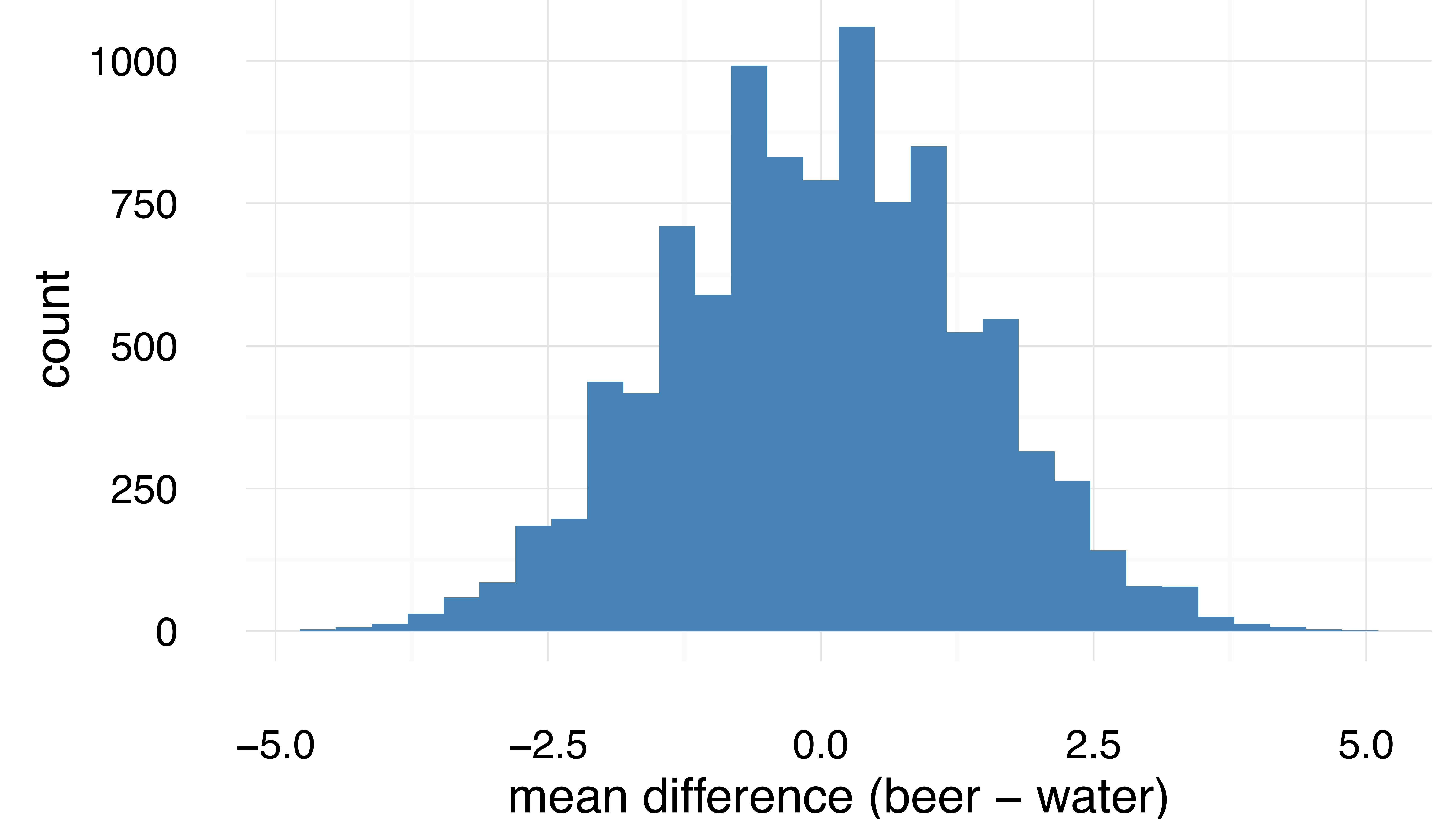
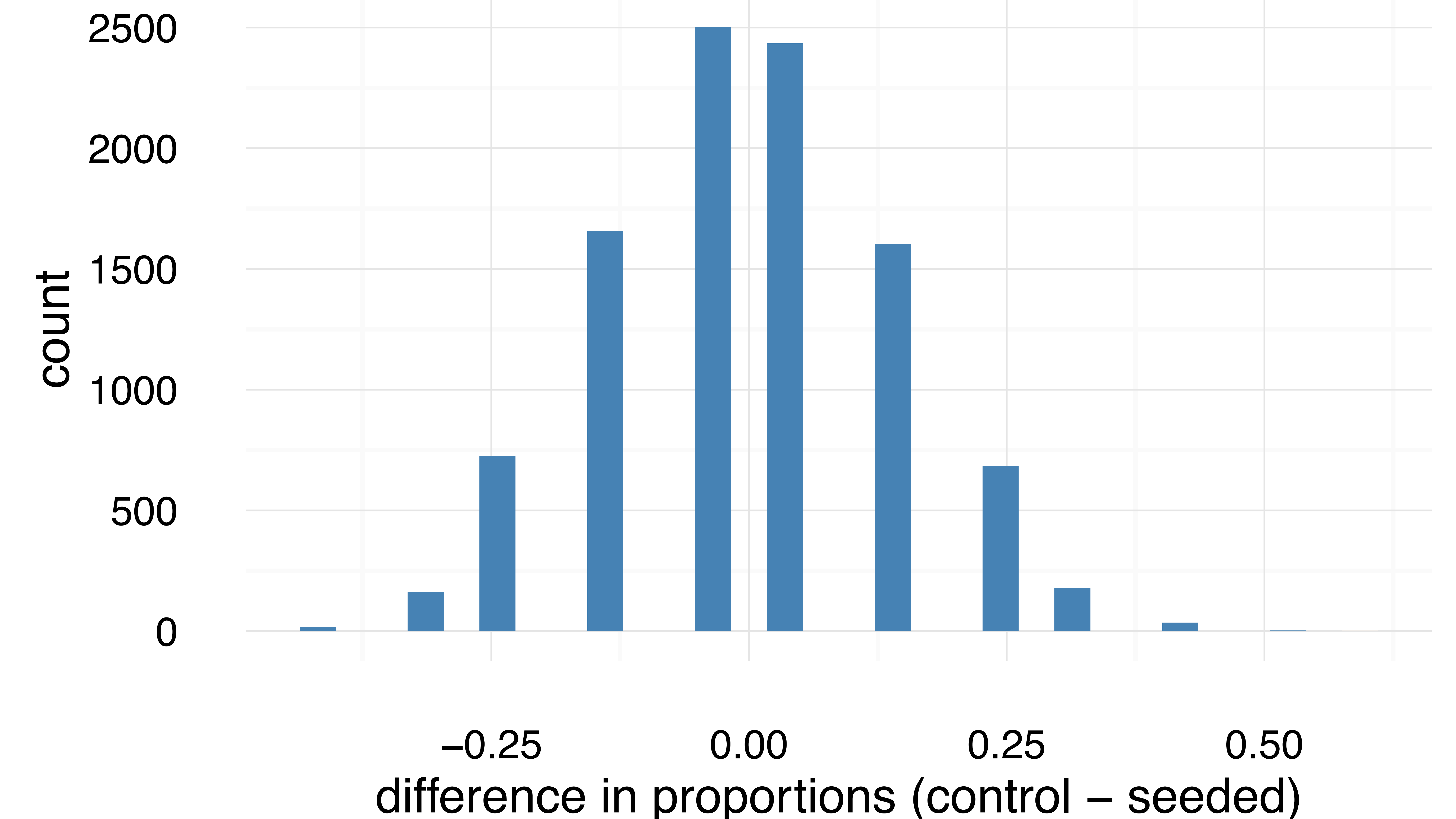
The Normal Distribution

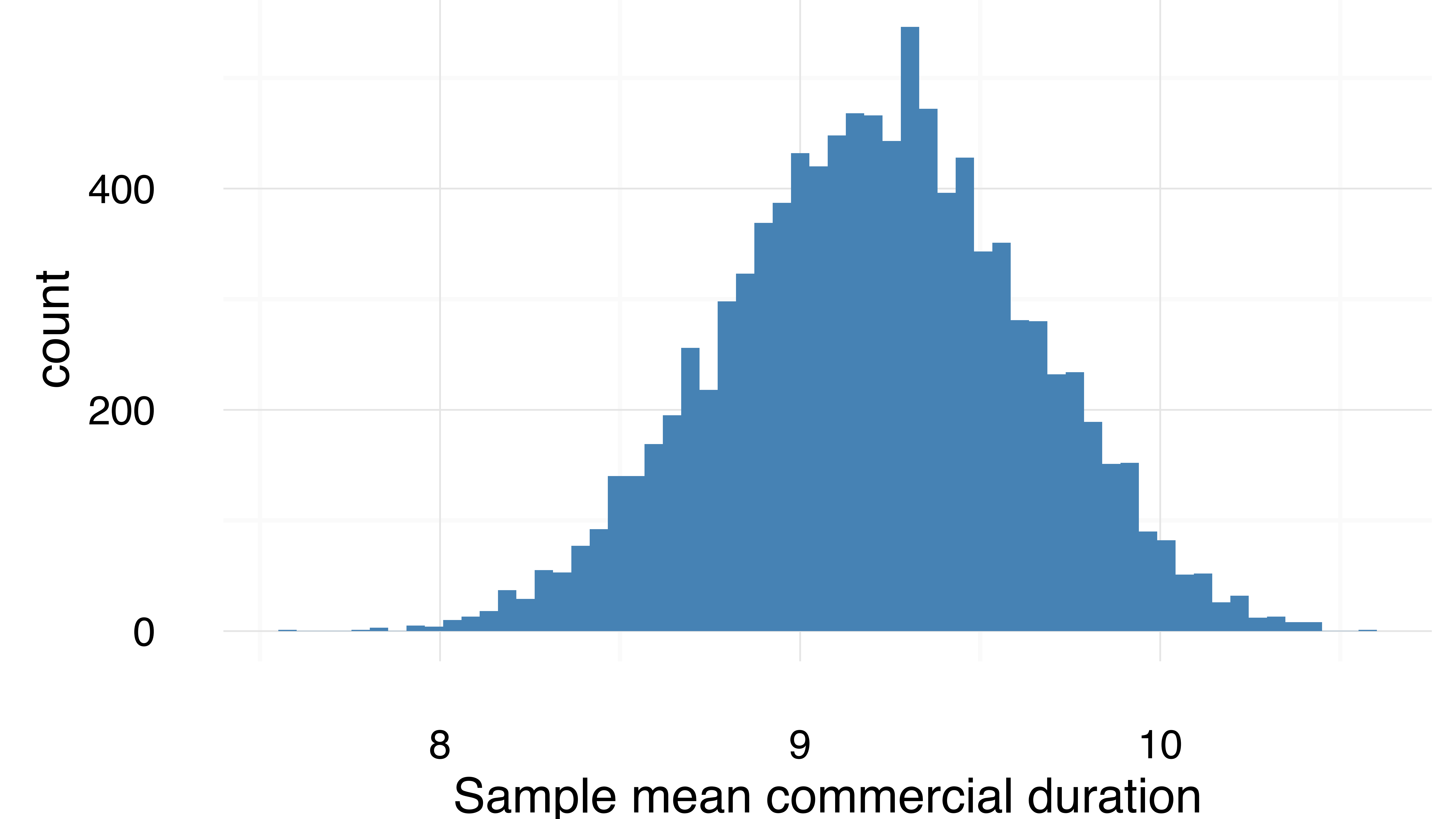
Let’s quickly look back at many of the randomization and bootstrap distributions that we have seen thus far in the course...

Think about what these distributions have in common.

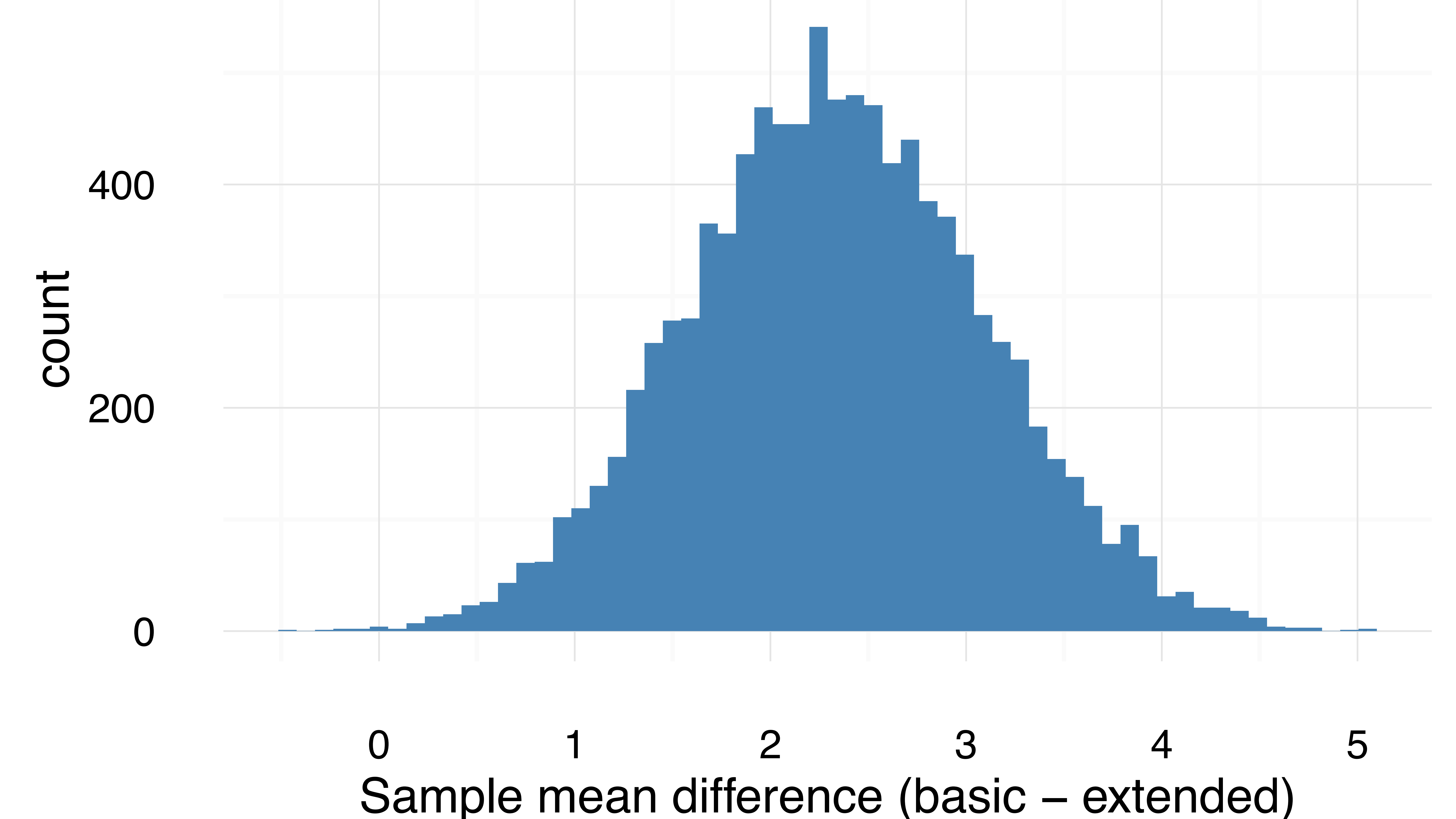


count
1000
750
500
250
0
−5.0
−2.5
0.0
2.5
5.0
mean difference (beer − water)

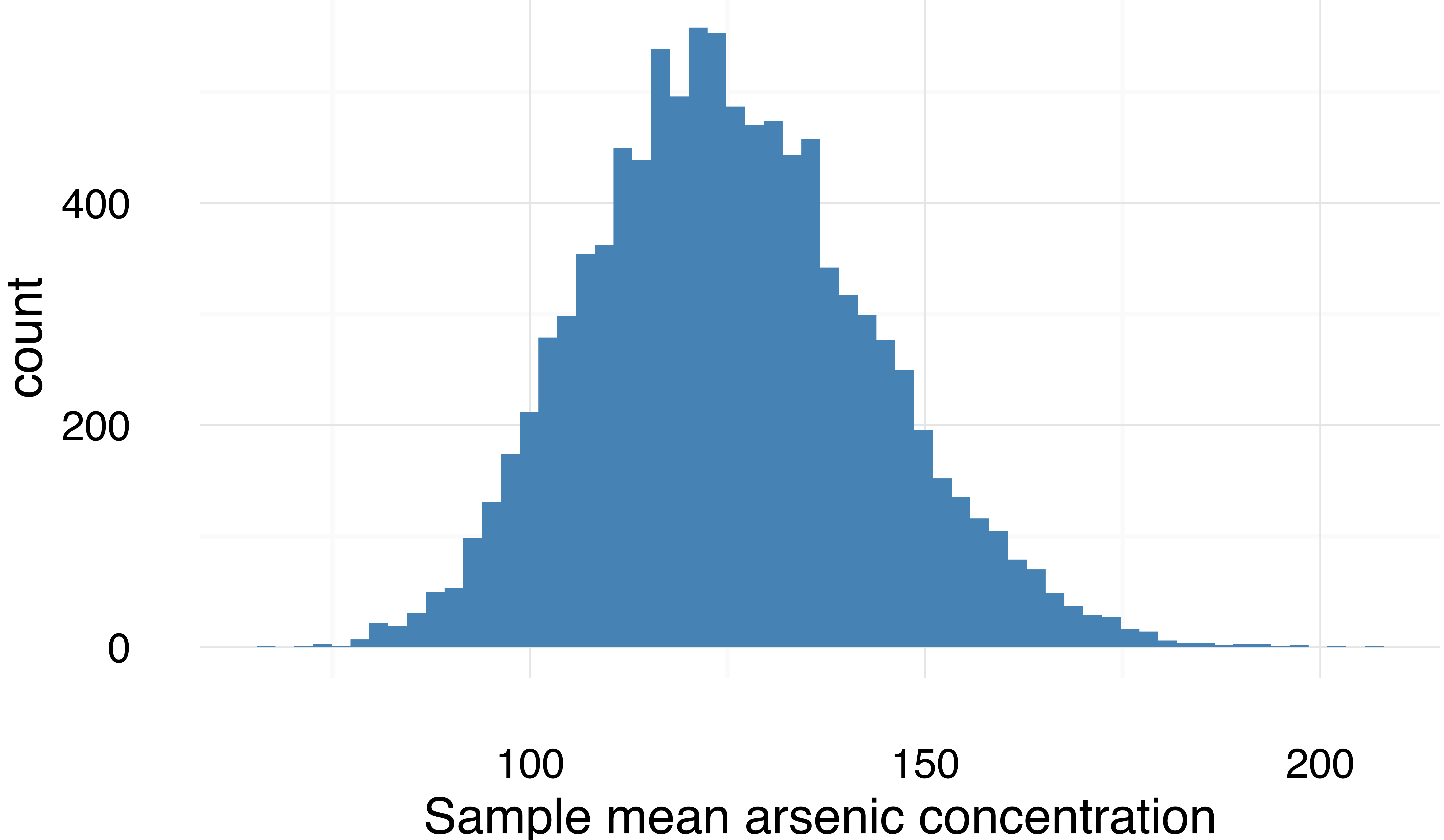

count
2500
2000
1500
1000
500
0
−0.25
0.00
0.25
0.50
difference in proportions (control − seeded)



count
400
200
0
8
9
10
Sample mean commercial duration

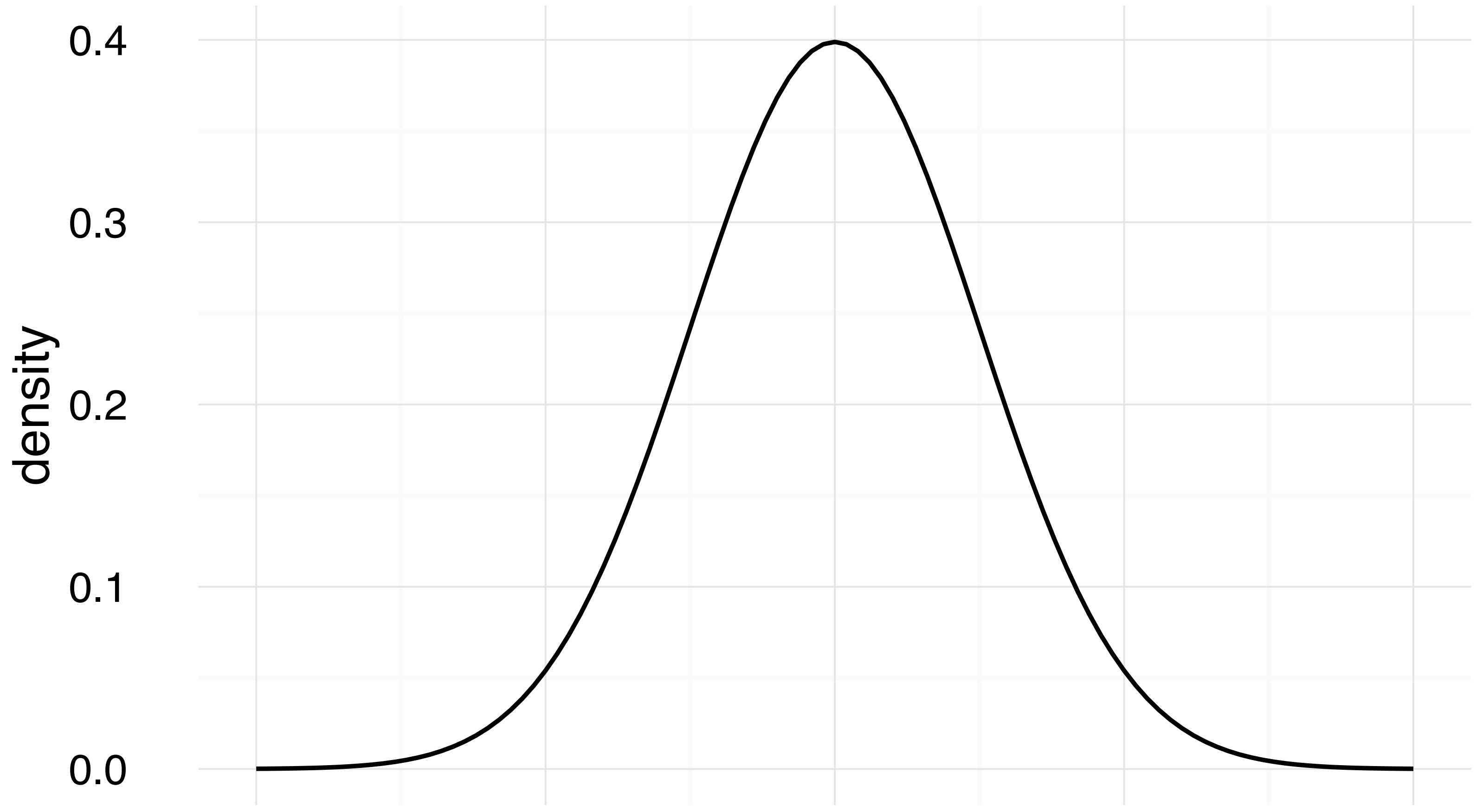


400
200
0
count
0
1
2
3
4
5
Sample mean difference (basic – extended)



count
400
200
0
100
150
200
Sample mean arsenic concentration

The normal distribution, $\mathcal{N}(\mu, \sigma)$



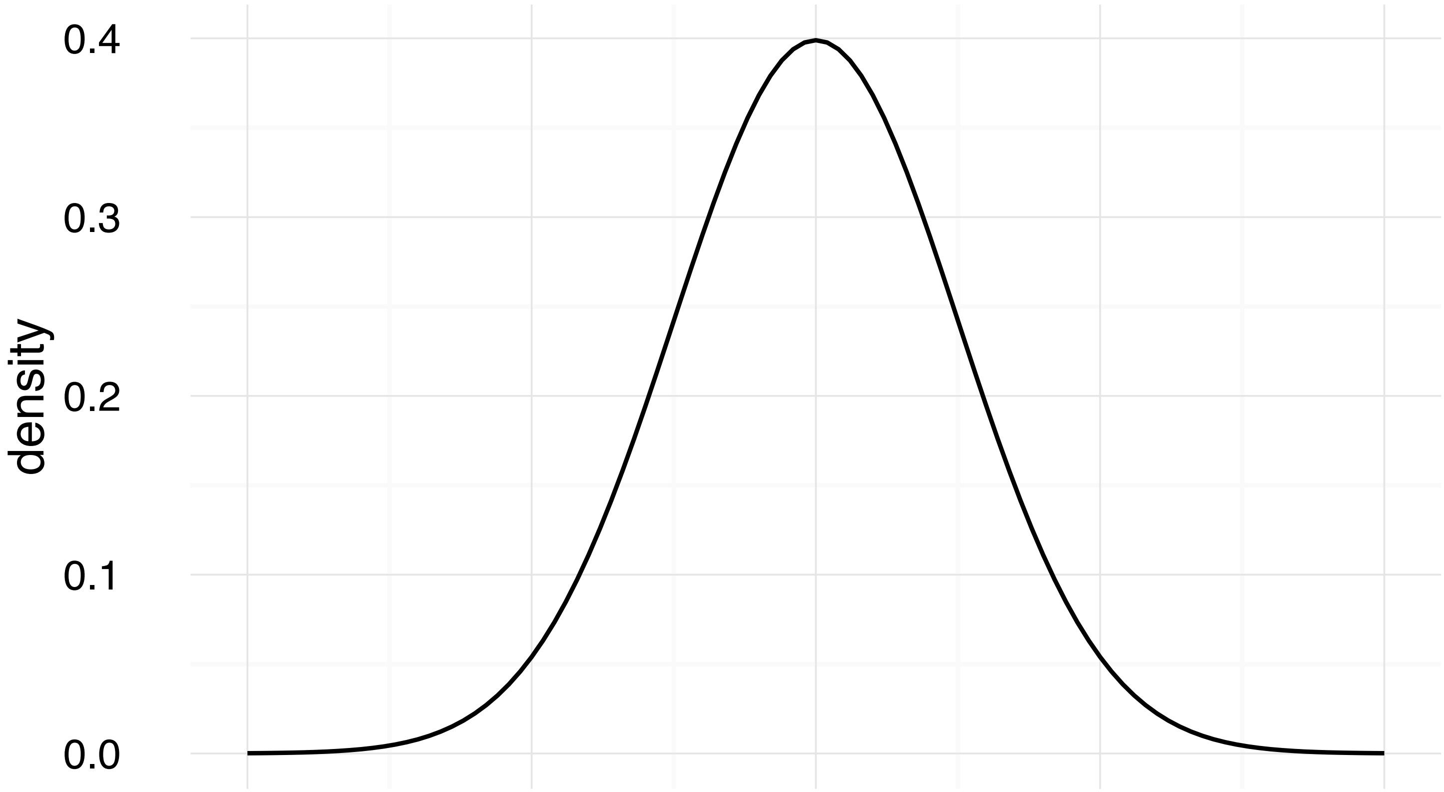
Unimodal and symmetric (bell shaped)

Total area under the curve is 1

Follows very strict guidelines about how the data are distributed around the mean

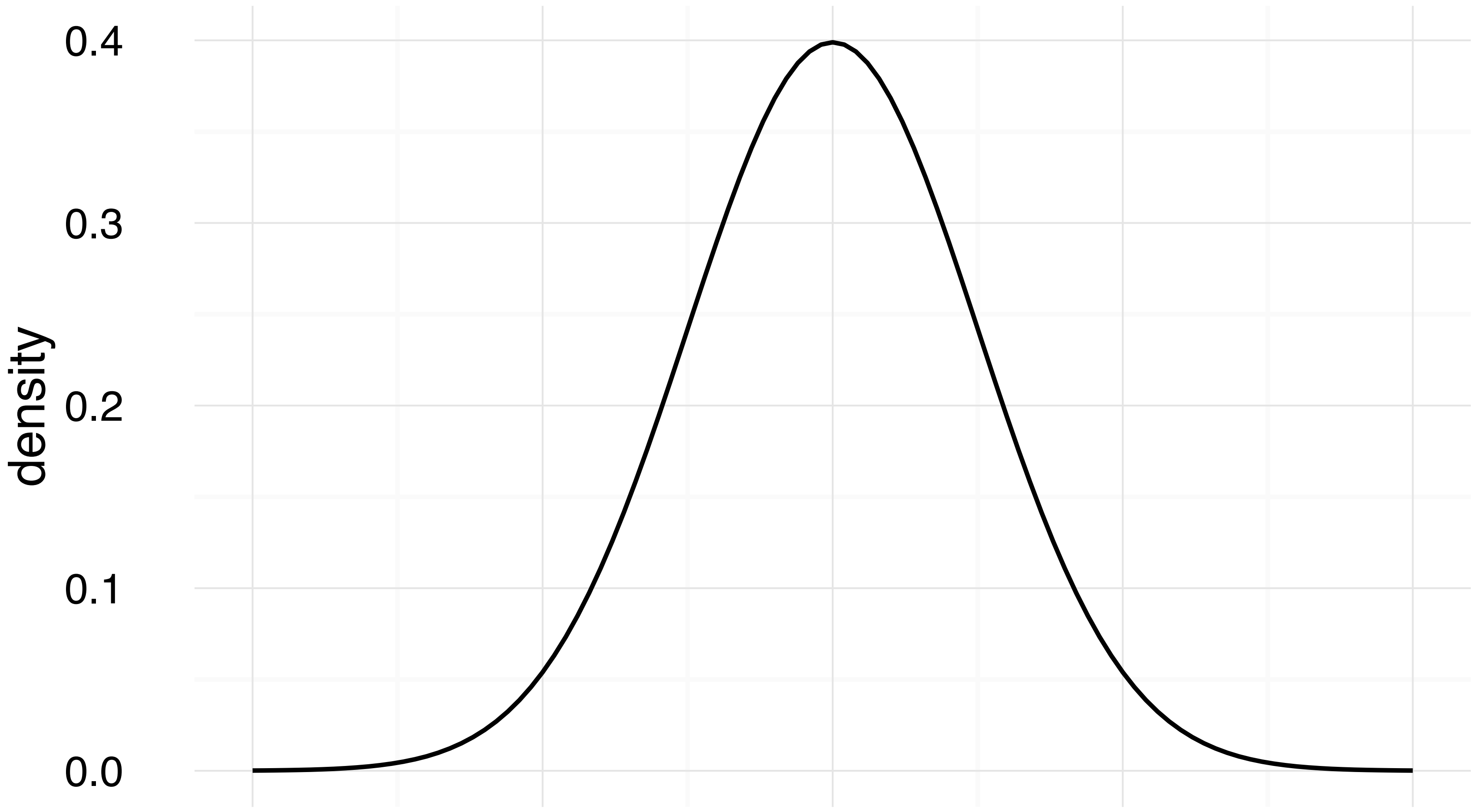
68-95-99.7 Rule

About 68% of the distribution falls within 1 SD of the mean



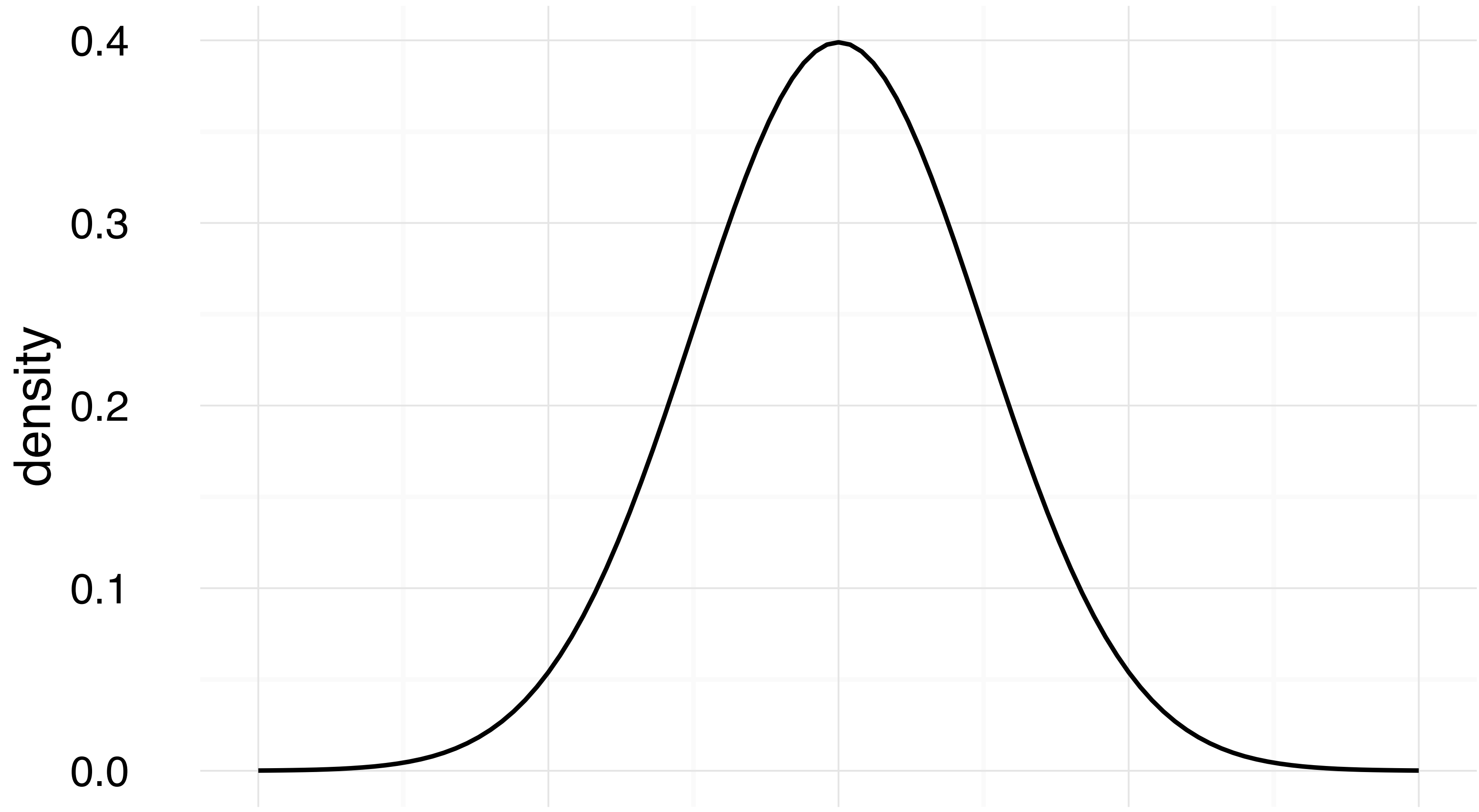
68-95-99.7 Rule

About 95% falls within 2 SD of the mean



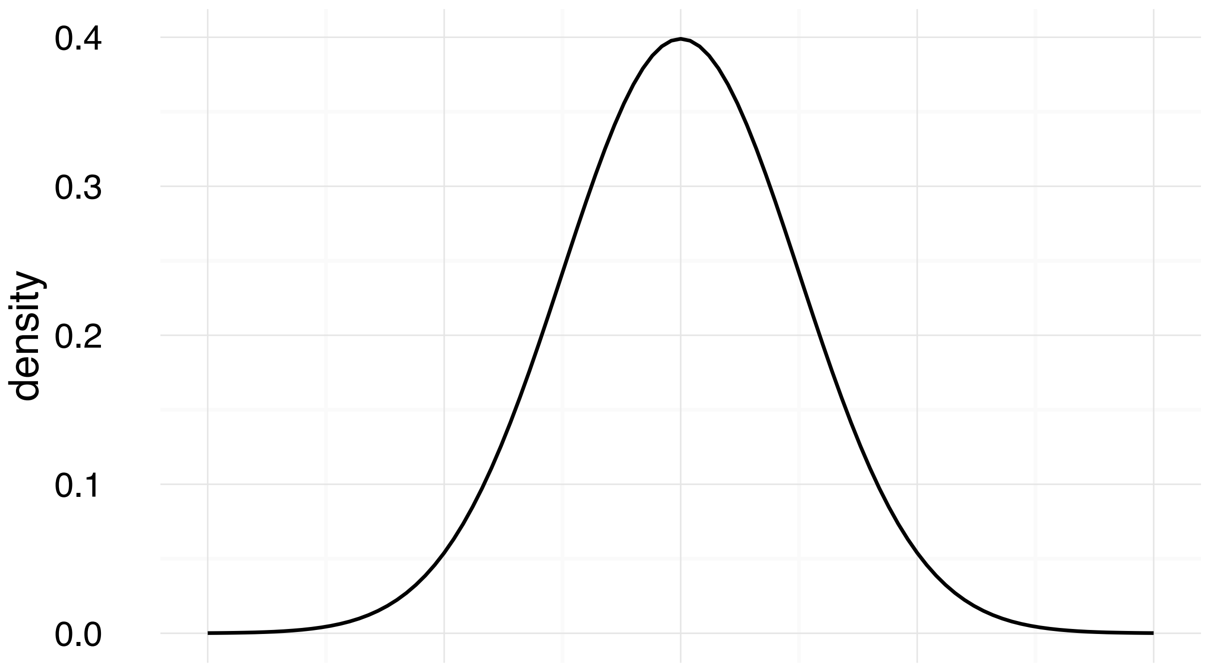
68-95-99.7 Rule

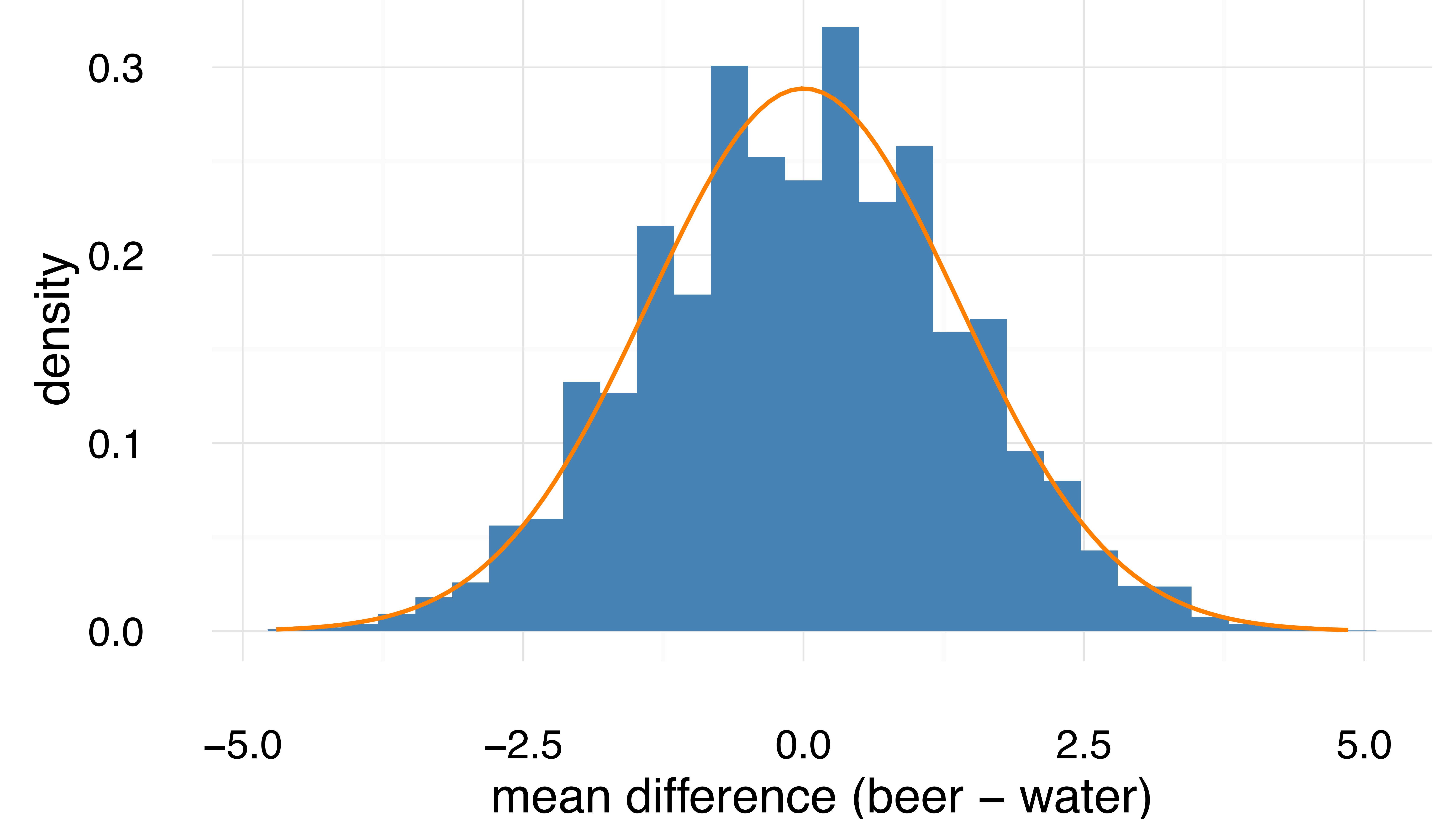
About 99.7% falls within 3 SD of the mean



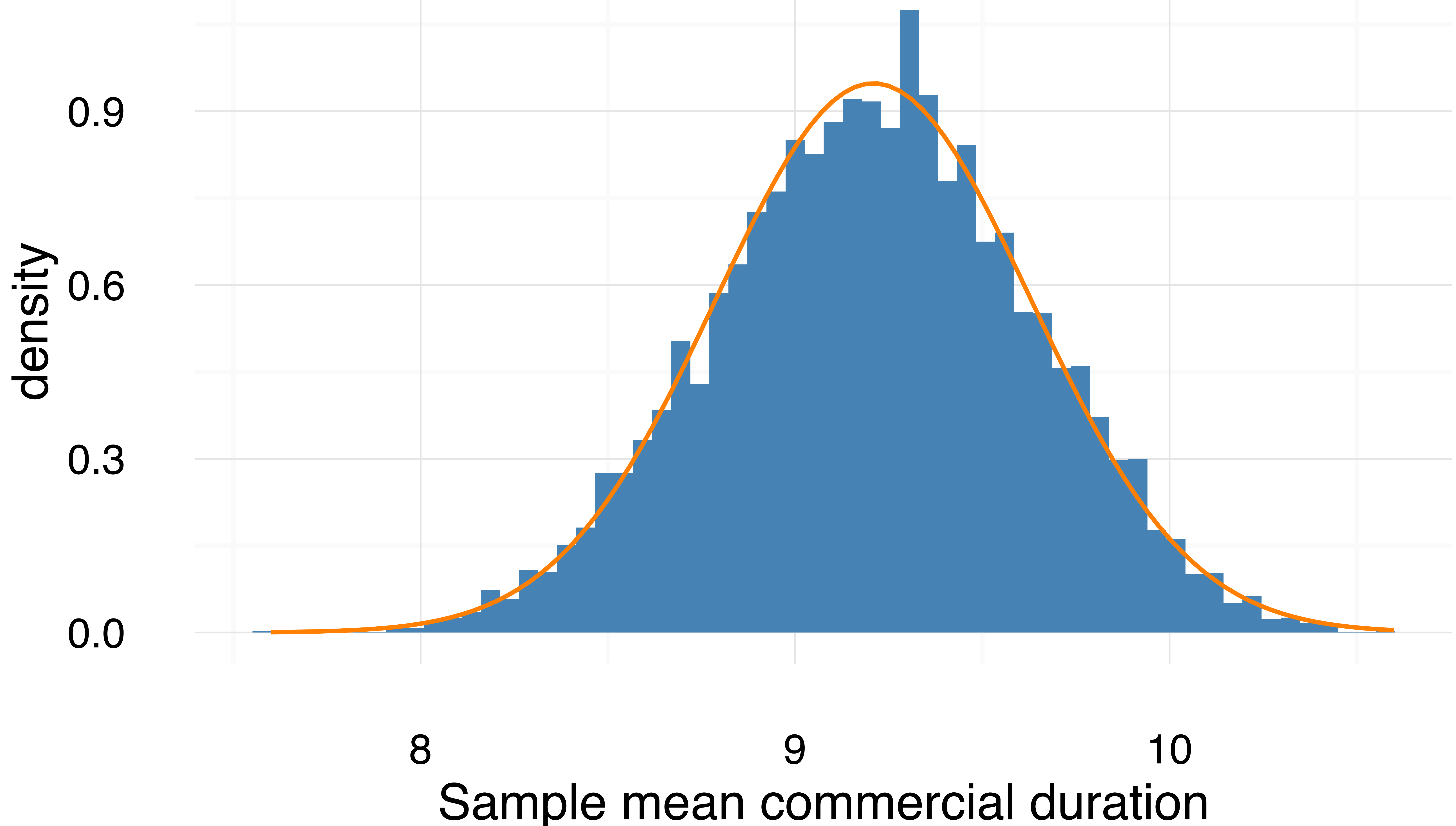
68-95-99.7 Rule

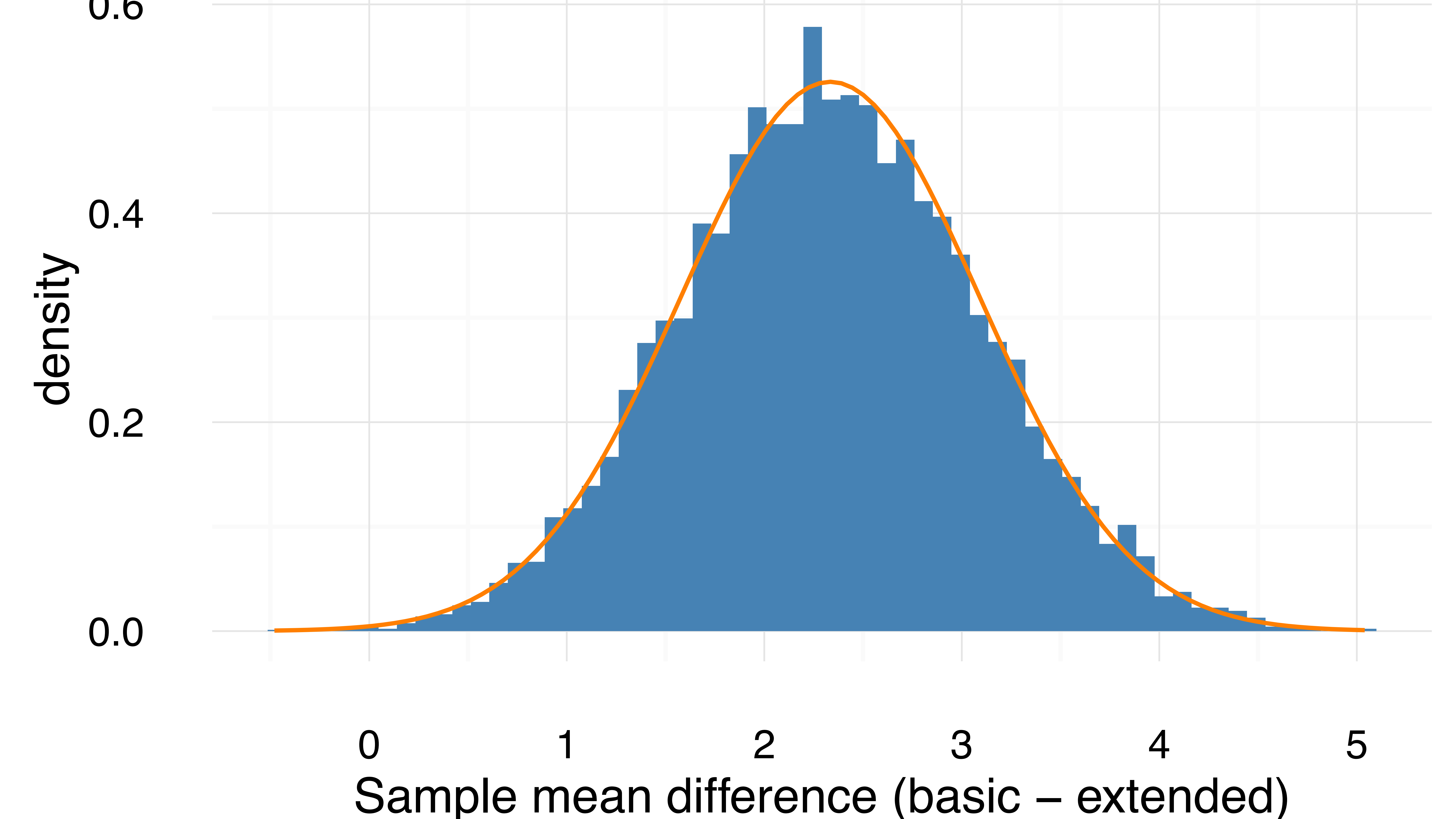
It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but this is very rare if the data are nearly normal



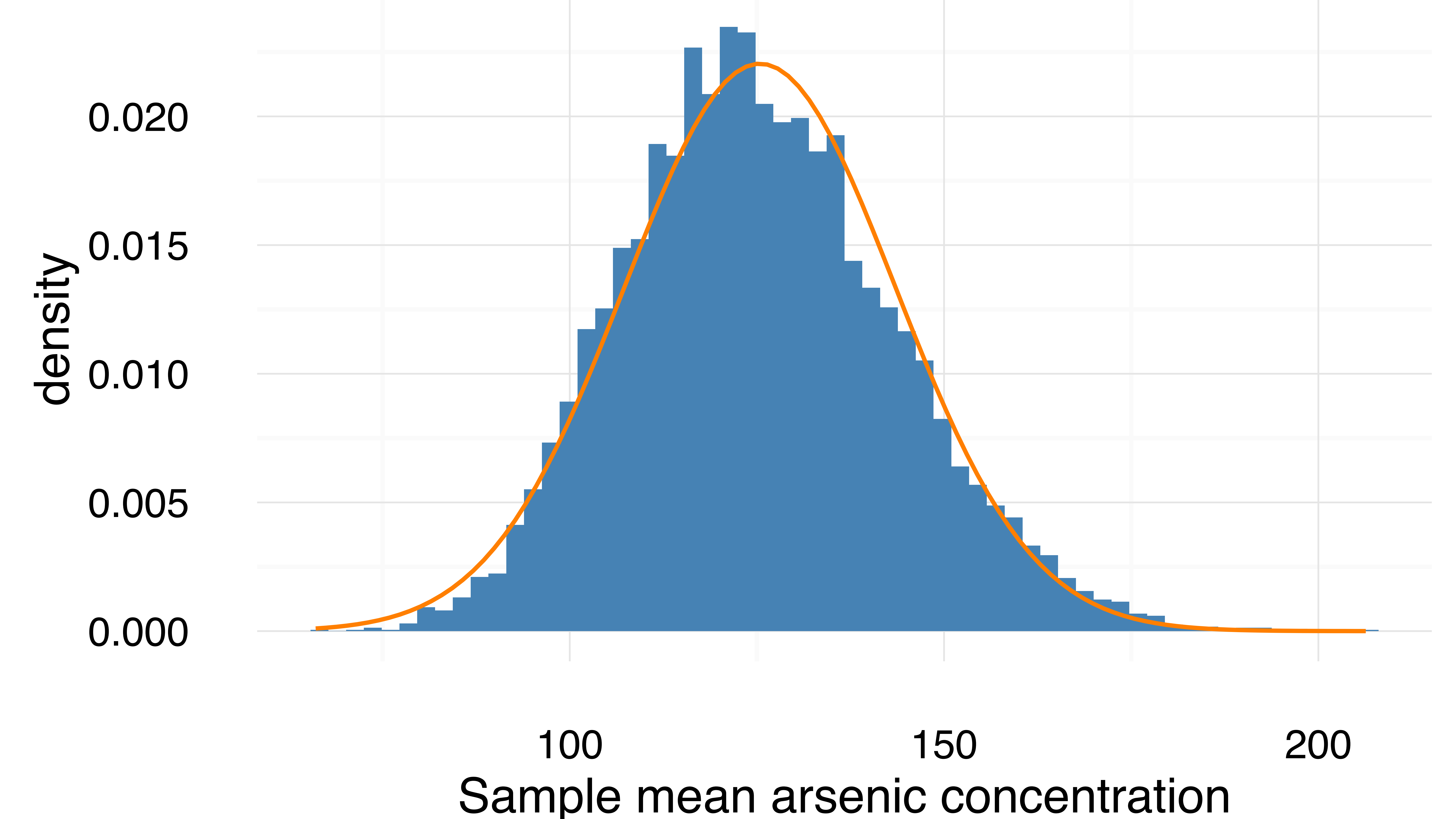


density
0.3
0.2
0.1
0.0
−5.0
−2.5
0.0
2.5
5.0
mean difference (beer − water)


density
0.0
0.3
0.6
0.9
8
9
10
Sample mean commercial duration



0.6
0.4
0.2
0.0
density
0
1
2
3
4
5
Sample mean difference (basic – extended)


0.020
0.015
0.010
0.005
0.000
density
100
150
200
Sample mean arsenic concentration

Working with $\mathcal{N}(\mu, \sigma)$

Knowing the mean and standard deviation of a normal distribution allows us to determine

1. **What proportion of individuals fall in a specified range**
2. **What percentile a given individual falls at if you know their data value**
3. **What data value corresponds to a given percentile**

Example: Scores on the Stanford-Binet IQ test follow a $\mathcal{N}(\mu = 100, \sigma = 16)$

Identify the central 95% of IQ scores.

What proportion of people have an IQ below 52?

What is the probability that a randomly selected person has an IQ of at least 100? Of at least 116?

What proportion of individuals have IQs between 100 and 132?

What is the 97.5th percentile of IQ scores?

Z scores

Formula: $Z = \frac{\text{observation} - \text{mean}}{\text{SD}}$

Interpretation: The number of SDs the observations falls above/below the mean

When the distribution is normal can we use Z scores to calculate proportions/probabilities/percentiles

Z scores

If $X \sim \mathcal{N}(\mu, \sigma)$ **and** $Z = \dfrac{\text{observation} - \text{mean}}{\text{SD}}$**,**

then $Z \sim \mathcal{N}(0, 1)$

Example: SAT Verbal scores for applicants at a college follow a $\mathcal{N}(580, 70)$

What percentage of applicants have a score higher than 650?

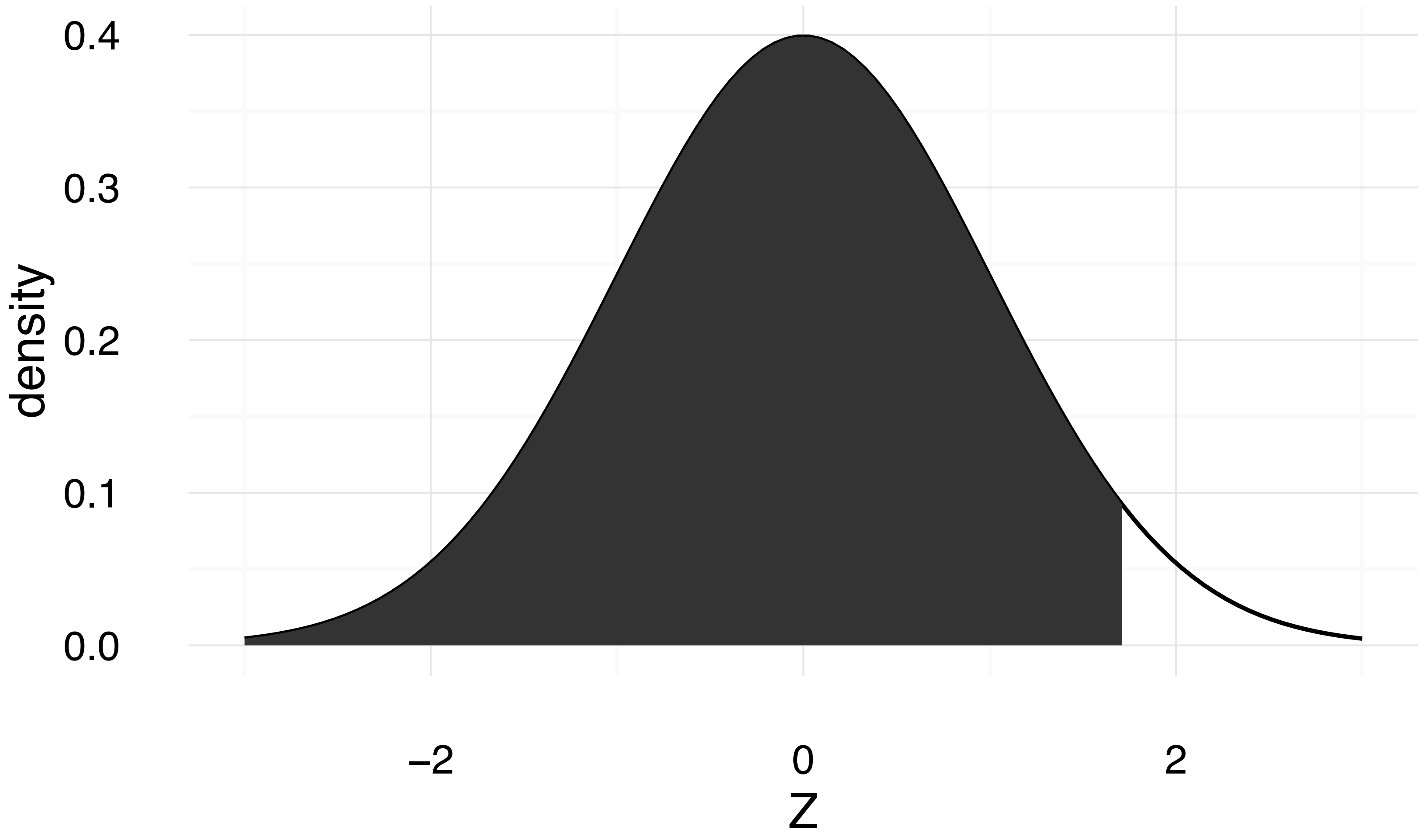
What percentage of applicants have a score lower than 440?

What percentage of applicants have a score higher than 700?

Normal probabilities in R

We can calculate lower tail probabilities in R using

```
pnorm(q, mean, sd)
```



Government data indicate that the average hourly wage for manufacturing workers in the United States is approximately $\mathcal{N}(\$18.61, \$1.35)$.

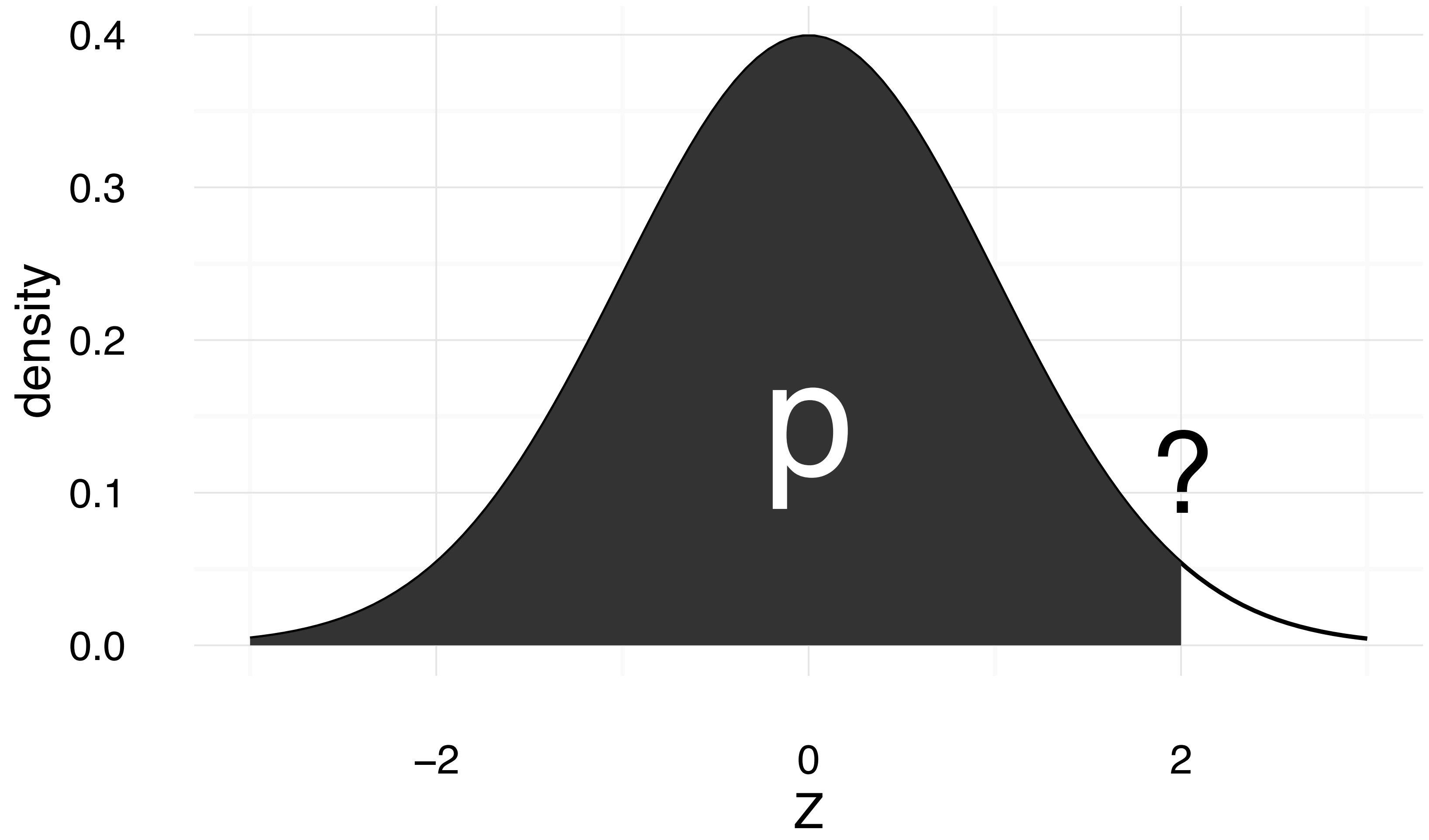
What proportion of manufacturing workers make more than $20/hour?

What proportion of manufacturing workers make between $18 and $20/hour?

Normal percentiles in R

We can find percentiles for normal distributions in R using

```
qnorm(p, mean, sd)
```



Government data indicate that the average hourly wage for manufacturing workers in the United States is approximately $\mathcal{N}(\$18.61, \$1.35)$.

Find the 90$^{\text{th}}$ percentile of hourly wages.

Find the IQR of hourly wages.