# ECON 0150 | Economic Data Analysis

*The economist's data analysis pipeline.*

## *Part 3.3 | Central Limit Theorem and Confidence Intervals*

# A Big Question

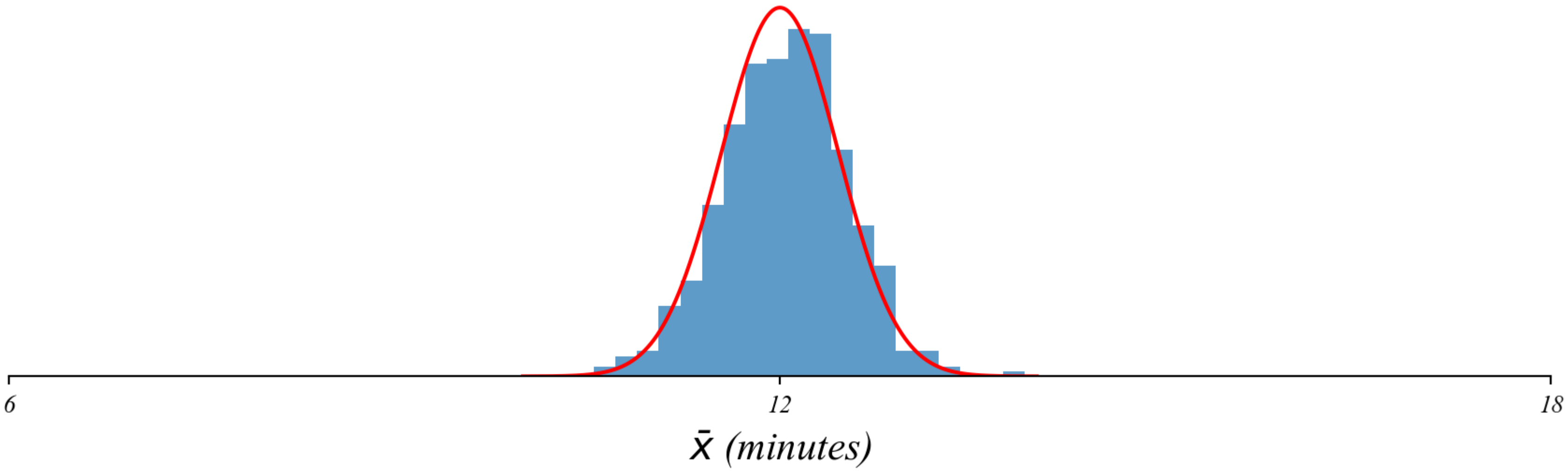
*We found $\bar{x}$ follows a normal distribution around $\mu$... now what?*

> *how can we use this to learn about the population?*

> *lets systematize how "close" $\bar{x}$ and $\mu$ are*

# The Distribution of $\bar{x}$

*Remember: sample means follow a normal distribution with mean ($\mu$) and standard error ($SE = \frac{\sigma}{\sqrt{n}}$).*




> $\mu = 12$ *and* $\sigma = 2.5$ *and* $n = 30$.

# Why SE = $\sigma/\sqrt{n}$?

*The standard error (SE) measures the precision of the estimate.*

Consider n independent observations, each with variance $\sigma^2$.

*1. The sum of* $n$ *samples has variance* $n\sigma^2$ $\left(VAR(a) + VAR(b) = VAR(a+b)\right)$

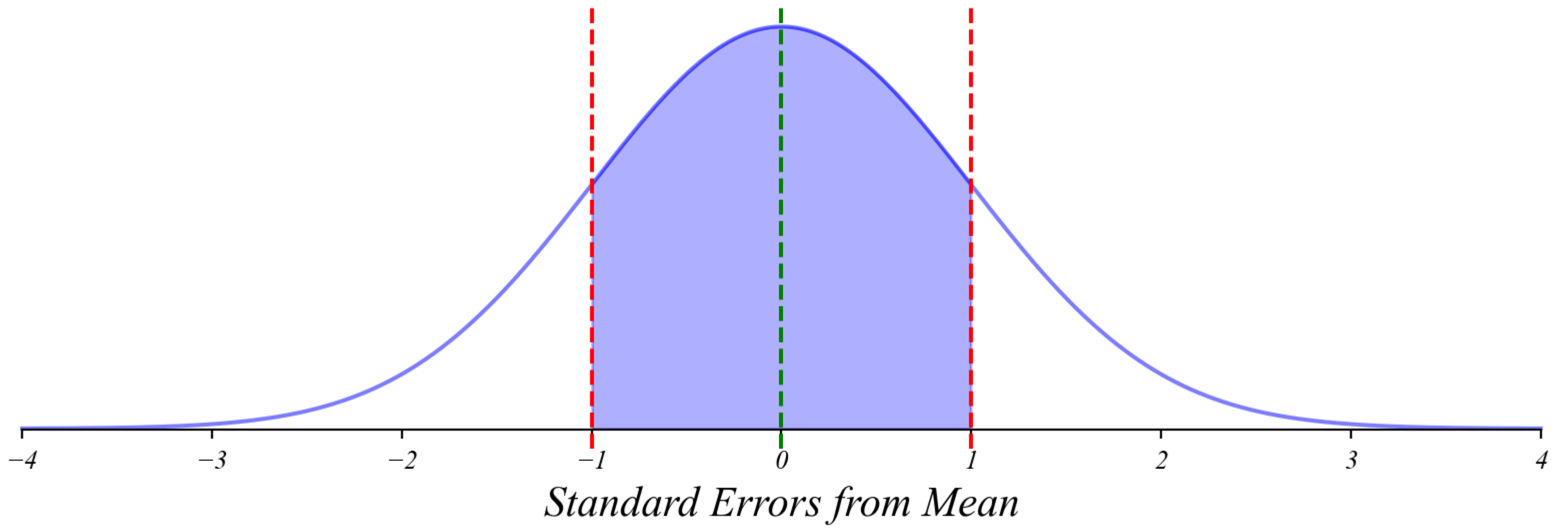
*2. Divide by* $n$ *to find that the mean of* $n$ *is* $\frac{\sigma^2}{n}$ $\left(nVAR(a) = VAR(n^2 a)\right)$

Therefore the standard error is $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

# Confidence Intervals

*If we know σ, we can calculate probabilities.*

> *what's the probability $\bar{x}$ is within one standard error of $\mu$?*




> $P(|\bar{x} - \mu| \leq \frac{\sigma}{\sqrt{n}}) \approx 0.68$

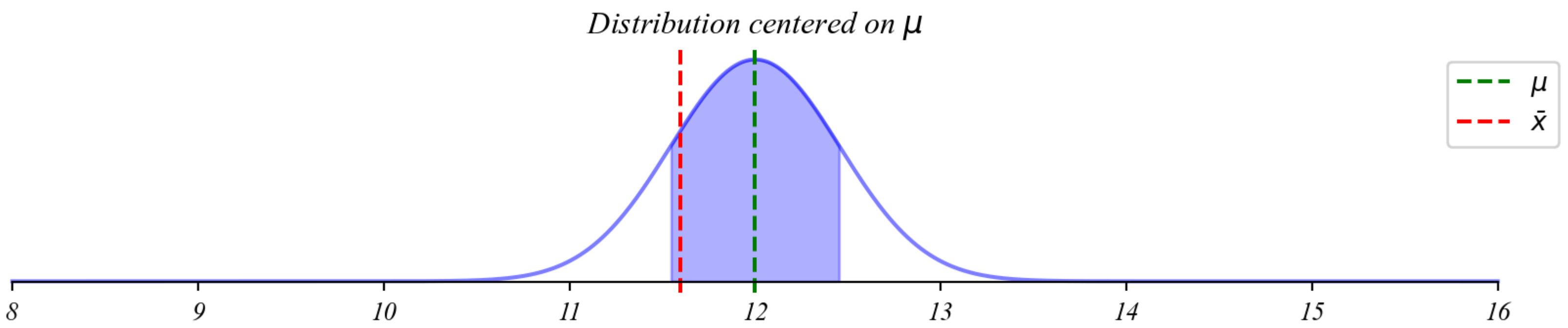
> *so 68% of the time $\bar{x}$ will fall within $[\mu - \frac{\sigma}{\sqrt{n}}, \mu + \frac{\sigma}{\sqrt{n}}]$*

> *we call $[\mu - \frac{\sigma}{\sqrt{n}}, \mu + \frac{\sigma}{\sqrt{n}}]$ a 68% confidence interval*

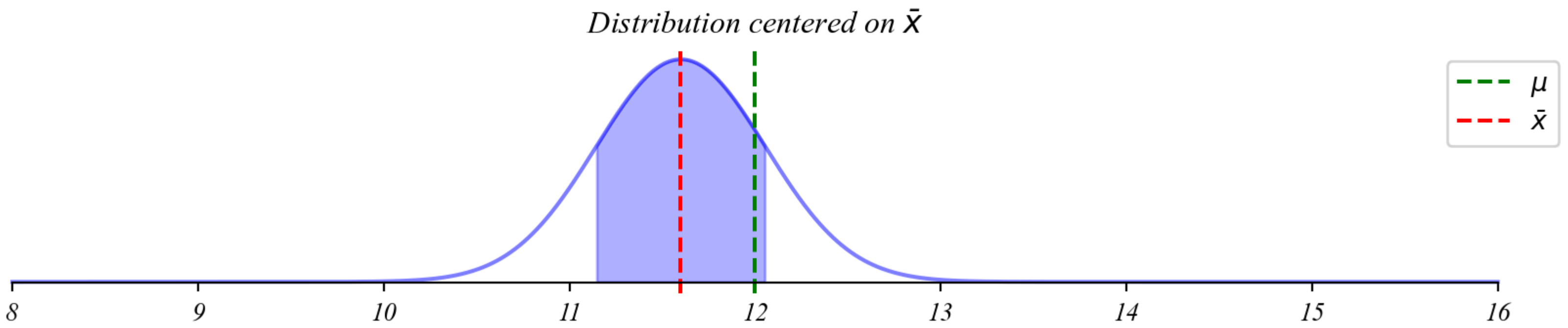
# Two Perspectives

*There are two mathematically equivalent perspectives to think about "closeness" between $\mu$ and $\bar{x}$.*

## Perspective 1: probability $\bar{x}$ is close to $\mu$

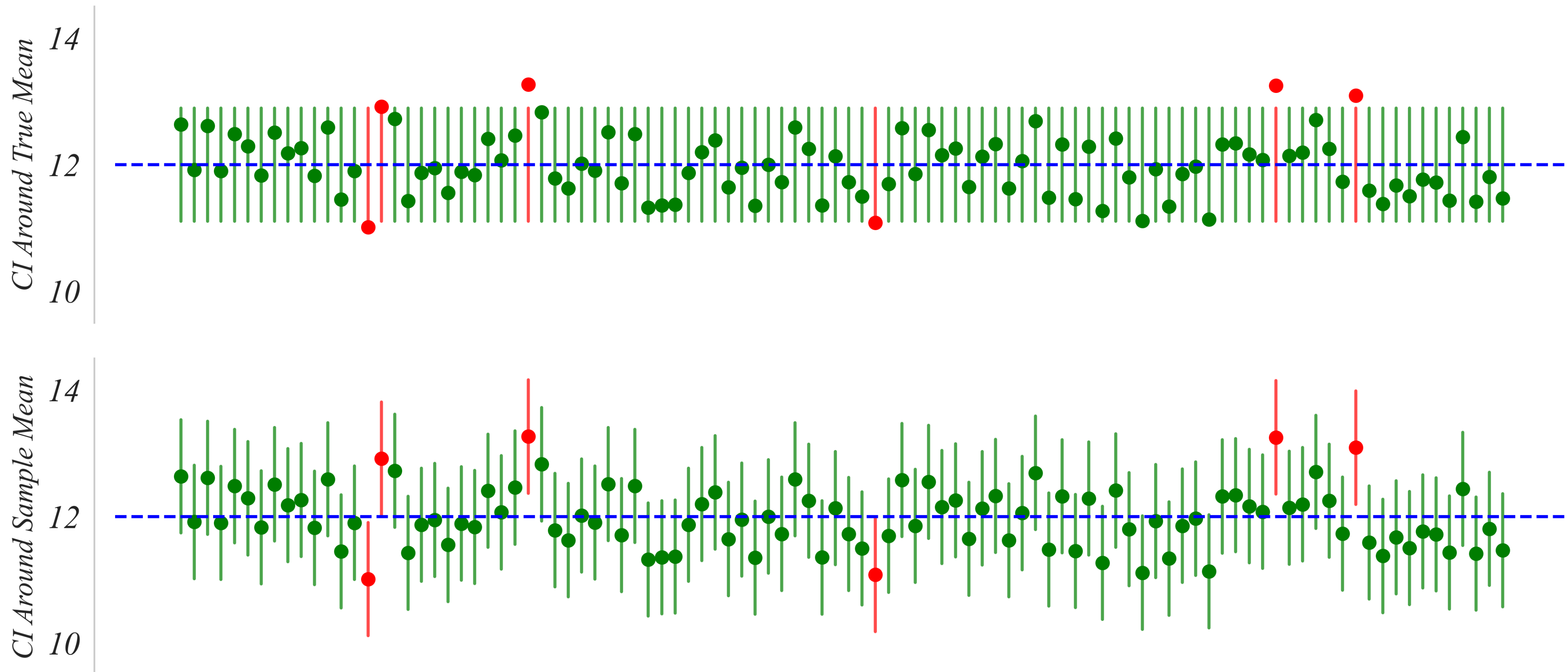



## Perspective 2: probability $\mu$ is close to $\bar{x}$



# Difference Center Points

*There are two mathematically equivalent perspectives to think about "closeness" between $\mu$ and $\bar{x}$.*




> *this is huge! we can center the confidence interval around $\bar{x}$ instead of $\mu$!*

# Using Confidence Intervals

*Example:* $\bar{x} = 102.3, \sigma = 1.6, \textit{and } n = 100$

**Question 1:** what's the probability $\mu$ is closer than 0.1 to $\bar{x}$?

```
distance = 0.1
se = sigma / np.sqrt(n)
probability = stats.norm.cdf(distance/se) - stats.norm.cdf(-distance/se)
```

**Question 2:** what's the 95% CI?

```
se = sigma / np.sqrt(n)
ci = stats.norm.interval(0.95, loc=x_bar, scale=se)
```

# One Problem Remains

*We don't know σ either!*

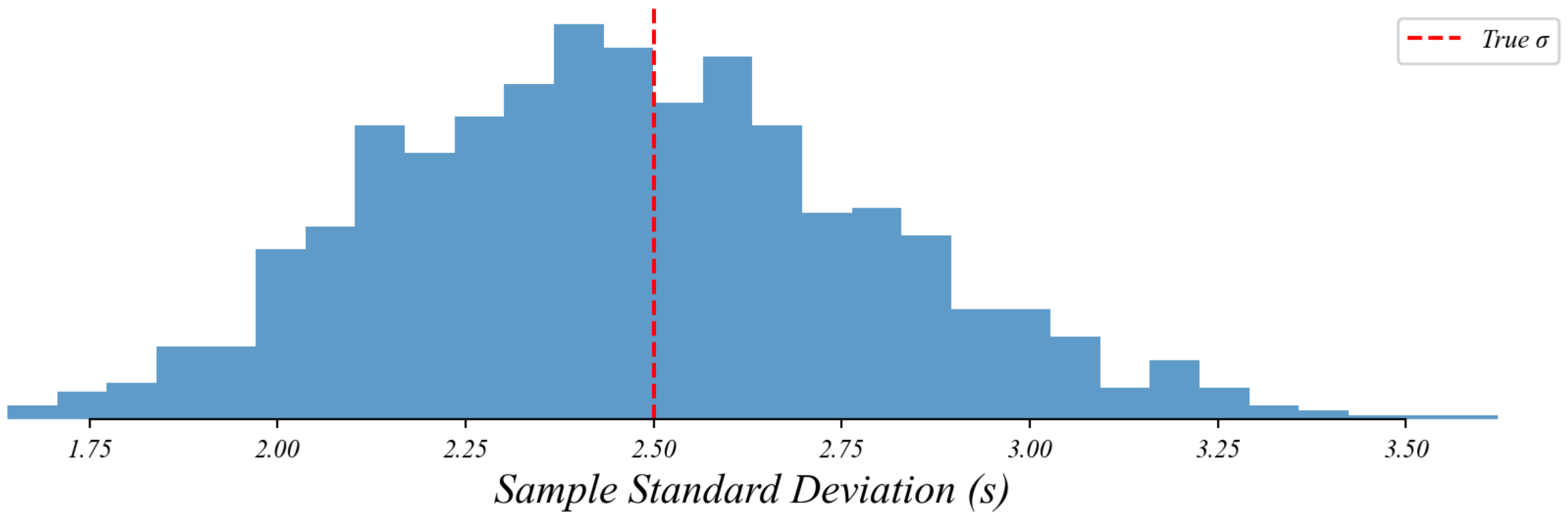
> *we used $\bar{x}$ to estimate $\mu$*

> *can we use s to estimate $\sigma$?*

> *yes, but there's a catch…*

# Using $s$ Instead of $\sigma$

*Sample standard deviation (s) has its own sampling variability.*




> *this adds extra uncertainty to our interval*

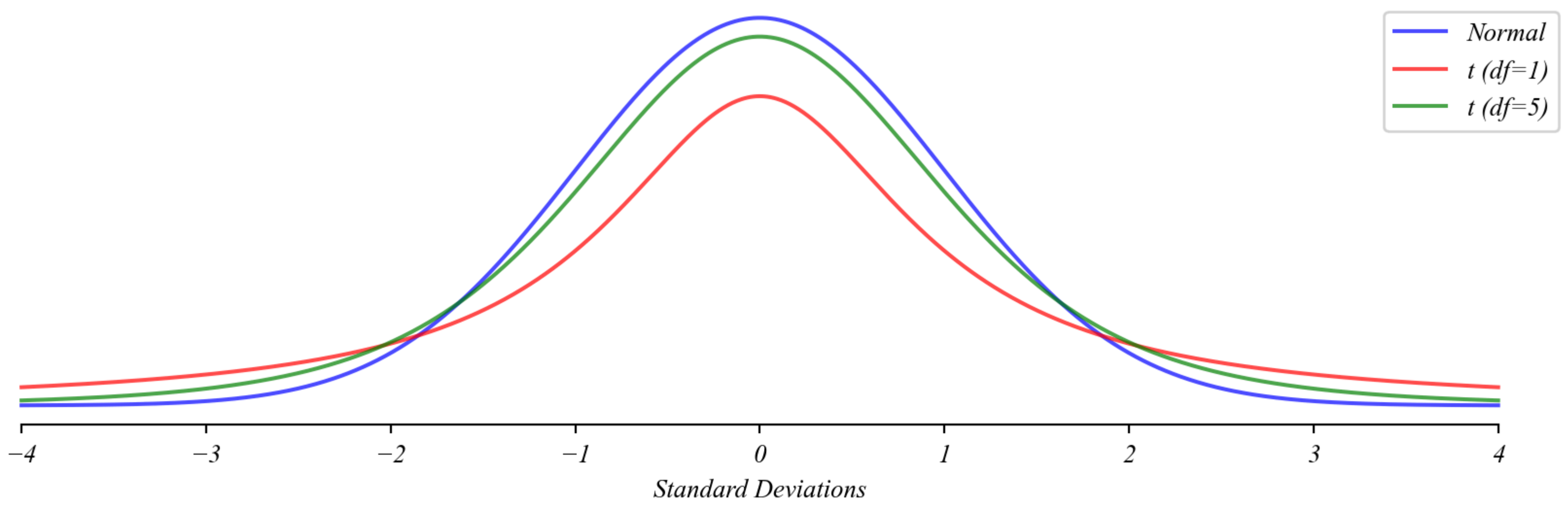
# Normal vs t-Distribution

*The t-distribution precisely accounts for the variation in $s$ around $\sigma$.*

> $\bar{x}$ follows a normal distribution with $\mu$ and $\sigma$

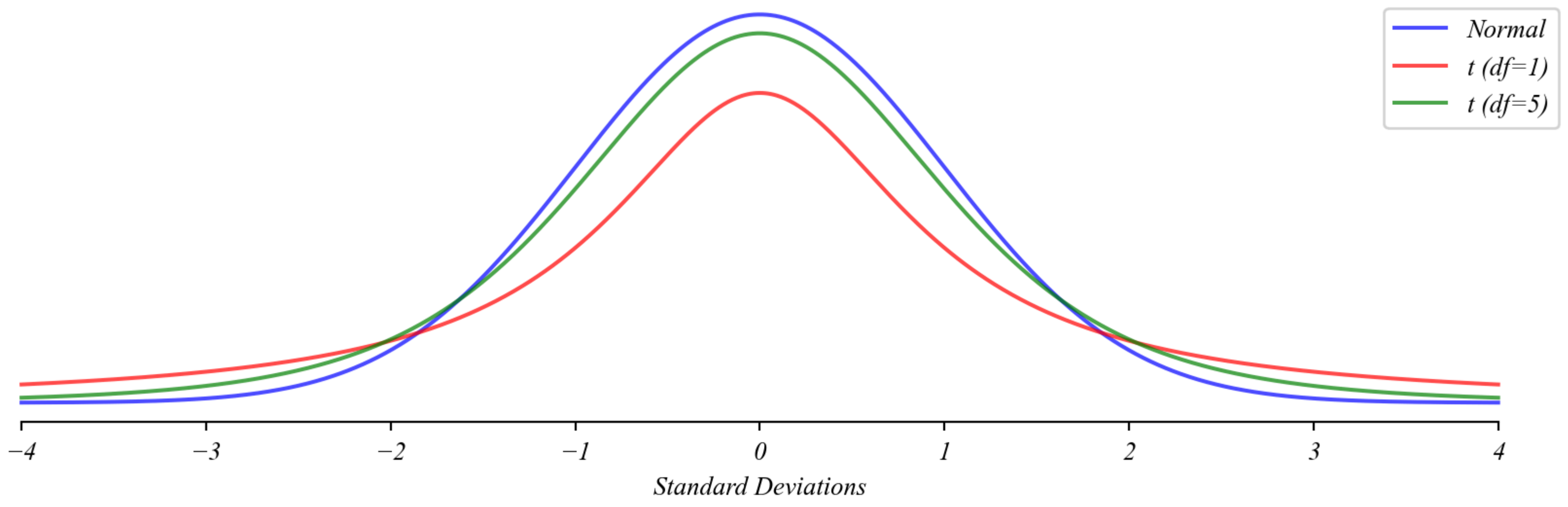
> key insight: since $s$ is random, using it instead $\sigma$ introduces another r.v.

> this gives us the t-distribution with n-1 degrees of freedom



# The t-Distribution

*... acounts for the extra uncertainty in $s$ around $\sigma$.*




*> t-distribution has heavier tails than normal*

*> approaches normal as sample size ($n$) increases*

# Putting It All Together

*Now we can quantify our uncertainty about an unknown $\mu$.*

*1. $\bar{x}$ follows a normal distribution around $\mu$.*

*2. We can center the distribution on $\bar{x}$ instead.*

*3. Using $s$ adds uncertainty, captured by t-distribution.*

*4. We can use the t-distribution to make probability statements about $\mu$.*

# Example: Wait Times

*Calculate the 95% confidence interval for waiting times.*

Generate some sample data.

```
sample = np.random.normal(12, 2.5, 30)
```

Calculate sample statistics.

```
x_bar = np.mean(sample)
s = np.std(sample, ddof=1)
n = len(sample)
se = s / np.sqrt(n)
```

Calculate how many standard errors the 95% CI is from $\bar{x}$.

```
t_crit = stats.t.ppf(0.975, n-1)
```

Calculate the CI from the critical value.

```
margin = t_crit * se
ci = [x_bar - margin, x_bar + margin]
```

> *if we took many samples, 95% of the time this interval would contain the truth*

> *we often just say: "we're 95% confident the truth is in this interval"*

# Extra Questions

1. *How would the confidence interval change if we:*
   - *Increased sample size?*
   - *Wanted 99% confidence instead?*
   - *Had a more variable population?*
2. *Why use t-distribution instead of normal?*
3. *What does "95% confident" really mean?*
4. *How could this help with economic decision-making?*