ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

INTRODUCTION TO INFERENCE

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Statistical Inference

- The process by which we estimate parameters of interest from the data, and quantify the uncertainty in our estimates.

- Example: Estimating a ***population proportion*** using a ***sample proportion***

- Tools we will introduce and use:

- **Point estimates**

- **Confidence Intervals** = range of plausible values for the true population value

- **Hypothesis Tests =** method to evaluate claims about the population

Point Estimates and Sources of Error

- Say you poll 1000 people on their voting intentions and 43% say they support candidate A.

- 43% would be a **point estimate, denoted $\hat{p}$,** of the voting intentions of the entire population (our parameter of interest), denoted p.

- $\hat{p}$ is usually different from p, and the difference is called the error.

- Sources of error:
  - **Sampling uncertainty/error**: due to variability between different groups of 1000 people in our example, usually related to **sample size**, n
  - **Bias:** due to systematic error (e.g., you collect your samples from Candidate A's family and friends, poorly worded questions in poll)

Point Estimates and Sources of Error

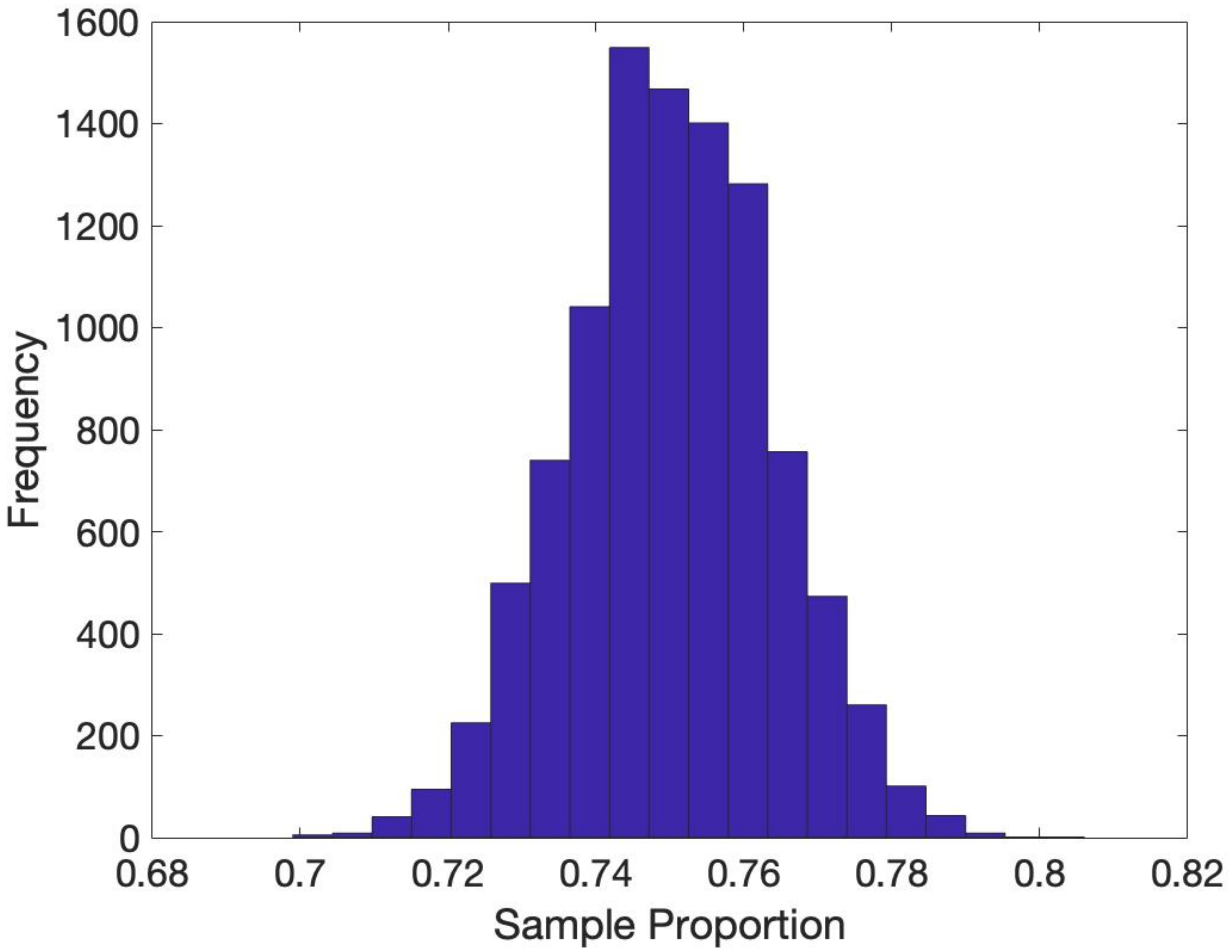
- **Thought experiment**: How does $\hat{p}$ behave when p is, say, 75%?
- Recall that $\hat{p}$ is an estimate you get from sampling a subset of the population, and p is the true population value (usually not available to us).
- To get a sense of how $\hat{p}$ behaves, we randomly sample 1000 "people", and find that

$$\hat{p}_1 = \frac{763}{1000} = 0.763 \implies \text{error} = 0.763 - 0.75 = 0.013$$

- We sample another 1000 people and get $\hat{p}_2 = 0.741$
- We sample another 1000 people and get $\hat{p}_3 = 0.749$

Sampling Distribution

- The **center** of the sampling distribution, denoted $\bar{x}_{\hat{p}}$ is 0.7501 (very close to $p = 0.75$)
- The "variability" of the point estimates, is denoted the **standard error,** denoted $SE_{\hat{p}}$. It is the standard deviation of $\hat{p}$. In our example, $SE_{\hat{p}} = 0.0136$.
- The distribution *looks like a normal distribution*
- ***Remember: We never observe the sampling distribution. This was a thought experiment.***

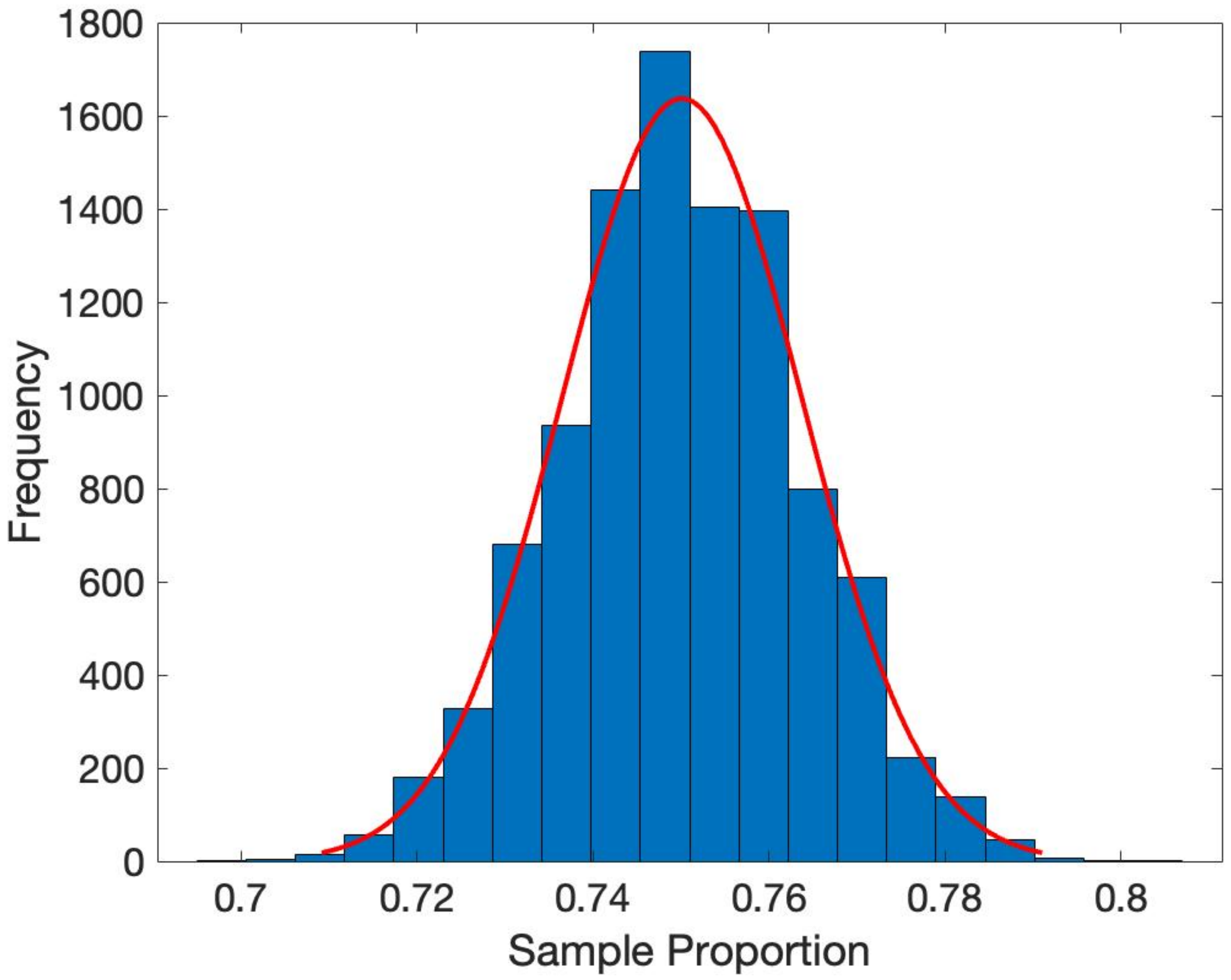


Central Limit Theorem

- Informally: The **Central Limit Theorem (CLT)** tells us that when the ***observations are independent*** and n is large, $\hat{p}$ will follow a normal distribution with

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

- **The Success/Failure Condition** is a rule of thumb: For the CLT to hold, we **need** $np \geq 10$ and $n(1-p) \geq 10$
- Exercise: Does the success/failure condition hold with our previous example?



Central Limit Theorem

- **Example (part 1)**: Compute the mean and standard error of $\hat{p}$ when $p = 0.75$ and $n = 1000$, using the CLT

- **Answer:** $\mu_{\hat{p}} = 0.75$ and $SE_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.75 \times 0.25}{1000}} \approx 0.0137$

Central Limit Theorem

- **Example (part 2):** How frequently does $\hat{p}$ falls within 2% of the true population value, $p = 0.75$?
  - The question is asking us for $\mathbb{P}(|\hat{p} - p| \leq 0.02) = \mathbb{P}(p - 0.02 \leq \hat{p} \leq p + 0.02)$.
  - We know (part 1, and the CLT) that $\hat{p}$ is approximately normal with: $\mu_{\hat{p}} = 0.75$ and $SE_{\hat{p}} \approx 0.0137$.
- We can use what we know about normal distributions!
  - Either we numerically calculate: $\mathbb{P}(0.73 \leq \hat{p} \leq 0.75) = \dfrac{1}{\sqrt{2\pi} \cdot SE_{\hat{p}}} \displaystyle\int_{0.73}^{0.77} e^{-\frac{(x-0.75)^2}{2 \cdot (SE_{\hat{p}})^2}} dx$
  - Or use Z-scores/tables: $Z_{0.73} = \dfrac{0.73 - 0.75}{0.0137} \approx -1.4599$ and $Z_{0.77} = \dfrac{0.77 - 0.75}{0.0137} \approx 1.4599$
- In either case, we get: $\mathbb{P}(0.73 \leq \hat{p} \leq 0.77) \approx 0.8599$

Central Limit Theorem

- **Interpretation:** If we run the experiment a very large number of times — each time taking n=1000 samples from a large population with p = 0.75, then
  - Approximately 86% of the time, we will get an estimate $\hat{p}$ that falls within 0.02 of the true value of p.
- **To summarize:** We approximated the sampling distribution, with a normal distribution (using the CLT), and used the normal distribution to estimate how frequently $\hat{p}$ that falls within 0.02 of the true p
- Finally, recall that applying the CLT requires
  - The samples to be ***independent:*** reasonable assumption if they are randomly assigned to control/ treatment groups, or if they are drawn randomly from a large population.
  - The **success/failure** condition to hold: $np \geq 10$ and $n(1-p) \geq 10$

Using the CLT in Practice

- In the real world, we don't have access to p, only to $\hat{p}$. We still would like to understand how well $\hat{p}$ approximates p.
- To use the CLT, we have to check the **success/failure condition,** and the **independence condition.**
- Not knowing p, has no effect on whether our sample is independent.
- For the S/F condition, simply replace p by $\hat{p}$, so just check that $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
- Similarly, to estimate $SE_{\hat{p}}$, we can simply replace p by $\hat{p}$ to get

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

Using the CLT in Practice

- **Example:** Suppose you randomly sample $n = 1000$ people, and find that 761 of them support candidate A.
- Calculate $\hat{p}$ and $SE_{\hat{p}}$
- **Answer:**
- $\hat{p} = \dfrac{761}{1000} = 0.761$
- $SE_{\hat{p}} \approx \sqrt{\dfrac{0.761(1 - 0.761)}{1000}} = 0.0135$
- **Remark:** If the true value of p was 0.75 as in our previous example, then $SE_{\hat{p}} = 0.0137$, as calculated before. This is quite close to our estimate of 0.0135 obtained using 0.761 instead of 0.75.