Statistical Thinking for the 21st Century

7.5 Confidence intervals

Most people are familiar with the idea of a “margin of error” for political polls. These polls usually try to provide an answer that is accurate within +/- 3 percent.

For example, when a candidate is estimated to win an election by 9 percentage points with a margin of error of 3, the percentage by which they will win is estimated to fall within 6 to 12 percentage points (9 minus 3 = 6 and 9 plus 6 = 12).

In statistics, we refer to this range of values as the confidence interval, which provides a measure of our degree of uncertainty about how close our estimate is to the population parameter.

We already know that with sufficient sample size, the sampling distribution of the mean is normally distributed, and the standard error describes the standard deviation of this sampling distribution.

Using this knowledge, we can ask: What is the range of values within which we would expect to capture 95% of all estimates of the mean? To answer this, we can use the normal distribution, for which we know the values between which we expect 95% of all sample means to fall.

We choose these points because we want to find the 95% of values in the center of the distribution, so we need to cut off 2.5% on each end in order to end up with 95% in the middle. Figure 7.3 shows that this occurs for $Z \pm 1.960$.

Using these cutoffs, we can create a confidence interval for the estimate of the mean:

X-bar is the sample mean

$$CI_{95\%} = \bar{X} \pm 1.960 * SEM$$

Let's compute the confidence interval for the NHANES height data.

Sample mean	SEM	Lower bound of CI	Upper bound of CI
169	1.400	166	172

Confidence intervals are notoriously confusing, primarily because they don't mean what we would hope they mean. It seems natural to think that the 95% confidence interval tells us that there is a 95% chance that the population mean falls within the interval.

However, as we will see throughout the course, concepts in statistics often don't mean what we think they should mean. In the case of confidence intervals, we can't interpret them in this way because the population parameter has a fixed value – it either is or isn't in the interval.

The proper interpretation of a 95% confidence interval is that 95 out of 100 samples (drawn from the same population) will contain the true population mean, but 5 of the 100 samples will not.

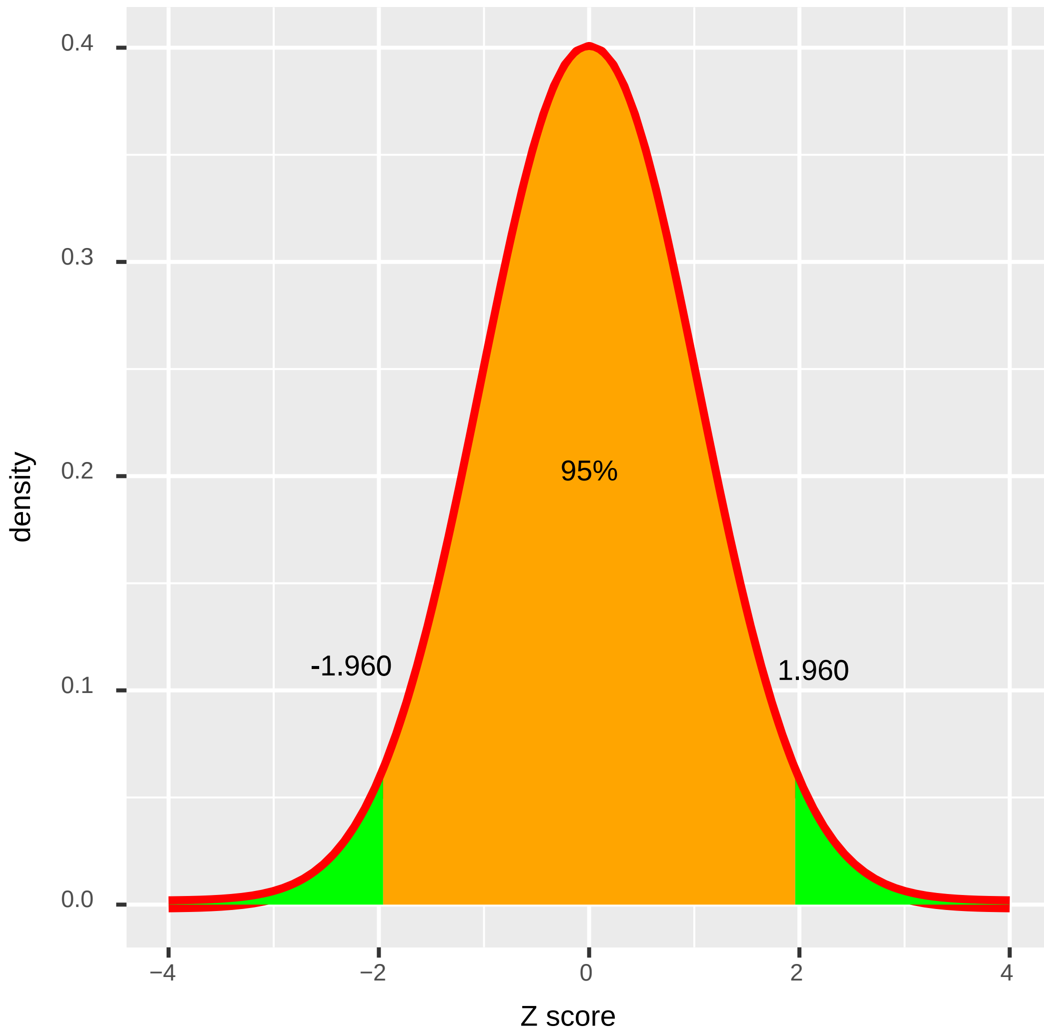

Figure 7.3: Normal distribution, with the orange section in the center denoting the range in which we expect 95 percent of all values to fall. The green sections show the portions of the distribution that are more extreme, which we would expect to occur less than 5 percent of the time.