Statistical Thinking for the 21st Century

7.4 The Central Limit Theorem

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, *even if the data within each sample are not normally distributed*.

We can see this in real data.

Let's work with the variable AlcoholYear in the NHANES distribution, which is highly skewed, as shown in the **left panel of Figure 7.2** on the next page. The distribution from only one sample is, for lack of a better word, funky – and definitely not normally distributed.

The right panel of Figure 7.2 shows the sampling distribution for this same variable, which was obtained by repeatedly drawing samples of size 50 from the NHANES dataset and taking the mean. Despite the clear non-normality of the original data (as shown in the left panel), the sampling distribution with 50 samples (as shown in the right panel) is remarkably close to the normal.

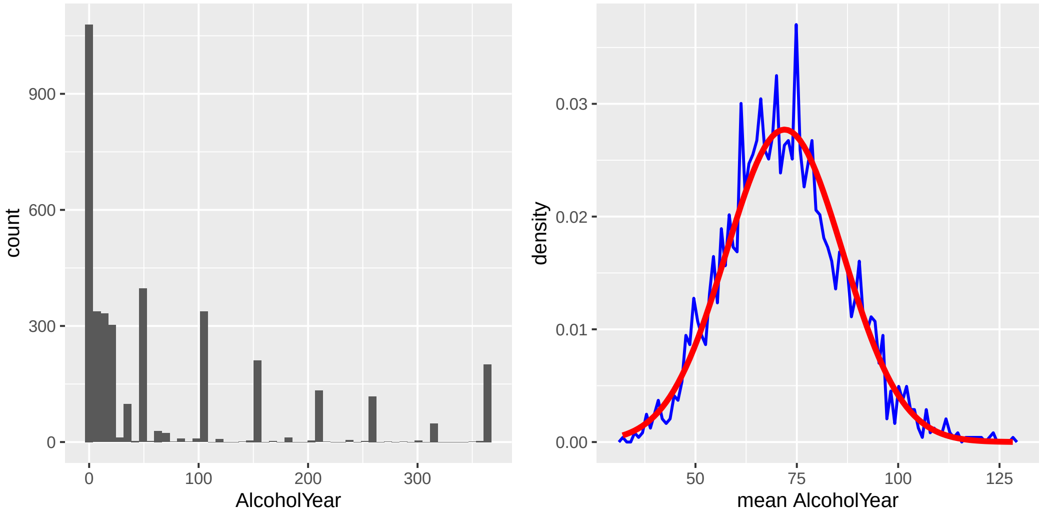

Figure 7.2:
Left: Distribution of the variable AlcoholYear in the NHANES dataset, which reflects the number of days that the individual drank in a year.

Right: The sampling distribution of the mean for AlcoholYear in the NHANES dataset, **obtained by drawing repeated samples of size 50**, in blue. The normal distribution with the same mean and standard deviation is shown in red.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. Because the distribution is normal, we can take advantage of statistical techniques that assume a normal distribution.