Applied Statistics for Data Analytics

Module 3: Confidence intervals

DeepLearning.AI

Confidence intervals

Module 3 introduction

DeepLearning.AI

Module 3 outline

- **Inferential statistics**
  - Define inferential statistics
- **Confidence intervals for means**
  - Construct intervals
  - Interpret intervals
  - Manage precision of the estimate
- **Confidence intervals for proportions**
  - Construct intervals
  - Interpret intervals
- **LLMs for confidence intervals**
  - Interpret results
  - Create simulation interfaces
  - Perform calculations
  - Create visualizations

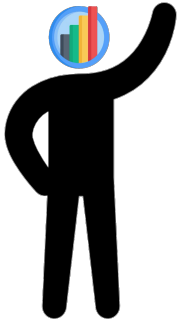
Diamond prices

Confidence intervals

Inferential statistics

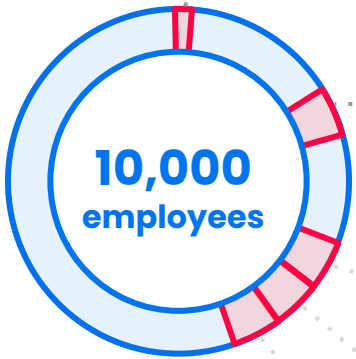
DeepLearning.AI

Scenario



You
Data Analyst

Figure out the rate of employee satisfaction:



Satisfaction is likely high

100 employees ➡ **82.0%** satisfied

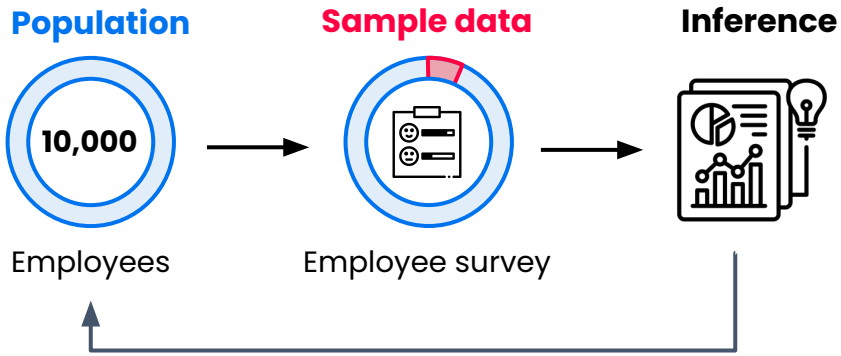
500 employees ➡ **91.0%** satisfied

500 employees ➡ **91.0%** satisfied

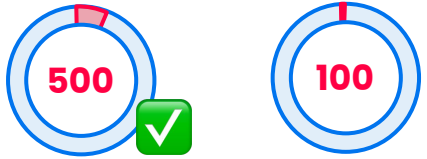
500 employees ➡ **92.1%** satisfied

500 employees ➡ **88.7%** satisfied

Inferential statistics



- Larger samples provide more reliable estimates than smaller samples



- Different samples show variability, even when drawn from the same population

- Allows you to quantify your level of confidence in your estimate

Descriptive statistics

State facts about sample

- "In a sample of 100 employees, 82 said they were satisfied."
- "A survey of 160 parents found that the parents with newborns got 6.1 hours of sleep while the parents with older children got 8.2 hours."

Describe the characteristics of a sample

Inferential statistics

Use sample data to draw conclusions about population

- Based on a sample of 100 employees, employee satisfaction among all 10,000 employees is likely between 88% and 91%.
- A survey of parents concluded that parents of older children sleep 2 hours more per night compared with parents of newborns.

Use characteristics of sample to make conclusions about population

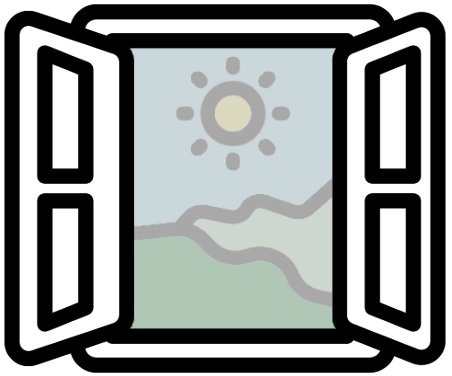
Inferential statistics

- Use probability to draw conclusions about population based on sample statistics
  - Taking into account:

Size of sample

Variability of the sample

Other factors



Sample

Inferential vs. descriptive statistics

Descriptive statistics

✅ "82 out of 100 employees are satisfied"

⛔ Can't generalize that to all employees

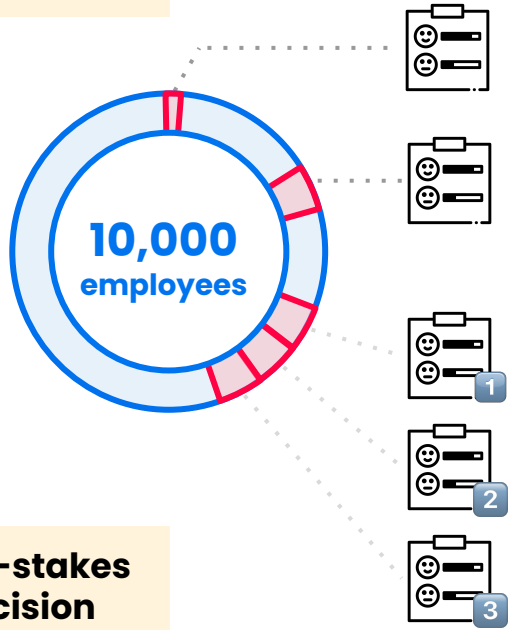
Low-stakes decision

Inferential statistics

✅ "The true proportion of satisfaction is 88-91%"

✅ Population parameter is likely to fall within those two values

High-stakes decision

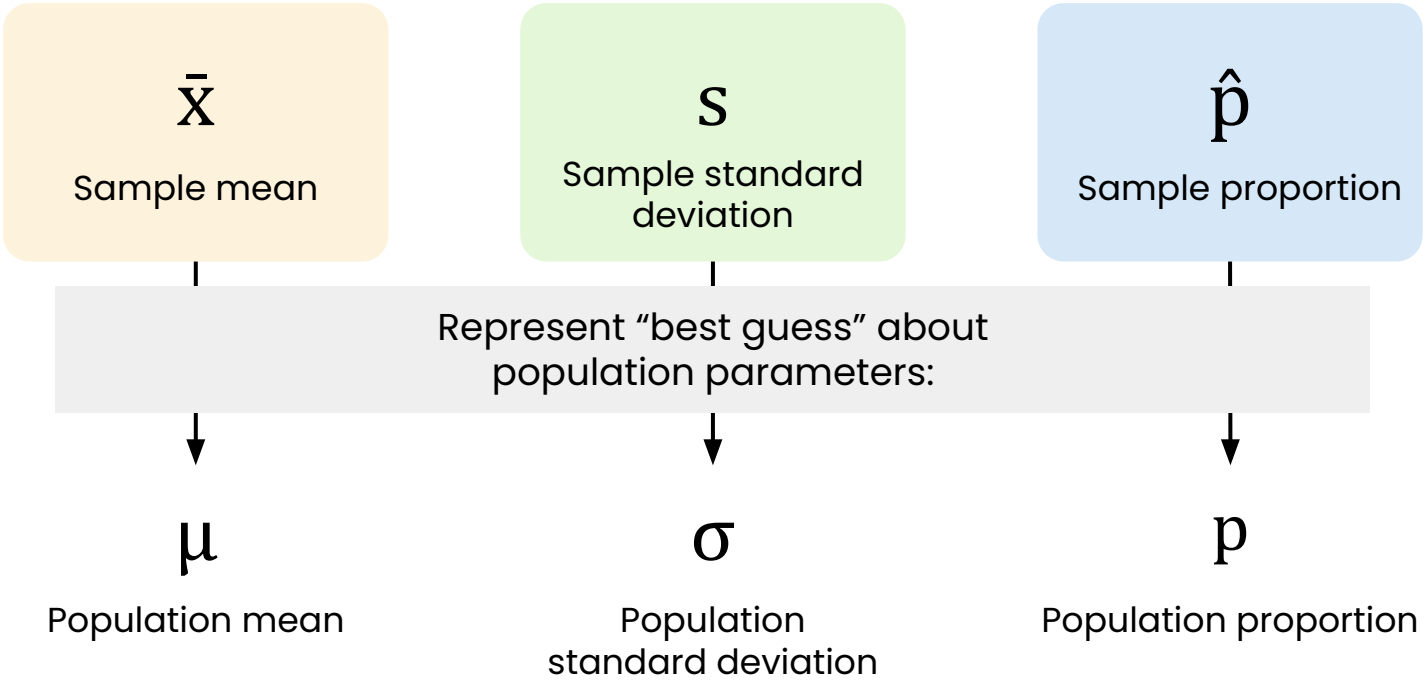


Confidence intervals

Point & interval estimates

DeepLearning.AI

Point estimates



Point estimates

❌ **Don't** contain information about the confidence of estimate

Example:

Random sample of 25 movies from 2013

↳ $\bar{x}$ of durations = 121 minutes

How certain can you be that true population mean is **exactly 121**? 🤔

Interval estimates

✅ Contain information about how confident you can be

Example:

Arriving within 10 to 15 mins 😎

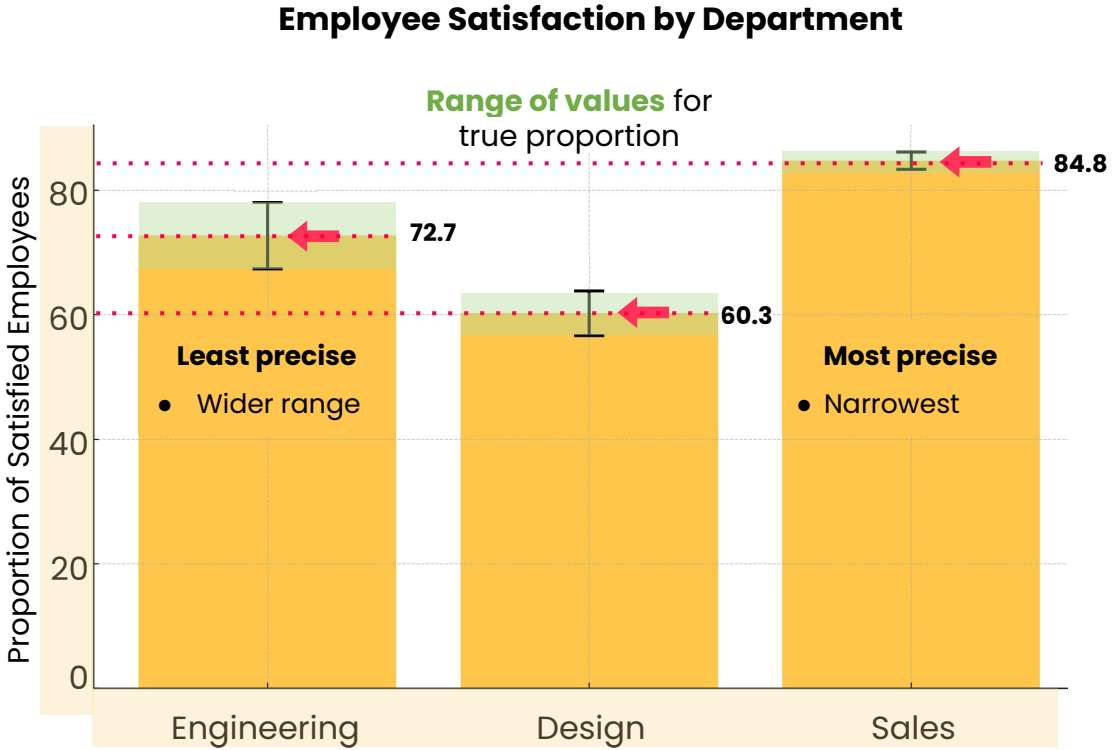
Arriving in exactly 10 mins 🤔

Visualizing intervals

Error bars: visual representation of an **interval estimate**

- Inference about true population proportion based on sample

🔁 If repeated many times, expect true population value to fall within this range

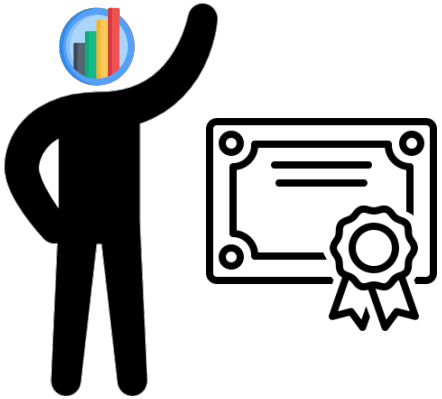


Confidence intervals

Sampling distributions &
the central limit theorem

DeepLearning.AI

Scenario



You
Data Analyst

💬 **Task**: Estimate average score on a professional certification exam

🏆 Possible scores: 0 to 100

Sample 1

📝 Ask **50 random people** their scores

Sample 2

📝 Ask **50 random people** their scores

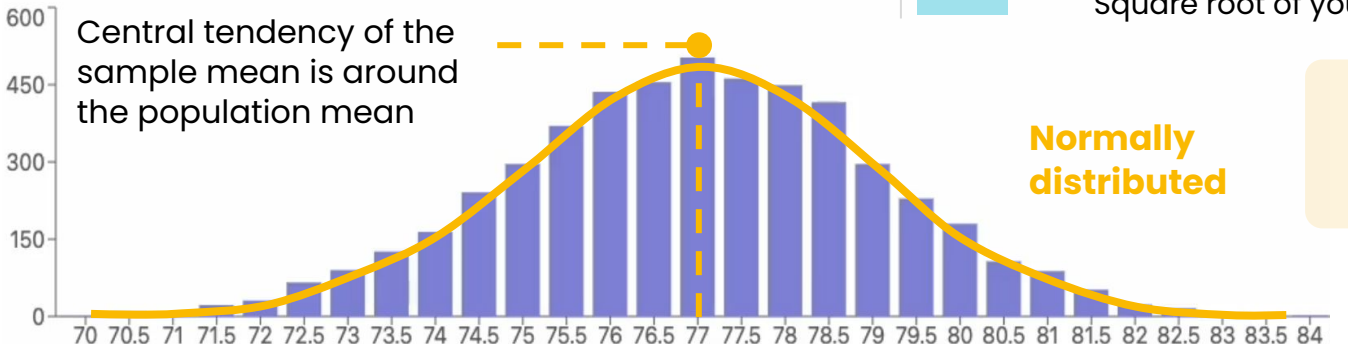
$\bar{x}$ **of scores** ⟵ Same value? ⟶ $\bar{x}$ **of scores**

Central limit theorem

$n > 30$

If you take sufficiently large samples from any distribution and calculate their **means**:

- Sample means will be normally distributed
- Mean of distribution will equal μ



Central tendency of the sample mean is around the population mean

Normally distributed

Increase your sample size

Estimate becomes more precise

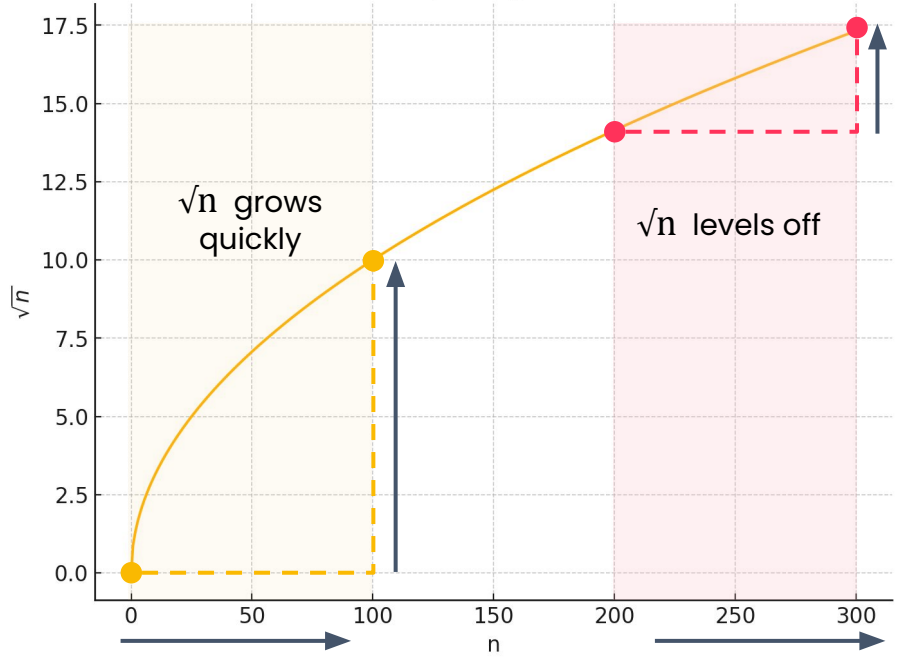
Standard deviation of your sample mean (**standard error** of the mean)

$$SE = \frac{\text{True population standard deviation}}{\text{Square root of your sample size}} = \frac{\sigma}{\sqrt{n}}$$

As n gets larger, $\sqrt{n}$ gets larger, but at a **slower rate**

Sample size and precision

Large gains in precision from adding a few more values

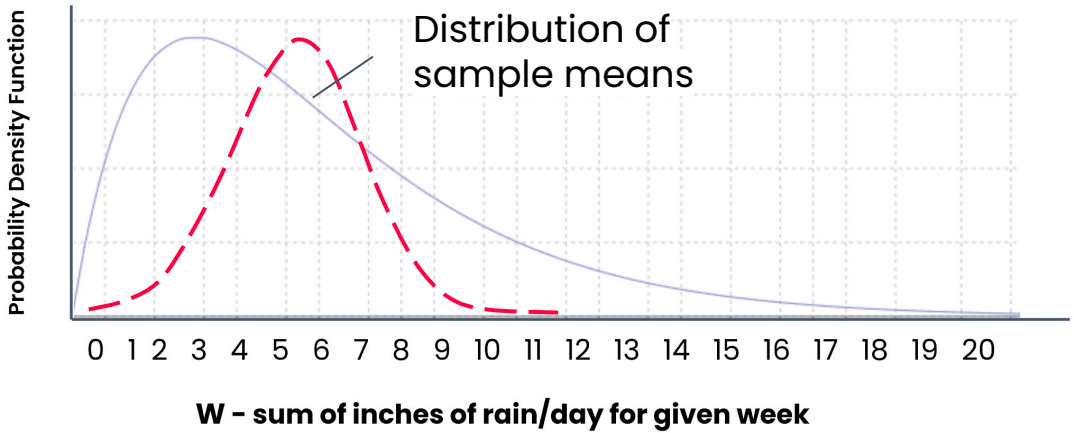


Diminishing returns for precision

Central limit theorem

Even if sample data is not normal:

- Sample means **will** be normal as long as $\mathbf{n}$ is sufficiently large



Central limit theorem applies to:

- Sample mean $\bar{x}$
- Sample proportion $\hat{p}$

In some cases:

- Sample variance s^2
- Sample standard deviation s

Confidence intervals

Demo: confidence intervals in action

DeepLearning.AI

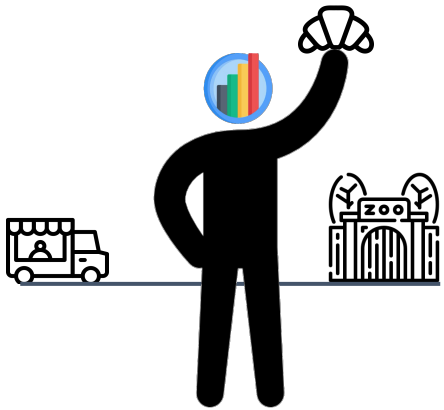
Confidence intervals

Confidence intervals

DeepLearning.AI

Scenario

🕖 Before 7am



You
Data Analyst

Task: Figure out how long it takes to deliver the pastries

Sample data:

🗓️ Monitor the delivery truck for 30 days

⏰ Record time it takes to get from bakery to zoo

$\bar{x} = 43$ minutes

$s = 11$ minutes

Maybe due to chance your deliveries were:

🐇 Unusually fast
🐢 Unusually slow

Confidence interval

$\bar{x} = 43$ minutes

$s = 11$ minutes

Best guess estimate for mean delivery time (μ)

43

$\bar{x} - 2 \times SE$

$= 43 - 2 \times 2$

$= 39$

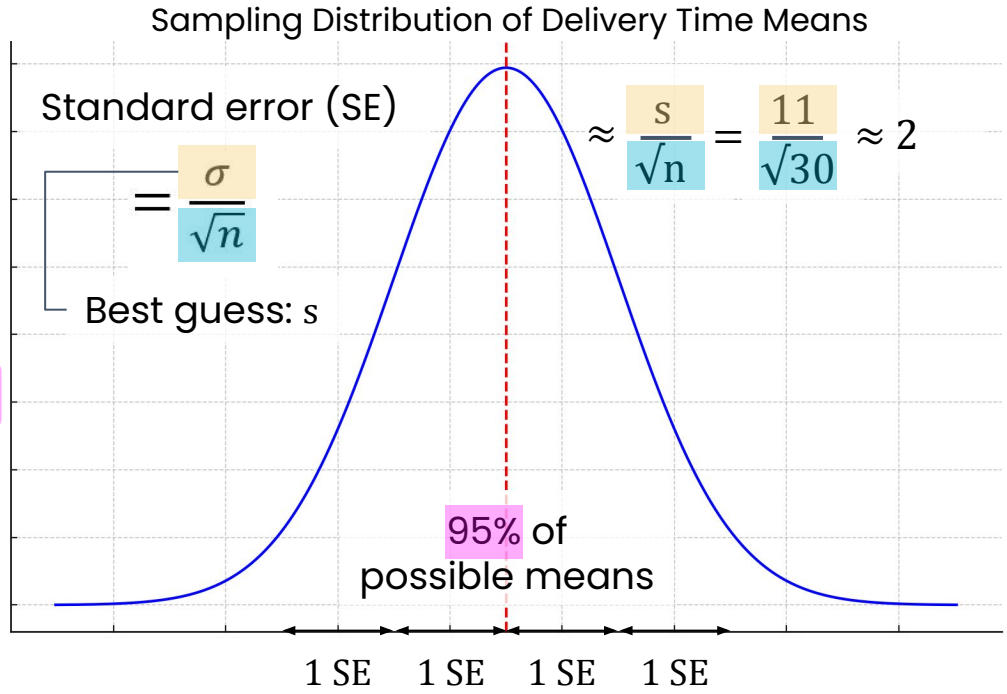
$\bar{x} + 2 \times SE$

$= 43 + 2 \times 2$

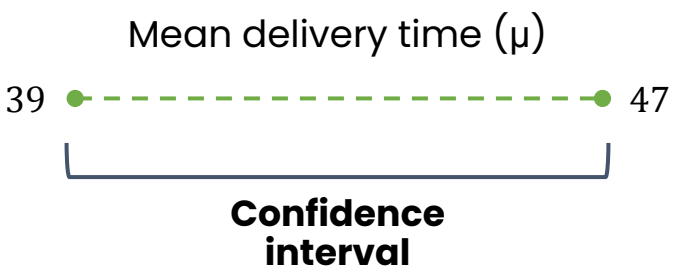
$= 47$

Variability | Sample size | Confidence

With 95% confidence, the true mean delivery time is between 39 and 47 minutes.



Confidence interval



- Range of values used to estimate a population parameter
- Quantifies the uncertainty of an estimate by the relative width of the range

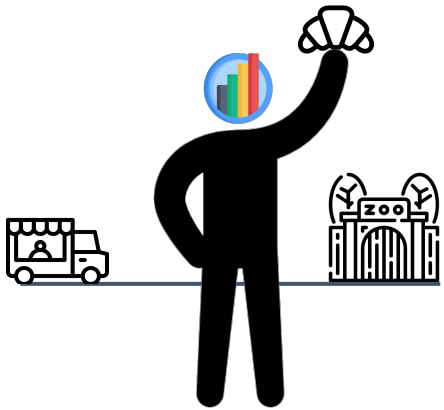
Broader range
High uncertainty

Narrow range
More precision

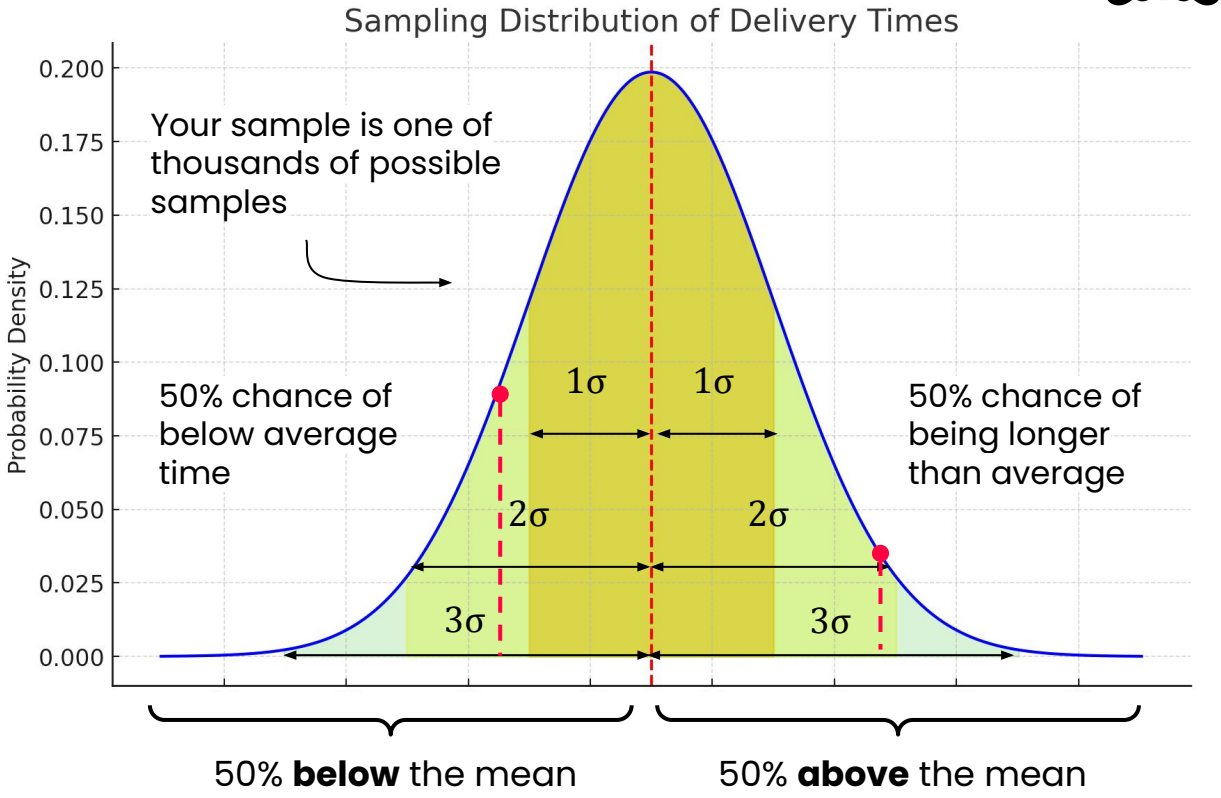
Understanding the possible average delivery time can help with precise scheduling

Scenario

Before 7am

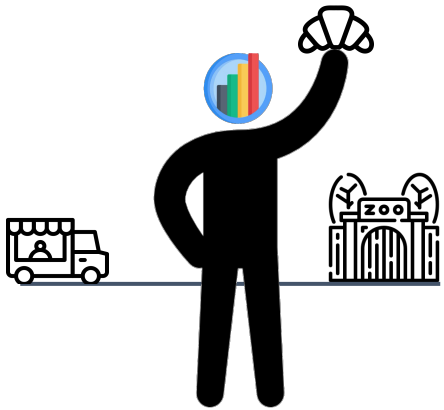


You
Data Analyst



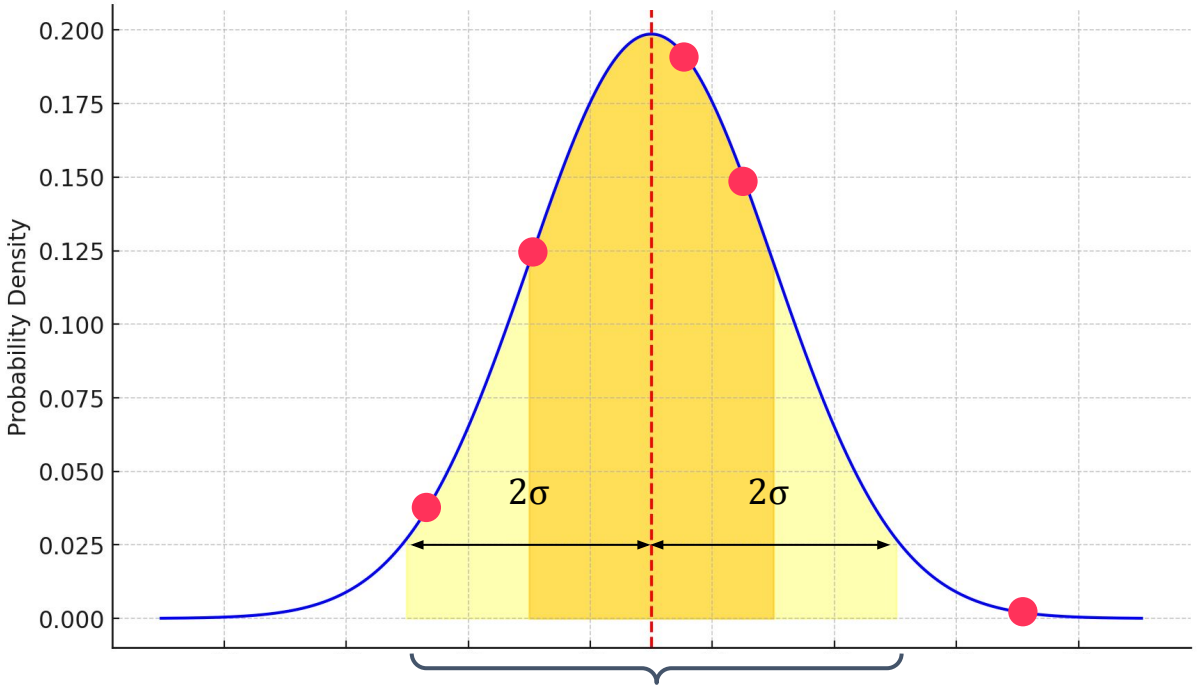
Scenario

Before 7am



You
Data Analyst

Probability that sample mean 43 is within **2 standard deviations** of μ?



Based on the two sigma rule: **95% chance**

Confidence intervals

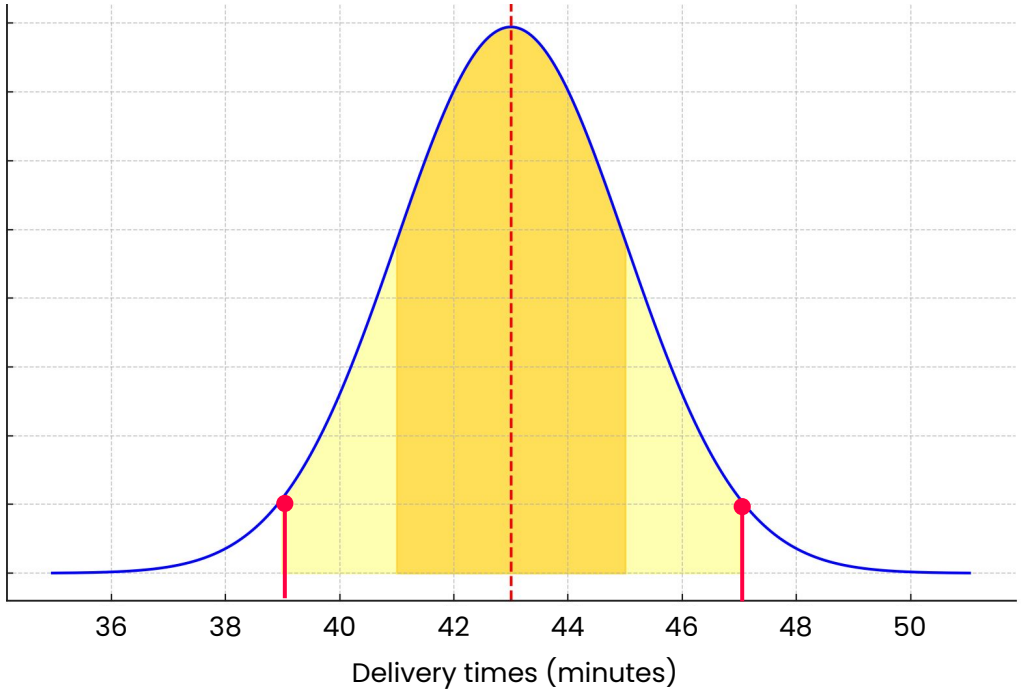
Mechanisms of confidence intervals

DeepLearning.AI

Example

- 95% confidence level reflects how reliable your estimation method is
- A method that's designed to be right about 95% of the time
- Population parameter is:
  - Fixed, but unknown to you
  - In this particular interval or it’s not

With 95% confidence, the true average delivery time is between 39 and 47 minutes



Confidence interval for the population mean

$$\bar{x} \pm z \times \left(\frac{s}{\sqrt{n}} \right)$$

Where:

- $\bar{x}$ - sample mean
- s - sample standard deviation
- n - sample size
- z - z-score value from the standard normal distribution

z-score

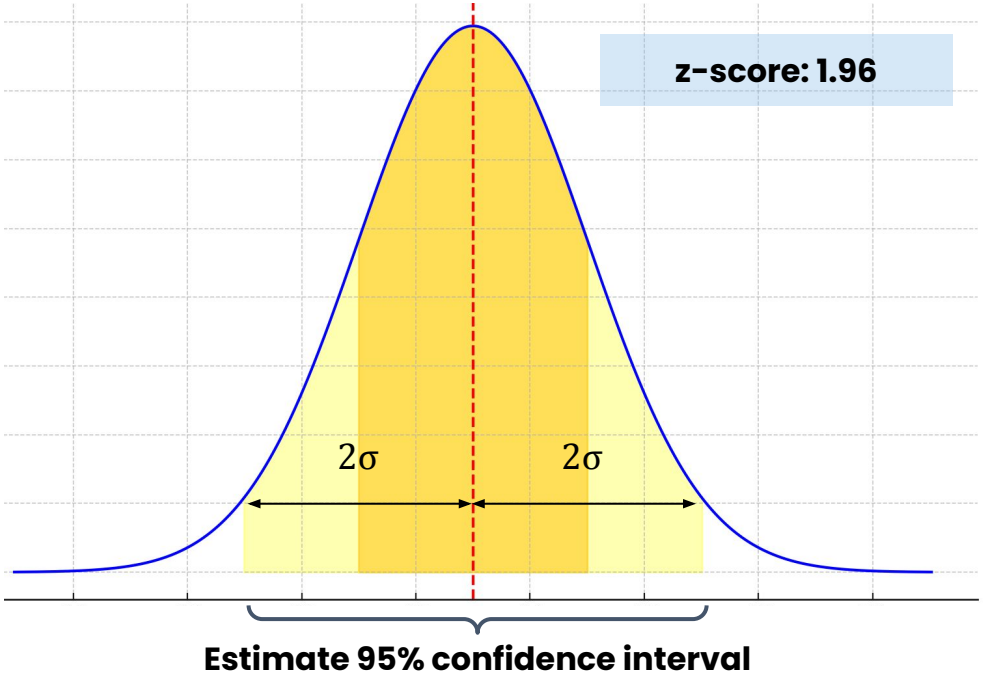
- Controls how "confident" you are that interval contains the population mean
- Number of standard deviations from mean in standard normal distribution

Confidence interval for the population mean

To calculate a **95% confidence interval**:

$$\bar{x} \pm 1.96 \times \left(\frac{s}{\sqrt{n}} \right)$$

- The two sigma rule is just an estimate
- $z = 2$ is associated with slightly higher confidence than 95%
- Use $z = 1.96$ to be more precise



Confidence interval for the population mean

To calculate a **95% confidence interval**:

$$\bar{x} \pm 1.96 \times \left(\frac{s}{\sqrt{n}} \right)$$

Margin of error

- Constructs the confidence interval
- Helps gauge the precision of the estimate

Confidence interval	z score
90%	1.645
95%	1.96
99%	2.576

⬆️ Higher level of confidence

⬆️ Use a higher z score

⬆️ Generate a wider confidence interval

Choosing a confidence level

90% confidence level

Used when missing true value is less important

Estimate avg. user rating

Users on average rate a new feature **7.2 out of 10**

✅ A less precise estimate is acceptable to help make initial decisions

95% confidence level

Balances confidence with potential for error

Estimate avg. delivery time

The average delivery time is between **39 and 47 minutes**

✅ Don't need to schedule too much buffer time

99% confidence level

Used when you want to minimize the risk of error

Estimate pollution in a river

Estimating pollution in a river

✅ Can help reduce risk of harmful impact

Confidence intervals

Understanding
margin of error

DeepLearning.AI

Margin of error

$$z \times \left(\frac{s}{\sqrt{n}} \right)$$

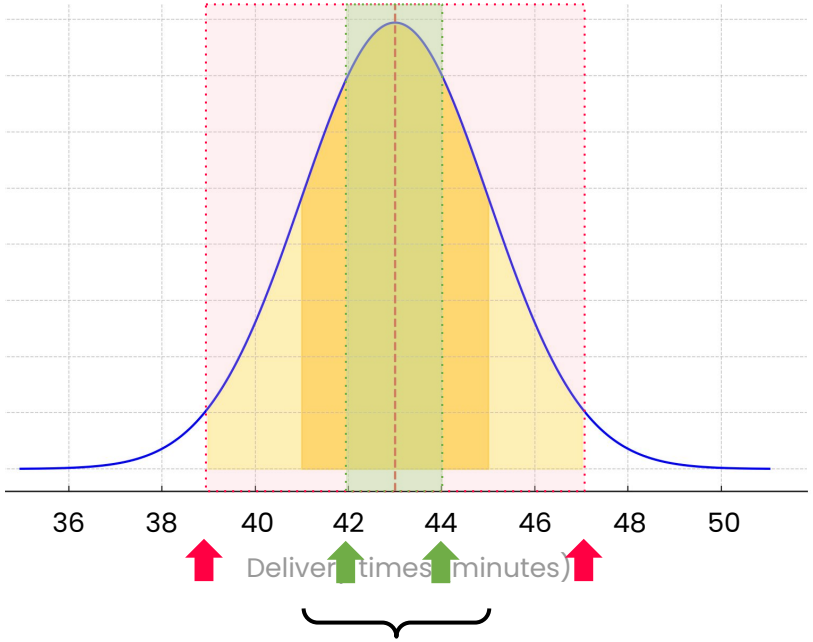
— Determines how wide your confidence interval is

✅ A **narrower confidence interval** is more desirable because it means you have a **more precise** estimate

Depends on three factors:

1. Desired confidence level
2. Standard deviation
3. Sample size

i.e. the amount of variability in your data



🗓️ Precision allows for better scheduling

Desired confidence level

$$z \times \left(\frac{s}{\sqrt{n}}\right) = z \times \left(\frac{10}{\sqrt{100}}\right) = z \times \left(\frac{10}{10}\right) = z \times 1$$

90% confidence level — $z = 1.645$

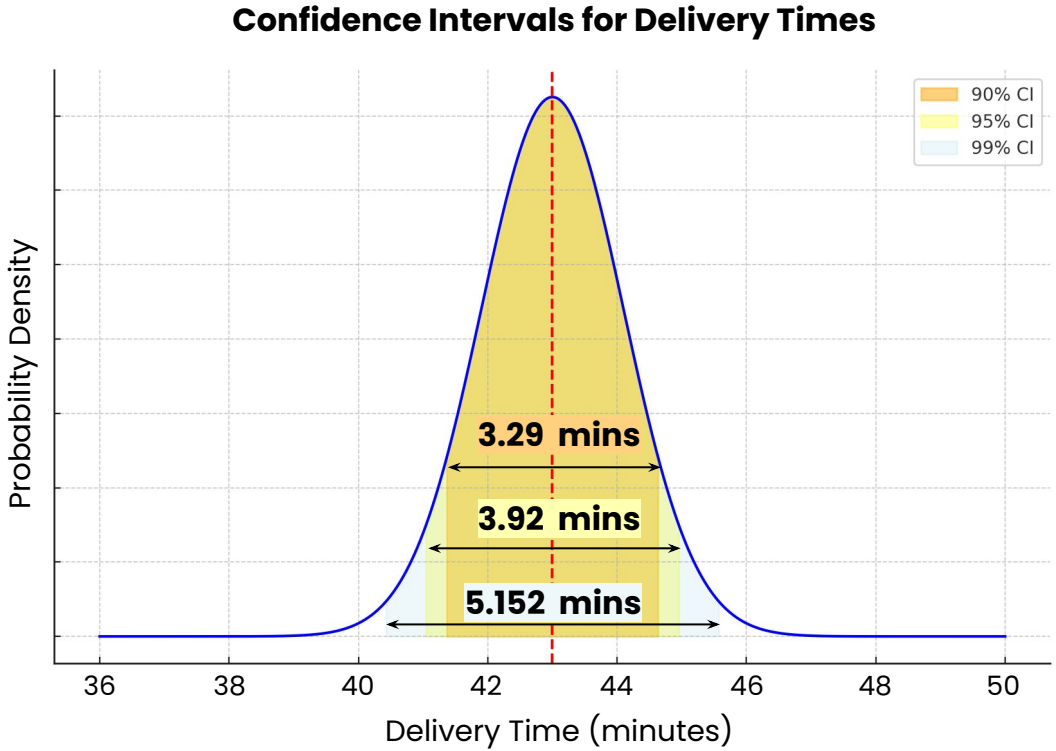
$z \times 1 = 1.645 \times 1 = 1.645$

95% confidence level — $z = 1.96$

$z \times 1 = 1.96 \times 1 = 1.96$

99% confidence level — $z = 2.576$

$z \times 1 = 2.576 \times 1 = 2.576$



Width vs. Confidence level

$$z \times \left(\frac{s}{\sqrt{n}}\right) = z \times \left(\frac{10}{\sqrt{100}}\right) = z \times \left(\frac{10}{10}\right) = z \times 1$$

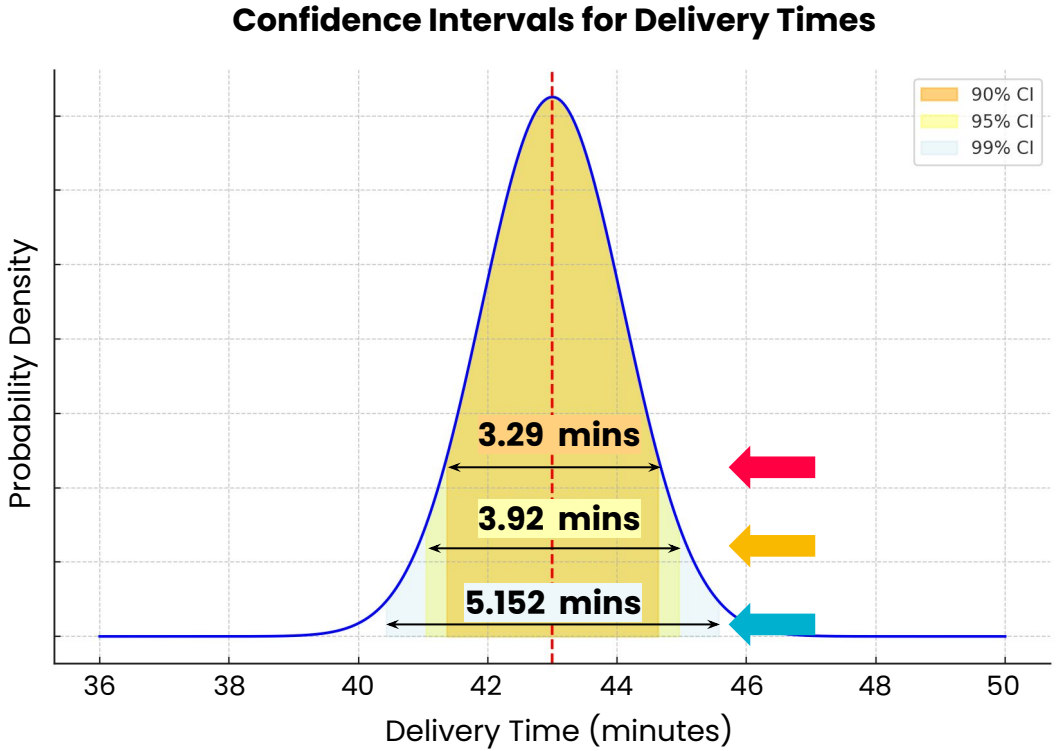
90% confidence level

95% confidence level

- 19% wider
- Gain 5% in relative confidence

99% confidence level

- 31% wider
- Gain 4% in relative confidence



Variability

95% confidence interval

$$z \times \left(\frac{s}{\sqrt{n}}\right) = z \times \left(\frac{s}{\sqrt{100}}\right) = 1.96 \times \left(\frac{s}{\sqrt{100}}\right)$$

s = 10

$$1.96 \times \left(\frac{10}{\sqrt{100}}\right) = 1.96 \times \left(\frac{10}{10}\right) = 1.96 \times 1 = 1.96$$

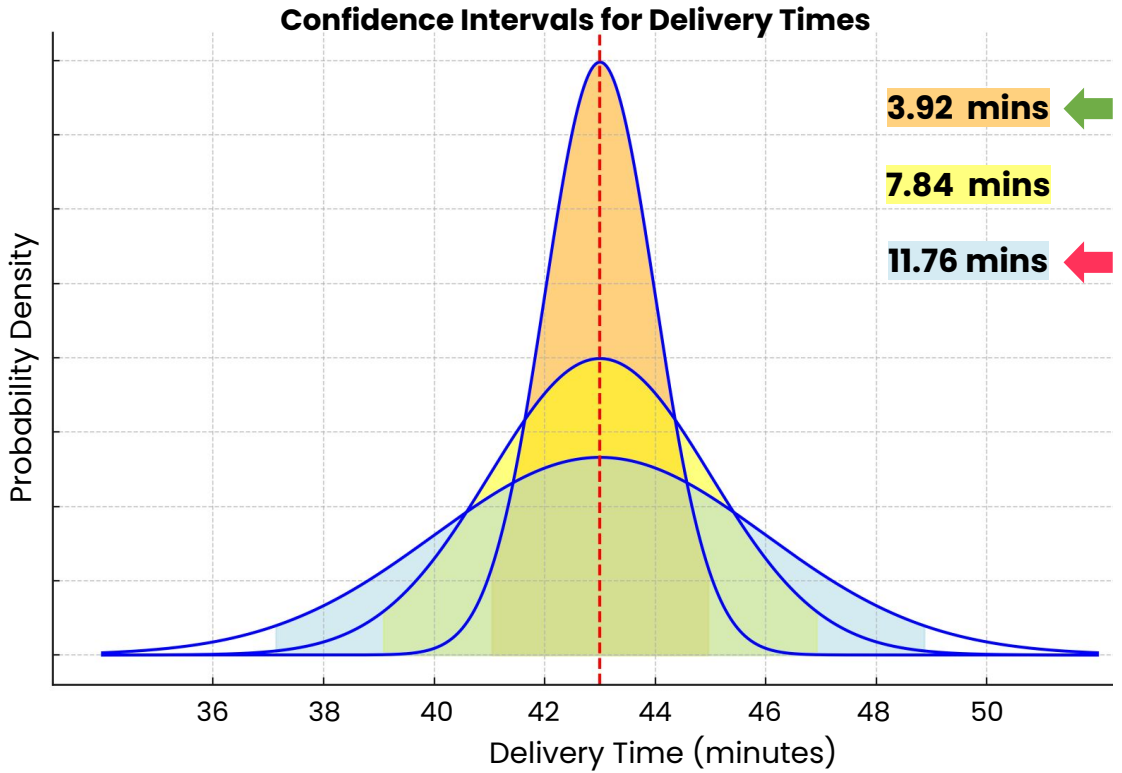
s = 20

$$1.96 \times \left(\frac{20}{\sqrt{100}}\right) = 1.96 \times \left(\frac{20}{10}\right) = 1.96 \times 2 = 3.92$$

s = 30

$$1.96 \times \left(\frac{30}{\sqrt{100}}\right) = 1.96 \times \left(\frac{30}{10}\right) = 1.96 \times 3 = 5.88$$

If data has a lot of variability, expect a less precise estimate.



DeepLearning.AI

Sean Barnes

City bus system

Rural bus system

✅ Smaller variability makes it easier to estimate the true average arrival time

Sample size

$$z \times \left(\frac{s}{\sqrt{n}} \right) \quad 1.96 \times \left(\frac{10}{\sqrt{n}} \right)$$

n = 100

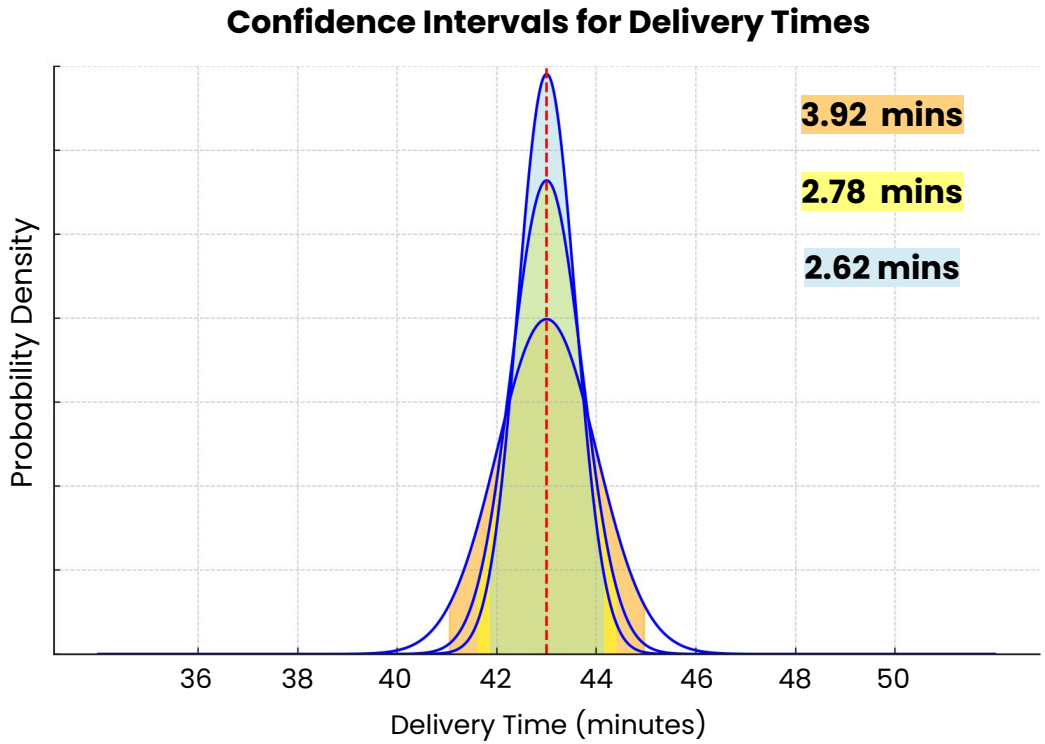
$$1.96 \times \left(\frac{10}{\sqrt{100}} \right) = 1.96 \times \left(\frac{10}{10} \right) = 1.96 \times 1 = 1.96$$

n = 200

$$1.96 \times \left(\frac{10}{\sqrt{200}} \right) = 1.96 \times \left(\frac{10}{14.1} \right) = 1.96 \times 0.71 \approx 1.39$$

n = 300

$$1.96 \times \left(\frac{10}{\sqrt{300}} \right) = 1.96 \times \left(\frac{10}{17.3} \right) = 1.96 \times 0.58 \approx 1.13$$



Sample size

$$z \times \left(\frac{s}{\sqrt{n}} \right)$$

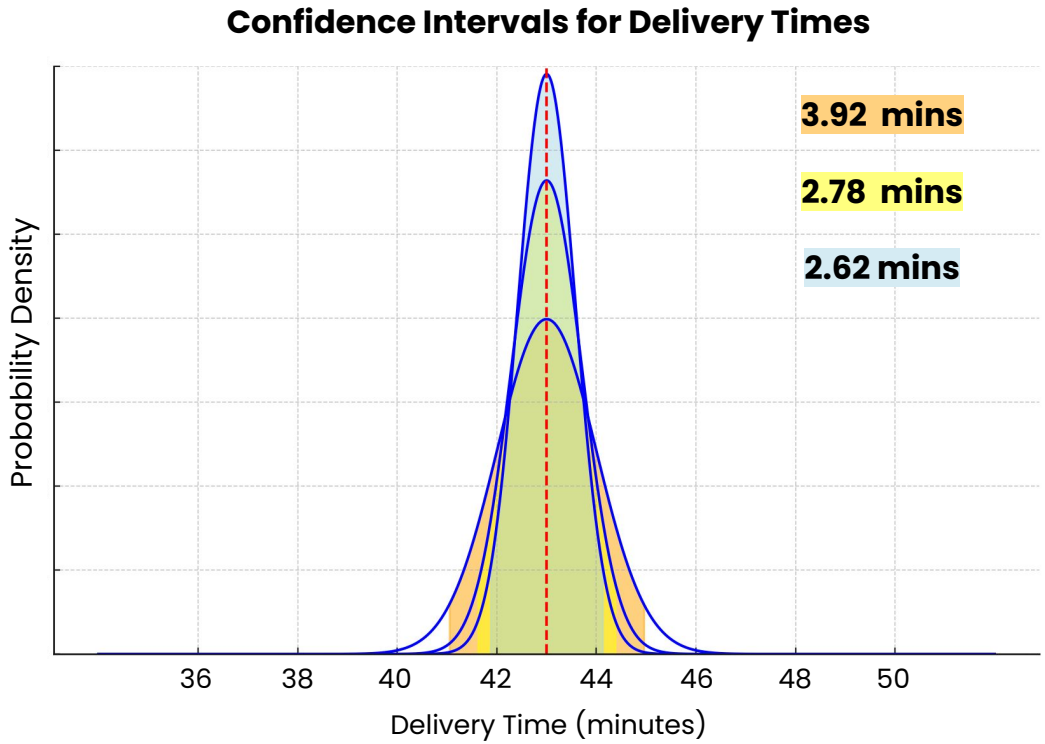
A larger sample:

- Allows you to construct a narrower confidence interval
- Reduces the size of the margin of error
- Produces diminishing returns

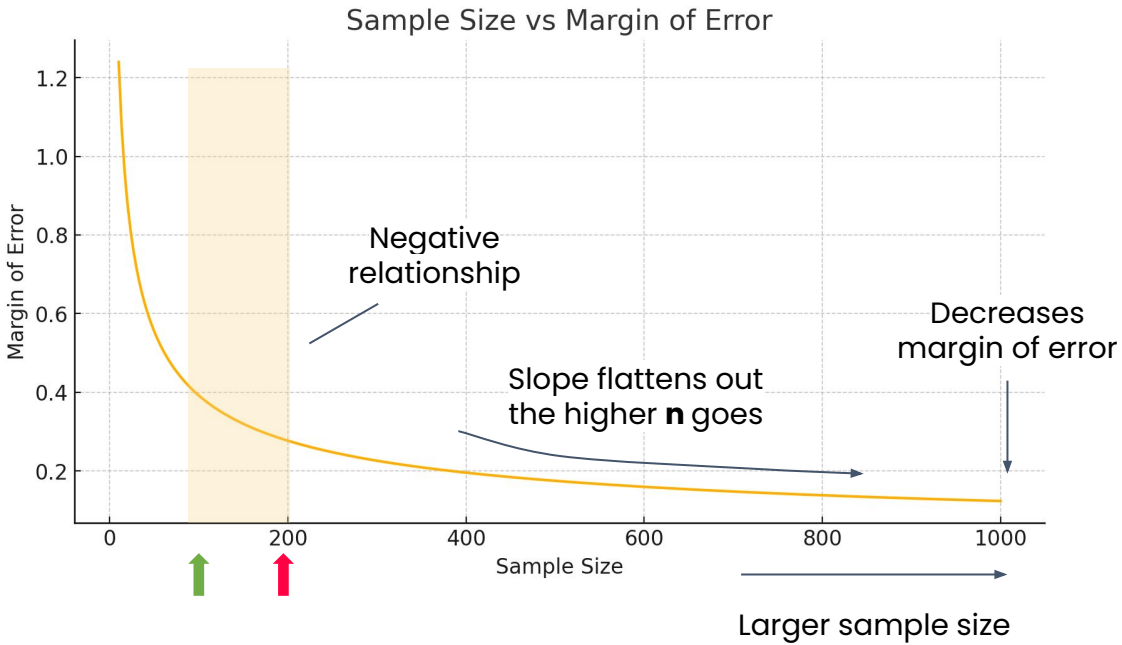
$\mathbf{n = 100}$

$\mathbf{n = 200}$ → Reduced interval by 29%

$\mathbf{n = 300}$ → Reduced interval by 6%



Sample size



- This relationship is a basic fact of statistics
- Often need a relatively small sample for very large population
- Adding more and more data will narrow interval
- In very high sample sizes, you will observe diminishing impact

Summary

You can achieve a narrower confidence interval in three ways:

1. By lowering your confidence level, which can increase your chances of missing the true value

2. By working with data that has less variability, which is generally out of your control

3. By increasing your sample size, though this approach has diminishing returns

Confidence intervals

Demo: confidence intervals for means

DeepLearning.AI

Confidence intervals

Confidence intervals
for proportions

DeepLearning.AI

Scenario

You
Data Analyst

💬 **Problem**: “What proportion of deliveries are on time?”

Deliveries that made it to the zoo by 7am → p

True proportion of on-time deliveries

🚚 Collect a sample of 30 deliveries

📝 Record if they were:

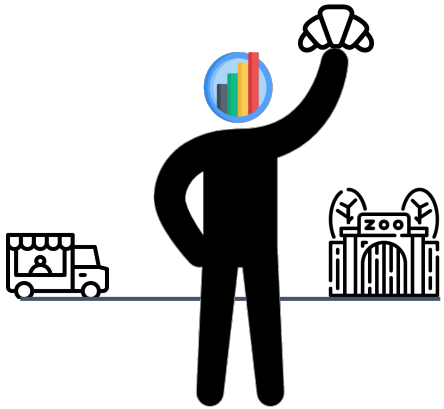
✅ On time

❌ Not on time

$\hat{p}$ ←

Estimate for the true proportion of p

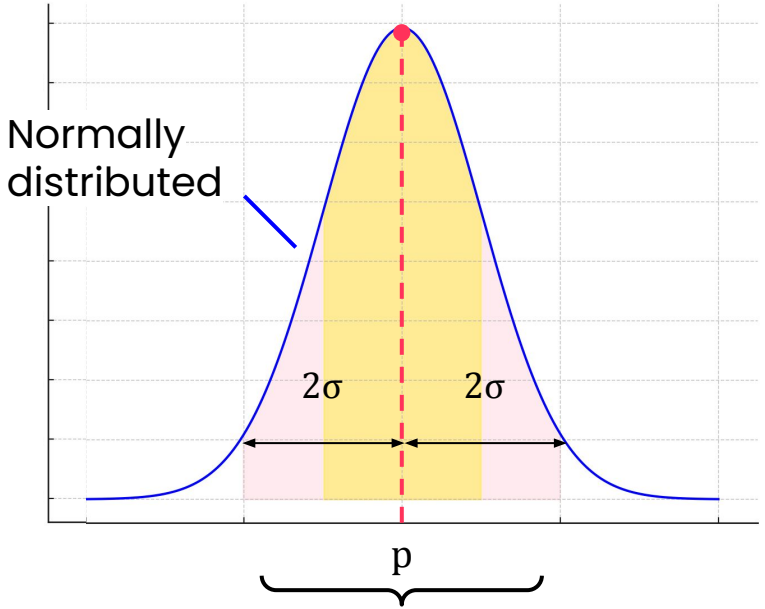
Scenario



You
Data Analyst

$\hat{p} = 0.6$ → **18 out of 30** on-time deliveries

Sampling distribution of $\hat{p}$



95% chance

Standard error

The standard error for the proportion is:

Square root

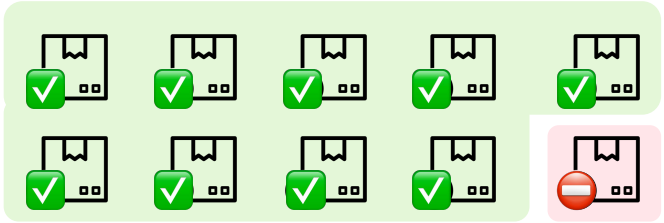
$$\sqrt{\frac{p(1-p)}{n}}$$

$p(1-p)$ — Use $\hat{p}$ to estimate

n — Sample size

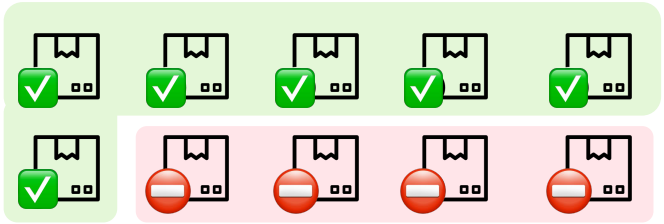
- $p(1-p)$ represents the variability
- Higher variability when $\hat{p}$ is closer to 0.5
- Take the square root so number is at the original scale

Low variability



$\hat{p} = 0.9 \longrightarrow \hat{p}\,(1 - \hat{p}) = 0.09$

Higher variability



$\hat{p} = 0.6 \longrightarrow \hat{p}\,(1 - \hat{p}) = 0.24$

More even mix of success and failures

Confidence interval

✅ $\hat{p}$ – estimate of true proportion of p

✅ $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ – standard error

Confidence interval for the mean:

$$\bar{x} \pm z \times \frac{s}{\sqrt{n}}$$

The interval is defined as:

Margin of error

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

represents the uncertainty in your estimate

Confidence interval

✅ $\hat{p}$ – estimate of true proportion of p

✅ $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ – standard error

Confidence interval for the mean:

$$\bar{x} \pm z \times \frac{s}{\sqrt{n}}$$

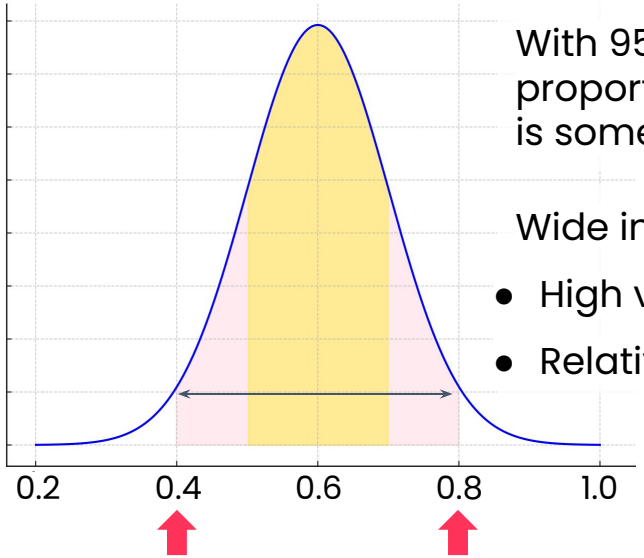
The interval is defined as:

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

Confidence interval

$\hat{p} = 0.6$ →

18 out of 30
on-time deliveries



With 95% confidence, the true proportion of on time deliveries is somewhere within this range

Wide interval is reflected in:

- High variability of the data
- Relatively small sample size

$$= \hat{p} \pm z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.6 \pm z \times \sqrt{\frac{0.6(1-0.6)}{n}}$$

$$= 0.6 \pm z \times \sqrt{\frac{0.6 \times 0.4}{n}}$$

$$= 0.6 \pm 1.96 \times \sqrt{\frac{0.6 \times 0.4}{n}}$$

$$= 0.6 \pm 1.96 \times \sqrt{\frac{0.6 \times 0.4}{30}}$$

$$= (0.4247, 0.7753)$$

Confidence intervals

Demo: confidence intervals for proportions

DeepLearning.AI

Confidence intervals

Interpretation with LLMs

DeepLearning.AI

Confidence intervals

Simulating random sampling with LLMs

DeepLearning.AI

Confidence intervals

Inference and visualization with LLMs

DeepLearning.AI