Estimating the Population Proportion

Statistical Inference

Point Estimation:

Estimate the population parameter with a ‘single number’ calculated from the sample.

Interval Estimation:

Estimate the population parameter with an interval of values that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

Hypothesis Testing:

we have some claim about the population, and we check whether or not the data obtained from the sample provide evidence against this claim.

Example

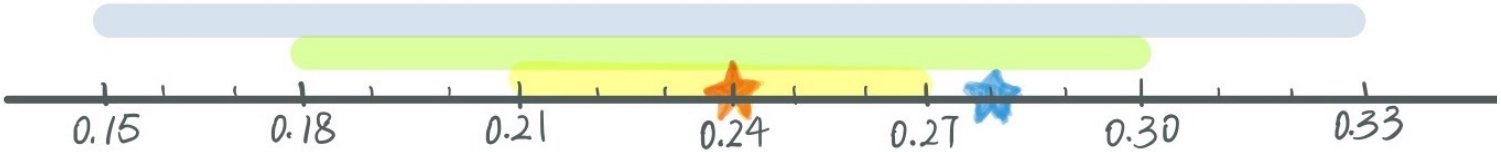
It is known that 24% of all American drivers admit to texting while driving. We randomly select 200 American drivers, and asked if they text while driving.

According to the central limit theorem (check the conditions), the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with a mean of p and a standard error of $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N(0.24, 0.03)$$

"_____% of all possible values of $\hat{p}$ we get from random samples falls within _____units of p"

Rephrase: "We are ___% **confident** that the population proportion p falls within _____ units of $\hat{p}$."



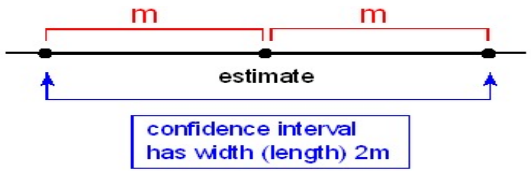
Suppose we happen to get a sample proportion of $\hat{p} = 0.28$, then "we are 95% confident that the population proportion p falls within 0.06 units (2*0.03 units) of 0.28", or "we are 95% confident that this interval (0.22, 0.34) contains the population proportion".

Confidence Intervals

A **confidence interval** is an interval/range of likely values for the population parameter.

The general form: $\mathbf{point\ estimate} \pm \mathbf{margin\ of\ error}$

For population proportion: $\hat{p} \pm m$



The margin of error is selected to produce the desired level of confidence. $m = z^* \times \mathbf{SE}$

(z^* -- confidence multiplier, the number of standard errors to include in the margin of error.)

The confidence interval for p is: $\hat{p} \pm z^* \times \sqrt{\frac{p(1-p)}{n}}$

For unknown population proportion, we replace p with $\hat{p}$:

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

Confidence Level and the Margin of Error

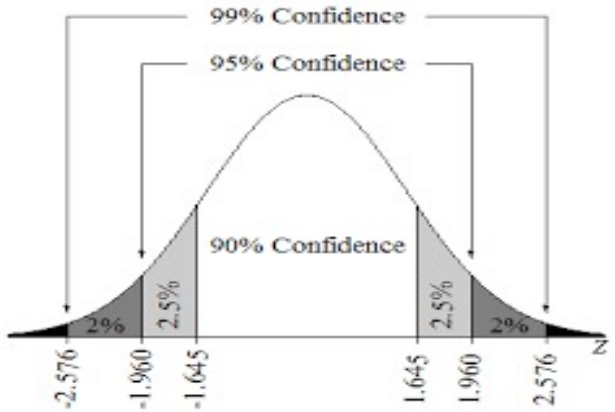
We can change the confidence level by changing the margin of error:

- **Increasing the margin of error increases our confidence level.**
- **Decreasing the margin of error decreases our confidence level.**

For a given confidence level, the margin of error is:

$$\mathbf{m} = z^* \times \mathbf{SE}$$

Confidence Level	Margin of Error
99%	2.576 SE
95%	1.96 SE
90%	1.645 SE
80%	1.28 SE



Confidence Level

If a same survey is conducted repeatedly on separate random samples, producing many confidence intervals for the population proportion, how often will the confidence intervals capture (or contain) the population proportion?

Confidence level can be thought of as the **capture rate,** measures the **success rate** of the method, not any one particular confidence interval.

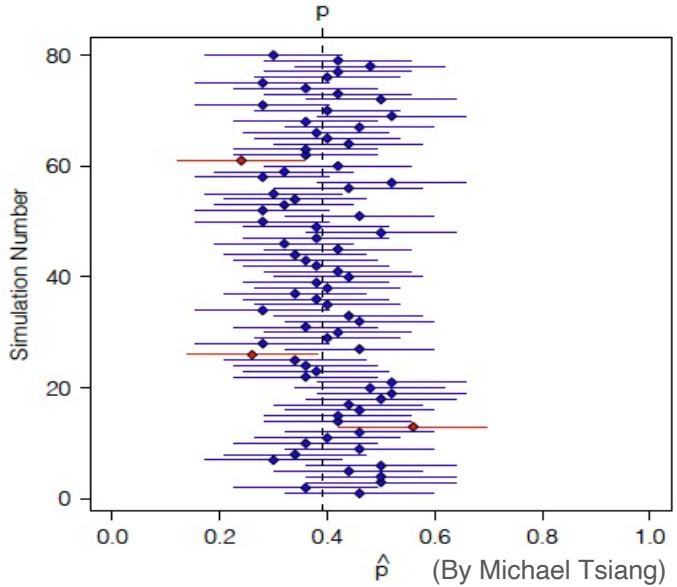
Simulation example:

The proportion of the 817 Major League Baseball players in 2015 that had a salary of less than $1 million is 0.3905.

1. Draw a simple random sample of 50 players.
2. Compute the sample proportion $\hat{p}$ of how many players out of 50 sampled players made less than $1 million in 2015 and the standard error of $\hat{p}$.
3. Construct a 95% confidence interval ($\hat{p} \pm 1.96\text{SE}$) for the population proportion.
4. Repeat steps 1-3 a total of 10,000 times. Each time, record $\hat{p}$ and the 95% confidence interval.

Simulation Example

- The plot below shows the confidence intervals for the first 80 simulations.
- The true proportion $p = 0.3905$ is indicated by the dashed vertical line.
- The blue and red diamonds represent the sample proportion $\hat{p}$ from each random sample.
- Confidence intervals which contain p are in blue.
- Confidence intervals which do not contain p are in red.



Of the 10,000 simulations, 9,471 of the constructed confidence intervals contain the population parameter p, so our interval estimation method succeeded 94.71% of the time, which is approximately 95%.

Interpreting Confidence Intervals

A common misinterpretation would be to say:
"There is a 95% chance that the population proportion is between p_1 and p_2."

This is completely wrong! The population proportion is a fixed value, so it is either always between p_1 and p_2 or it is not. There is no probability involved.

Correct Interpretation

A confidence interval gives a set of values that are plausible for the population proportion.

We often say "We are **95% confident** that the population proportion is between p_1 and p_2 " to mean that if we were to repeat our estimation method on many random samples, 95% of the intervals would contain the true population proportion.

Example: Finding CI

In a random sample of 100 UCLA students, 35 have travelled outside the U.S.

Estimate the true population proportion of UCLA students who have traveled outside the US with a 95% confidence level.

1. Check the conditions for CLT.

2. Find $\hat{p},\ SE,\ z^*$.

3. Calculate the confidence interval.

$$\hat{p} \pm z^* SE$$

Example Continued

Which of the following interpretations of the confidence interval is correct? (select all that apply)

A. If we took another random sample of 100 UCLA students, there is a 95% chance that the sample proportion would be between 0.26 and 0.44.

B. There is a 95% chance that the true population proportion is a value between 0.26 and 0.44.

C. We are 95% confident that this interval captured the true population proportion of UCLA students who have travelled outside the US.

D. If we send out 200 pollsters and each construct an interval from separate independent samples, we can expect about 190 of those intervals to contain the true proportion of UCLA students who have travelled outside the U.S.

Comparing Proportions from Two Populations

Confidence Interval for Two Proportions

	Population 1	Population 2
Population proportion	p_1	p_2
Sample statistic	$\hat{p}_1$	$\hat{p}_2$
Difference in proportions	$p_1 - p_2$	
Estimator of difference in proportions	$\hat{p}_1 - \hat{p}_2$	

The estimated standard error is:

$$SE_{\text{est}} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The confidence interval for the difference between proportions is:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z* is the critical z-score that depends on the confidence level.

Conditions for Confidence Intervals

z* is derived from using the Central Limit Theorem to assume that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ has an approximate normal distribution.

1. **Random samples.** The samples are randomly selected from the appropriate populations.
2. **Independent samples.** The samples are independent of each other.
3. **Independent within samples.** The observations within each sample are independent of one another.
4. **Large Samples.** Both sample sizes must be large enough.
$$n_1\hat{p}_1 \geq 10, \qquad n_1(1-\hat{p}_1) \geq 10, \qquad n_2\hat{p}_2 \geq 10, \qquad n_2(1-\hat{p}_2) \geq 10$$
5. **Big Populations.** Each population is at least 10 times as big as its sample
$$N_1 \geq 10n_1, \qquad N_2 \geq 10n_2$$

Compare Proportions through CI

- If the confidence interval **contains 0**, then it is possible that the difference between the two population proportions equals 0, i.e., that there is no significant difference between the two population proportions.

- The confidence interval **does not contain 0**, which means that we can conclude that the two population proportions are not equal with some level of confidence.

It does not matter if you change the order of proportions when you calculate the difference. The confidence interval will change but the implication or the conclusion from the confidence interval will remain same.

Example: Medication for Hives

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions.

20 out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. 12 out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication.

Is there a significant difference in patient reactions of the two medications?

(using a 95% confidence level)

	Medication A	Medication B	Total
Still had hives after 30 mins	20	12	32
Hives disappeared after 30 mins	180	188	368
Total	200	200	400

Example – Confidence Interval