Set 24 - Confidence Intervals for Population Proportions

March 22, 2023 1:23 PM

Stat 260 Lecture Notes

Set 24 - Confidence Intervals for Population Proportions

↳ percentages

Consider a binomial experiment with n trials where the probability of success is p.
Here we know p. In the real world, maybe we don't. For example:

Setup: 10% of the population has a certain disease. If 20 people are selected at random what is the probability that exactly 3 have the disease?

binomial, $n=20$, $p=0.10$, $P(x=3)$

How would it have been possible to measure $p = 0.10$ here? We would have to guess by looking at a sample of the population.

$$\hat{p} = \frac{X}{n} = \frac{\#\text{ in sample with trait}}{\#\text{ in sample space}}$$

$\hat{p}$ = proportion in the sample

"p hat"

$\hat{p}$ is a point estimate for the population proportion p. (Just like $\overline{x}$ is an estimate of μ.)

Rule: If $np \geq 5$, and $n(1-p) \geq 5$, then the sample proportion $\hat{p} = \dfrac{X}{n}$ is approximately normally distributed. Furthermore, the expected value of $\hat{p}$ is p and the variance is $\dfrac{p(1-p)}{n}$.

Rule: The random variable $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ is approximately standard normal when n is big. (Specifically, when $np \geq 5$, and $n(1-p) \geq 5$.)

$Z = \dfrac{\text{r.v.} - \text{expected value}}{\text{std. dev. of r.v.}}$ ← standard error

Note:

If we know population p, use this in $\sqrt{\dfrac{p(1-p)}{n}}$

(if don't know then use $\hat{p}$)

A $(1-\alpha)\%$ confidence interval for estimating p is

$$[L, U] = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

This can be shortened to

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

estimate ± (c.v.)(e.s.e.)

proportion questions always use z.

Note: We use $\hat{p}$ in $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for the CI because the CI is being used to estimate p, meaning we don't know a value for p yet.

Example 1: A university wants to know what proportion of its students are vegetarian. In a random sampling of 200 students, it was found that 41 students were vegetarian. Find a 95% CI for p, the true proportion of students at the university who are vegetarian.

Follow-up: Is is reasonable to say that 20% of students at the university are vegetarian?

Another follow-up: Is it reasonable to say that 30% of students at the university are vegetarian?

$n = 200 \quad \hat{p} = \frac{41}{200}$

$z_{\alpha/2} = z_{0.025} = 1.96$

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \frac{41}{200} \pm 1.96 \sqrt{\frac{\frac{41}{200}\left(1 - \frac{41}{200}\right)}{200}}$$

$= [0.149, 0.261]$

The CI gives reasonable estimates for p. (anything in interval is reasonable)

Yes, it is reasonable to say that 20% of students are vegetarian, since 20% = 0.20 is in the CI

No, it is not reasonable to say that 30% are vegetarian, since 30% is not in the CI.

Finding the Sample Size for the p Confidence Interval

Let's revisit our $(1-\alpha)\%$ confidence interval for estimating the population proportion p:

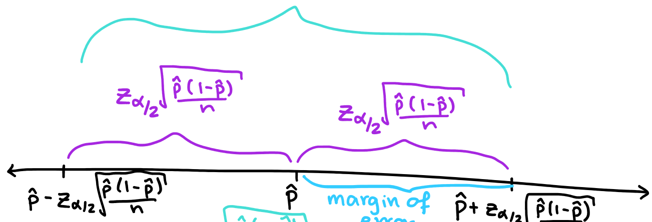
Note: $\hat{p}$ is always at the centre of this CI (we don't know about p) → p can be anywhere

$$\hat{p} \pm \underbrace{\boxed{z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}}_{d}$$

estimate ± (C.V.)(e.s.e.)

C.V. = critical value; e.s.e. = estimated standard error (formula sheet)

A picture of the interval:



width of interval $= 2 \cdot z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

margin of error of interval $= z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = d$ (for distance)

margin of error "within"

How big should n be in order to build the confidence interval so that our estimate is within a desired amount d?

not on formula sheet

$$d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

$$\sqrt{n} = \frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{d}$$

$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{d}\right)^2$$

$$= \left(\frac{z_{\alpha/2}}{d}\right)^2 (\hat{p}(1-\hat{p}))$$

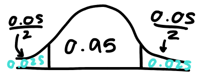
We need at least this value of n to stay within distance d.

Notes: n has to be an integer.
If we round n down, we have a smaller n and so our confidence interval is wider.
We would rather have a narrower interval (since we have to stay within the amount d), so we **always** round up. $n = 123.21 \rightarrow n = 124$

Example 2: Vegetarians at university revisited. A university wants to know what proportion of its students are vegetarian. In a random sampling of 200 students, it was found that 41 students were vegetarian. Find the sample size needed to estimate the true proportion of vegetarians at the university to within 1% with 95% confidence.

d is everything after ± in confidence interval

$\hat{p} = \frac{41}{200}$ $\quad d = 1\% = 0.01$



$Z_{\alpha/2} = Z_{0.025}$
$= 1.96$

$$d = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \left(\frac{Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{d}\right)^2 = \left(\frac{1.96\sqrt{41/200(1-41/200)}}{0.01}\right)^2$$

$= 6260.8476$

round up, $n = 6261$

What if we don't have a good measured guess for $\hat{p}$?

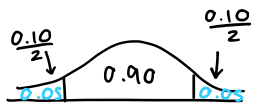
In this case, use $\hat{p} = 0.5$ since $\hat{p}(1-\hat{p})$ is maximized when $\hat{p} = 0.5$. (i.e. This will give us the biggest possible value for the minimum n value needed.)

Example 3: The EPA has identified some waste dumping sites in the US as being potentially dangerous. How large a sample size is needed to estimate the true proportion of sites that are dangerous to within 2% with 90% confidence?

$d = 2\% = 0.02$

don't have $\hat{p}$, use $\hat{p} = 0.5$

$$d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



$z_{\alpha/2} = z_{0.05}$
$= 1.645$

$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{d}\right)^2 = \left(\frac{1.645\sqrt{0.5(0.5)}}{0.02}\right)^2$$

$$= 1691.2656 = 1692$$

always round up for n

Finding the Sample Size for the μ Confidence Interval

Let's revisit our $(1-\alpha)\%$ confidence interval for μ:

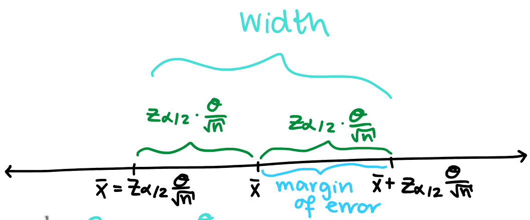
Note: $\bar{x}$ is always the centre of this CI

population mean (μ) can be anywhere

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$ (if don't know θ)

A picture of the interval:



width of interval $= 2 \cdot z_{\alpha/2} \cdot \frac{\theta}{\sqrt{n}}$

margin of error of interval $= z_{\alpha/2} \cdot \frac{\theta}{\sqrt{n}} = d$ margin of error "within"

How big should n be in order to build the confidence interval so that our estimate is within a desired amount d?

$$d = z_{\alpha/2} \cdot \frac{\theta}{\sqrt{n}}$$

$$\sqrt{n} = \frac{z_{\alpha/2} \cdot \theta}{d}$$

$$n = \left(\frac{z_{\alpha/2} \cdot \theta}{d}\right)^2 \quad \text{or} \quad n = \left(\frac{z_{\alpha/2} \cdot s}{d}\right)^2$$

Notes: n has to be an integer.
If we round n down, we have a smaller n and so our confidence interval is wider.
We would rather have a narrower interval (since we have to stay within the amount d), so we **always** round up.

proportion or average?

Example 4: A professor's grades from the past 25 years are stored in a large filing cabinet. The professor is asked to report the historical average final grade of their students, however the professor is lazy and doesn't want to have to input all their marks in a spreadsheet to do the calculation. Their plan is to take a random sample of the grades from the filing cabinet and calculate the average of just these students. Find the sample size needed to estimate the average within 2 marks with 99% confidence. Suppose the professor did a quick small sample calculation to estimate that the standard deviation is 9 marks. $s = 9$ $d = 2$ $z_{0.005} = 2.575$

$$d = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{d}\right)^2 = \left(\frac{2.575(9)}{2}\right)^2 = 134.374$$

$n = 135$

always round up

Remember: here the confidence interval is giving us estimates of the population average μ.
In these questions how is it possible that we have a measured s or σ, but we don't know the number of samples needed n, and we also don't have a guess for μ? A few options:

- We could use s from a previous study as an estimate.
- We could run a small preliminary or pilot study and use the value of s found there to help plan the larger experiment.
- The Normal Probability Rule (also called the Empirical Rule) guarantees that for an approximately normal distribution X will be within 2 standard deviations of the mean 95% of the time,
  i.e. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$. This would mean that the range of our X values is roughly 4 standard deviations. We could use the range divided by 4 to approximate s.

Why don't we use $t_{\nu,\alpha/2}$ in these calculations?

One reason is that we need to know n in order to find the degrees of freedom $\nu = n - 1$. Another reason is that most of the time, the n value we find is so large that it indicates we could use the normal distribution anyways. A third reason is that for a fixed value of $\alpha/2$, the value of $z_{\alpha/2}$ is smaller than the value of $t_{\nu,\alpha/2}$ and so using $z_{\alpha/2}$ in our calculations gives us a minimum value of the n we should be using. (Using the $t_{\nu,\alpha/2}$ would give a larger n value.)