Chapter 3 Confidence Intervals and Percentiles

3.1 Confidence Intervals

In statistics, we usually don't know the exact values of the population parameters so we use statistical methods to approximate them. For example, we might take a random sample from the population, calculate the sample mean, and use that value as an estimate for the population mean. The sample mean is an example of a **point estimator** because it is a single value, or a point. There are also **interval estimators**, which instead offer a range, or interval, of values which are likely to contain the value we are trying to estimate. Perhaps the most common interval estimator is the **confidence interval**.

3.1.1 Definition and Interpretation

A confidence interval is always associated with some percentage, or a probability. For example, a 95% confidence interval. Second, we find confidence intervals for values. Perhaps the most common confidence interval is a **95% confidence interval for the mean**. The correct interpretation of this 95% confidence interval would be: "We are 95% confident that the true mean lies within the lower and upper bounds of the confidence interval."

Notice that with this interpretation we aren't saying anything about the exact value of the population mean, we are simply giving a range of values that the mean is very likely to lie in.

3.1.2 Finding a Confidence Interval

More often than not, confidence intervals will be 95% confidence intervals. Think back to the **68-95-99.7% Rule** from last chapter, especially note the 95. 95% of the data lies within two standard deviations of the mean, so by computing two standard deviations we compute the 95%

confidence interval; with the upper bound of the CI being the mean plus two standard deviations ($\mu + 2\sigma$), and the lower bound being the mean minus two standard deviations ($\mu - 2\sigma$).

3.1.3 Example

Consider the 95% confidence interval for the true mean of 25 rolls of a fair die. We found the 95% confidence interval to be: (2.37,3.71). When we interpret this confidence interval, we say, “We are 95% confident that the true mean is between 2.37 and 3.71.”

The word, “confident” implies that if we repeated this process many, many times, 95% of the confidence intervals we would get would contain the true mean μ. It does not imply anything about whether or not one specific confidence interval will contain the true mean.

We do not say that “there is a 95% probability (or chance) that the true mean is between 2.37 and 3.71.” The probability that the true mean μ is between 2.37 and 3.71 is either 1 or 0.

3.2 Percentiles

Refer back to Chapter 1 for a brief refresher on the definition of a percentile, but this section will go into greater detail.

Imagine a very long street with houses on one side. The houses increase in value from left to right. At the left end of the street is a small cardboard box with a leaky roof. Next door is a slightly larger cardboard box that does not leak. The houses eventually get larger and more valuable. The rightmost house on the street is a huge mansion.

Notice that if there was a fence between each house, it would take 99 fences to separate the houses.

house 1 | house 2 | … | house 99 | house 100

The home values are representative of data. If we have a list of data, sorted in increasing order, and we want to divide it into 100 equal groups, we only need 99 dividers (like fences) to divide up the data. The first divider is as large or larger than 1% of the data. The second divider is as large or larger than 2% of the data, and so on. The last divider, the 99th, is the value that is as large or larger than 99% of the data. These “dividers” (i.e. the fences) are called percentiles. A percentile is a number such that a specified percentage of the data are at or below this number. For example, the 99th percentile is a number such that 99% of the data are at or below this value. As another example, half (50%) of the data lie at or below the 50th percentile. The word “percent” means “÷100.” This can help you remember that the percentiles divide the data into 100 equal groups.

Quartiles are special percentiles. The word “quartile” is from the Latin quartus, which means “fourth.” The quartiles divide the data into four equal groups. The quartiles correspond to specific percentiles. The first quartile, Q1, is the 25th percentile. The second quartile, Q2, is the same as the 50th percentile or the median. The third quartile, Q3, is equivalent to the 75th percentile.