WEEK 02

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

2.1
POPULATION AND SAMPLE

POPULATION PARAMETER

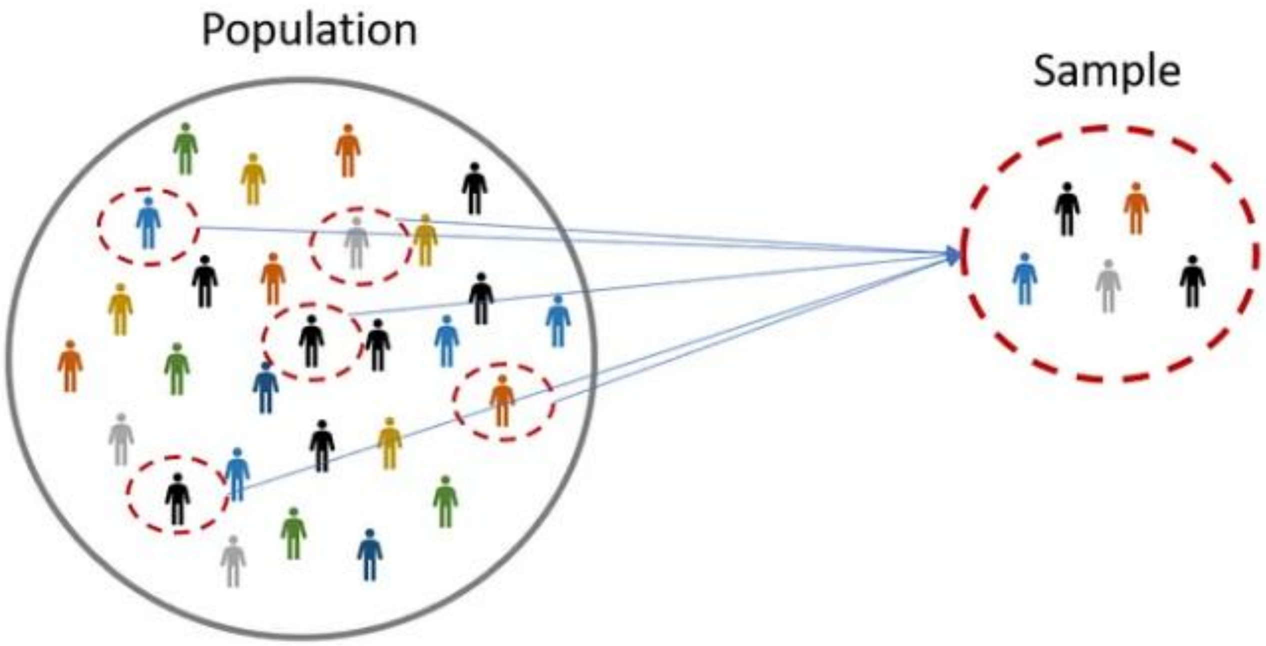
- A **population parameter** is a fixed, but often unknown, numerical value that describes a characteristic of an entire population.

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.



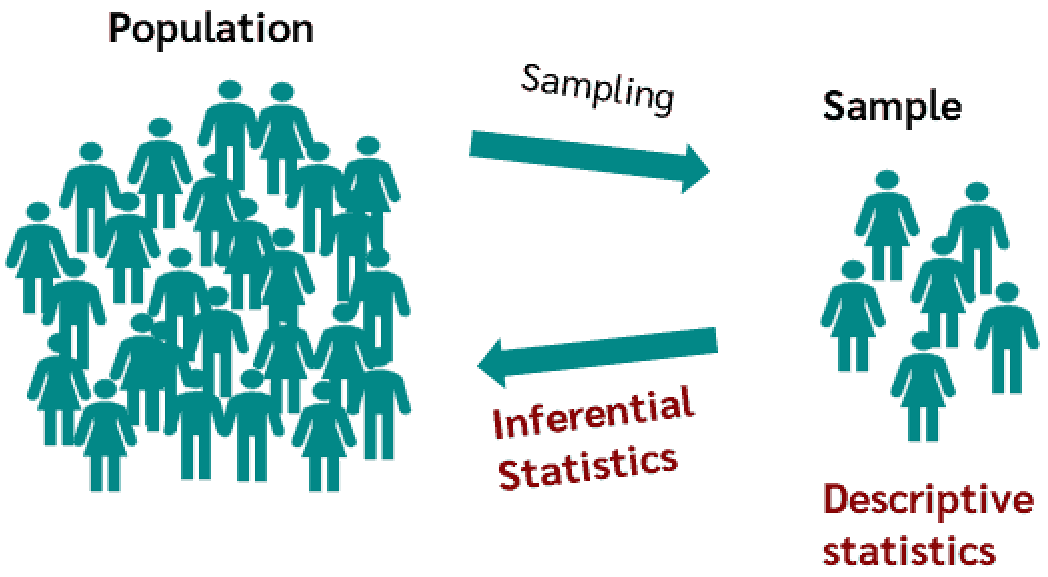
SAMPLE STATISTICS (POINT ESTIMATES)

- Sample statistics: A value that describes your sample.
- Point estimates: A single value used to approximate a population parameter

Sample Statistics	Formulation
Sample Mean ($\bar{x}$)	$\bar{x} = \frac{\sum x_i}{n}$
Sample Proportion ($\hat{p}$)	$\hat{p} = \frac{X}{n}$
Sample Variance (s^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.



DESCRIPTIVE? OR INFERENTIAL?

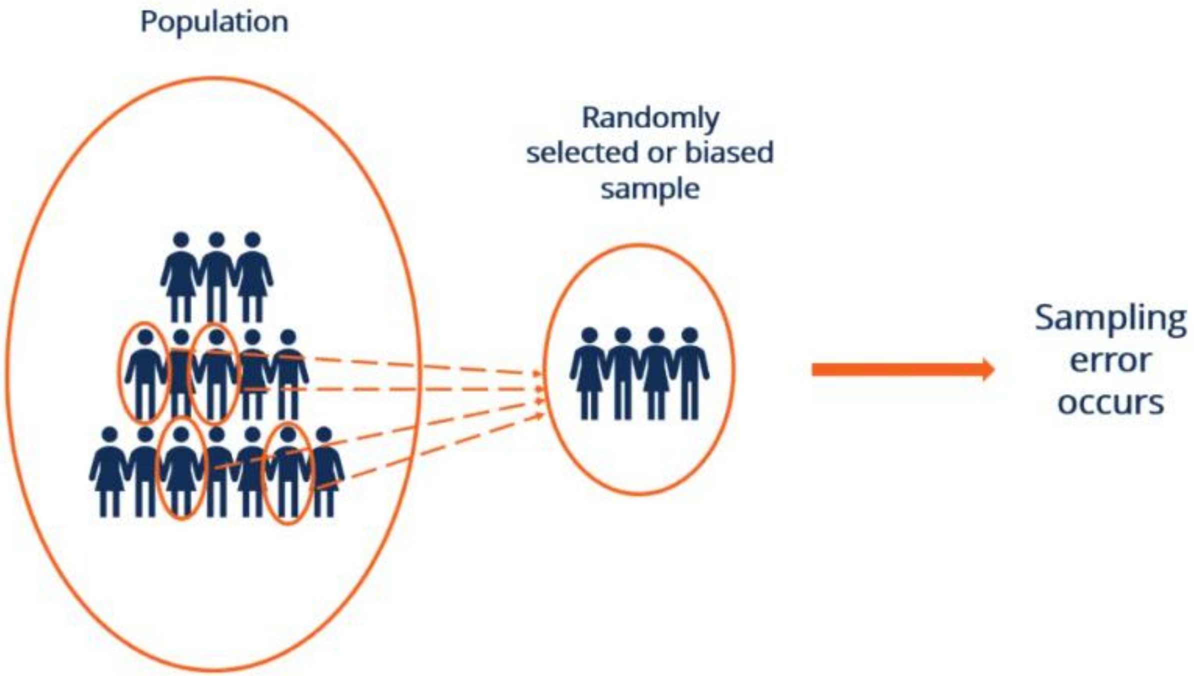
- Mean
- Standard Deviation
- Hypothesis Testing
- Variance
- Confidence Interval

TIME TO THINK 🤔

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

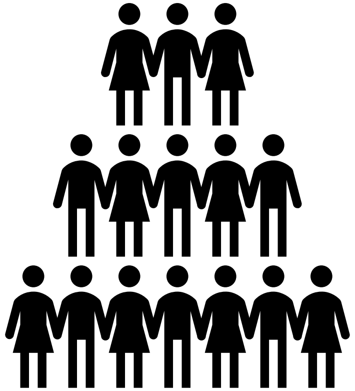
ERROR AND SAMPLING ERROR

- **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.

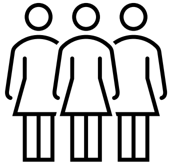


EXPERIMENT: VARIABILITY OF THE SAMPLE

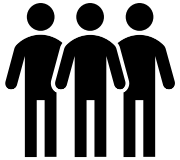
p: 88% support energy expansion



$\hat{p}_1$: ? % support energy expansion



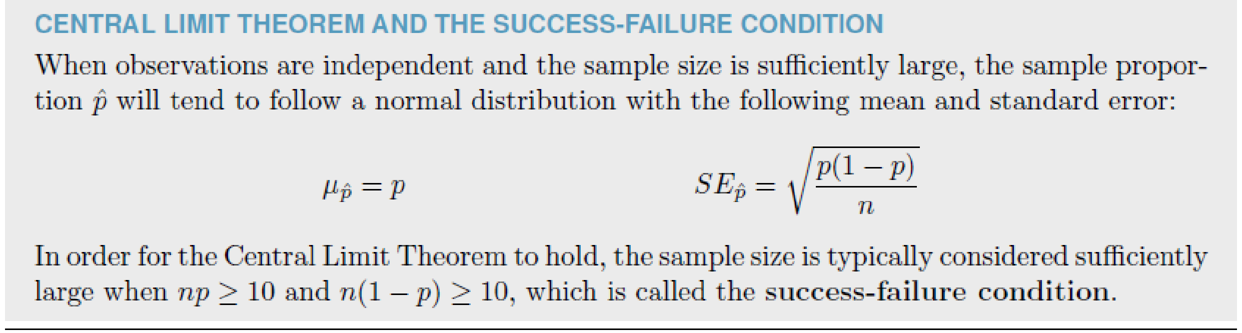
$\hat{p}_2$: ? % support energy expansion



…

$\hat{p}_{1000}$: ? % support energy expansion

CENTRAL LIMIT THEOREM

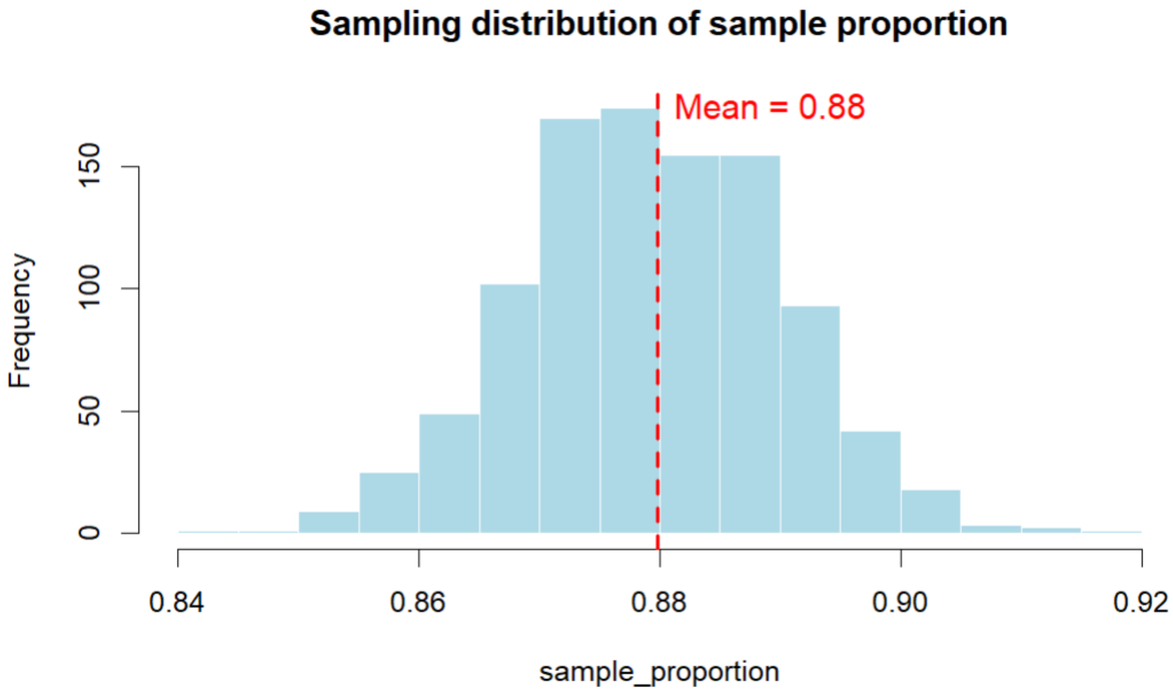


CENTRAL LIMIT THEOREM AND THE SUCCESS-FAILURE CONDITION

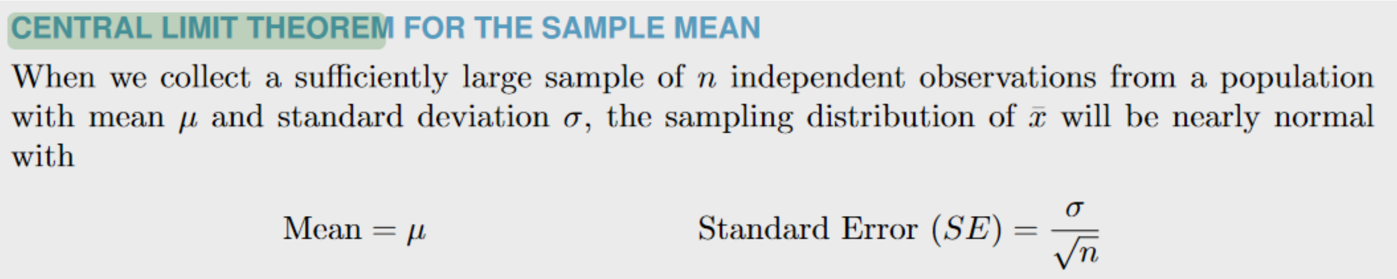
When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

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

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.



CENTRAL LIMIT THEOREM ACCORDING TO OPENINTROSTATISTICS VERSION4



CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

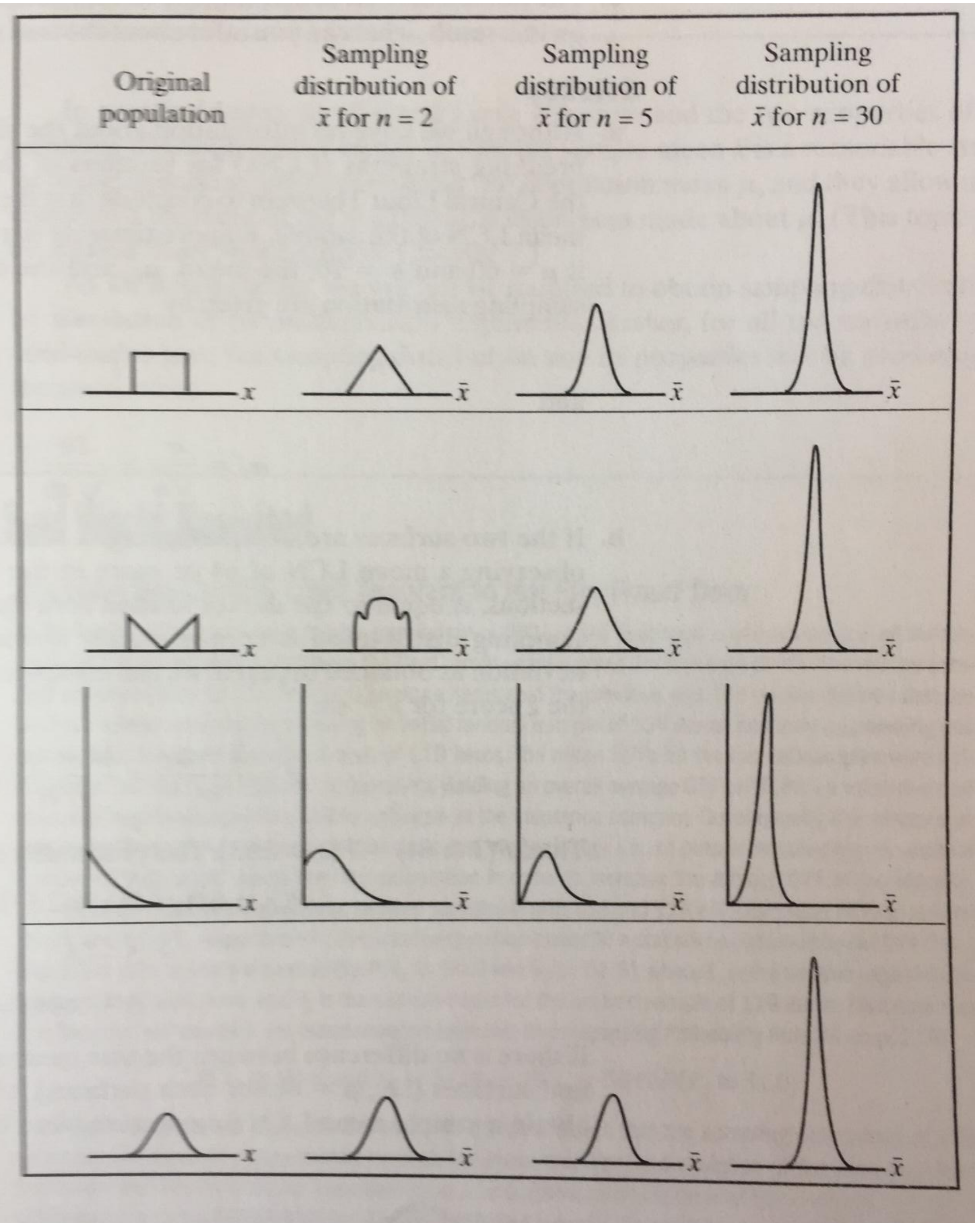
When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$

The distribution refers to a histogram of x

The sampling distribution refers to a histogram of $\bar{x}$

Standard Error is the standard deviation of $\bar{x}$

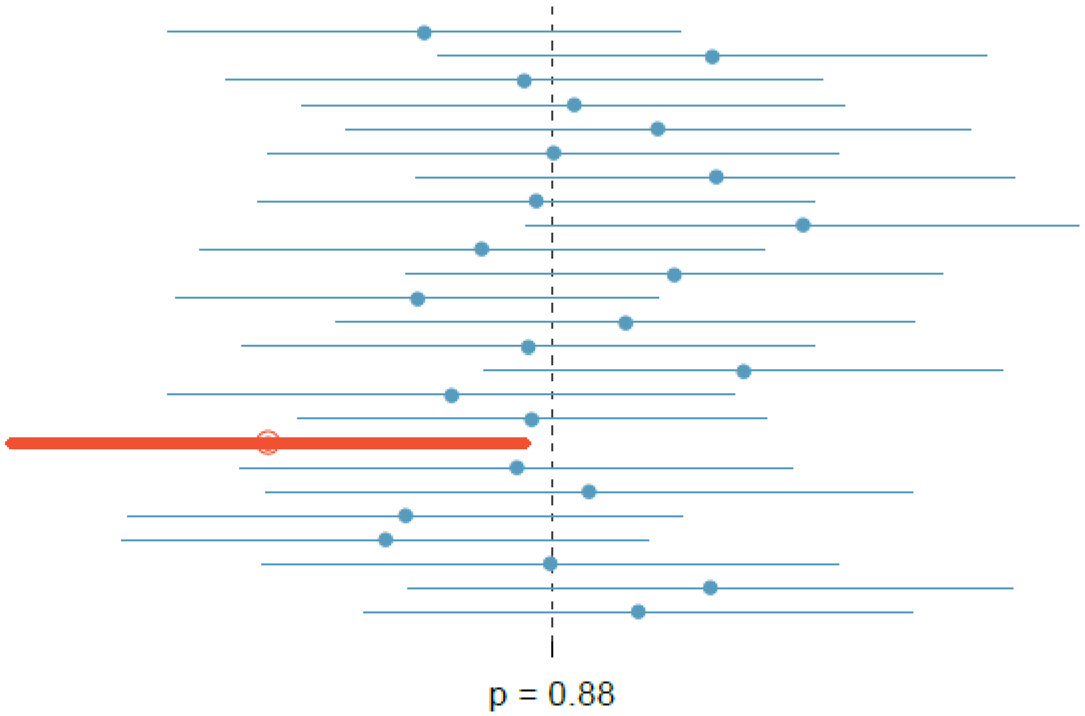

Original population
Sampling distribution of $\bar{x}$ for $n = 2$
Sampling distribution of $\bar{x}$ for $n = 5$
Sampling distribution of $\bar{x}$ for $n = 30$
x
$\bar{x}$

CENTRAL LIMIT THEOREM APPLICATION

2.2 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



WHAT DOES 95% CONFIDENT MEAN?

Some people tell Pontius that a (1-alpha)% confidence interval means that we are (1-alpha)% *confident* that the interval contains the population parameter. Pontius asks "What is the meaning of *confident*?"

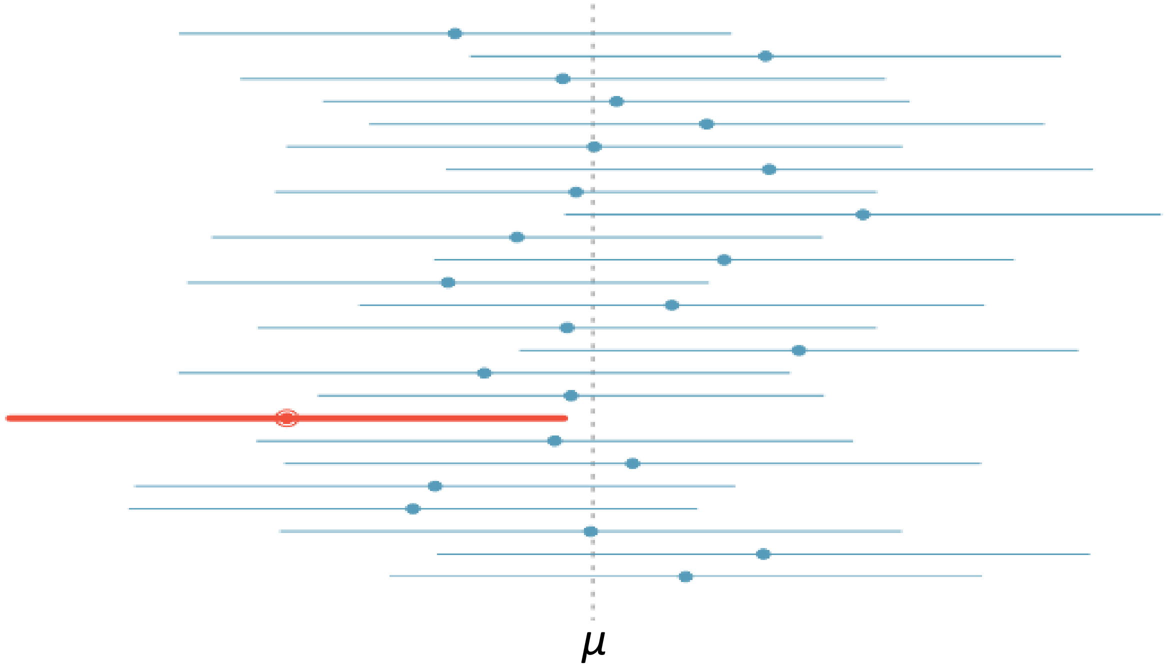
Suppose we repeated an experiment several times and built a 95% confidence interval from each experiment using the expression:

point estimate ± $Z_{\alpha/2}$ *(Standard Error)*

Then we expect about 95% of those intervals would contain the population parameter.

The figure shows 25 confidence intervals, each of which derives from an experiment that takes several random samples. Each experiment examines different data because of random sampling.

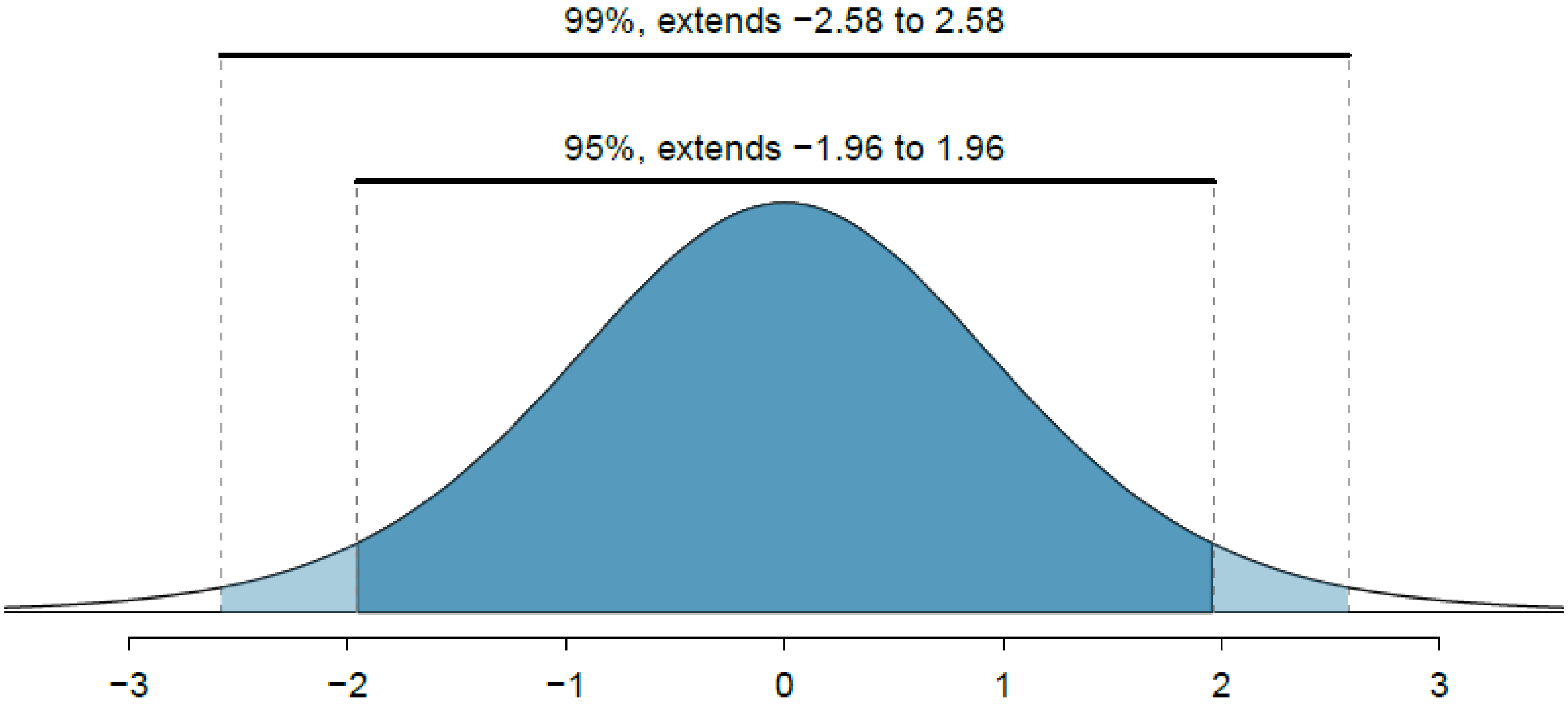
The dashed line shows the population parameter. We see that 24 of the 25 resulting confidence intervals contain the population parameter while one does not.



CONFIDENCE LEVEL IN CONFIDENCE INETERVAL

CI = point estimate $\pm z_{\frac{\alpha}{2}} * se =$ point estimate $\pm z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}-(1-\hat{p})}{n}}$

z correspond to the confidence level selected



2.3 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.

$$H_0: p = 0.5$$

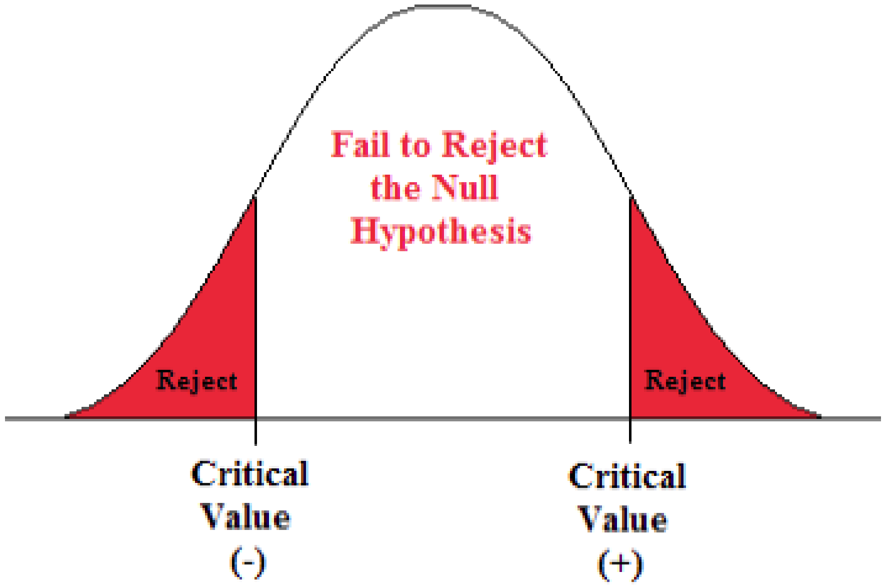
$$H_1: p \neq 0.5$$

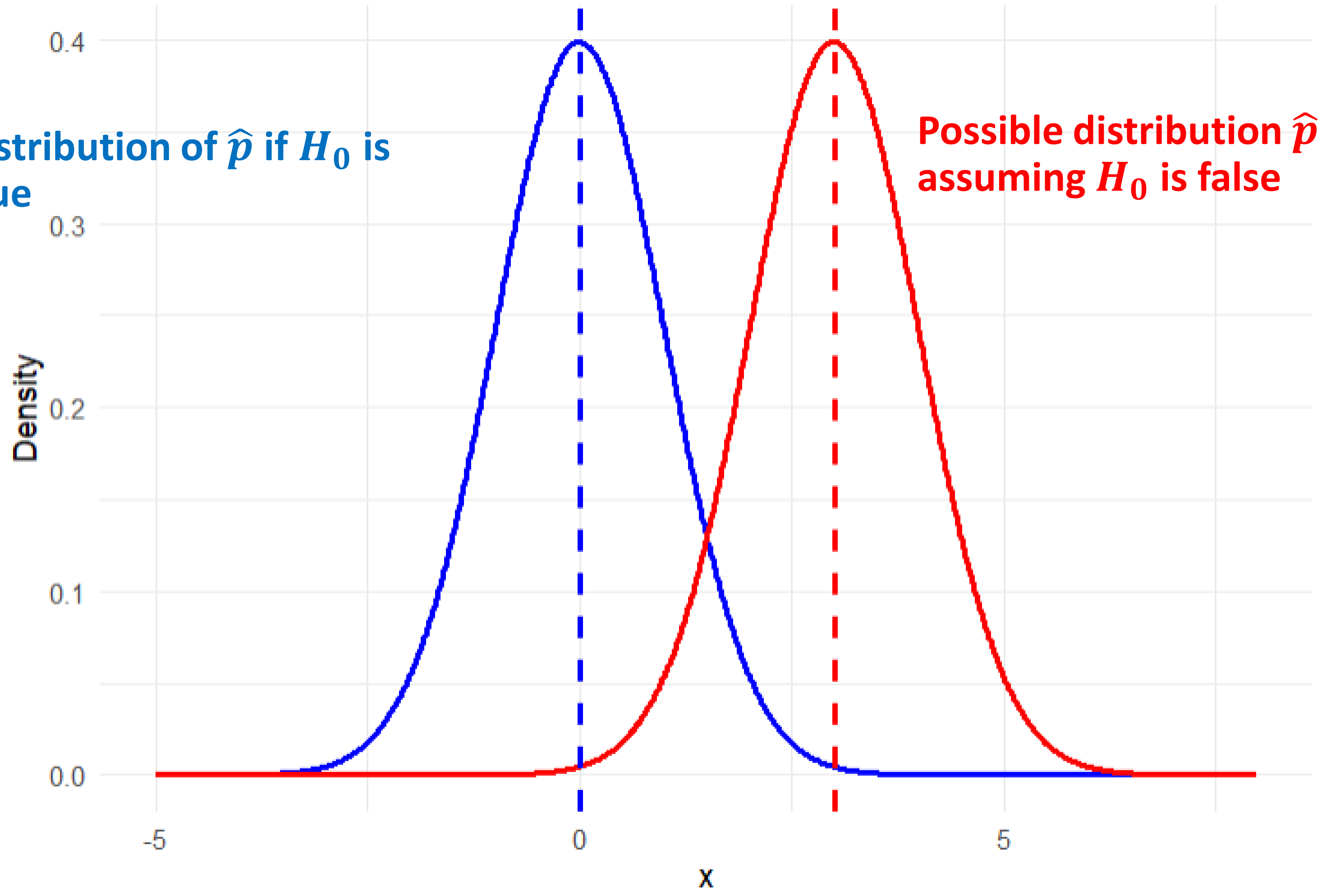
P-VALUE

- 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.



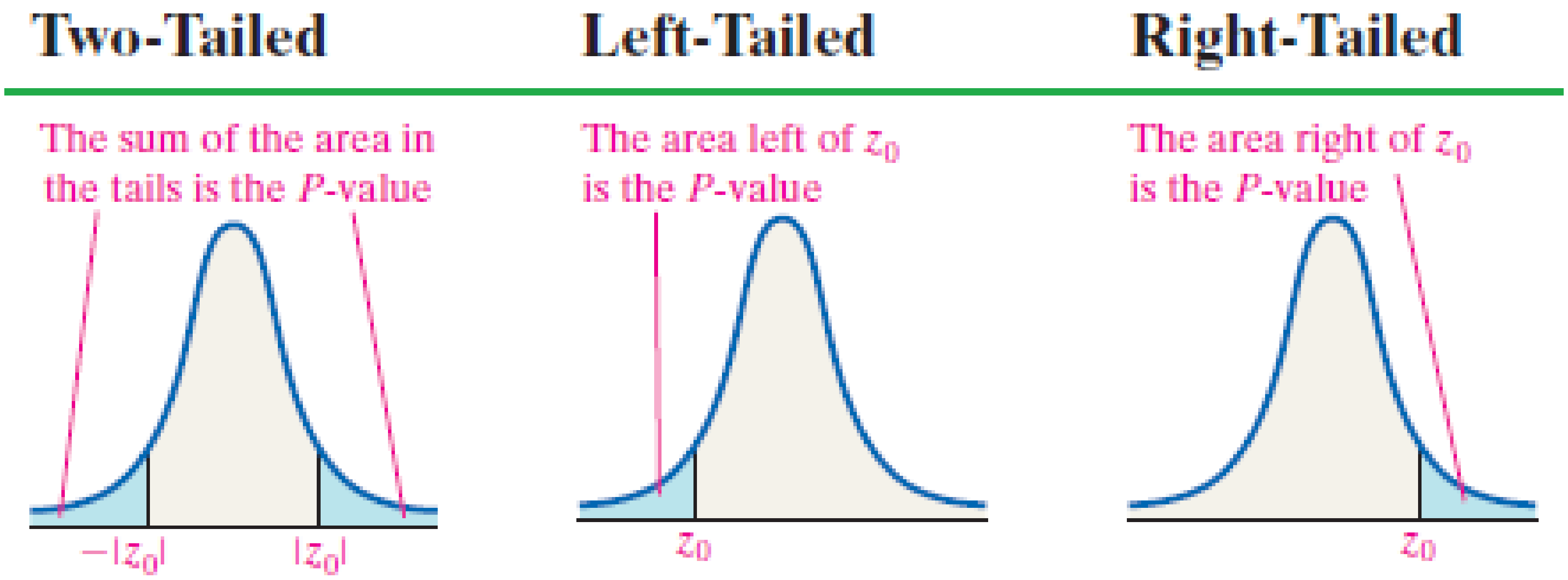


Distribution of $\widehat{p}$ if H_0 is true

Possible distribution $\widehat{p}$ if assuming H_0 is false

CALCULATION OF P-VALUE

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



Two-Tailed

The sum of the area in the tails is the *P*-value

Left-Tailed

The area left of z_0 is the *P*-value

Right-Tailed

The area right of z_0 is the *P*-value

WEEK 02

DEMO SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025