

The binomial distribution

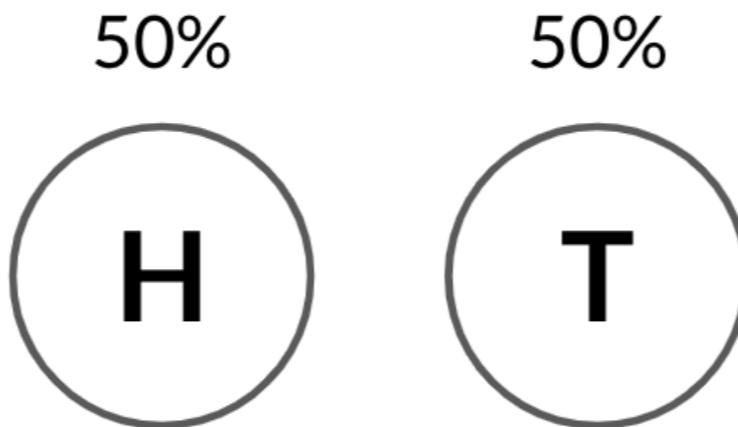
INTRODUCTION TO STATISTICS



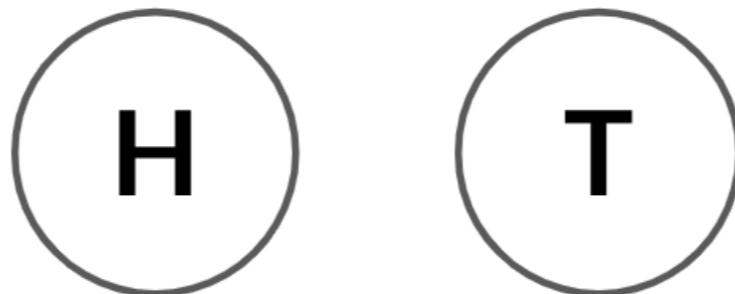
George Boorman

Curriculum Manager, DataCamp

Coin flipping



Binary outcomes



1

0

Success

Failure

Win

Loss

One coin flip many times

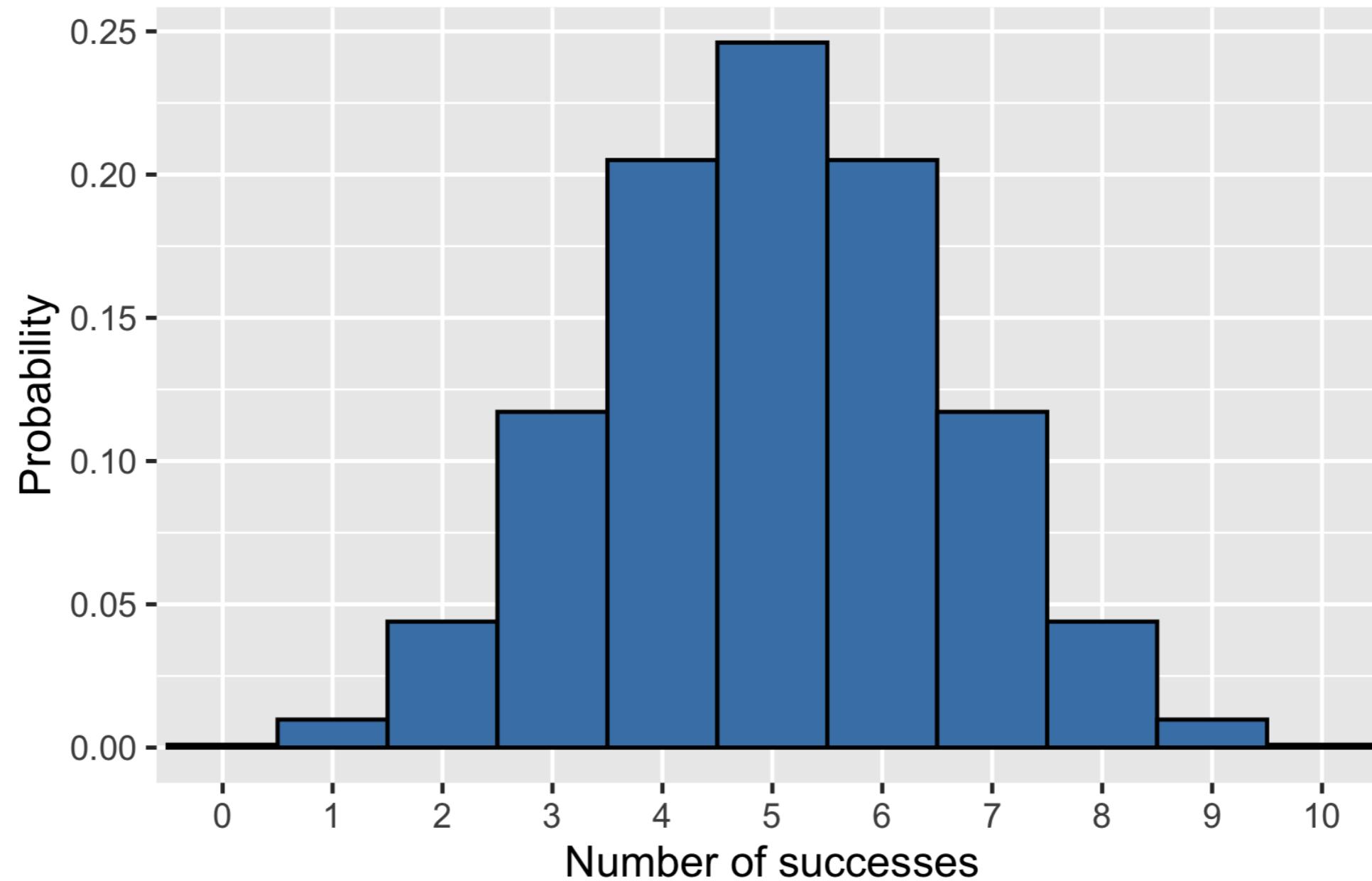
Coin Flip	Result
1	0
2	0
3	0
4	1
5	0
6	0
7	1
8	0
9	1
10	1

Binomial distribution

- Probability distribution of the **number of successes** in a sequence of **independent events**
- For example, the number of heads in a sequence of coin flips
- Described by n and p
 - n : total number of events
 - p : probability of success

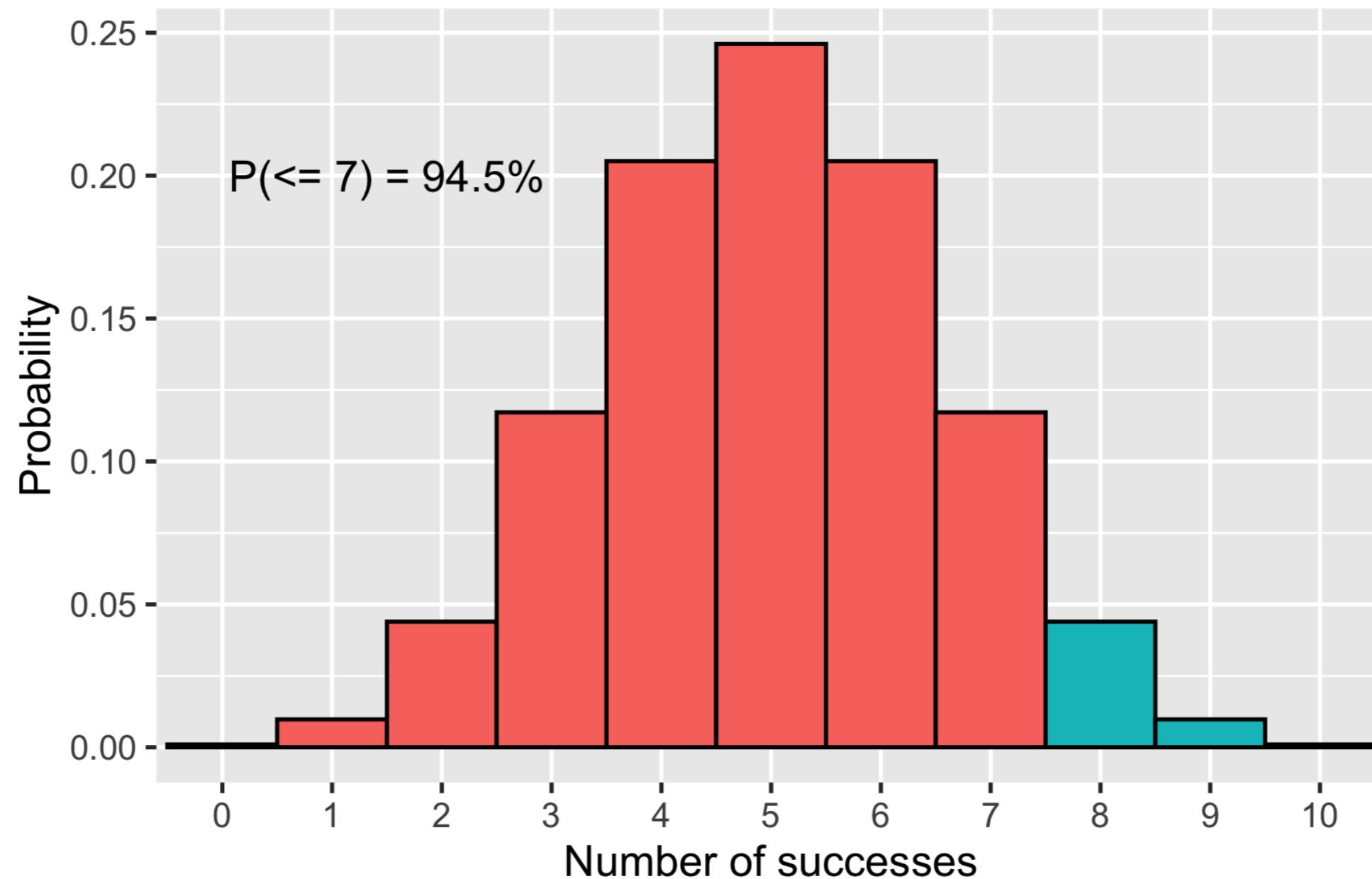
Binomial distribution

Binomial Distribution ($n=10, p=0.5$)



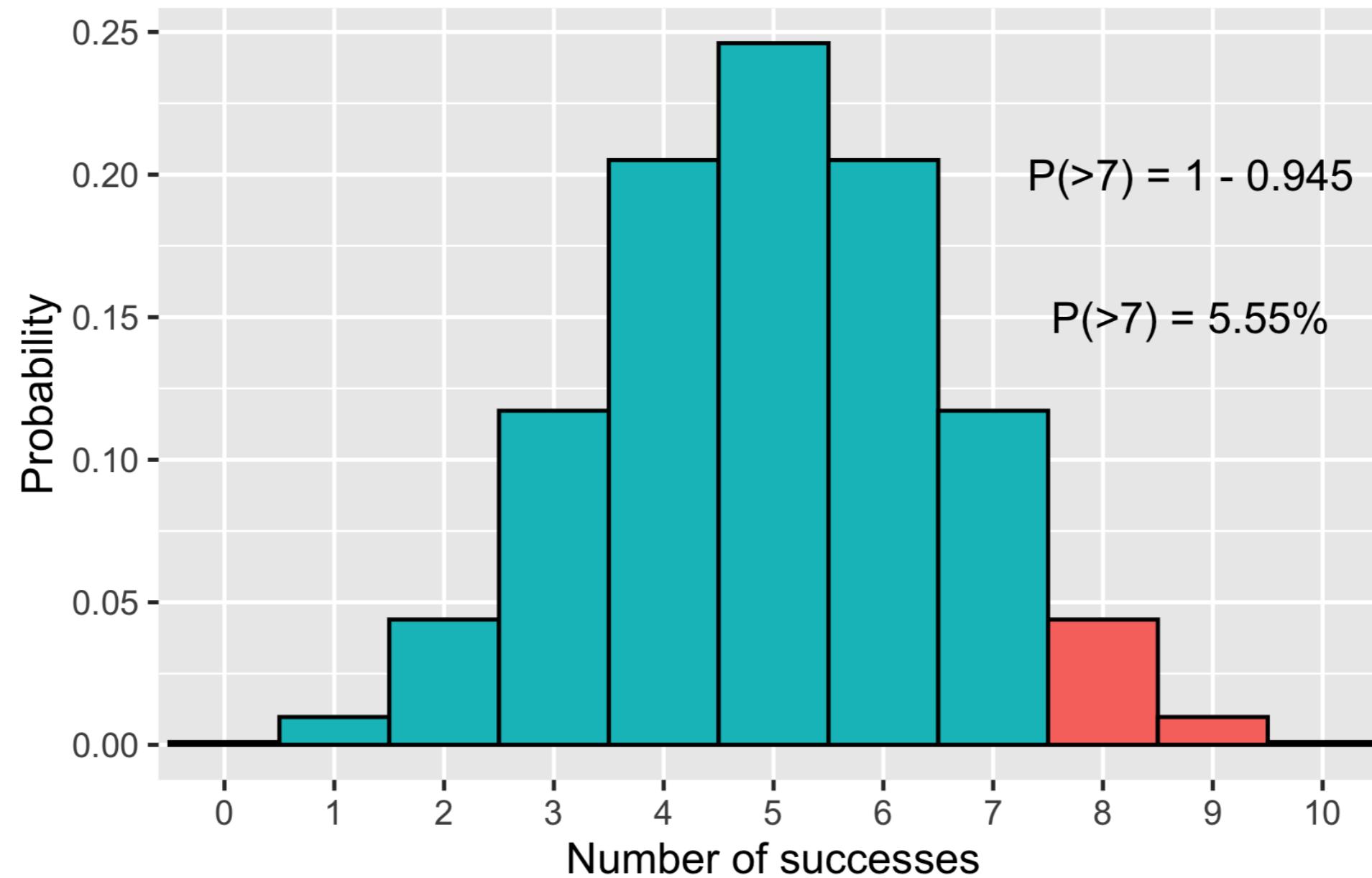
Probability of 7 or fewer heads

Binomial Distribution ($n=10, p=0.5$)



Probability of 8 or more heads

Binomial Distribution ($n=10, p=0.5$)



Expected value

Expected value = $n \times p$

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

If we don't know p , but know n and the expected value:

$$p = \frac{\text{expected value}}{n}$$

Independence

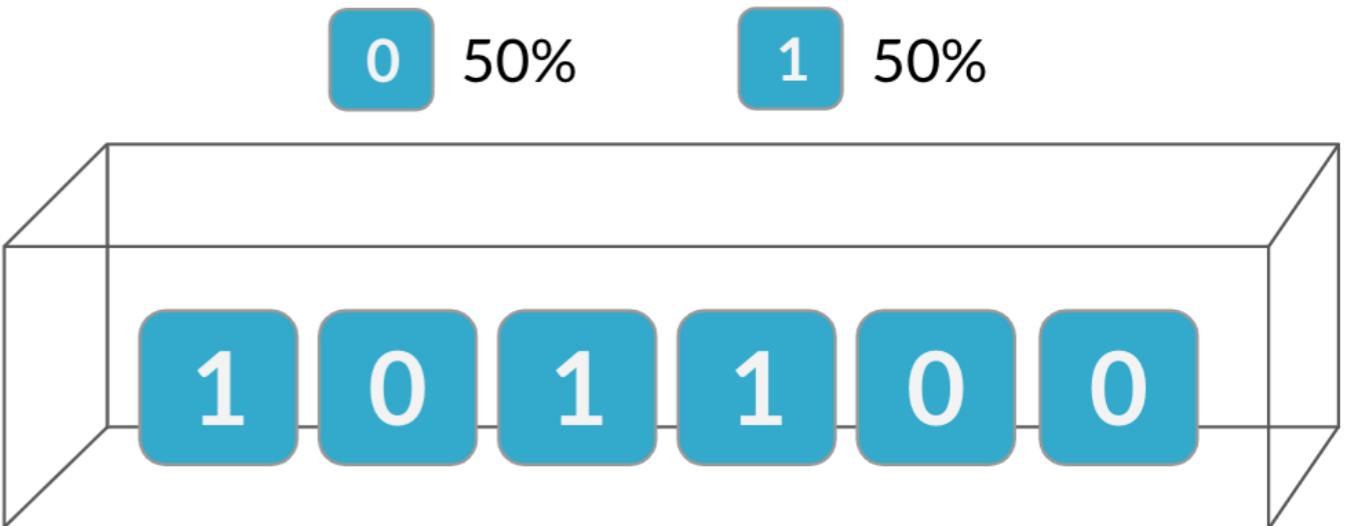
*The binomial distribution is a probability distribution of the number of successes in a sequence of **independent** events*

0 50% 1 50%

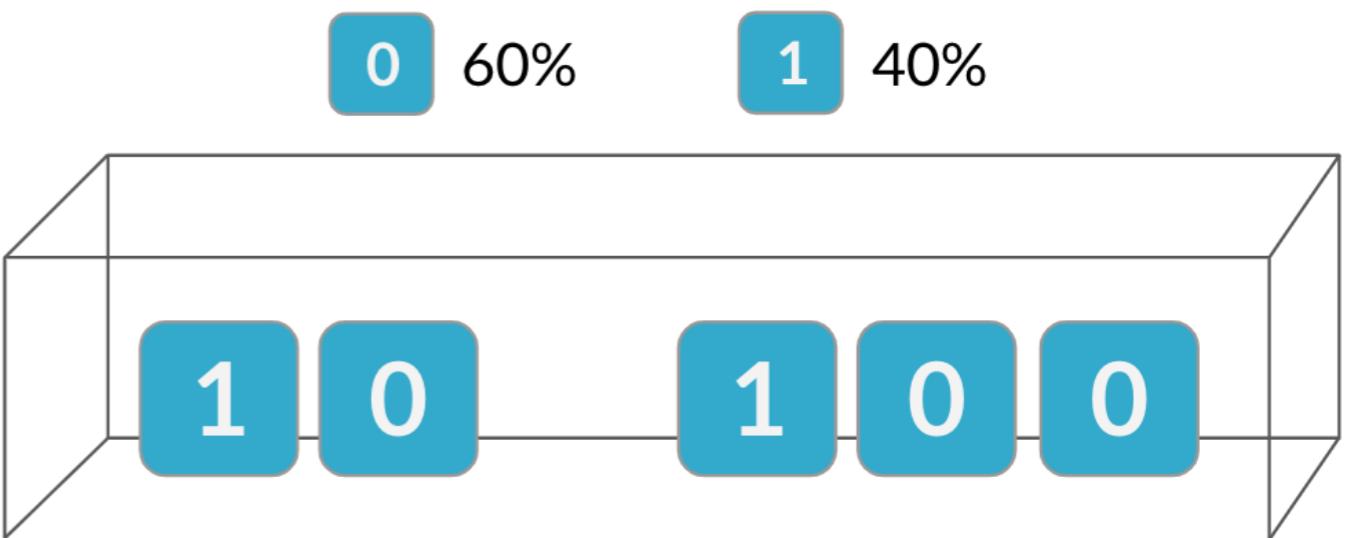


Independence

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



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



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

General applications

The binomial distribution can be used for independent events producing binary outcomes

- Clinical trial measuring drug effectiveness
 - Effective or not
- Betting on the result of a sports match
 - Bettor can win or lose



¹ Image credit: <https://unsplash.com/@towfiq99999>

Let's practice!

INTRODUCTION TO STATISTICS

The normal distribution

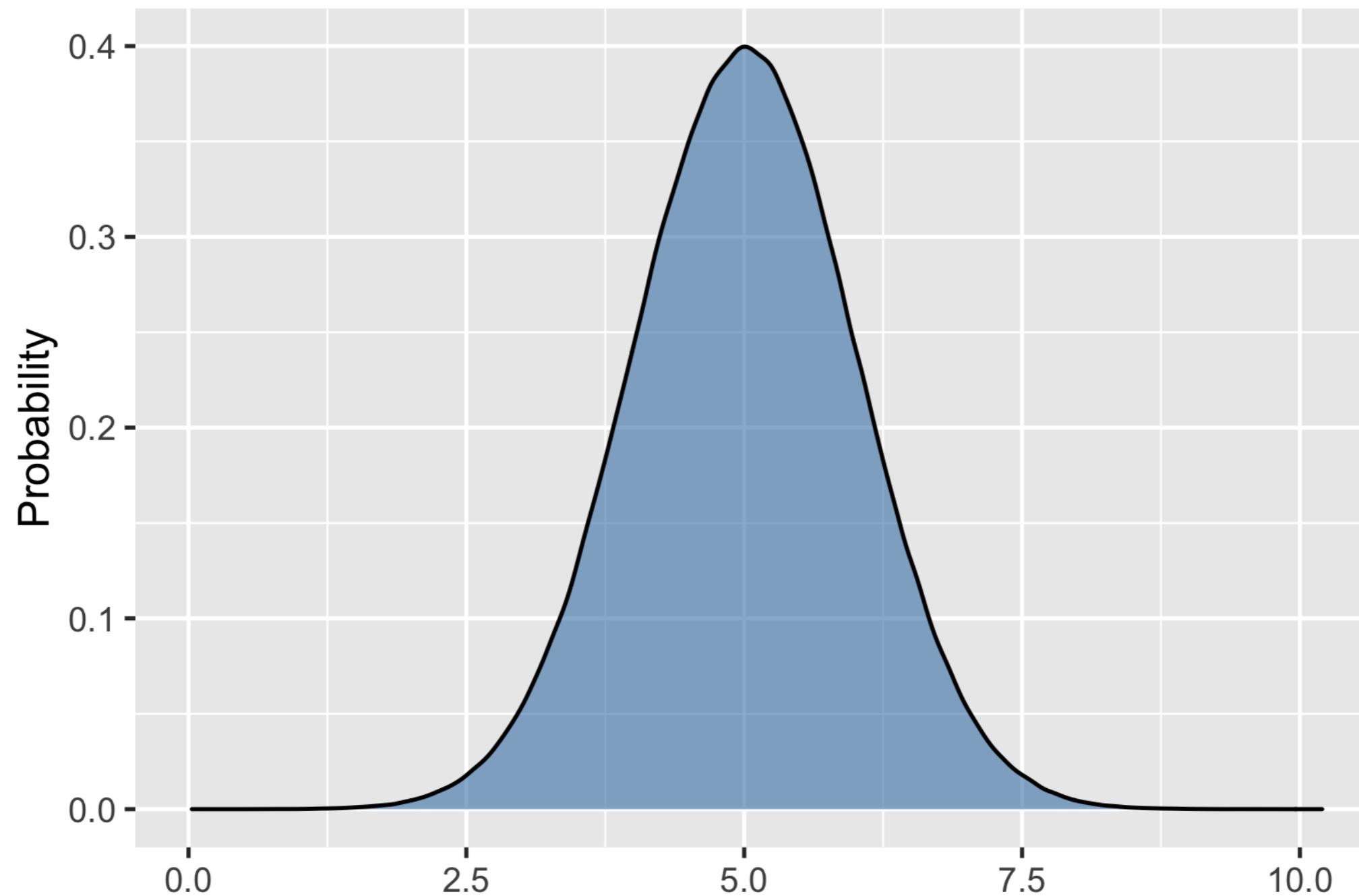
INTRODUCTION TO STATISTICS



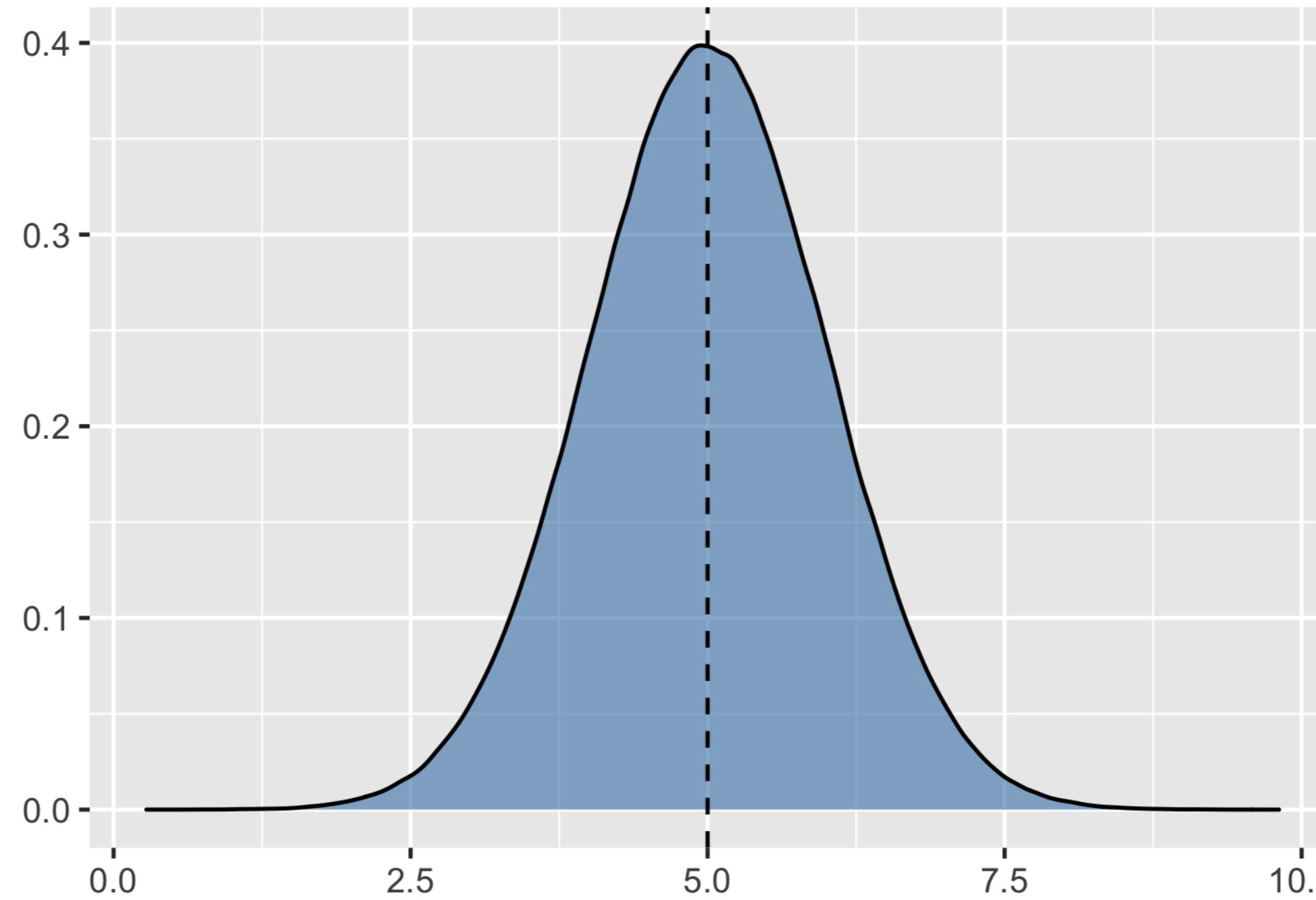
George Boorman

Curriculum Manager, DataCamp

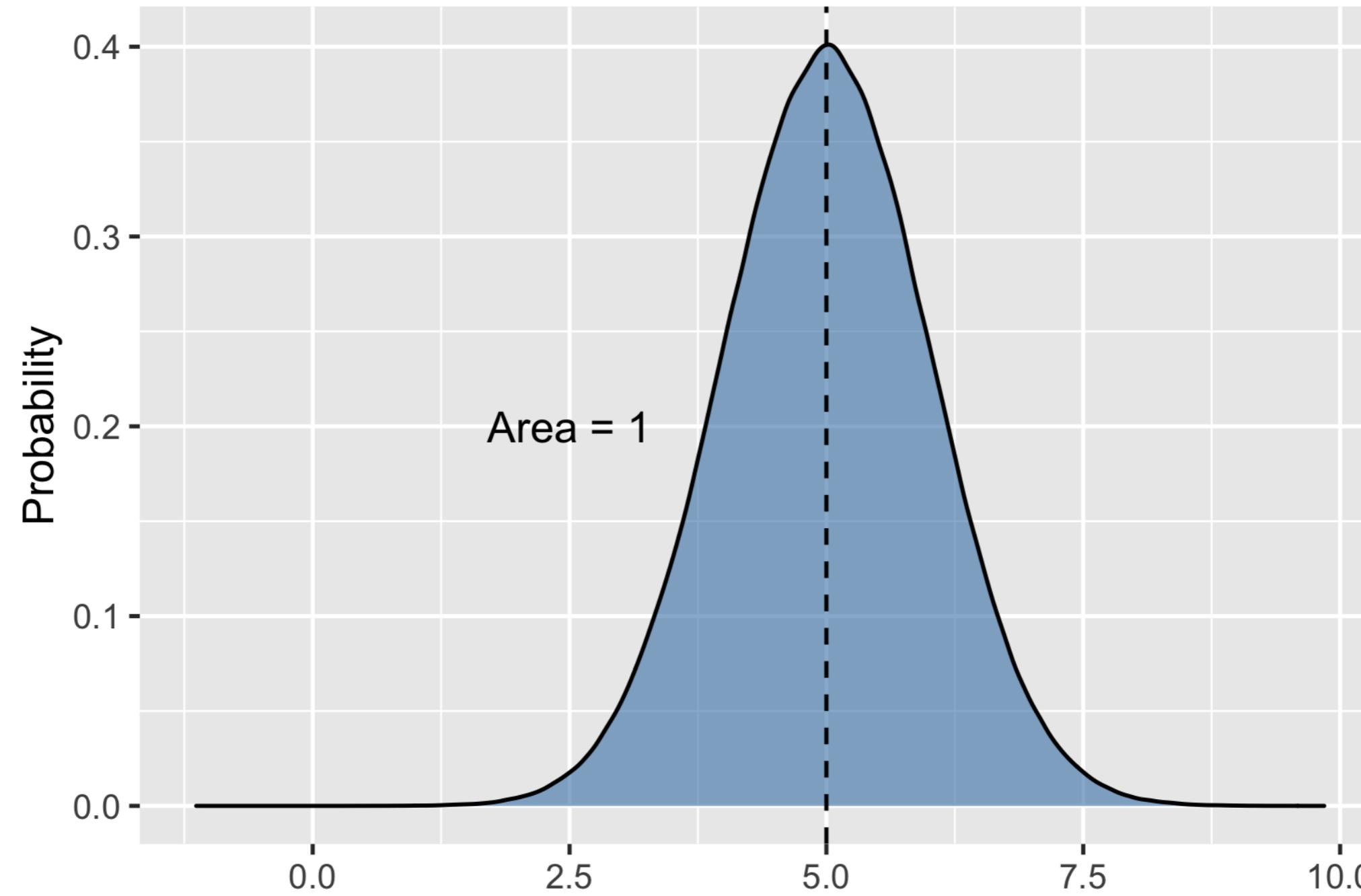
What is the normal distribution?



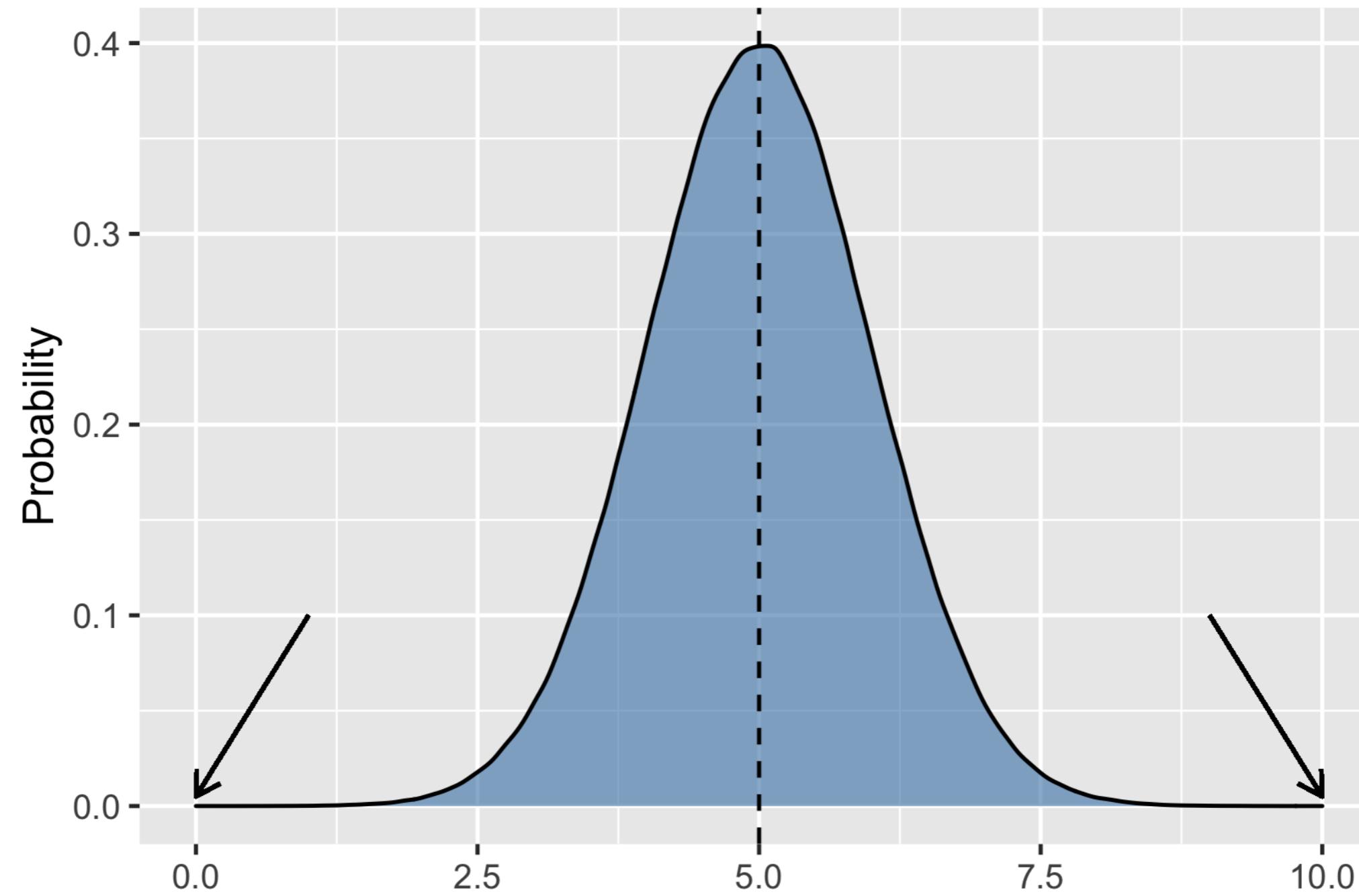
Symmetrical



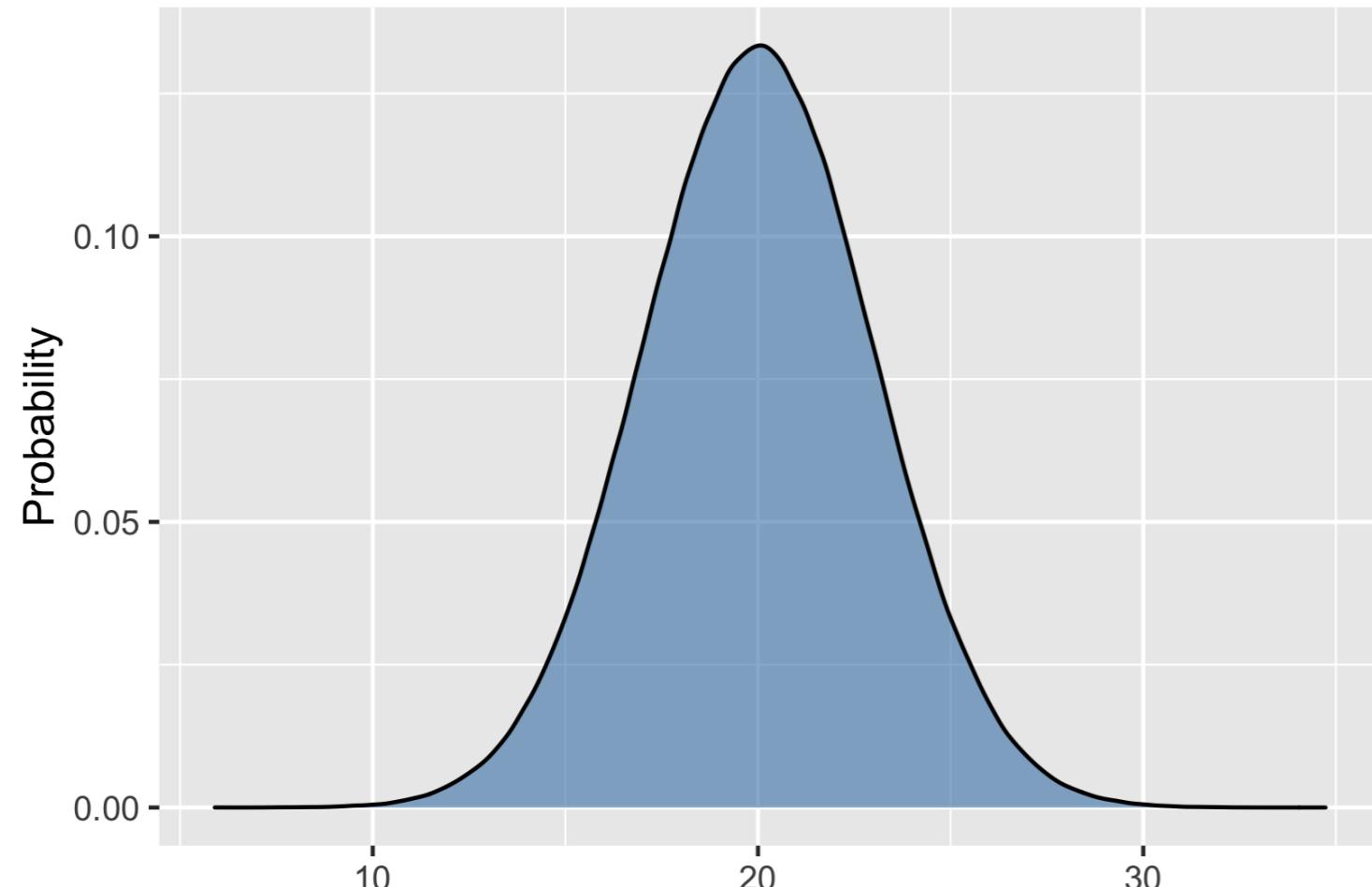
Area = 1



Curve never hits 0

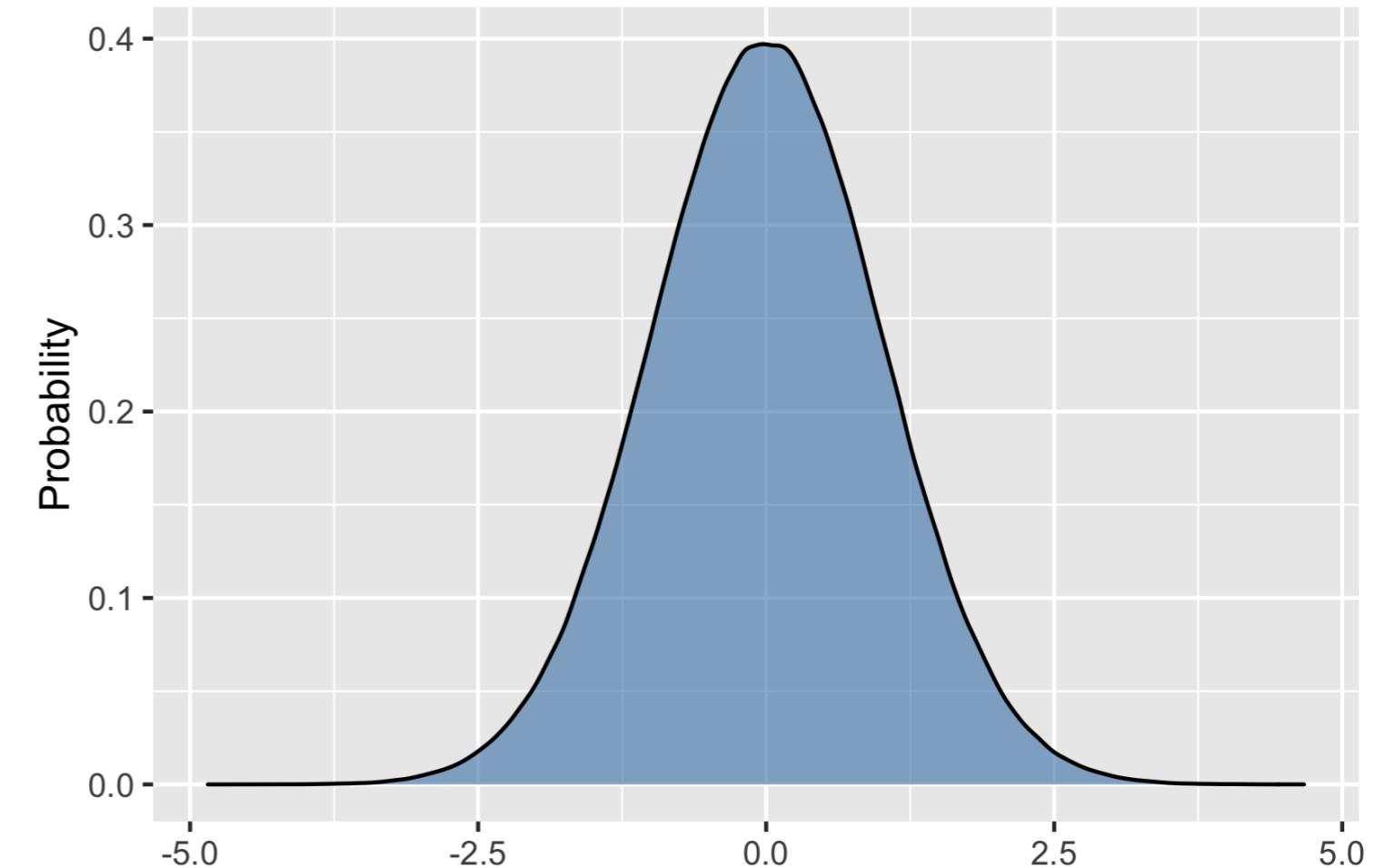


Described by mean and standard deviation



Mean = 20

Standard Deviation = 3

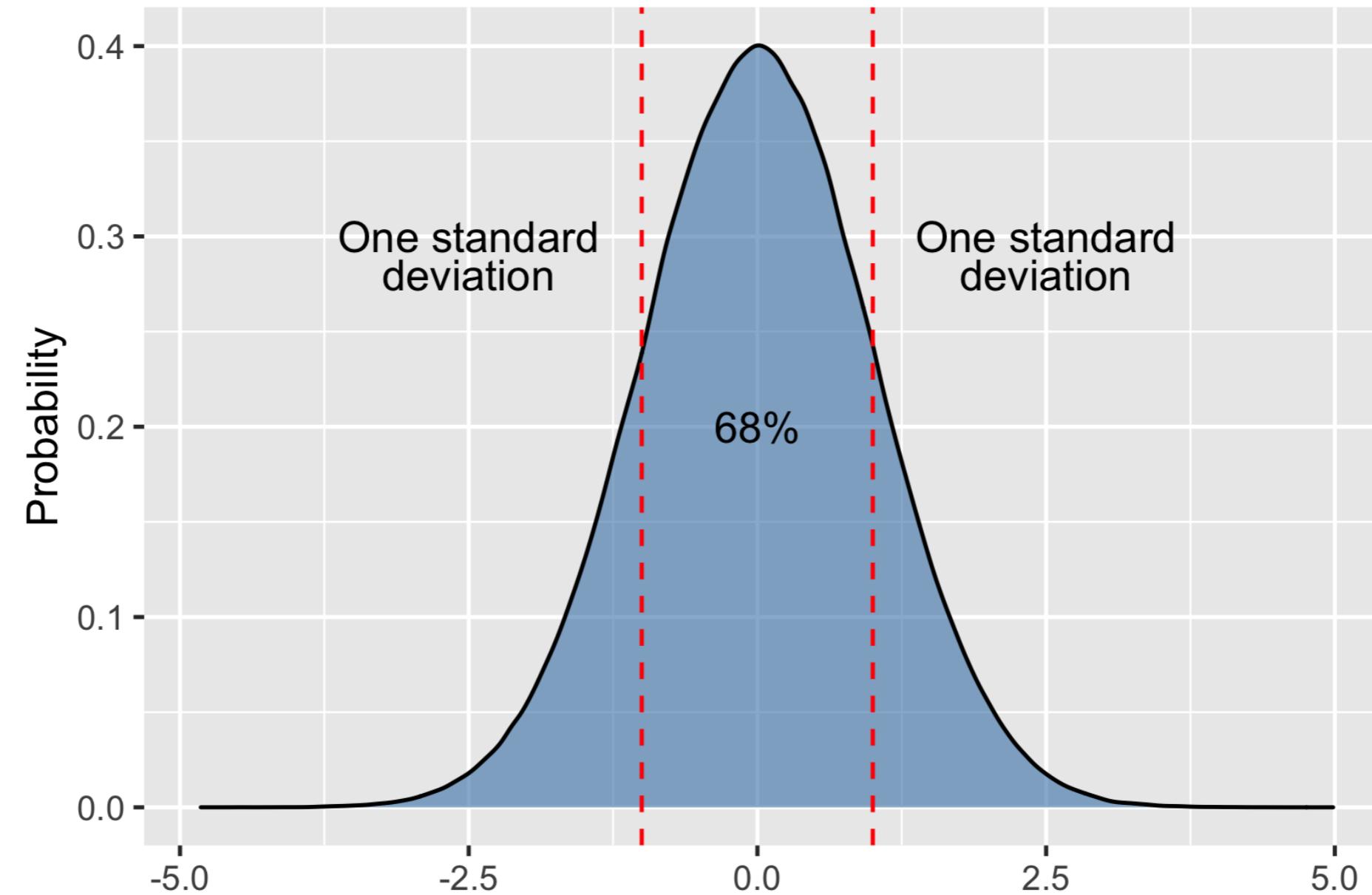


Mean = 0

Standard Deviation = 1

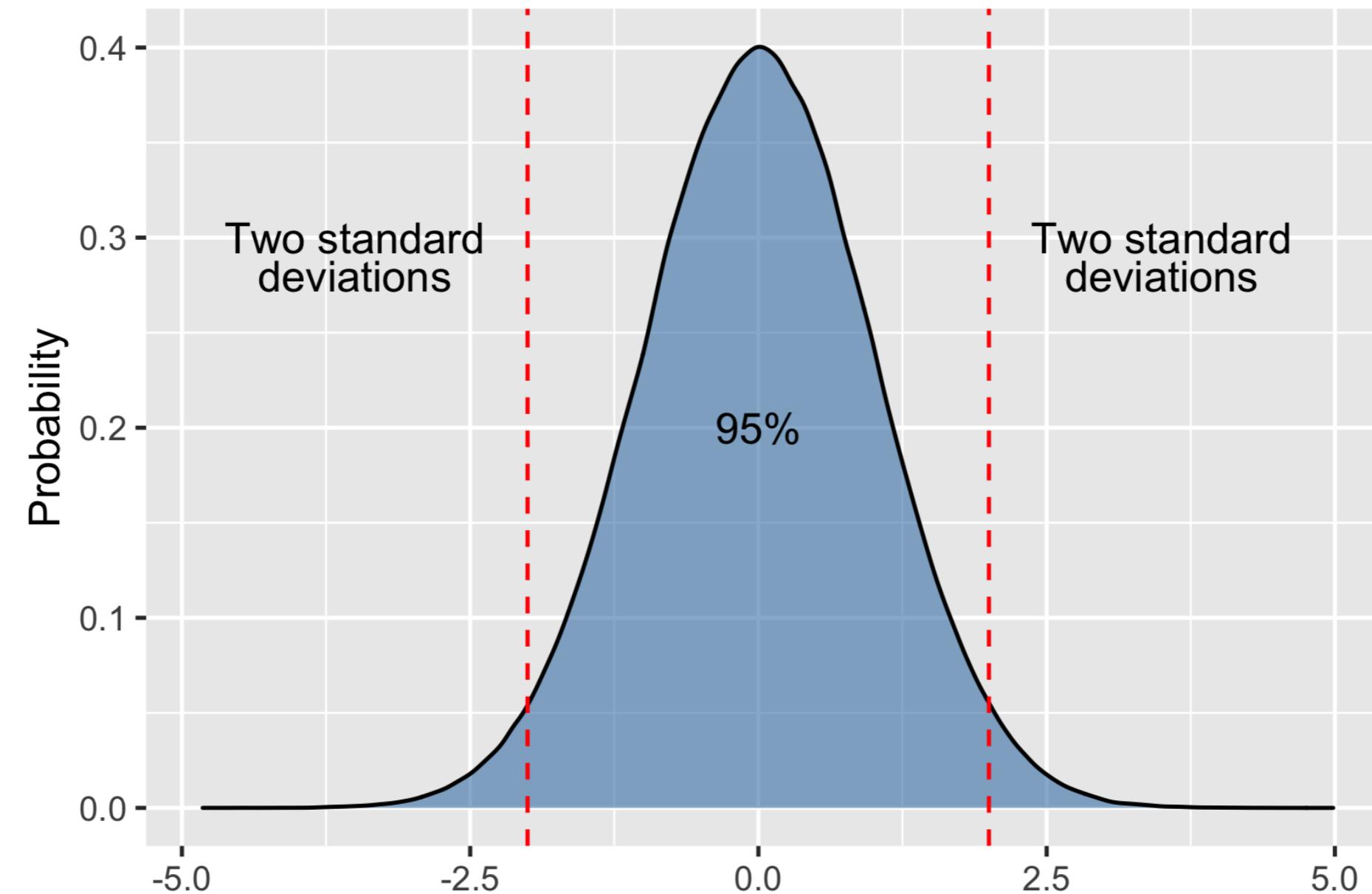
Areas under the normal distribution

68% falls within one standard deviation



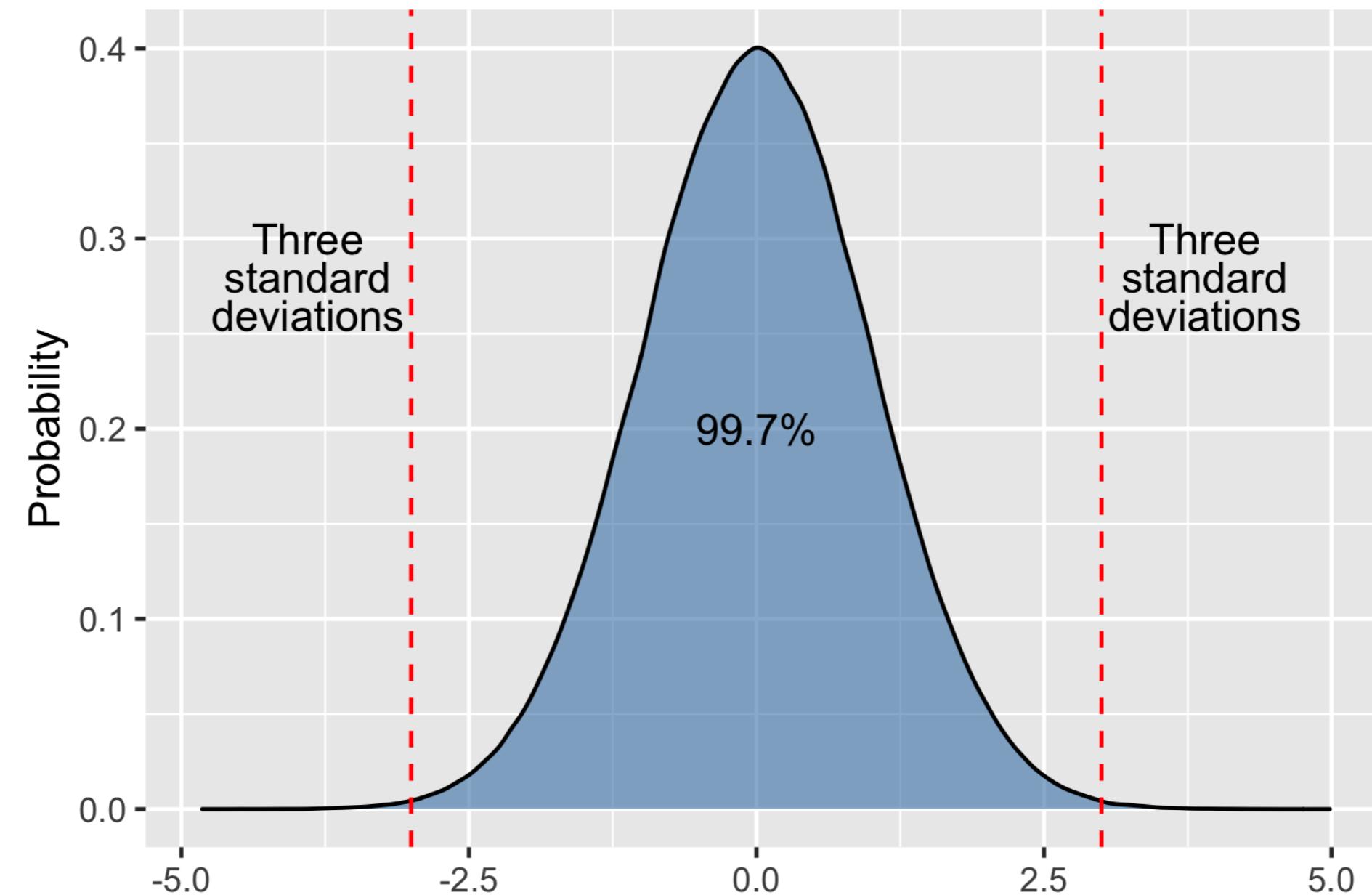
Areas under the normal distribution

95% falls within two standard deviations



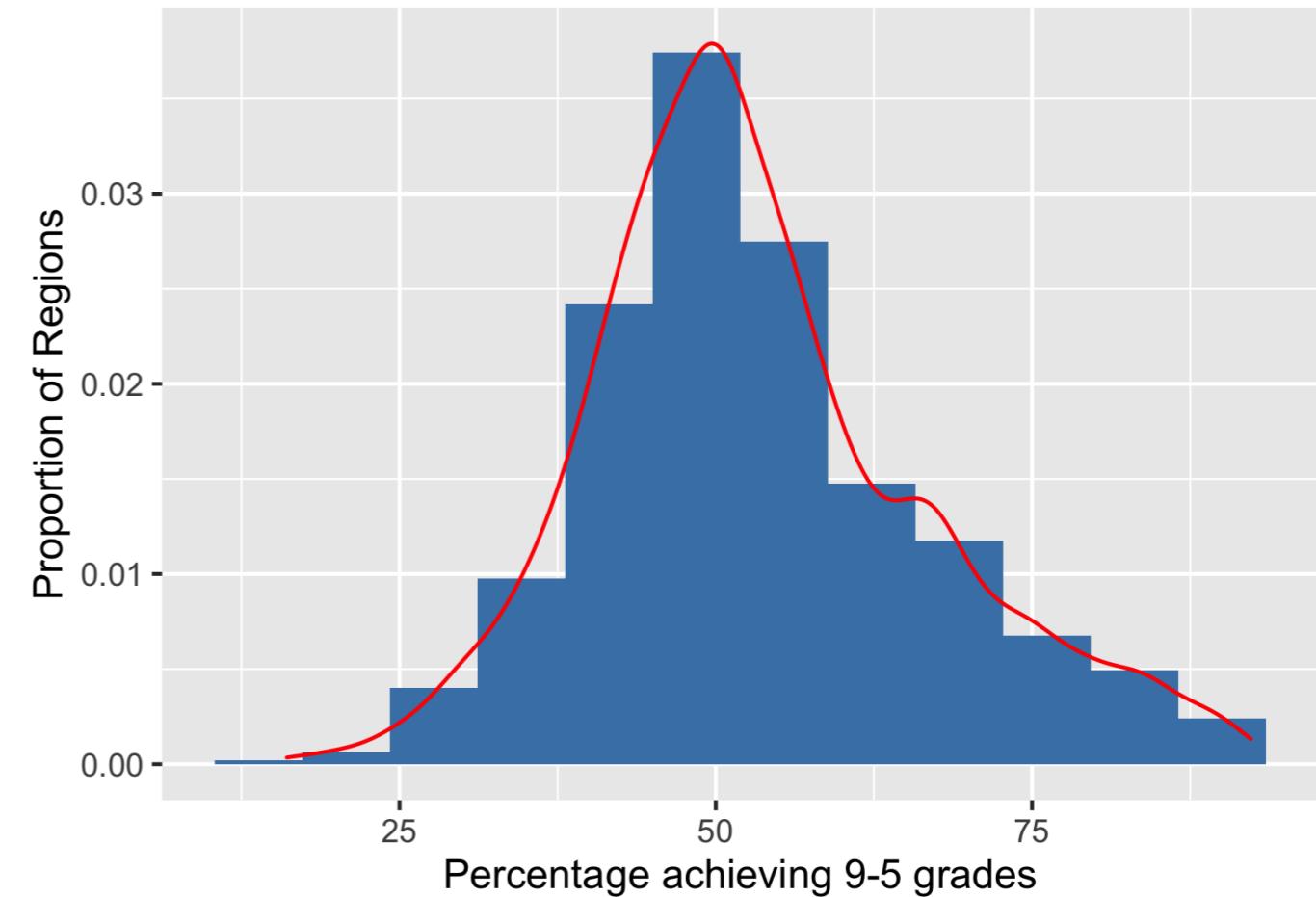
Areas under the normal distribution

99.7% falls within three standard deviations



Why is the normal distribution important?

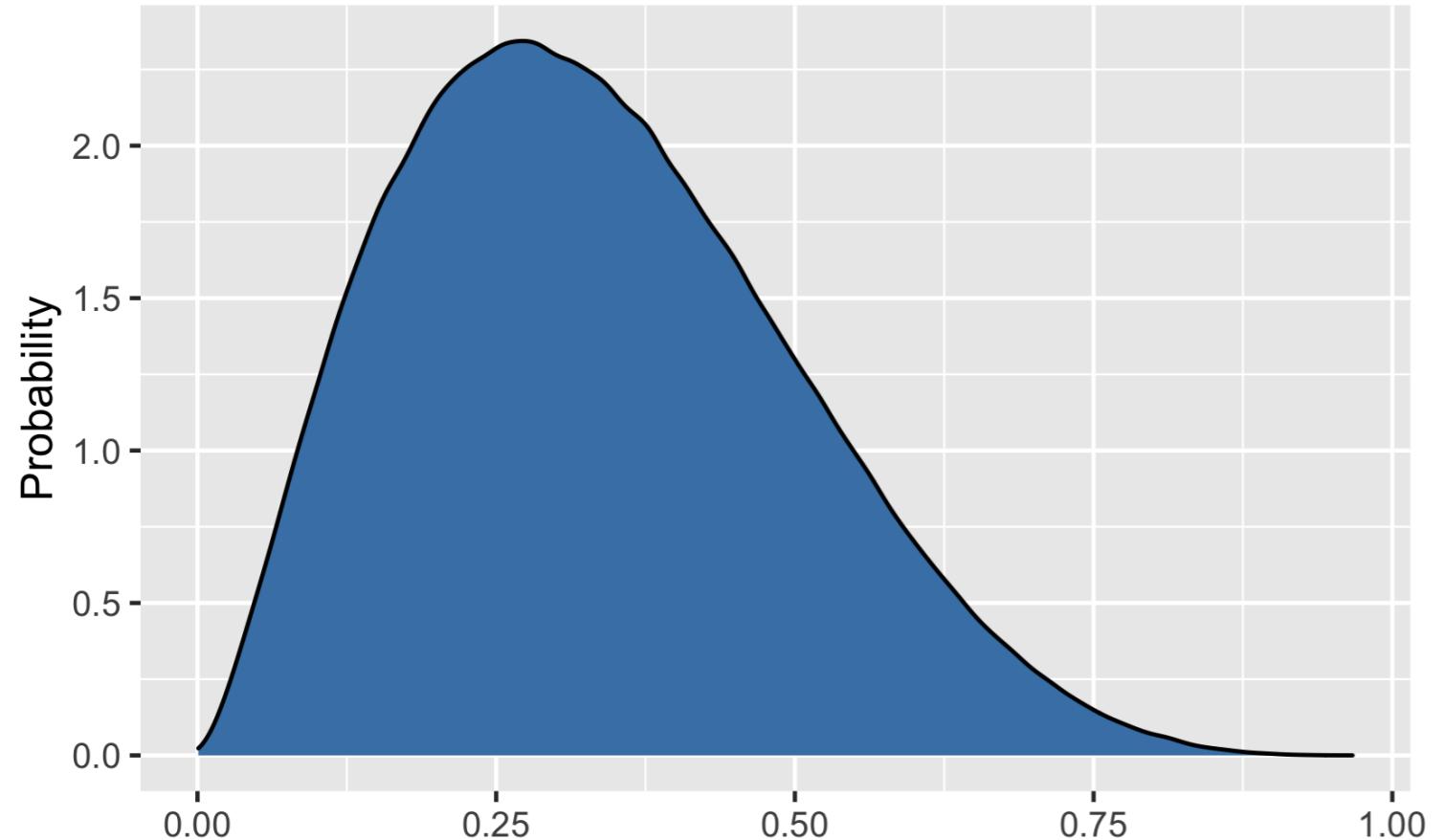
- Lots of real-world data resembles a normal distribution.
- A normal distribution is required for many statistical tests.



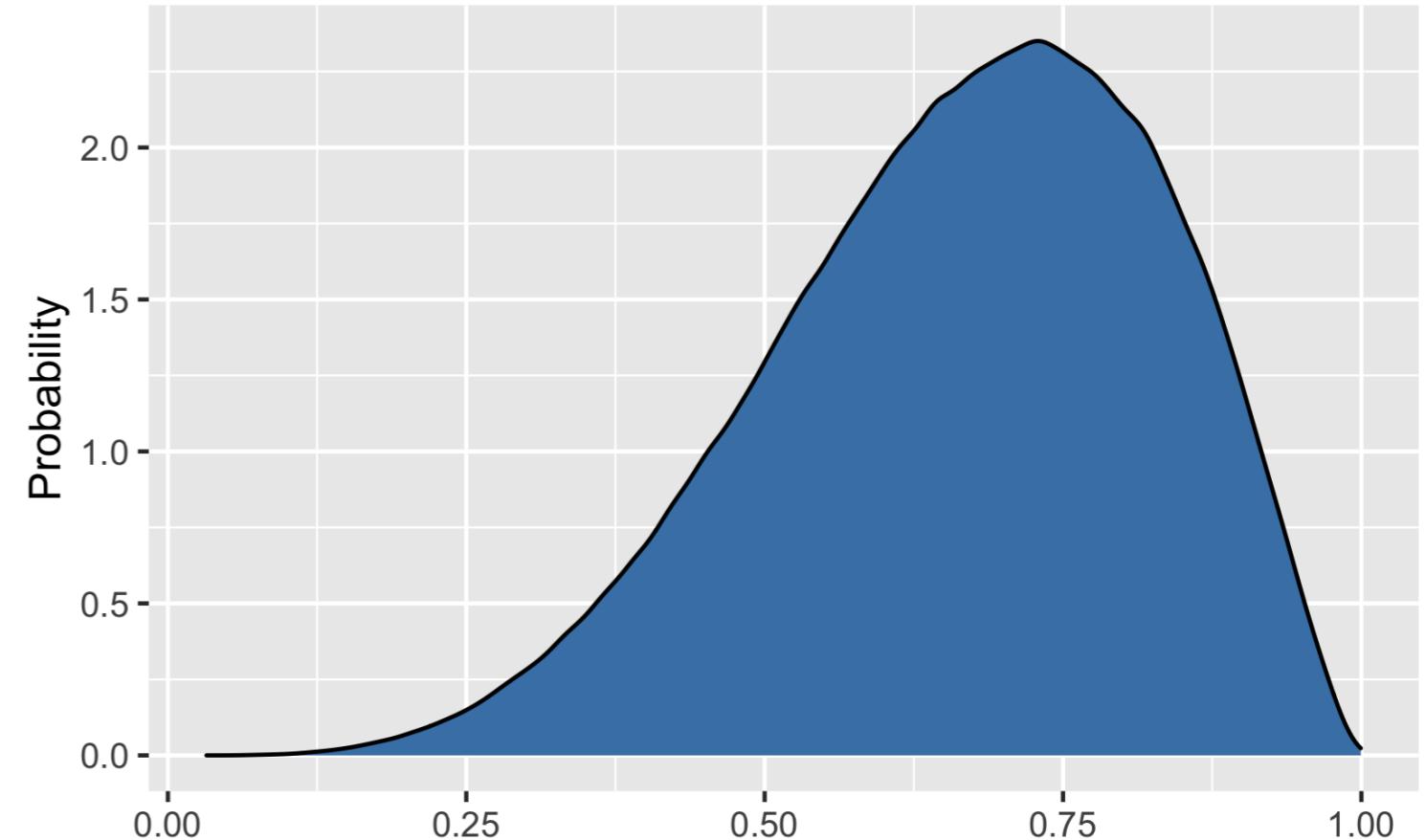
¹ Data source: <https://data.gov.uk/dataset/ec1efd76-d6ad-4594-9b4d-944aa4170e63/gcse-english-and-maths-results-by-ethnicity>

Skewness

Positive Skewed



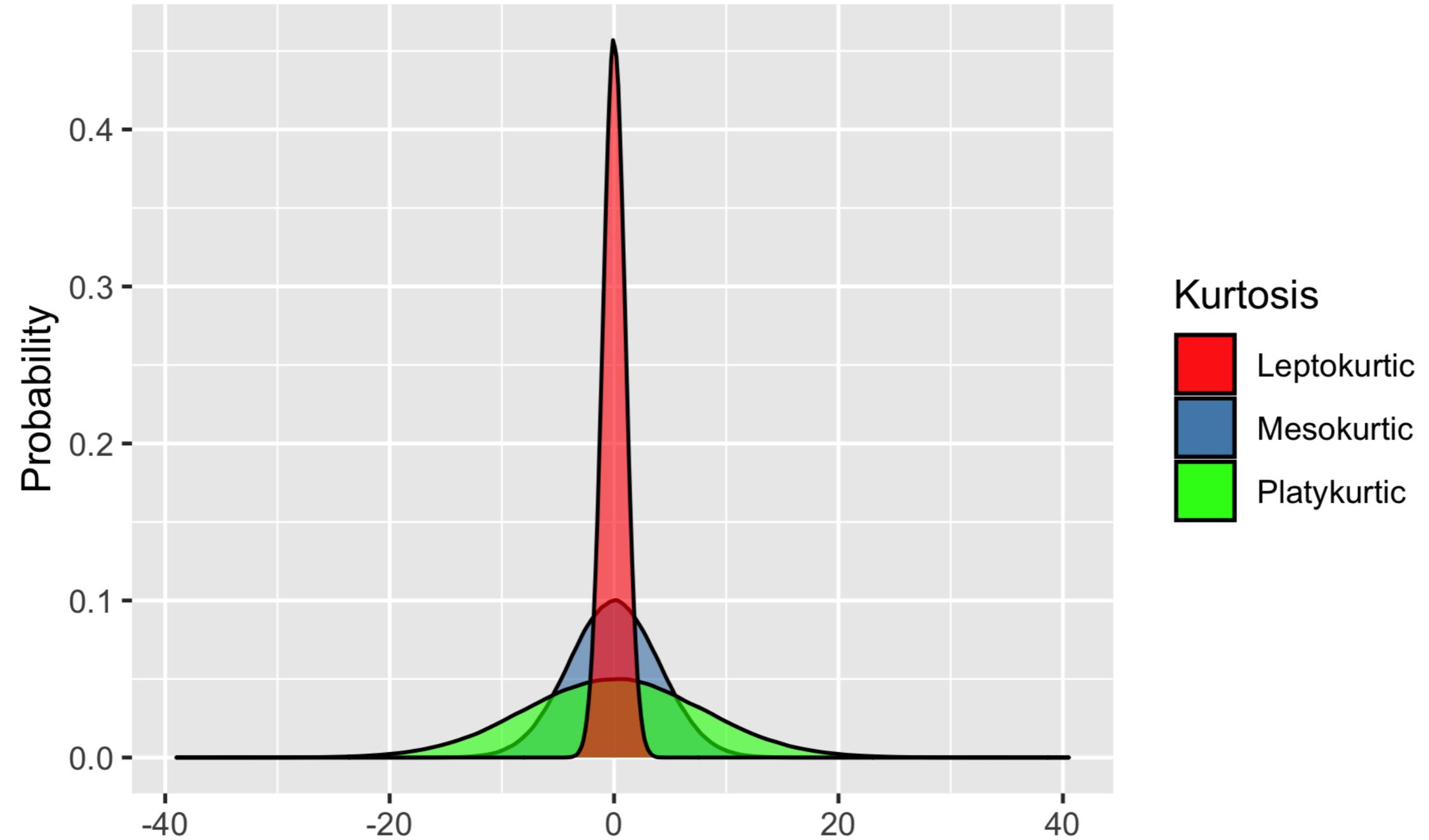
Negative Skewed



Kurtosis

- Kurtosis - a way of describing the occurrence of extreme values in a distribution.
- Three types of kurtosis

Kurtosis



Let's practice!

INTRODUCTION TO STATISTICS

The central limit theorem

INTRODUCTION TO STATISTICS



George Boorman

Curriculum Manager, DataCamp

Rolling a die five times



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

$$Mean(Results) = 2$$

Rolling a die five times

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

$$\text{Mean}(\text{Results}) = 4.4$$

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

$$\text{Mean}(\text{Results}) = 3.2$$

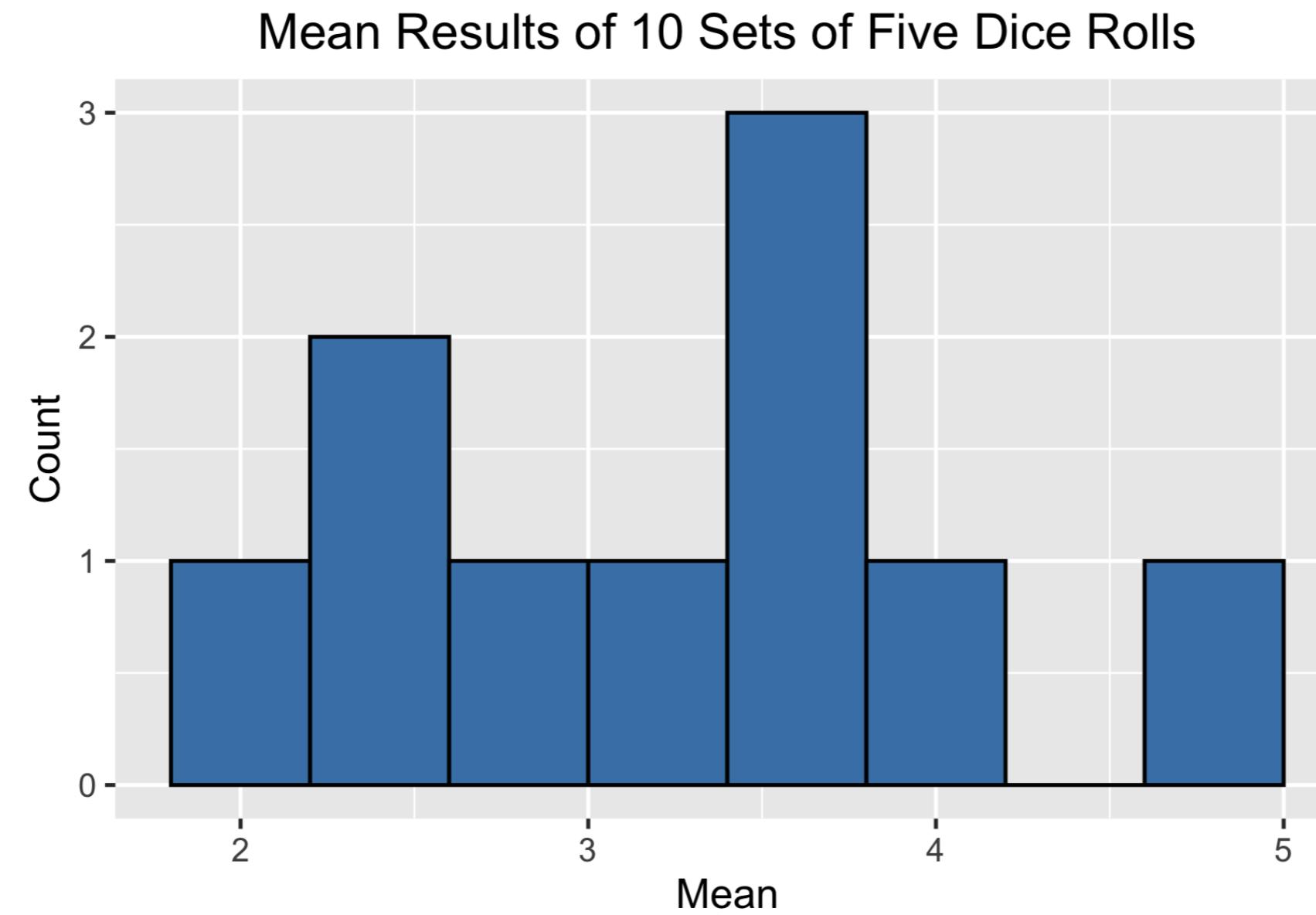
10 sets of five die rolls

- Roll a die five times
- Record the mean
- Repeat 10 times

Set	Mean
1	3.8
2	4.0
3	3.8
4	3.6
5	3.2
6	4.8
7	2.6
8	3.0
9	2.6
10	2.0

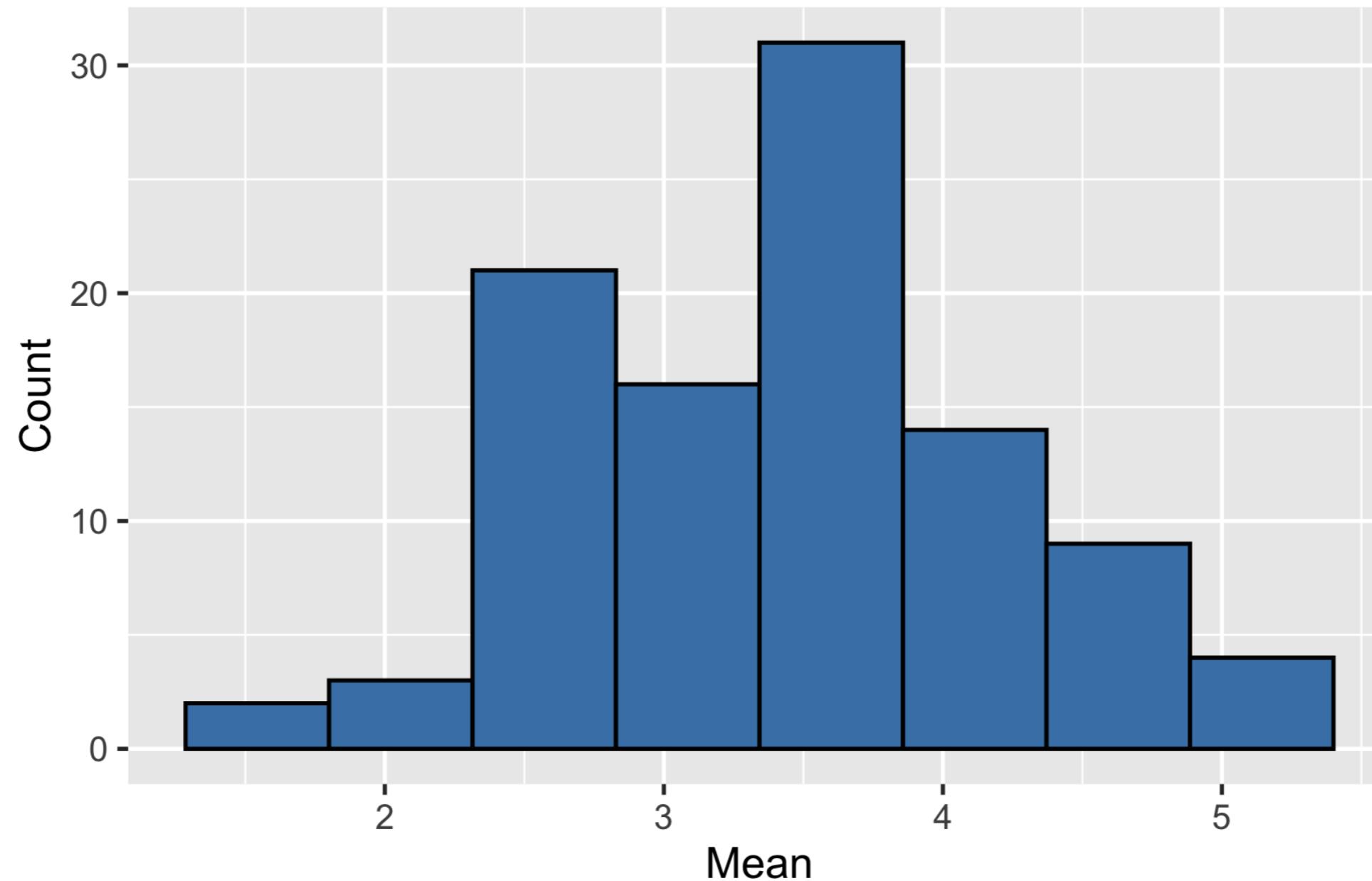
Sampling distributions

Sampling distribution of the sample mean



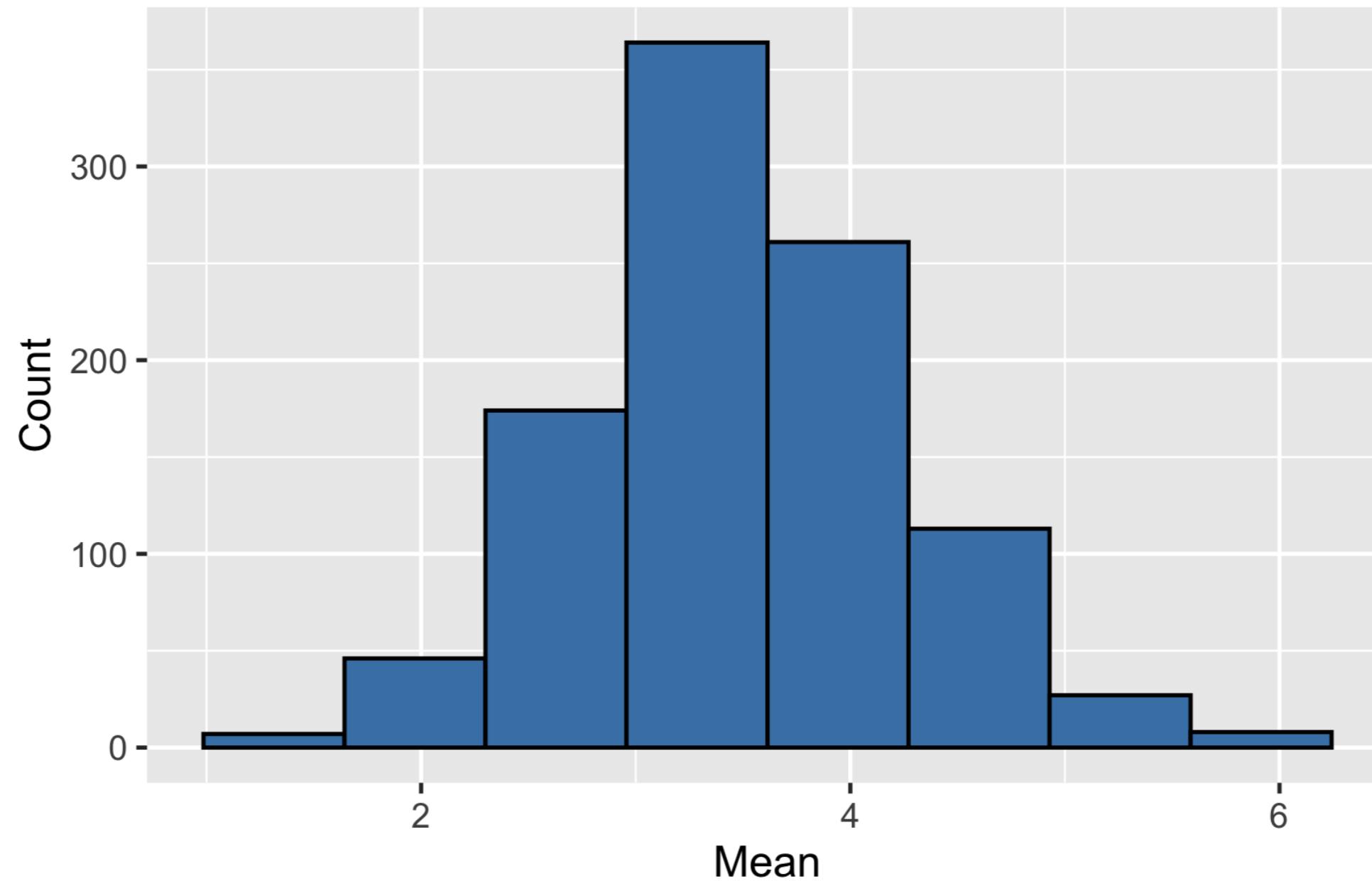
100 sample means

100 Sample Means



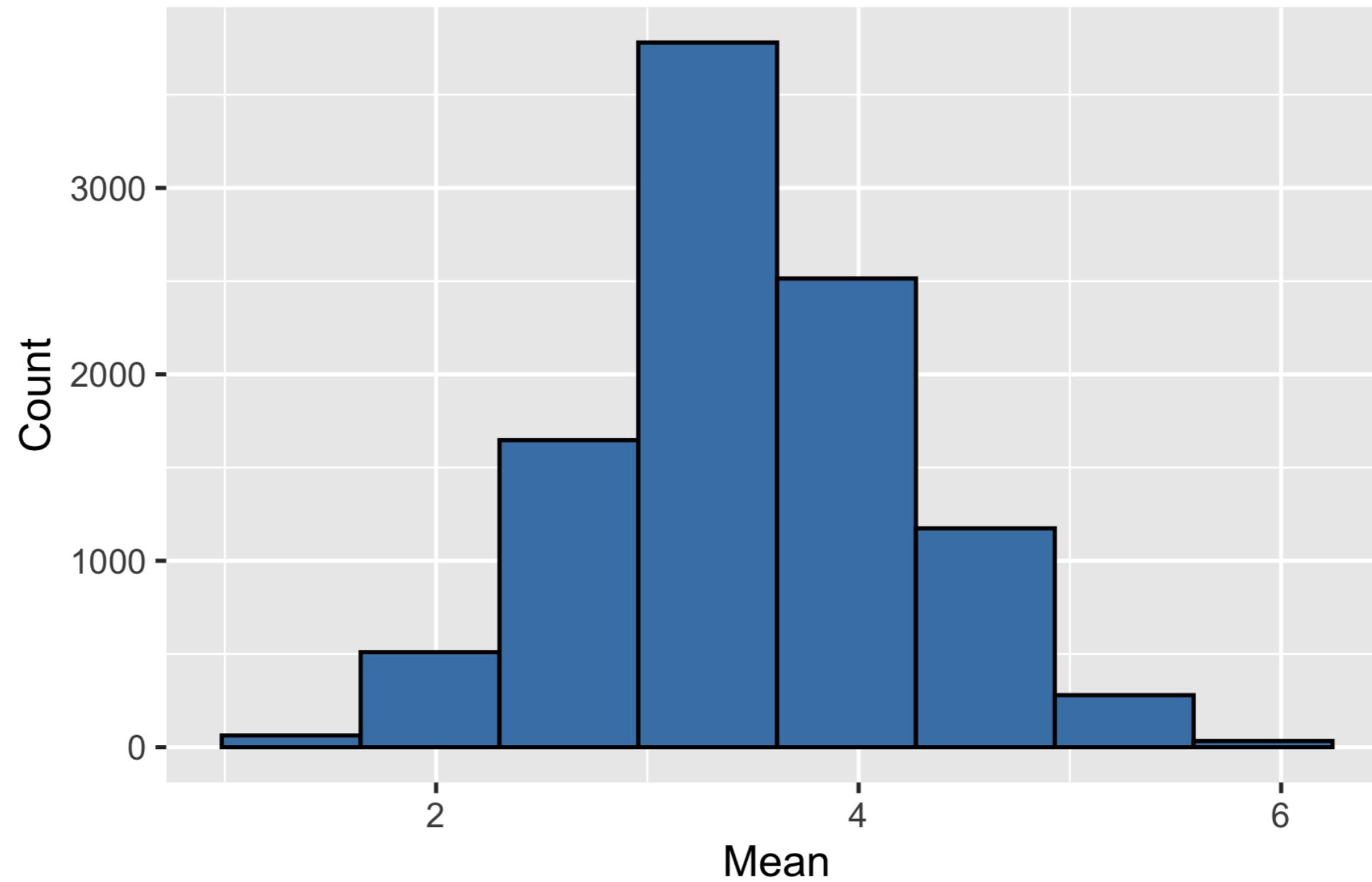
1000 sample means

1,000 Sample Means



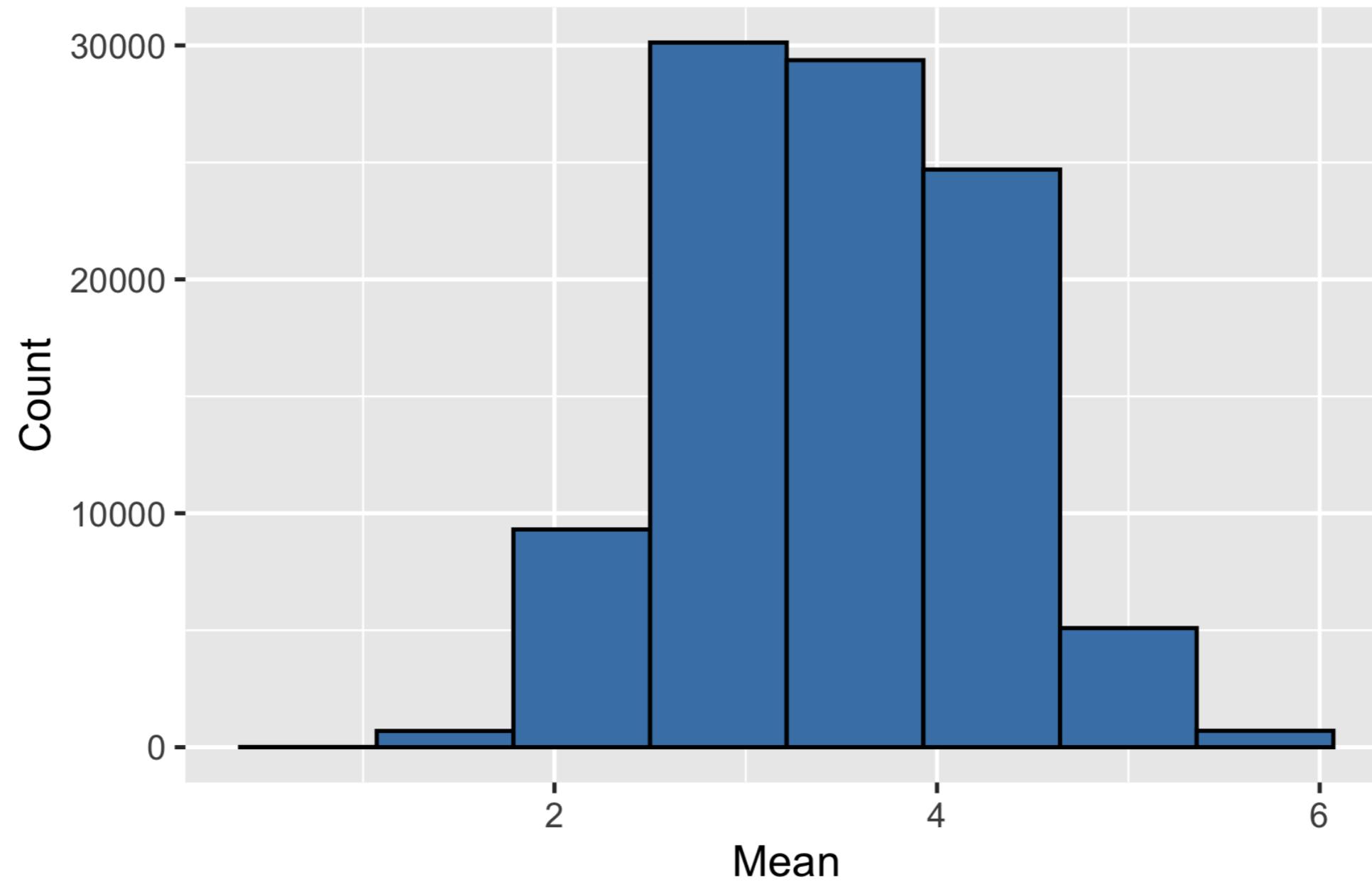
10000 sample means

10,000 Sample Means



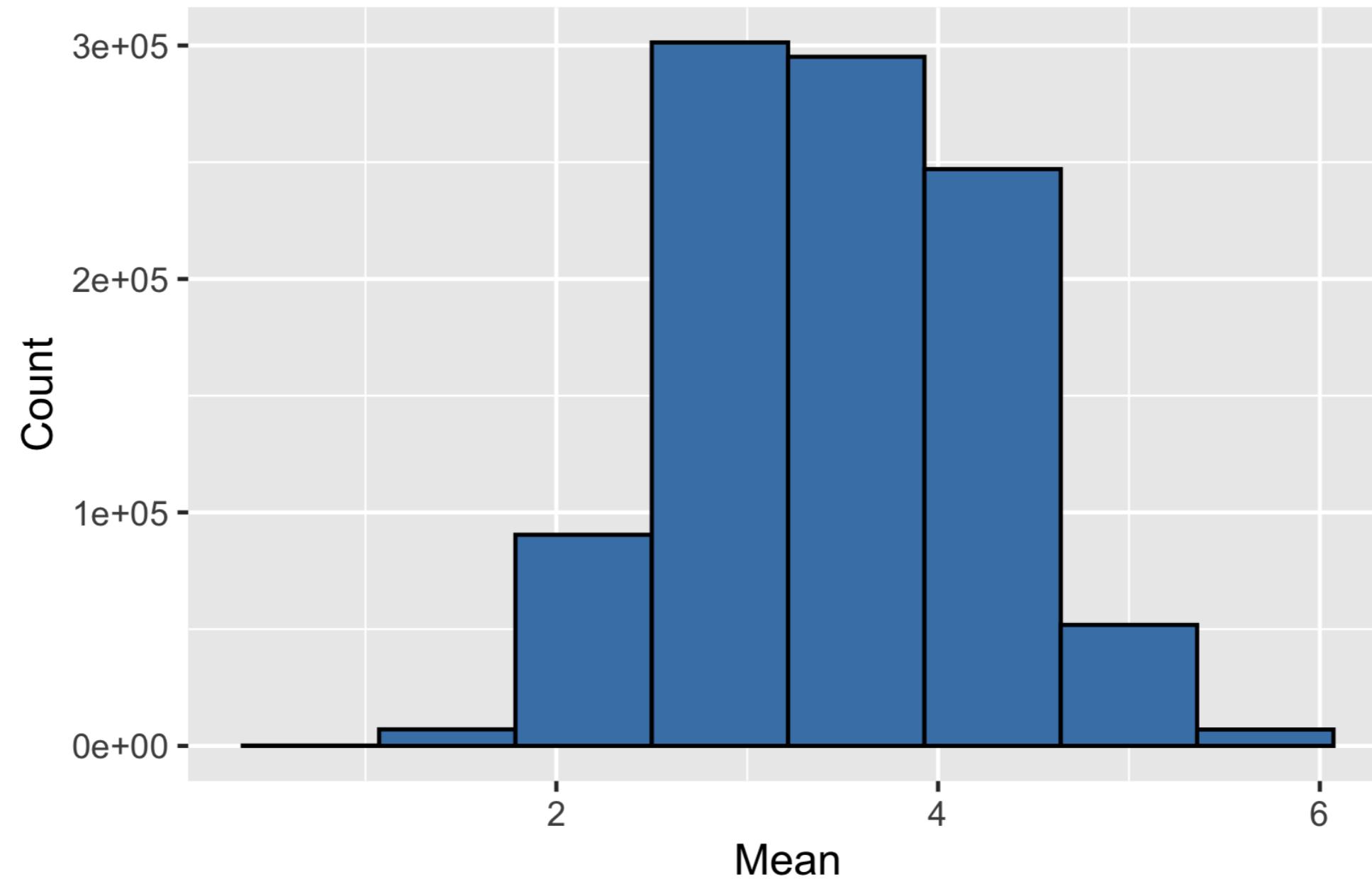
100000 sample means

100,000 Sample Means



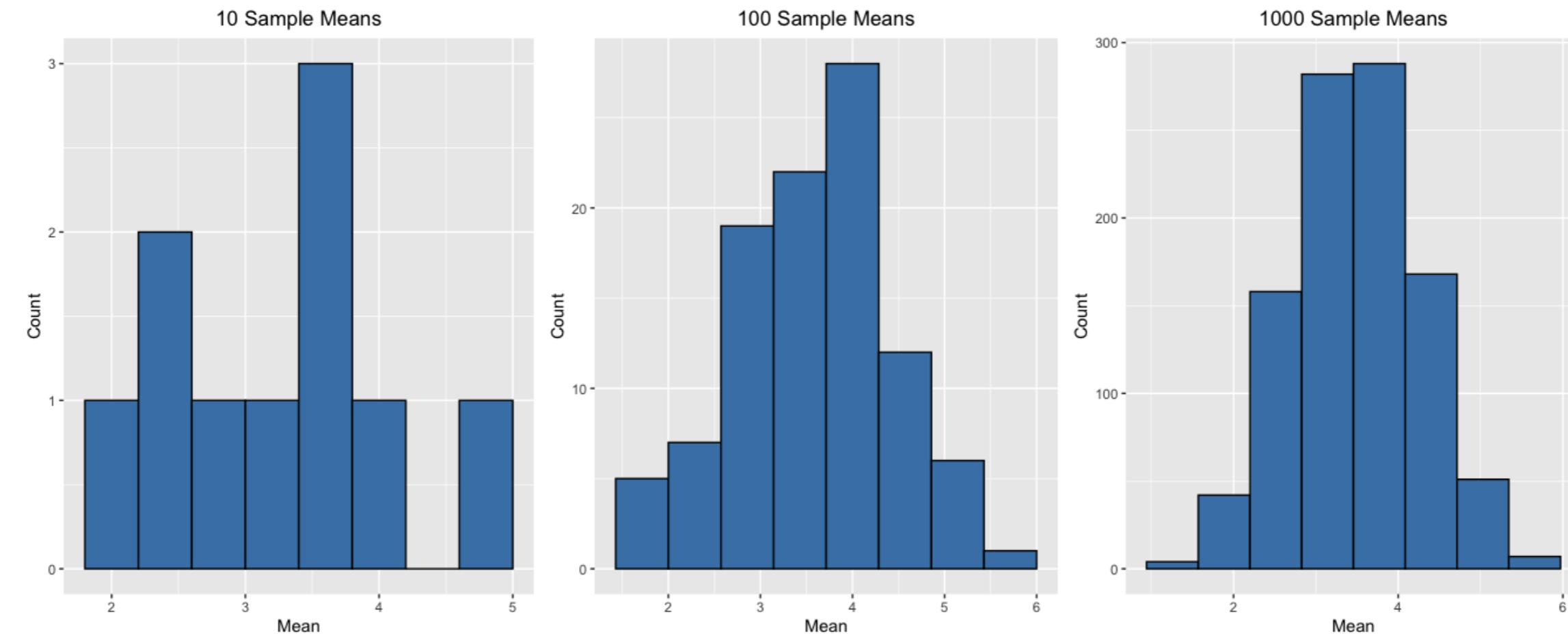
One million sample means

1,000,000 Sample Means



Central limit theorem

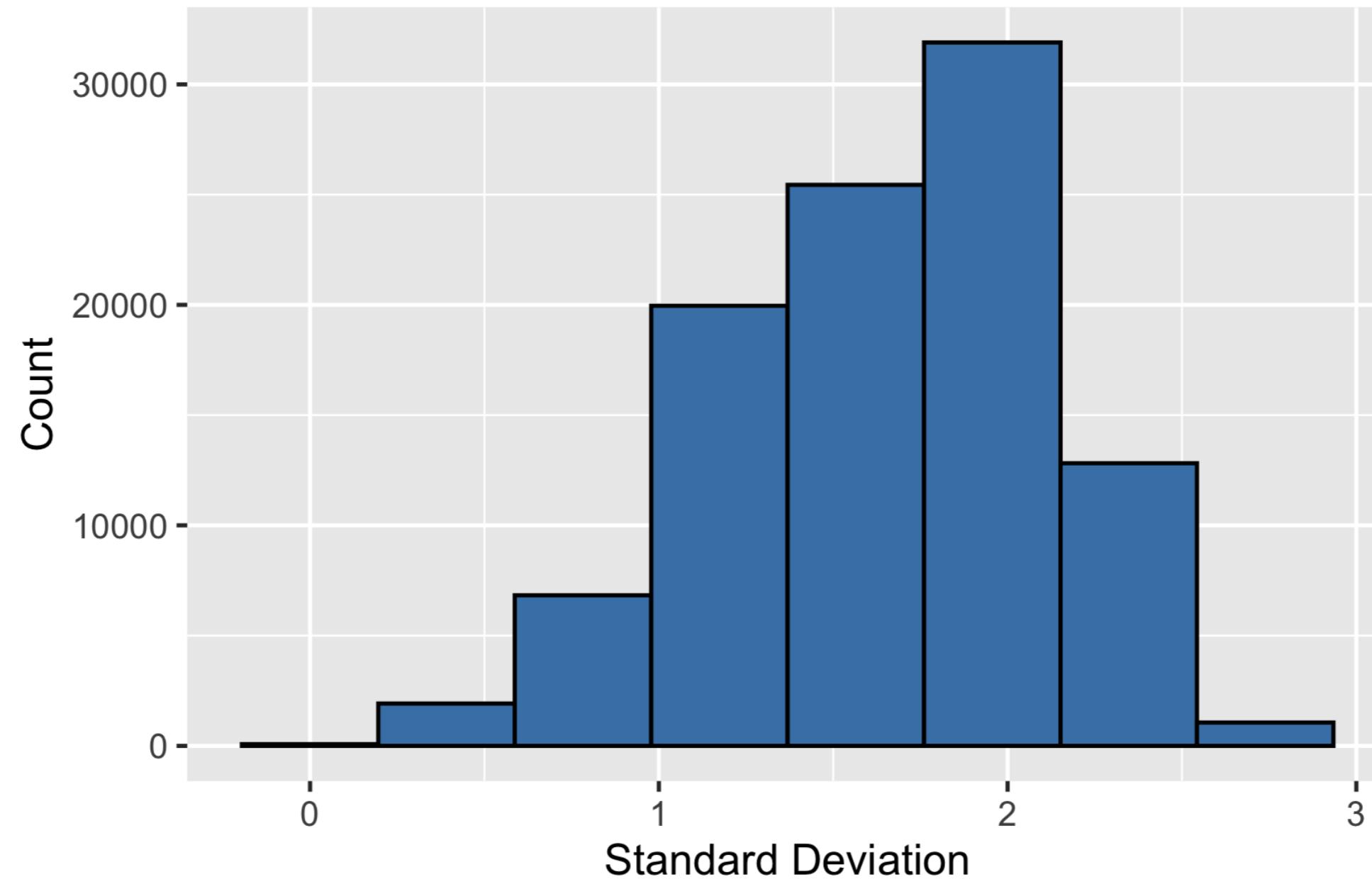
The sampling distribution of a statistic becomes closer to the normal distribution as the size of the sample increases.



* *Samples should be random and independent*

Standard deviation and the CLT

Sampling Distribution of the Standard Deviation



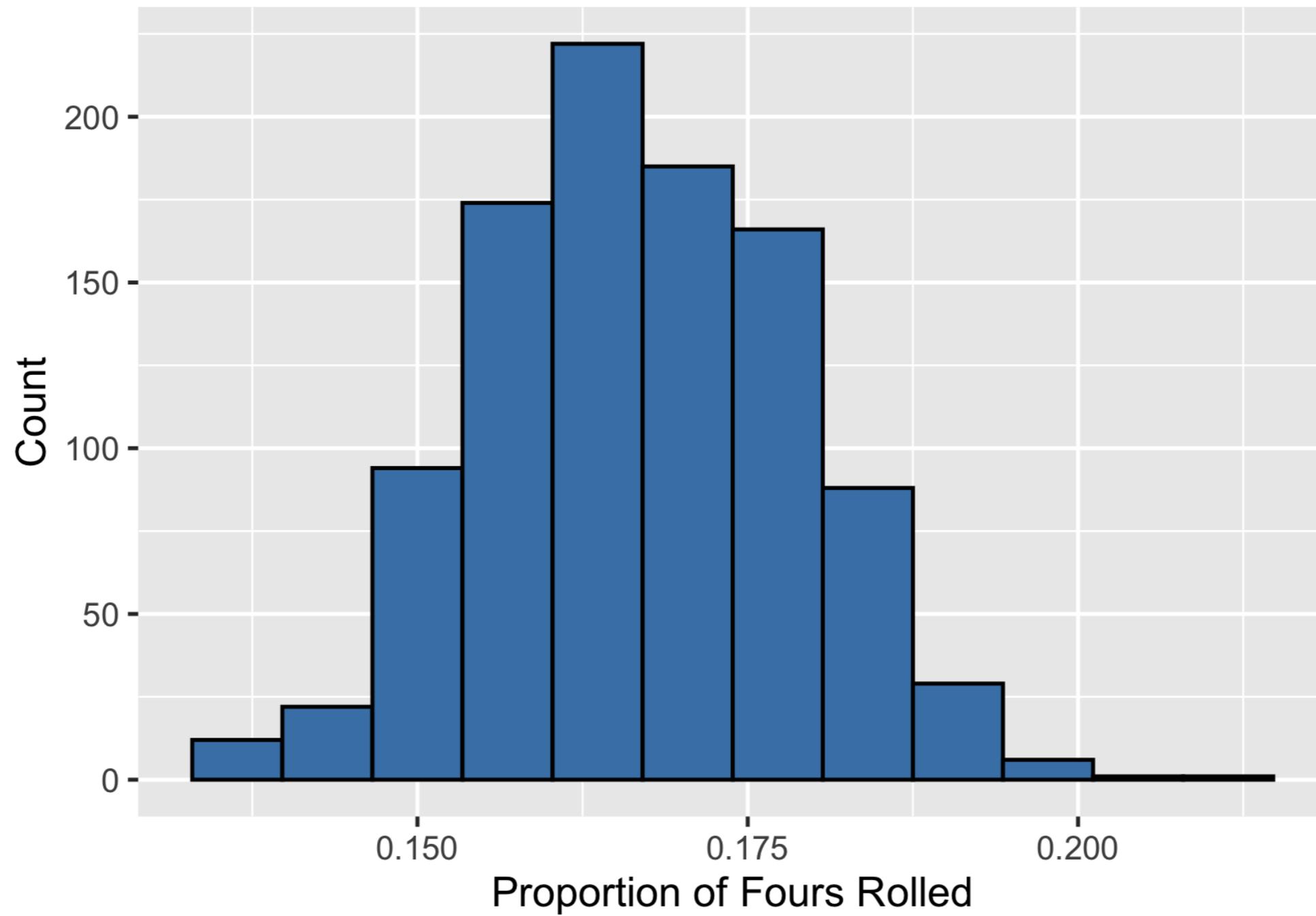
Proportions and the CLT

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

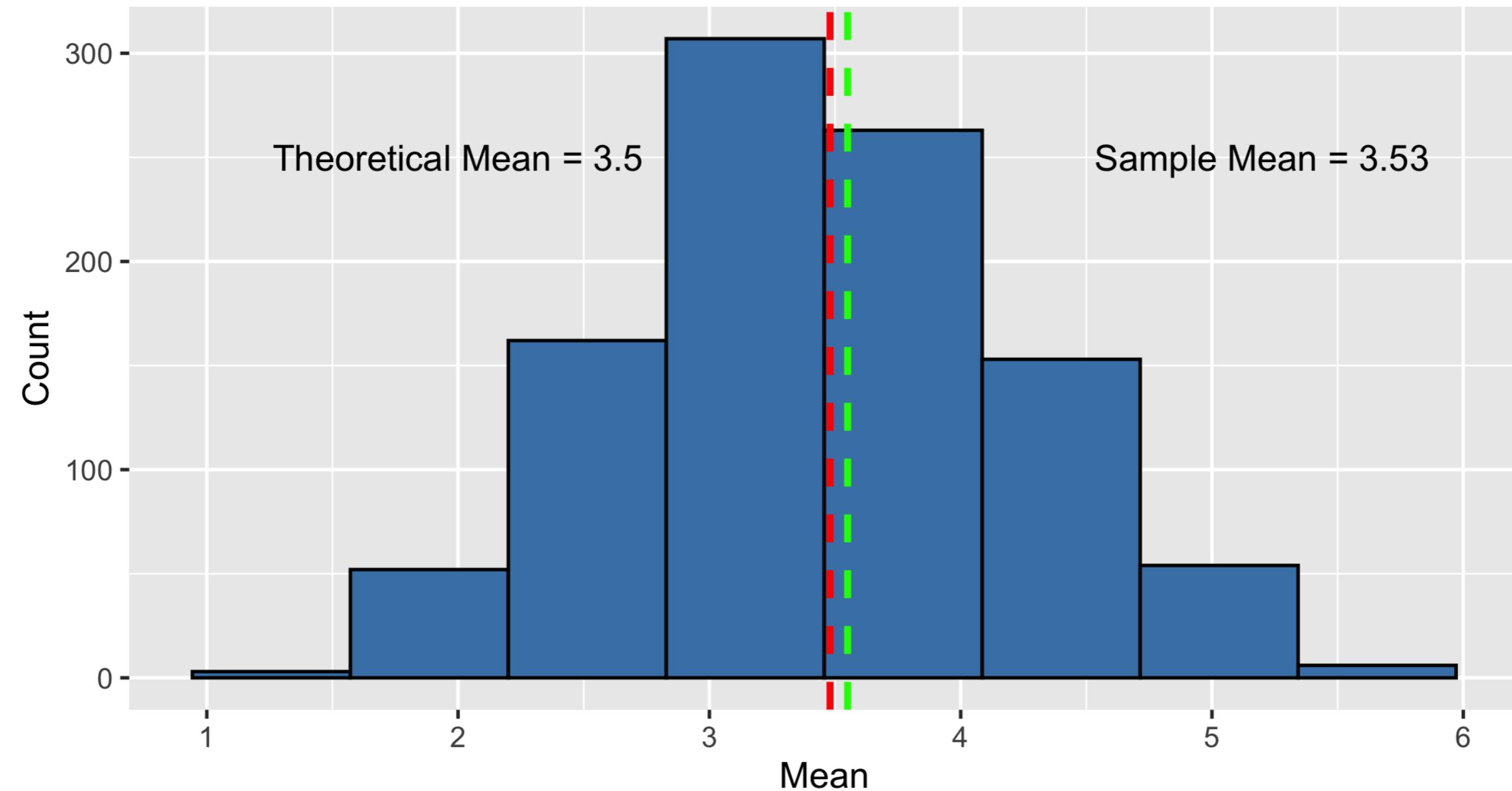
Set	Mean
1	4
2	4
3	1
4	4
5	3

- $\frac{1}{5}$ or 20% are a 4
- $\frac{3}{5}$ or 60% are a 4

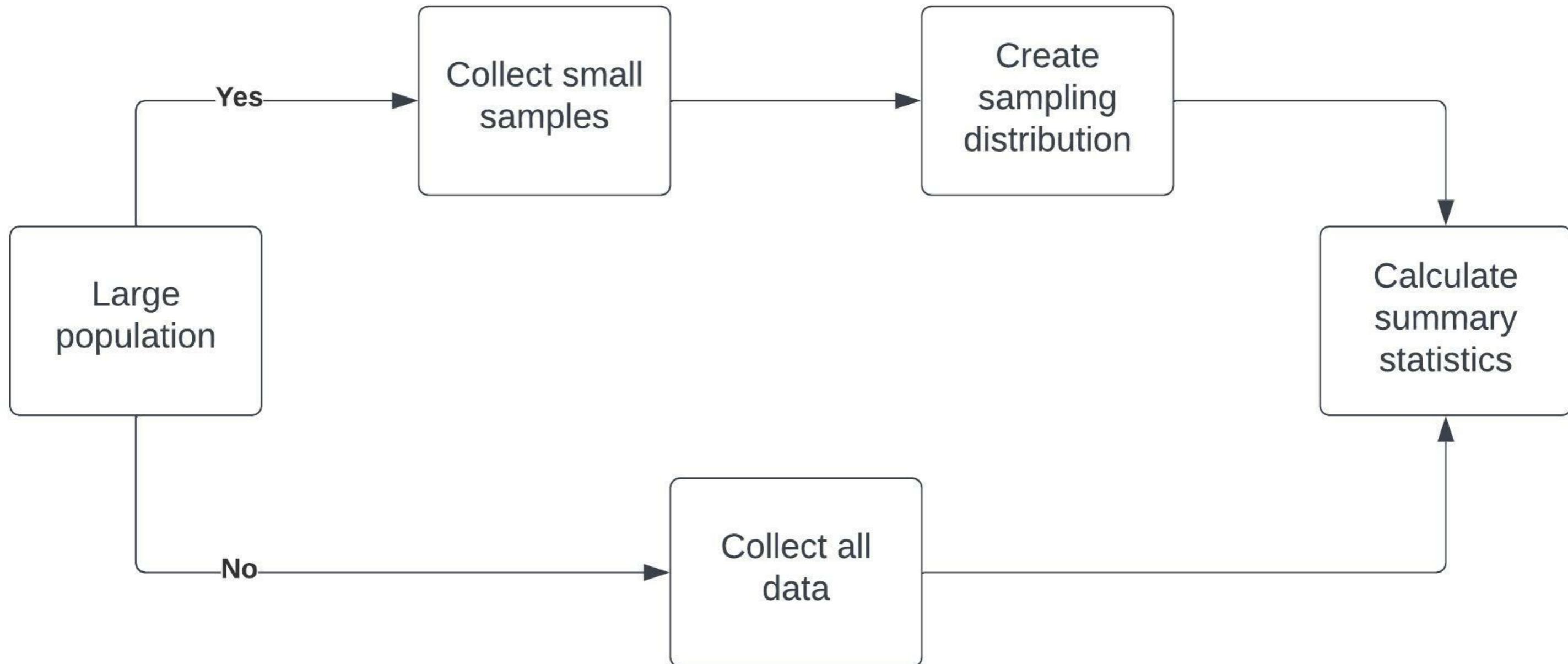
Sampling distribution of proportion



Mean of the sampling distribution



Benefits of the central limit theorem



Let's practice!

INTRODUCTION TO STATISTICS

The Poisson distribution

INTRODUCTION TO STATISTICS



George Boorman

Curriculum Manager, DataCamp

Poisson processes

- Average # of events in a period is known
 - but the time or space between events is random
- Poisson processes
 - Number of animals adopted from an animal shelter per week
 - Number of people arriving at a restaurant per hour
 - Number of website visits in a day



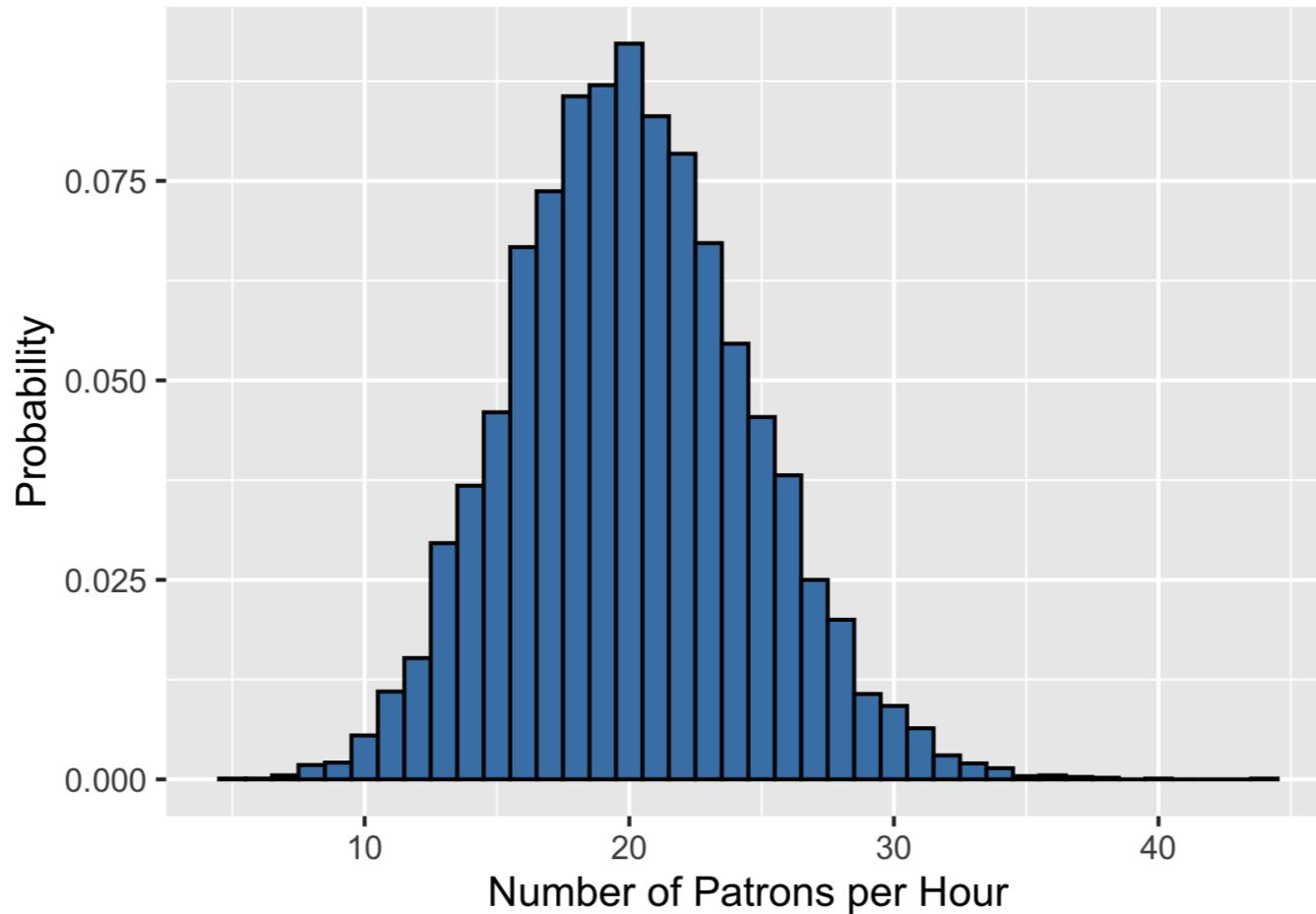
¹ Image credit: <https://unsplash.com/@rodlong>

Poisson distribution

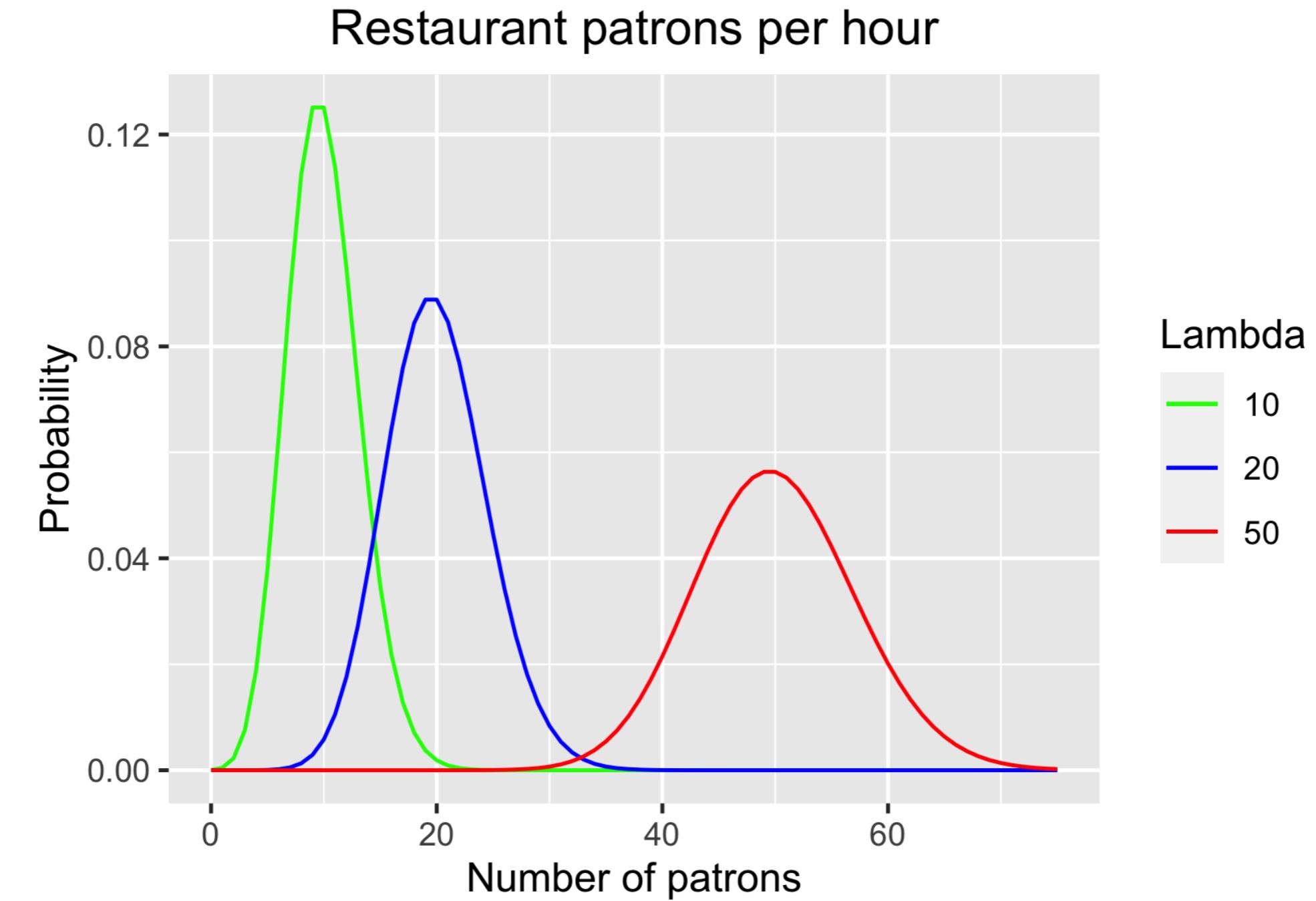
- Probability of some # of events occurring over a fixed period of time
- Examples
 - Probability of at least five animals adopted from an animal shelter per week
 - Probability of 12 people arriving at a restaurant per hour
 - Probability of less than 200 visits to a website in a day

Lambda (λ)

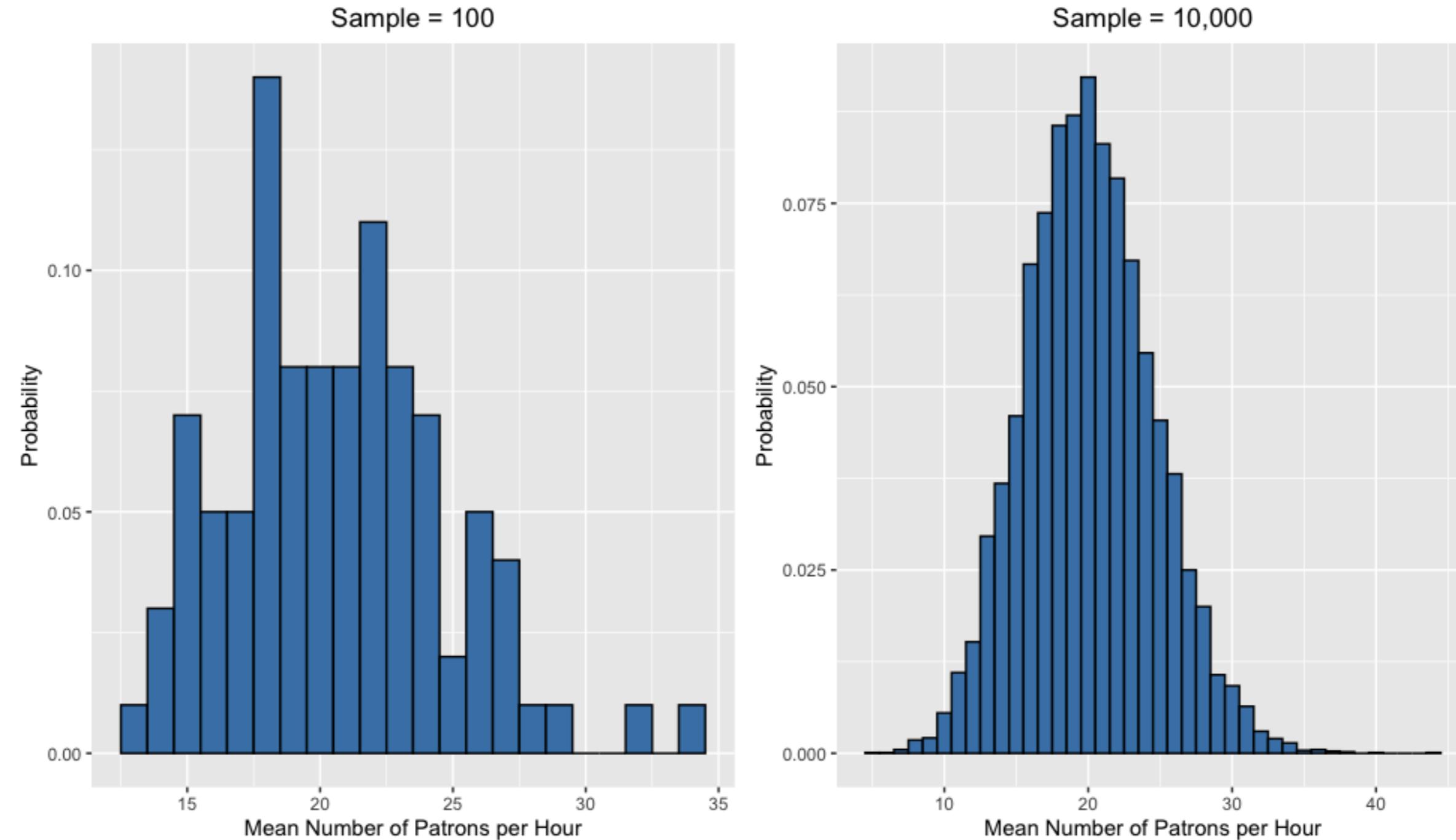
- λ = average number of events per time interval
 - Average number of patrons per hour = 20
 - λ = the expected value of the distribution!



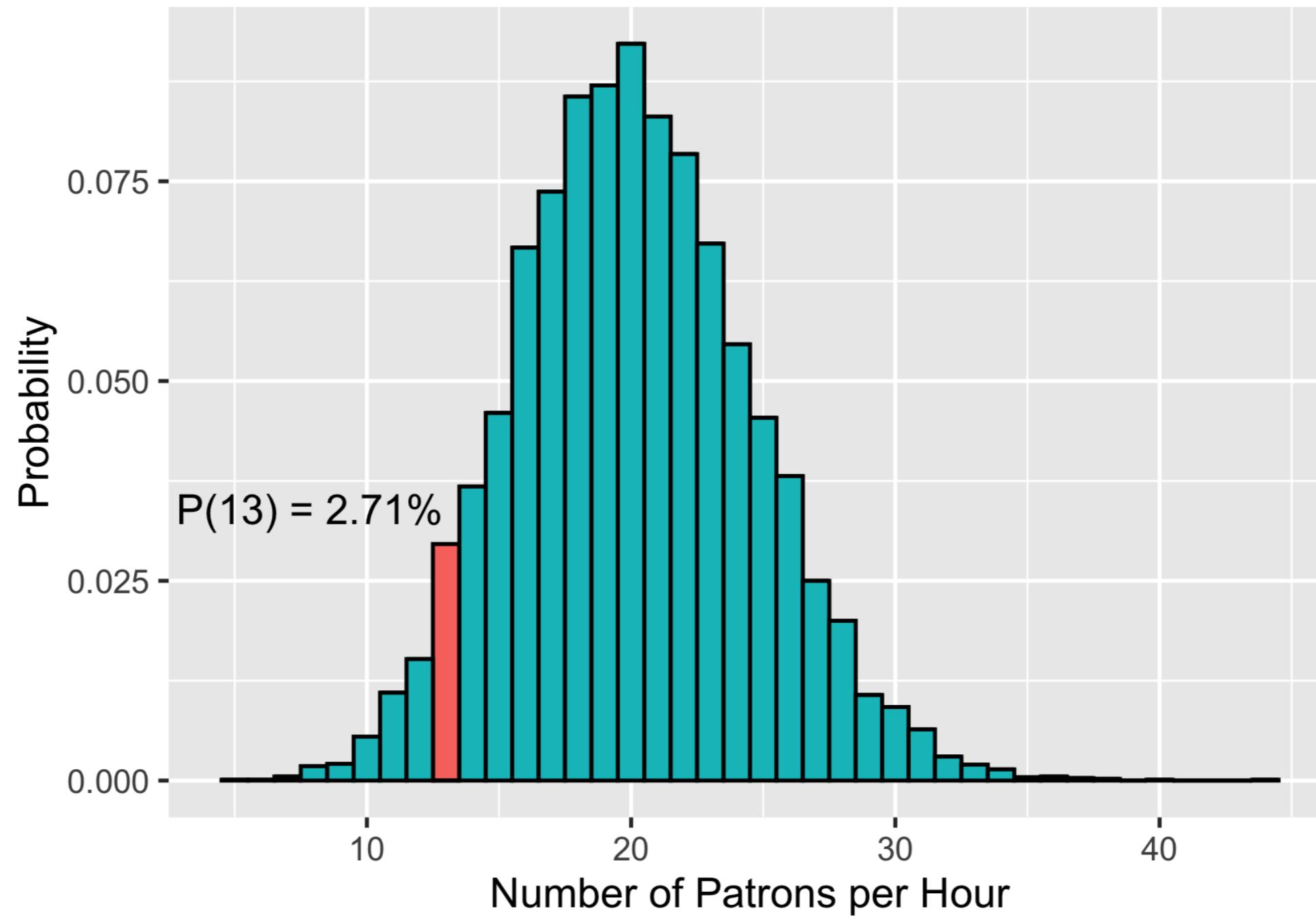
Lambda is the distribution's peak



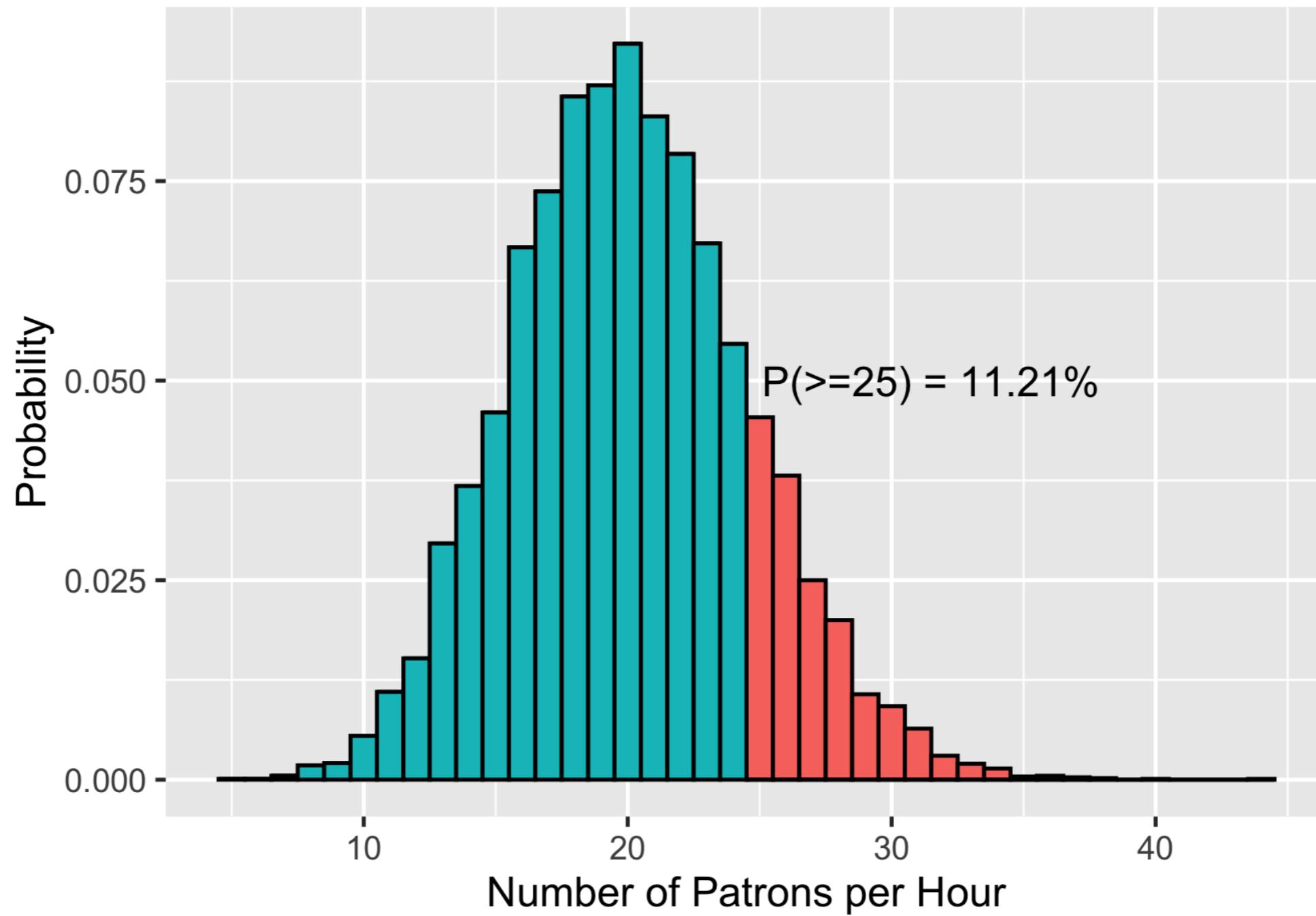
Central limit theorem still applies!



Probability of 13 patrons in an hour



Probability of 25 or more patrons



Let's practice!

INTRODUCTION TO STATISTICS