Social Statistics

Sampling Distributions

October 17, 2023

Thinking Probabilistically

- So far: Summarizing observed values of variables
  - → Descriptions about the centers and shapes of distributions
- Now: Estimating probability of observing value in the sample
  - → Or, quantifying chance that a value is different from what is observed
- Up next: Inference
  - → What is the probability that a sample statistic is different from a population parameter?

Thinking Probabilistically

- Using what we observe in our sample to estimate what we want to know about the population
  - → The population value exists but we don't observe it. We'll use what we do know to estimate the range of possible values for the population measure.
- From describing precision…
  - → 25% of all 741 commuting zones spend more than the national average on school expenditures

Thinking Probabilistically

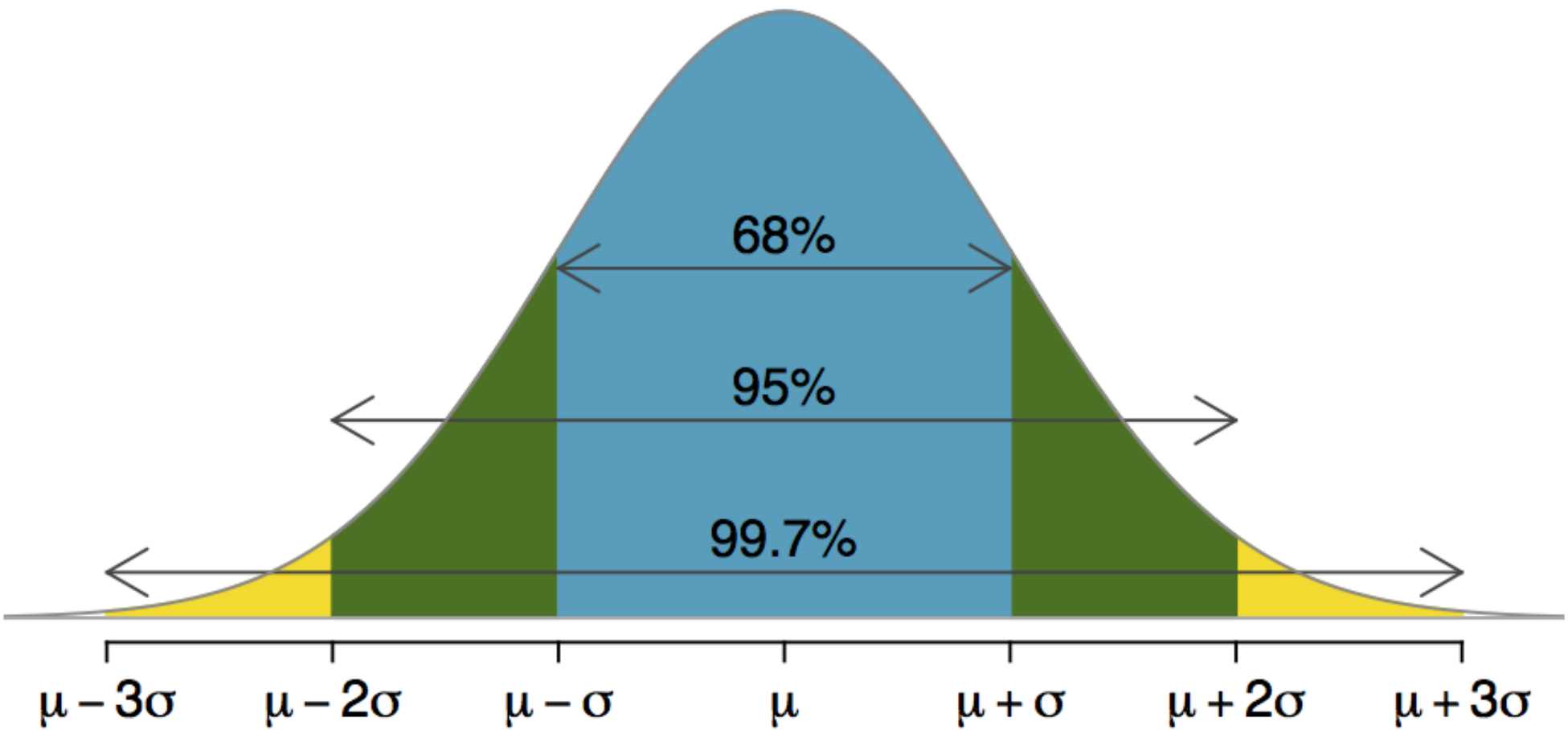
- To quantifying likelihood…

  → In a sample of 300 commuting zones, the average spent on school expenditures is $1000 per student. How likely is it that the average across all 741 commuting zones is higher or lower than that?

Thinking Probabilistically

- For now, assume variable is normally distributed, and apply what we know about means and standard deviations in normal distributions…

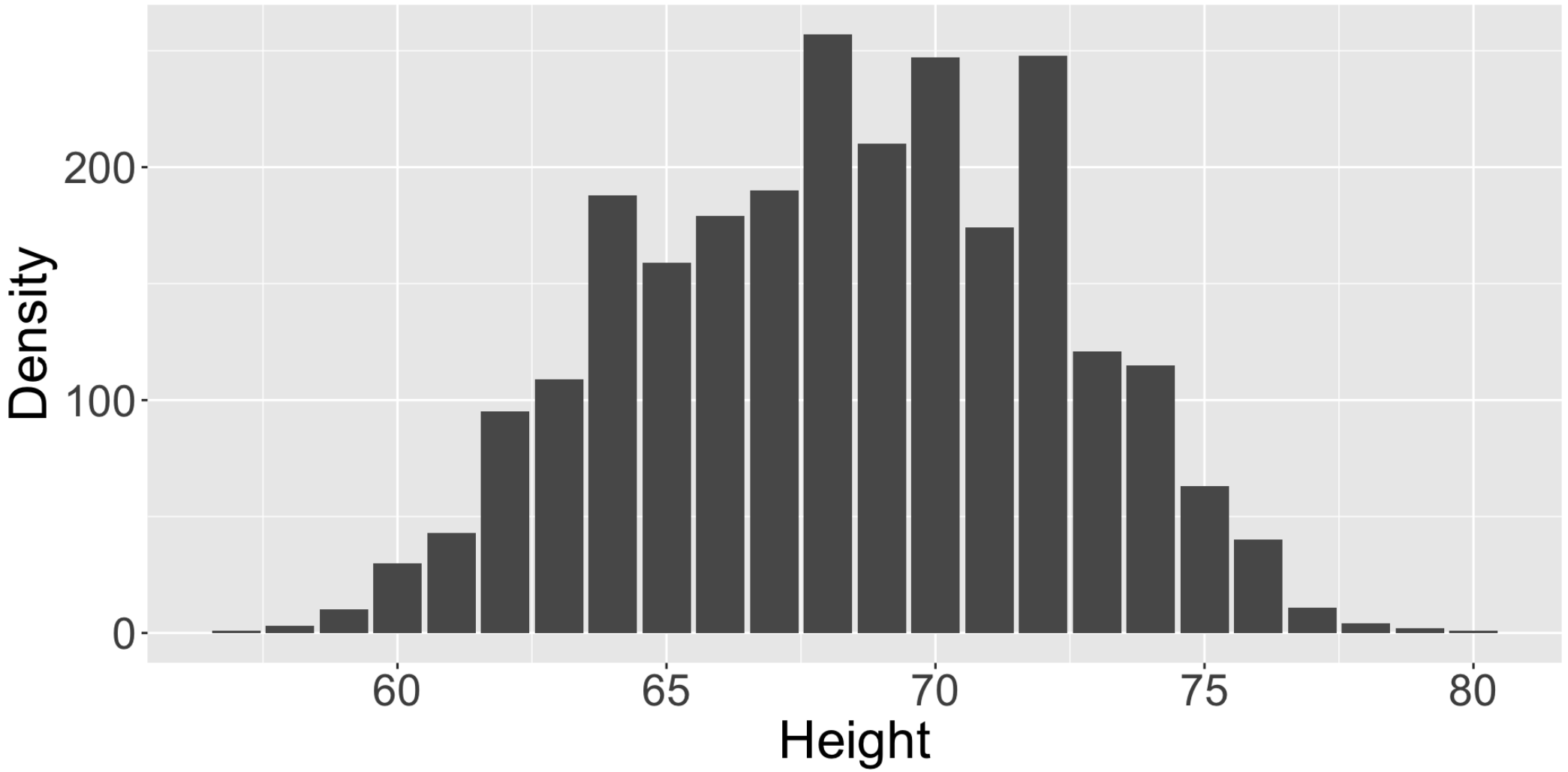
Normal Distribution



Back to R

- Strategy with probability distributions is to find a value's distance from the mean in standard deviations. This measure is called the `z-score`.
- $$z = \frac{x - \mu}{\sigma}$$
- Positive z-score is a value's distance above the mean in standard deviations
- Negative z-score is a value's distance below the mean in standard deviations

Example With OK Cupid Dataset



Calculating Z Scores

- For each observation, we need to find the difference from the mean and then divide that difference by the standard deviation.

```
cupid <- cupid |>
  mutate(height_z = (height - mean(height)) /
           sd(height))
```

Calculating Z Scores

- Z-scores should be normally distributed with a mean of 0 and a standard deviation of 1. Were we successful?

```
1 round(mean(cupid$height_z),3)
```

```
[1] 0
```

```
1 sd(cupid$height_z)
```

```
[1] 1
```

Interpreting Z-Scores

- What is the z-score for a height of 71 inches?

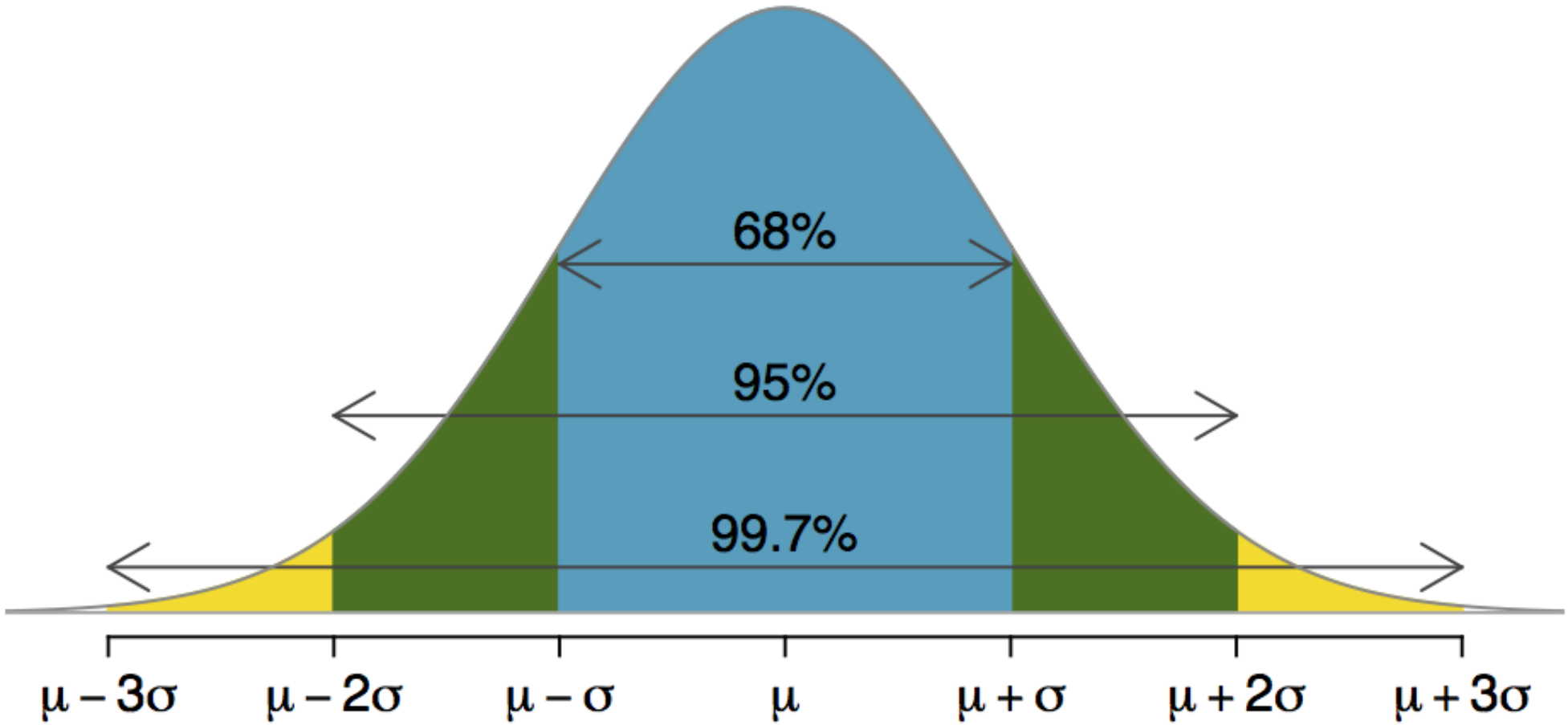
```
1 mean(cupid$height_z[cupid$height==71])
```

```
[1] 0.6818372
```

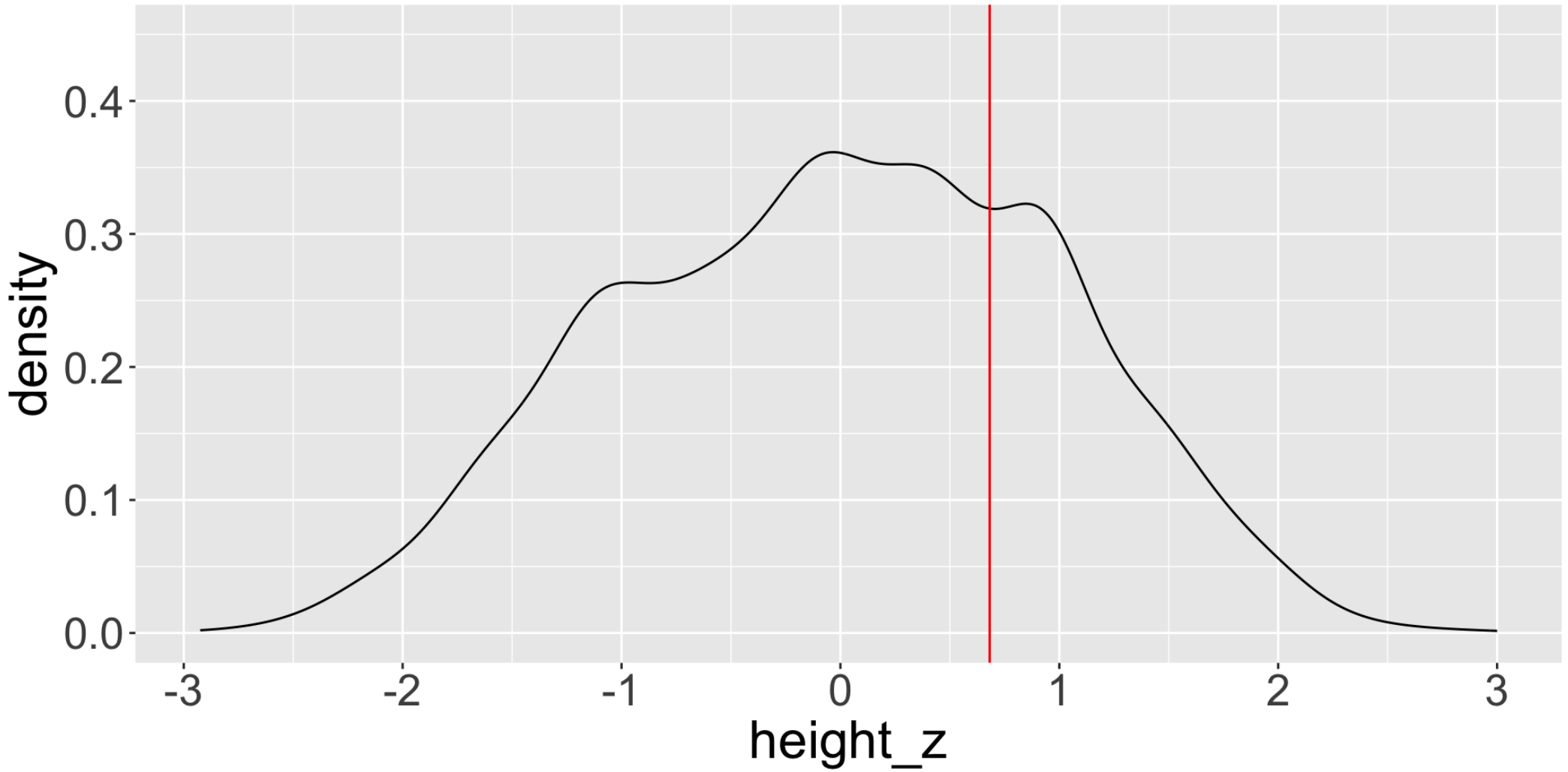
- In words, someone who is 71 inches tall is .68 standard deviations taller than the mean.

Interpreting Z-Scores

- When we plot standardized values that are approximately normal, we now know a lot about how many observations fall along different points of the distribution…


68%
95%
99.7%
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ

Interpreting Z-Scores



Interpreting Z-Scores

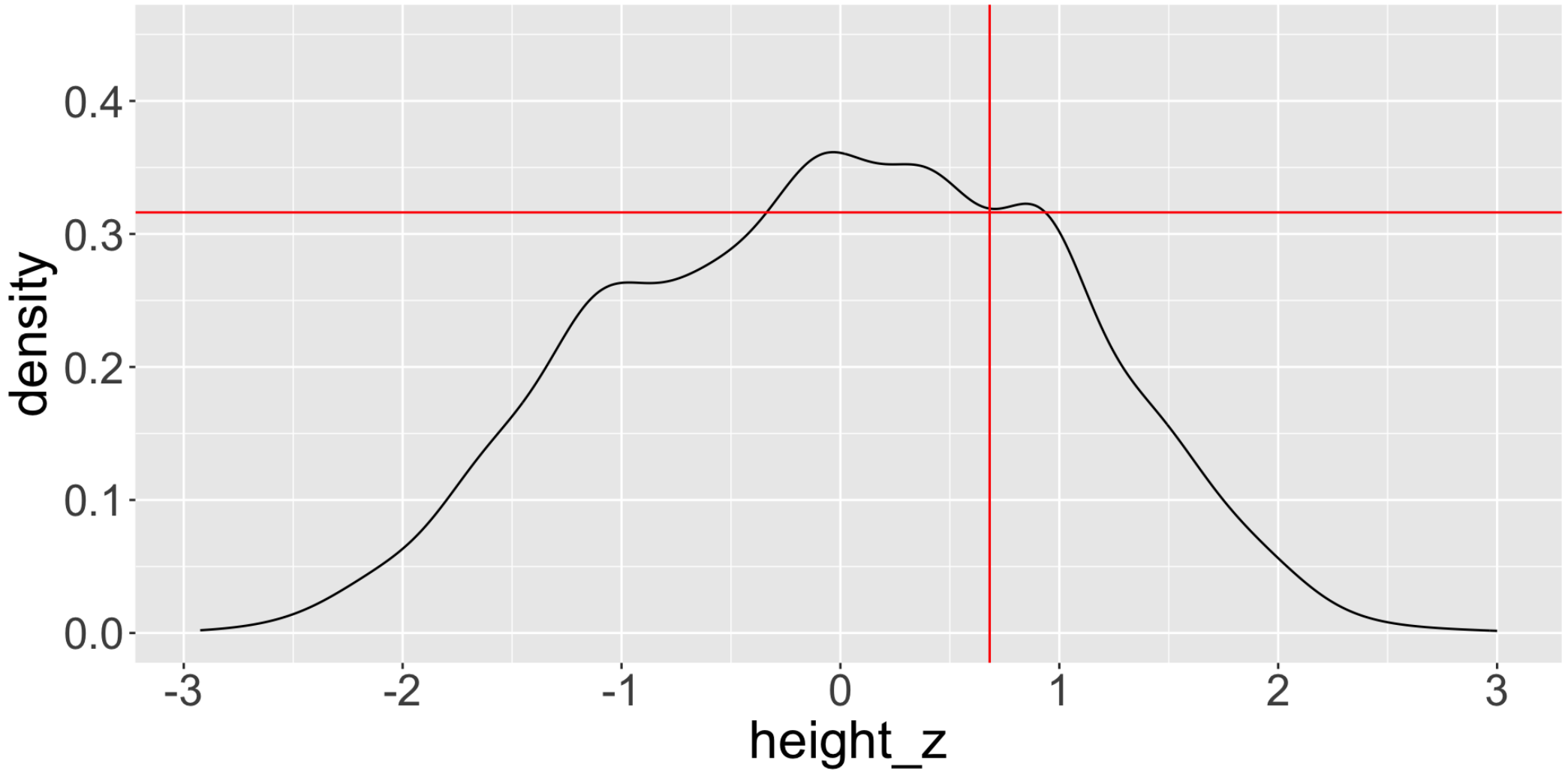
- To find the height on the density plot for an observation with a height of 71 inches, use `dnorm()` with the z-score for that observation's value:

```
1 dnorm(0.6818372) # d for density
```

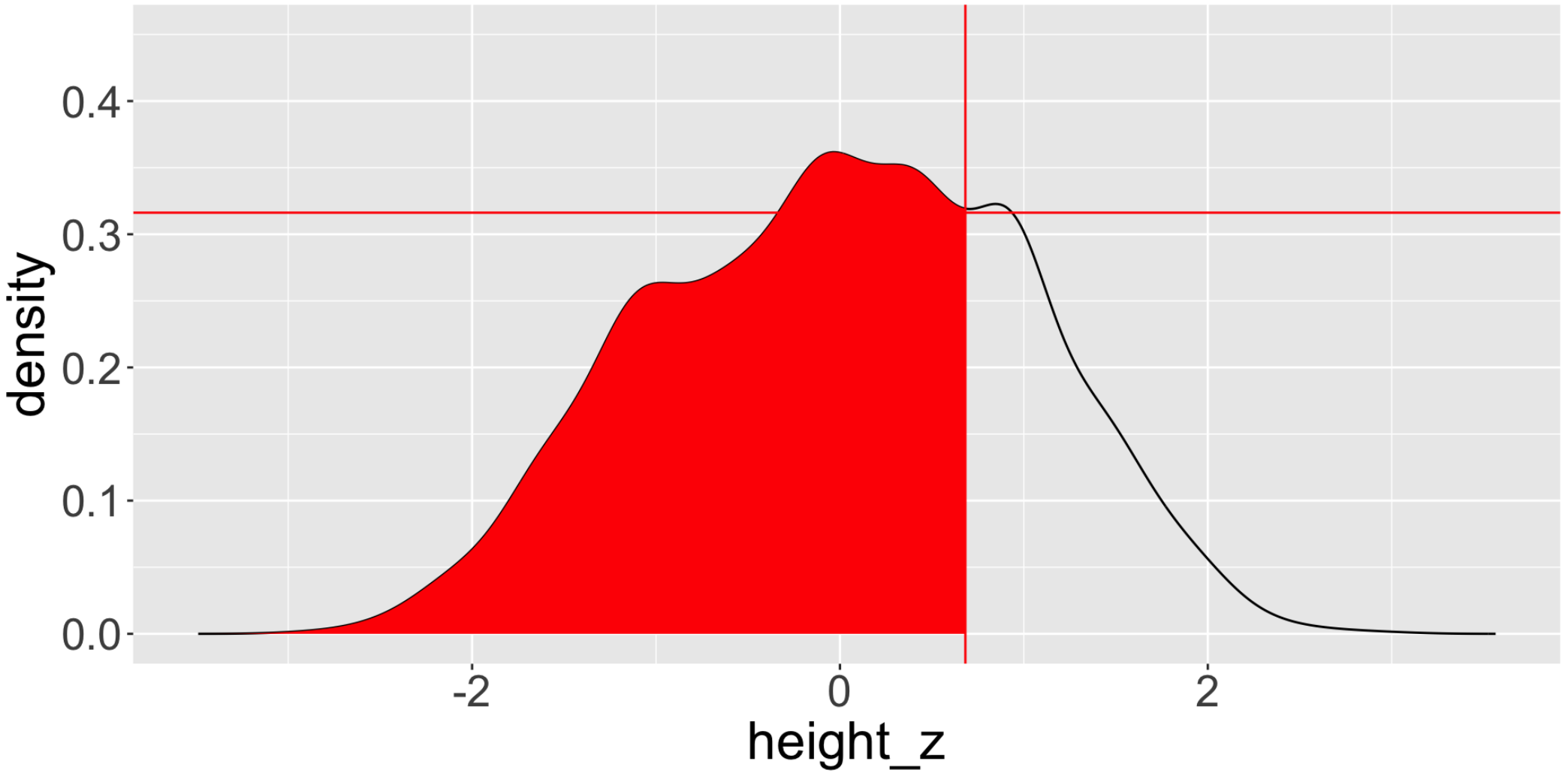
```
[1] 0.3161971
```

Looks pretty close in our example…

Interpreting Z-Scores



Cumulative Probabilities and Densities



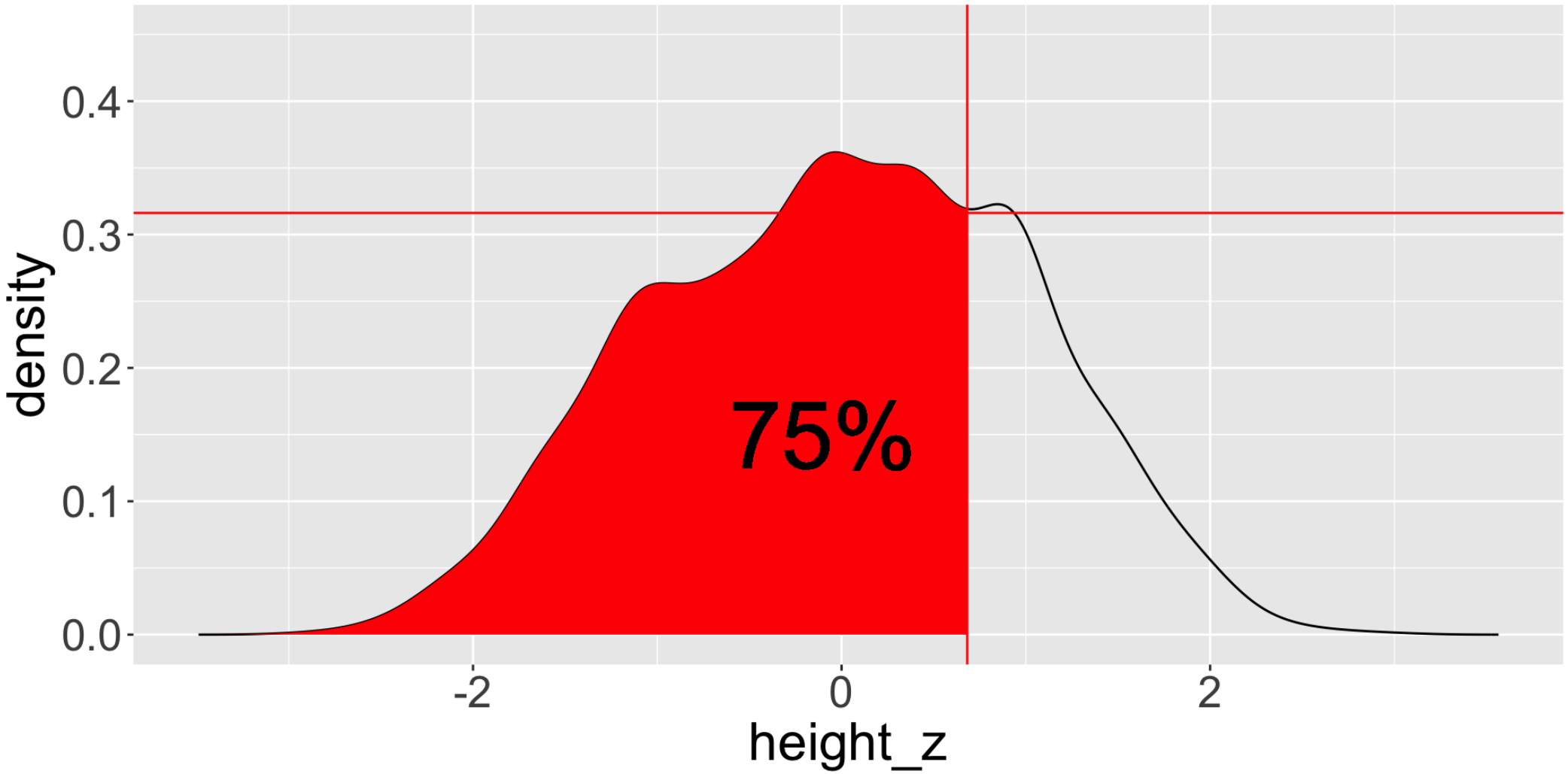
Cumulative Probabilities and Densities

- To find the area under the curve, we need to know the `cumulative density` not the density. The cumulative density is the same as the percentile.
- If you have the z-value and want the percentile associated with it, use `pnorm()` which gives you the proportion of the distribution ***to the left*** of your z-value.
- For Height of 71 Inches:

```
1 pnorm(.6818372) # p for percentile
```

```
[1] 0.7523291
```

Cumulative Probabilities and Densities



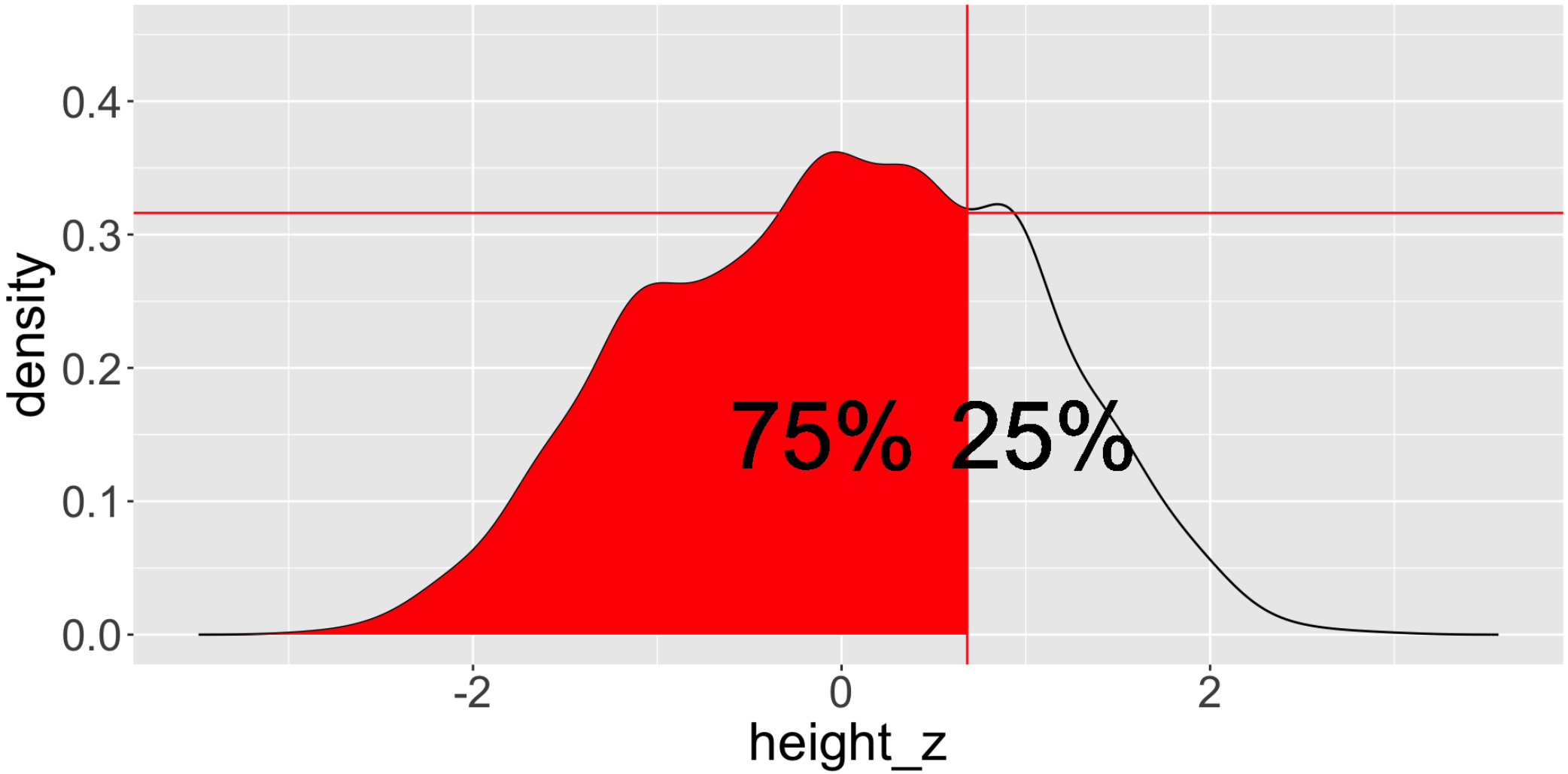
Use What You Know

- What is the probability of another respondent being taller than 71 inches?

```
1 1 - pnorm(.682)
```

```
[1] 0.2476195
```

Use What You Know



Exercise

- What is the z-score for 64 inches?
- What is the probability that someone in our sample is shorter and taller than 64 inches?

Exercise

- What is the z-score for 64 inches?

```
1 mean(cupid$height_z[cupid$height==64])
```

```
[1] -1.12142
```

- What is probability that someone in our sample is shorter than 64 inches?

```
1 pnorm(-1.12)
```

```
[1] 0.1313569
```

- What is the probability that someone in our sample is taller than 64 inches?

```
1 1 - pnorm(-1.12)
```

```
[1] 0.8686431
```

What's the point?

- The key bridge to inference is thinking of the x-axis not as observed values of height in our sample but as possible values of the true mean of height in the population.
- We want to know how close the mean in our observed sample is to the true (unobserved) population mean. Knowing where it falls in the distribution of all the possible sample means is how we infer how similar the sample mean and the population are.
- Remember our new language: what is the probability of another randomly drawn sample mean being more extreme than our sample mean *simply by chance*.

Measuring Sampling Variation

- So we need lots of samples. Bootstrapping is one approach. It gives us repeated samples of our actual sample so we have more possible values of our statistic.
- We need more samples because as sample size increases, distribution of z-values of repeated sample means is normally distributed around standardized population mean of 0 with a standard deviation of 1.
- The key insight: since the means from repeated samples are normally distributed, now it is not a problem if the distribution in one sample is not normally distributed. If our sample size is big enough, we can think of our observed values as possible estimates of the population mean!

Measuring Sampling Variation

- We won't use bootstrapping to pull repeated samples. But we'll use the *standard deviation* of our sample to calculate the *standard error* of the sampling distribution.

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{sd}{\sqrt{samplesize}}$$

An Example

- In the `cupid` data set, the age variable is *not* normally distributed, but we still want to use probability to estimate the population mean from our sample mean.
- Let's find the standard error of the age variable. We'll save this as an object, not as a new variable (since it is the same for the entire sample):

```
age_se <- sd(cupid$age) / sqrt(length(cupid$age))
```

```
age_se
```

```
[1] 0.1847327
```

From SE To Confidence Intervals

- We use the standard error, the sample mean and what we know about the distribution of z-scores to build a range of possible values for the population mean
- The most common range is a *95% confidence interval*
  - → That is the range in which the true population mean will be found in 95% of sampling distributions
  - → We are ***not*** saying we are 95% sure that our sample mean is the population mean!

From SE To Confidence Intervals

- To build that 95% interval, we need to define a range that captures 95% of the normal distribution. In other words, outside this range there will be only a 2.5% chance that another sample will have a mean above our mean and a 2.5% chance that another sample will have a mean below our mean.
- That should sound like z-scores!

From SE To Confidence Intervals

- We need the z-scores that are associated with .025 and .975. To find them, we use `qnorm()`.

```
1 qnorm(.025)
```

```
[1] -1.959964
```

```
1 qnorm(.975)
```

```
[1] 1.959964
```

Z-Scores For Confidence Intervals

- 95% (most common) = 1.96
- 99% = 2.58
- 90% (less common) = 1.65

From SE To Confidence Intervals

- The z-score for the confidence level we want multiplied by our standard error is the *margin of error*

```
# Margin of Error:

1.96*age_se
```

```
[1] 0.3620762
```

From SE To Confidence Intervals

- Our sample mean plus and minus the margin of error is our confidence interval
- Find both the *lower limit* and the *upper limit*

```
# Lower Limit of Confidence Interval
age_ll <- mean(cupid$age) - 1.96*age_se

# Upper Limit of Confidence Interval
age_ul <- mean(cupid$age) + 1.96*age_se
```

From SE To Confidence Intervals

- Often helpful to save the lower limit, mean, and upper limit as a vector

```
age_ci <- c(age_ll, mean(cupid$age), age_ul)
```

```
age_ci
```

```
[1] 32.02192 32.38400 32.74608
```

From SE To Confidence Intervals

- Interpretation?
  - → 95% of the repeated samples we might imagine pulling would be expected to have means within this range, giving us 95% confidence that the true population mean falls within this range

Exercise

- What is the 99% confidence interval for height?
- Find the standard error
- Find the margin of error
- Construct the confidence interval

Exercise

```r
# Find the standard error:

height_se <- sd(cupid$height) / sqrt(length(cupid$height))

height_se
```

```
[1] 0.0776373
```

```r
# For the margin of error, we need .005 on each side of our mean:

qnorm(.995)
```

```
[1] 2.575829
```

```r
# Margin of error =

2.58 * height_se
```

Exercise

```r
# Construct the 99% Confidence Interval

height_ll <- mean(cupid$height) - 2.58*height_se
height_ul <- mean(cupid$height) + 2.58*height_se

height_ci <- c(height_ll, mean(cupid$height), height_ul)
```

```r
# Display

height_ci
```

```
[1] 68.1529 68.3532 68.5535
```

- Interpretation?