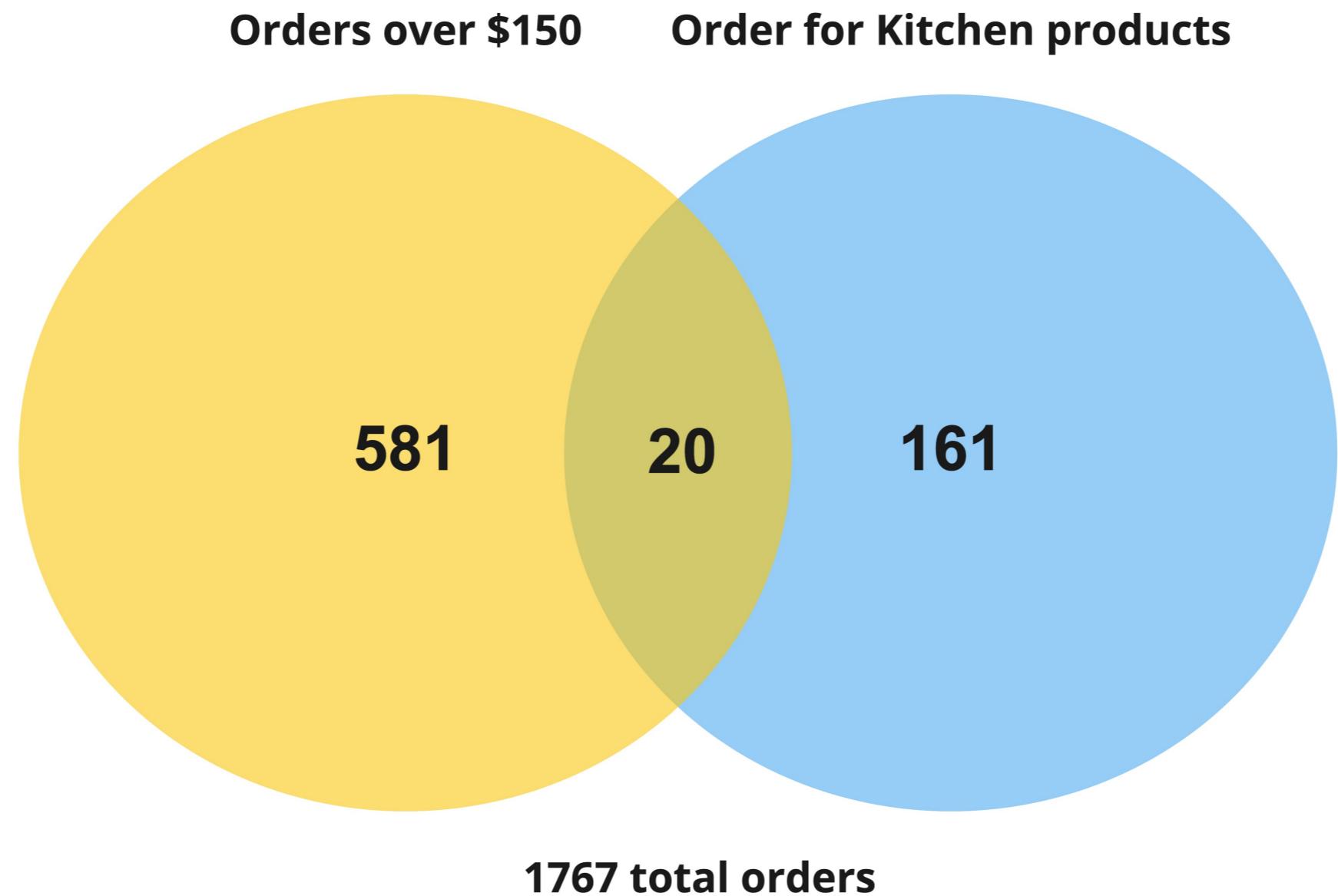


<sup>1</sup> Image credit: <https://unsplash.com/@pixeldan>

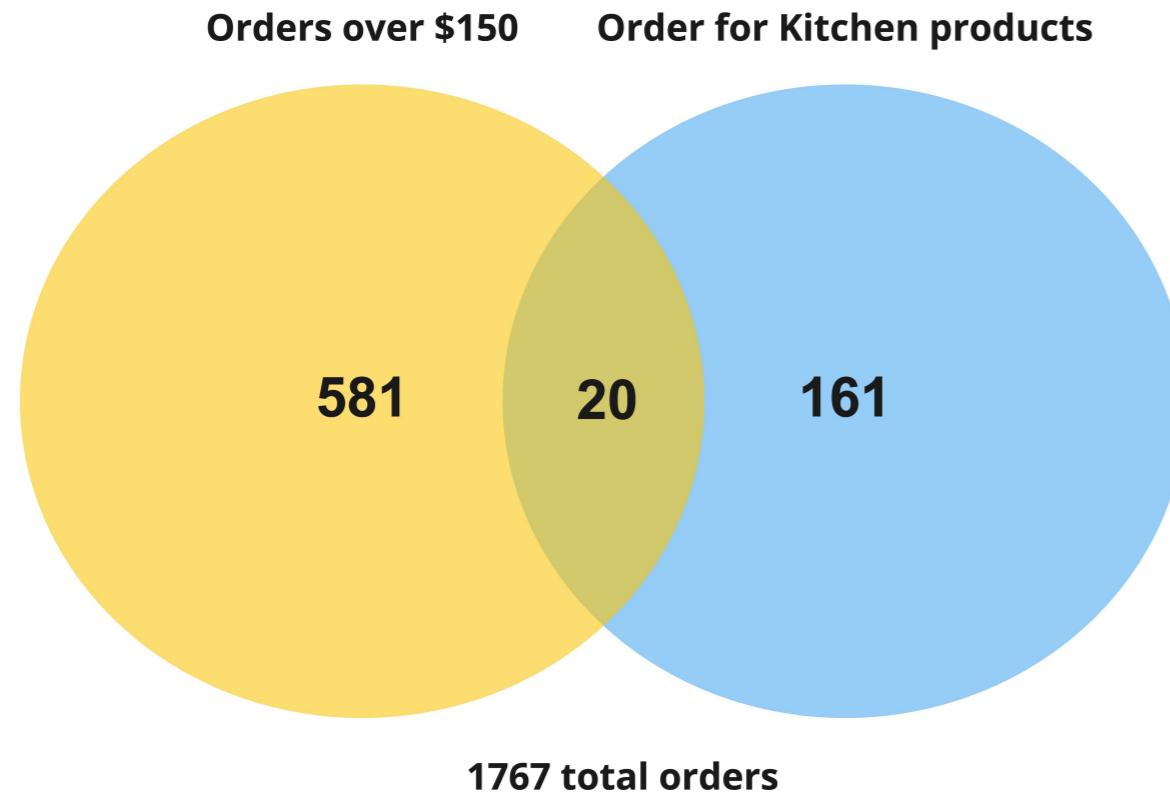
# Venn diagrams



# Kitchen sales over \$150



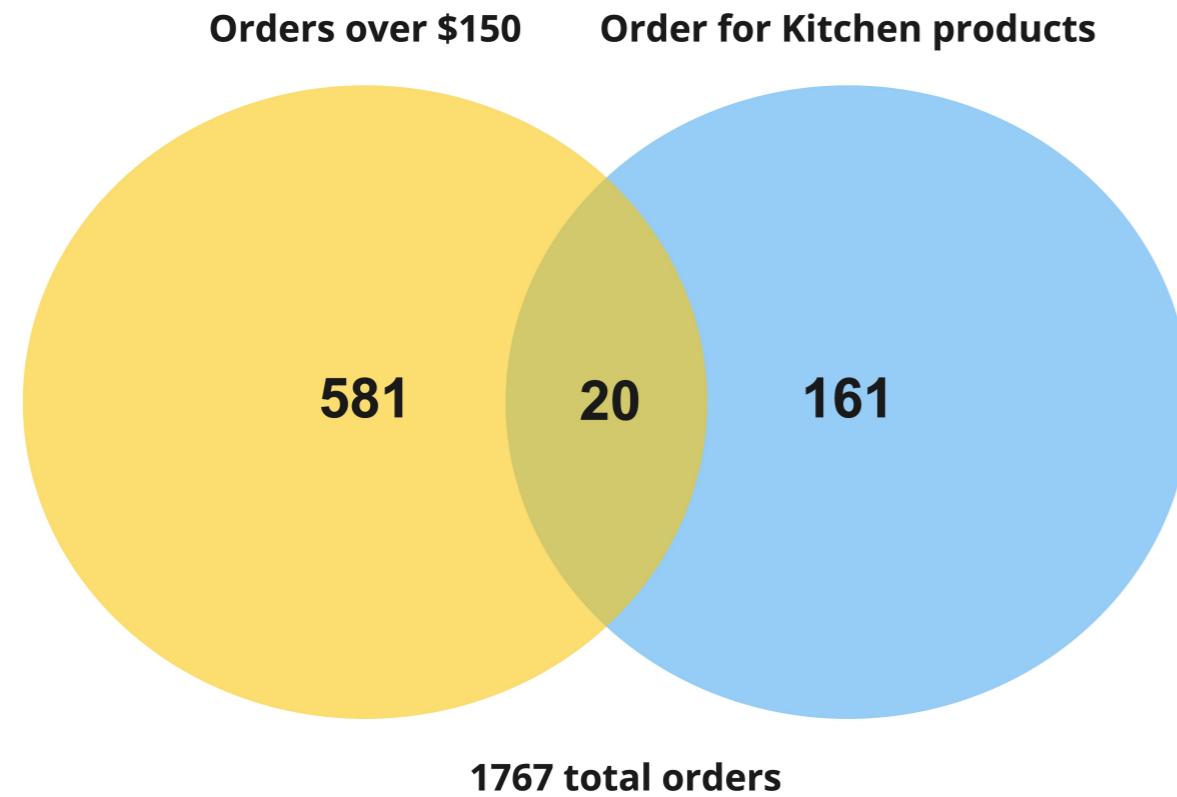
# Kitchen sales over \$150



$$P(\text{Order} > 150 | \text{Kitchen}) = \frac{20}{\frac{1767}{181}}$$

$$P(\text{Order} > 150 | \text{Kitchen}) = \frac{20}{181}$$

# The order of events matters



$$P(Kitchen | Order > 150) = \frac{20}{\frac{1767}{601}}$$

$$P(Kitchen | Order > 150) = \frac{20}{601}$$

# Conditional probability formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- $P(A|B)$  → Probability of event A, given event B
- $P(A \cap B)$  → Probability of event A **and** event B
  - Divided by the probability of event B →  $P(B)$

# **Let's practice!**

## **INTRODUCTION TO STATISTICS**

# Discrete distributions

INTRODUCTION TO STATISTICS

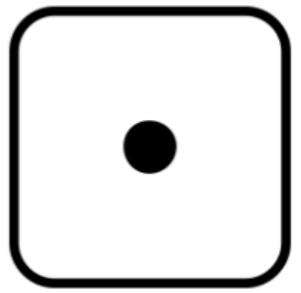


George Boorman

Curriculum Manager, DataCamp



# Rolling the dice



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$

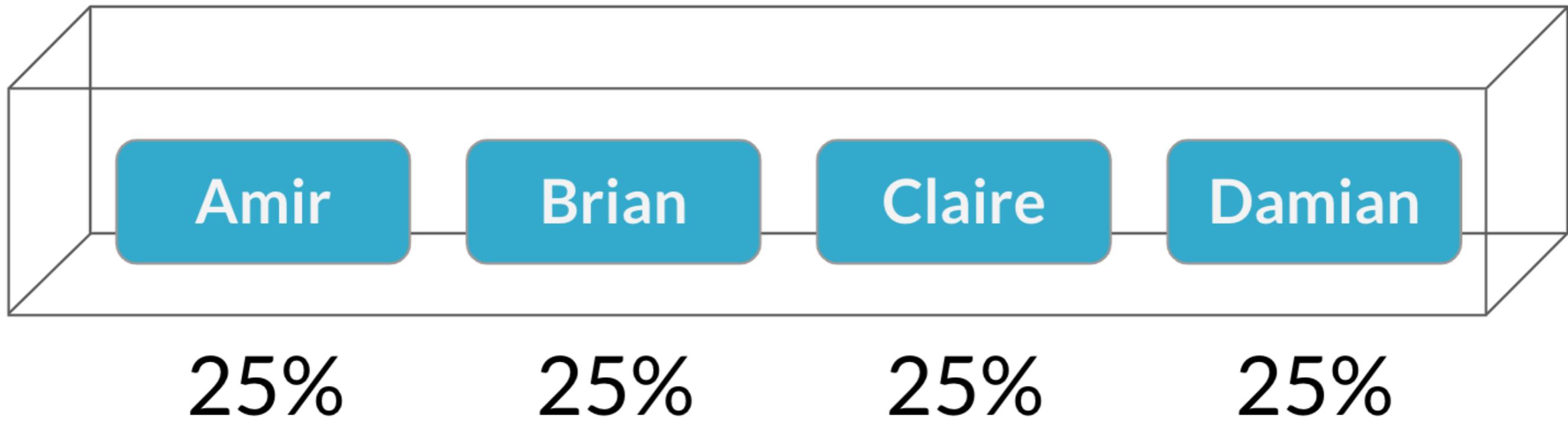


$\frac{1}{6}$



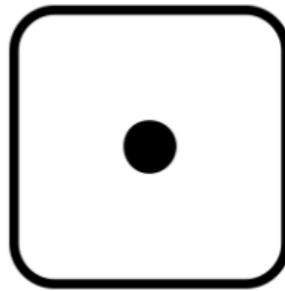
$\frac{1}{6}$

# Choosing salespeople

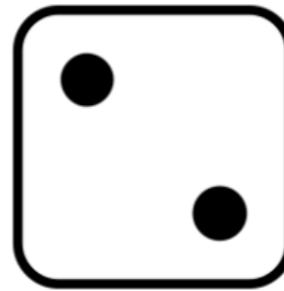


# Probability distribution

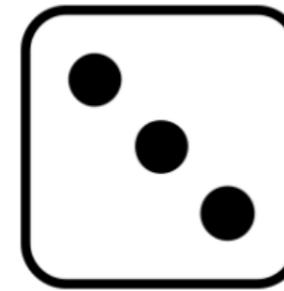
*Describes the probability of each possible outcome in a scenario*



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$

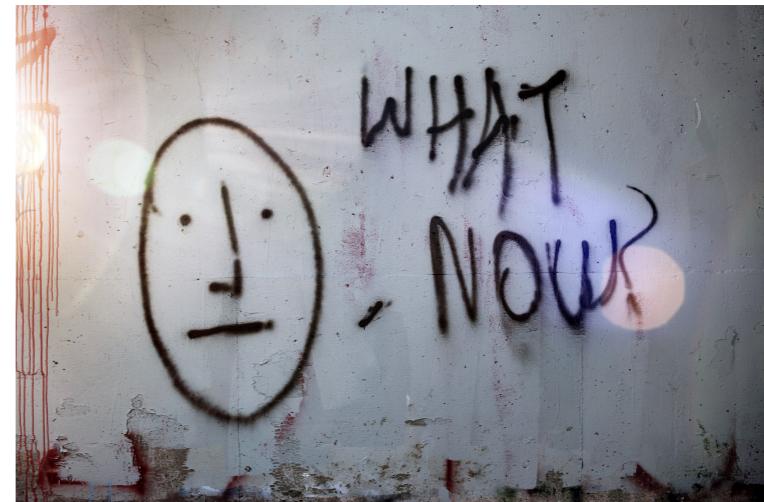
**Expected value:** The *mean* of a probability distribution

Expected value of a fair die roll =

$$(1 \times \frac{1}{6}) + (2 \times \frac{1}{6}) + (3 \times \frac{1}{6}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) = 3.5$$

# Why are probability distributions important?

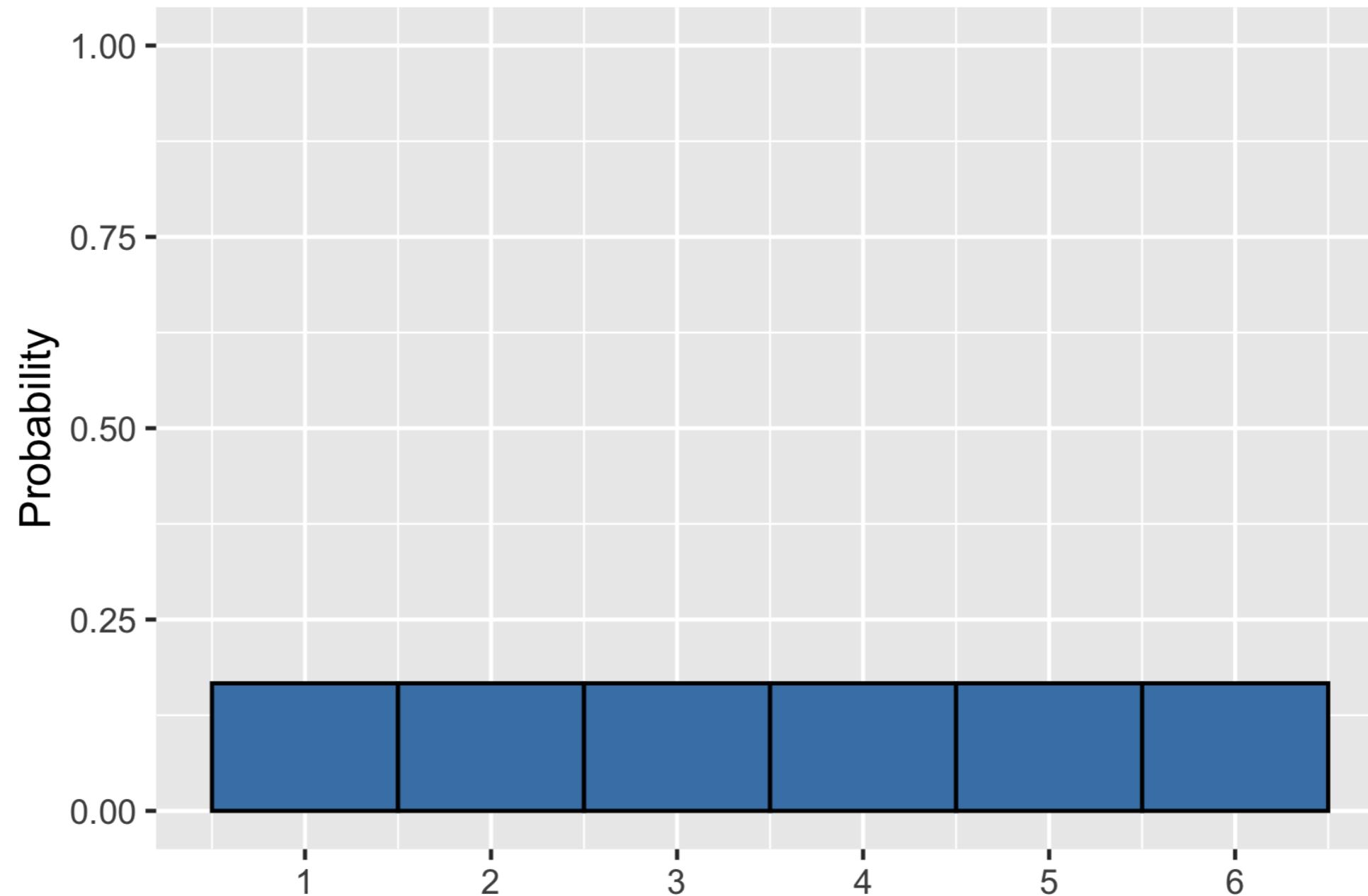
- Help us to quantify risk and inform decision making



- Used extensively in hypothesis testing
  - Probability that the results occurred by chance

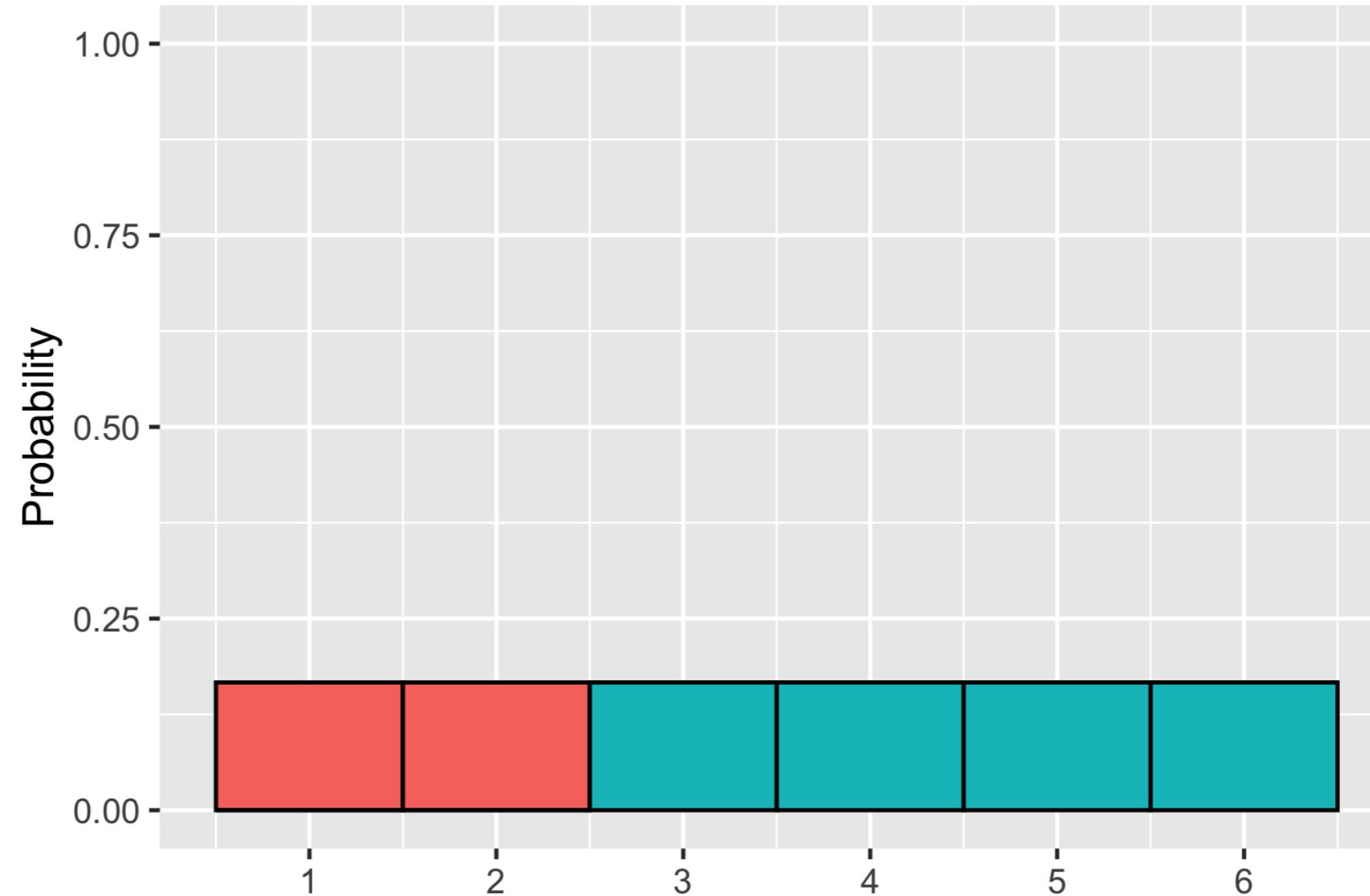
<sup>1</sup> Image credit: <https://unsplash.com/@timmosholder>

# Visualizing a probability distribution



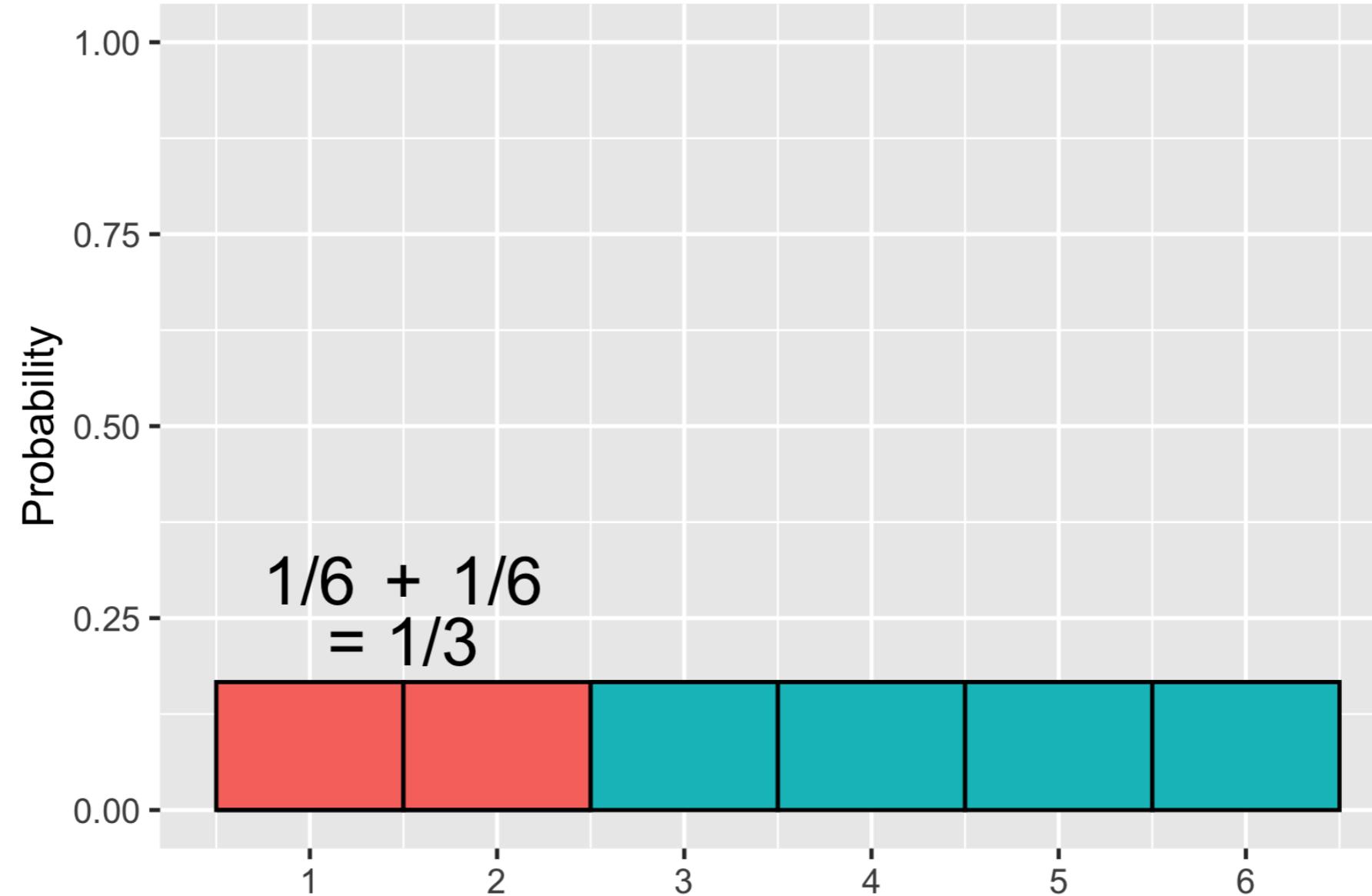
# Probability = area

$$P(\text{die roll}) \leq 2 = ?$$

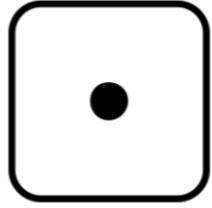


# Probability = area

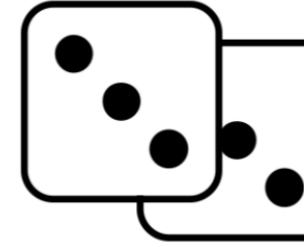
$$P(\text{die roll}) \leq 2 = 1/3$$



# Uneven die



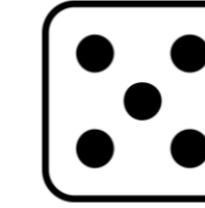
$\frac{1}{6}$



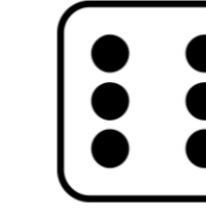
$\frac{1}{3}$



$\frac{1}{6}$



$\frac{1}{6}$

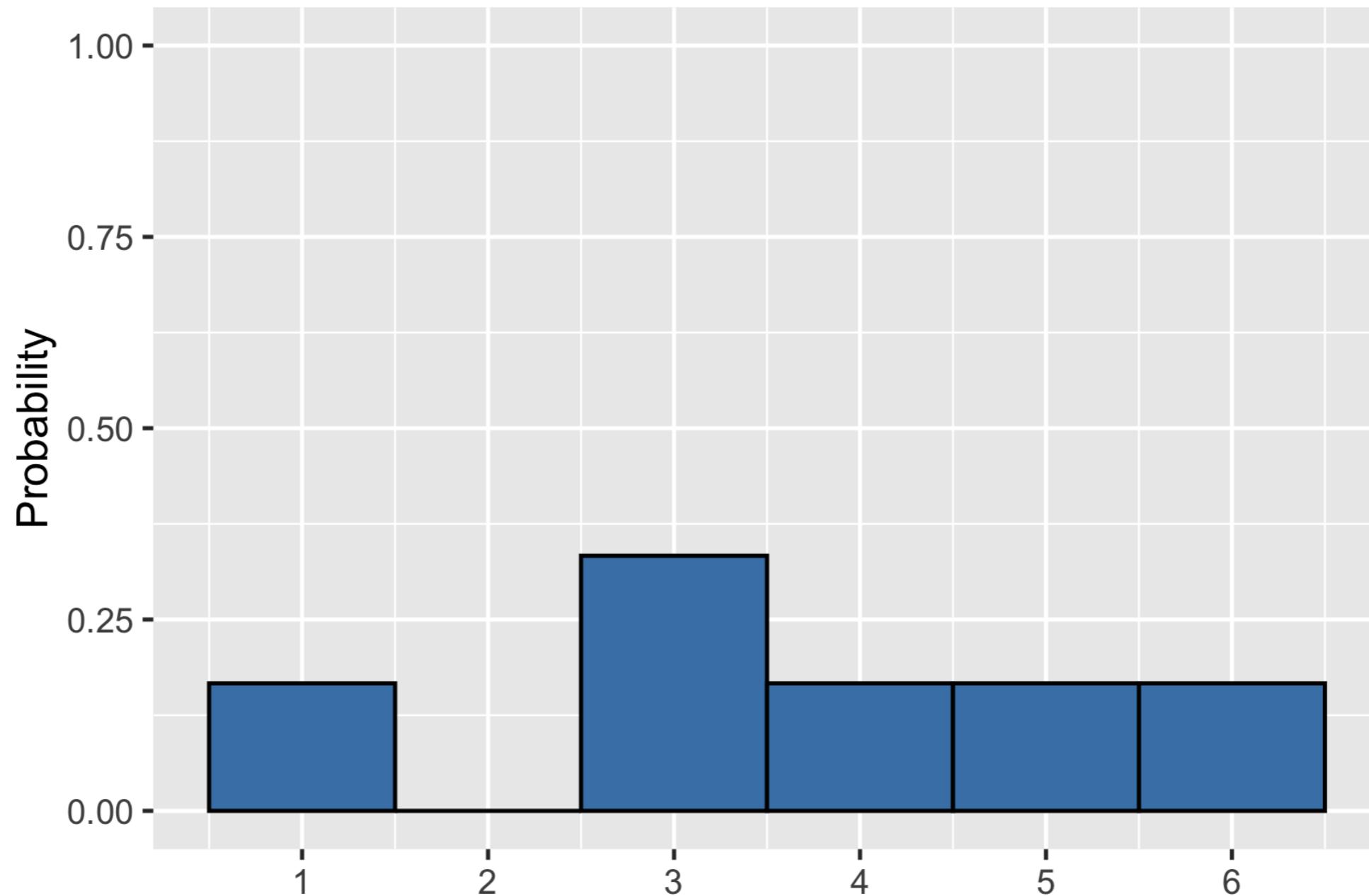


$\frac{1}{6}$

Expected value of uneven die roll =

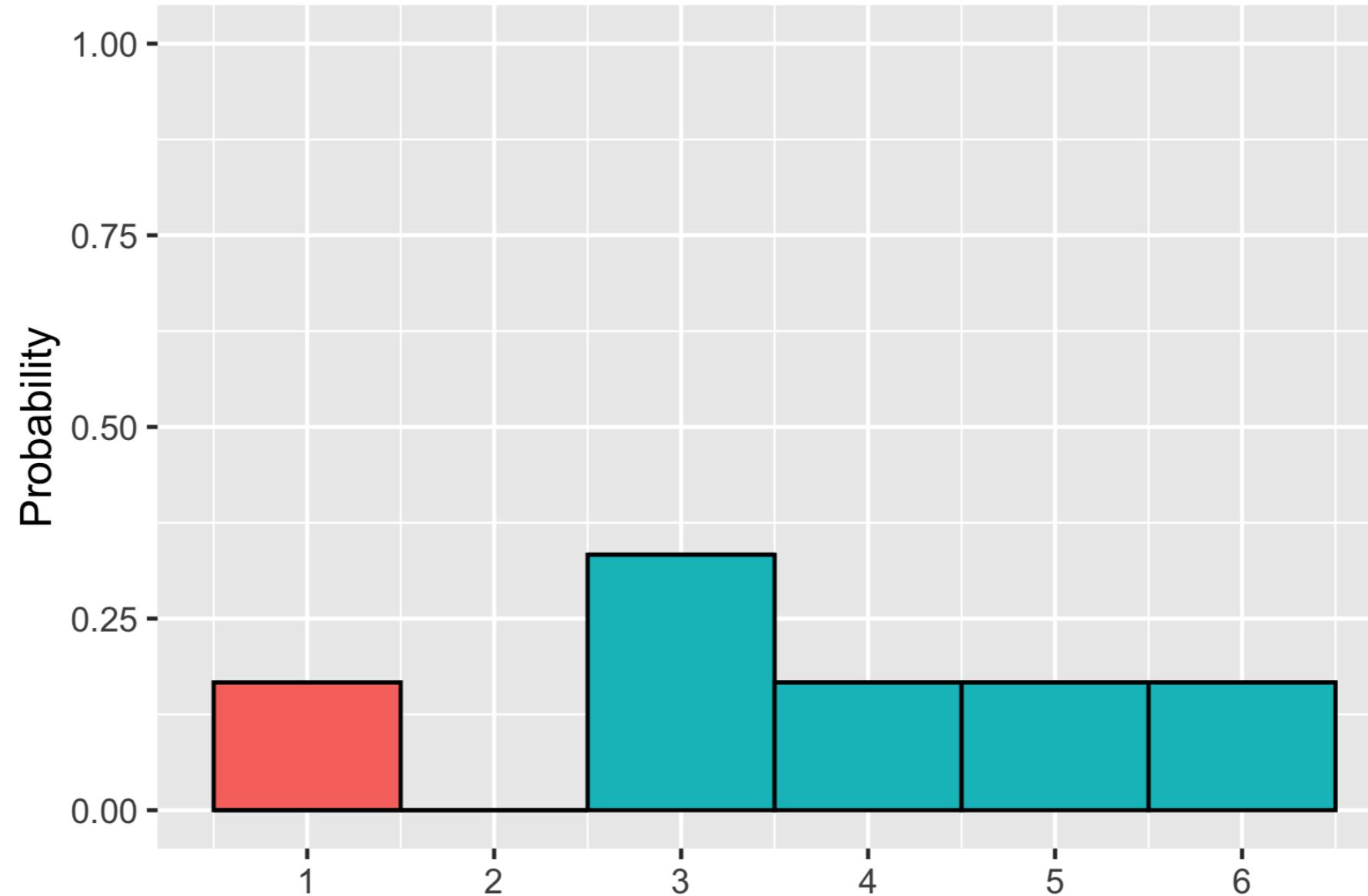
$$(1 \times \frac{1}{6}) + (2 \times 0) + (3 \times \frac{1}{3}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) = 3.67$$

# Visualizing uneven probabilities



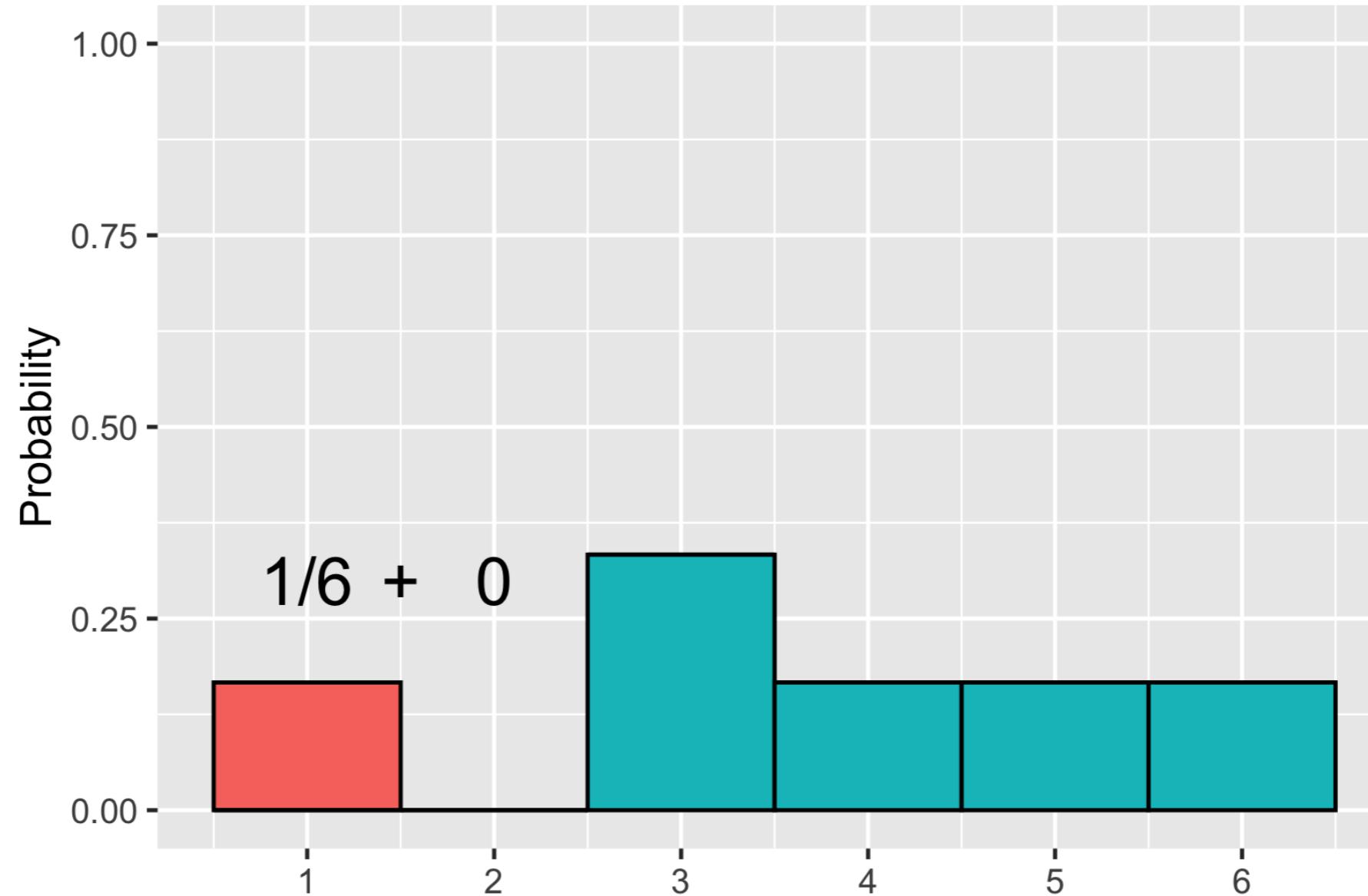
# Adding areas

$P(\text{uneven die roll}) \leq 2 = ?$



# Adding areas

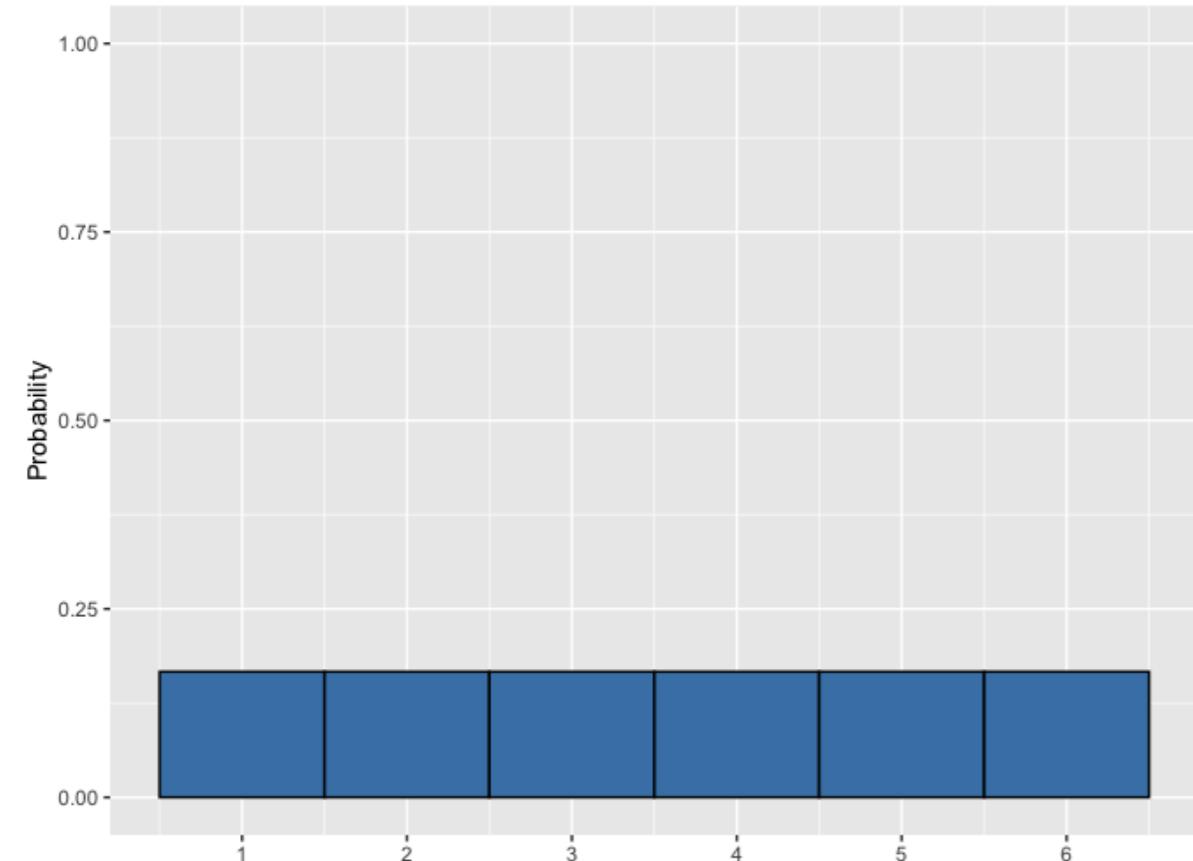
$$P(\text{uneven die roll}) \leq 2 = 1/6$$



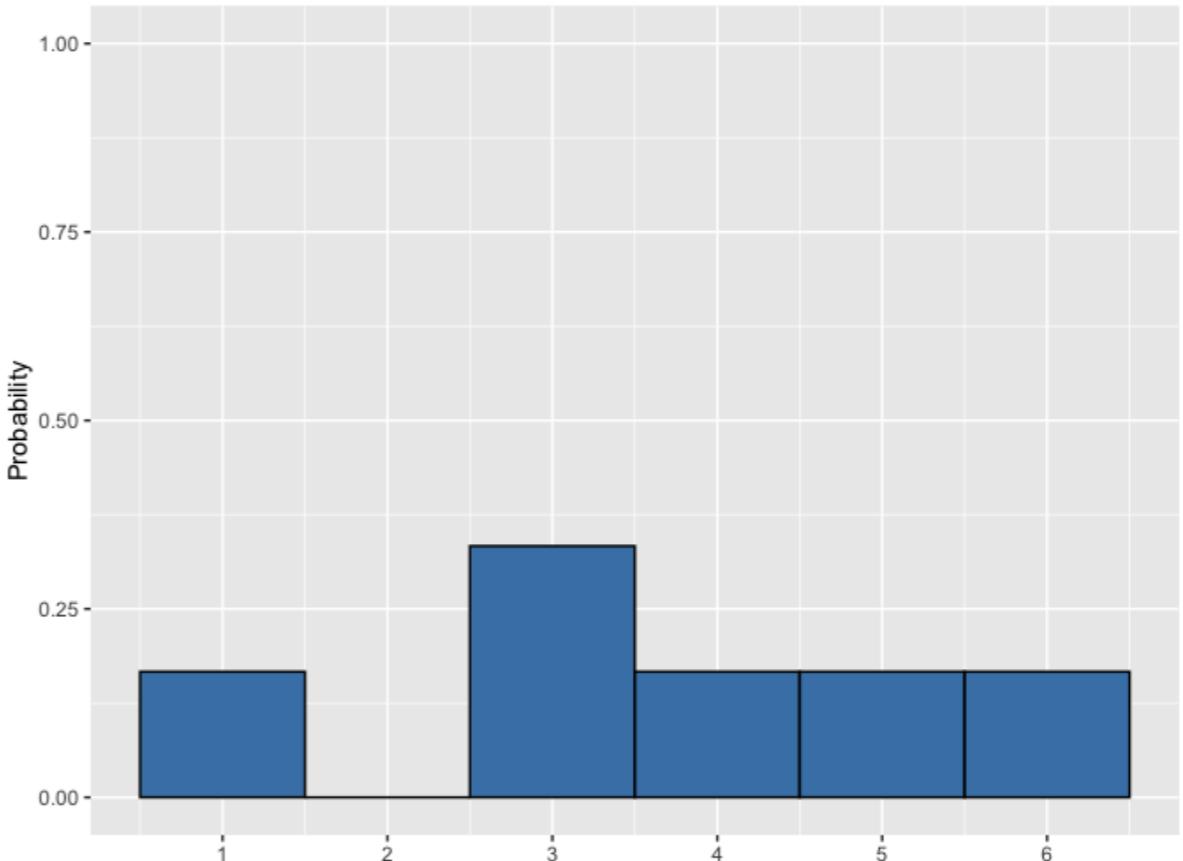
# Discrete probability distributions

*Describe probabilities for discrete outcomes*

Fair die



Uneven die



*Discrete uniform distribution*

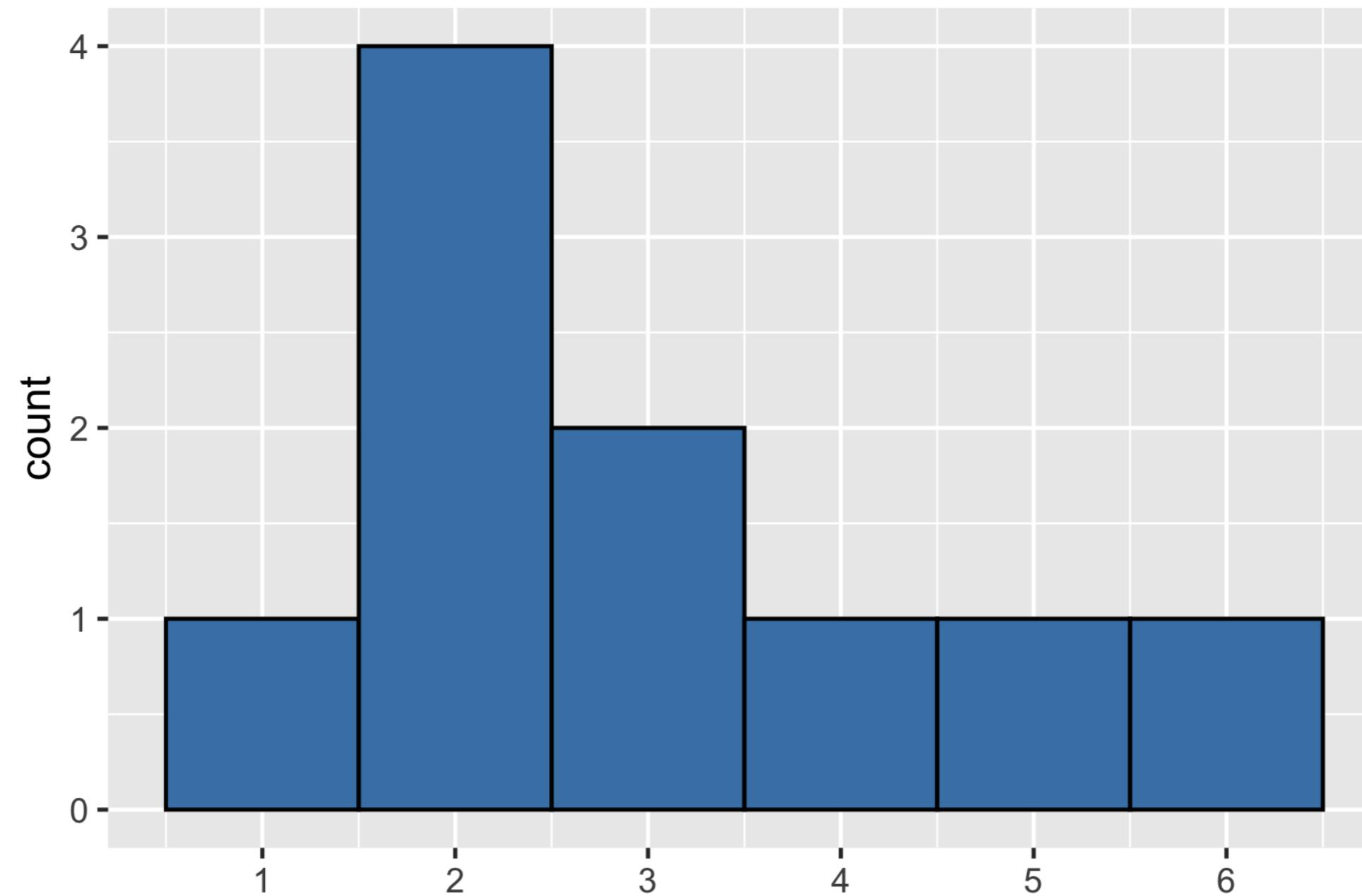
# Sampling from a discrete distribution

Roll	Result
1	1
2	2
3	3
4	4
5	5
6	6

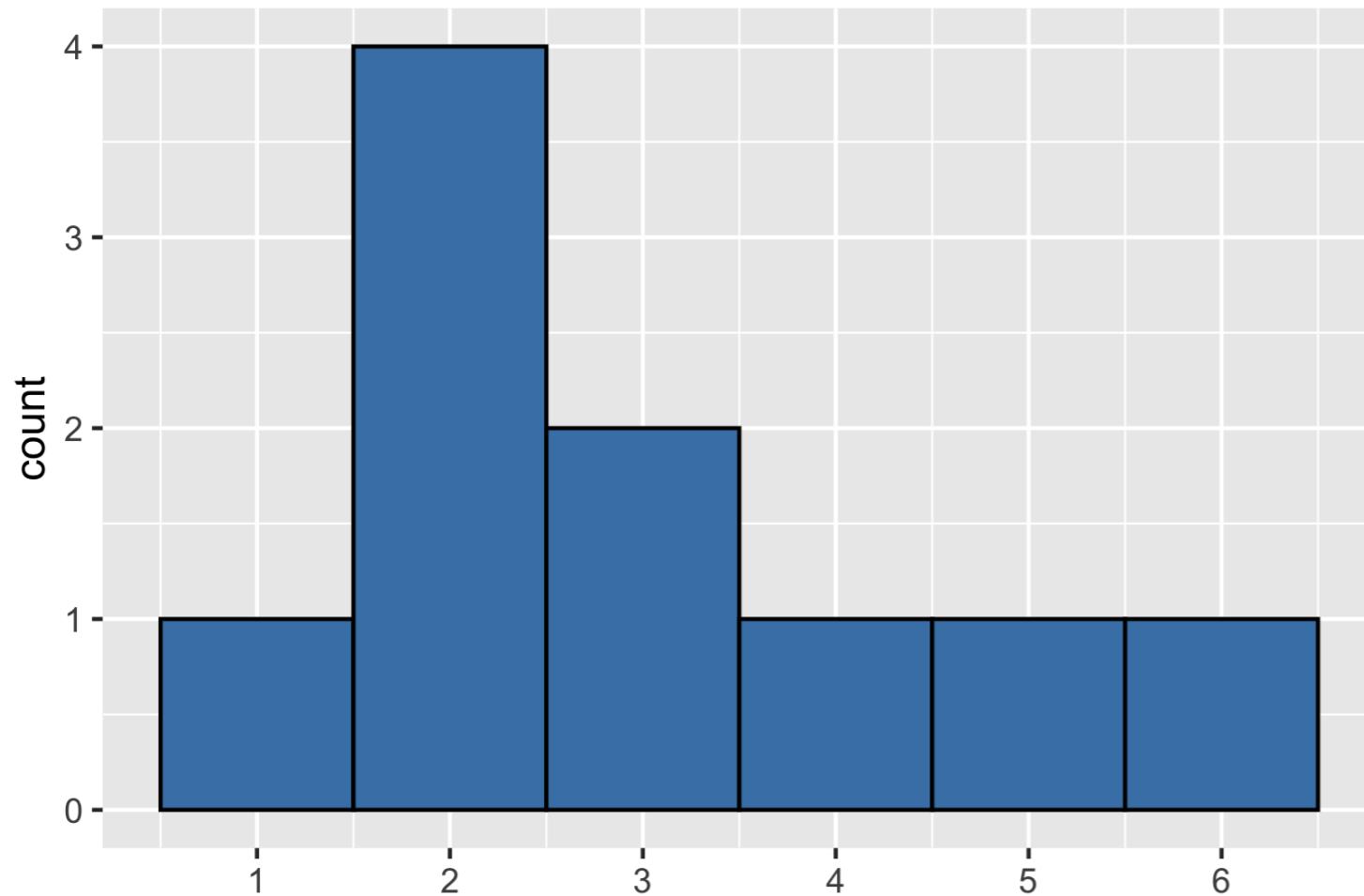
$$\text{Mean} = 3.5$$

Roll	Result
1	3
2	1
3	2
4	4
5	6
6	3
7	2
8	2
9	2
10	5

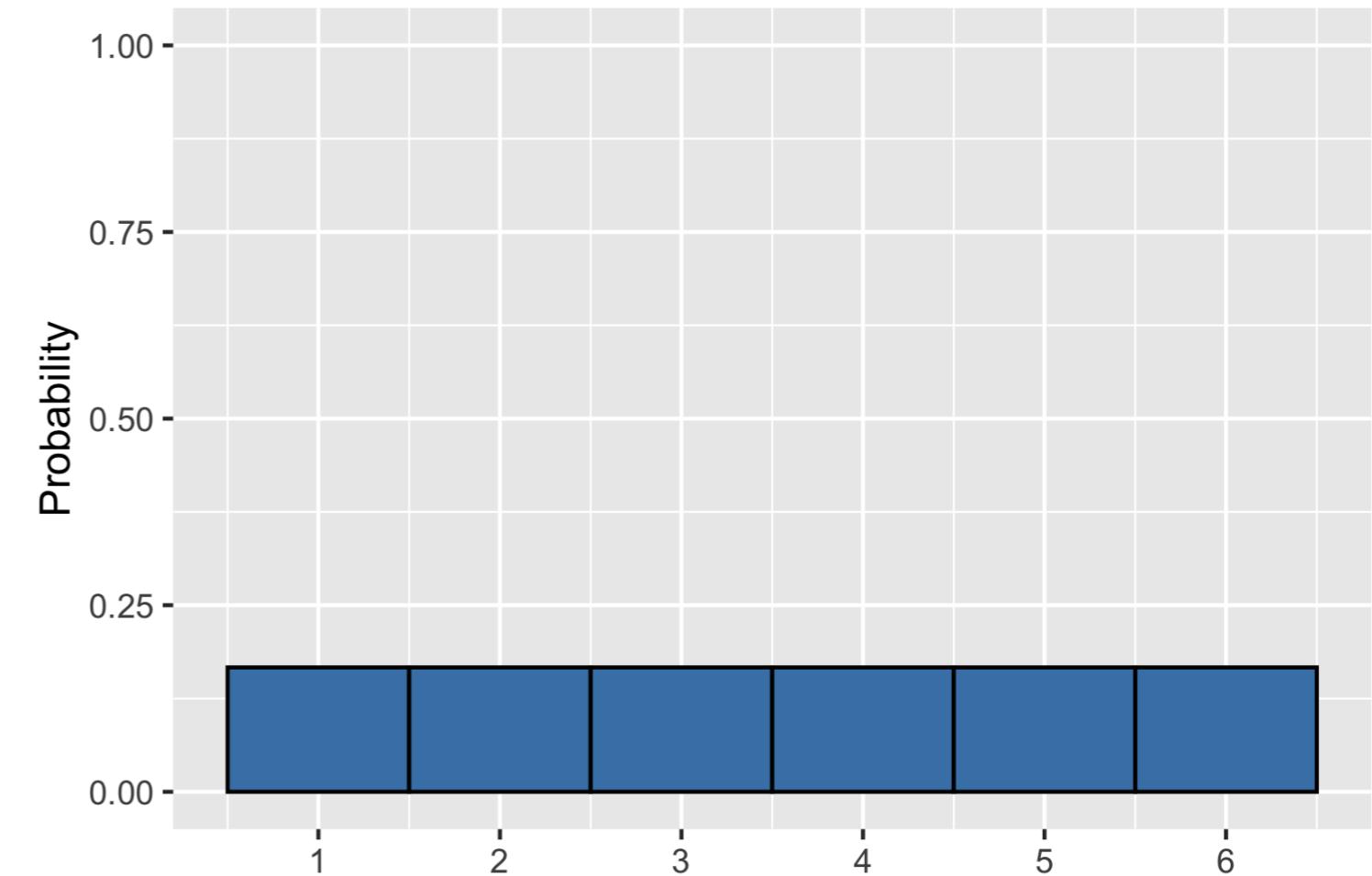
# Visualizing a sample



# Sample distribution vs theoretical distribution



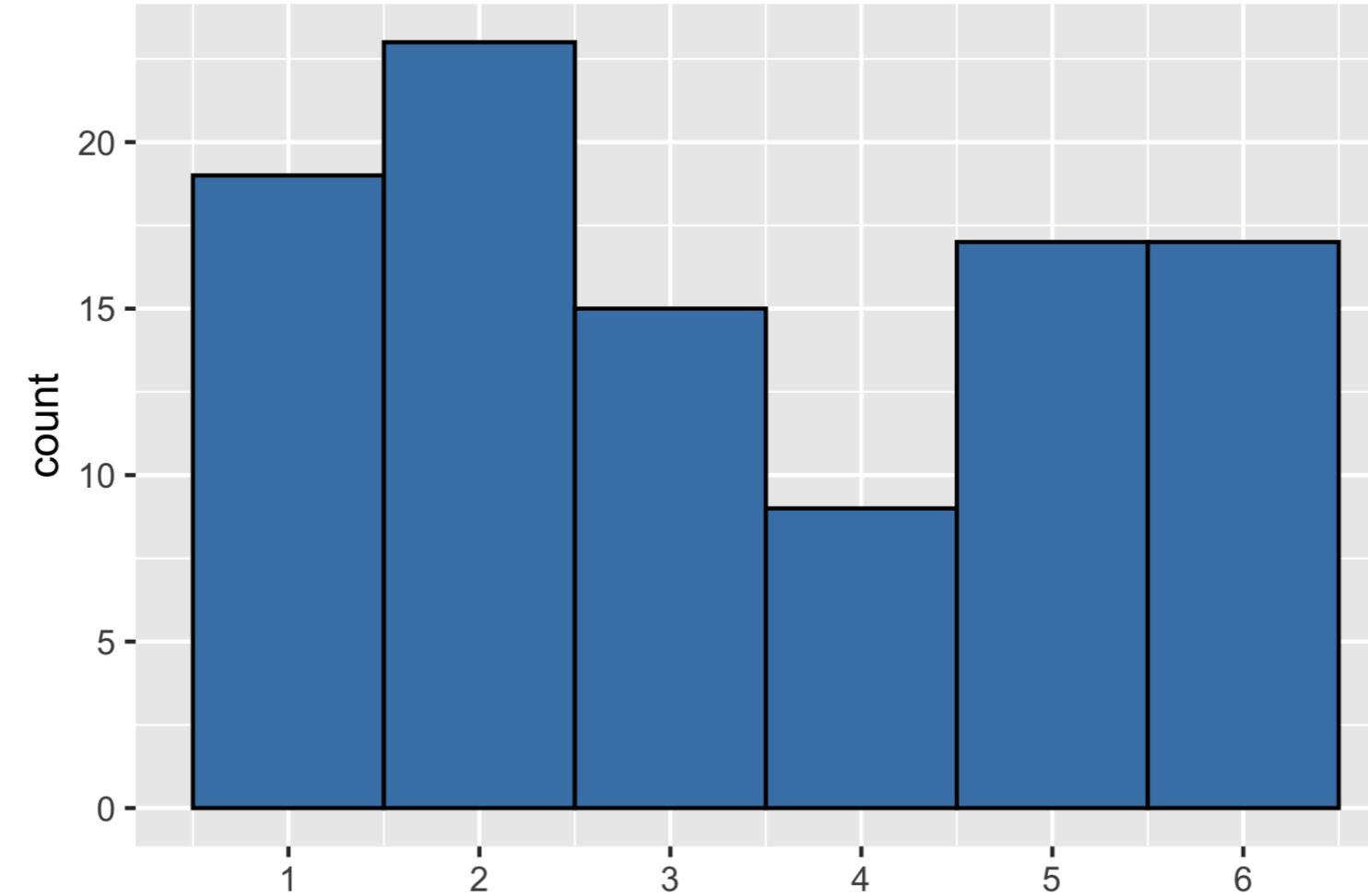
*Mean = 3.0*



*Mean = 3.5*

# A bigger sample

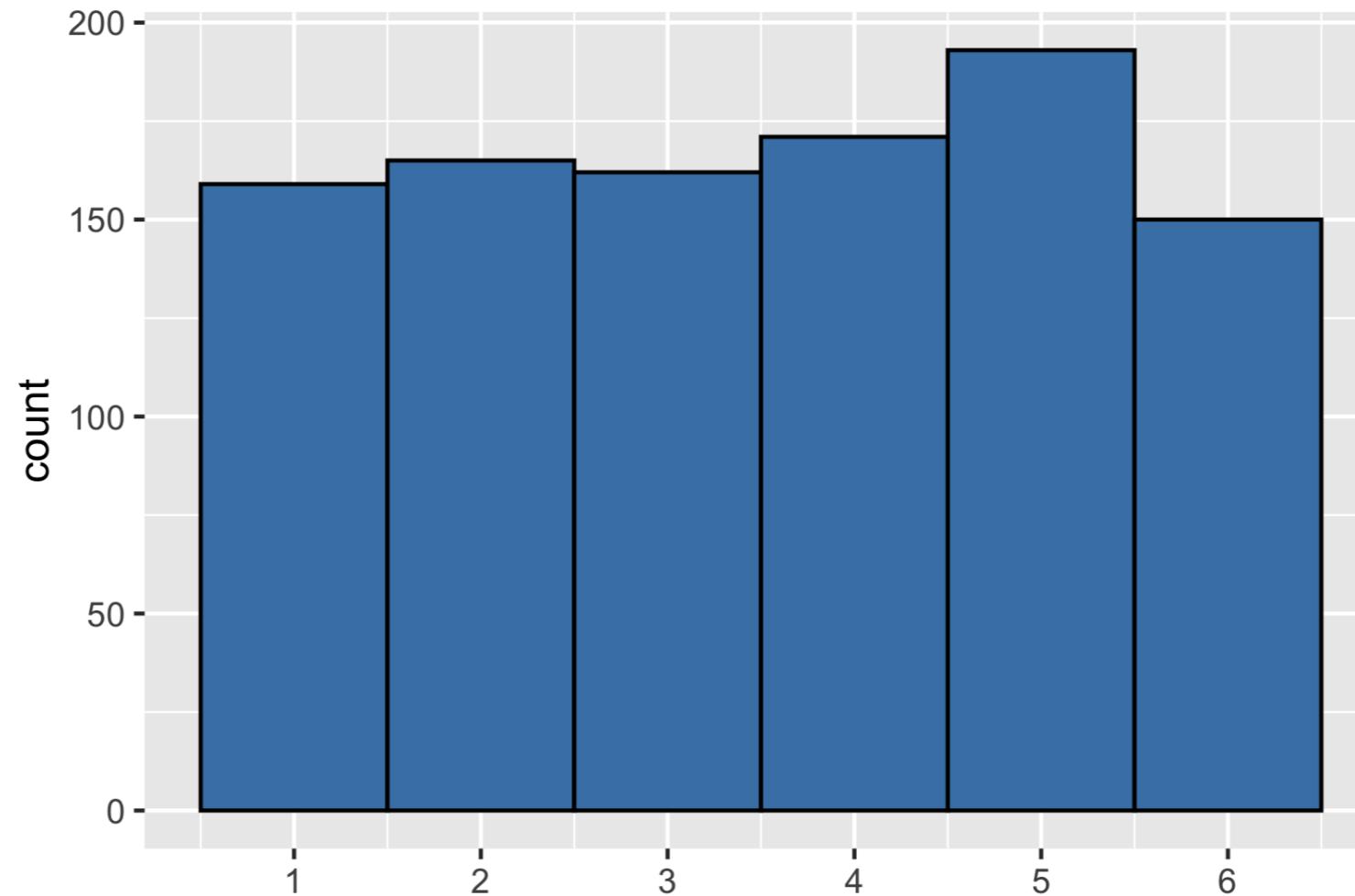
Sample of 100 rolls



$$\text{Mean} = 3.33$$

# An even bigger sample

Sample of 1000 rolls



$$\text{Mean} = 3.52$$

# Law of large numbers

*As the size of your sample increases, the sample mean will approach the expected value.*

Sample size	Mean
10	3.00
100	3.33
1000	3.52

# **Let's practice!**

## **INTRODUCTION TO STATISTICS**

# Continuous distributions

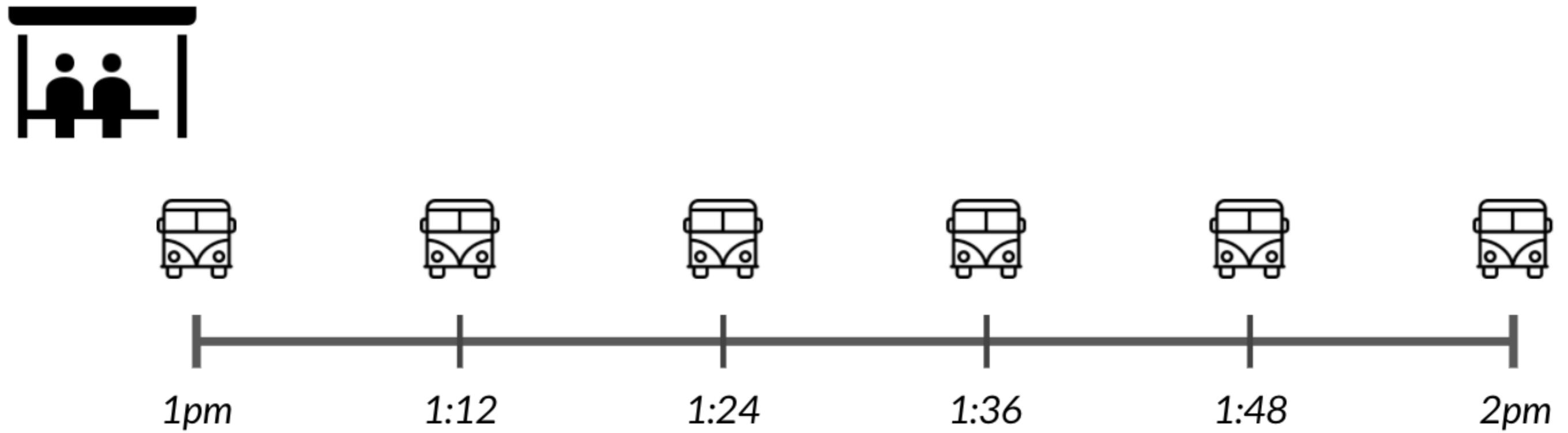
INTRODUCTION TO STATISTICS



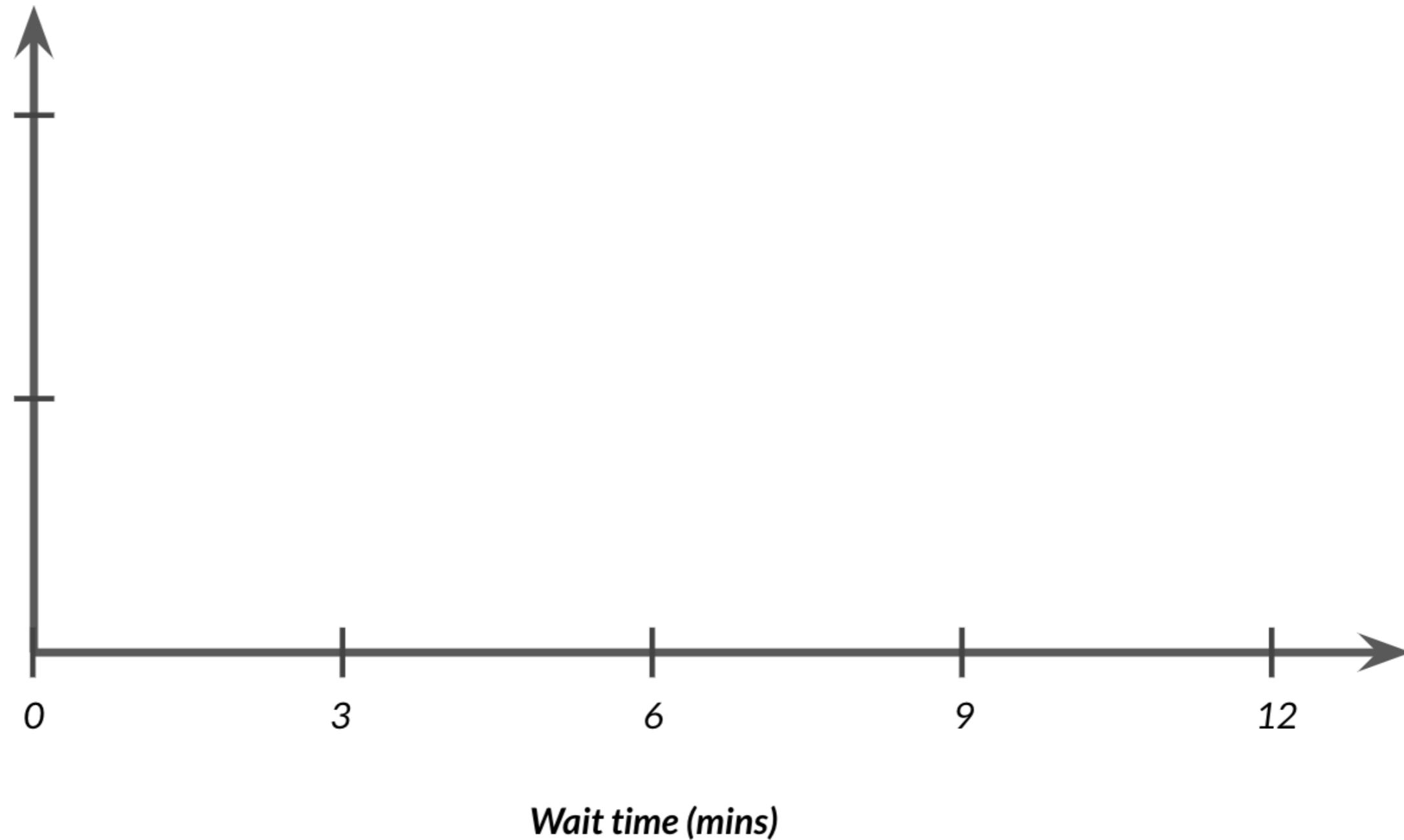
George Boorman

Curriculum Manager, DataCamp

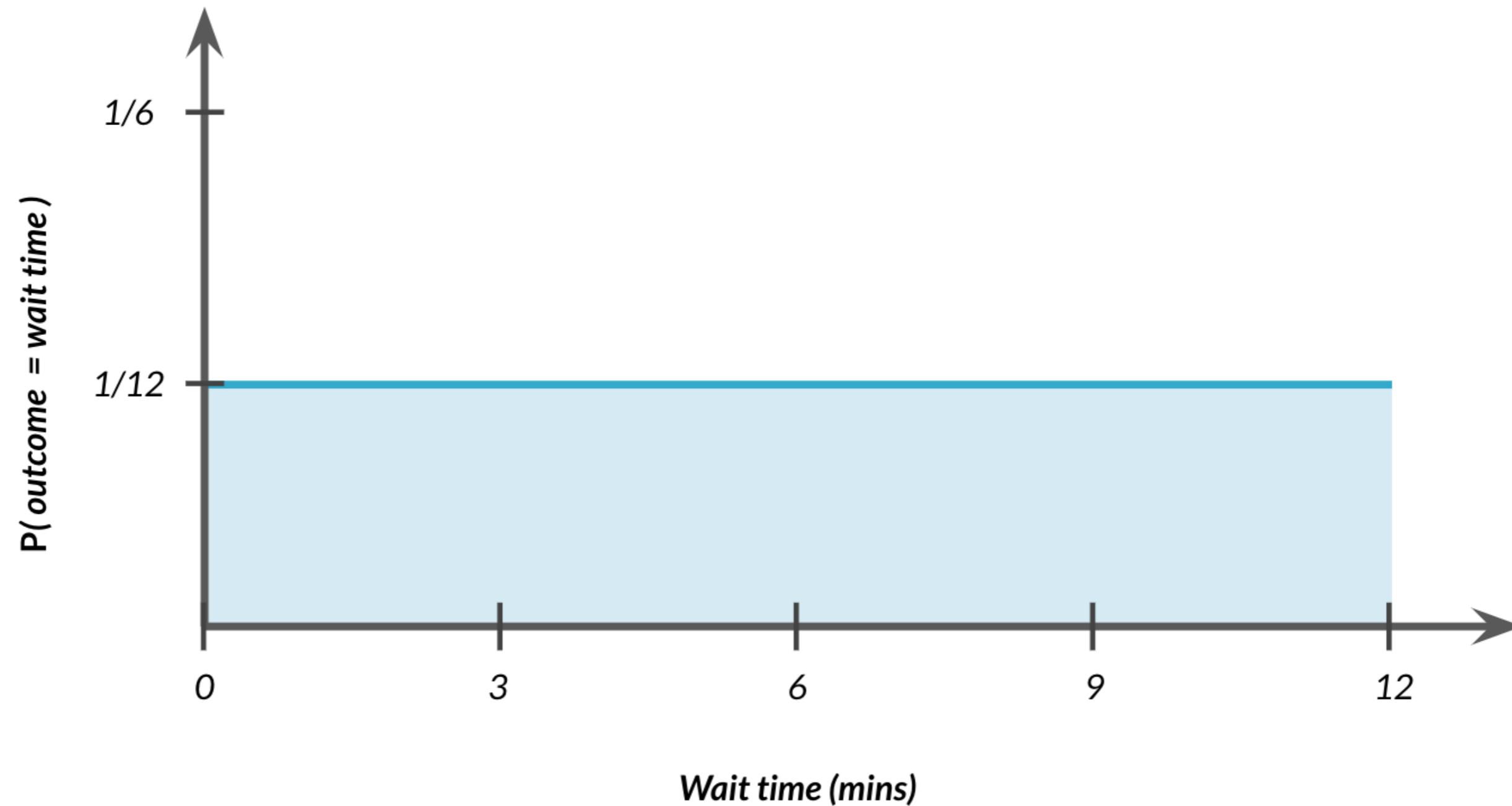
# Waiting for the bus



# Continuous uniform distribution

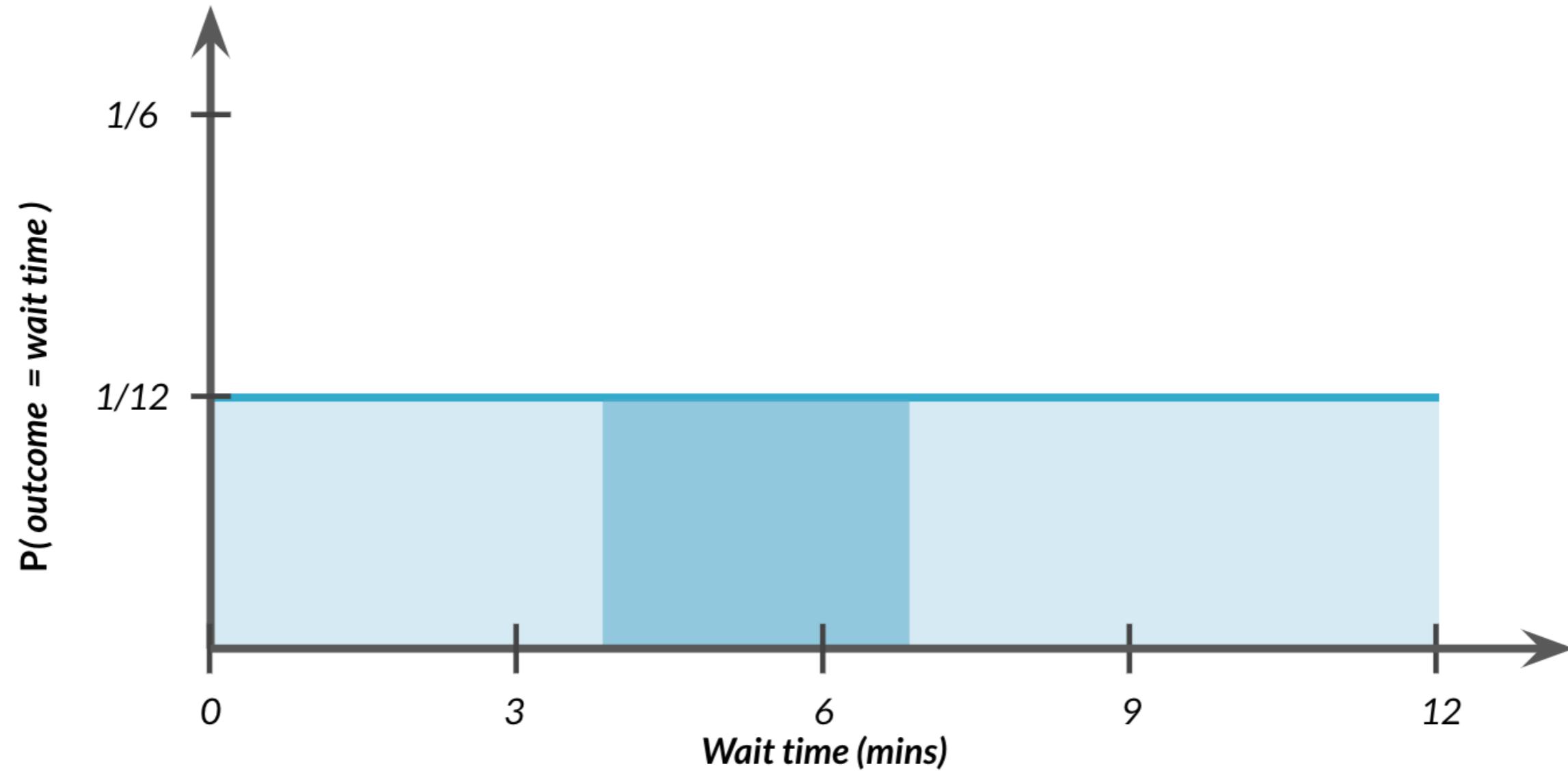


# Continuous uniform distribution



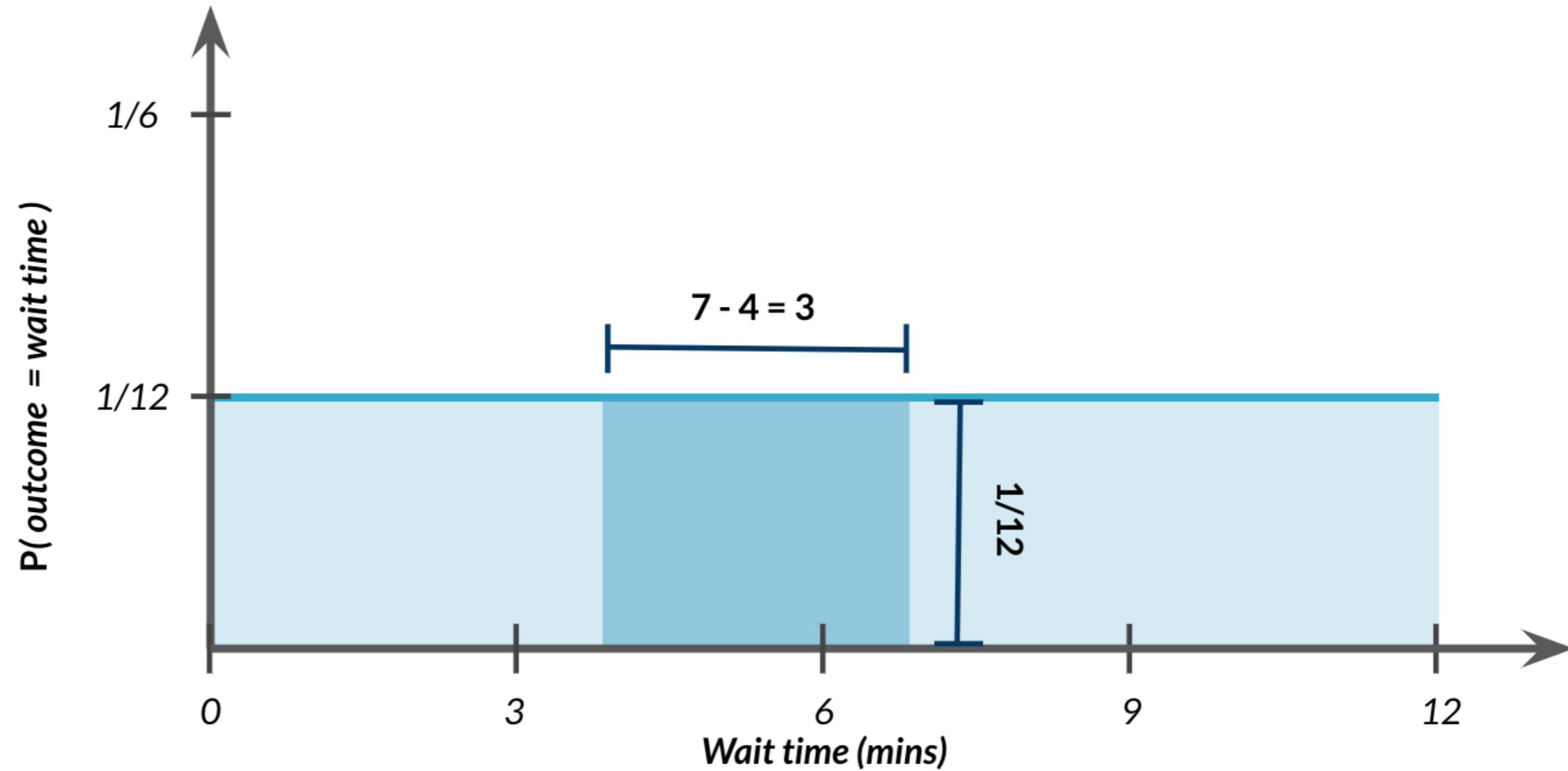
# Probability still = area

$$P(4 \leq \text{wait time} \leq 7) = ?$$



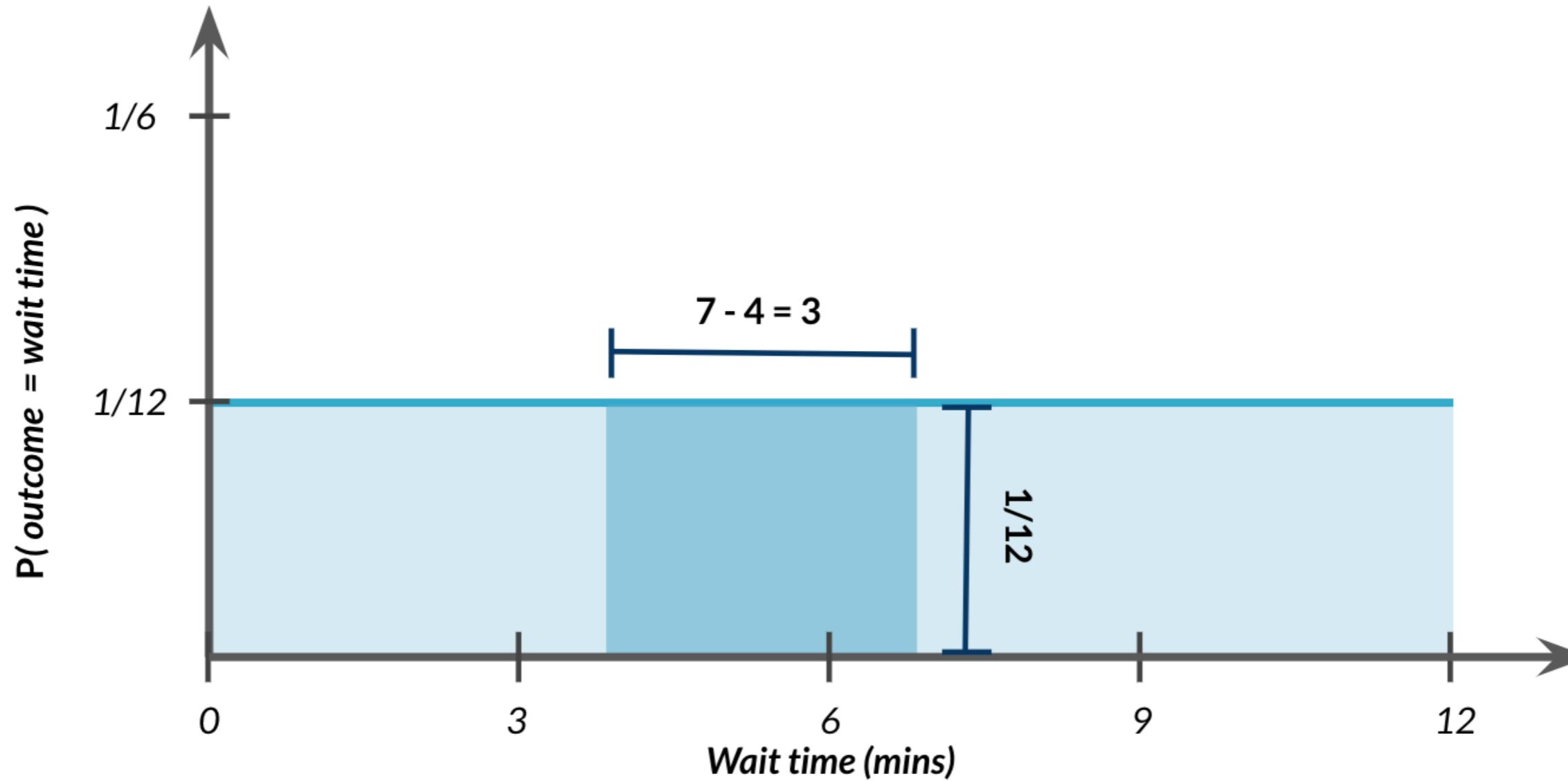
# Probability still = area

$$P(4 \leq \text{wait time} \leq 7) = ?$$



# Probability still = area

$$P(4 \leq \text{wait time} \leq 7) = 3 \times 1/12 = 3/12$$

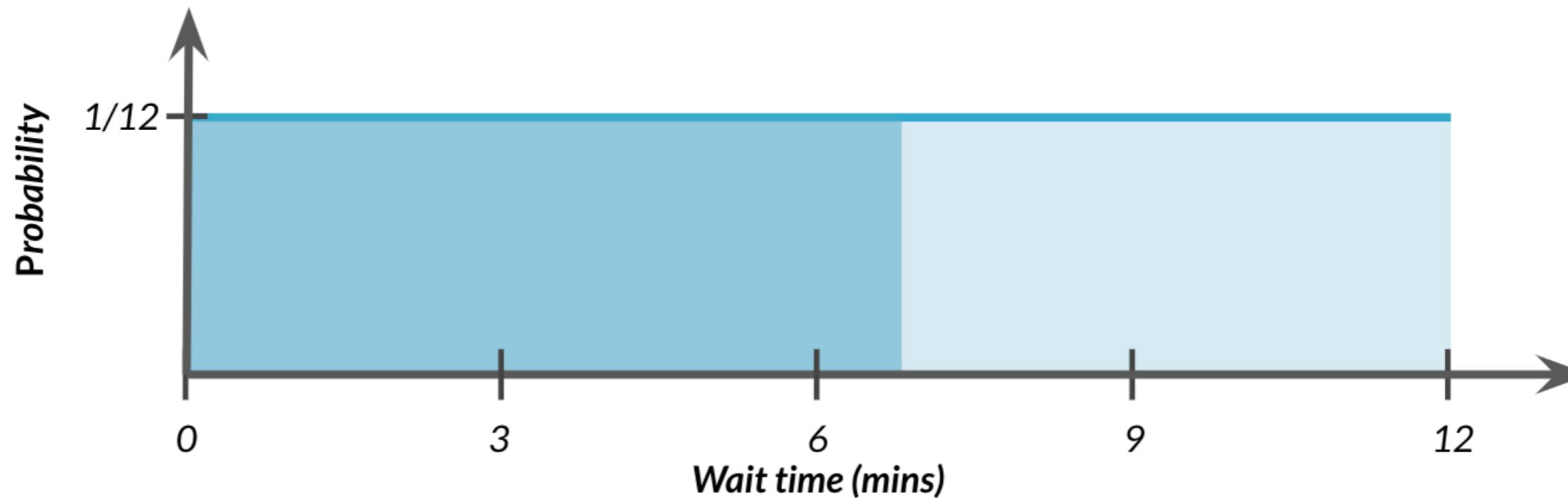


# Waiting seven minutes or less

$$P(\text{wait time} \leq 7) = ?$$

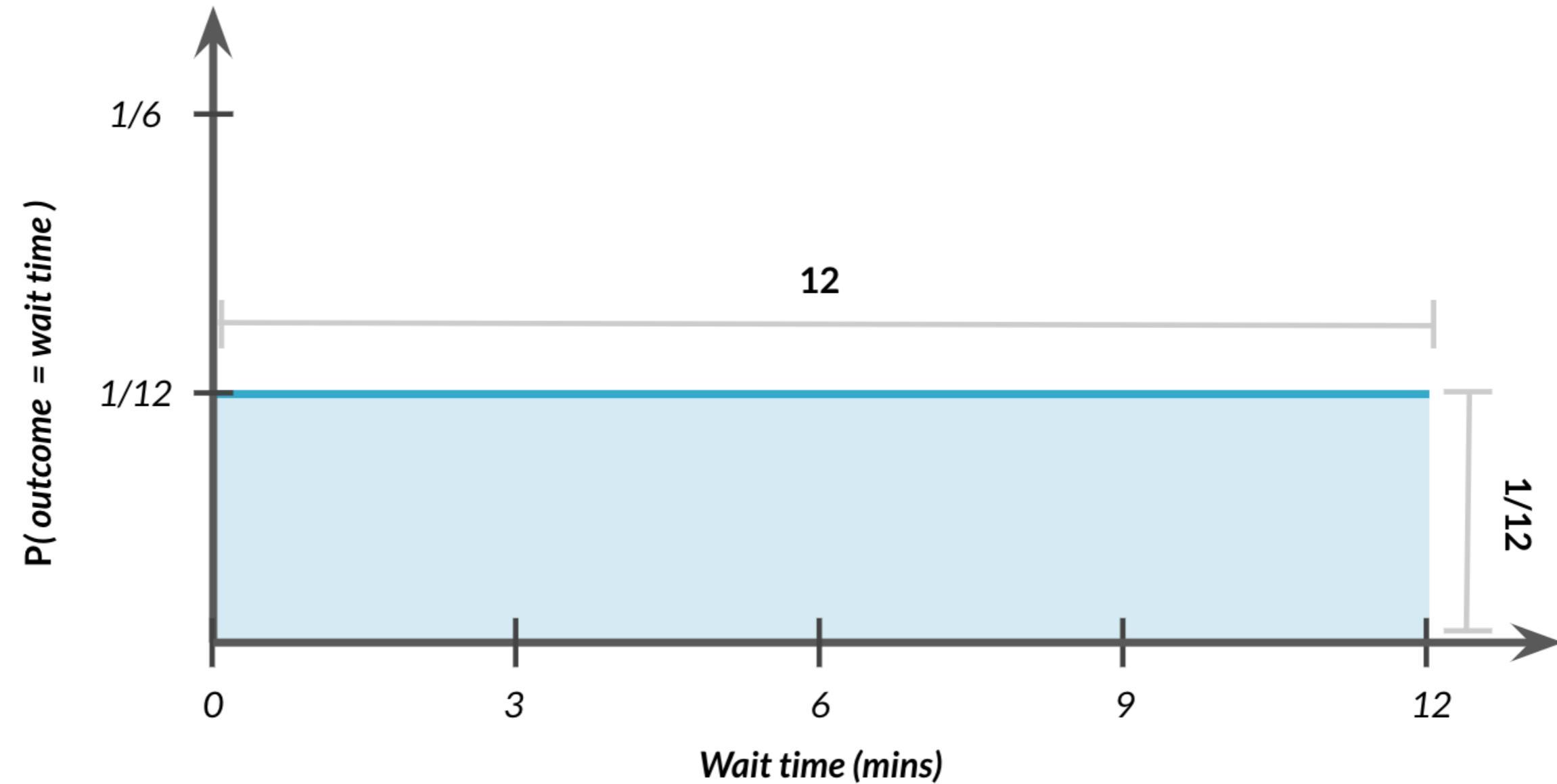
$$P(\text{wait time} \leq 7) = \frac{7 - 0}{12}$$

$$P(\text{wait time} \leq 7) = \frac{7}{12} = 58.33\%$$



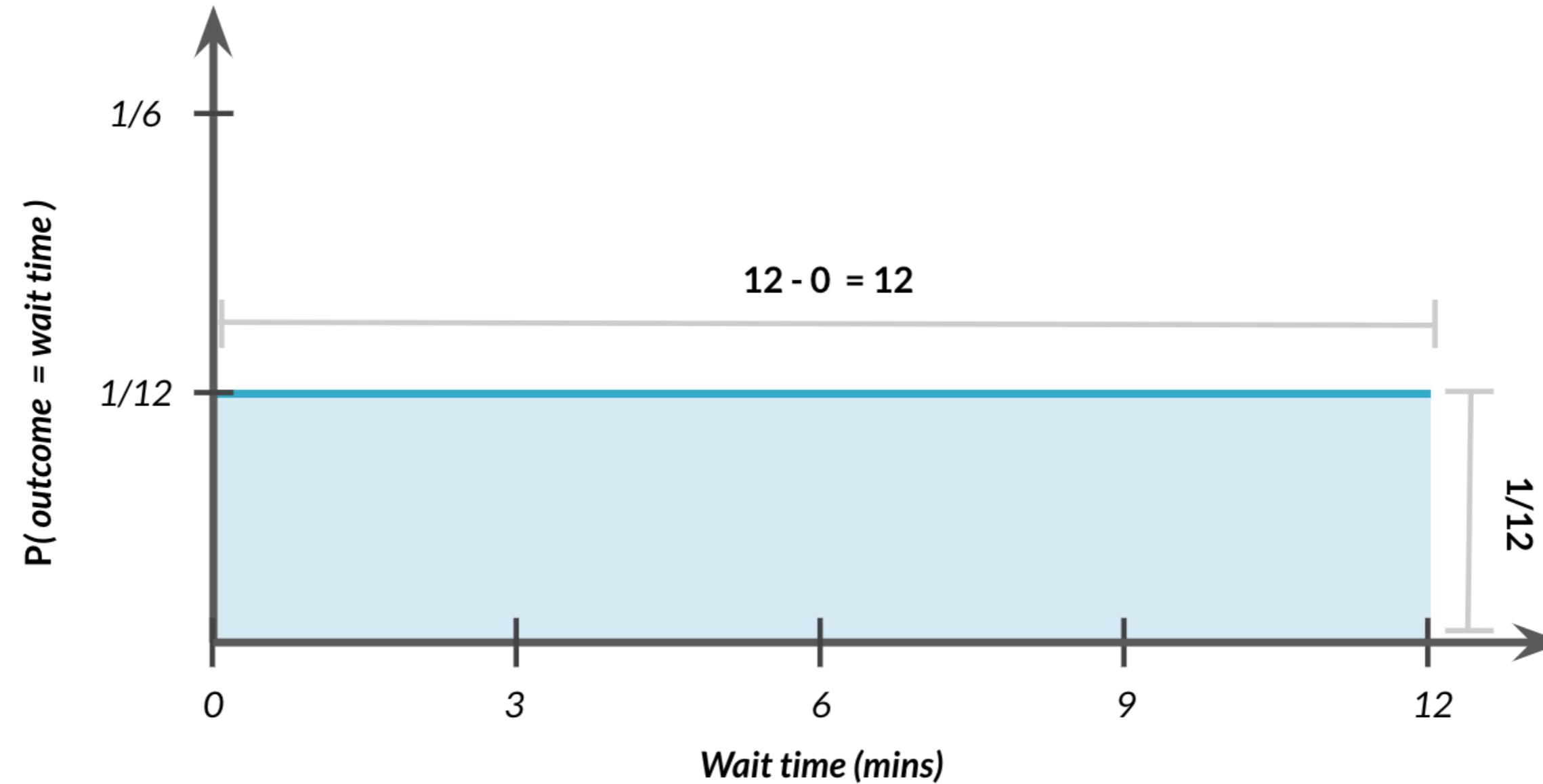
**Total area = 1**

$$P(0 \leq \text{wait time} \leq 12) = ?$$



# Total area = 1

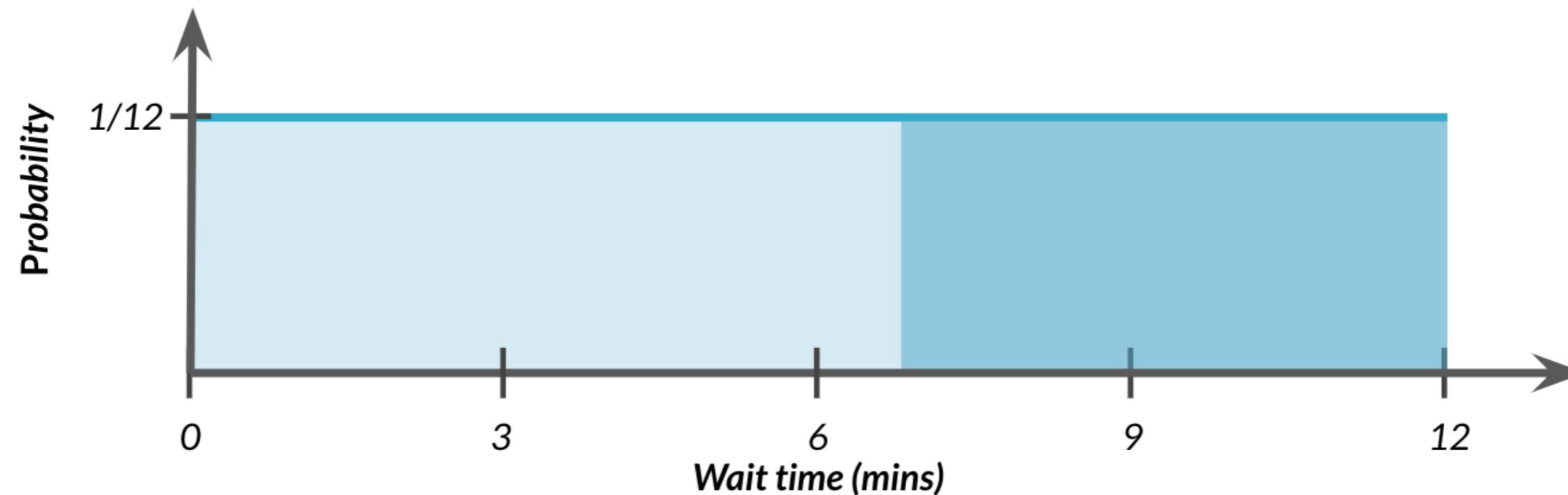
$$P(0 \leq \text{outcome} \leq 12) = 12 \times 1/12 = 1$$



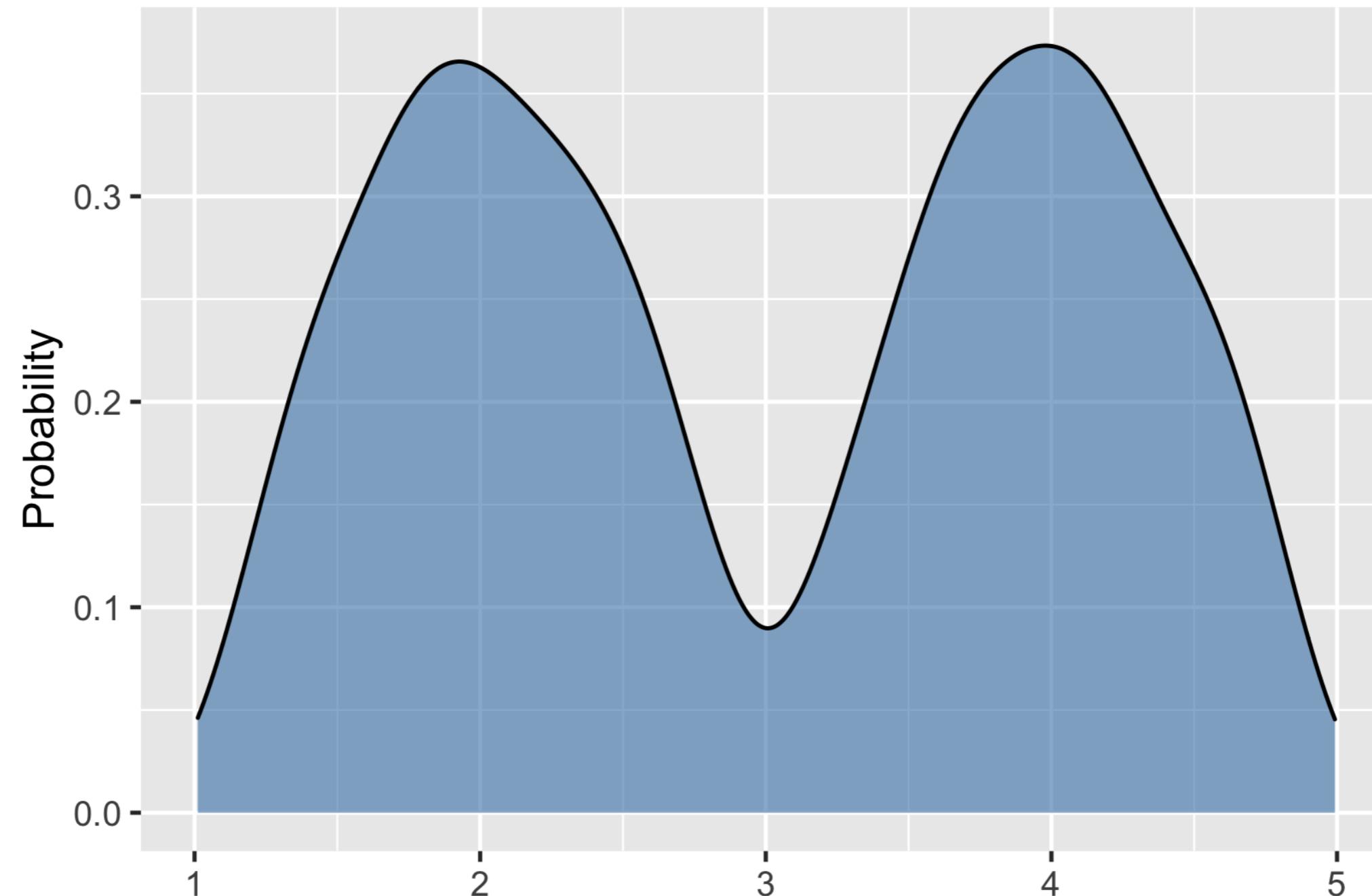
# Probability of waiting more than seven minutes

$$P(\text{wait time} \geq 7) = 1 - \frac{7}{12}$$

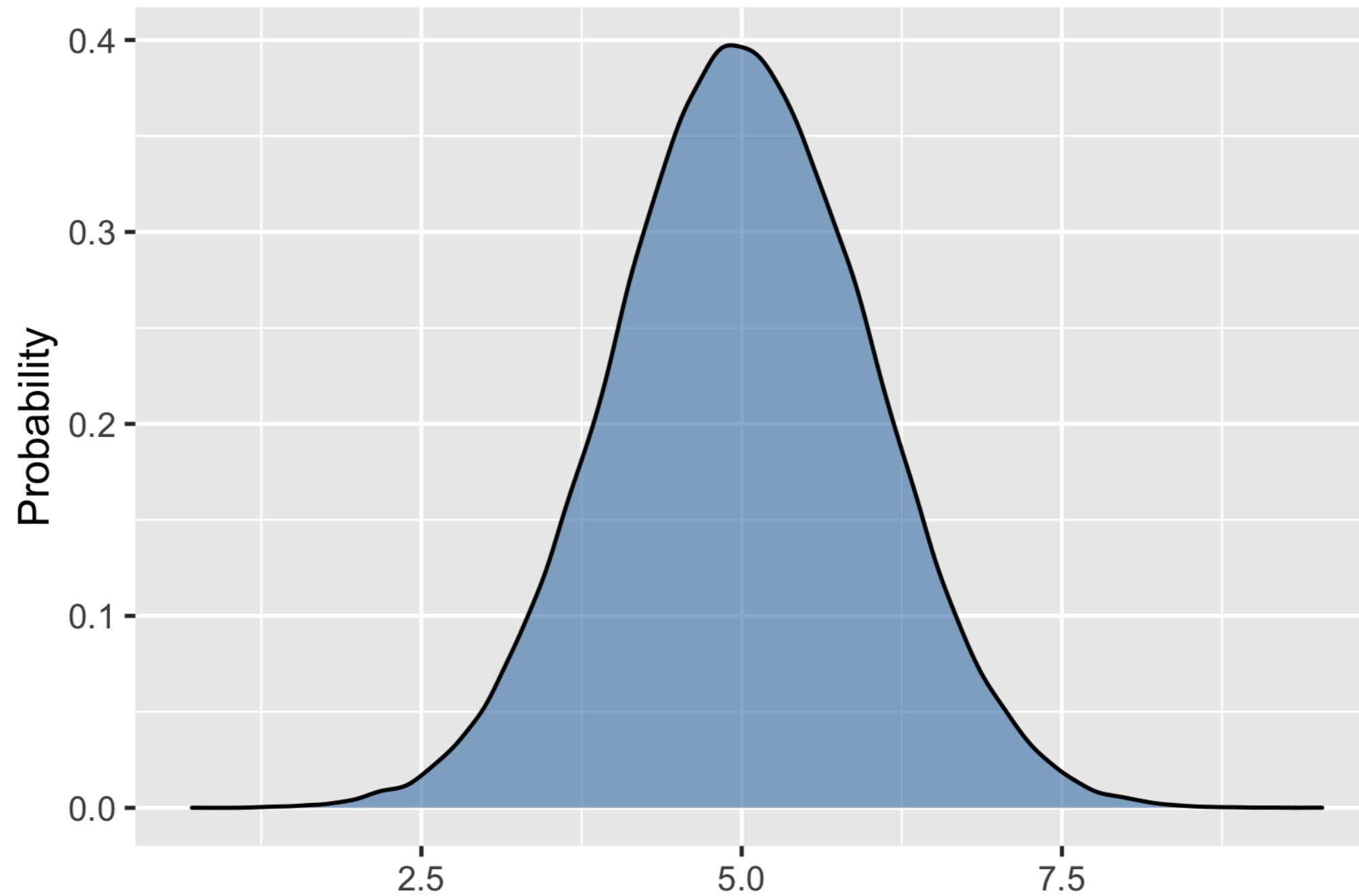
$$P(\text{wait time} \geq 7) = \frac{5}{12} = 41.67\%$$



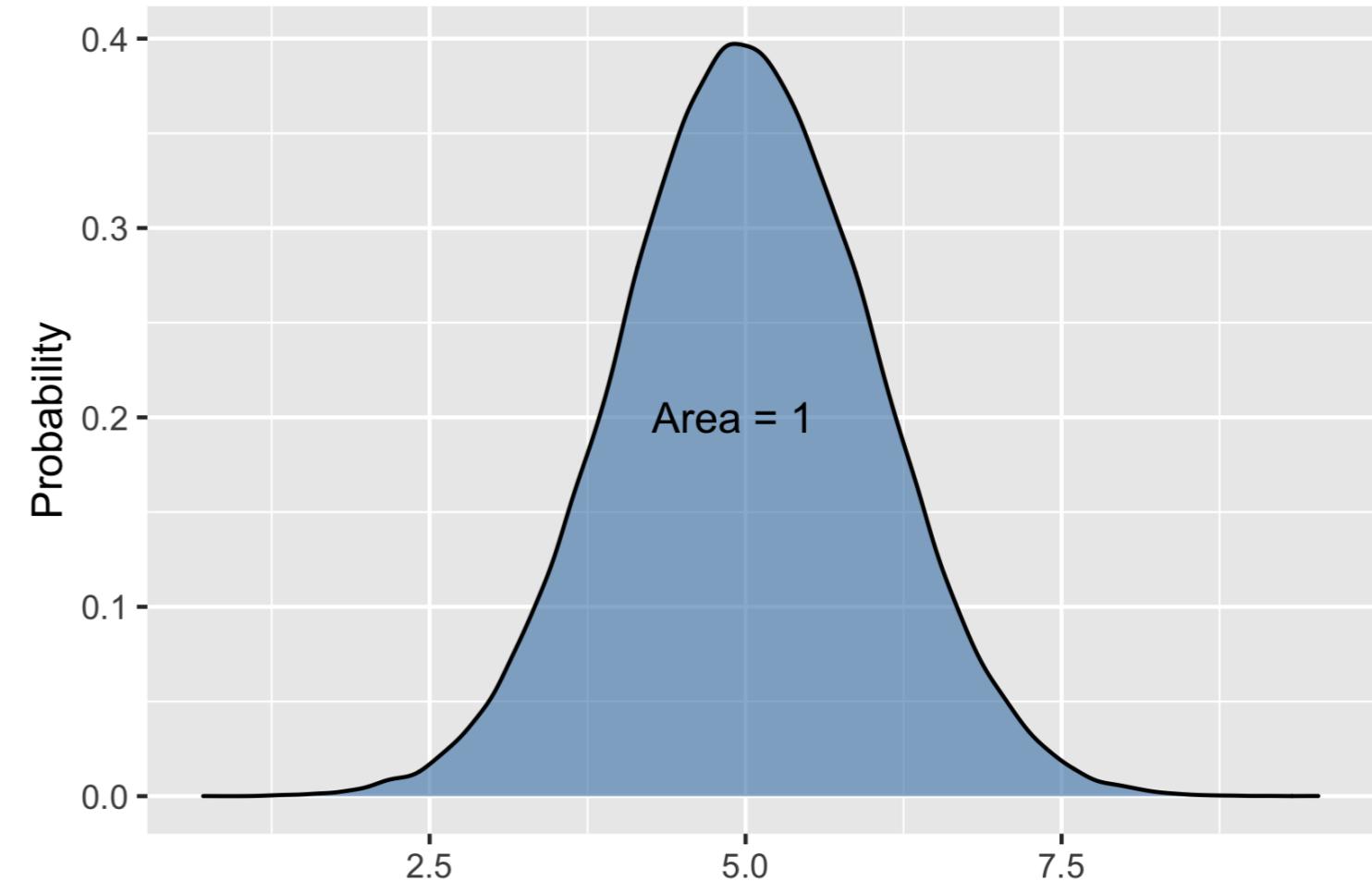
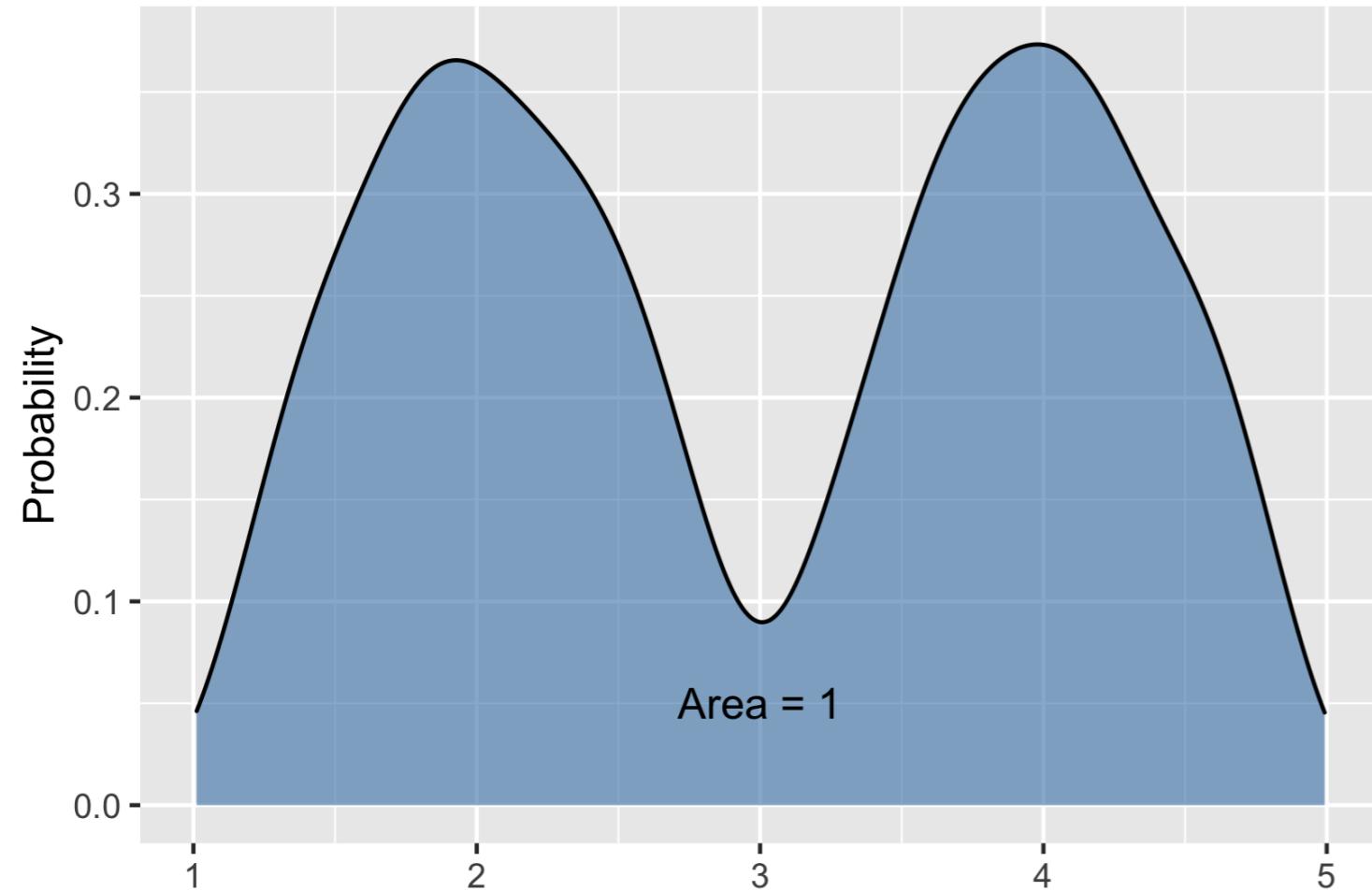
# Bimodal distribution



# The normal distribution



# Total area still = 1



# **Let's practice!**

## **INTRODUCTION TO STATISTICS**