ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

# CONFIDENCE INTERVALS FOR A PROPORTION

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

# Confidence Intervals Using the CLT

- The sample proportion $\hat{p}$, is a point estimate for the population proportion $p$.
- As such, it is a single plausible value for $p$. It is **not a perfect estimate,** and has a standard error associated with it — as we have seen.
- Instead of just reporting our imperfect estimate $\hat{p}$, we can report a **range of plausible values for $p$.**
- How do we do this?
  - CLT to the rescue!
  - When CLT conditions are satisfied: $\hat{p}$ follows a **normal distribution**, so if we repeat an experiment many many times, $95\,\%$ of the time, $\hat{p}$ would fall within 1.96 standard deviations of $p$.

## Confidence Intervals Using the CLT

- We can say *we are 95% confident that the interval*

$$I = (\hat{p} - 1.96 \cdot SE_{\hat{p}} \quad , \quad \hat{p} + 1.96 \cdot SE_{\hat{p}})$$

*captures $p$.*

- **Interpretation:** If we **repeatedly** collect $n$ random samples and, each time, use them to calculate the sample proportions $\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots,$ and construct the 95% confidence intervals

$$I_j = (\hat{p}_j - 1.96 \cdot SE_{\hat{p}} \quad , \quad \hat{p}_j + 1.96 \cdot SE_{\hat{p}}), \quad j = 1,2,3,\ldots$$

then, $\approx 95\,\%$ of the intervals would contain the population mean $p$.

- **Importantly:** $\approx 5\,\%$ of our intervals would not contain $p$!

# Confidence Intervals Using the CLT

- But where did the 1.96 come from?

- Mathematically, we are looking for an interval $I$, so that $\mathbb{P}(p \in I) \approx 0.95$. Here the randomness is on the draw of the samples.

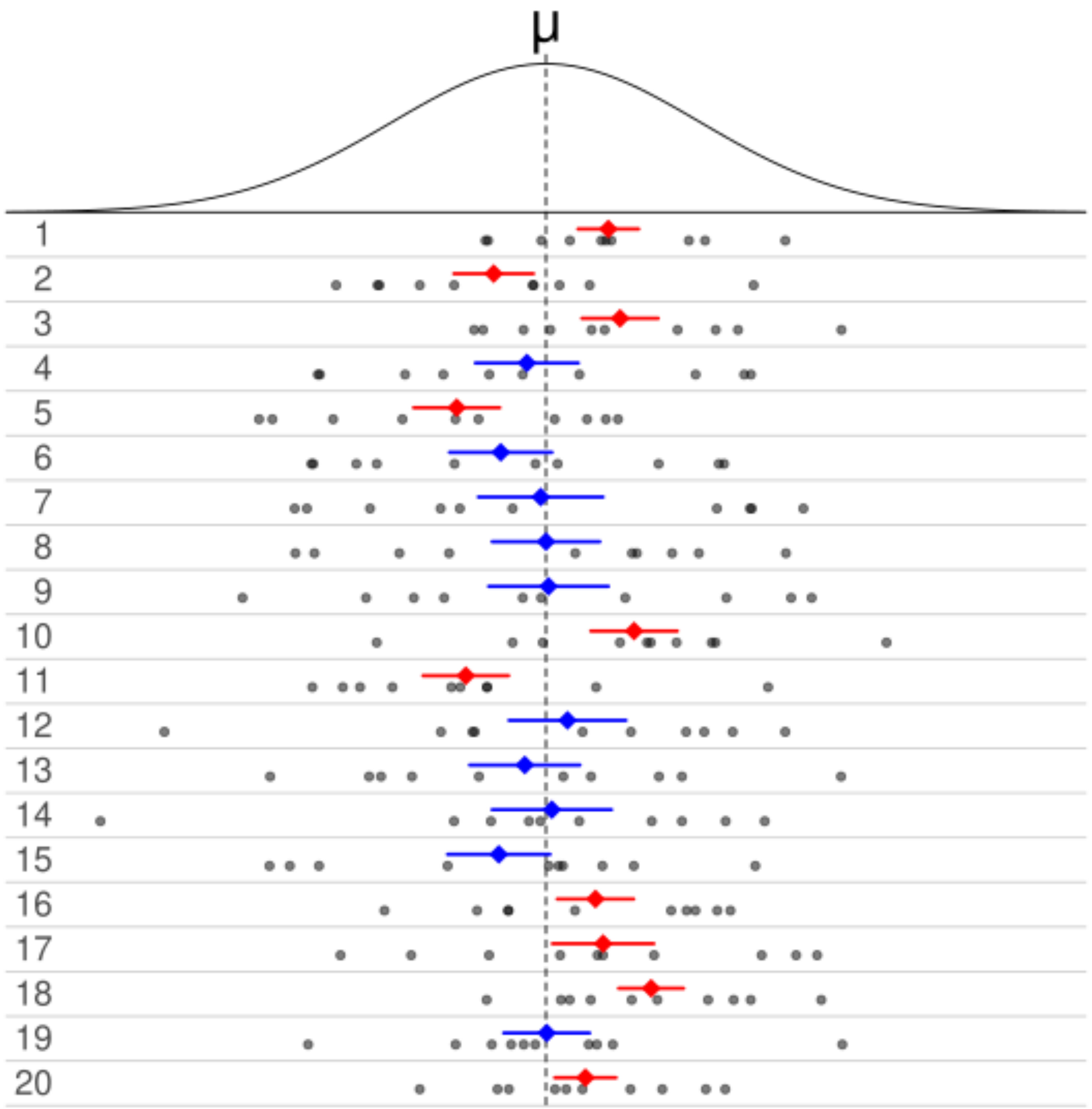
- Using the CLT approximation, we can choose $I = (\hat{p} - z^\star \cdot SE_{\hat{p}} \quad , \quad \hat{p} + z^\star \cdot SE_{\hat{p}})$. What we need is a value of $z^\star$, so that $\mathbb{P}(p \in I) \approx 0.95$.

- In other words solve for $z^\star$ in: $\dfrac{1}{\sqrt{2\pi} \cdot SE_{\hat{p}}} \displaystyle\int_{\hat{p}-z^\star \cdot SE_{\hat{p}}}^{\hat{p}+z^\star \cdot SE_{\hat{p}}} e^{\frac{(x-\hat{p})^2}{SE_{\hat{p}}^2} dx} \approx 0.95$

- Turns out, the right value of $z^\star$ to achieve $95\,\%$ confidence, is $z^\star = 1.96$. (Use probability tables, or software to find $z^\star$).

## Interpretation of CI

- On the right is a graph, showing 20 experiments.
- In each one, $n$ random samples are drawn and a sample proportion is calculated, and a $50\ \%$ Confidence Interval is constructed (the colored lines).
- The distribution on top represents the underlying normal distribution with mean $\mu$.
- As these are $50\ \%$ CI, you can see that the "blue intervals" which contain $\mu$, constitute about $50\ \%$ of all intervals.



# Example: Confidence Intervals Using the CLT

- **Example:** In our favorite example, 761 out of 1000 people sampled support candidate A. Compute and interpret a $95\,\%$ confidence interval for the population proportion.
- **Answer:**

$$\hat{p} = 0.761, SE_{\hat{p}} \approx 0.0135$$

$$I = (\hat{p} - 1.96 \cdot SE_{\hat{p}} \quad , \quad \hat{p} + 1.96 \cdot SE_{\hat{p}})$$

$$\implies I = (0.761 - 1.96 \times 0.0135,\ 0.761 + 1.96 \times 0.0135)$$

$$\implies I = (0.7346,\ 0.7874)$$

## Variations with Different Confidence Levels

- Suppose we want to construct a Confidence Interval with a $99\ \%$ confidence level.
- First, let's notice that this interval should be *wider* than the interval at the $95\ \%$ level.
- We are now looking for an interval, so that $\mathbb{P}(p \in I) \approx 0.99$.
- Using the CLT approximation, we can write $I = (\hat{p} - z^\star \cdot SE_{\hat{p}} \quad , \quad \hat{p} + z^\star \cdot SE_{\hat{p}})$. What we need is a value of $z^\star$, so that $\mathbb{P}(p \in I) \approx 0.99$.
- Turns out (probability tables/software), choosing $z^\star = 2.58$ gives

$$\frac{1}{\sqrt{2\pi} \cdot SE_{\hat{p}}} \int_{\hat{p}-z^\star \cdot SE_{\hat{p}}}^{\hat{p}+z^\star \cdot SE_{\hat{p}}} e^{\frac{(x-\hat{p})^2}{SE_{\hat{p}}^2}} dx \approx 0.99$$

## Summary: Confidence Interval for a Single Proportion

To construct a confidence interval for a single proportion:

1. **Identify $\hat{p}$ and $n$** and determine what confidence level you wish to use.
2. **Verify the independence and suceess/failure conditions** to ensure $\hat{p}$ is nearly normal. Use $\hat{p}$ in place of $p$ to check the success/failure condition.
3. If the conditions hold, compute $SE$ using $\hat{p}$, find the relevant $z^\star$, and **construct the interval.**
4. **Interpret the confidence interval** in the context of the problem.