Chapter 15: Confidence Intervals for Proportions

Methods for making inferences about population parameters fall into one of two categories:

- **Estimation:** estimating or predicting the value of the parameter.
- **Hypothesis testing:** making a decision about the value of a parameter based upon some preconceived idea about what the value might be.

There are two types of estimates for population parameters:

- Point estimates
- Interval estimates

A **point estimate** of a population parameter is a single number calculated from sample data that is used to estimate the parameter.

Parameter	p	μ	σ
Point estimate	$\hat{p}$	$\bar{y}$	s

Recall: For a population of size N and a random sample of size n,

$p =$ proportion of the population that has a specific characteristic

$$= \frac{\text{number of successes in the population}}{N}$$

binary categorical variable.

$\hat{p} =$ proportion of the sample that has a specific characteristic

$$= \frac{\text{number of successes in the sample}}{n}$$

Example: An online club that offers monthly specials wants to test out a new item. It randomly selected 250 members from its list of over 9000 subscribers and asked them if they wanted to purchase the item. In this sample, 70 members decided to purchase the new item. Give a point estimate of the proportion of all club members that could be expected to purchase the item.

$$\hat{p} = \frac{70}{250} = 0.28$$

Note: Although a point estimate represents our "best guess" for the value of a population parameter, a point estimate on its own conveys no information about how close it is to the value of the parameter.

An **interval estimate** is an interval of numbers around a point estimate in which the value of the parameter is likely to lie.

A **confidence interval** for a population parameter is an interval estimate for the parameter that has an associated level of confidence. The confidence level provides information on how much "confidence" we can have in the **method** used to construct the interval estimate.

Recall: For random samples of size n drawn from a population with proportion p (binary categorical variable), if

- $np \geq 10$ and
- $n(1-p) \geq 10$,

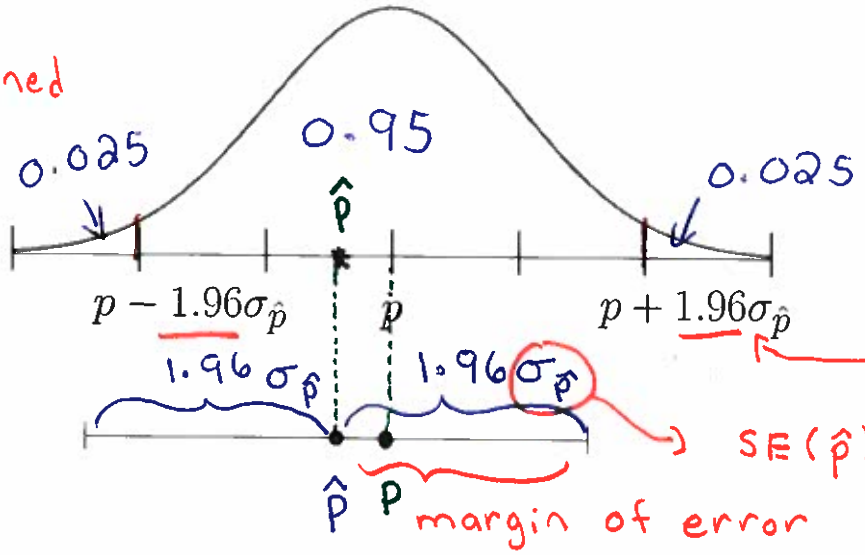
then the sampling distribution for $\hat{p}$ with sample size n is approximately normally distributed with

- mean $\mu_{\hat{p}} = p$
- standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

68 - 95 - 99.7 rule
↑ ~ $2\sigma_{\hat{p}}$ → 95.4% using z-table

Consider a 95% confidence level:

1.96 determined by confidence level



from z-table

→ SE($\hat{p}$)

margin of error → extent of interval on either side of $\hat{p}$

Problem: since p is unknown, the standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ is also unknown.

Solution: Estimate the value of $\sigma_{\hat{p}}$ using $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

When we estimate the standard deviation of a sampling distribution using statistics computed from sample data, the estimate is called a **standard error**.

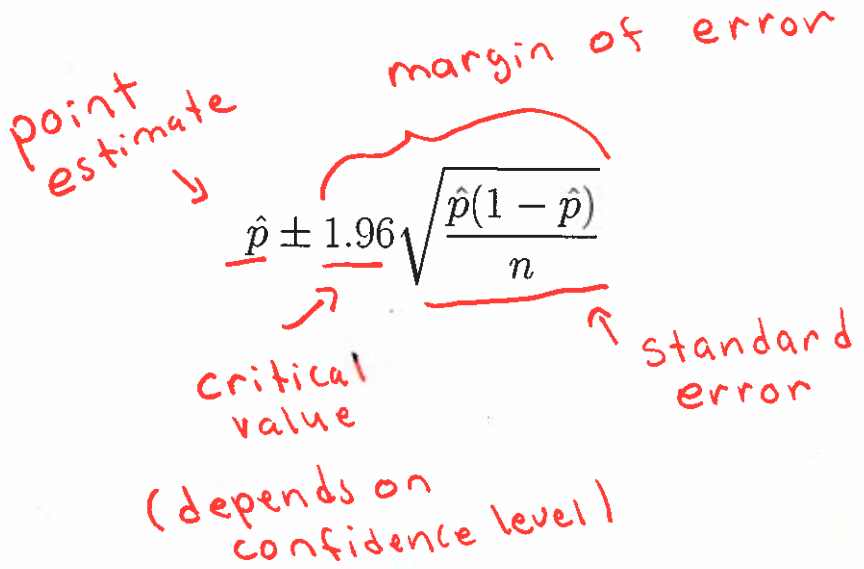
For the sampling distribution of $\hat{p}$, the standard error is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A large sample 95% confidence interval for a population proportion p has the form → $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

which we often write as



Example: A government agency wants to assess the prevailing unemployment rate in its country. They randomly selected 500 adults from their labour force and found that 41 of them were unemployed. Find a 95% confidence interval for the rate of unemployment in the country.

$\hat{p} = \frac{41}{500} = 0.082$ $\quad n = 500$ $\quad n\hat{p} = 41 \geq 10$

$1 - \hat{p} = \frac{459}{500} = 0.918$ $\quad n(1-\hat{p}) = 459 \geq 10$

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.082 \pm 1.96\sqrt{\frac{(0.082)(0.918)}{500}}$$

$$= 0.082 \pm 1.96(0.012227)$$

$$= 0.082 \pm 0.024$$

$$= (0.058, 0.106)$$

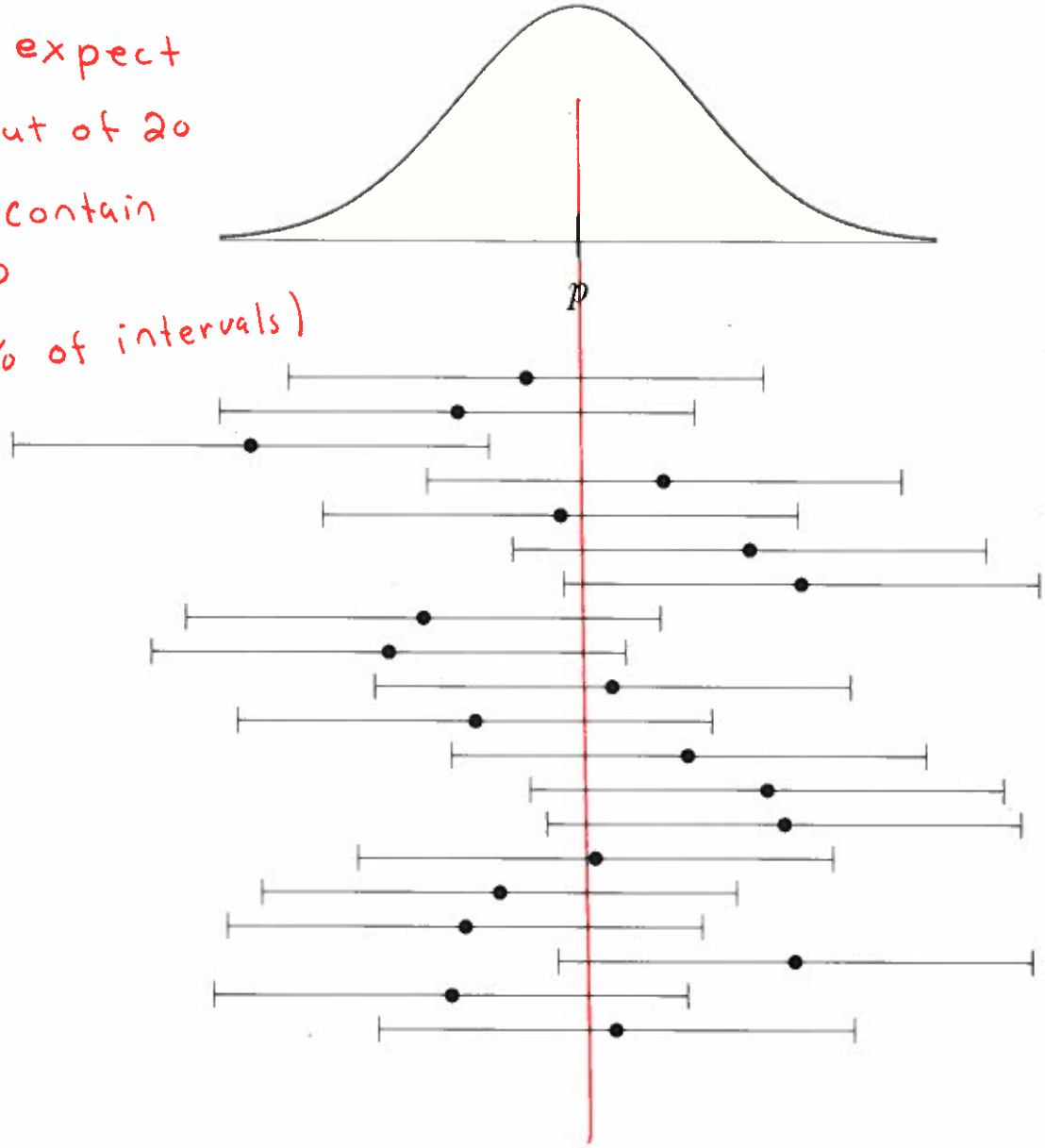
or (5.8%, 10.6%)

$\therefore$ we are 95% confident that

$p \in (0.058, 0.106)$

What does "95% confidence" mean?

- we expect 19 out of 20 to contain p (95% of intervals)



The 95% in "95% confidence" refers to the percentage of **all possible** samples of size n that result in an interval that contains p.

If we take samples of size n over and over again from the population and use each one separately to compute a 95% confidence interval, then about 95% of these intervals will capture p in the long run.

We **cannot** make chance/probability statements about particular intervals. A particular interval either includes p or it does not.

↳ never know

The confidence level refers to the **method** used to construct the interval rather than any particular interval.

↘ ↙

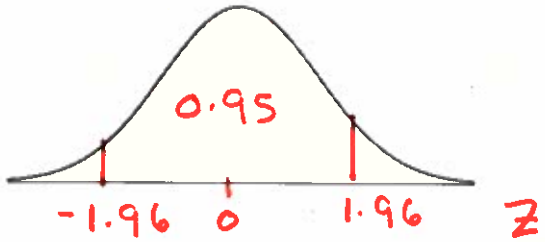
Example: For the unemployment rate example, we cannot say that there is a 95% chance that p is between 0.058 and 0.106. The interval $(0.058, 0.106)$ either includes p or it does not. Instead, we say "we are 95% confident that $(0.058, 0.106)$ contains p."

In the formula for a large sample 95% confidence interval for a population proportion p,

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- the value 1.96 is called the **critical value**.
- the value $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the **standard error**.
- the product $ME = 1.96 \times SE(\hat{p})$ is called the **margin of error**.

If we consider the standard normal distribution Z, the critical value 1.96 is the z-value that captures the central area of 0.95 under the z-curve. This value is only used for 95% confidence intervals.



We can obtain any other confidence level by replacing 1.96 with an appropriate z critical value.

↳ last line of t-table

Confidence Levels and Critical Values

$0 \le \alpha \le 1$

The critical value needed to find a $100(1-\alpha)\%$ confidence interval is the z-score, denoted $z_{\alpha/2}$ or z^*, that has an area of $\alpha/2$ to its right under the standard normal curve, that is,

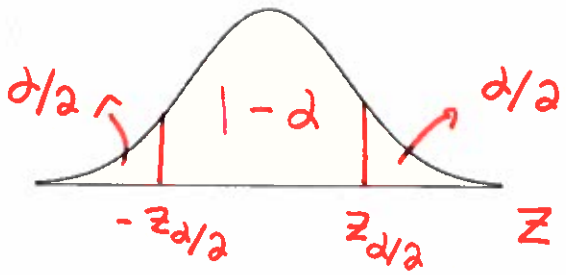
$$P(z > z_{\alpha/2}) = \alpha/2$$

or equivalently, it is the z-score that encloses an area of $1-\alpha$ between $-z_{\alpha/2}$ and $z_{\alpha/2}$, that is,

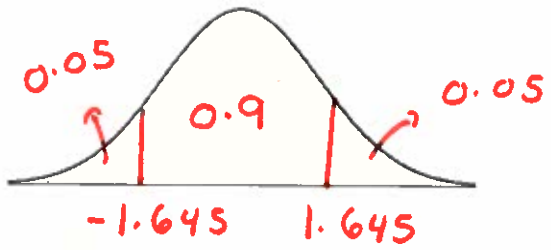
$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1-\alpha$$

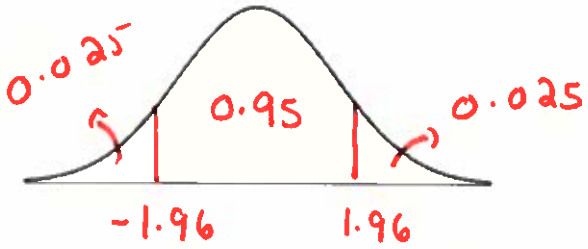
↑ confidence level → ↑ $z_{\alpha/2}$

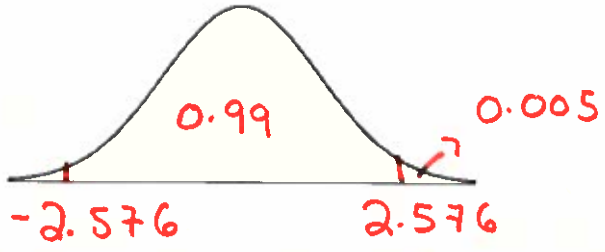
↓ confidence level → ↓ $z_{\alpha/2}$



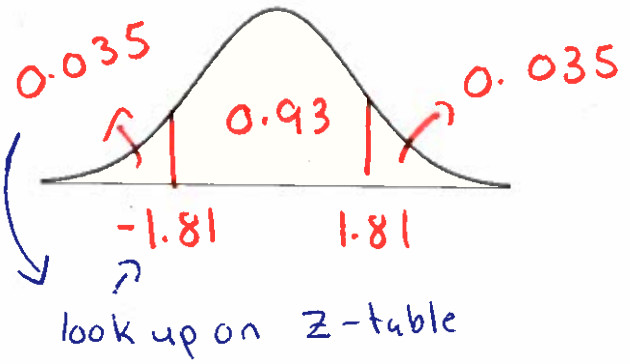
$100(1-\alpha)\%$ confidence level	critical value $z_{\alpha/2}$ or z^*
90%	1.645
95%	1.96
99%	2.576 ↑ 2.58, 2.575







Example: Find z^* required for a 93% confidence interval.



$1 - \alpha = 0.93$

$\alpha = 1 - 0.93 = 0.07$

$\alpha/2 = 0.035$

$z_{\alpha/2}$ or $z^* = 1.81$

Large Sample Confidence Intervals for Population Proportion

Assumptions and Conditions:

a) **Independence Assumption:** values in sample must be independent of each other.

i) **Randomization Condition:** data obtained from a simple random sample from the population or from a properly randomized experiment. → participants randomly selected.

ii) **10% Condition:** sample size should be less than 10% of population size. (without replacement)

b) **Sample Size Assumption:** sample size must be sufficiently large to be able to use the CLT

i) **Success/Failure Condition:** there should be at least ten successes and ten failures in the sample, that is,

$$n\hat{p} \geq 10 \qquad \text{and} \qquad n(1-\hat{p}) \geq 10$$

If these conditions are met, then a $100(1-\alpha)\%$ confidence interval for the population proportion p is

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

or equivalently,

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

- the value z^* is the **critical value** corresponding to the confidence level $100(1-\alpha)\%$.
- the value $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the **standard error**.
- the product $ME = z^* \times SE(\hat{p})$ is the **margin of error**.

Thus, a confidence interval for p has the form

$$\text{point estimate} \pm \text{margin of error}$$

$$= \text{point estimate} \pm \left(\text{critical value} \times \text{standard error of the estimate}\right)$$

The margin of error for a confidence interval:

- increases as the confidence level increases → wider interval
- decreases as the sample size increases → narrower interval

$n = 1000$

Example: In a survey of 1000 randomly selected Canadian workers, 586 said that they take their lunch to work with them.

$n\hat{p} = 586 \geq 10$

$n(1-\hat{p}) = 414 \geq 10$

$\hat{p} = \frac{586}{1000} = 0.586$

$1-\hat{p} = \frac{414}{1000} = 0.414$

a) Find a 95% confidence interval for the proportion of all Canadian workers who take their lunch to work with them. $z^* = 1.96$

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

$$= 0.586 \pm 1.96 \sqrt{\frac{(0.586)(0.414)}{1000}}$$

$$= (0.5555, 0.6165)$$

b) Find a 99% confidence interval for p. $z^* = 2.576$

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

$$= 0.586 \pm 2.576 \sqrt{\frac{(0.586)(0.414)}{1000}}$$

$$= (0.5459, 0.6261)$$ → wider than 95% CI