WEEK 02

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

2.1.1
POPULATION AND SAMPLE

POPULATION PARAMETER

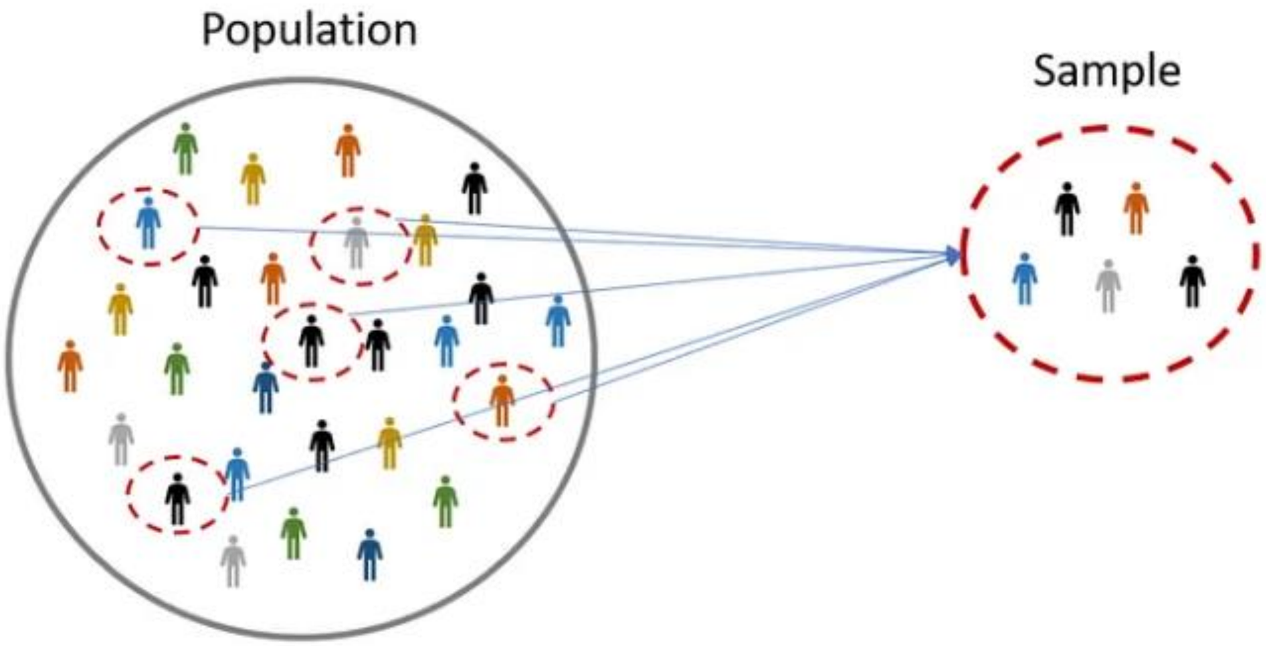
- A **population parameter** is a fixed, but often unknown, numerical value that describes a characteristic of an entire population.

POPULATION PARAMETER	FORMUALTION
Population Mean (μ)	$\mu = \frac{\sum x_i}{N}$
Population Proportion (p)	$p = \frac{X}{N}$
Population Variance (σ²)	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$
…	…

- But complete populations are difficult to collect data on, so we use **sample statistics** as **point estimates** for the unknown population parameters of interest

POPULATION & SAMPLE

- A **sample** is a **subset of individuals** or **data points** taken from a **larger population** to make inferences about the population as a whole.



SAMPLE STATISTICS (POINT ESTIMATES)

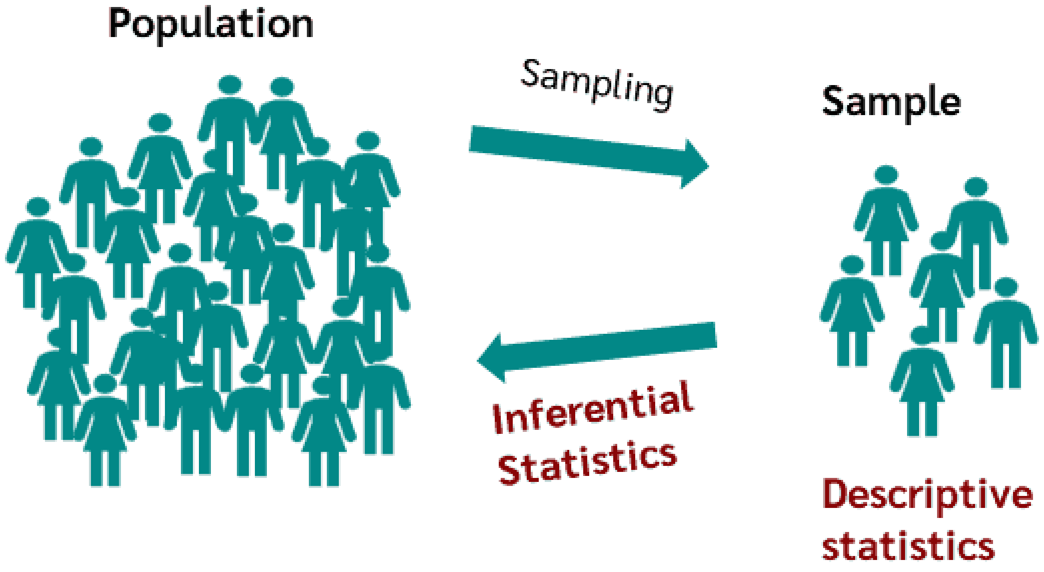
- Sample statistics: A value that describes your sample.
- Point estimates: The use of a sample statistic to estimate a population parameter

Sample Statistics	Formulation
Population Mean (μ)	$\bar{x} = \frac{\sum x_i}{n}$
Population Proportion (p)	$\hat{p} = \frac{X}{n}$
Population Variance (σ^2)	$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n}$
…	…

DESCRIPTIVE VERSUS INFERENTIAL STATISTICS

Descriptive statistics describe data.

Inferential statistics compute metrics from a sample to make inferences concerning parameters in a population. Inferential statistics is a subset of descriptive statistics.

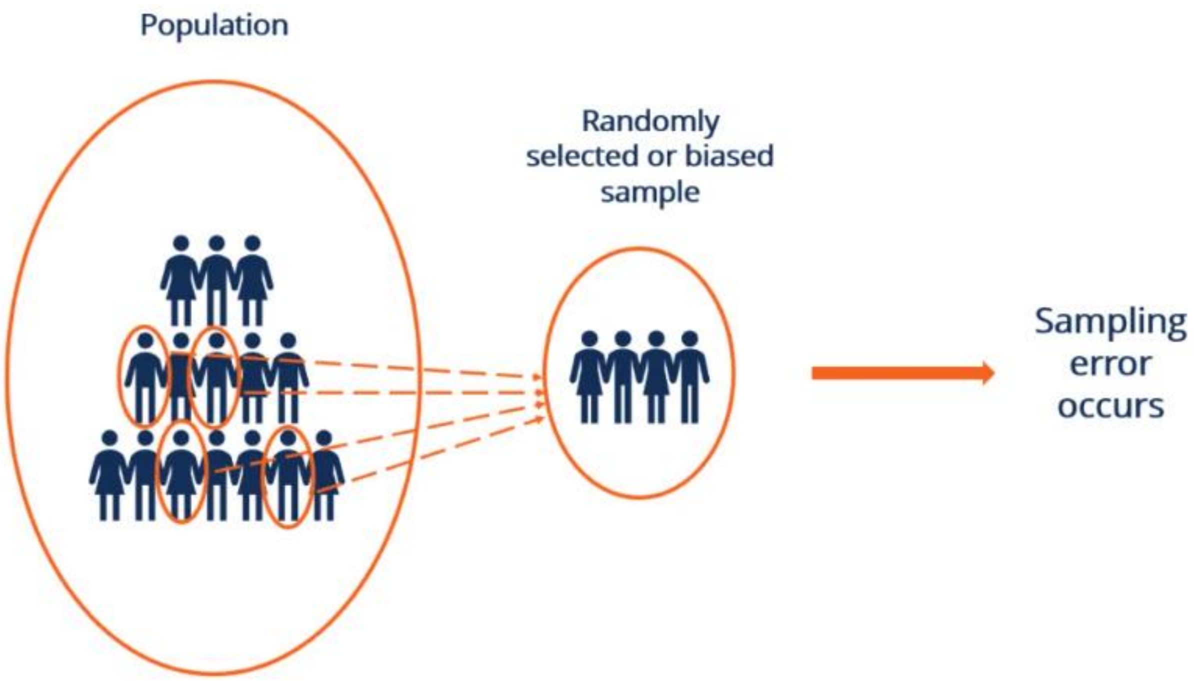


TIME TO THINK 🤔

If you have collected data from an entire population (a census), do you need to perform inferential statistics?

ERROR AND SAMPLING ERROR

- **Error** in the estimate = difference between population parameter and sample statistic
- Sampling error describes how much an estimate will tend to vary from one sample to the next.



SAMPLING DISTRIBUTION - PROPORTION

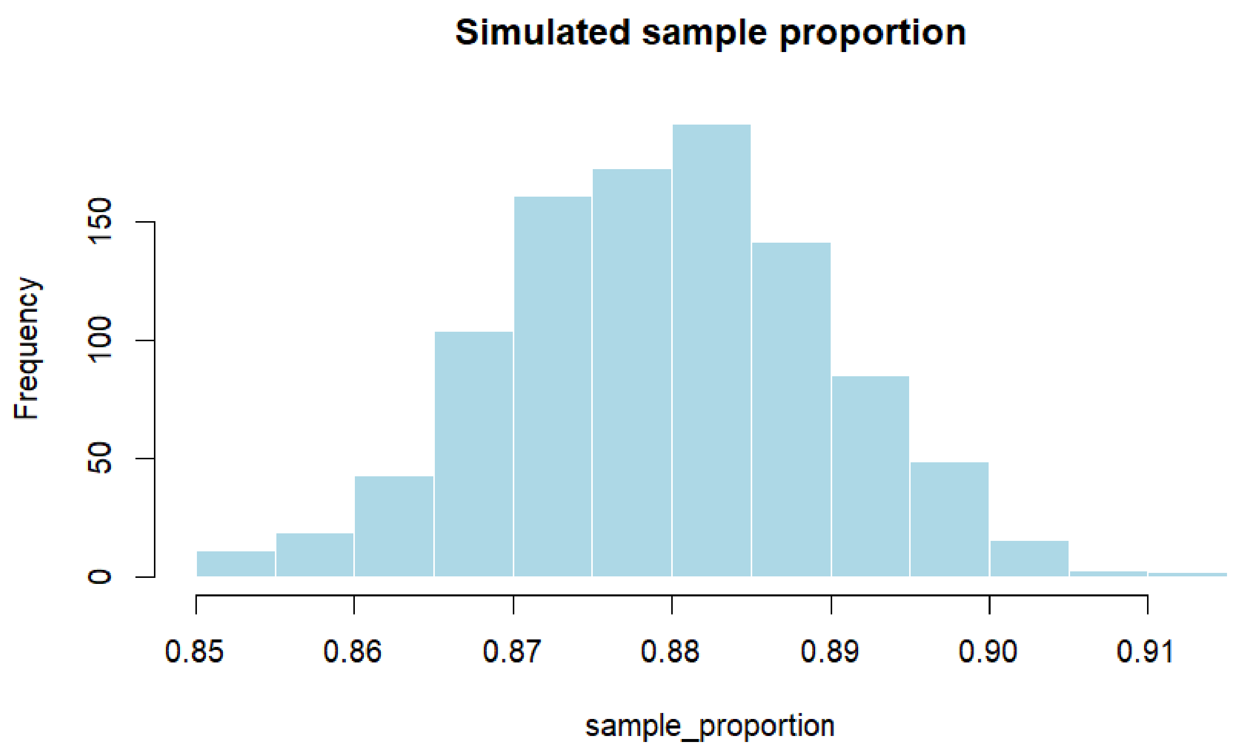
you might sample from the population and use your sample proportion as the best guess for the unknown population proportion.

Suppose you want to estimate the proportion of university students who regularly use public transportation, but you don’t have access to the entire student population. In this case, you could take a sample of students from different campuses and use the proportion of students in your sample who use public transportation as your best guess for the unknown proportion in the overall student population.

1) Sample, with replacement, 1000 students from the population, and record whether they use public transportation or not
2) Find the sample proportion.
3) Plot the distribution of the sample proportions obtained

SAMPLING DISTRIBUTION - PROPORTION

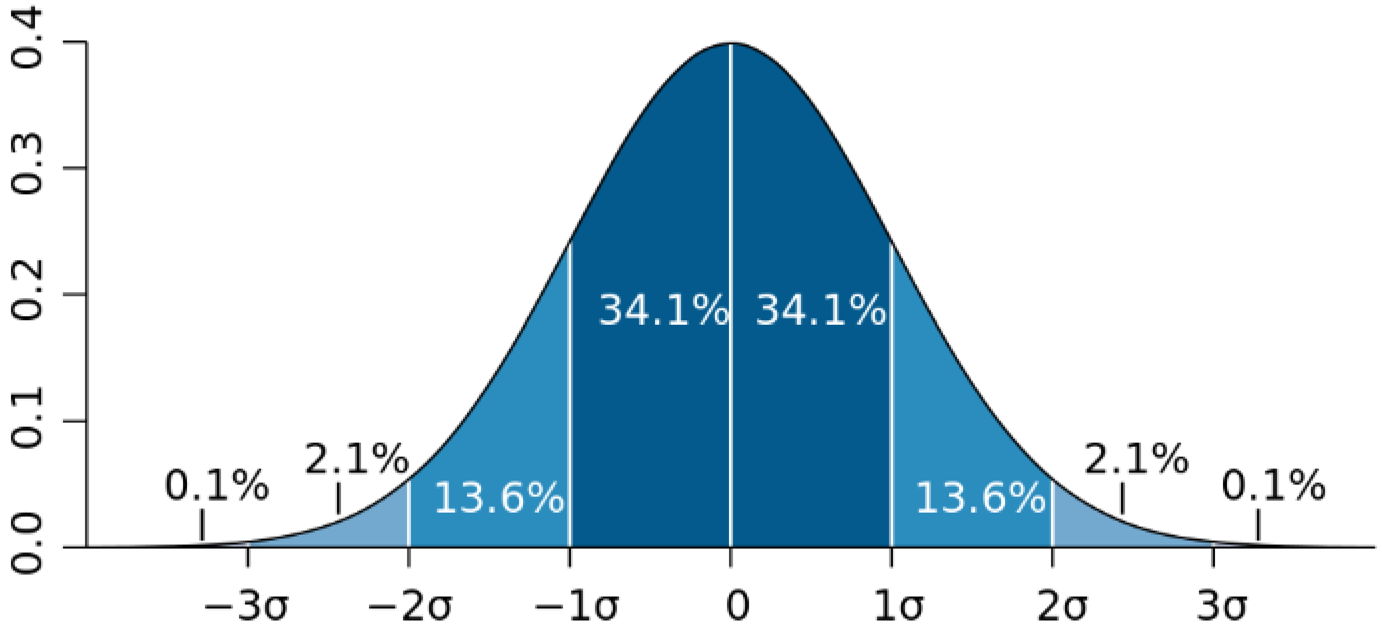
- A reasonable guess for the true population proportion is the center of this distribution, approximately 0.88.



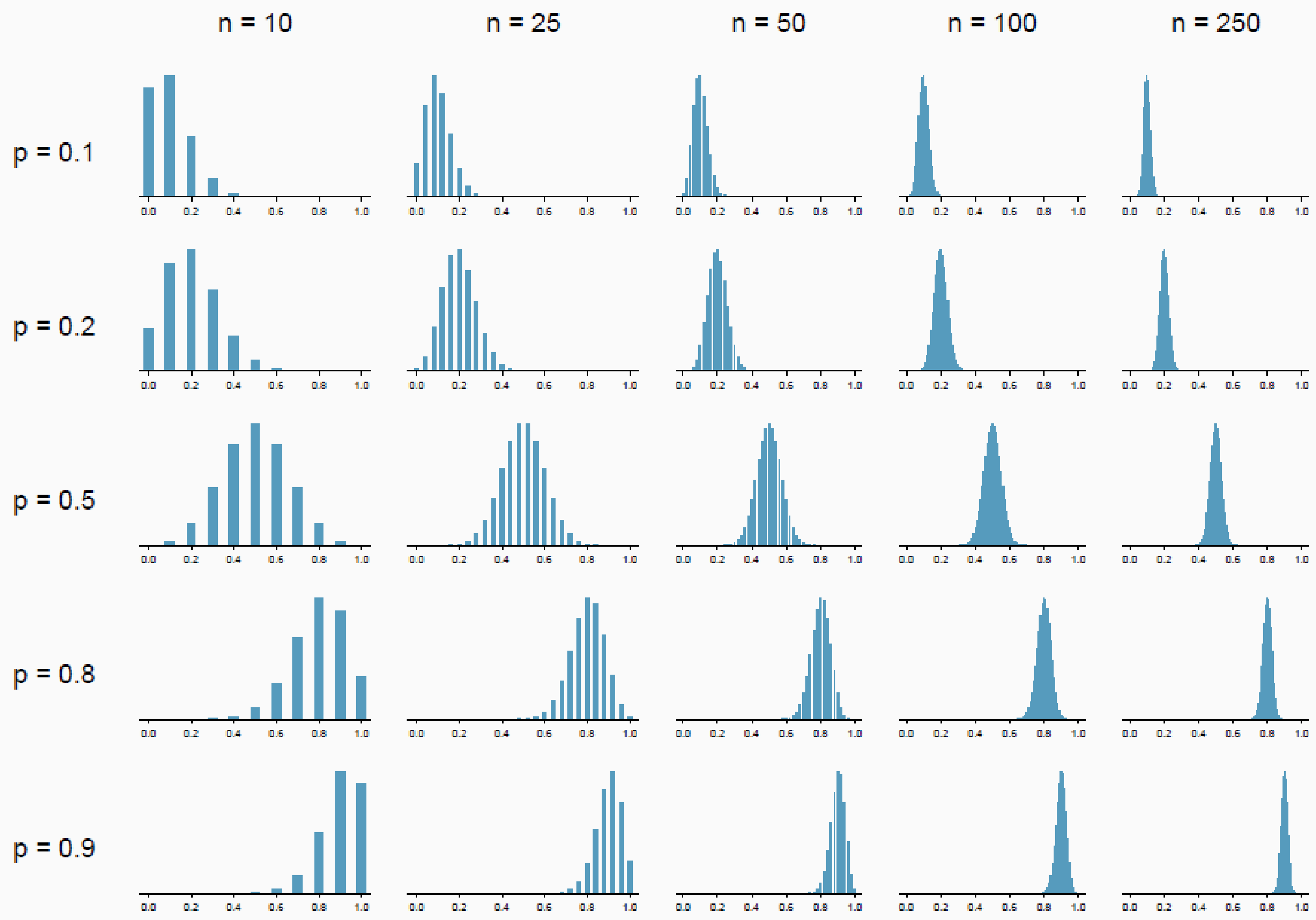
CENTRAL LIMIT THEOREM

- The sample size is sufficiently large, the sample proportion $\hat{P}$ will tend to follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$



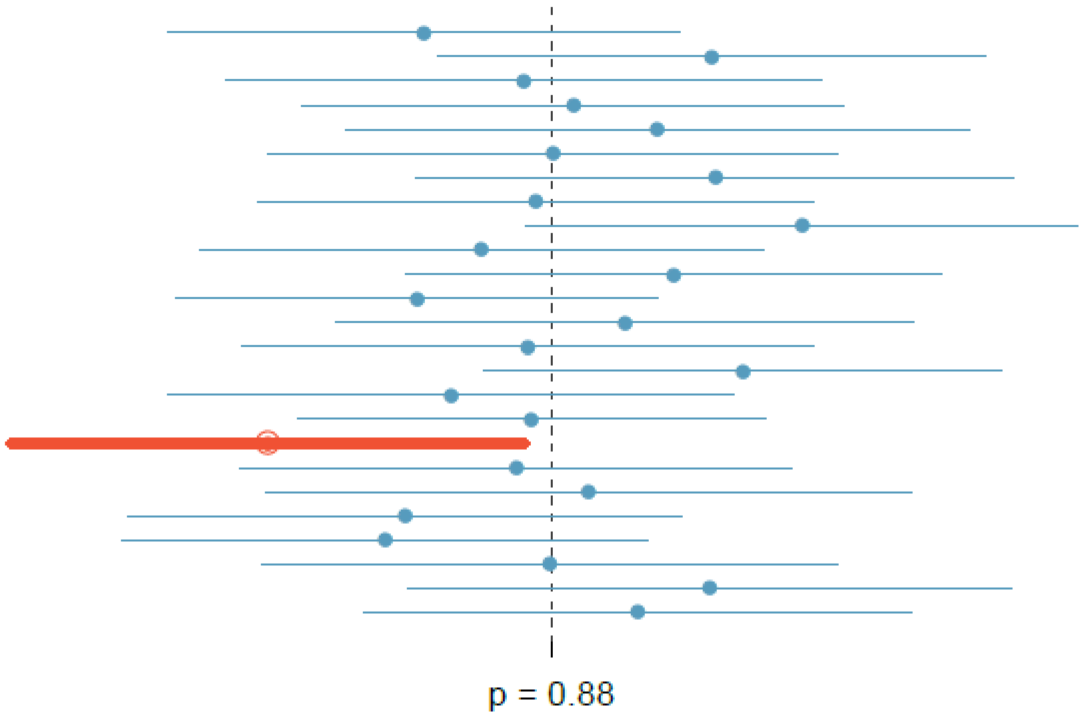
- What happens when np and/or $n(1-p) < 10$?



2.1.2 CONFIDENCE INTERVAL

CONFIDENCE INTERVAL OF PROPORTION

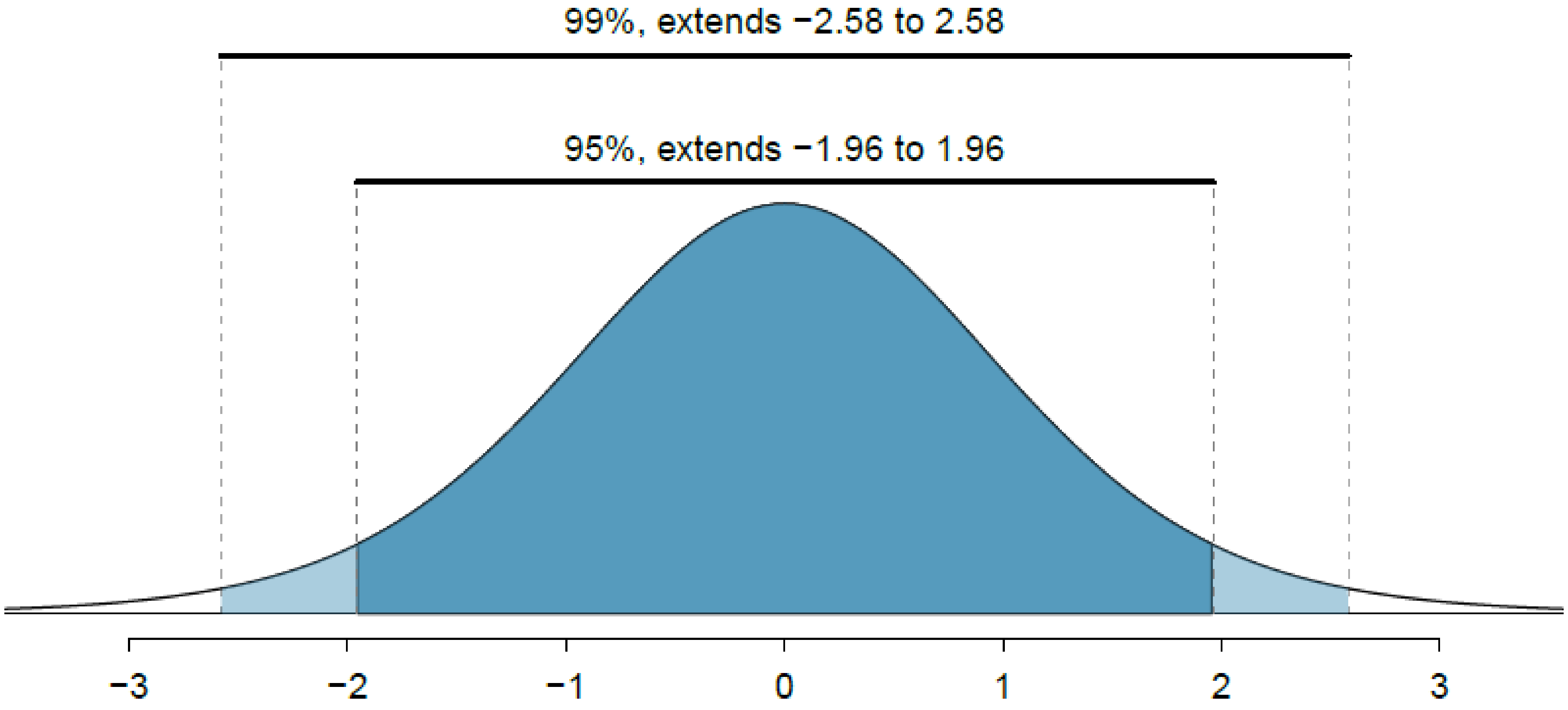
- A **confidence interval (CI)** is a **range of values** that is likely to contain the **true population parameter** (such as the mean or proportion) with a certain level of confidence



CONFIDENCE LEVEL IN CONFIDENCE INETERVAL

point estimate $\pm z \ * SE$

z correspond to the confidence level selected



2.1.2 HYPOTHESIS TESTING

NULL AND ALTERNATIVE HYPOTHESIS

- The **null hypothesis ($\boldsymbol{H_0}$)** is the default assumption that there is no effect, no difference, or no relationship in the population.

- The **alternative hypothesis ($\boldsymbol{H_A}$)** is the statement you want to prove. It suggests that there is a real effect, difference, or relationship in the population.

$$H_0: p = 0.5$$

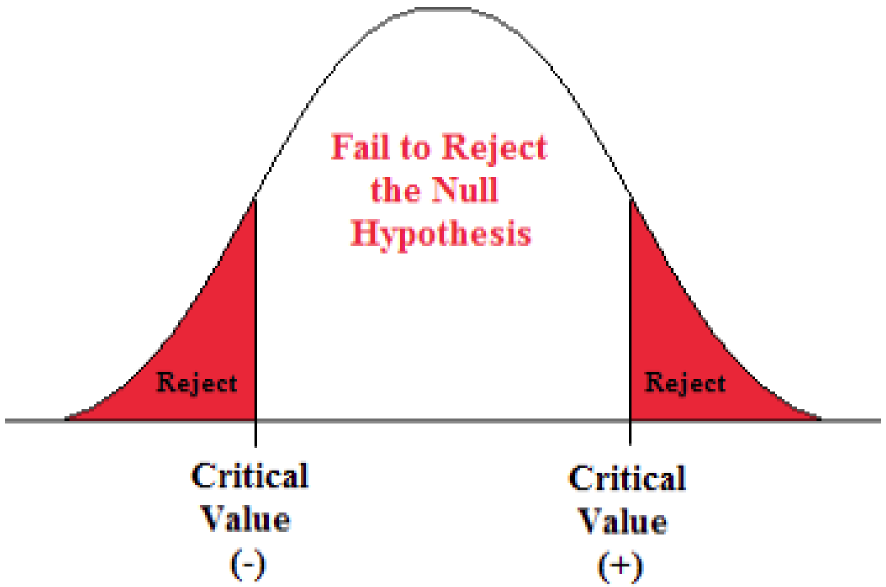
$$H_1: p \neq 0.5$$

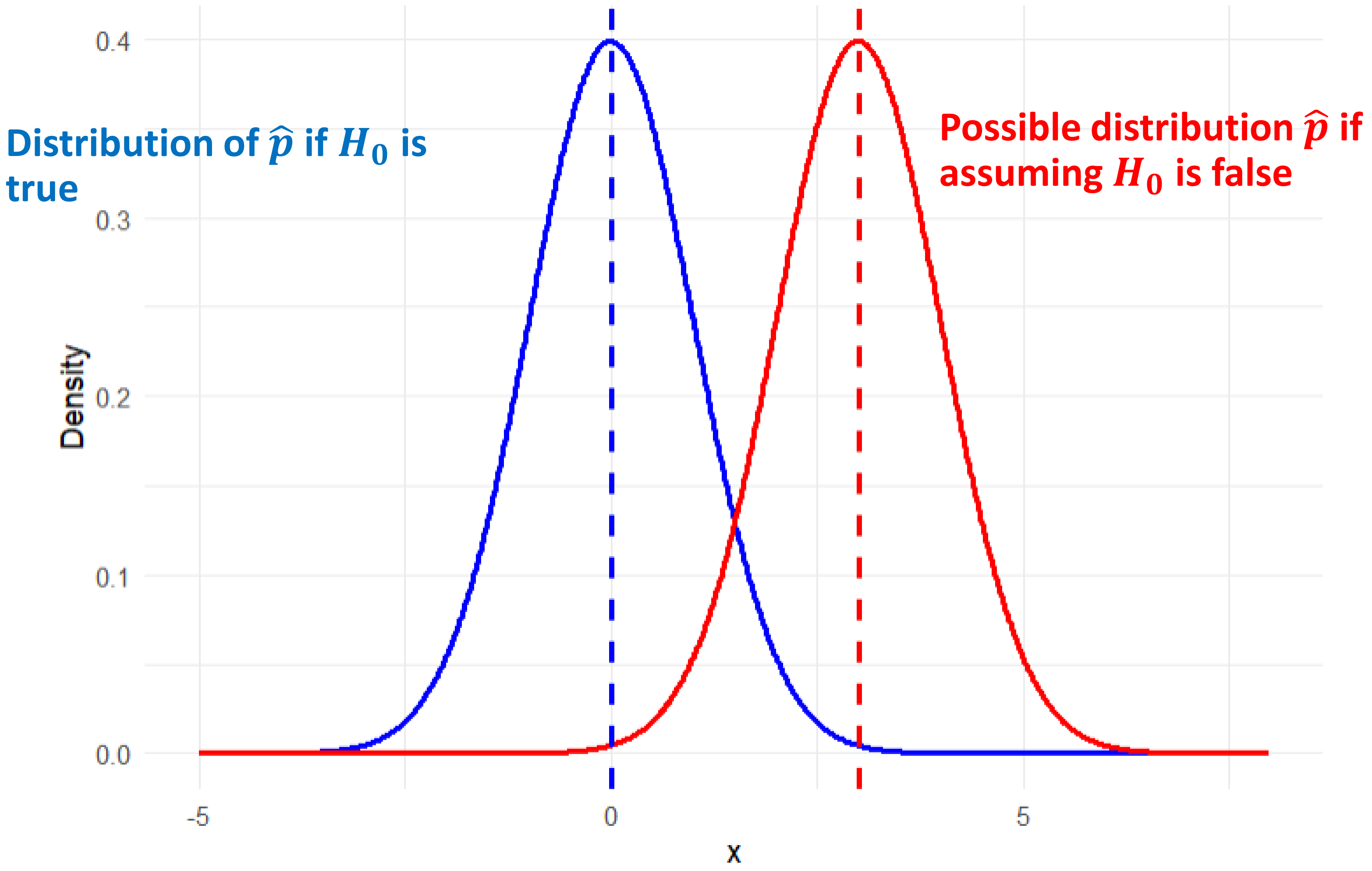
P-VALUE

- The **p-value approach** is the likelihood or probability that a sample will result in a statistic such as the one obtained if the null hypothesis is true.

If p-value ≤ α, reject the null hypothesis.

If p-value > α, fail to reject the null hypothesis.

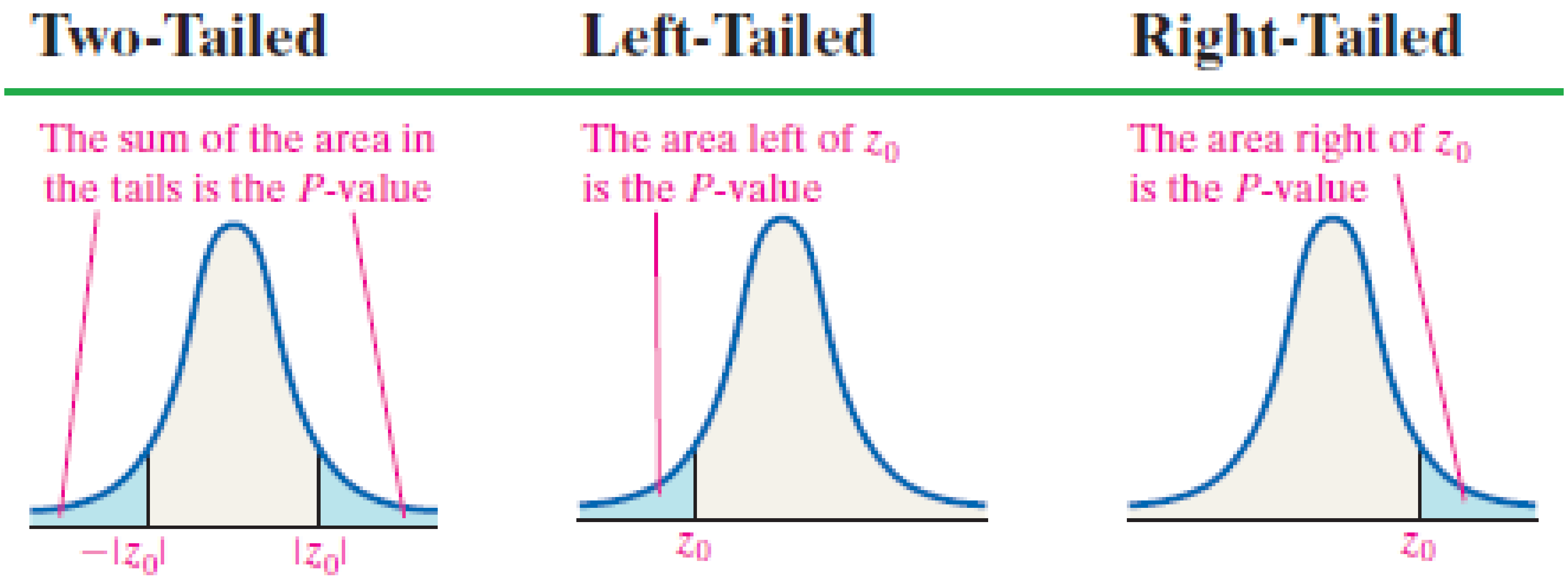




Distribution of $\widehat{p}$ if H_0 is true
Possible distribution $\widehat{p}$ if assuming H_0 is false
0.4
0.3
0.2
0.1
0.0
Density
-5
0
5
x

CALCULATION OF P-VALUE

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$



WEEK 02

LAB SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025