6.1: Introduction to Confidence Intervals

Learning Objectives

- Motivate the need for interval estimates
- Introduce and interpret confidence intervals
- Introduce margin of error
- Deduce information from confidence intervals knowing the general form

Review and Preview

Recall that in conducting inferential statistics, we are interested in understanding parameters, facts about a population, without having to do the work of studying the entire population. We want to study a sample and use the facts about it, the sample statistics, to estimate a population parameter. In the last chapter, we developed sampling distributions that connect the possible values of sample statistics with their probabilities. We found that, in general, we can approximate sampling distributions fairly well using continuous random variables. As such, we solidified our growing intuition that we should not expect a sample statistic computed from a simple random sample to be exactly equal to the population parameter. We expect there to be a difference between the two. The actual difference between a computed sample statistic and the population parameter must be unknown because the population parameter is unknown. Simply using a sample statistic as an estimate of the population parameter is insufficient. Instead, we estimate the population parameter by developing an interval estimate, called a **confidence interval**, based on the sample statistics and the sampling distribution.

In constructing an interval estimate, we hope to provide a meaningful range of values in which we feel confident that the population parameter is located. There are two competing desires here: the confidence that the population parameter is in the interval estimate, which we would like to be fairly high, and the length of the interval which we would like to be fairly small. But, as one might guess, increasing the confidence results in larger intervals. So, it is a balancing act, but luckily there is another factor at play that can help us manage both desires that we will study throughout this chapter. Let us begin our development of confidence intervals.

The Differences and Their Probabilities

Recall the text exercise about the grade distribution of statistics students for a particular instructor over several semesters. We computed the probability that the difference between the population mean 71% and the sample mean from a random sample of past students was less than 2%. For a sample size of 36, the probability that the sample mean fell within 2% of the population mean was 45.1494%.

We now make an important but seemingly trivial note. If the sample mean is within a certain distance of the population mean, then the population mean is within that same distance of the sample mean. We can translate the previous probability statement as such: the probability of randomly selecting a sample of size 36 so that the population mean is within 2% of the sample mean is about 45%. So, for about 45% of the samples of size 36, the population mean lies within 2% of the sample mean. Which is to say, for about 45% of the samples of size 36, the population mean lies in the interval $(\bar{x}-2, \bar{x}+2)$, where $\bar{x}$ is the computed sample mean from the randomly chosen sample. We might randomly select a sample of size 36 and compute a sample mean of 67%. In which case, the interval $(\bar{x}-2, \bar{x}+2)=(65, 69)$ does not contain 71 (see the red interval below). On the other hand, we might randomly select a sample of size 36 and compute a sample mean of 72.5, resulting in an interval of $(70.5, 74.5)$(see the dark blue interval below). We note that the population mean 71 does fall in this interval. Using a sample size of 36 with a difference of 2% results in the population parameter lying in the constructed interval about 45% of the time. We want to construct an interval that we are more confident in its ability to catch the parameter.

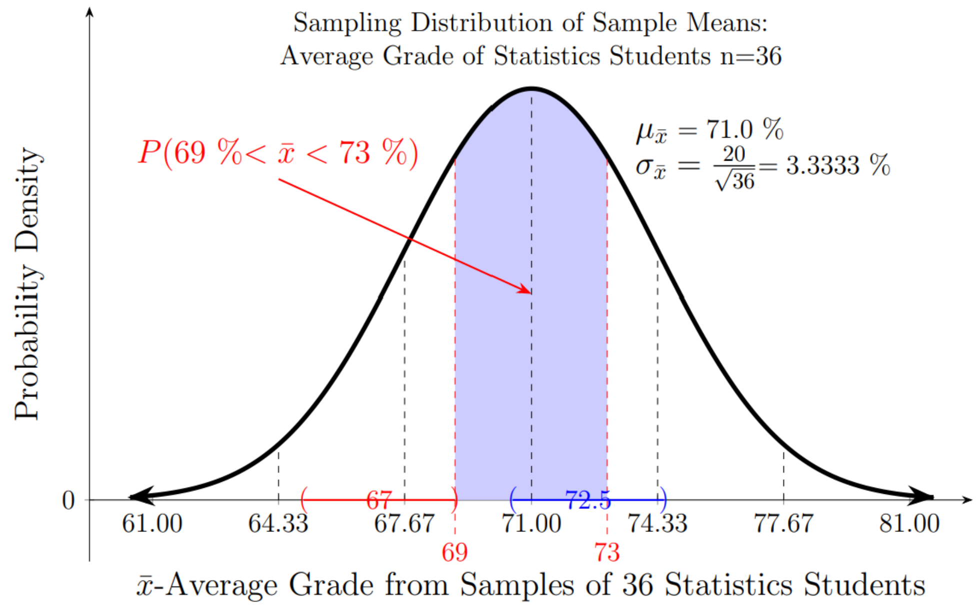

Figure $6.1.1$: Sampling distribution of sample means $n = 36$

The text exercise progressed through various sample sizes to see the effect sample size had on the probability of the sample mean falling within 2% of the population mean. We noticed that the probability increased as the sample size increased. For a sample size of 288, the probability that the sample mean fell within 2% of the population mean was 91.0314%.Making a similar translation as before, the probability of randomly selecting a sample of size 288 so that the population mean is within 2% of the sample mean is about 91%. This means that for about 91% of the possible random samples of size 288 taken from our population, the interval constructed from our computed sample mean, $(\bar{x} - 2, \bar{x} + 2)$, will contain the population mean. If we did not know the population mean, we could not be sure which intervals successfully caught the population mean, but knowing that $91.0314\%$of the possible random samples of size 288 produce an interval containing the population mean elicits a certain confidence that most of the time we are successful in catching the population mean.

Herein lies an understanding of the name **confidence interval.** The **confidence level (CL)** of a confidence interval is the percentage of times (if we conducted random sampling repeatedly) that we would expect the population parameter to fall in our constructed interval. In general, the confidence level is set first and then the confidence interval is constructed to ensure the level of confidence. We now begin to fix the discrepancy in ordering.

Suppose, we want to construct a confidence interval at the level of 95% using a sample of size 288. Being within 2% of the population mean was insufficient to produce such a level of confidence. To increase our confidence, we must increase the distance the sample means are within. Will 3% be enough or 4%? Can we find the precise distance? Consider these questions in conjunction with the following figure.

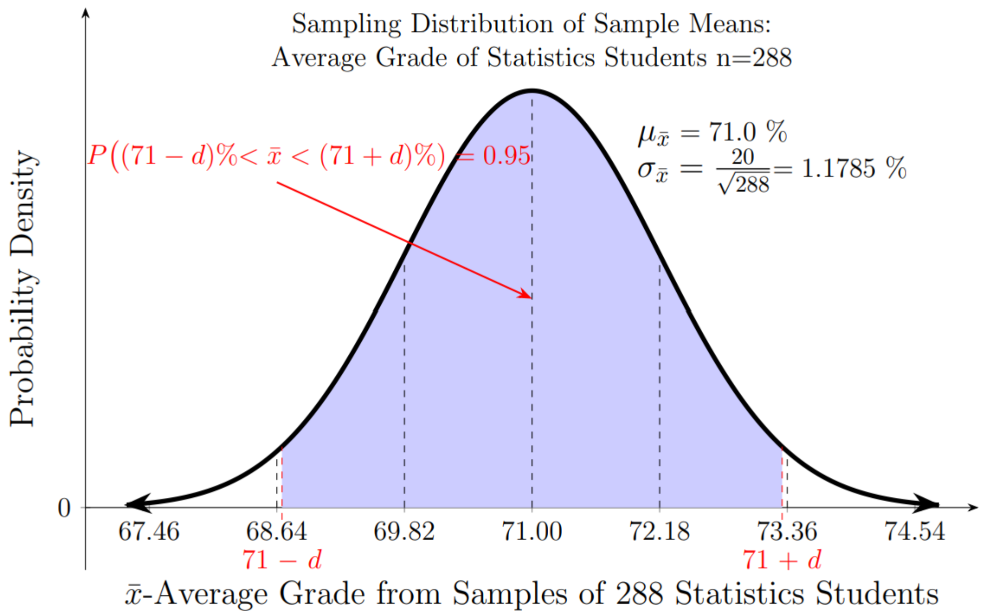

Figure $6.1.2$: Sampling distribution of sample means $n = 288$

We are trying to find the distance d such that 95% of all samples of size 288 produce sample means that are within $d\%$ of the population mean. The shaded region has an area of 0.95, and the lower and upper bounds must be $d\%$ below and above the population mean, respectively. Using the fact that the total area under a probability density function is 1. We know that the total area of the white regions, the two tails, is $1-0.95 = 0.05.$. Notice that the two tails each have the same area since our probability density curve and our shaded region are symmetric about the mean. From this, we can deduce that the left tail, the white region on the left, has an area of $\frac{0.05}{2} = 0.025$. Now we know that $\text{NORM.INV}(0.025, 71, \frac{20}{\sqrt{288}}) \approx 68.6902$ is the lower bound of our shaded region, which also happens to be $71-d$. We have $d = 71 - \text{NORM.INV}(0.025, 71, \frac{20}{\sqrt{288}}) \approx 71 - 68.6902 = 2.3098$. Both 3% and 4% produce an interval wider than is necessary to attain 95% confidence.

95% of all samples of size 288 from our population produce sample means that are within 2.3098% of the population mean. If we construct an interval $(\bar{x} - 2.3098, \bar{x} + 2.3098)$ using the computed sample mean from a random sample of size 288, we would expect to catch the population mean in the constructed interval, 95% of the time. We would call 2.3098% our margin of error.

Confidence Intervals

Since the confidence level is a major driving force in constructing the confidence interval, the confidence level is given in conjunction with the confidence interval; a 95% confidence interval is a confidence interval constructed at the 95% confidence level. Recall that this indicates that the method of constructing the confidence interval, 95% of the time, produces an interval containing the population parameter. The remaining 5% of the time, the method fails to catch the population parameter; this rate of expected failure is often referred to as the **α value** (lowercase Greek letter alpha) of the confidence interval. The confidence level CL and α values are related to each other. When we construct a confidence interval, we either successfully catch the parameter or fail to catch the parameter. There are no other options. As such, the success rate plus the fail rate must be 1. Hence, $\text{CL} + \alpha = 1$. Common confidence levels are 90%, 95%, and 99%, but confidence levels can theoretically be any positive value less than 1 (100%). We can determine α for these common confidence levels since CL and α are complementary. So, we have α values of $0.1, 0.05,$ and $0.01,$ respectively.

If our confidence level is CL, our task is to answer the question: CL of the sample statistics are within what distance of the population parameter? This distance is called the **margin of error $(\mathbf{ME})$**, and it depends on the sampling distribution. An attentive reader will recognize a possible issue here. We knew the population mean was 71 when we computed d, the margin of error. How are we to compute such a distance if we do not know what the population parameter is? The solution is quite simple. We focused our study of sampling distributions primarily on two statistics: sample mean and sample proportion. Both of these statistics are unbiased estimators of their associated population parameters. This is important because we know that each respective sampling distribution's mean (expected value) is the associated population parameter. When the sampling distribution is approximately normal, computing the margin of error reduces to computing the number of standard deviations from the mean that are necessary to gain the given confidence. We will discuss the specifics of computing the margin of error in later sections. For now, we recognize that the margin of error depends on the confidence level and the standard deviation of the sampling distribution (commonly referred to as the standard error). This method of confidence interval construction results in confidence intervals of the form: $(\text{sample statistic} - \text{margin of error}, \text{sample statistic} + \text{margin of error})$.

Text Exercise 6.1.1

Use this information to answer the following questions. The state senate needs a two-thirds majority vote to override a governor's veto. A large random sample of senators was taken to estimate the percentage of senators who support overriding the veto. The 90% confidence interval for proportions $(0.67, 0.73)$ was constructed using the sample data and the method discussed above.

1. Explain the meaning of the 90% confidence level and the resulting confidence interval in the context of the problem.

Answer

The 90% confidence level indicates that the method of constructing the confidence interval catches the population proportion 90% of the time. So, we are 90% confident that the population proportion p falls somewhere between 67% and 73%.

2. Based on the confidence interval, would you recommend that a proponent of the override motion initiate a vote?

Answer

Since the lower boundary of the confidence interval is 67%**,** which is greater than two-thirds, the proponent of the override can feel confident that the Senate can override the governor's veto, although it appears that the margin of victory will not be very large. We, therefore, recommend that the proponent initiate the vote.

3. Determine the computed sample proportion $\hat{p}$ and margin of error (ME).

Answer

Because this line of reasoning applies to any confidence interval constructed using the method discussed above, we answer this question first in generality and then in particular. Since the confidence interval was constructed using the method discussed above, we know that the lower bound is equal to the sample statistic minus the margin of error and that the upper bound is equal to the sample statistic plus the margin of error. So we have the following system of equations.

$$\begin{aligned}\text{lower bound } &= \text{sample statistic} - \text{margin of error}\\ \text{upper bound } &= \text{sample statistic} + \text{margin of error}\end{aligned}$$

We can eliminate the margin of error by adding the two equations together.

$$\text{lower bound} + \text{upper bound} = 2 \cdot \text{sample statistic}$$

We can eliminate the sample statistic by taking the difference of the two equations.

$$\text{upper bound} - \text{lower bound} = 2 \cdot \text{margin of error}$$

We have developed the following formulas.

$$\text{sample statistic} = \frac{\text{lower bound} + \text{upper bound}}{2} \qquad \text{margin of error} = \frac{\text{upper bound} - \text{lower bound}}{2}$$

This confirms that the sample statistic is the midpoint of the confidence interval and that the margin of error is half of the length of the confidence interval. So $\hat{p}$ is the midpoint between 0.67 and 0.73 which is 0.70, and ME is the distance from the midpoint to an end or half of the interval length which is 0.03.

> **Note: Reading Symbols with Meaning**
>
> Reading mathematical symbols with meaning is an important skill to develop in a quantitatively and symbolically driven society. As we have seen in the development of confidence intervals, the margin of error, ME, is the distance such that the confidence level, CL, produces sample statistics within that distance of the population parameter. We can understand this equivalently as the maximum distance a sample statistic could be from the population parameter and still be captured in the confidence interval. Since we are conducting simple random sampling in the context of this course, we can understand that statement as the probability of randomly selecting a sample of a given size that produces a sample statistic within ME of the population parameter is CL. This can be expressed with symbols quite elegantly. The symbols for statistics and parameters differ based on the context; so, for now, let us remain in the context of means. The statements made above can be expressed as $P\left(|\bar{x}-\mu|<\text{ME}\right)=\text{CL}$.
>
> Read $|\bar{x}-\mu|$, the absolute value of the difference between the sample mean and population mean, as the distance between the sample mean and population mean.
>
> Read $|\bar{x}-\mu|<\text{ME}$ as the distance between the sample mean and population mean is less than the margin of error.
>
> Read $P\left(|\bar{x}-\mu|<\text{ME}\right)$ as the probability that the distance between the sample mean and population mean is less than the margin of error.
>
> Read $P\left(|\bar{x}-\mu|<\text{ME}\right)=\text{CL}$ as the probability that the distance between the sample mean and population mean is less than the margin of error is the confidence level.
>
> It is important to remember that the random variable at play is the sample statistic because the underlying random experiment is conducting a random sample of a given size.
>
> > **Text Exercise 6.1.2**
> >
> > Read the following mathematical expression with meaning: $P\left(|\hat{p}-p|>\text{ME}\right)=1-\text{CL}=\alpha$
> >
> > **Answer**
> >
> > The probability that the distance between the sample proportion and population proportion is greater than the margin of error is one minus the confidence level, which is the alpha value.
