Confidence interval for the proportion

In real life, when we're interested in a proportion, we usually won't know the population proportion p, because we won't be able to survey or test every subject within our population. For instance, we might want to know the proportion of people in our country who support a particular political candidate, but, since we can't ask every person in the country, we can't truly know the population proportion, p.

Instead, we have to take a smaller sample of our larger population, and then compute the sample proportion $\hat{p}$. Once we find $\hat{p}$, we can use it to make inferences about the value of p.

We'll find the sample proportion $\hat{p}$ by taking a sample with n subjects and surveying the number of those subjects that meet our criteria. Out of that sample, the percentage of subjects that meet our criteria will be $\hat{p}$.

$$\hat{p} = \frac{\text{number of subjects that meet our criteria}}{n}$$

The confidence interval for the sample proportion is given by

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* comes from the z-table, $\hat{p}$ is the sample proportion, and n is the sample size. We use $\hat{p}$ in the confidence interval formula (instead of p) because, if we're constructing a confidence interval, by definition that means we're trying to use a sample proportion to estimate the population proportion, which means we don't know the population proportion. Since

we don't know the population proportion, we use the sample proportion $\hat{p}$ in our formula, instead of the population proportion p.

In order to be able to use this formula, we need to have $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, which will almost always be the case in real life, as long as our sample size is reasonably large.

The finite population correction factor applies here as well, so if we're sampling without replacement from more than $5\,\%$ of a population of finite size N ($n/N > 0.05$), then the confidence interval for the population proportion is given by

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

So if we know how we're sampling, what confidence level we want to use, and we know the sample proportion, then we can plug these values into the correct formula, find the critical value associated with the confidence level, and then calculate the confidence interval directly.

Let's do an example so that we can see how to calculate a confidence interval for the population proportion.

Example

There are 500 sea turtles that live in a bay off of Maui, Hawaii, and we want to estimate the proportion that are male. Let's say we take a random sample of 50 turtles and find that 20 of them are male.

Based on this sample, what is a $90\,\%$ confidence interval for the proportion of male sea turtles in the bay.

As always, we have to first check for normality. We were told that the sample we took was random.

Based on the sample proportion $\hat{p} = 20/50 = 0.4$, we'll get at least 5 "successes" ($50 \cdot 0.4 = 20$) and at least 5 "failures" ($50 \cdot 0.6 = 30$), so we've met the normal condition. It looks like we're sampling without replacement, using $50/500 = 10\,\%$ of the population, so we'll need to use the finite population correction factor in our confidence interval formula.

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$(a, b) = 0.4 \pm z^* \sqrt{\frac{0.4(0.6)}{50}} \sqrt{\frac{500-50}{500-1}}$$

For a $90\,\%$ confidence interval, we're looking in a normal distribution at the middle $90\,\%$ of probability, which means we'll only have $10\,\%$ probability in the two tails, or just $5\,\%$ in the upper tail, which means we're interested in the z-score that puts us at $95\,\%$ probability. If we look for approximately 0.9500 in the z-table, we get about 1.65. So the critical value z^* is approximately 1.65, and we can say that our confidence interval is

$$(a, b) = 0.4 \pm 1.65 \sqrt{\frac{0.4(0.6)}{50}} \sqrt{\frac{500-50}{500-1}}$$

$$(a, b) = 0.4 \pm 1.65 \sqrt{\frac{0.24}{50}} \sqrt{\frac{450}{499}}$$

$$(a, b) \approx 0.4 \pm 1.65 \sqrt{0.0048} \sqrt{0.9018}$$

$$(a, b) \approx 0.4 \pm 0.1086$$

$$(a, b) \approx (0.4 - 0.1086, 0.4 + 0.1086)$$

$$(a, b) \approx (0.2914, 0.5086)$$

We interpret this to mean that about $90\,\%$ of the confidence intervals we construct this way (with 50-turtle samples) will contain the actual population proportion p of male sea turtles in the bay.

Margin of error

Just like for the confidence interval for the mean, the margin of error for the proportion is simply the part of the confidence interval formula that comes after the $\pm$ sign.

$$z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we want to keep our margin of error at or below a certain value, then we can set up an inequality that will allow us to find the minimum possible sample size we'd need to use in order to keep the margin of error fixed to that preset maximum.

Example

We want the margin of error in our sea turtle study (from the previous example) to be no more than $\pm 4\,\%$ at a $90\,\%$ confidence level. Find the

smallest possible sample size we can use to stay within that margin of error.

First, we'll set up the inequality.

$$z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}} \leq 0.04$$

If we want to find the smallest possible sample size n that keeps us within this margin of error, then we need to optimize $\hat{p}(1-\hat{p})$, since making the numerator of a fraction as large as possible will make the entire fraction as large as possible. In turn, that will make the value of the square root as large as possible, which will make the entire value on the left side of the inequality as large as possible, thereby minimizing the value of n.

We could prove this algebraically, but the value of $\hat{p}$ that optimizes $\hat{p}(1-\hat{p})$ is always $\hat{p} = 0.5$. Therefore, we'll plug everything we know into the margin of error inequality, remembering that a $90\,\%$ confidence level has a critical value of approximately 1.65.

$$1.65\sqrt{\frac{0.5(0.5)}{n}}\sqrt{\frac{500-n}{500-1}} \leq 0.04$$

$$\frac{\sqrt{0.5^2}}{\sqrt{n}} \cdot \frac{\sqrt{500-n}}{\sqrt{499}} \leq \frac{0.04}{1.65}$$

$$\frac{0.5}{\sqrt{n}} \cdot \frac{\sqrt{500-n}}{\sqrt{499}} \leq \frac{0.04}{1.65}$$

$$\frac{0.5}{\sqrt{499}} \cdot \frac{\sqrt{500-n}}{\sqrt{n}} \leq \frac{0.04}{1.65}$$

Solve the inequality for n.

$$\frac{\sqrt{500-n}}{\sqrt{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500-n}{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500}{n} - \frac{n}{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500}{n} - 1} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\frac{500}{n} - 1 \leq \left(\frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2$$

$$\frac{500}{n} \leq \left(\frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2 + 1$$

We can invert both sides if we flip the inequality sign.

$$\frac{n}{500} \geq \frac{1}{\left(\frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2 + 1}$$

$$n \geq \frac{500}{\left(\frac{0.04\sqrt{499}}{1.65 \cdot 0.5}\right)^2 + 1}$$

$$n \geq 230.092$$

If we need to sample more than 230.092 members of our population, that means we need to sample at least 231 of them, because sampling only 230 wouldn't quite be enough to meet our threshold. Therefore, our minimum sample size has to be

$$n \geq 231$$