

What are the chances?

INTRODUCTION TO STATISTICS IN R



Maggie Matsui

Content Developer, DataCamp

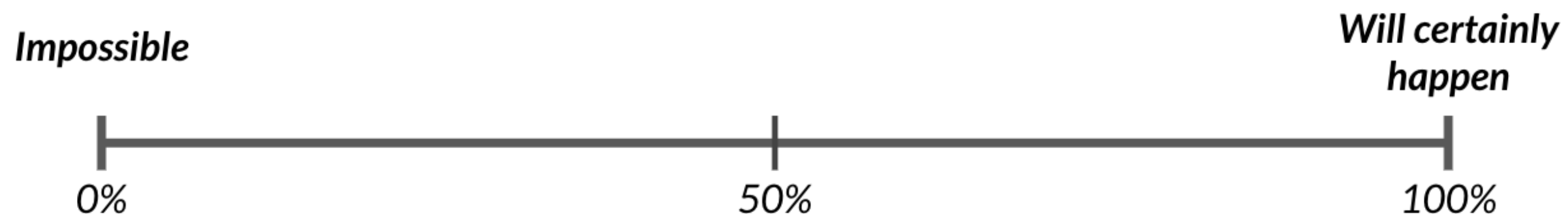
Measuring chance

What's the probability of an event?

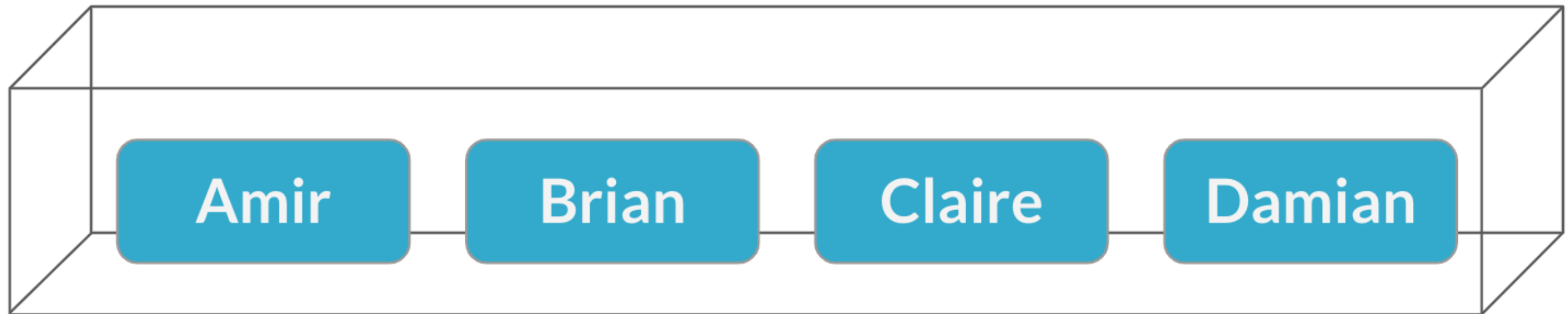
$$P(\text{event}) = \frac{\# \text{ ways event can happen}}{\text{total } \# \text{ of possible outcomes}}$$

Example: a coin flip

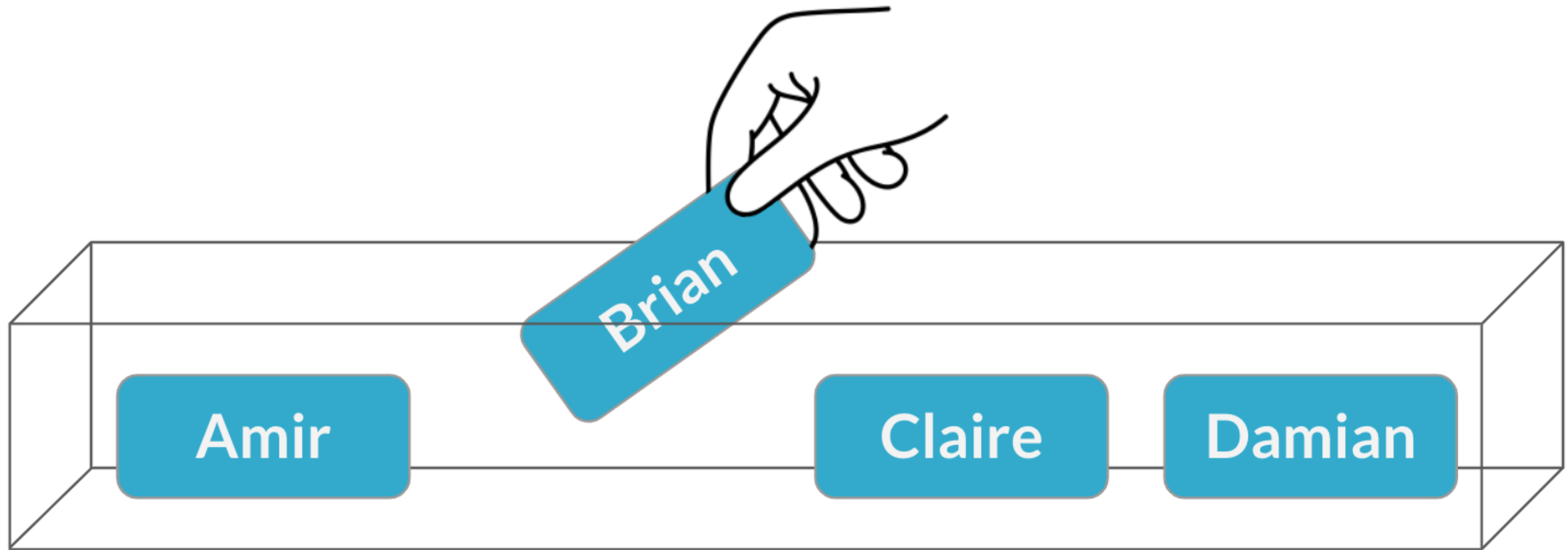
$$P(\text{heads}) = \frac{1 \text{ way to get heads}}{2 \text{ possible outcomes}} = \frac{1}{2} = 50\%$$



Assigning salespeople



Assigning salespeople



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

Sampling from a data frame

```
sales_counts
```

```
  name  n_sales  
1 Amir     178  
2 Brian    126  
3 Claire     75  
4 Damian     69
```

```
sales_counts %>%  
  sample_n(1)
```

```
  name  n_sales  
1 Brian     126
```

```
sales_counts %>%  
  sample_n(1)
```

```
  name  n_sales  
1 Claire     75
```

Setting a random seed

```
set.seed(5)  
sales_counts %>%  
  sample_n(1)
```

```
  name  n_sales  
1 Brian     126
```

```
set.seed(5)  
sales_counts %>%  
  sample_n(1)
```

```
  name  n_sales  
1 Brian     126
```

A second meeting

Sampling without replacement



A second meeting



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

Sampling twice in R

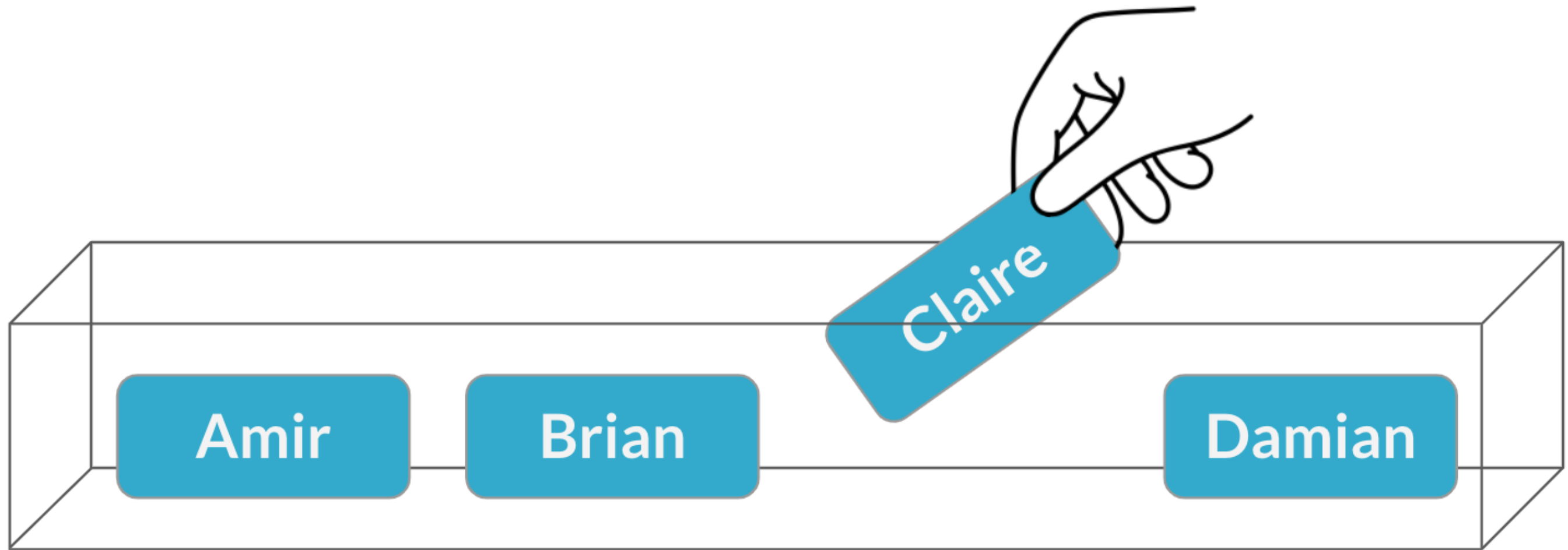
```
sales_counts %>%  
  sample_n(2)
```

```
  name  n_sales  
1 Brian    126  
2 Claire    75
```

Sampling with replacement



Sampling with replacement



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

Sampling with replacement in R

```
sales_counts %>%  
  sample_n(2, replace = TRUE)
```

	name	n_sales
1	Brian	126
2	Claire	75

5 meetings:

```
sample(sales_team, 5, replace = TRUE)
```

	name	n_sales
1	Brian	126
2	Claire	75
3	Brian	126
4	Brian	126
5	Amir	178

Independent events

Two events are *independent* if the probability of the second event *isn't* affected by the outcome of the first event.

Sampling with Replacement

First pick

Second pick

Amir

Brian

Claire

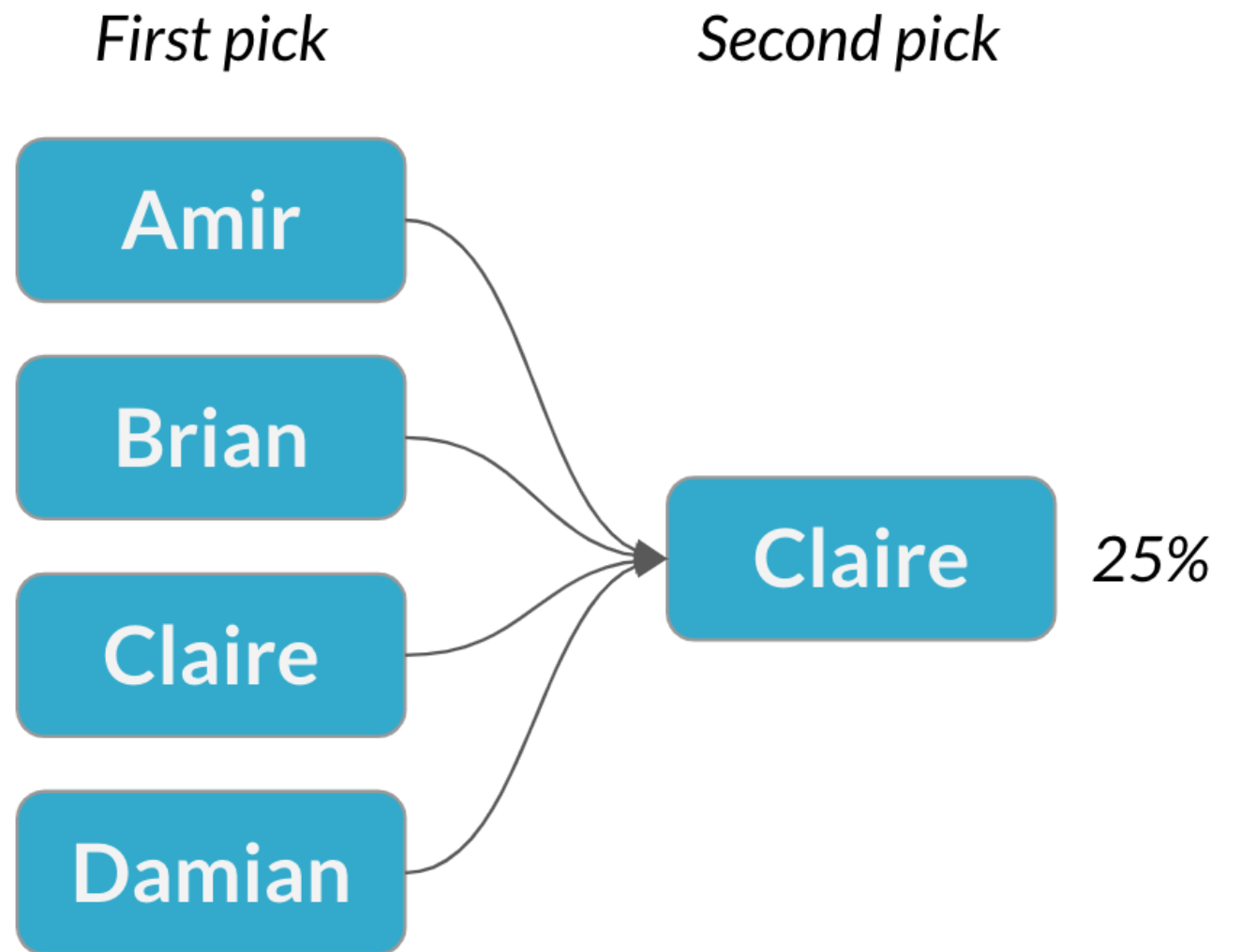
Damian

Independent events

Two events are *independent* if the probability of the second event *isn't* affected by the outcome of the first event.

Sampling with replacement = each pick is independent

Sampling with Replacement



Dependent events

Two events are *dependent* if the 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

Two events are *dependent* if the 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

Claire

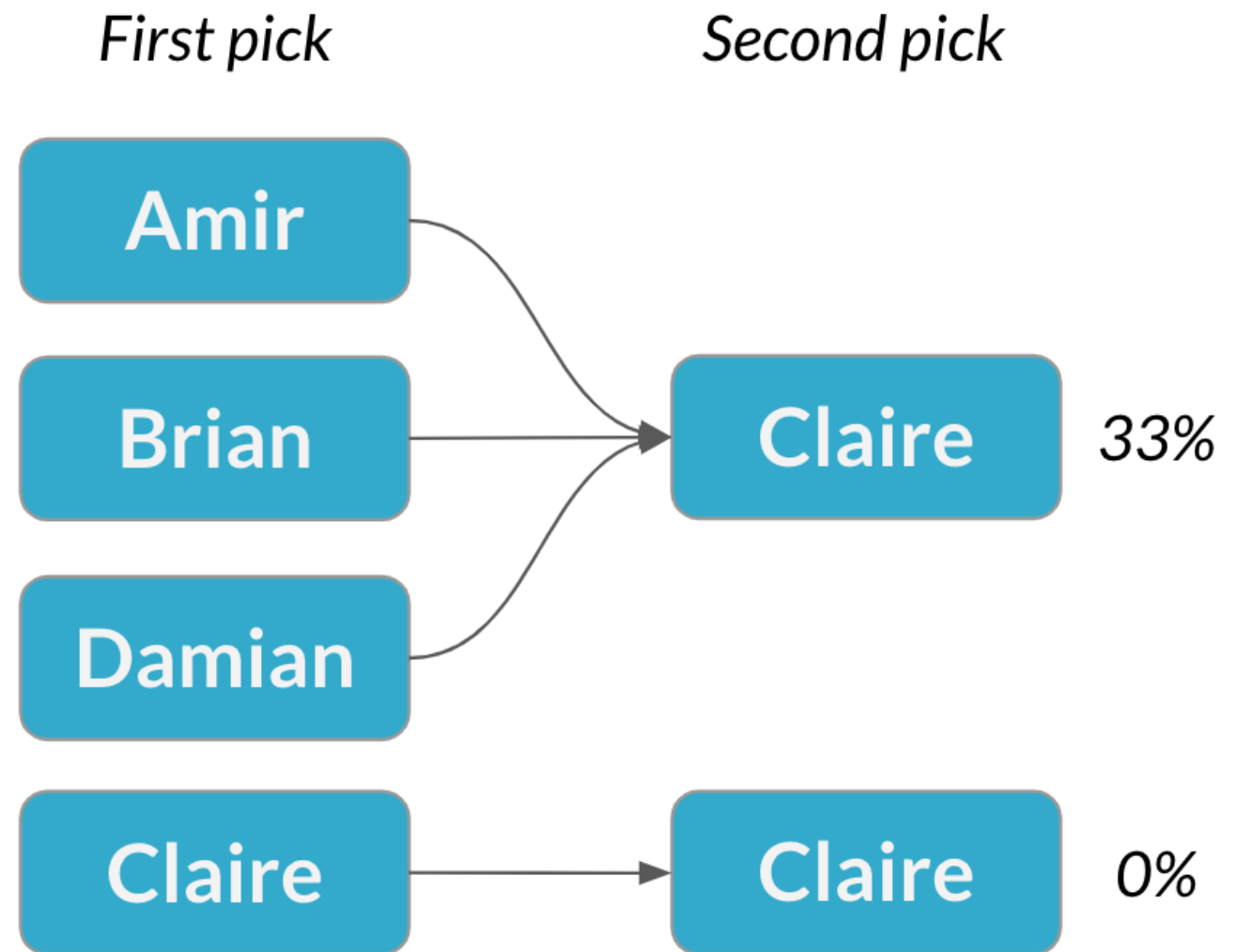
0%

Dependent events

Two events are *dependent* if the probability of the second event *is* affected by the outcome of the first event.

Sampling without replacement = each pick is dependent

Sampling without Replacement



Let's practice!

INTRODUCTION TO STATISTICS IN R

Discrete distributions

INTRODUCTION TO STATISTICS IN R



Maggie Matsui

Content Developer, DataCamp

Rolling the dice



Rolling the dice



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



$\frac{1}{6}$



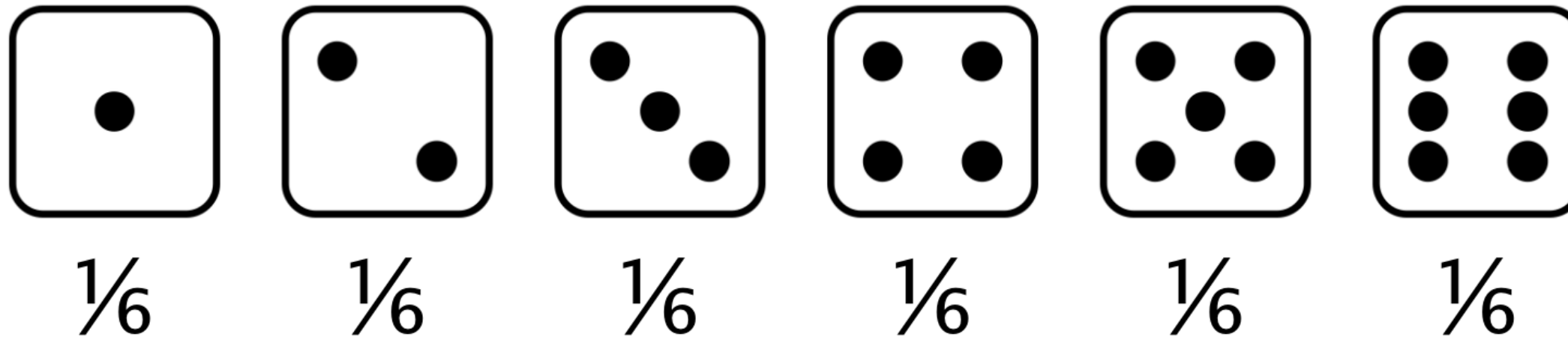
$\frac{1}{6}$

Choosing salespeople



Probability distribution

Describes the probability of each possible outcome in a scenario



Expected value: mean of a probability distribution

Expected value of a fair die roll =

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

Visualizing a probability distribution



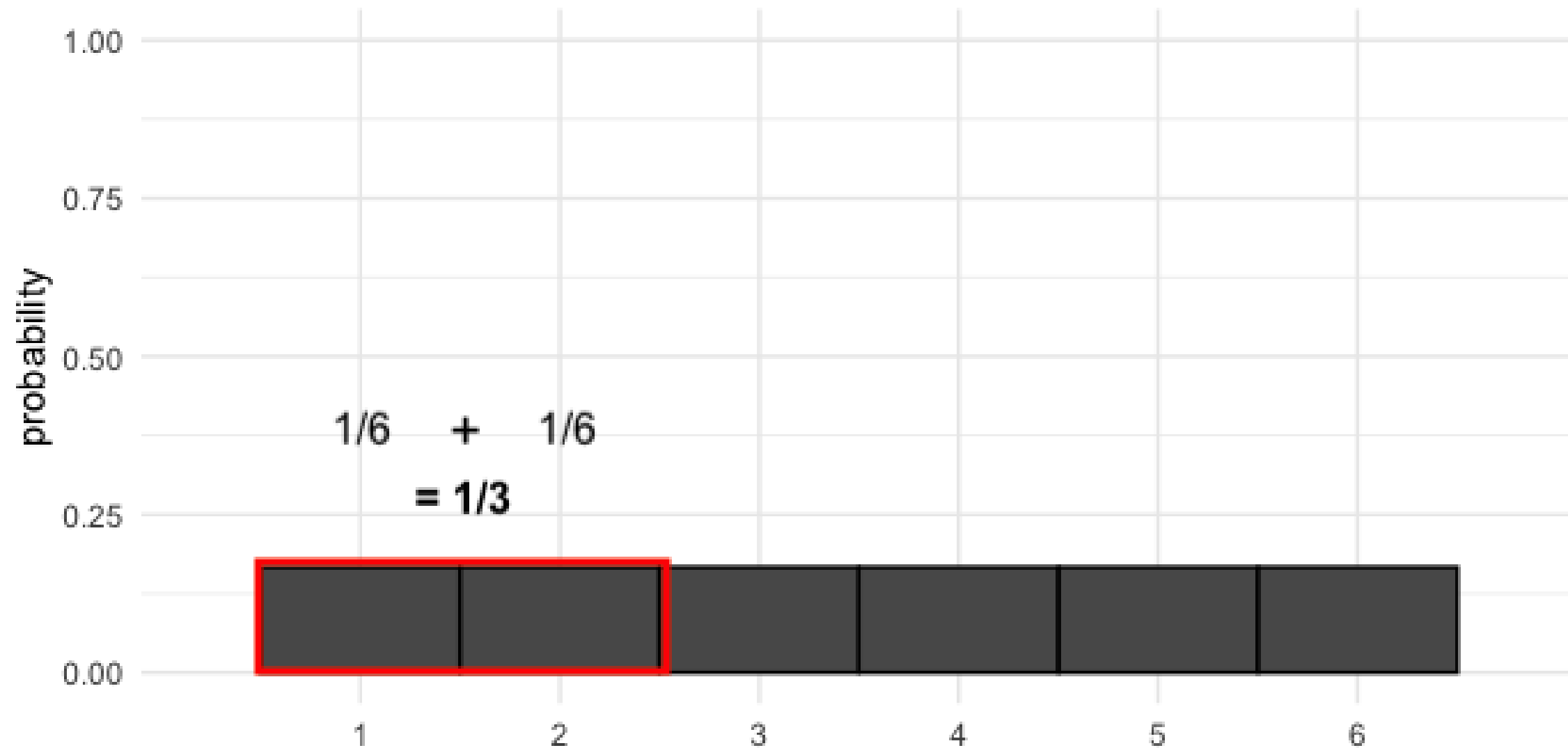
Probability = area

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

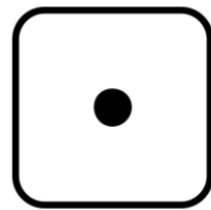


Probability = area

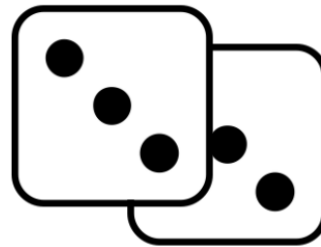
$$P(\text{die roll}) \leq 2 = 1/3$$



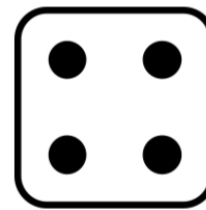
Uneven die



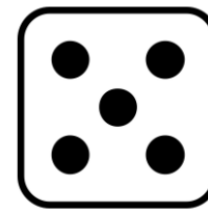
$\frac{1}{6}$



$\frac{1}{3}$



$\frac{1}{6}$



$\frac{1}{6}$

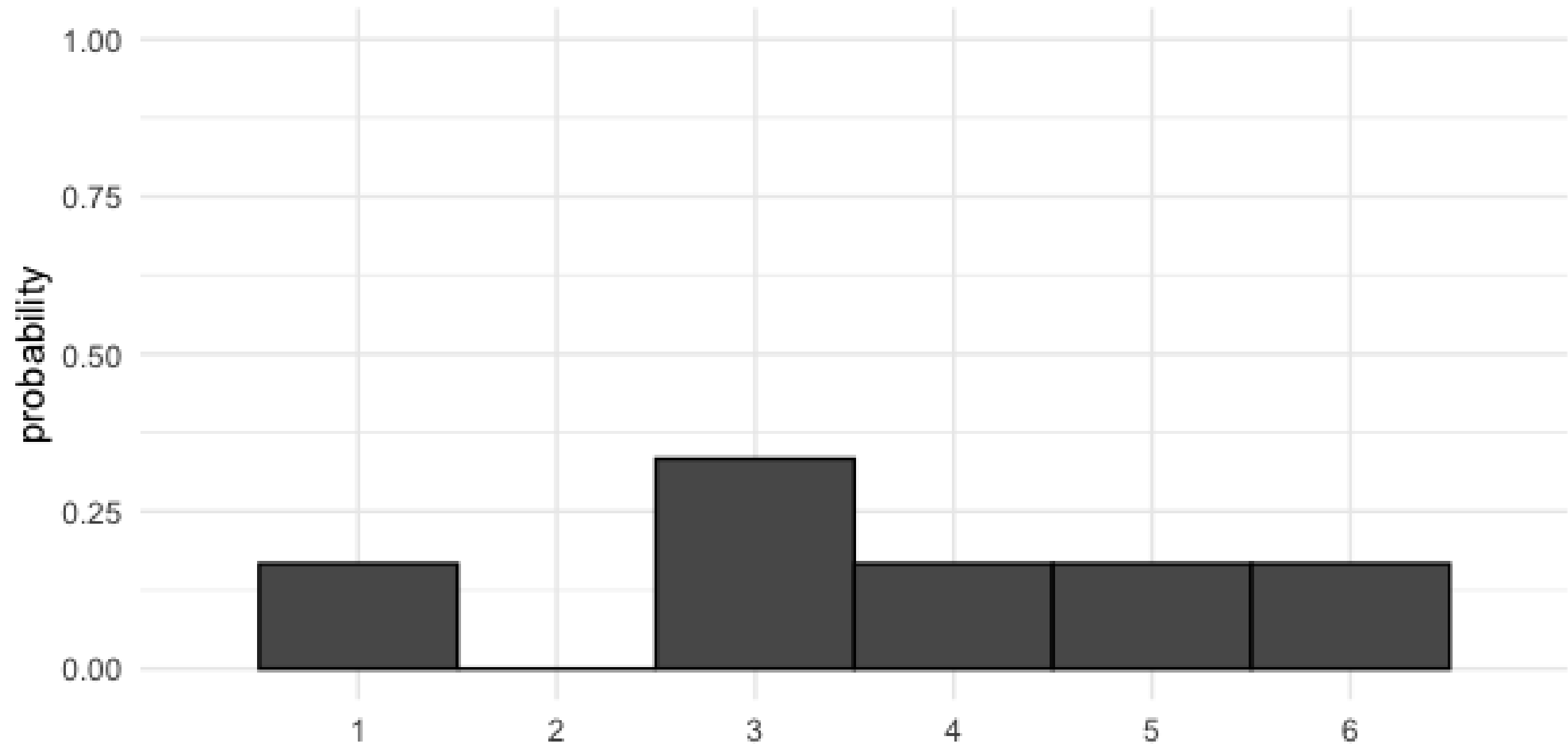


$\frac{1}{6}$

Expected value of uneven die roll =

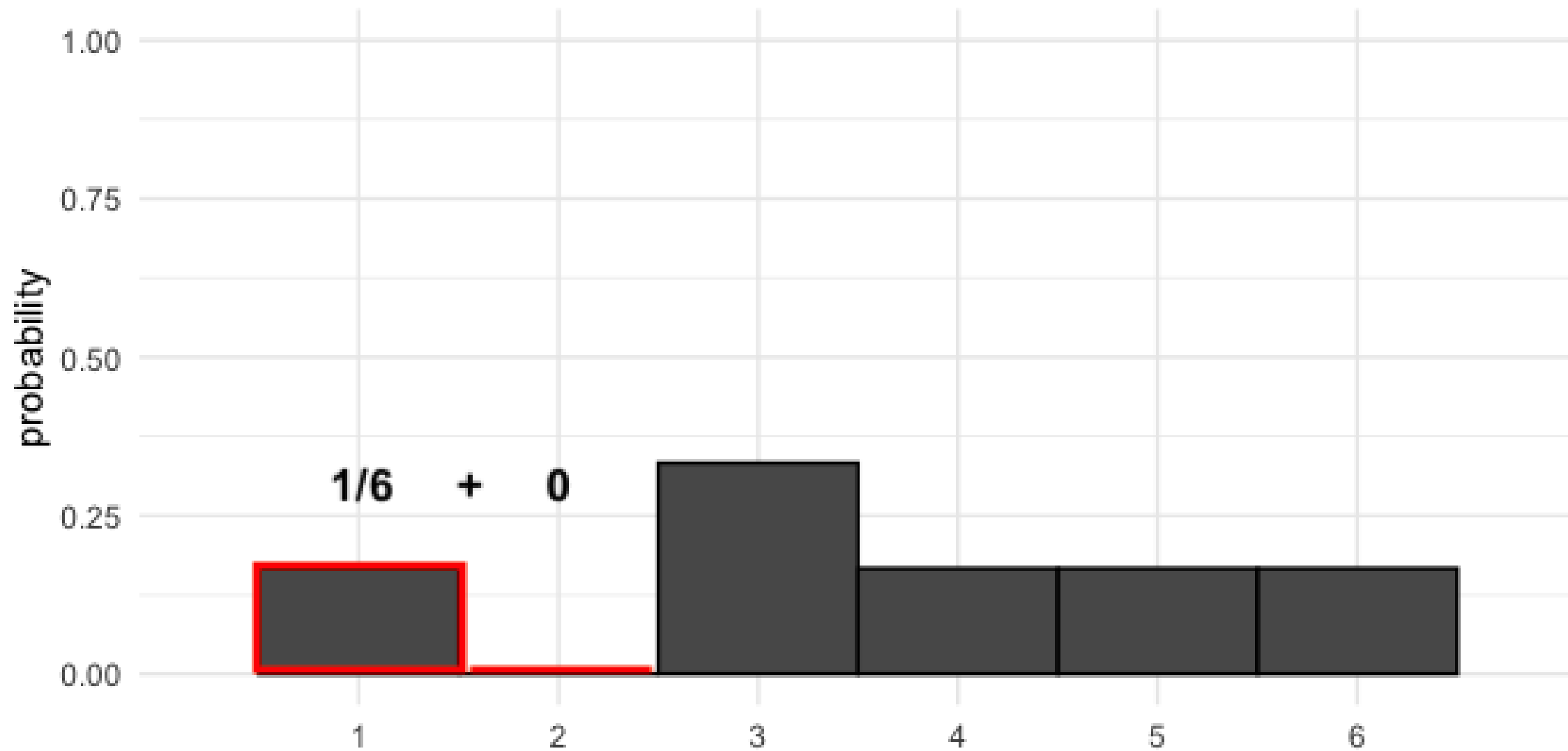
$$(1 \times \frac{1}{6}) + (2 \times 0) + (3 \times \frac{1}{3}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) = 3.67$$

Visualizing uneven probabilities



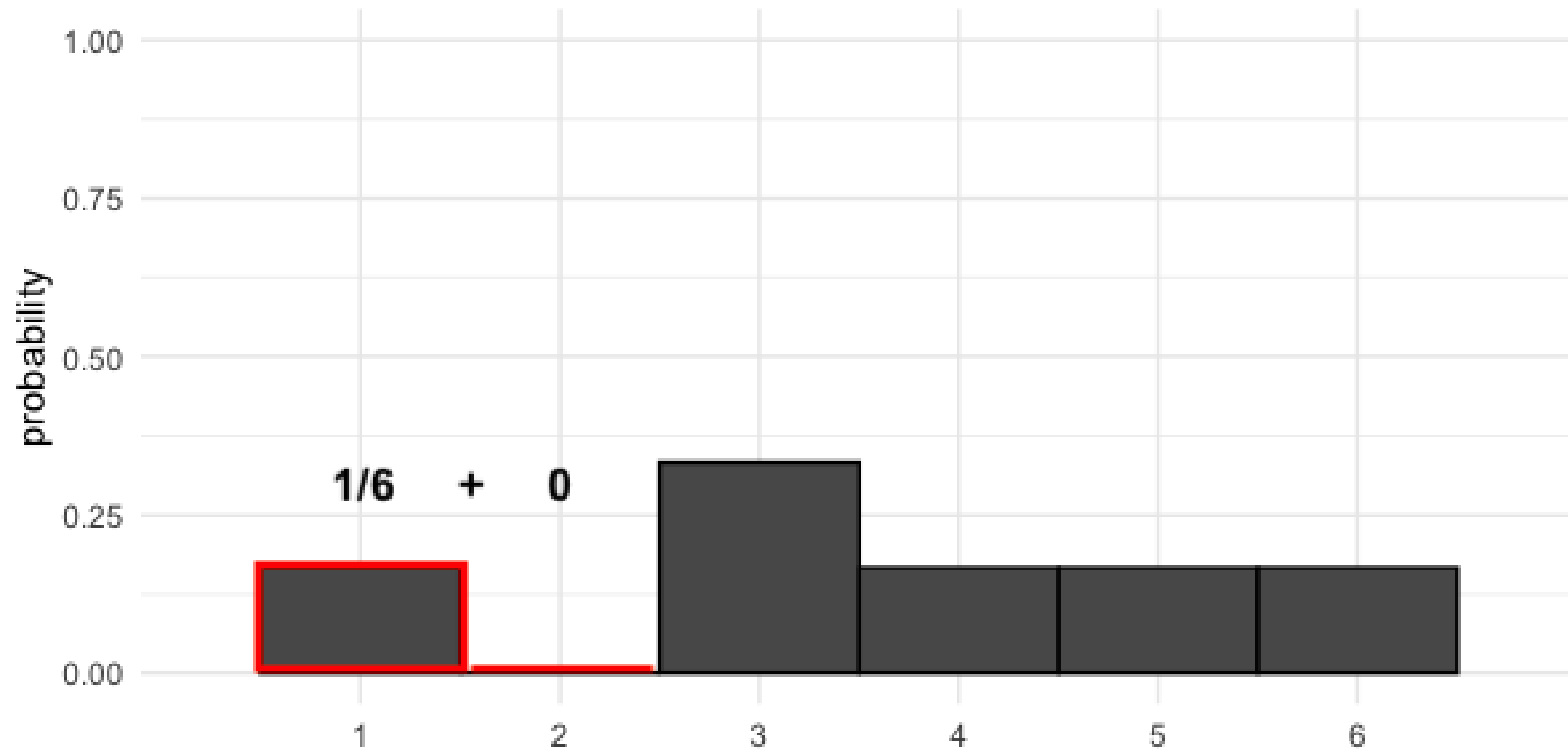
Adding areas

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



Adding areas

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



Discrete probability distributions

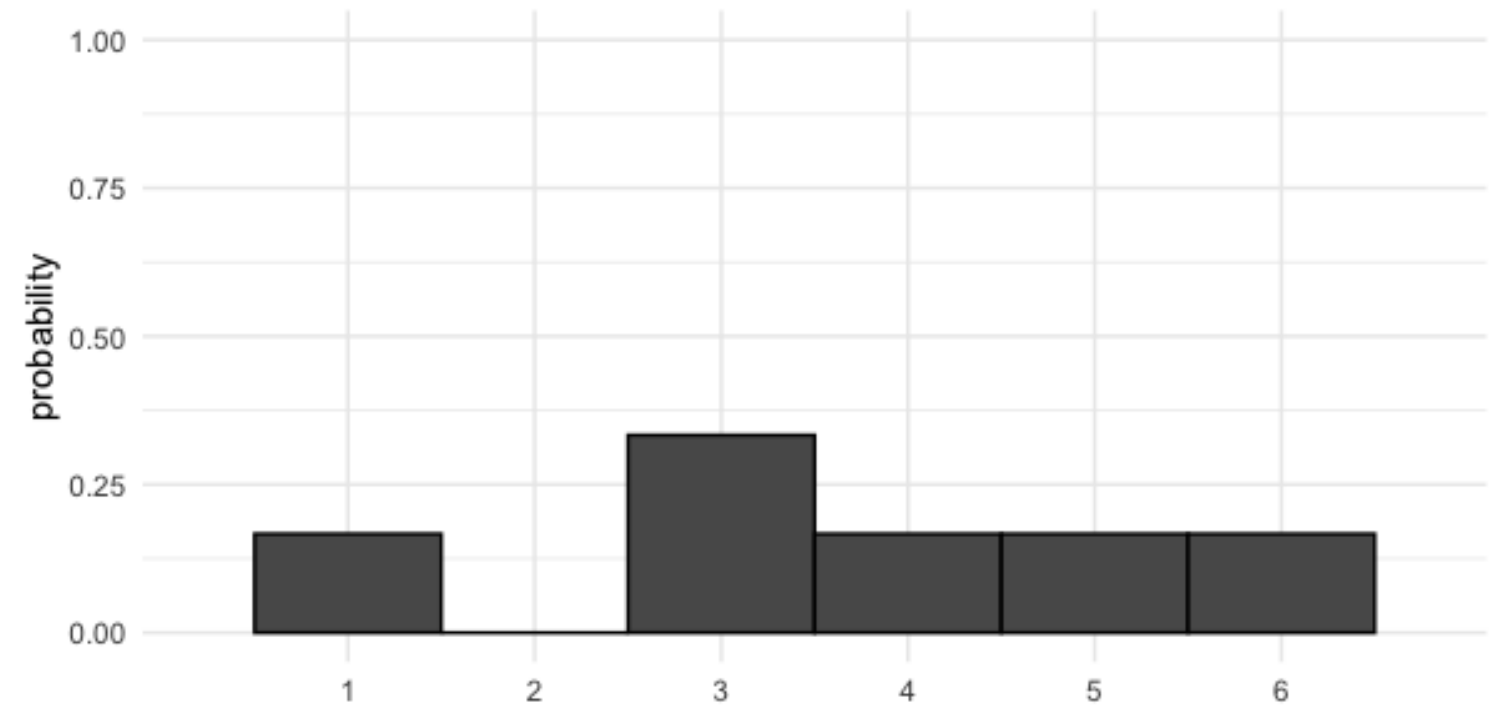
Describe probabilities for discrete outcomes

Fair die



Discrete uniform distribution

Uneven die



Sampling from discrete distributions

```
die
```

```
  n  
1 1  
2 2  
3 3  
4 4  
5 5  
6 6
```

```
mean(die$n)
```

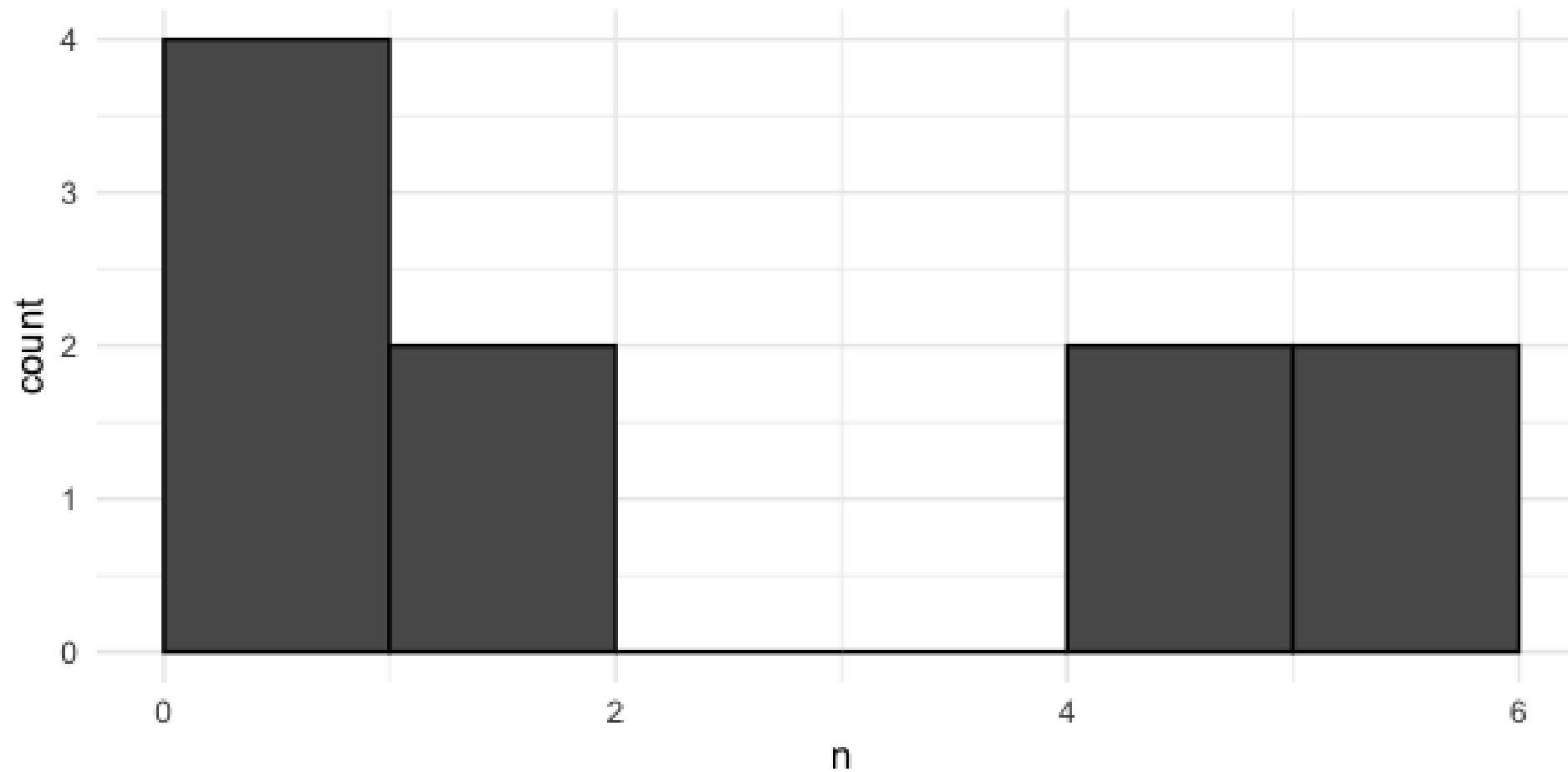
```
3.5
```

```
rolls_10 <- die %>%  
  sample_n(10, replace = TRUE)  
rolls_10
```

```
  n  
1 1  
2 1  
3 5  
4 2  
5 1  
6 1  
7 6  
8 6  
...
```

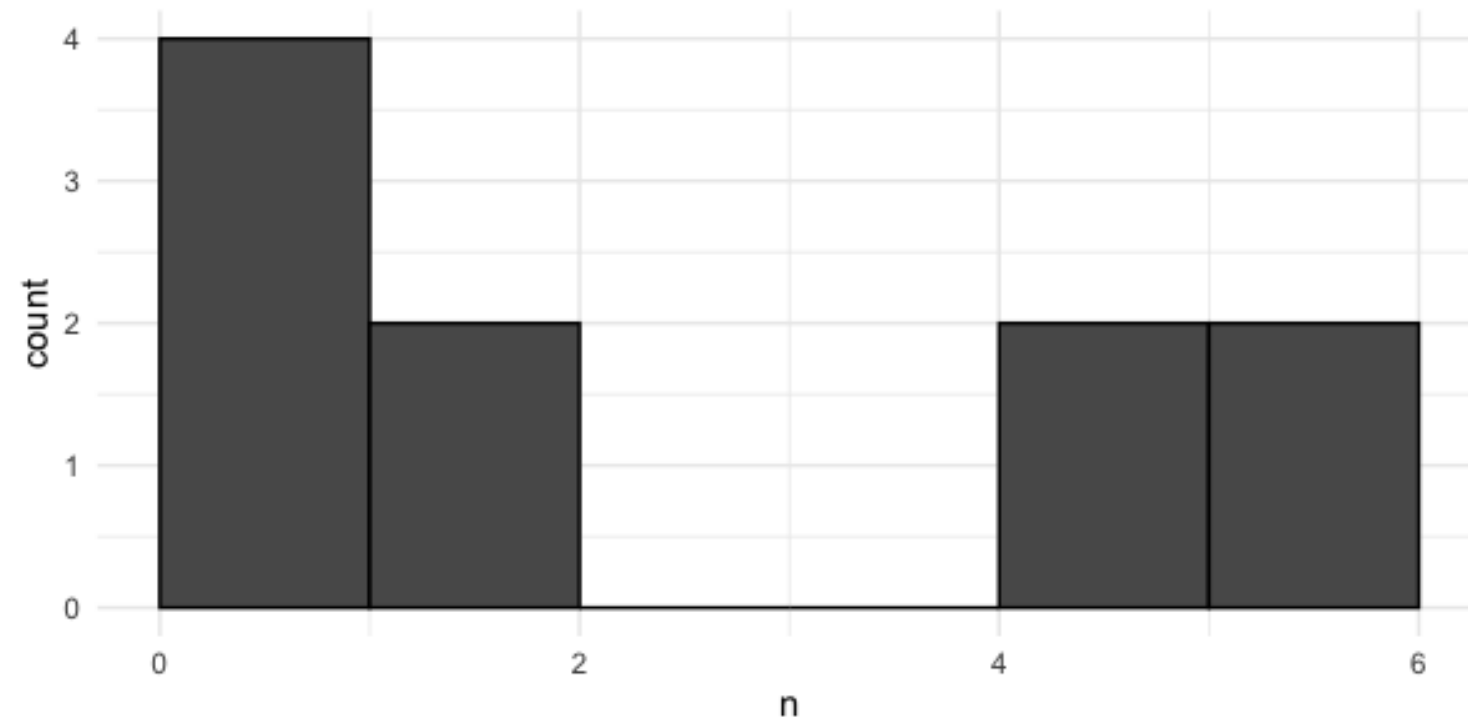

Visualizing a sample

```
ggplot(rolls_10, aes(n)) +  
  geom_histogram(bins = 6)
```



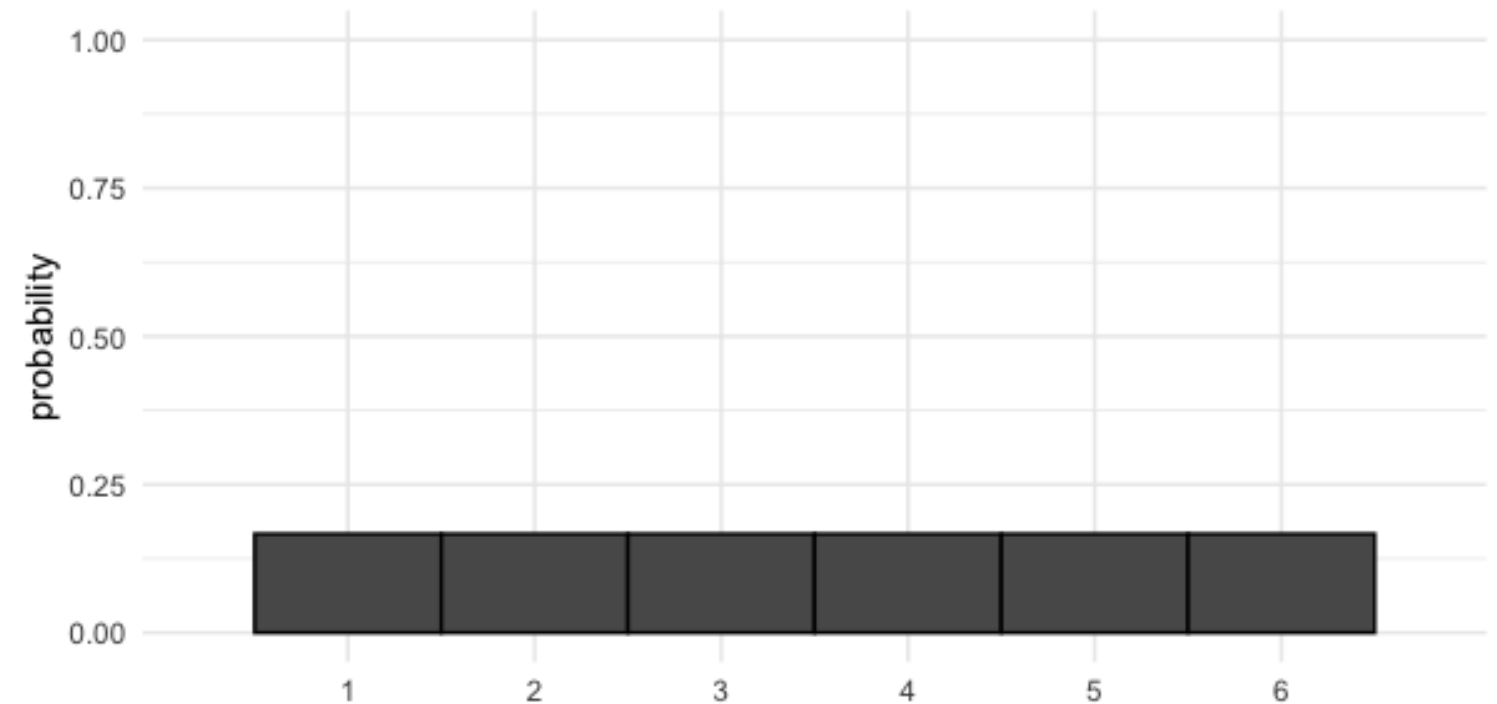
Sample distribution vs. theoretical distribution

Sample of 10 rolls



```
mean(rolls_10$n) = 3.0
```

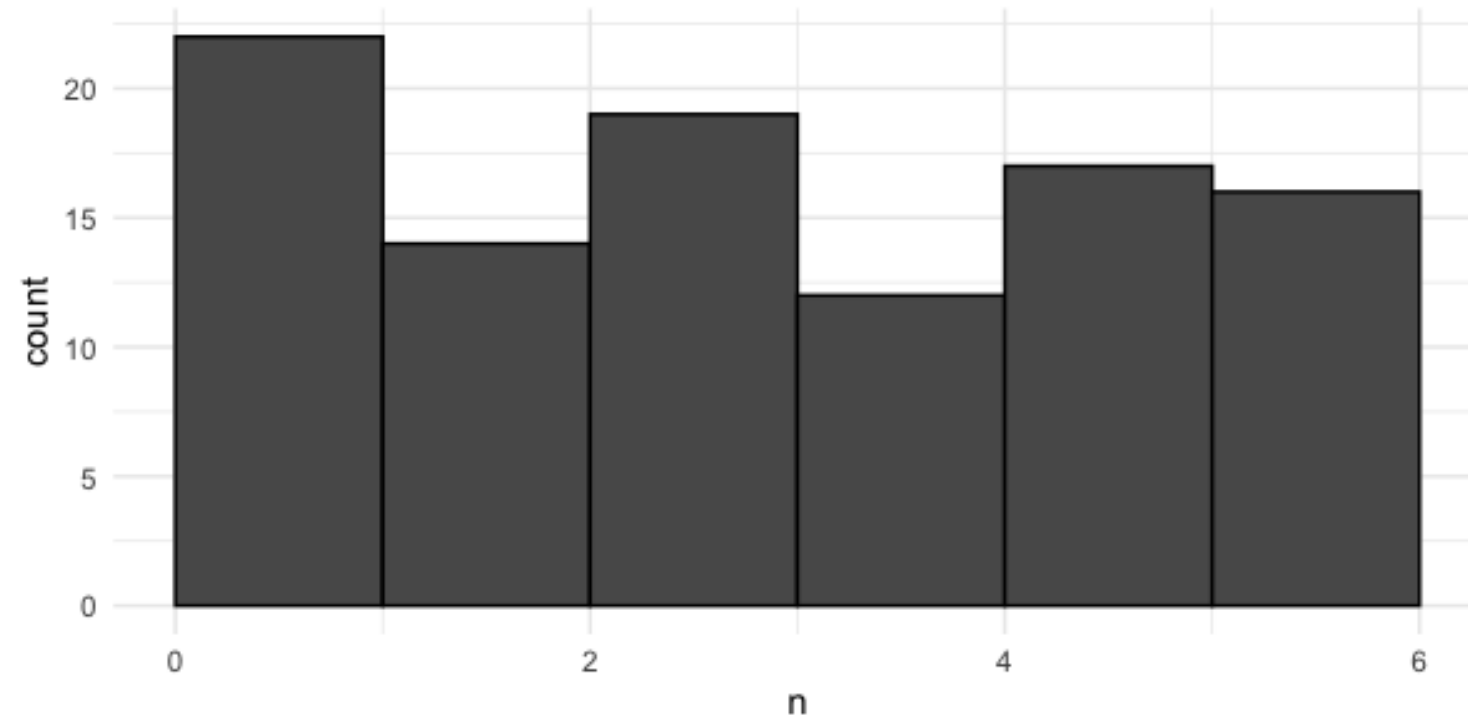
Theoretical probability distribution



```
mean(die$n) = 3.5
```

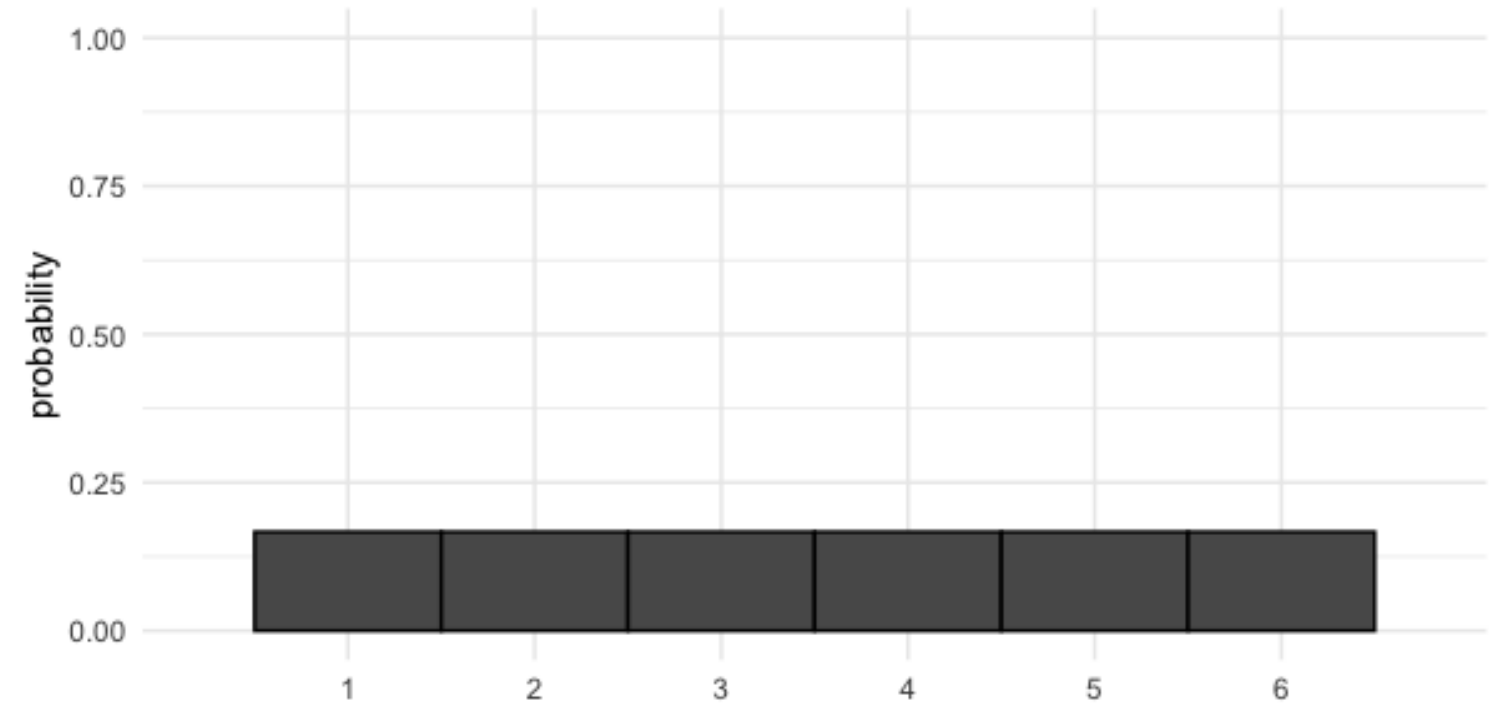
A bigger sample

Sample of 100 rolls



```
mean(rolls_100$n) = 3.36
```

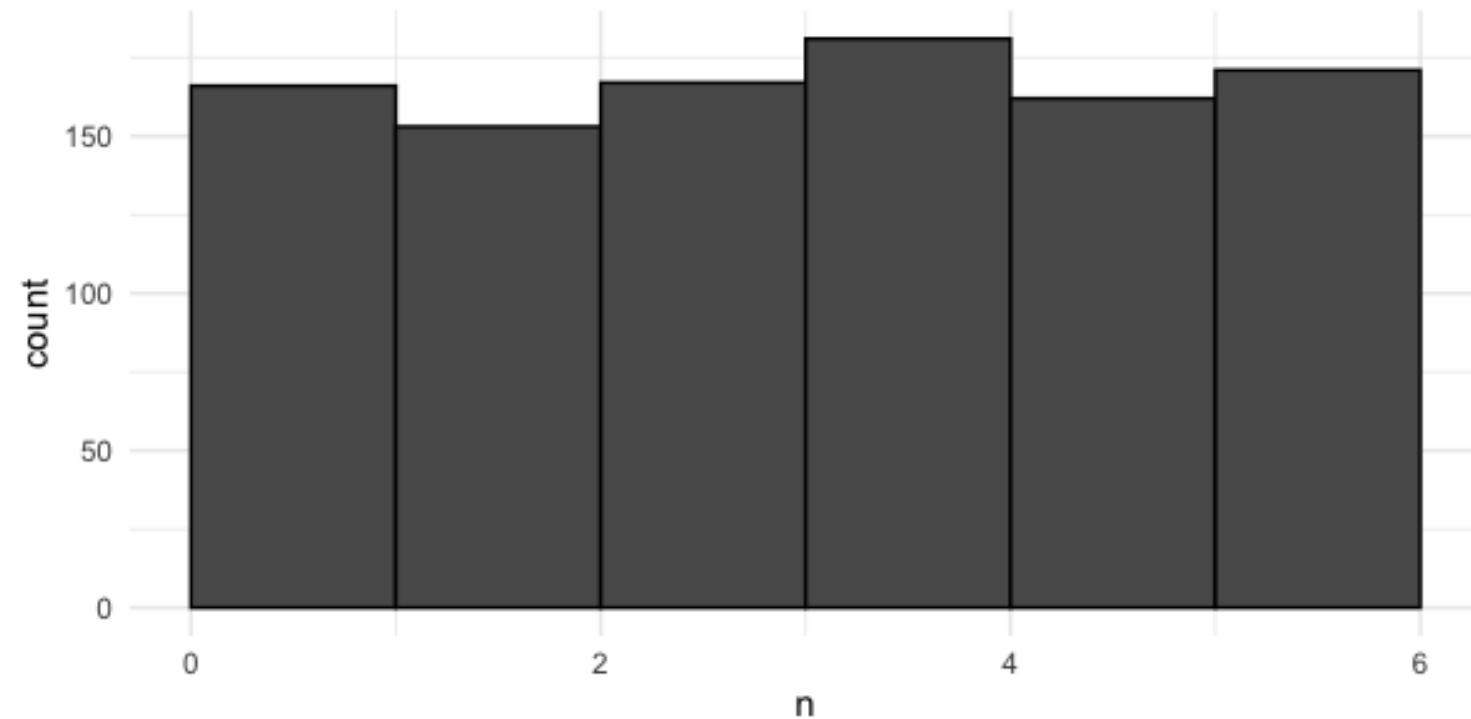
Theoretical probability distribution



```
mean(die$n) = 3.5
```

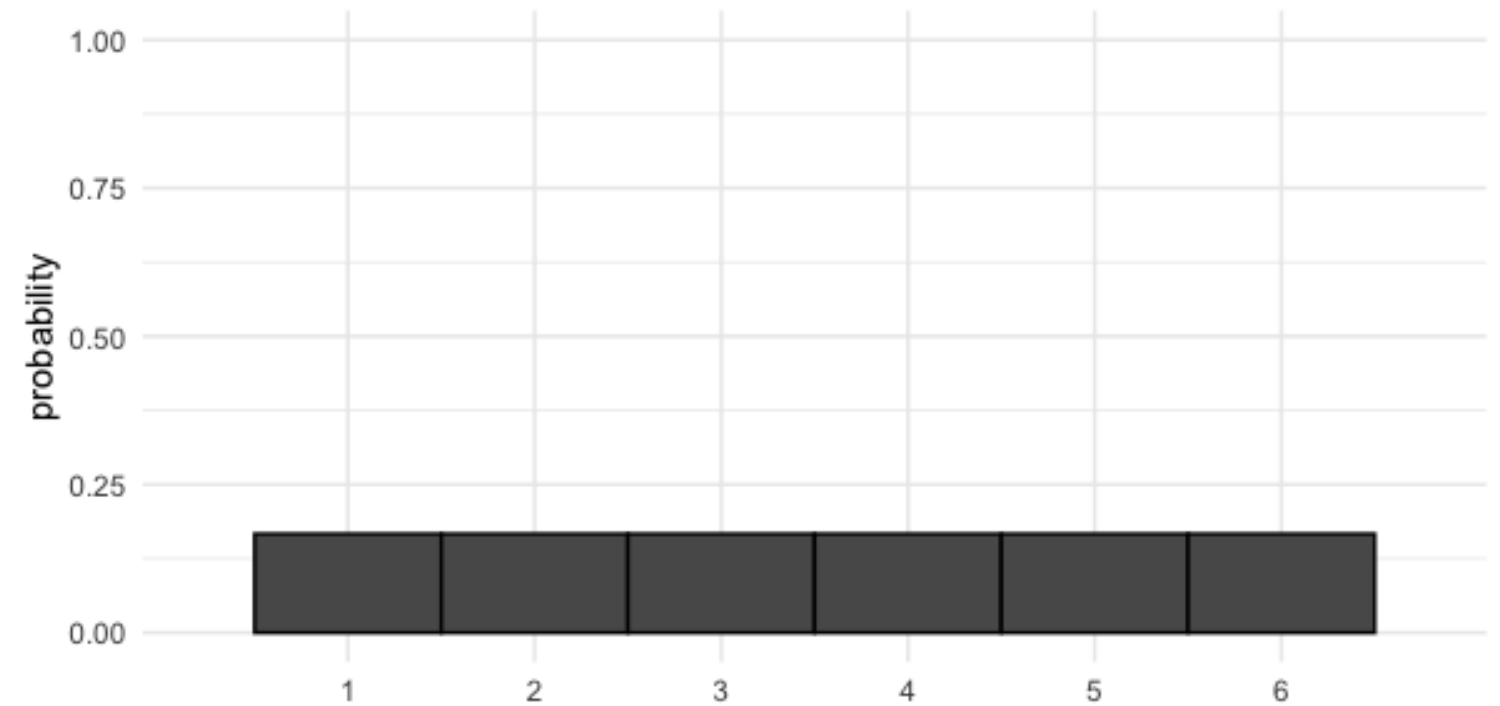
An even bigger sample

Sample of 1000 rolls



```
mean(rolls_1000$n) = 3.53
```

Theoretical probability distribution



```
mean(die$n) = 3.5
```

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	3.36
1000	3.53

Let's practice!

INTRODUCTION TO STATISTICS IN R

Continuous distributions

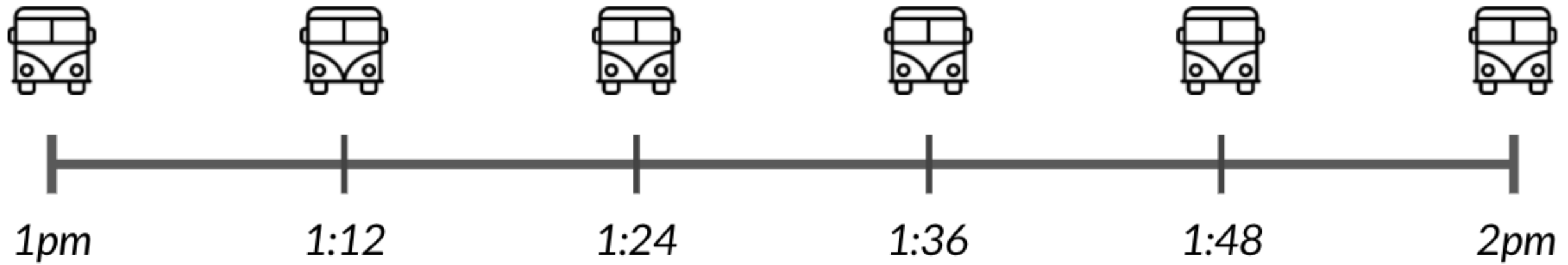
INTRODUCTION TO STATISTICS IN R



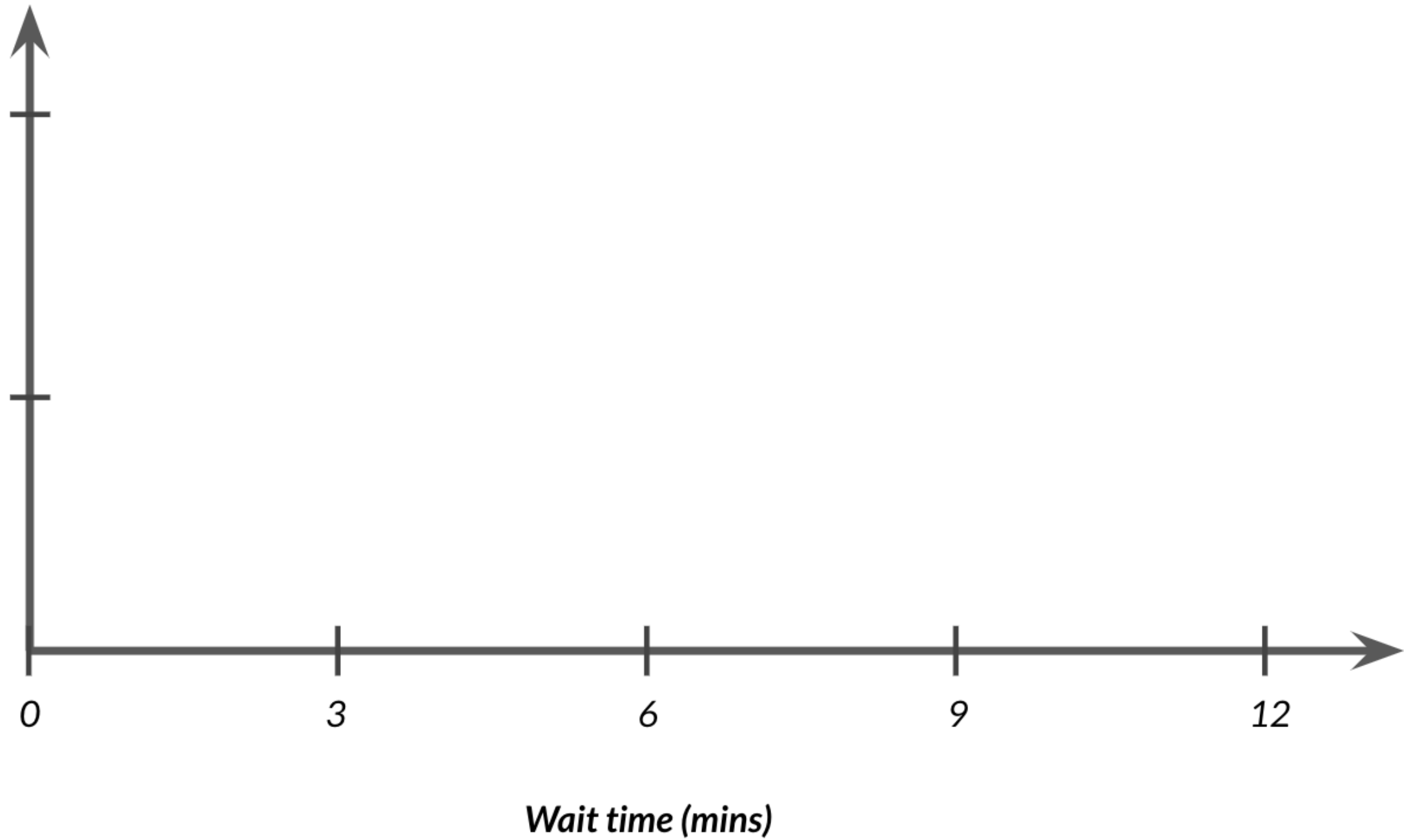
Maggie Matsui

Content Developer, DataCamp

Waiting for the bus



Continuous uniform distribution



Continuous uniform distribution



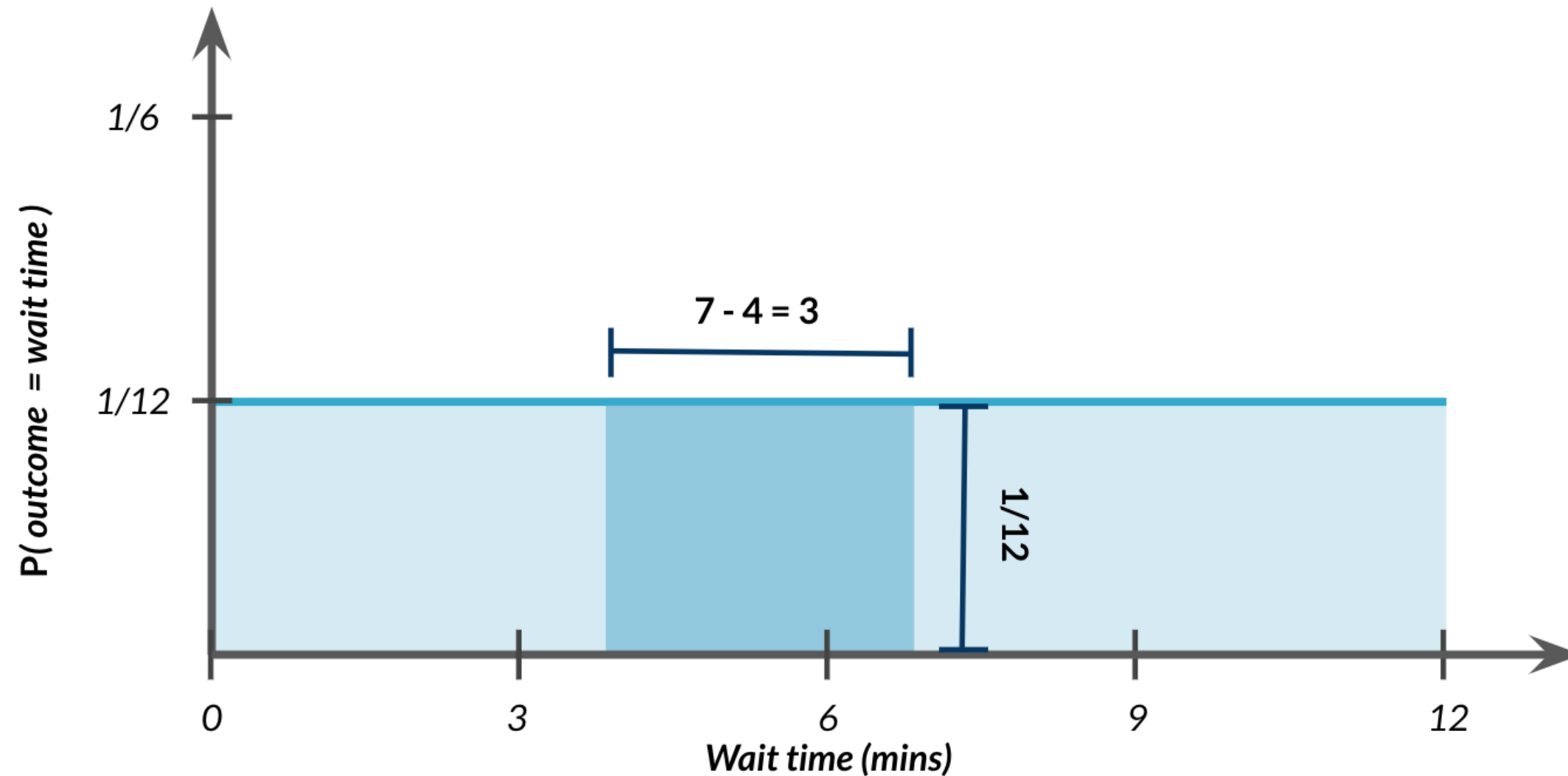
Probability still = area

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



Probability still = area

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



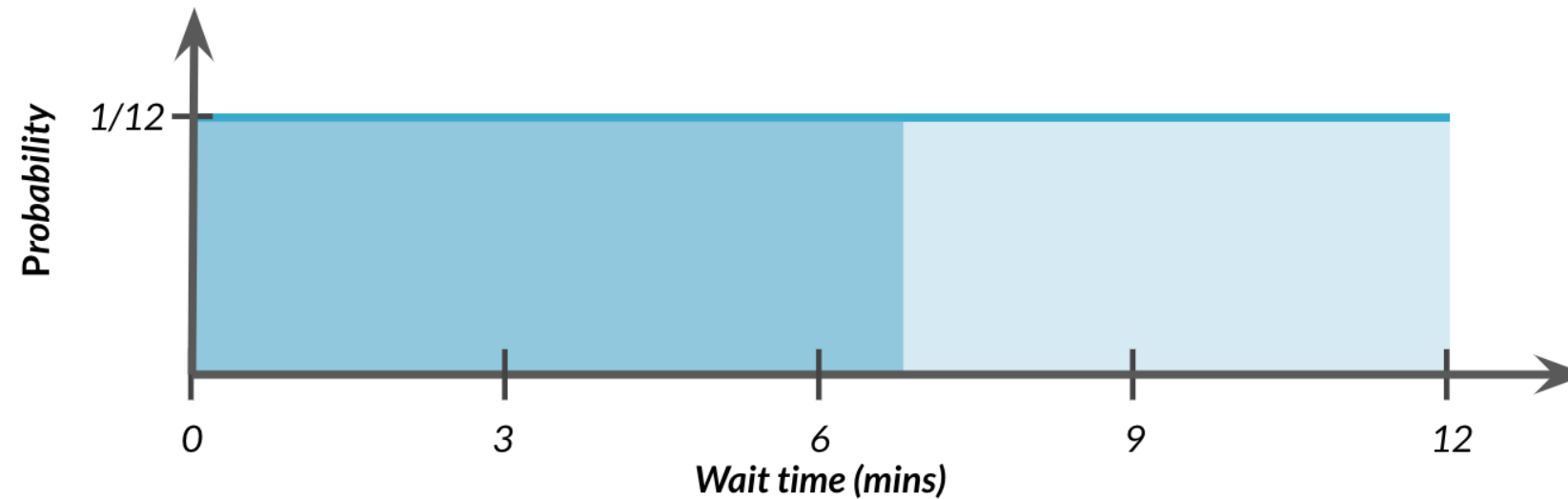
Probability still = area

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



Uniform distribution in R

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

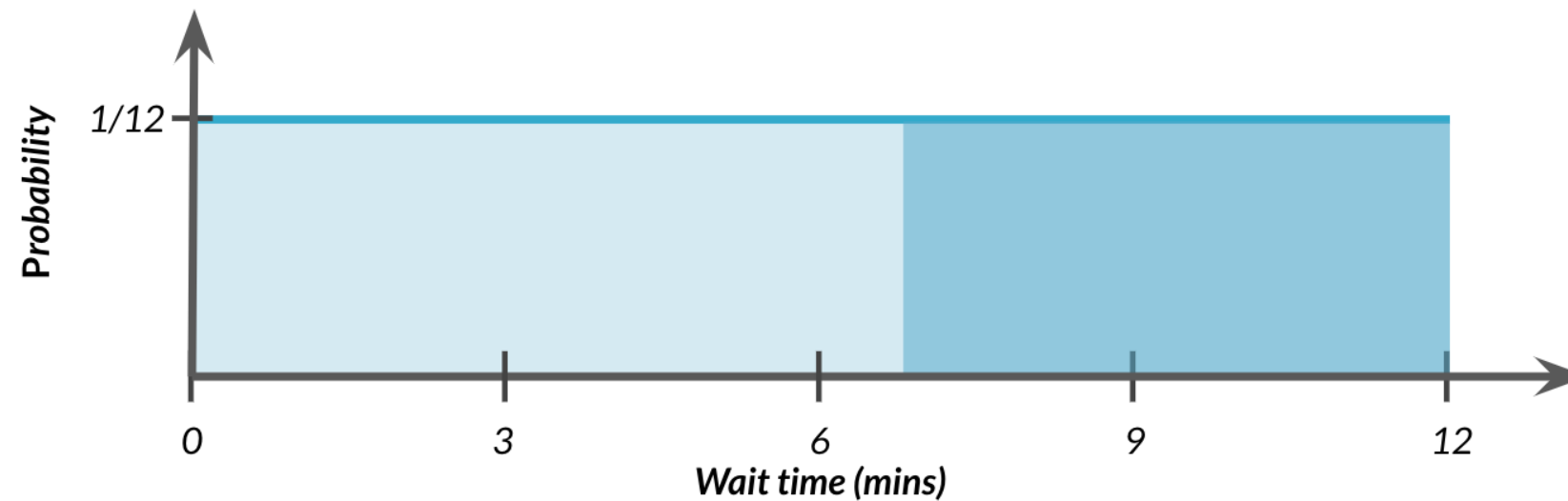


```
punif(7, min = 0, max = 12)
```

```
0.5833333
```

lower.tail

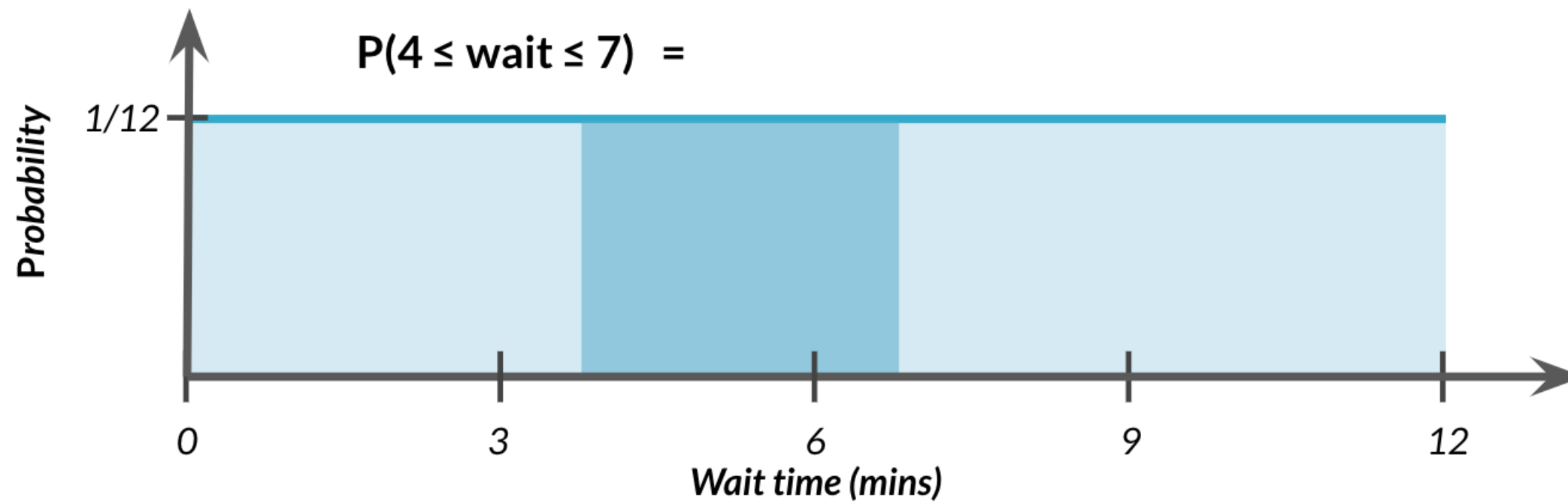
$$P(\text{wait time} \geq 7)$$



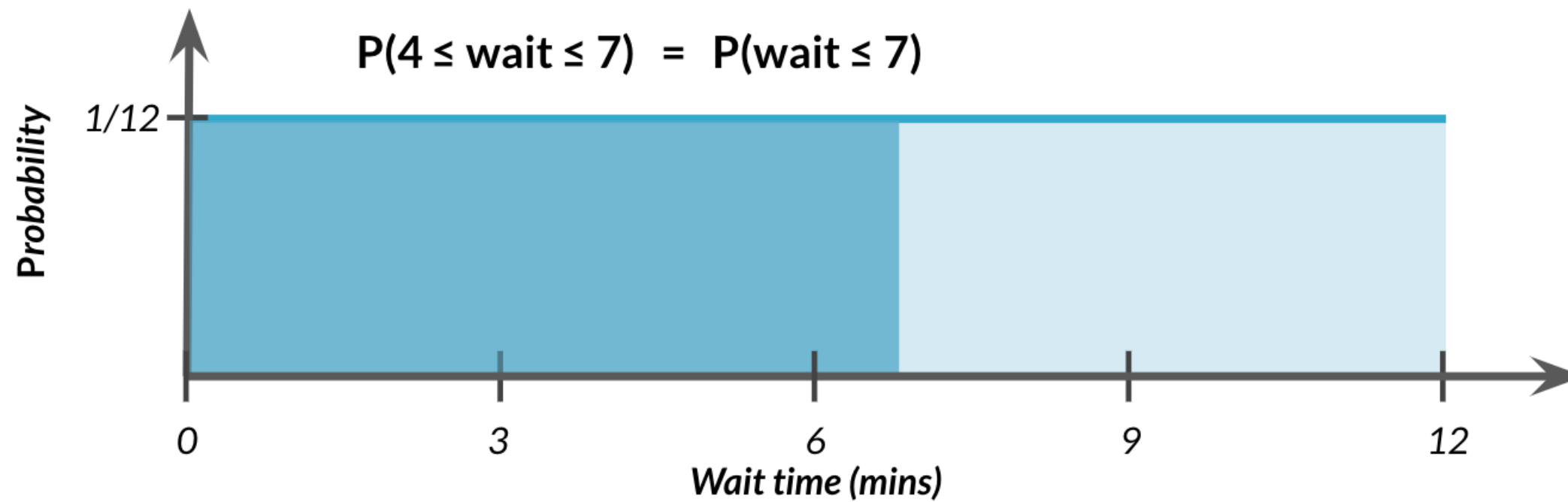
```
punif(7, min = 0, max = 12, lower.tail = FALSE)
```

```
0.4166667
```

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



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



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

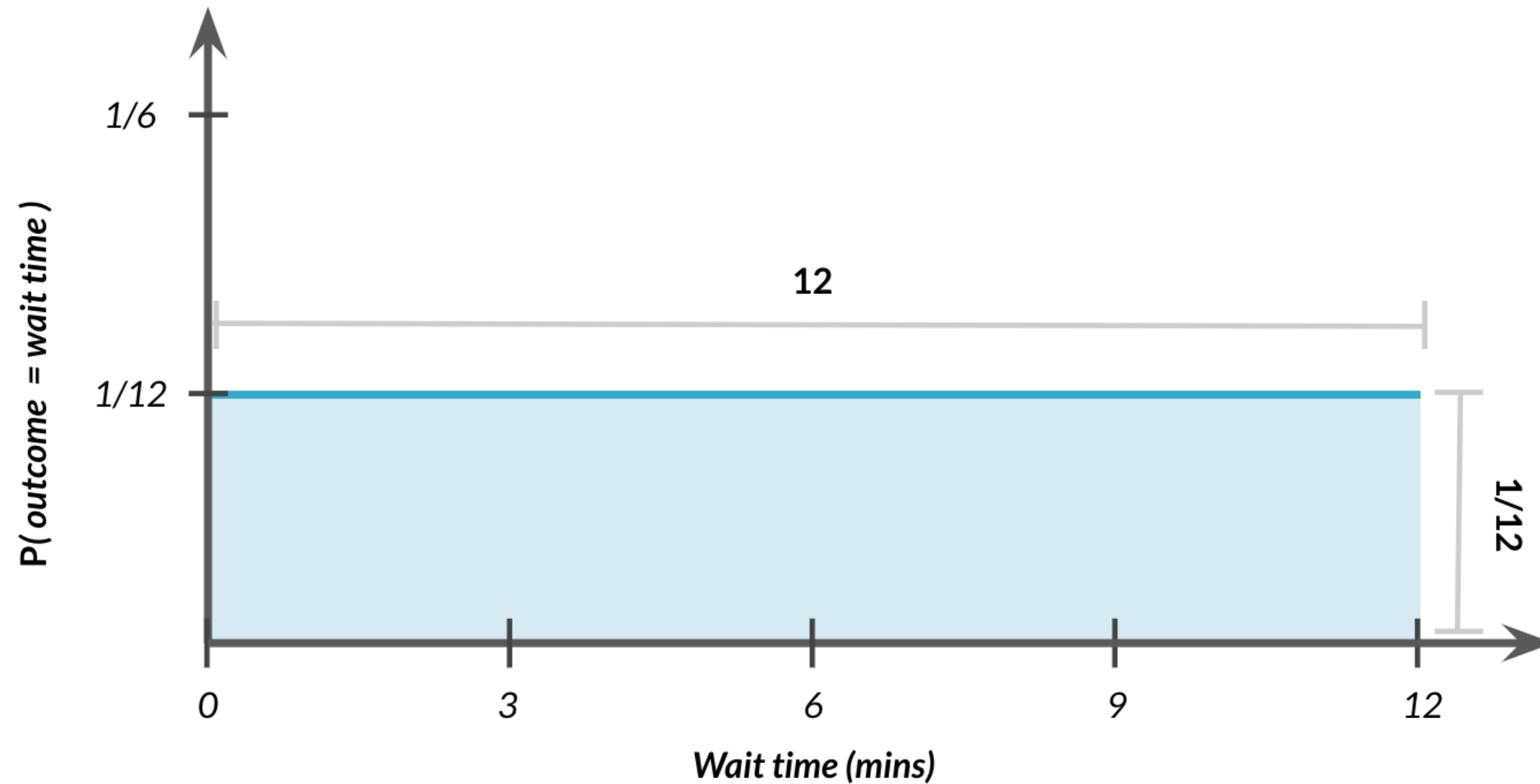


```
punif(7, min = 0, max = 12) - punif(4, min = 0, max = 12)
```

```
0.25
```

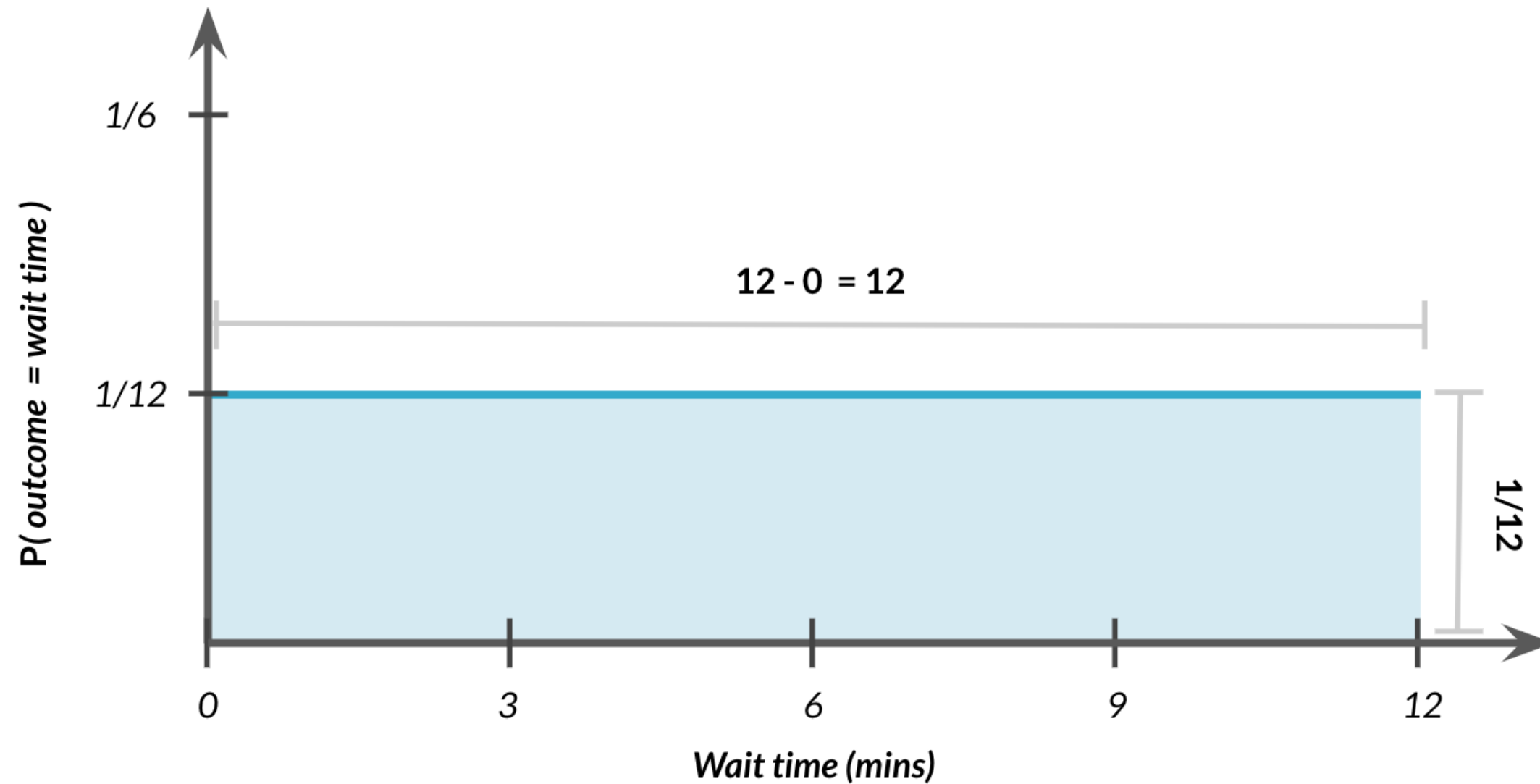
Total area = 1

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

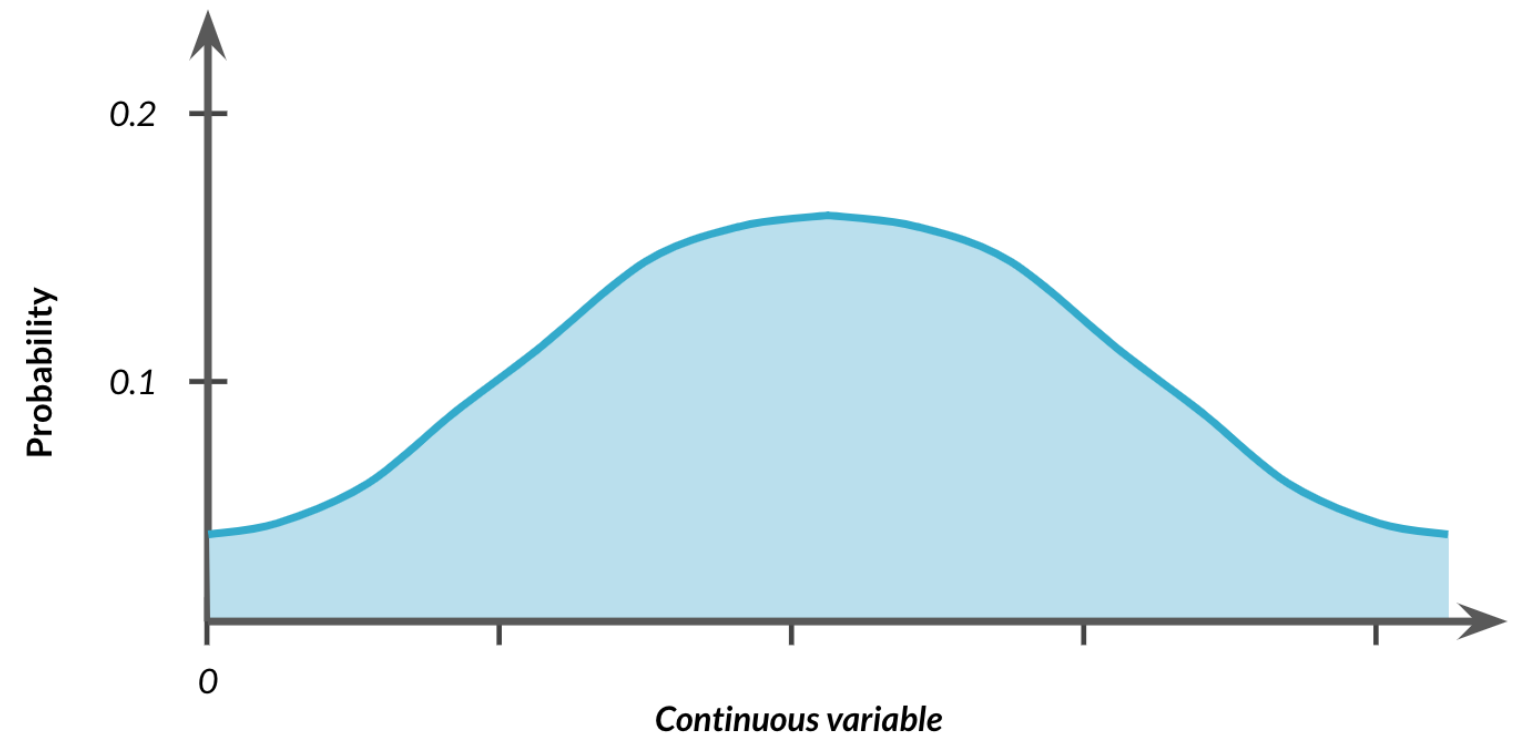
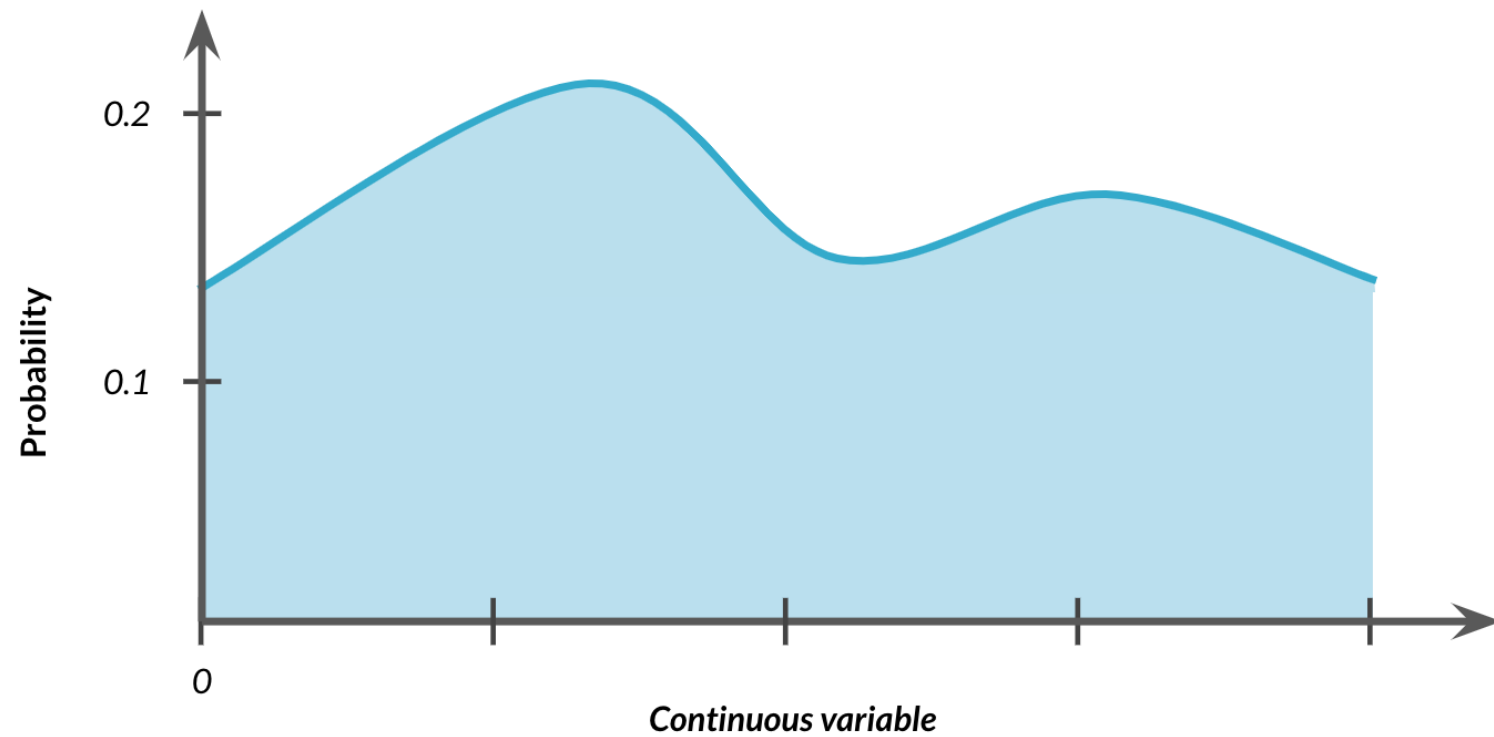


Total area = 1

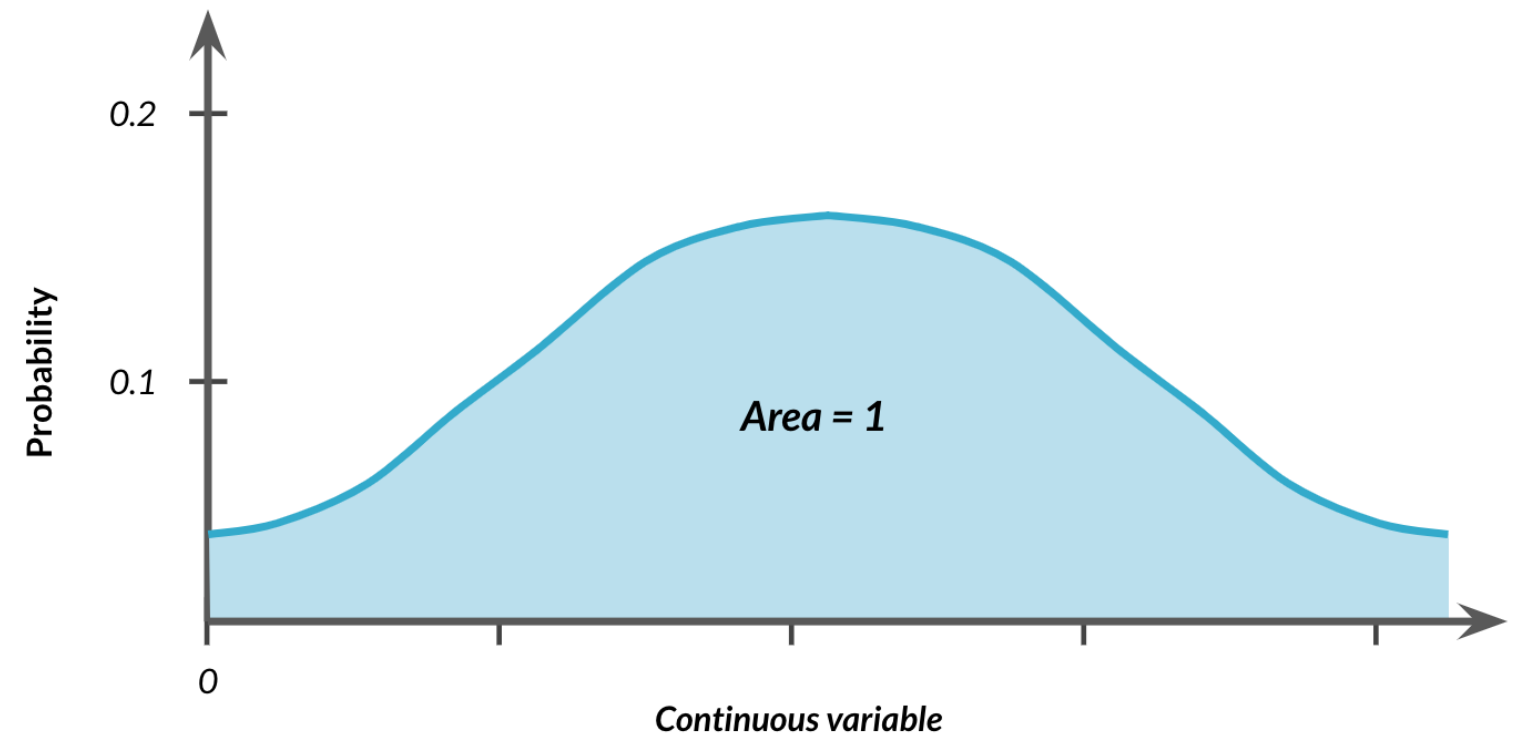
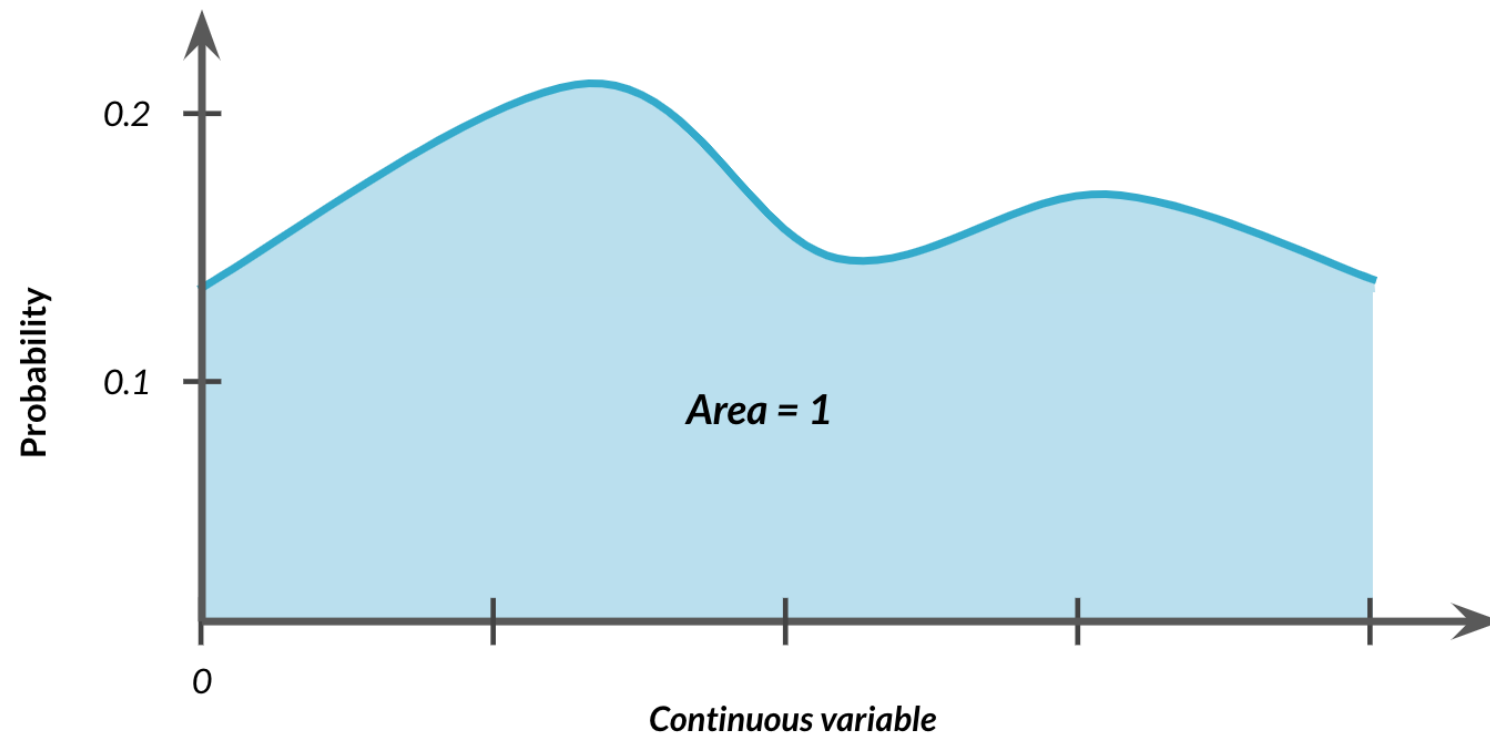
$$P(0 \leq \text{outcome} \leq 12) = 12 \times 1/12 = 1$$



Other continuous distributions

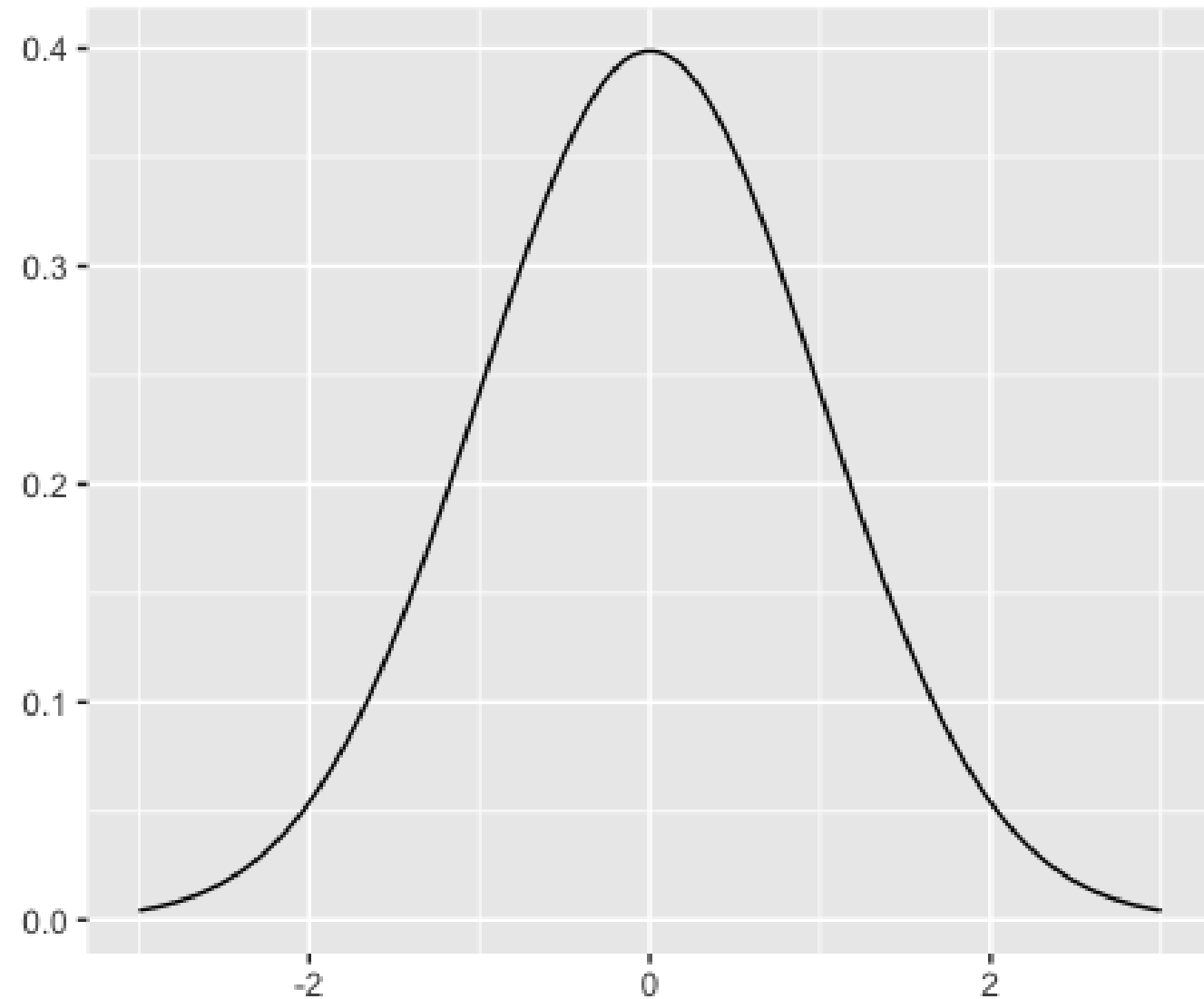


Other continuous distributions

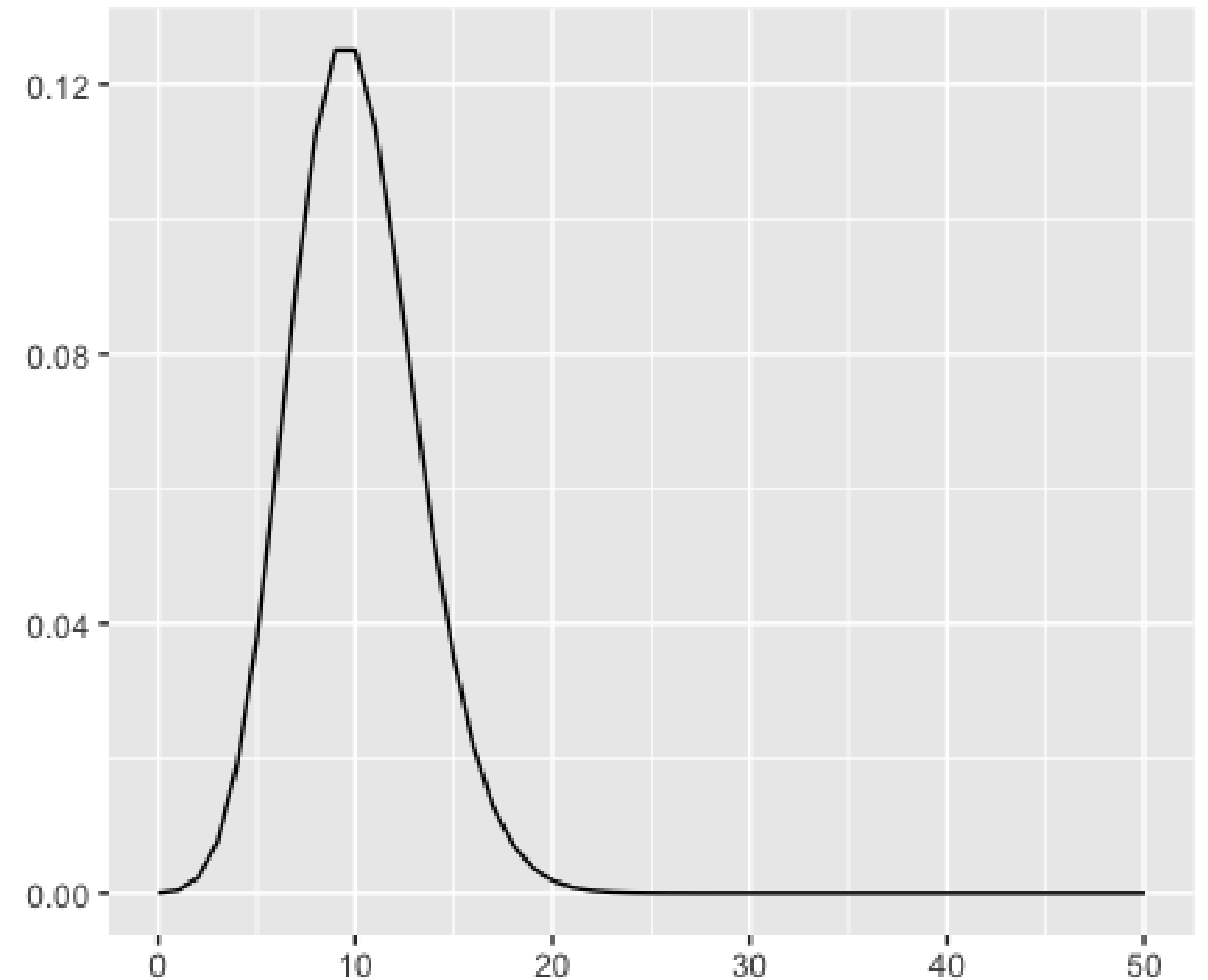


Other special types of distributions

Normal distribution



Poisson distribution



Let's practice!

INTRODUCTION TO STATISTICS IN R

The binomial distribution

INTRODUCTION TO STATISTICS IN R



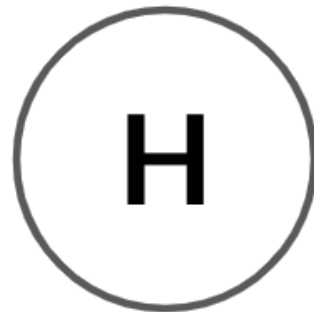
Maggie Matsui

Content Developer, DataCamp

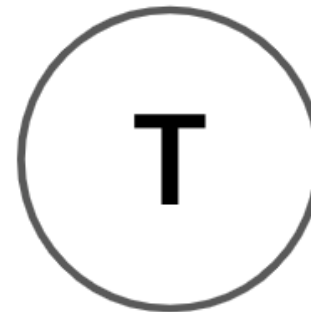
Coin flipping



50%



50%



Binary outcomes

H

T

1

0

Success

Failure

Win

Loss

A single flip

```
rbinom(# of trials, # of coins, # probability of heads/success)
```

1 = head, 0 = tails

```
rbinom(1, 1, 0.5)
```

```
1
```

```
rbinom(1, 1, 0.5)
```

```
0
```

One flip many times

```
rbinom(8, 1, 0.5)
```

```
1 0 0 1 0 0 1 0
```

```
rbinom(8, 1, 0.5)
```

8 flips of 1 coin with 50%
chance of success

Many flips one time

```
rbinom(1, 8, 0.5)
```

```
3
```

```
rbinom(1, 8, 0.5)
```

1 flip of 8 coins with 50%
chance of success

Many flips many times

```
rbinom(10, 3, 0.5)
```

```
2 0 1 0 1 1 3 3 3 1
```

```
rbinom(10, 3, 0.5)
```

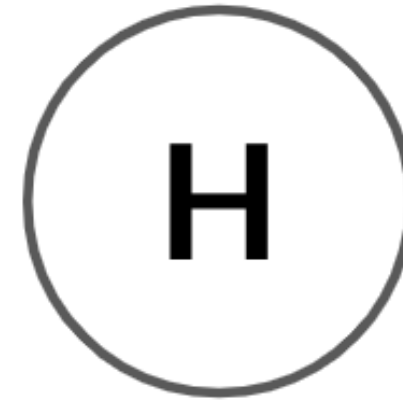
10 flips of 3 coins with 50%
chance of success

Other probabilities

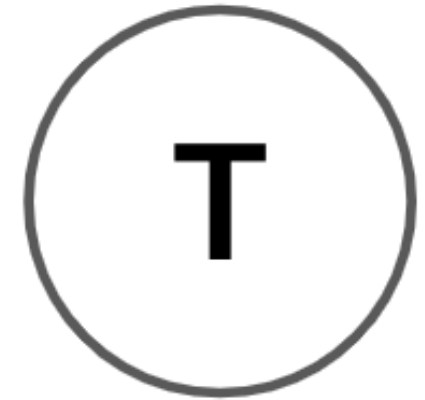
```
rbinom(10, 3, 0.25)
```

```
1 1 0 0 1 1 1 1 2 1
```

25%



75%



Binomial distribution

Probability distribution of the number of successes in a sequence of independent trials

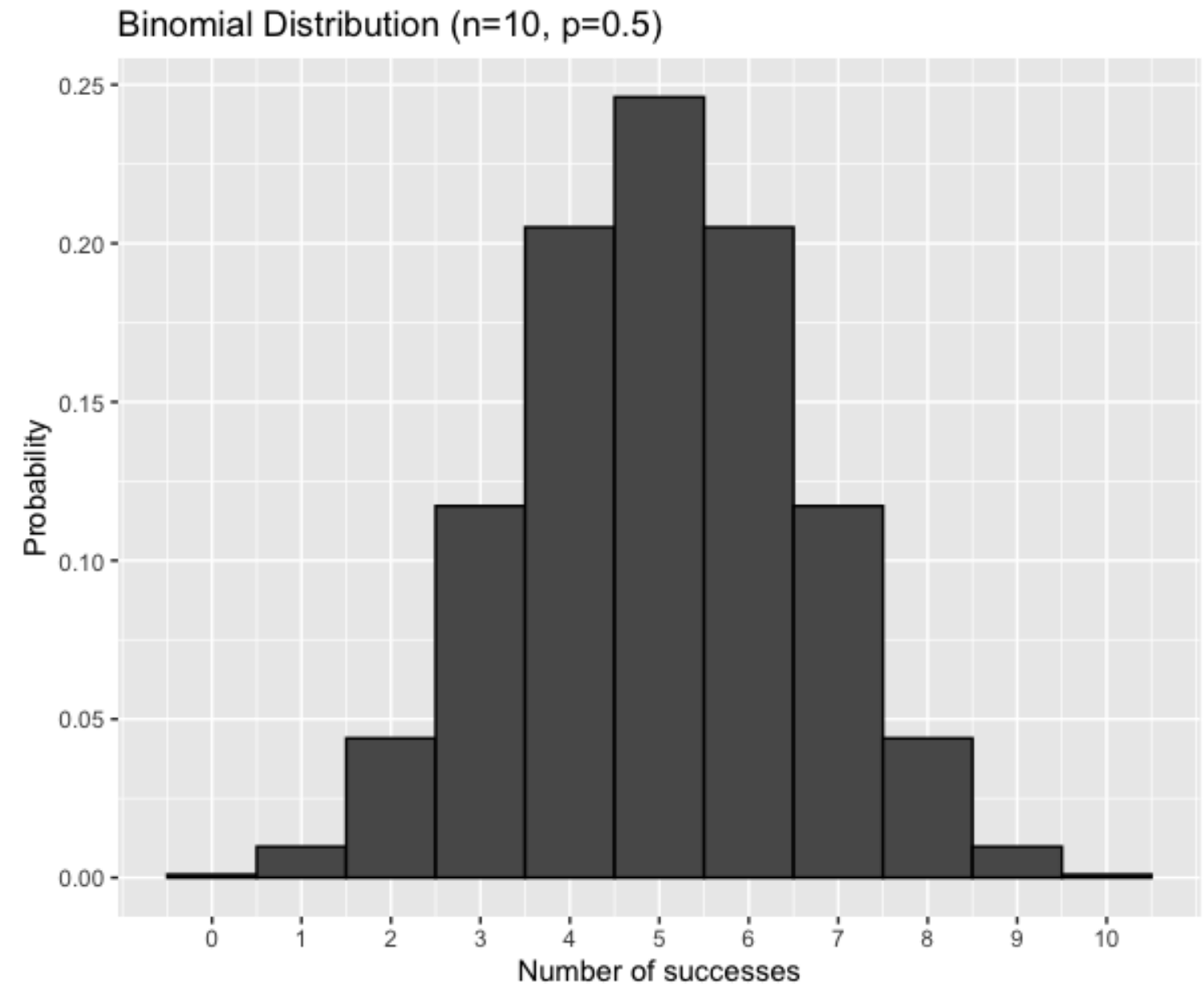
E.g. Number of heads in a sequence of coin flips

Described by n and p

- n : total number of trials
- p : probability of success

`rbinom(3, 10, 0.5)`

(Note: The original image contains a diagram with colored boxes and lines mapping the values 3, 10, and 0.5 to the parameters n and p . The value 3 is in a pink box, 10 is in a blue box, and 0.5 is in a yellow box. Lines connect these boxes to the n and p labels above them.)



What's the probability of 7 heads?

$P(\text{heads} = 7)$

```
# dbinom(num heads, num trials, prob of heads)
dbinom(7, 10, 0.5)
```

```
0.1171875
```

What's the probability of 7 or fewer heads?

$P(\text{heads} \leq 7)$

```
pbinom(7, 10, 0.5)
```

```
0.9453125
```

What's the probability of more than 7 heads?

$P(\text{heads} > 7)$

```
pbinom(7, 10, 0.5, lower.tail = FALSE)
```

```
0.0546875
```

```
1 - pbinom(7, 10, 0.5)
```

```
0.0546875
```

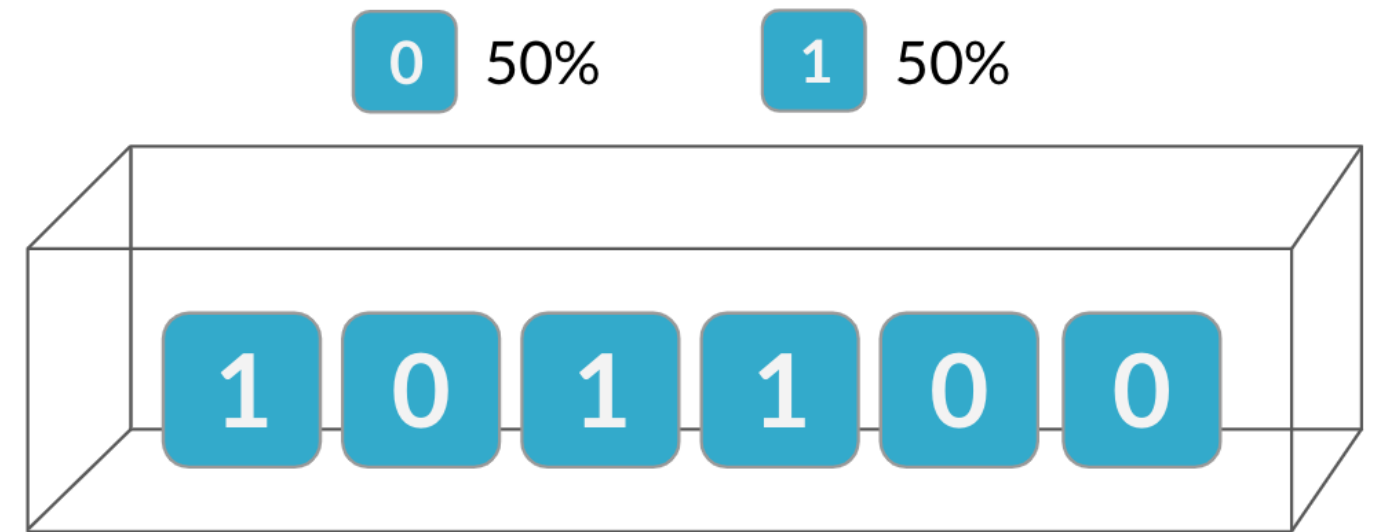
Expected value

$$\text{Expected value} = n \times p$$

$$\text{Expected number of heads out of 10 flips} = 10 \times 0.5 = 5$$

Independence

*The binomial distribution is a probability distribution of the number of successes in a sequence of **independent** trials*

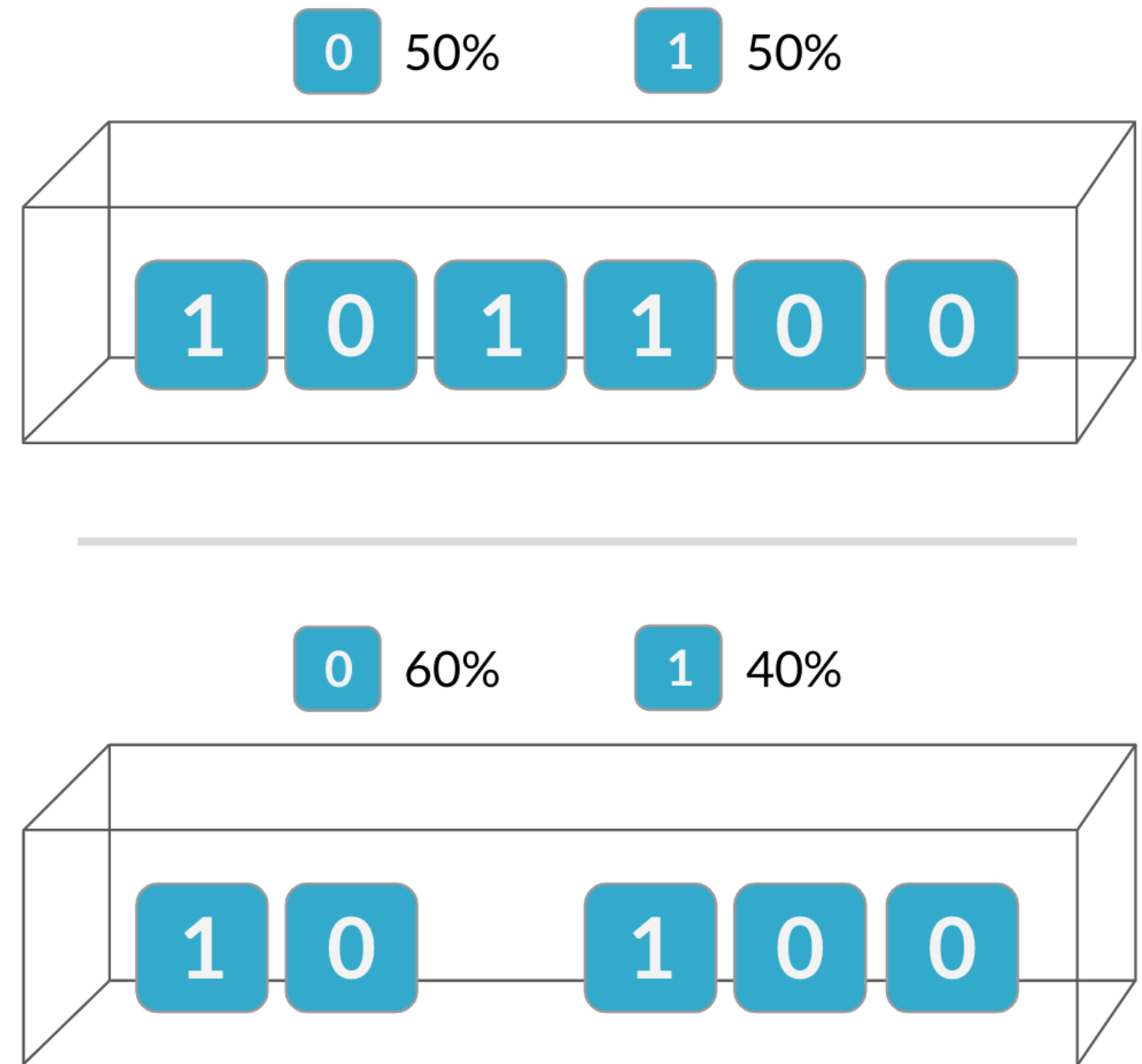


Independence

*The binomial distribution is a probability distribution of the number of successes in a sequence of **independent** trials*

Probabilities of second trial are altered due to outcome of the first

If trials are not independent, the binomial distribution does not apply!



Let's practice!

INTRODUCTION TO STATISTICS IN R