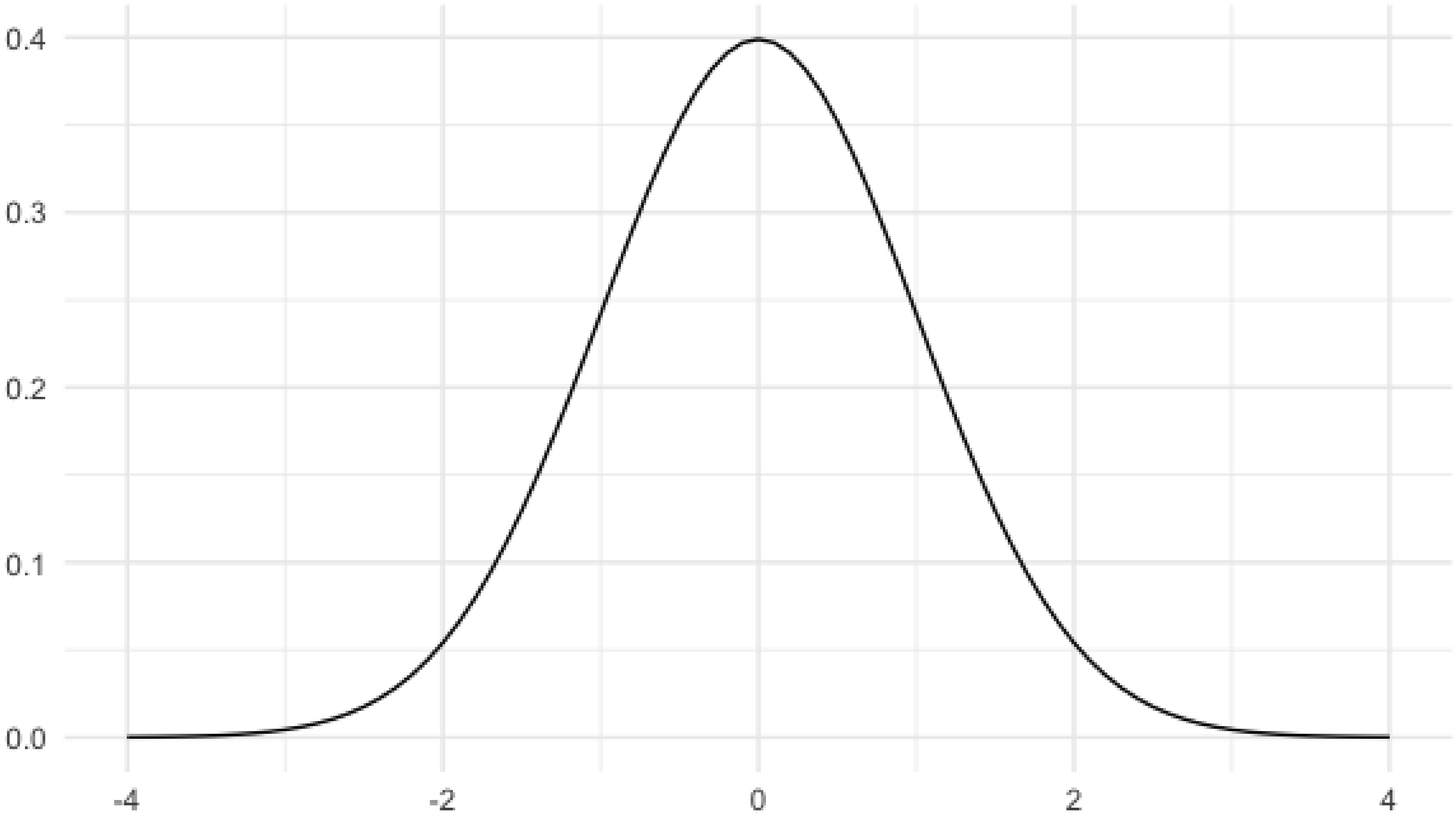
The normal distribution

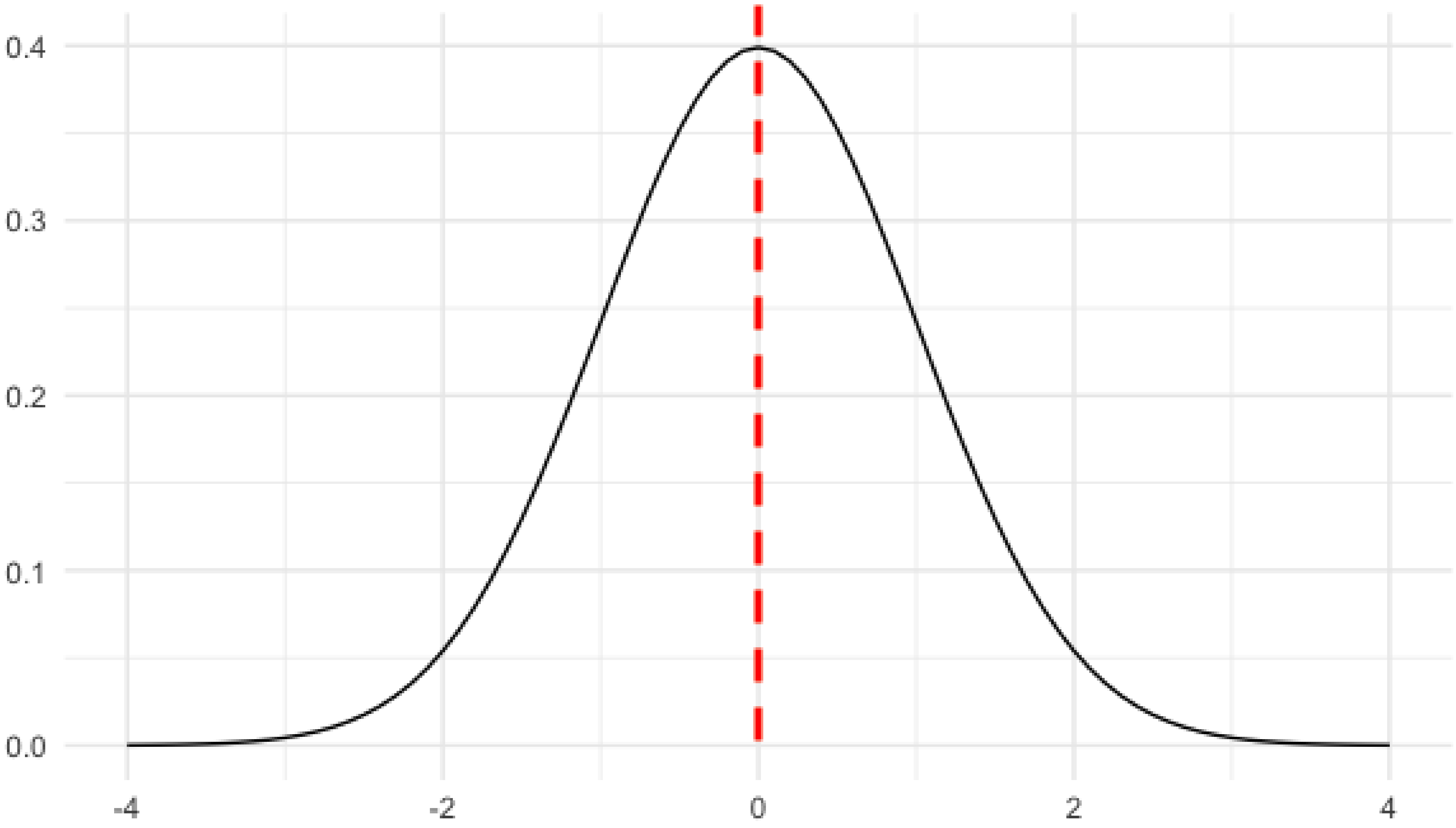
INTRODUCTION TO STATISTICS IN R

Maggie Matsui
Content Developer, DataCamp

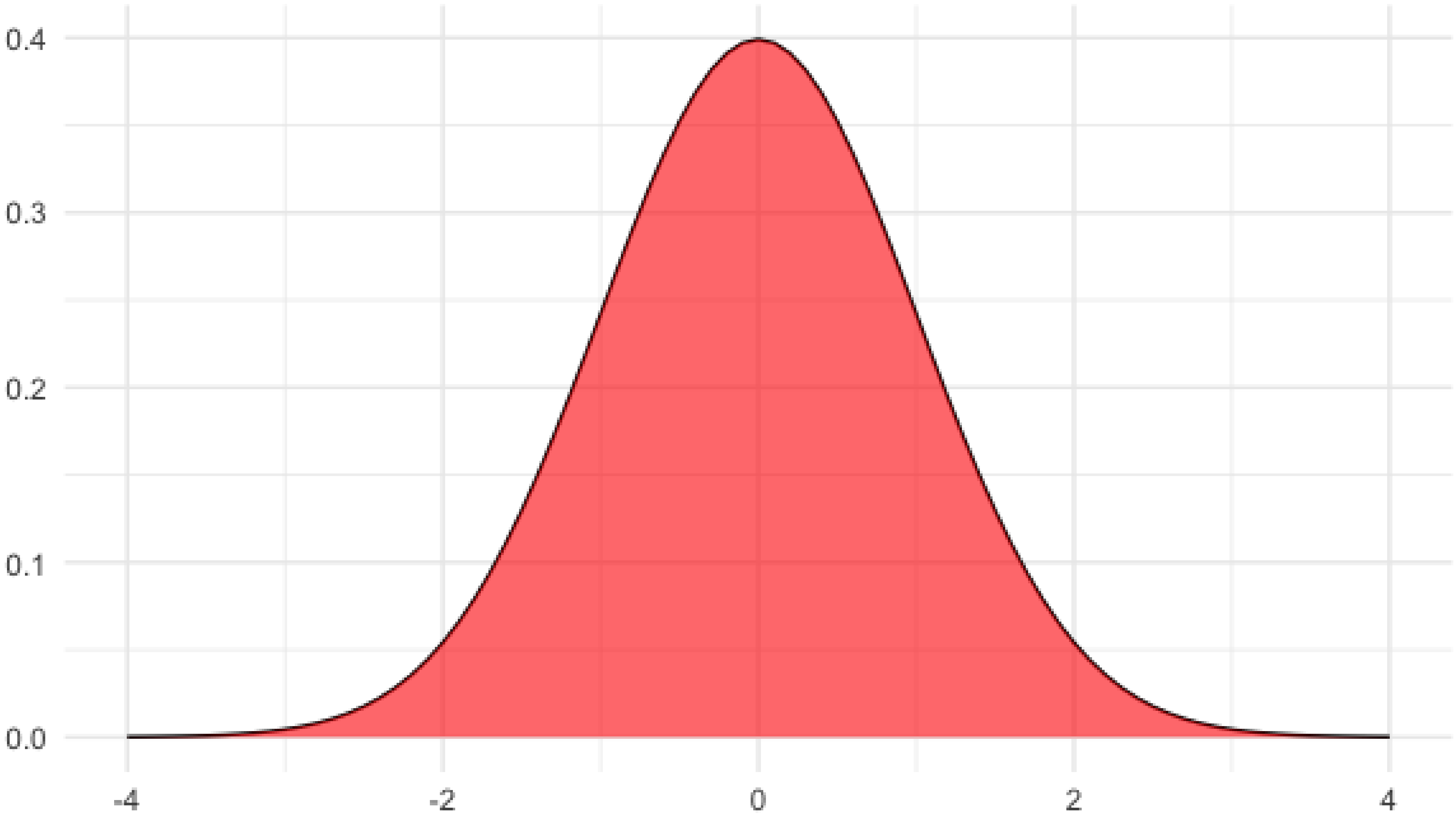
What is the normal distribution?



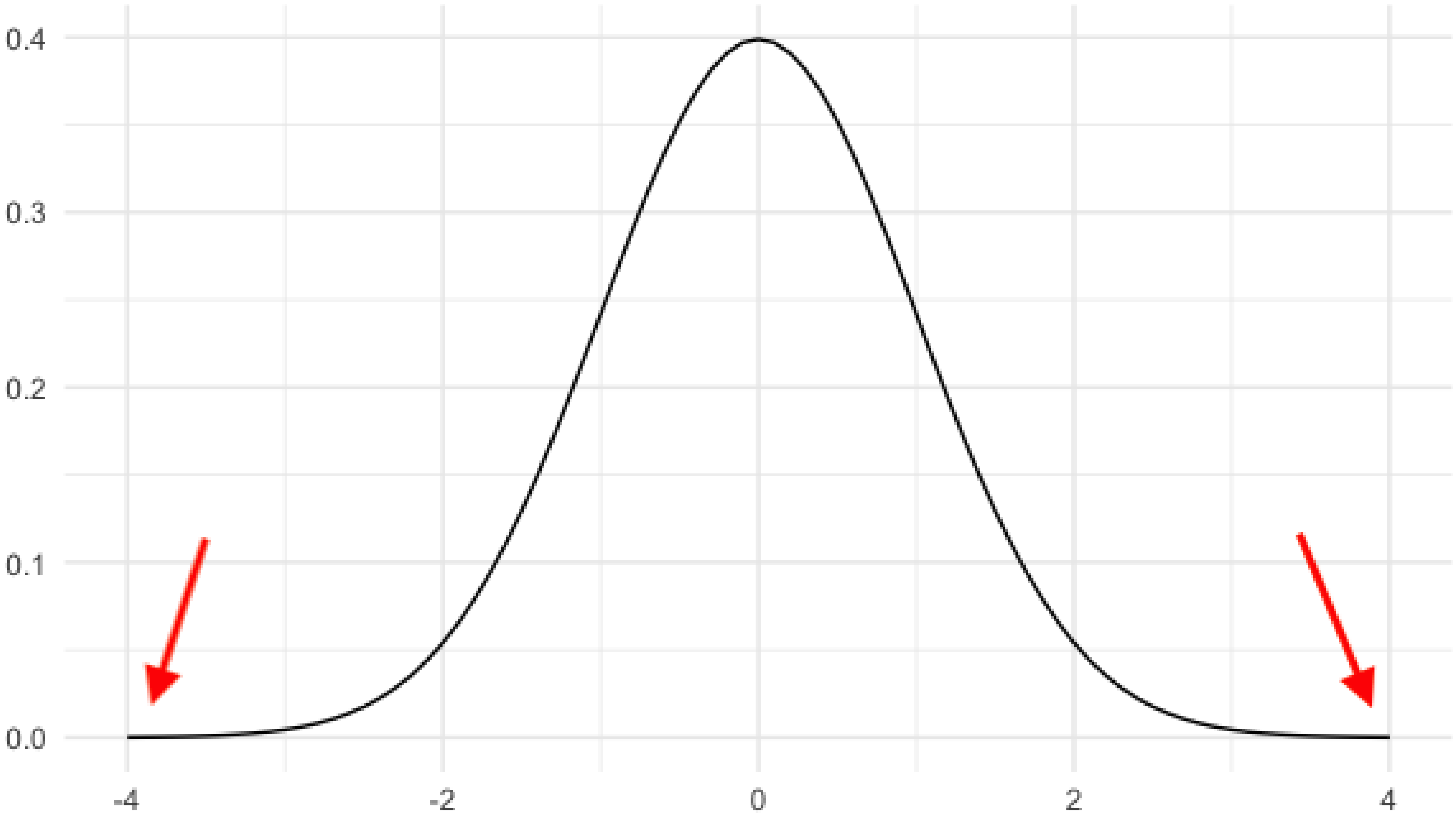
Symmetrical



Area = 1



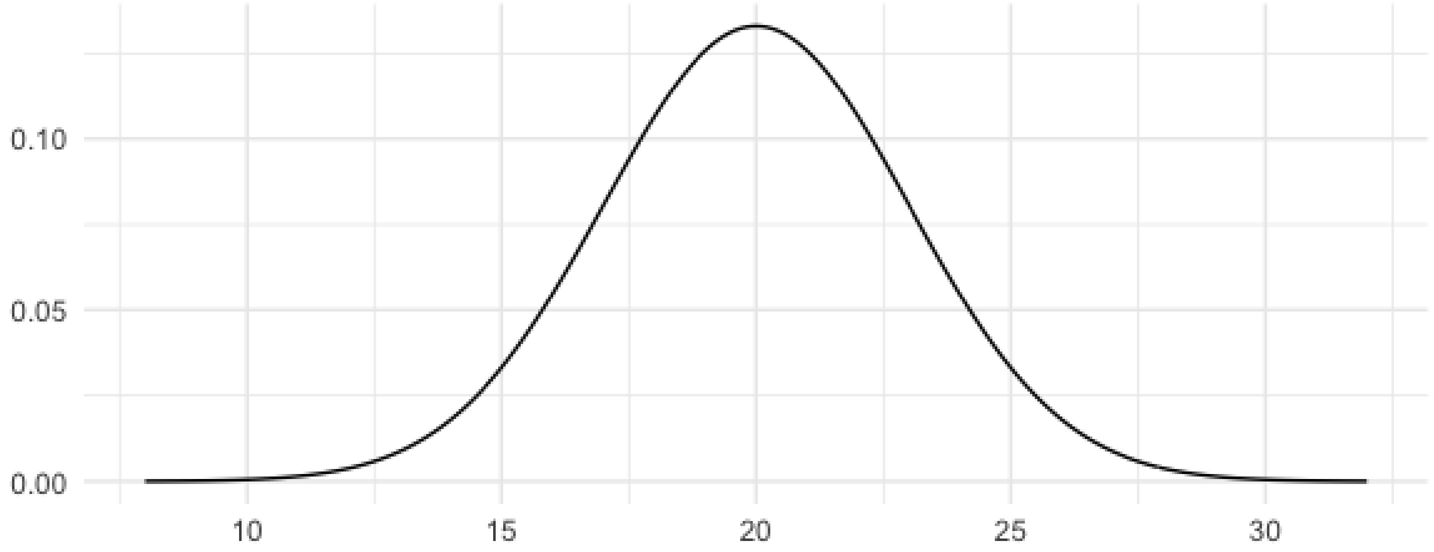
Curve never hits 0



Described by mean and standard deviation

Mean: 20

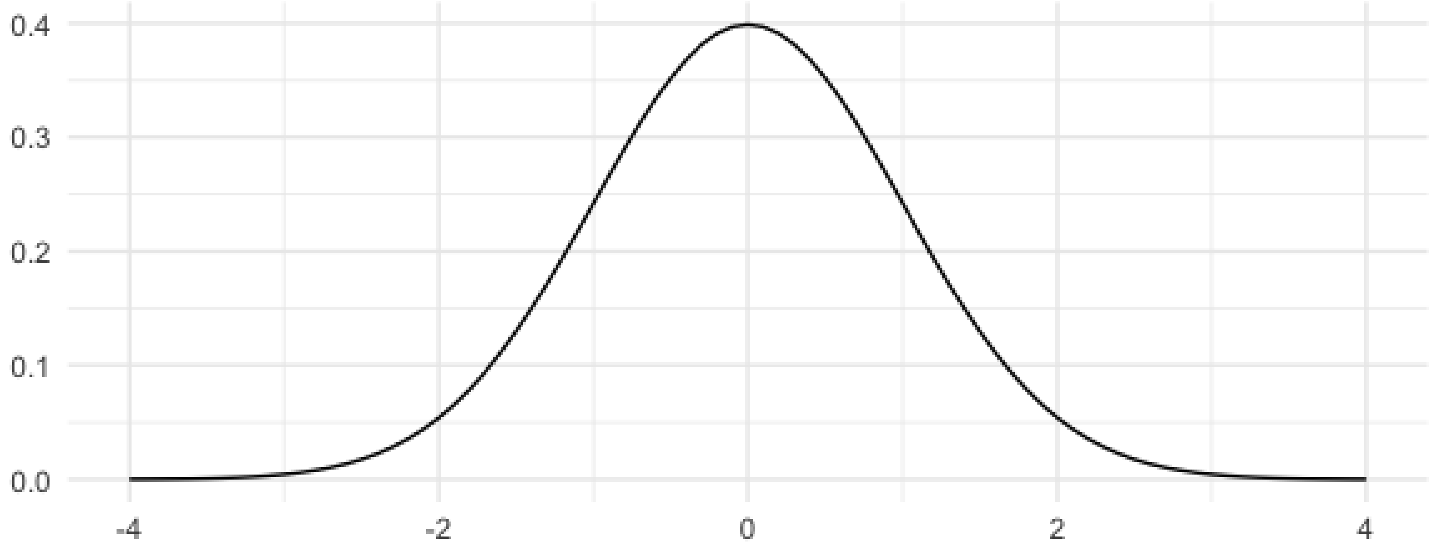
Standard deviation: 3



Standard normal distribution

Mean: 0

Standard deviation: 1



Described by mean and standard deviation

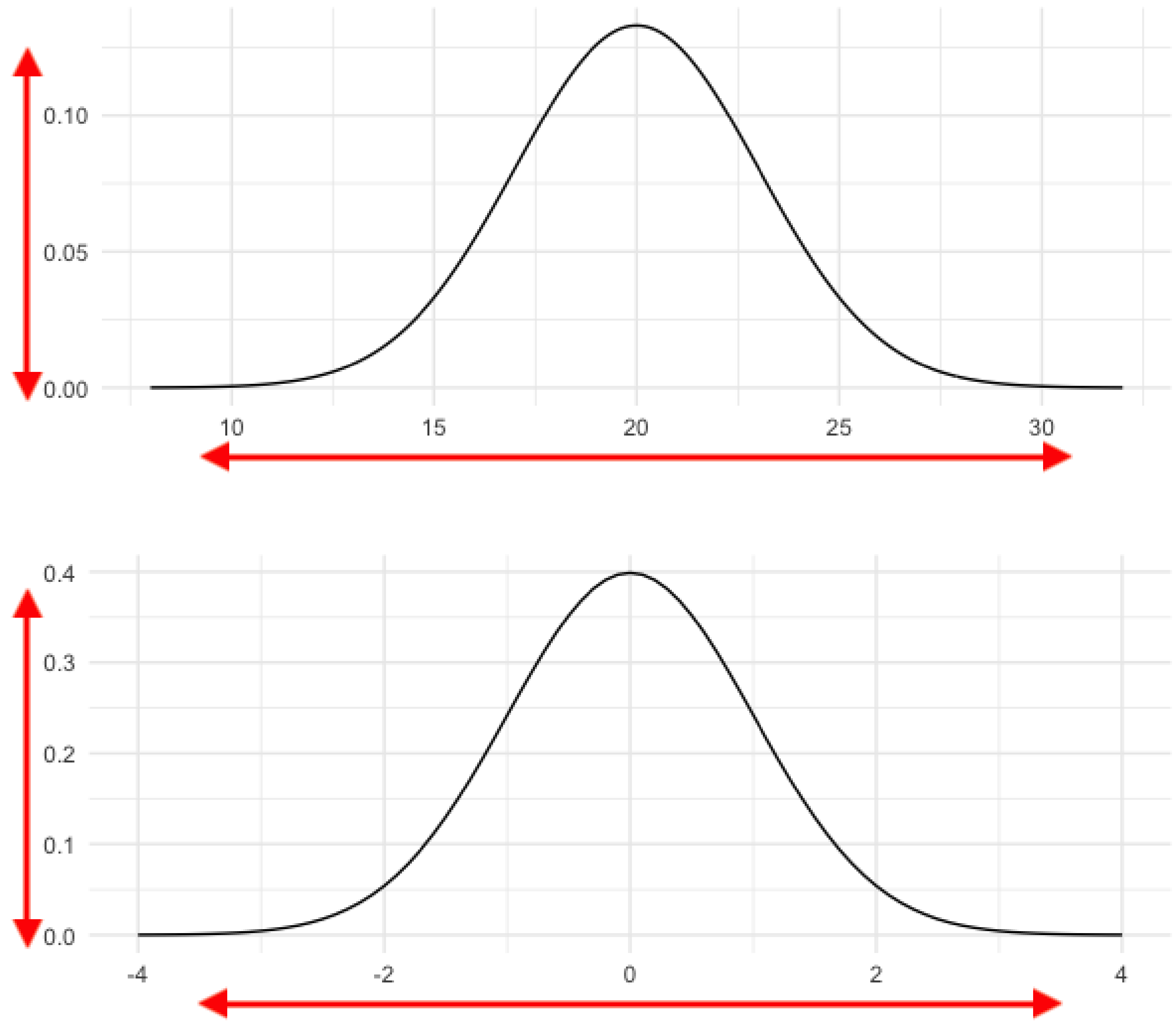
Mean: 20

Standard deviation: 3

Standard normal distribution

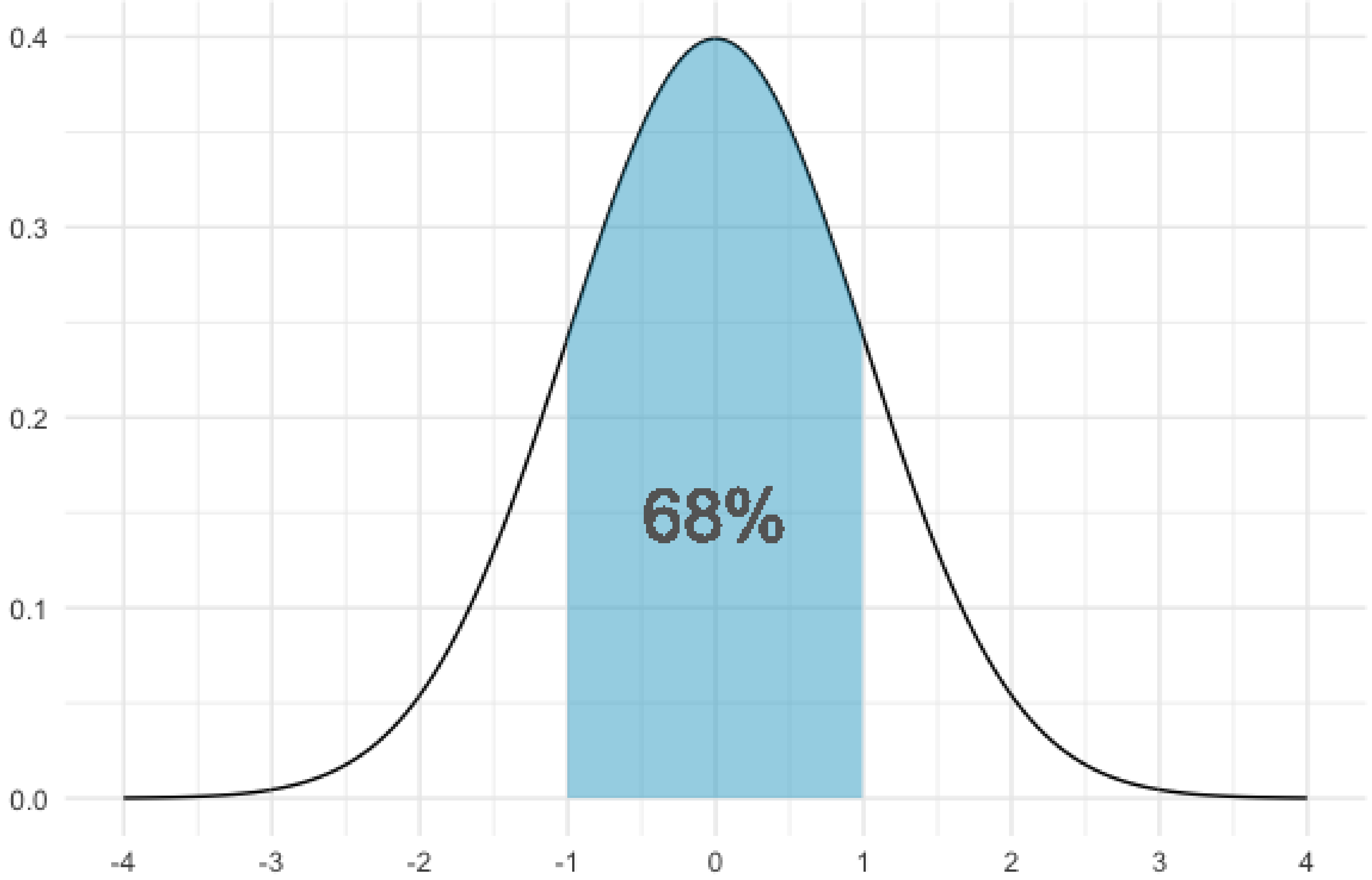
Mean: 0

Standard deviation: 1



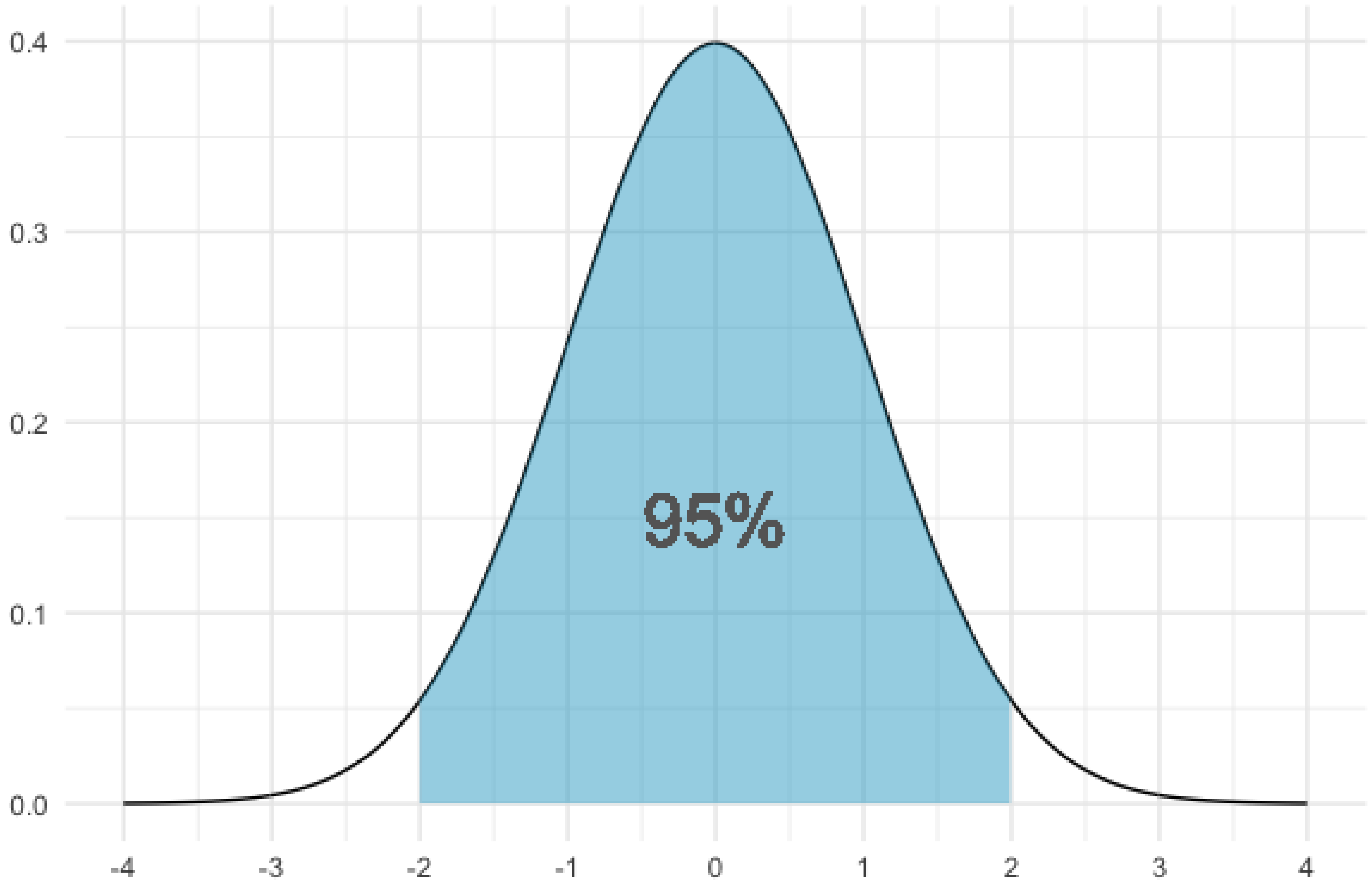
Areas under the normal distribution

68% falls within 1 standard deviation



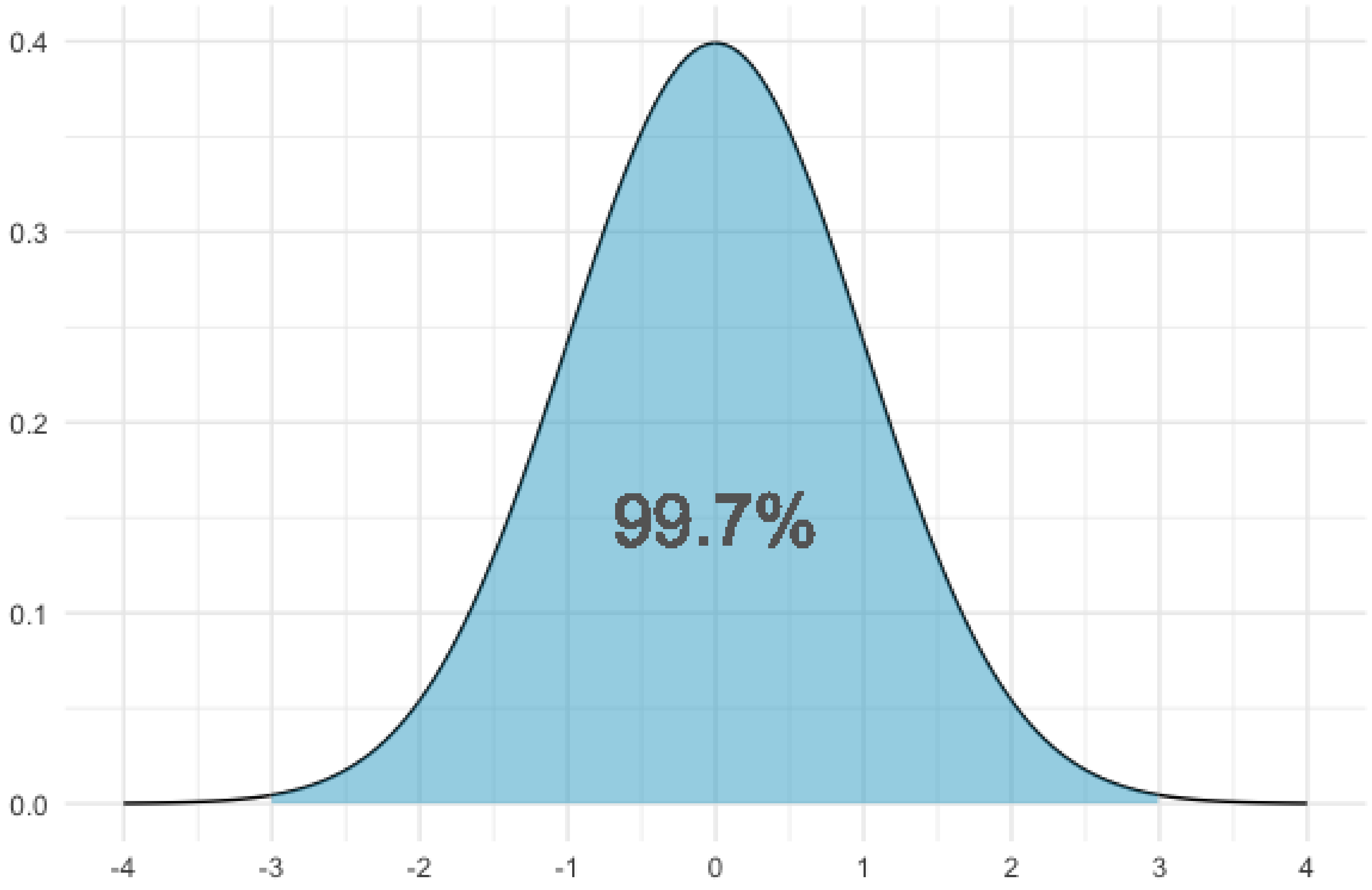
Areas under the normal distribution

95% falls within 2 standard deviations



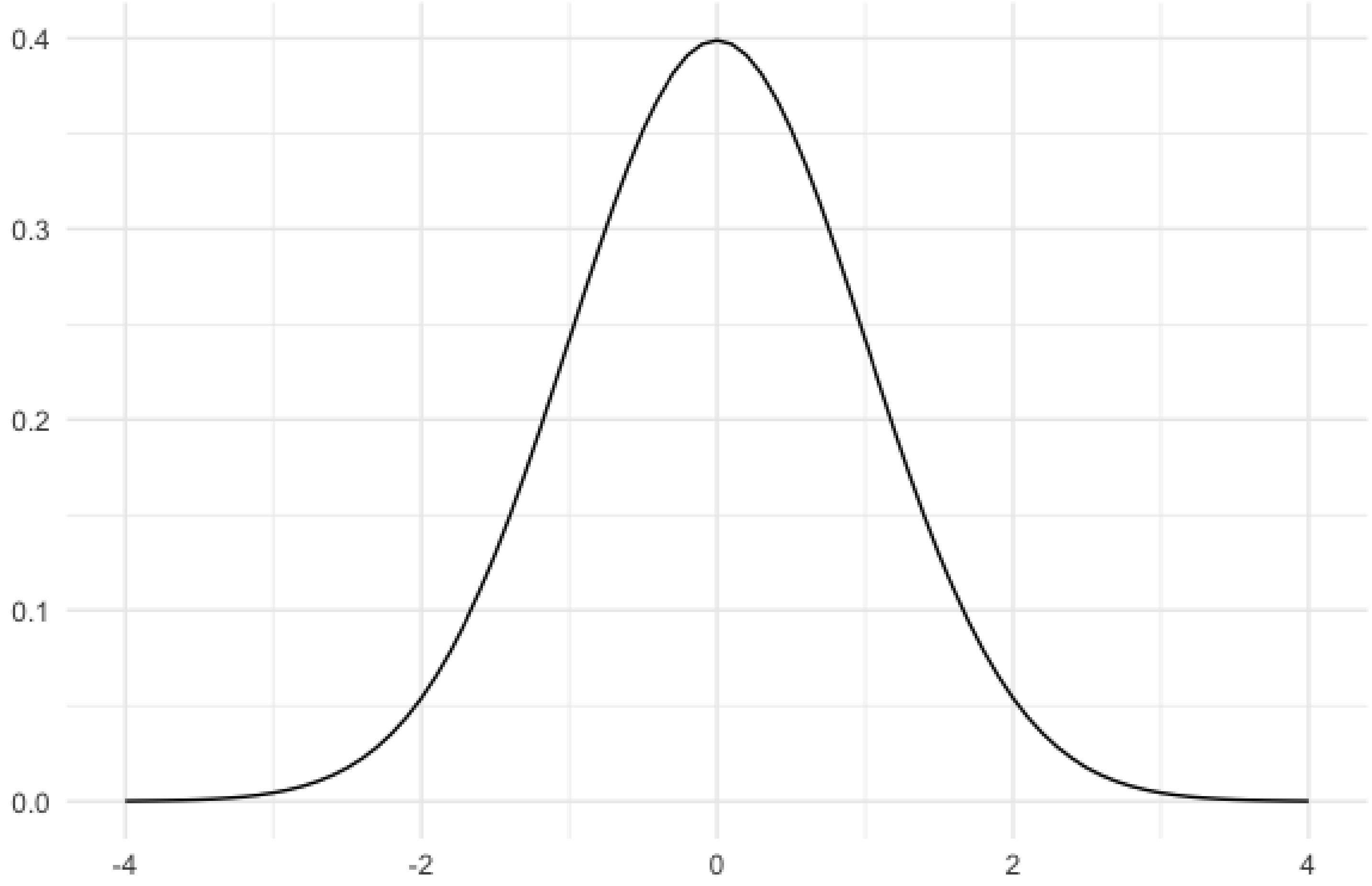
Areas under the normal distribution

99.7% falls within 3 standard deviations

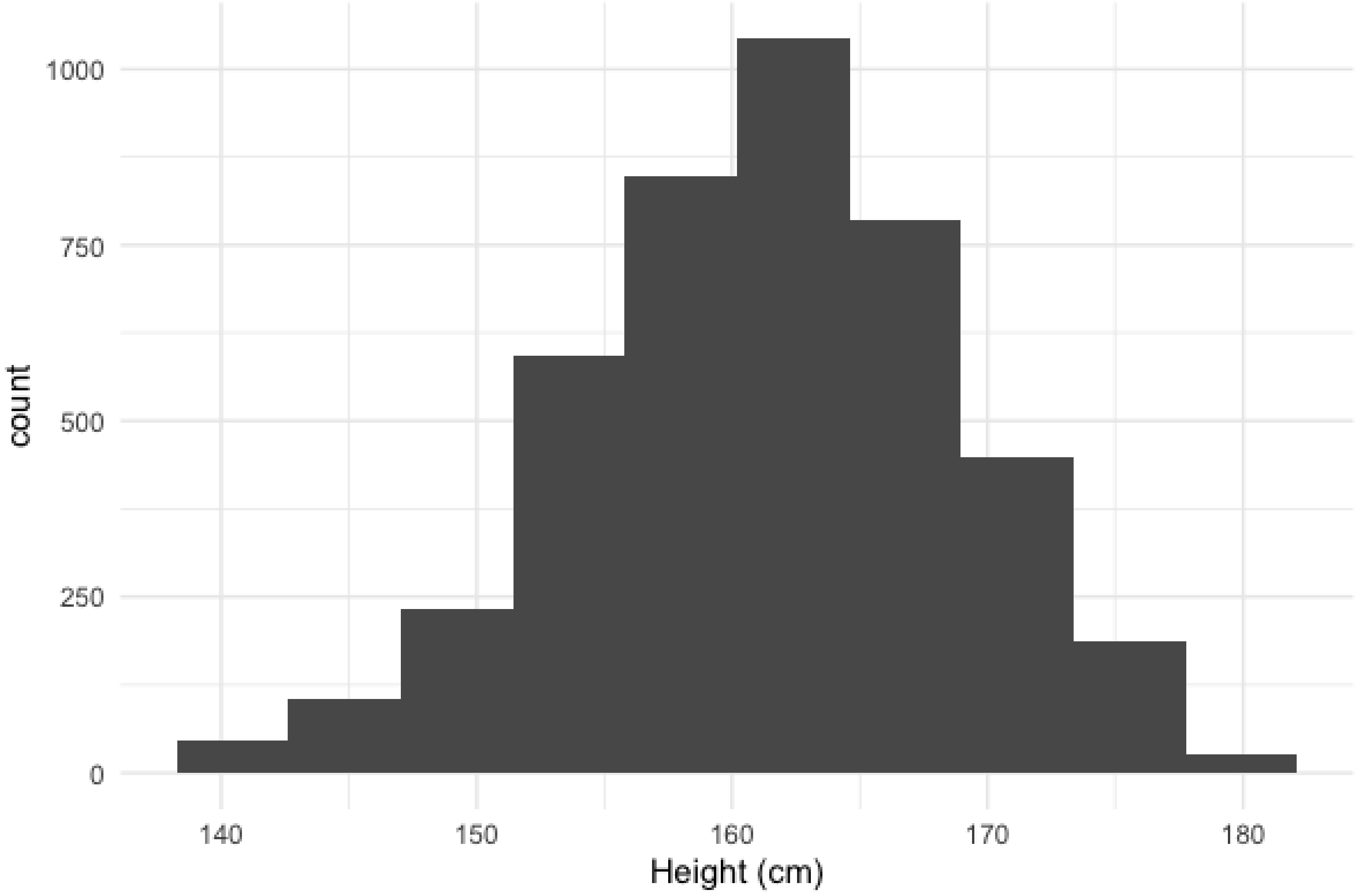


Lots of histograms look normal

Normal distribution

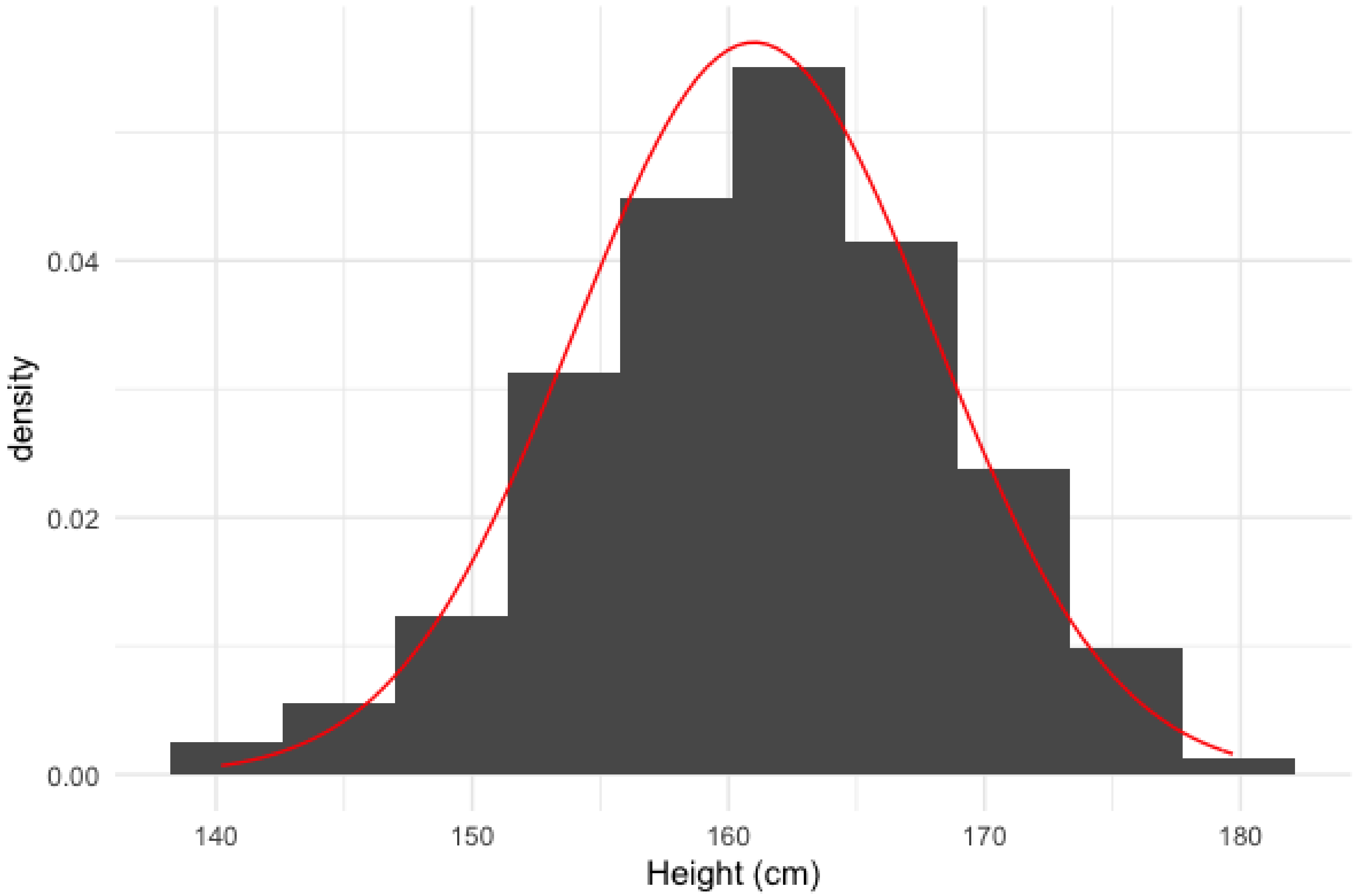


Women's heights from NHANES

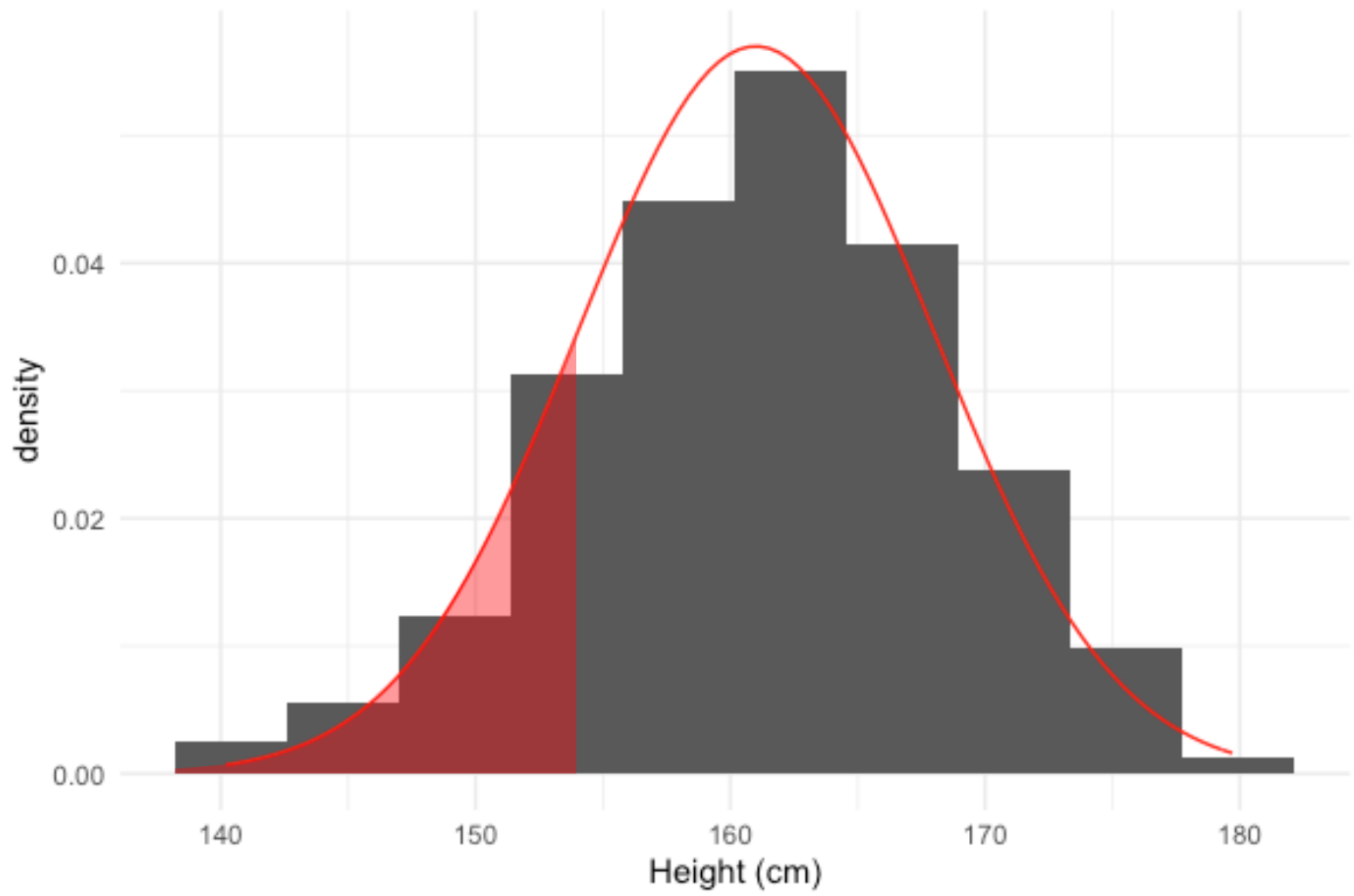


Mean: 161 cm ***Standard deviation: 7 cm***

Approximating data with the normal distribution



What percent of women are shorter than 154 cm?

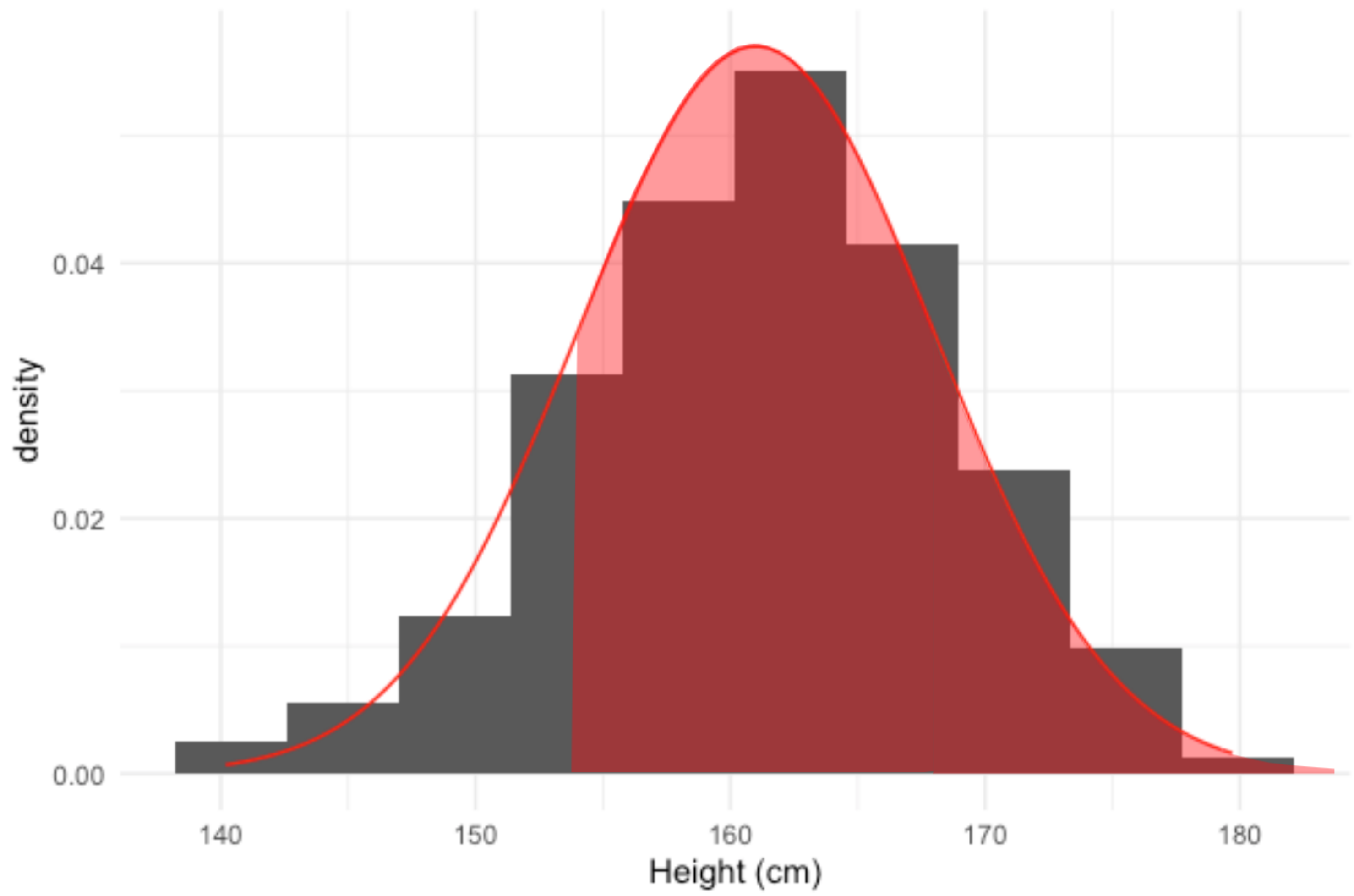



```r
pnorm(154, mean = 161, sd = 7)
```

```
0.159
```

16% of women in the survey are shorter than 154 cm

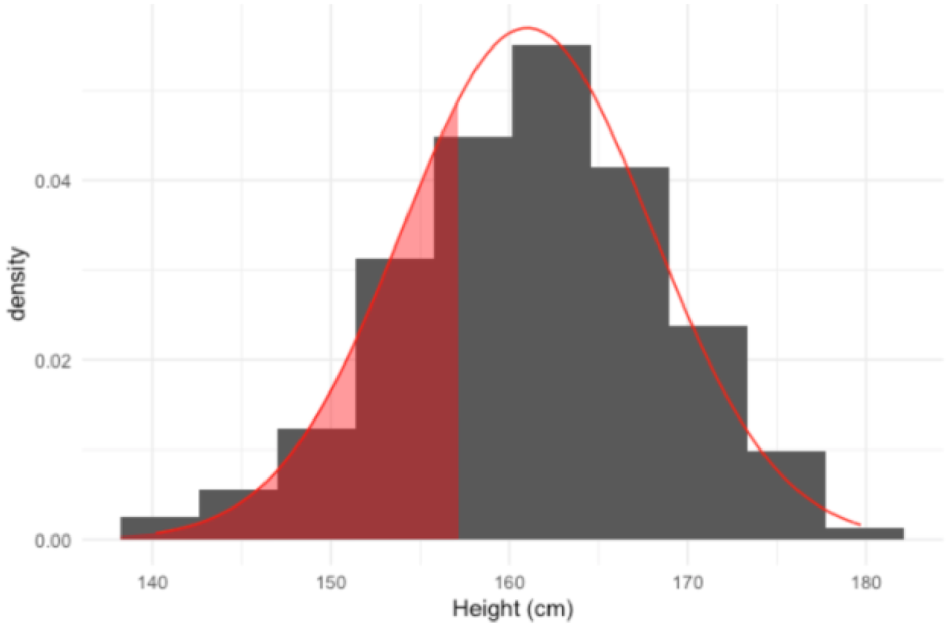
What percent of women are taller than 154 cm?



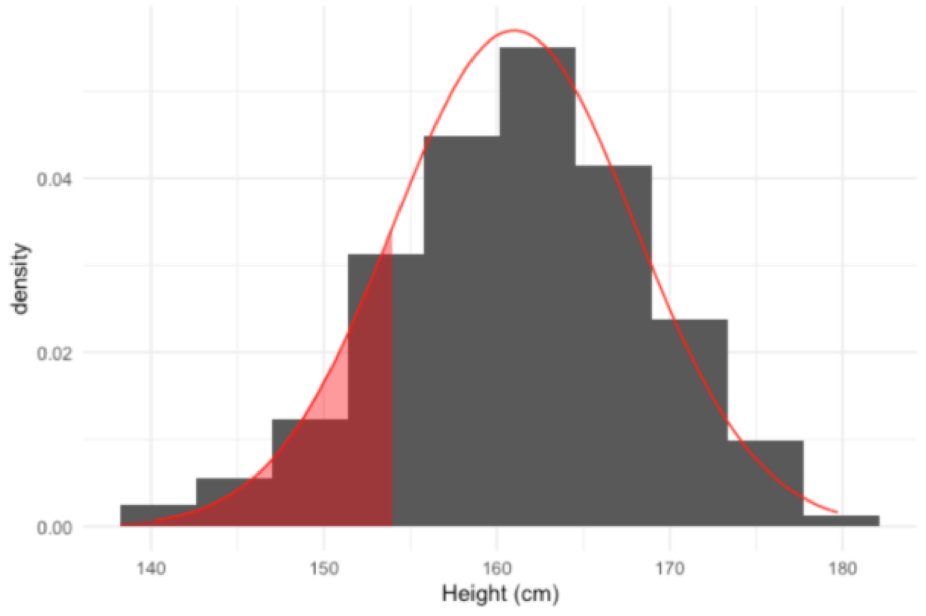

```r
pnorm(154, mean = 161, sd = 7,
      lower.tail = FALSE)
```

```
0.8413447
```

What percent of women are 154-157 cm?

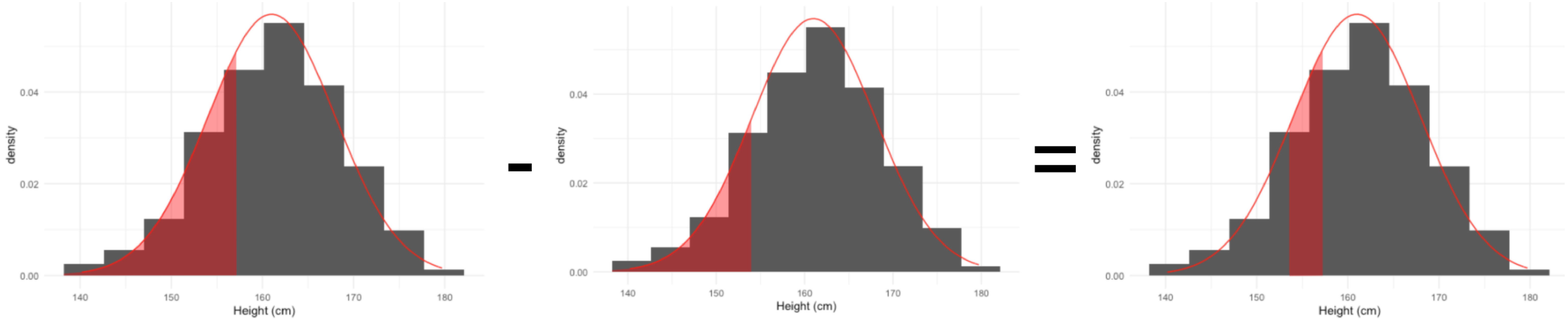


-




```
pnorm(157, mean = 161, sd = 7) - pnorm(154, mean = 161, sd = 7)
```

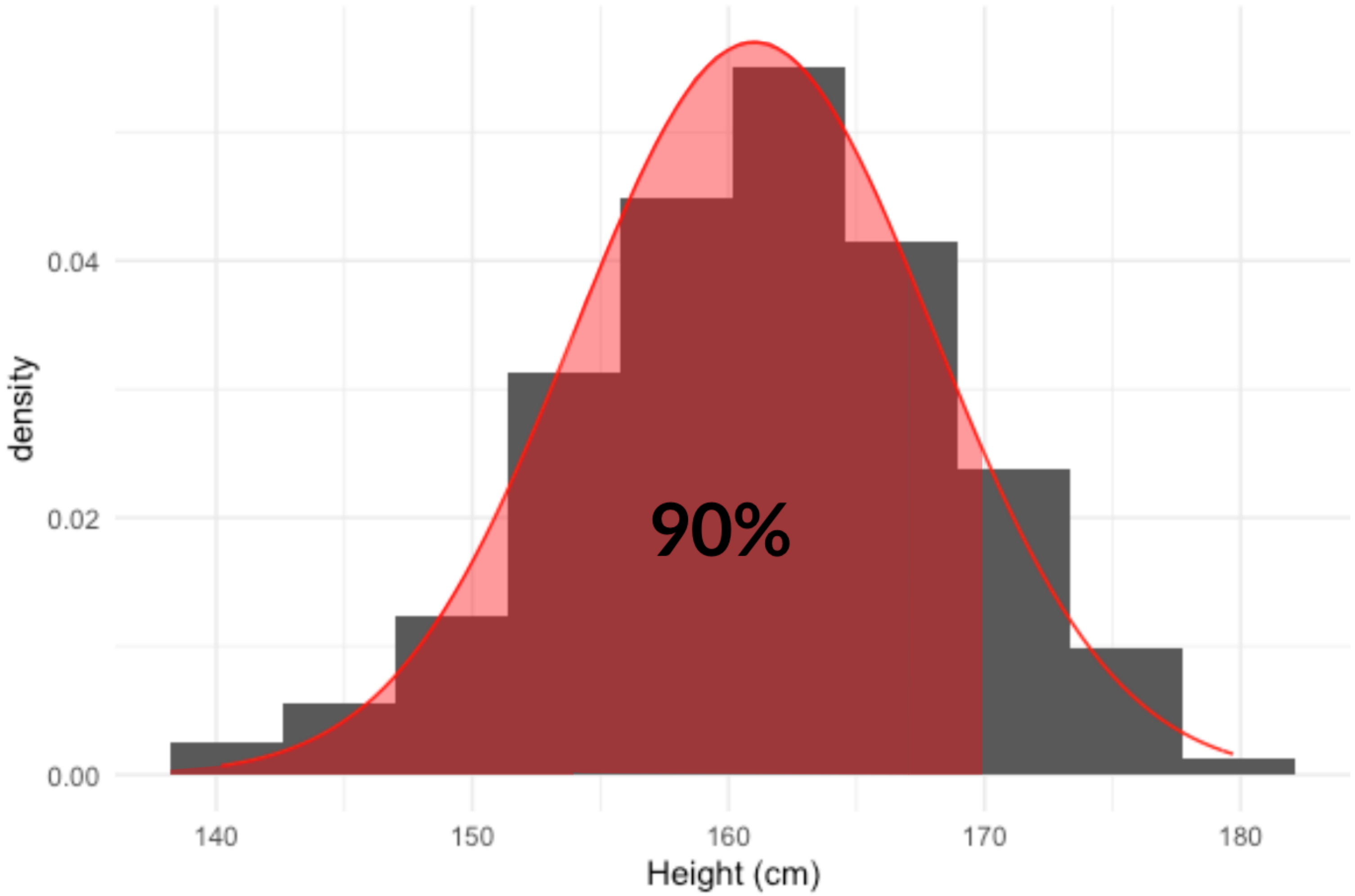
What percent of women are 154-157 cm?




```
pnorm(157, mean = 161, sd = 7) - pnorm(154, mean = 161, sd = 7)
```

```
0.1252
```

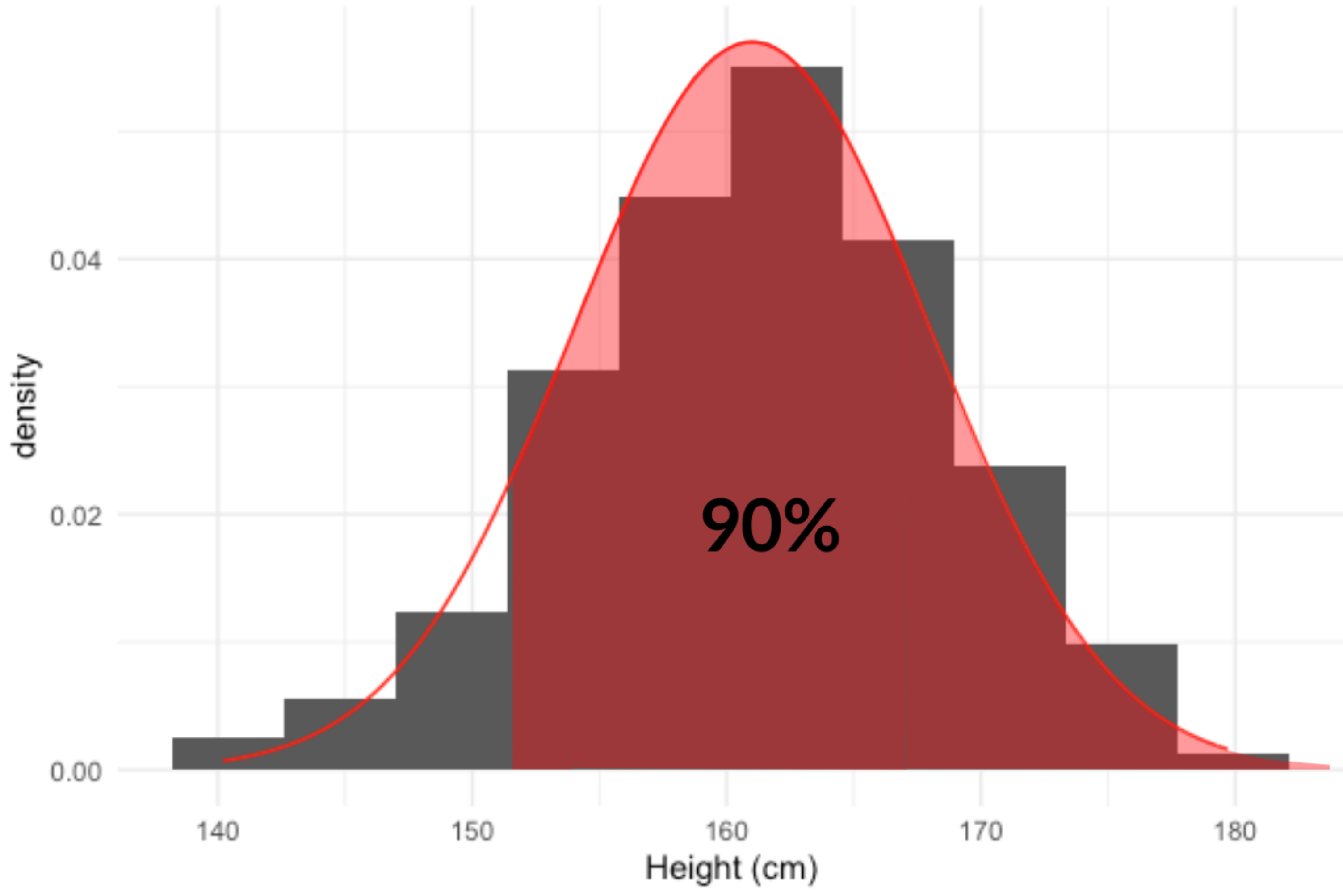
What height are 90% of women shorter than?




```
qnorm(0.9, mean = 161, sd = 7)
```

```
169.9709
```

What height are 90% of women taller than?




```r
qnorm(0.9,
      mean = 161,
      sd = 7,
      lower.tail = FALSE)
```

```
152.03
```

Generating random numbers

```
# Generate 10 random heights
rnorm(10, mean = 161, sd = 7)
```

```
159.35 157.34 149.85 156.75 163.53 156.33 157.22 171.44 158.10 170.12
```

Let's practice!

The central limit theorem

INTRODUCTION TO STATISTICS IN R

Maggie Matsui
Content Developer, DataCamp

Rolling the dice 5 times

```
die <- c(1, 2, 3, 4, 5, 6)
# Roll 5 times
sample_of_5 <- sample(die, 5,
                      replace = TRUE)
sample_of_5
```

```
1 3 4 1 1
```

```
mean(sample_of_5)
```

```
2.0
```

Rolling the dice 5 times

```
# Roll 5 times and take mean
sample(die, 5, replace = TRUE) %>% mean()
```

```
4.4
```

```
sample(die, 5, replace = TRUE) %>% mean()
```

```
3.8
```

Rolling the dice 5 times 10 times

Repeat 10 times:

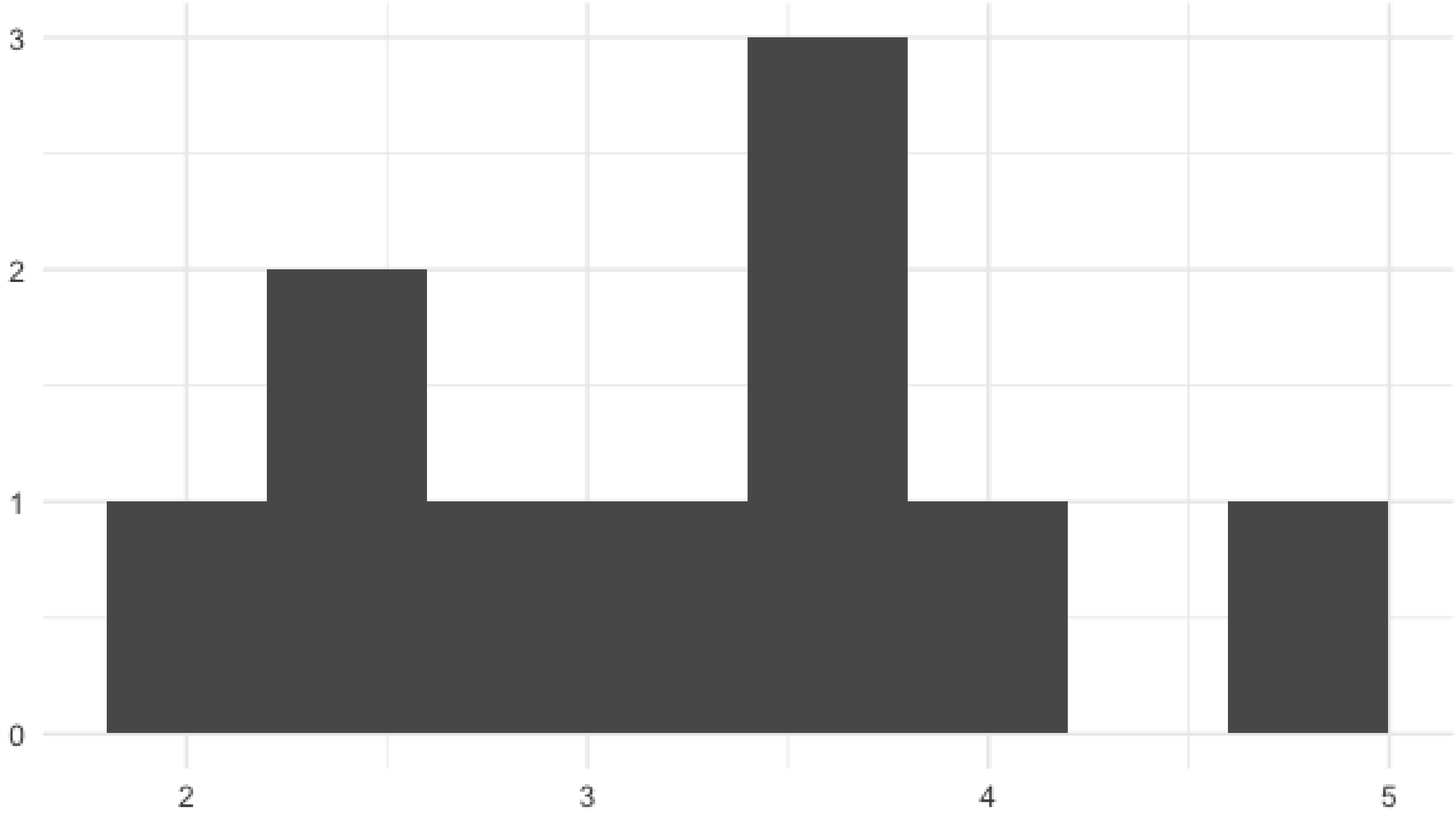
- Roll 5 times
- Take the mean

```
sample_means <- replicate(10, sample(die, 5, replace = TRUE) %>% mean())
sample_means
```

```
3.8 4.0 3.8 3.6 3.2 4.8 2.6 3.0 2.6 2.0
```

Sampling distributions

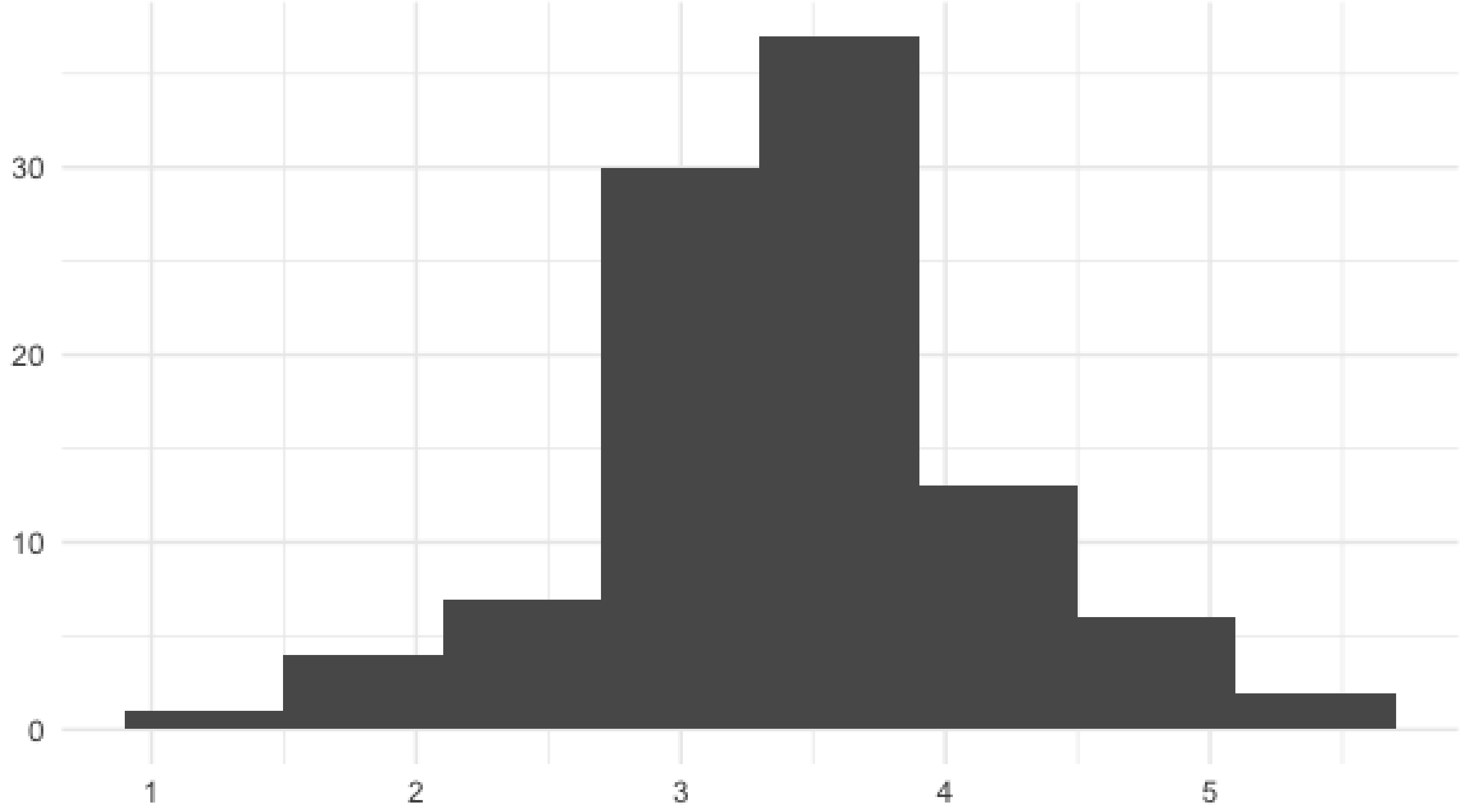
Sampling distribution of the sample mean



100 sample means

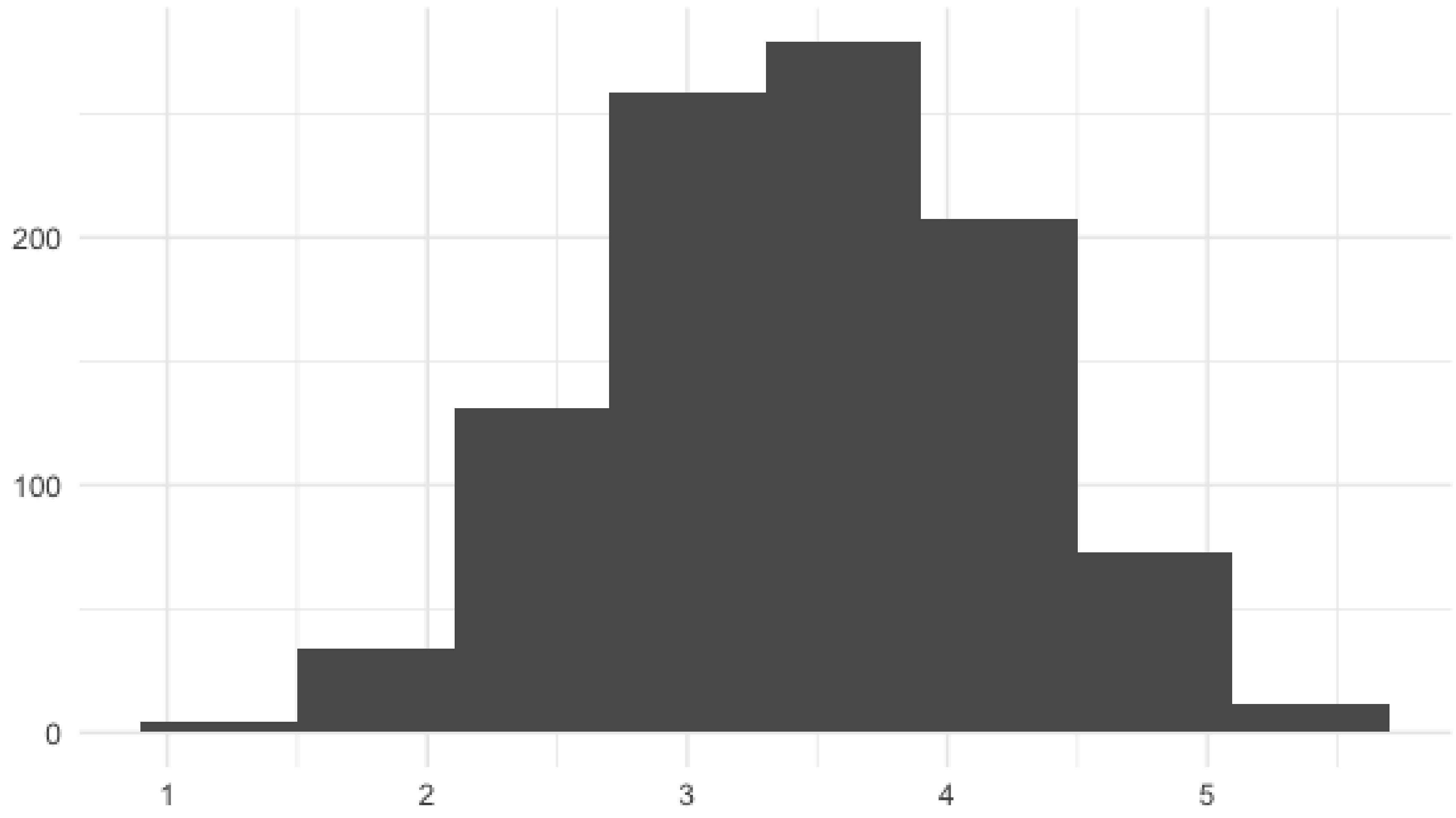
```
replicate(100, sample(die, 5, replace = TRUE) %>% mean())
```

```
2.8 3.2 1.8 4.6 4.0 2.8 4.4 2.4 3.4 2.8 4.2 3.4 ... 2.2 3.8 3.6 3.8 4.4 4.8 2.4
```



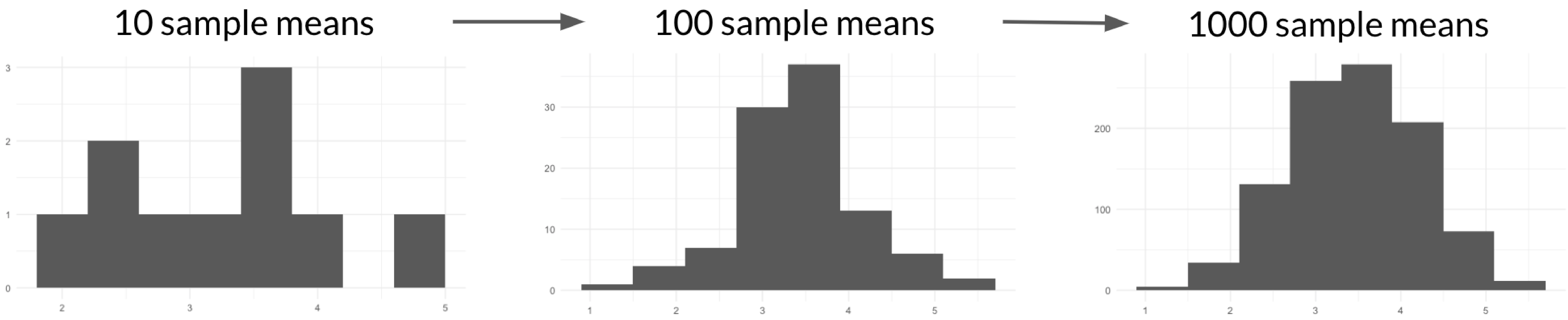
1000 sample means

```
sample_means <- replicate(1000, sample(die, 5, replace = TRUE) %>% mean())
```



Central limit theorem

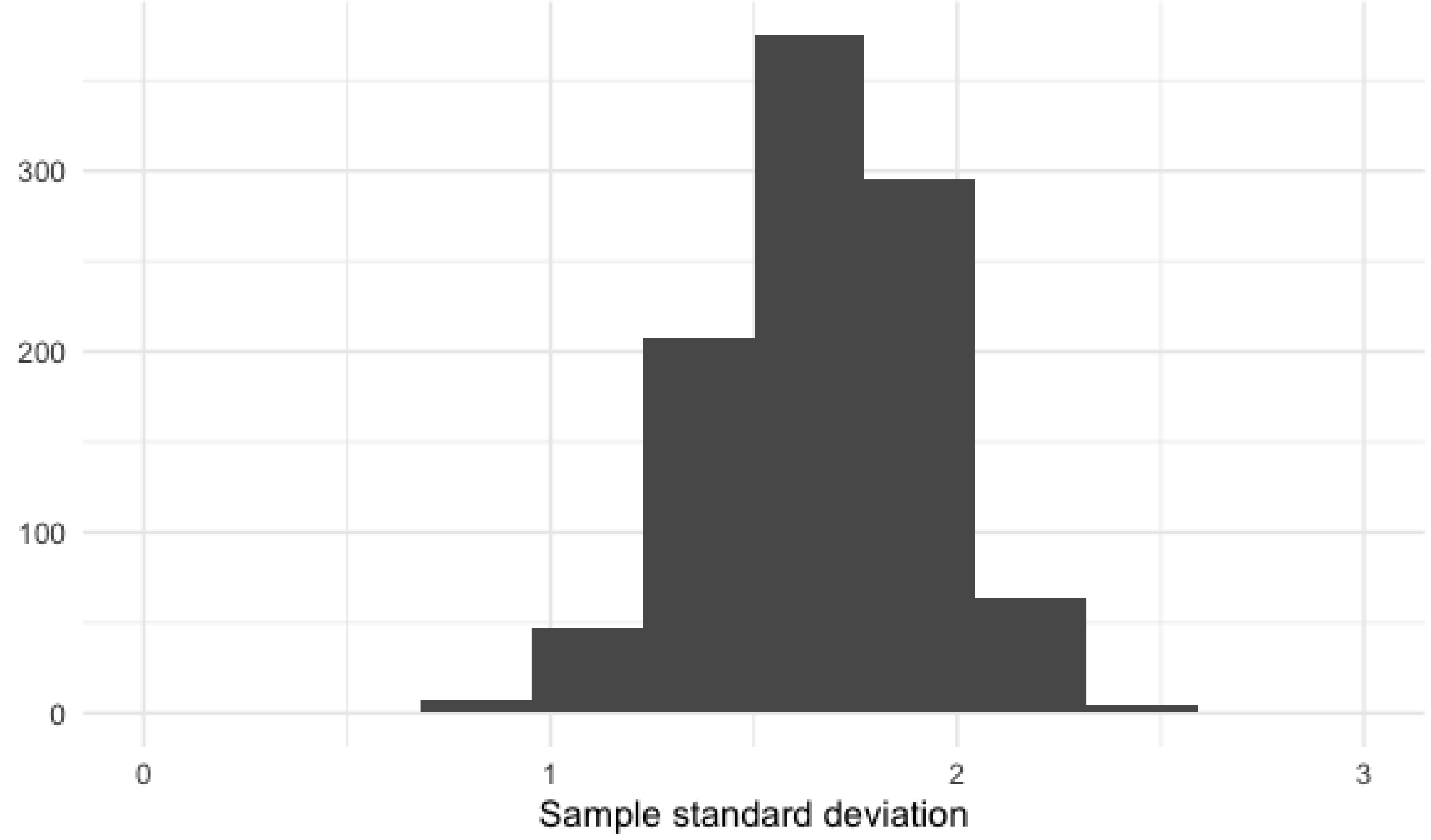
The sampling distribution of a statistic becomes closer to the normal distribution as the number of trials increases.



** Samples should be random and independent*

Standard deviation and the CLT

```
replicate(1000, sample(die, 5, replace = TRUE) %>% sd())
```



Proportions and the CLT

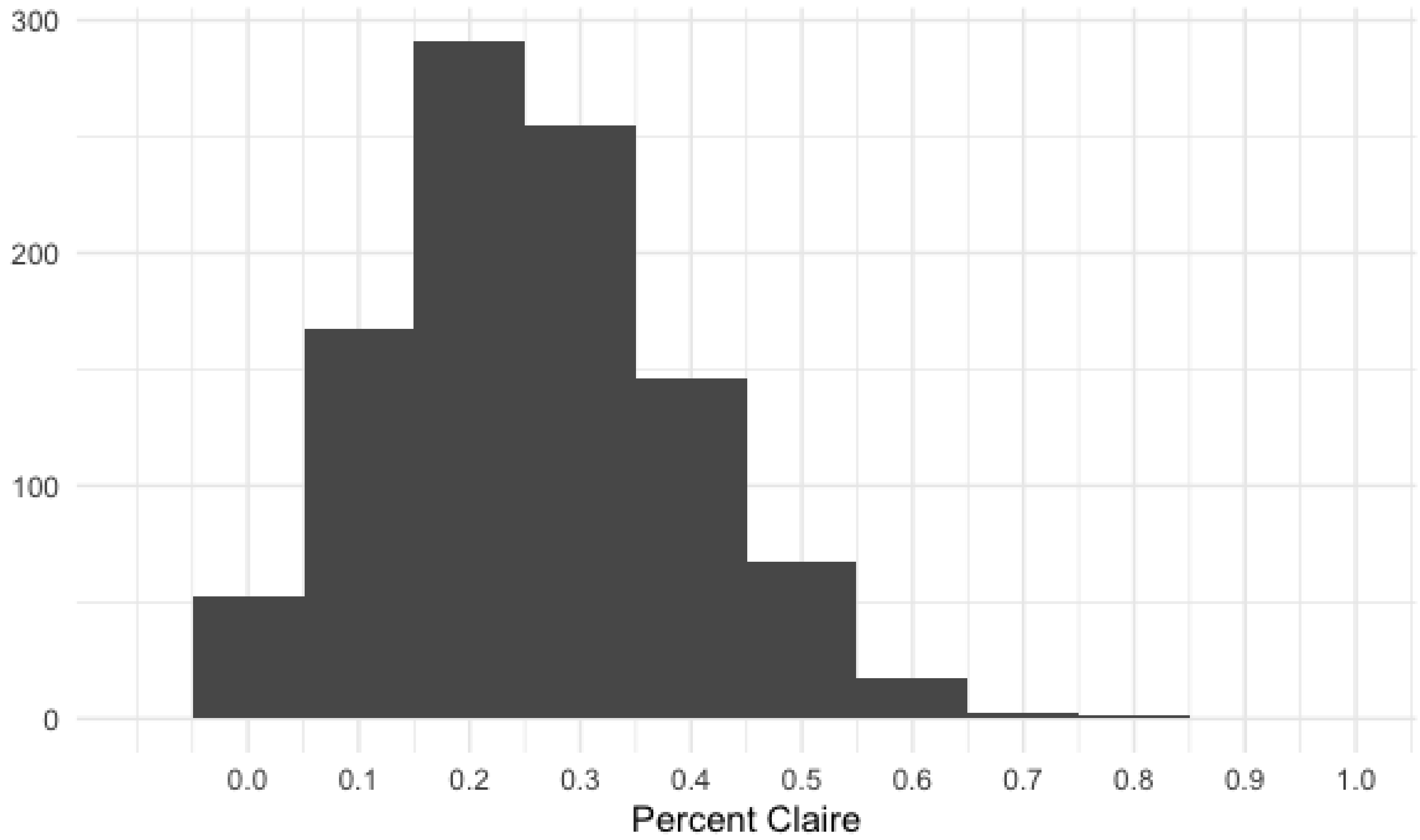
```
sales_team <- c("Amir", "Brian", "Claire", "Damian")
sample(sales_team, 10, replace = TRUE)
```

```
"Claire" "Brian"  "Brian"  "Brian"  "Damian" "Damian" "Brian"  "Brian"
"Amir"   "Amir"
```

```
sample(sales_team, 10, replace = TRUE)
```

```
"Amir"   "Amir"   "Claire" "Amir"   "Amir"   "Brian"  "Amir"   "Claire"
"Claire" "Claire"
```

Sampling distribution of proportion



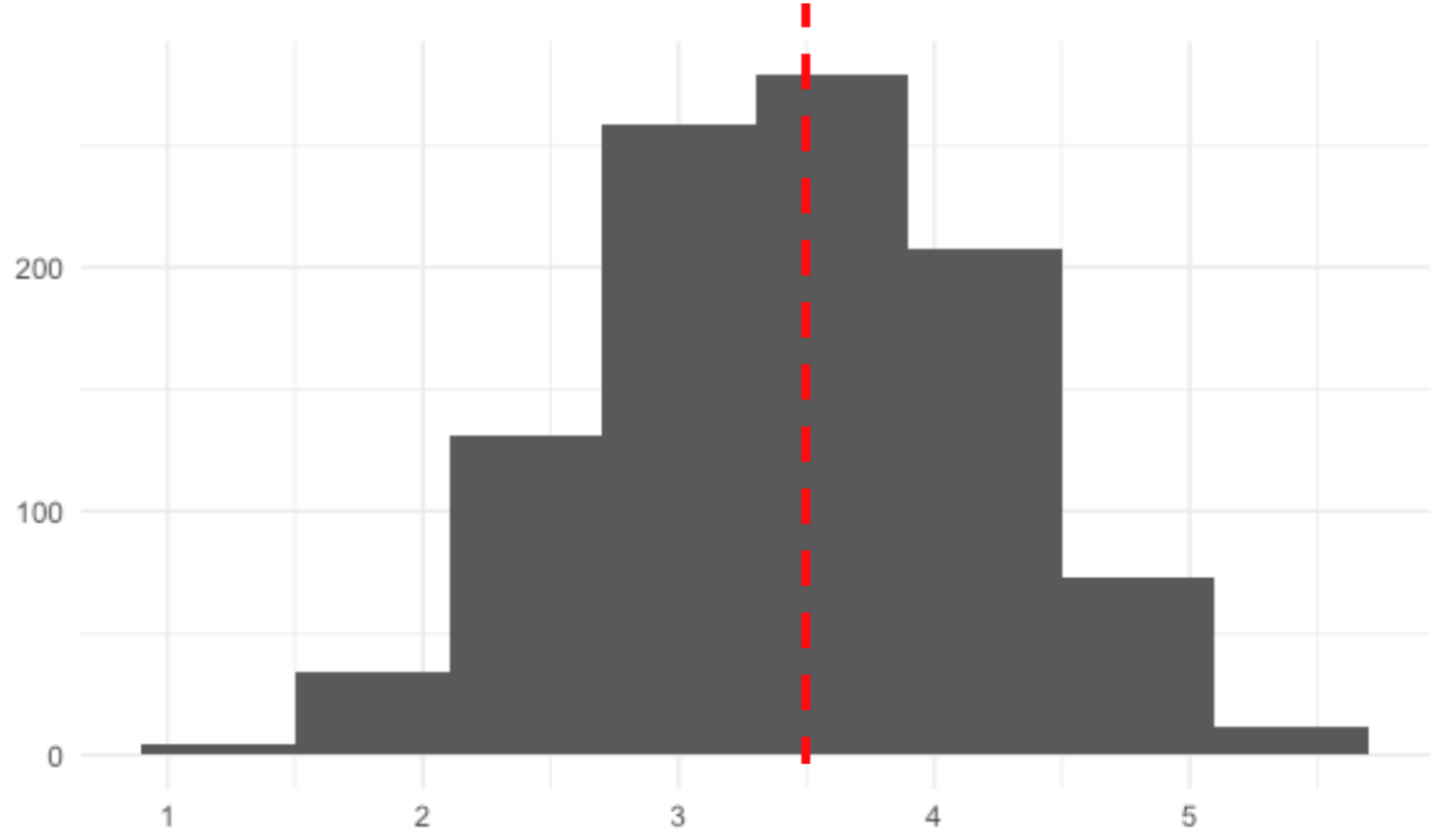
Mean of sampling distribution

```
# Estimate expected value of die
mean(sample_means)
```

```
3.48
```

```
# Estimate proportion of "Claire"s
mean(sample_props)
```

```
0.26
```



- Estimate characteristics of unknown underlying distribution
- More easily estimate characteristics of large populations

Let's practice!

The Poisson distribution

INTRODUCTION TO STATISTICS IN R

Maggie Matsui
Content Developer, DataCamp

Poisson processes

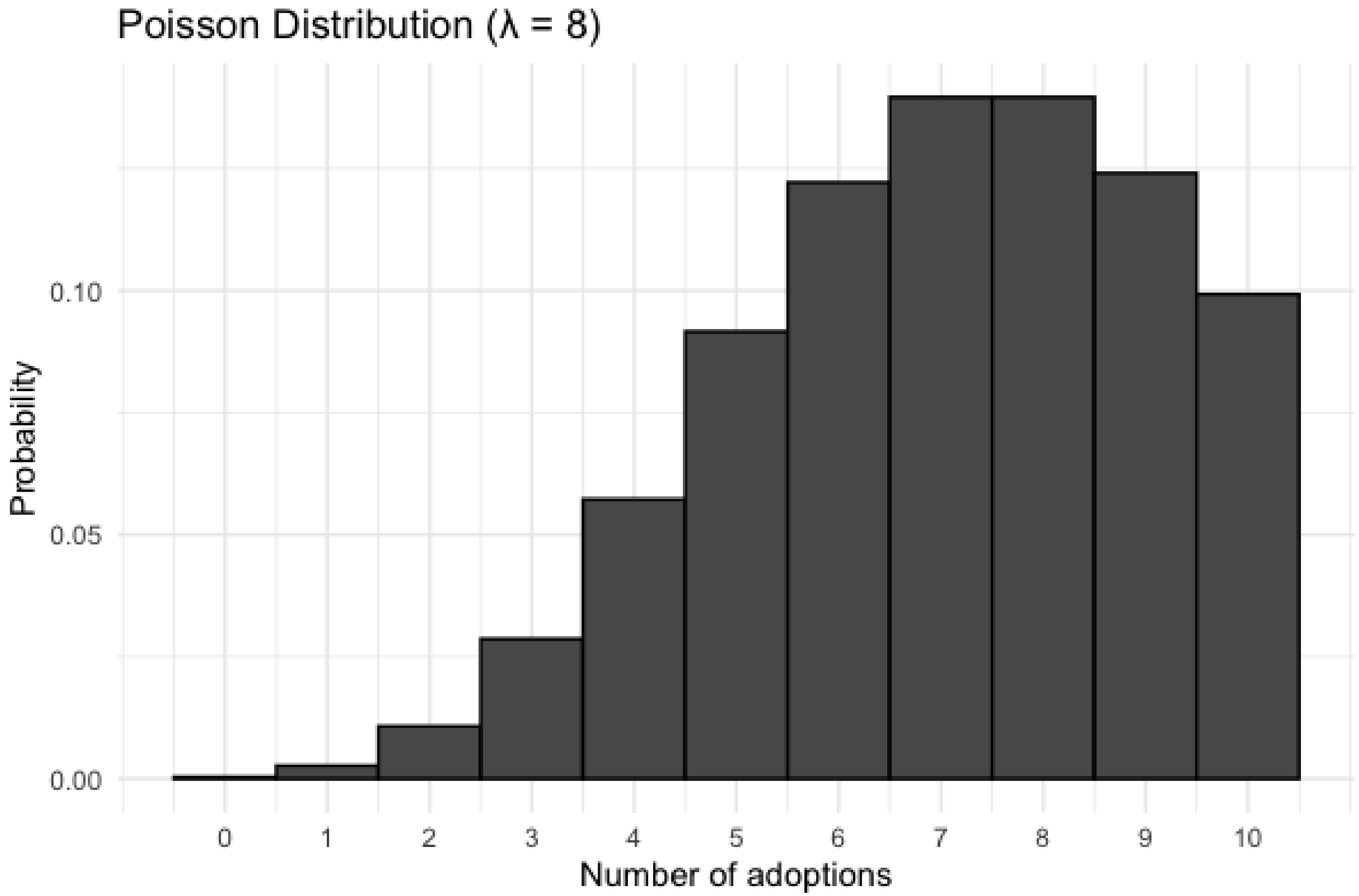
- Events appear to happen at a certain rate, but completely at random
- Examples
  - Number of animals adopted from an animal shelter per week
  - Number of people arriving at a restaurant per hour
  - Number of earthquakes in California per year

Poisson distribution

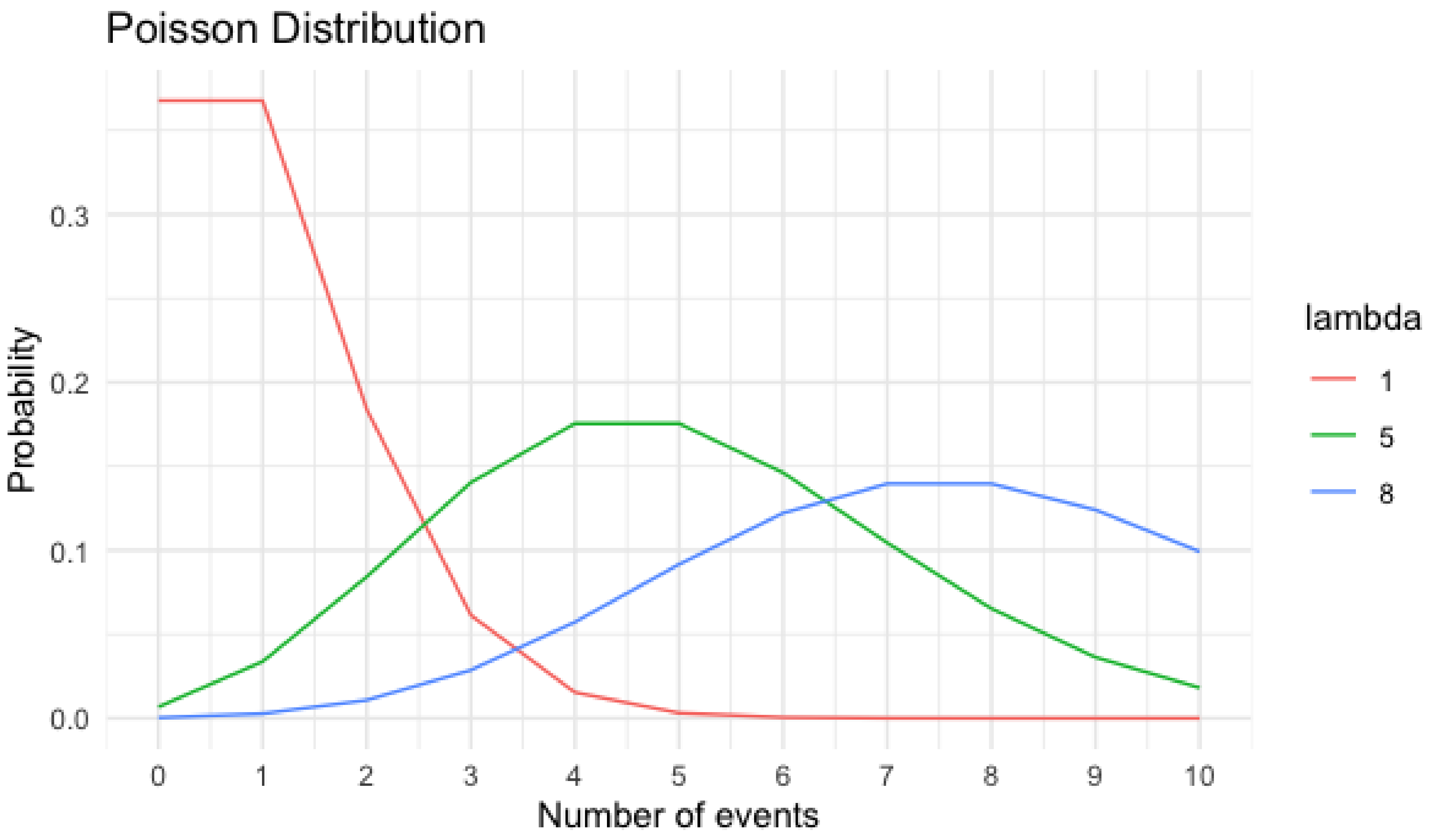
- Probability of some # of events occurring over a fixed period of time
- Examples
  - Probability of ≥ 5 animals adopted from an animal shelter per week
  - Probability of 12 people arriving at a restaurant per hour
  - Probability of < 20 earthquakes in California per year

Lambda (λ)

- λ = average number of events per time interval
  - Average number of adoptions per week = 8



Lambda is the distribution's peak



Probability of a single value

If the average number of adoptions per week is 8, what is $P(\# \text{ adoptions in a week} = 5)$?

```
dpois(5, lambda = 8)
```

```
0.09160366
```

Probability of less than or equal to

If the average number of adoptions per week is 8, what is $P(\#\text{ adoptions in a week} \leq 5)$?

```
ppois(5, lambda = 8)
```

```
0.1912361
```

Probability of greater than

If the average number of adoptions per week is 8, what is $P(\# \text{ adoptions in a week} > 5)$?

```
ppois(5, lambda = 8, lower.tail = FALSE)
```

```
0.8087639
```

If the average number of adoptions per week is **10**, what is $P(\# \text{ adoptions in a week} > 5)$?

```
ppois(5, lambda = 10, lower.tail = FALSE)
```

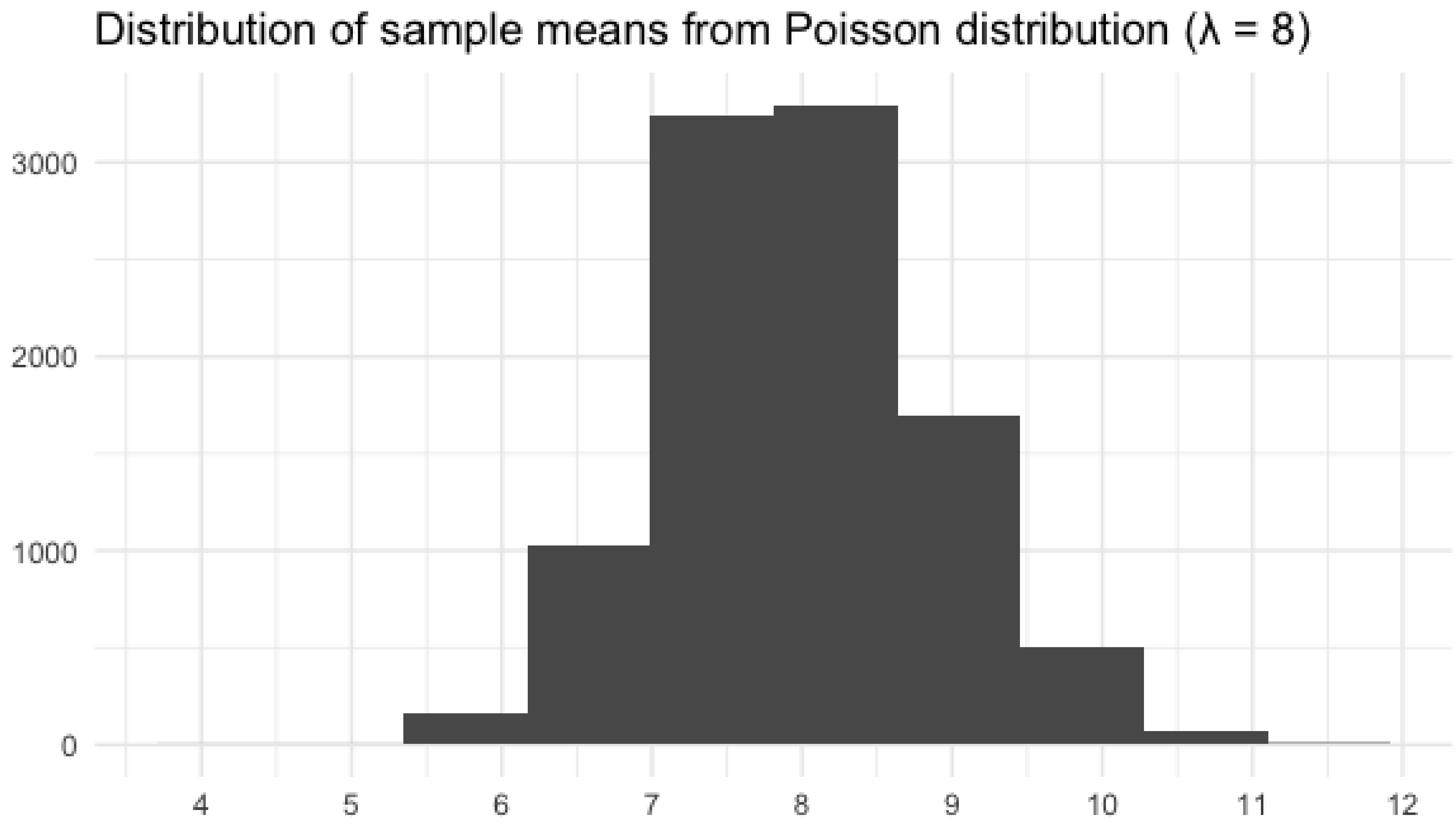
```
0.932914
```

Sampling from a Poisson distribution

```
rpois(10, lambda = 8)
```

```
13  6 11  7 10  8  7  3  7  6
```

The CLT still applies!



Let's practice!

INTRODUCTION TO STATISTICS IN R

More probability distributions

INTRODUCTION TO STATISTICS IN R

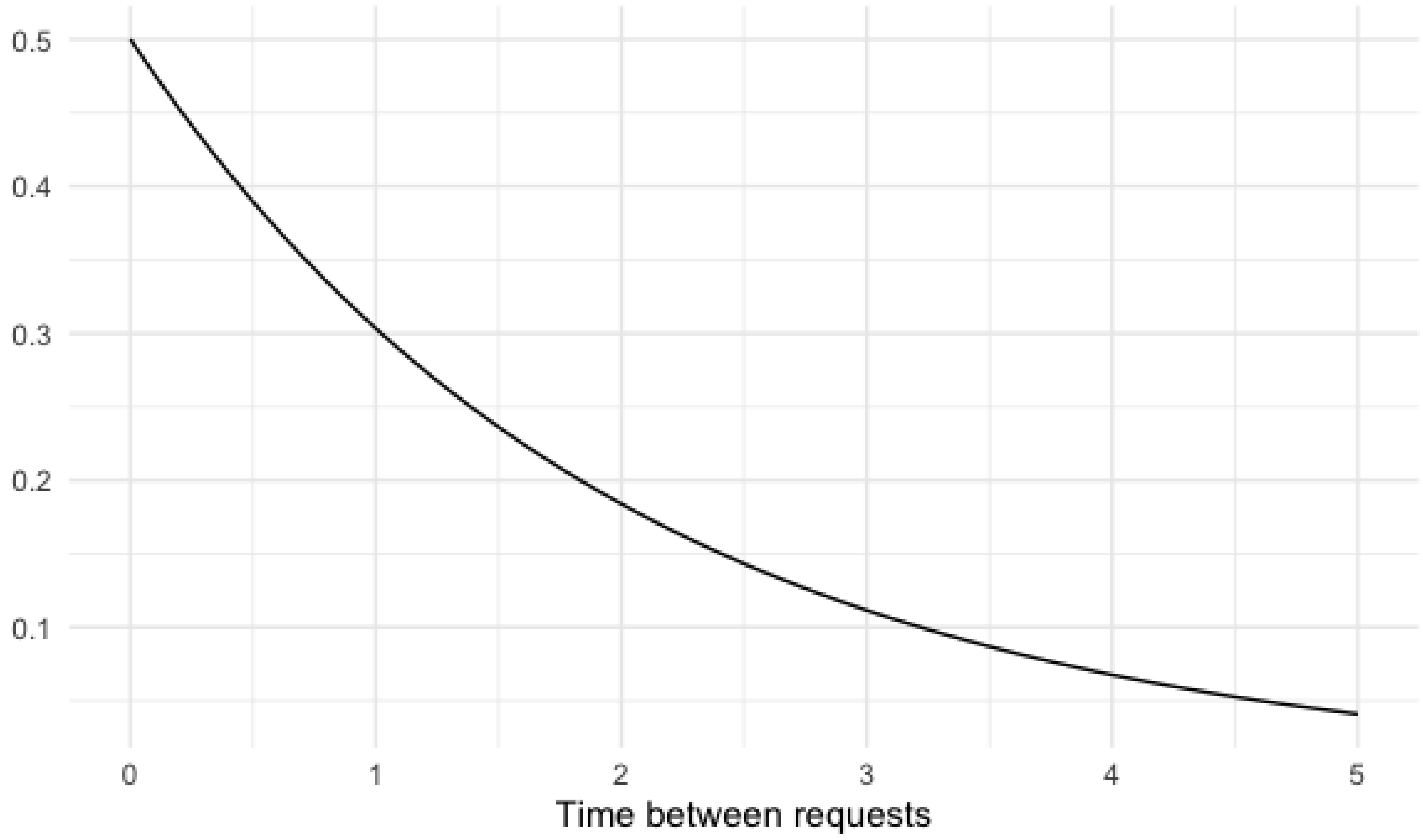
Maggie Matsui
Content Developer, DataCamp

Exponential distribution

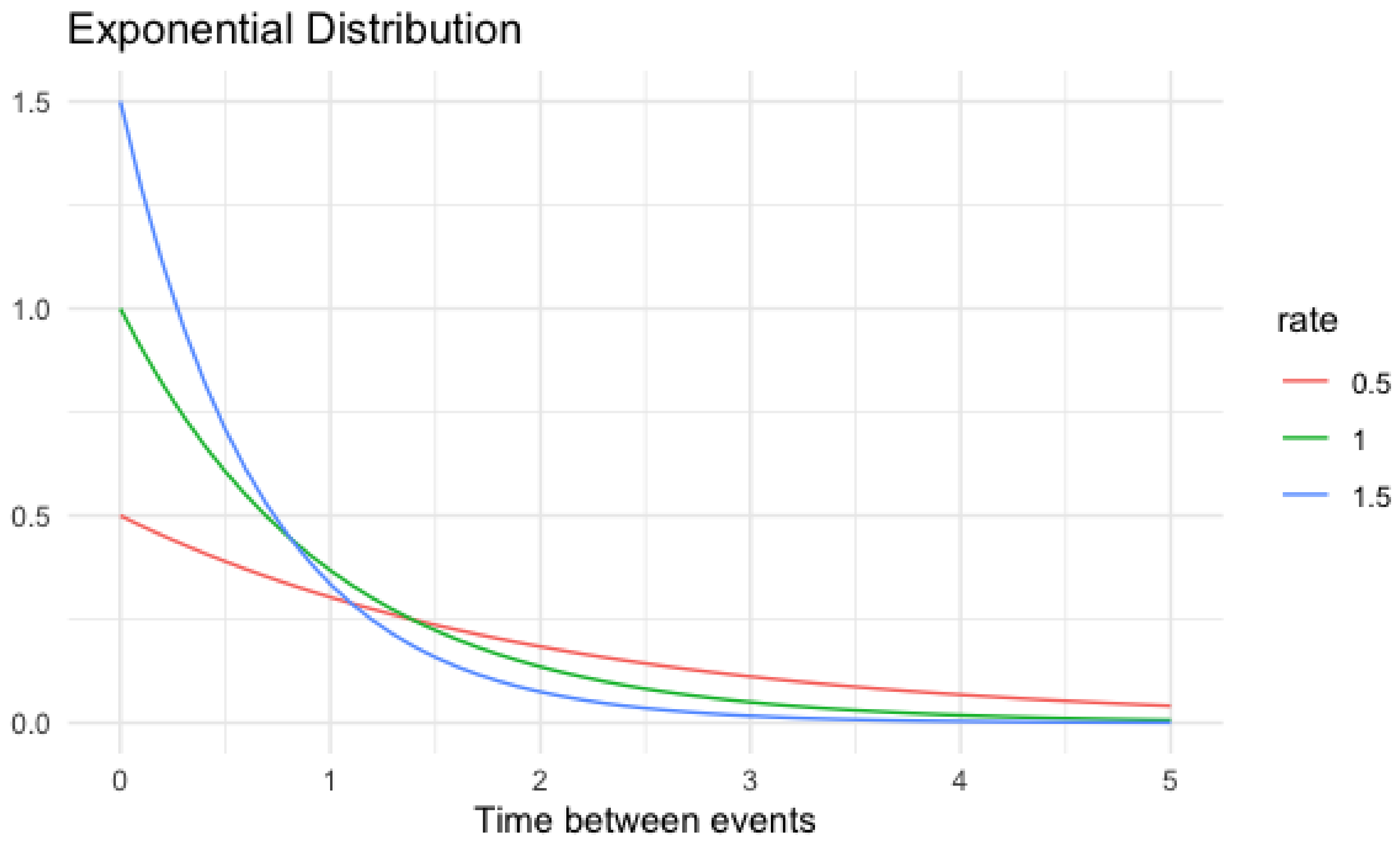
- Probability of time between Poisson events
- Examples
  - Probability of > 1 day between adoptions
  - Probability of < 10 minutes between restaurant arrivals
  - Probability of 6-8 months between earthquakes
- Also uses lambda (rate)
- Continuous (time)

Customer service requests

- On average, one customer service ticket is created every 2 minutes
  - λ = 0.5 customer service tickets created each minute



Lambda in exponential distribution



How long until a new request is created?

$P(\text{wait} < 1 \text{ min})$ =

```
pexp(1, rate = 0.5)
```

```
0.3934693
```

$P(\text{wait} > 4 \text{ min})$ =

```
pexp(4, rate = 0.5, lower.tail = FALSE)
```

```
0.1353353
```

$P(1 \text{ min} < \text{wait} < 4 \text{ min})$ =

```
pexp(4, rate = 0.5) - pexp(1, rate = 0.5)
```

```
0.4711954
```

Expected value of exponential distribution

In terms of rate (Poisson):

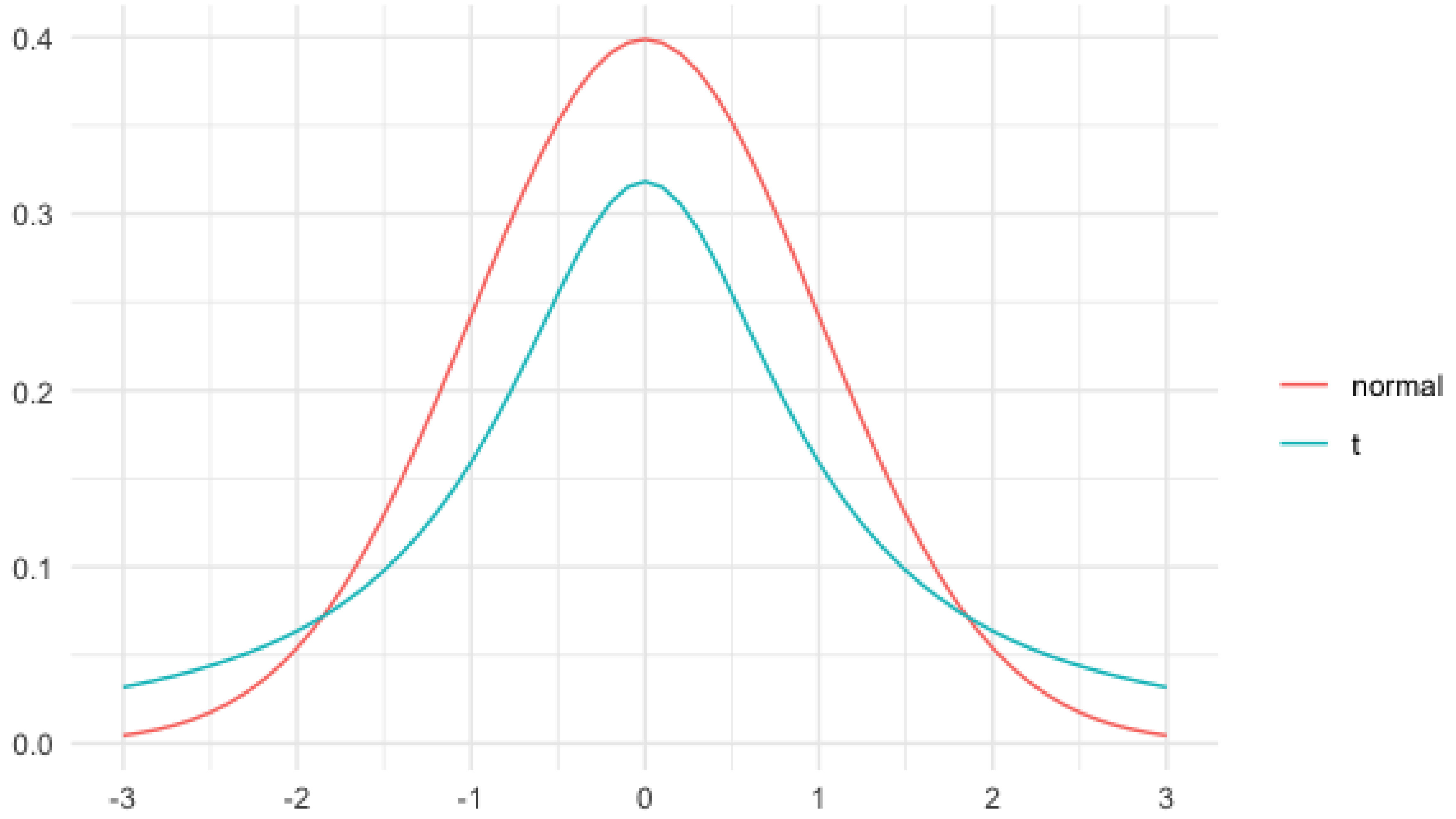
- λ = 0.5 requests per minute

In terms of time (exponential):

- $1/\lambda$ = 1 request per 2 minutes

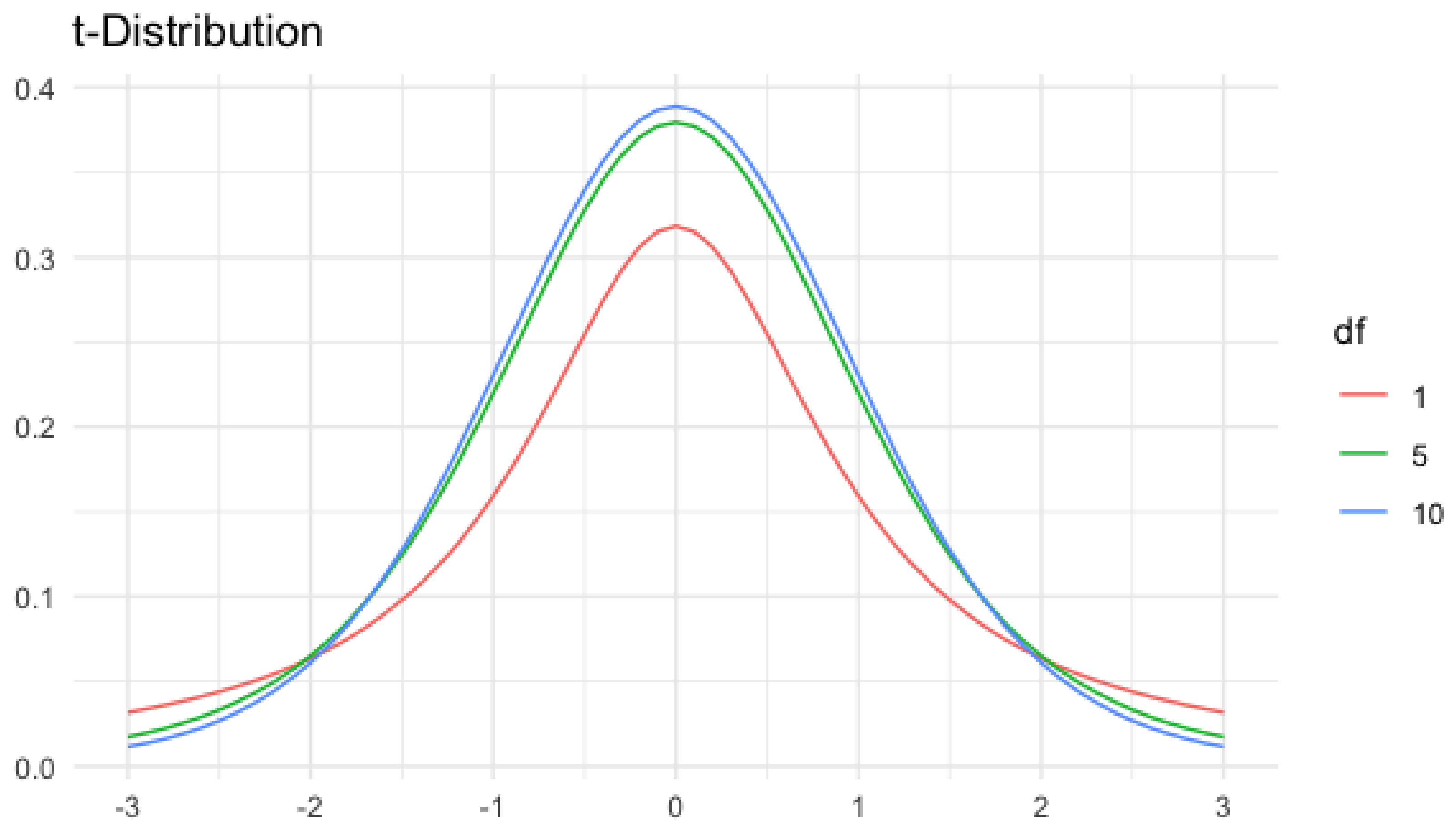
(Student's) t-distribution

- Similar shape as the normal distribution



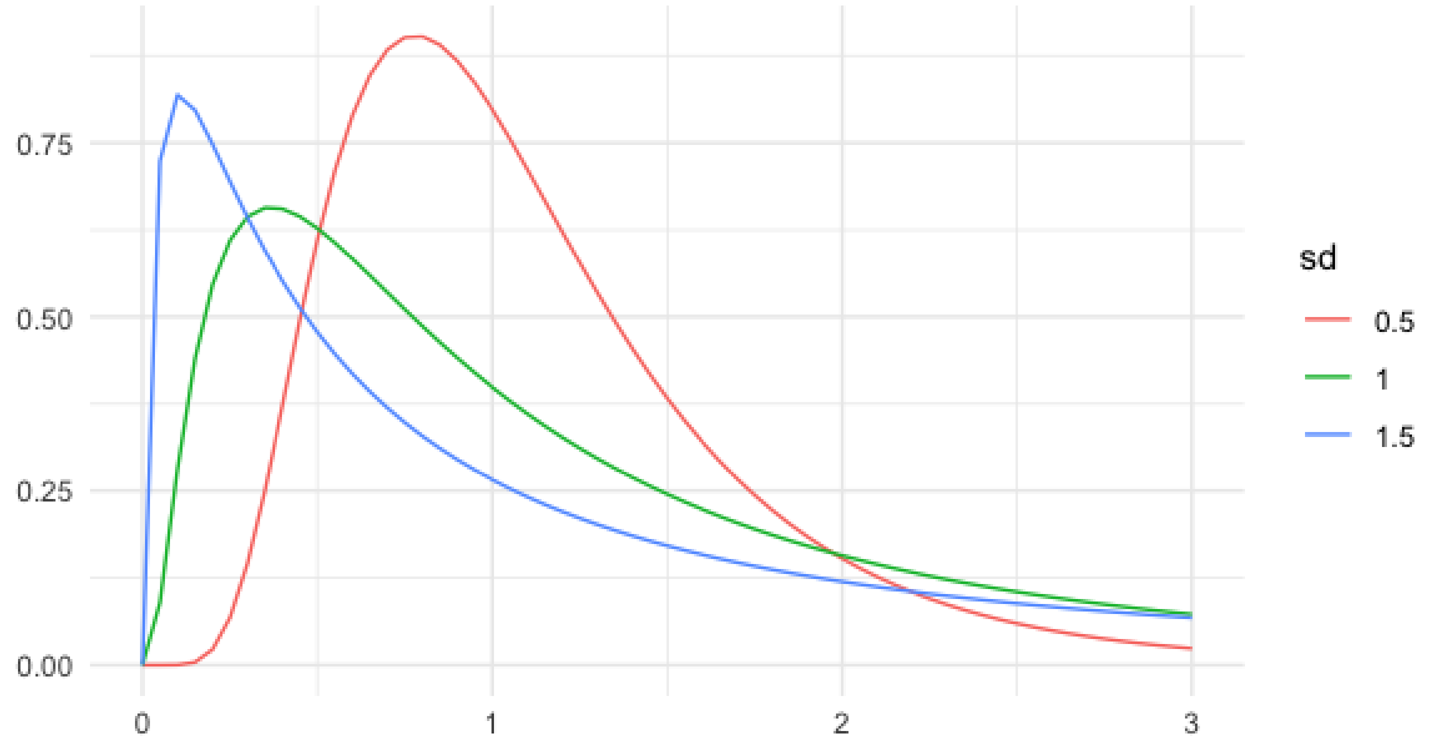
Degrees of freedom

- Has parameter degrees of freedom (df) which affects the thickness of the tails
  - Lower df = thicker tails, higher standard deviation
  - Higher df = closer to normal distribution



Log-normal distribution

- Variable whose logarithm is normally distributed
- Examples:
  - Length of chess games
  - Adult blood pressure
  - Number of hospitalizations in the 2003 SARS outbreak



Let's practice!