12.1: Confidence Intervals for One Population Proportion

Recall: A *parameter* is a numerical summary of a population; a *statistic* is a numerical summary of a sample. (For example, the population mean and population standard deviation are parameters; the sample mean and sample standard deviation are statistics.)

The *sampling distribution of a statistic* is the probability distribution of all possible values for that statistic computed from all possible samples of fixed size n.

The sampling distribution of the sample proportion:

In this section, the parameter we are interested in is the *population proportion*, usually denoted p.

The proportion is the percentage p (in decimal form) of the population that possesses some characteristic of interest.

For example, we may be interested in the proportion of children who have a certain medical condition, the proportion of U.S. citizens who received a tax refund, the proportion of students at a certain high school that decide to go to college, or the proportion of nurse candidates who pass the nursing licensure exam.

The *sampling distribution of the sample proportion* is the probability distribution of all possible values for the sample proportion, denoted $\hat{p}$, computed from all possible samples of fixed size n.

$\hat{p}$ (p-hat)

If x is the number of data points in a sample of size n that have the characteristic of interest, then the sample proportion is

$$\hat{p} = \frac{x}{n}.$$

In the same manner as for the sample mean, we use the sample proportion $\hat{p}$ to make inferences about the population proportion p.

Shape, mean and standard deviation of the sampling distribution of the sample proportion:

Sampling distribution of the sample proportion:

Suppose random samples of size n are taken from a population with population proportion p.

Also suppose that the sample size is small compared to the size of the population.
(Rule of thumb: The sample must be less than 5% of the population size; otherwise we must use a finite population correction factor, which is beyond the scope of this class.)

Then:

The shape of the sampling distribution of $\hat{p}$ is approximately normal, provided that the sample sizes are sufficiently large.

Rule of thumb: to assume the sample proportion is normally distributed, we need both $np \geq 5$ and $n(1-p) \geq 5$.

If $q = 1-p$,

$np \geq 5$ and $nq \geq 5$

The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

The standard deviation of the sampling distribution of : $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Point estimates for the population proportion:

Recall:

Definition: A *point estimate* is the value of a statistic that estimates the value of a parameter.

Definition: A confidence interval for an unknown parameter is an interval of numbers generated by a point estimate for that parameter.

Definition: The *confidence level* (usually given as a percentage) represents how confident we are that the confidence interval contains the parameter.

If a large number of samples is obtained, and a separate point estimate and confidence interval are generated from each sample, then a 95% confidence level indicates that 95% of all these confidence intervals contain the population parameter.

A confidence interval is obtained by placing a *margin of error* on either side of the point estimate of the parameter.

In other words, the confidence interval consists of: Point estimate $\pm$ margin of error

The point estimate of the population proportion p is the sample proportion $\hat{p}$.
The point estimate of the mean of the sampling distribution of the sample proportions is $\mu_{\hat{p}} = \hat{p}$.

The point estimate of the standard deviation of the sampling distribution of the sample proportions is

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \qquad \text{(standard error)}$$

So, for every sample, the sample proportion will be in the center of the confidence interval. If we use E to indicate the margin of error, the confidence interval is $\hat{p} \pm E$, or $(\hat{p}-E, \hat{p}+E)$

If we use the sample proportion $\hat{p}$ as a starting point, we should be able to write the confidence interval as $(\hat{p} - z_c\sigma_{\hat{p}}, \hat{p} + z_c\sigma_{\hat{p}})$, where $\sigma_{\hat{p}}$ is the standard deviation of the sampling distribution of the sample proportions, and z_c is a multiplier that tells us how many standard deviations (of the sampling distribution of the sample proportions) lie between the sample proportion $\hat{p}$ and the edge of the confidence interval. We call this z_c the *critical value* for a *z*-score in the sampling distribution of the sample proportions.

Constructing the confidence interval for the proportion:

Procedure:

1. Verify that $np \geq 5$ and $n(1-p) \geq 5$ and that the sample is no more than 5% of the population.

2. Determine the confidence level, $1-\alpha$.

3. Determine the critical value $z_{\alpha/2}$ (using the standard normal table).

4. Use the sample proportion to estimate the standard deviation of the sampling distribution of the sample proportions:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

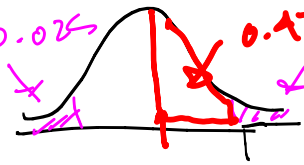
5. Multiply the critical value $z_{\alpha/2}$ by the estimated standard deviation $\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to obtain the margin of error.

6. Add and subtract the margin of error from the sample proportion to obtain the lower and upper bounds of the confidence interval:

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

   Upper bound: $\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\alpha = 0.05 \Rightarrow$



Looking up Area $= 0.475$ in table $\Rightarrow z_{0.025} = 1.96$

Useful critical values of z:
(You can use these instead of looking in the table every time).

Confidence Level	α	Area in each tail, $\alpha/2$	Critical value $z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.575

Example 1: In a random sample of 537 Americans, 173 indicated that they frequently ate peanut butter. Construct and interpret the 90% and the 95% confidence intervals for the proportion of Americans who frequently eat peanut butter. p = proportion of people who frequently eat PB.

Sample info

$n = 537$

$\hat{p} = \frac{x}{n} = \frac{173}{537} \approx 0.3222$

$\hat{q} = 1 - \hat{p} = 1 - 0.3222 = 0.6778$

Note: $n\hat{p} = 537(0.322) = 172.9 > 5$

$n\hat{q} = 537(0.6778) = 363.97 > 5$

ⓐ Construct the 90% CI.

$\alpha = 0.10$

From normal table, the critical value is $z_{0.05} = 1.645$

Find std error:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.3222(0.6778)}{537}} = 0.020166$$

Lower bound: $\hat{p} - z_{0.05}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.3222 - 1.645(0.020166)$

$= 0.289$

Upper bound: $\hat{p} + z_{0.05}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.3222 + 1.645(0.020166)$

$= 0.355$

See next page

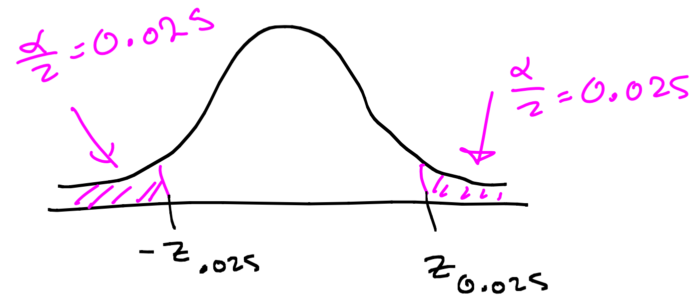
$\boxed{\text{90\% CI: } (0.289, 0.355)}$

we're 90% confident that the proportion of Americans freq. eating PB is between 0.289 and .355

Example 1 cont'd:

Construct 95% CI:

$\alpha = 1 - 0.95 = 0.05$



Use normal table to get critical value $z_{0.025}$:

$z_{0.025} = 1.96$

Lower bound: $\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

$= 0.3222 - 1.96(0.020166)$

$= 0.2827$

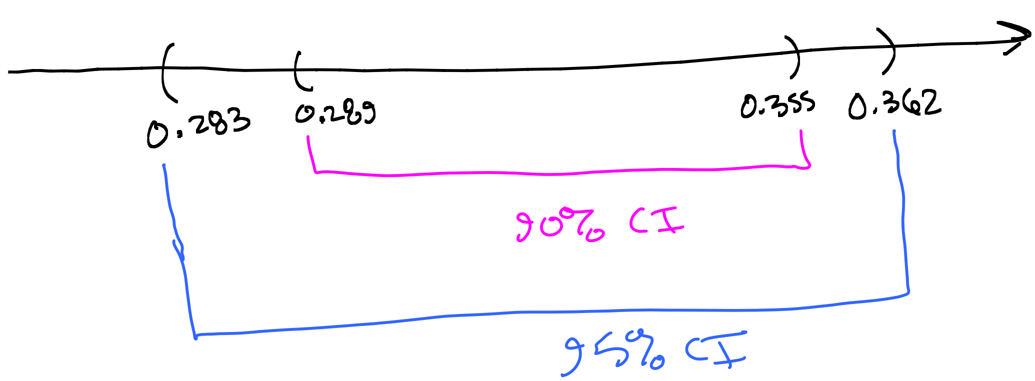
Upper bound: $\hat{p} + z_{0.025}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

$= 0.3222 + 1.96(0.020166)$

$= 0.3617$

95% CI: $\boxed{(0.283, 0.362)}$

Notice: 95% CI is wider than 90% CI: We sacrifice precision for increased certainty.



Sample size needed to estimate the population proportion within a given margin of error:

When constructing a confidence interval about the sample proportion $\hat{p}$, the margin of error is

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

We can solve this for *n*:

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

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

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

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

In order to calculate the *n* needed, we need an educated guess for the population proportion *p*. If such an educated guess is available (perhaps from a prior study), we can use the above formula to calculate *n*. If not, we use the very conservative assumption that $\hat{p} = 0.5$, which gives us the maximum possible value for $\hat{p}(1-\hat{p})$, which is (0.5)(0.5) = 0.25.

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

Required sample size for estimation of the population proportion:

a) For a specified α associated with a confidence level, the sample size required to estimate the population proportion within a margin of error E is

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

where $\hat{p}_g$ is an educated guess for the population proportion p.

b) If you know a likely range of values for the sample proportion, choose the value in that range that is closest to 0.5. Use this value as the educated guess $\hat{p}_g$ in the above formula. (The above formula will be at its maximum when $\hat{p}_g = 0.5$. Thus a larger sample is required when $\hat{p}_g$ is close to 0.5, compared to when $\hat{p}_g$ is further away. To be sure we have a big enough sample, we look at all the possible values for $\hat{p}$ and choose the one closest to 0.5.)

c) If no estimate for the population proportion is available, we should use a sample size of at least

$$n = 0.25\left(\frac{z_{\alpha/2}}{E}\right)^2.$$

In all cases, because the calculated n is considered a minimum threshold, we round the calculated value of n up to the nearest whole number *above*.

Example 2: A pollster wishes to estimate the percentage of likely voters who support Candidate A. Based on earlier polls, the pollster expects the candidate's level of support to be approximately 38%. What sample size should be obtained if the pollster wishes to estimate the candidate's support level within a margin of error of 3 percentage points, with 95% confidence?

$\hat{p}_g = 0.38$

$1 - \hat{p}_g = \hat{q} = 1 - 0.38 = 0.62$

E = Margin of Error: 0.03 (3 percentage points)

$\alpha = 1 - 0.95 = 0.05$

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

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

$$= 0.38(0.62)\left(\frac{1.96}{0.03}\right)^2 = 1005.65 \Rightarrow \text{use } 1006$$

When calculating required sample size, always round up.

Get a sample of at least 1006

Example 3: a) Estimate the minimum sample size required to estimate the population proportion within a margin of error of 0.02, if the proportion is expected to be between 0.1 and 0.3. Use a confidence level of 95%.
b) Suppose the sample size from part (a) is obtained, and that the proportion of the characteristic of interest turns out to be 0.25. Construct the 95% confidence interval for the population proportion. What is the margin of error?

(a) $E = 0.02$

$\hat{p}$ expected to be between 0.1 and 0.3 $\Rightarrow$ use $\hat{p}_g = 0.3$
(choose the proportion closest to 0.5)

$1 - \hat{p}_g = 1 - 0.3 = 0.7$

$\alpha = 0.05 \Rightarrow z_{\alpha/2} = 1.96$

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

$= 0.3(0.7)\left(\frac{1.96}{0.02}\right)^2 = 2016.84$

$\Rightarrow$ $\boxed{\text{required sample size is } 2017}$

Example 4: Estimate the minimum sample size required to estimate the population proportion within a margin of error of 0.03, if you have no idea what the proportion will turn out to be.