ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 3.2 | Sampling, Sample Means, Central Limit Theorem

A Big Question

If all we see is the sample, how do we learn about a population?

> *in general, we can't know a population distribution*

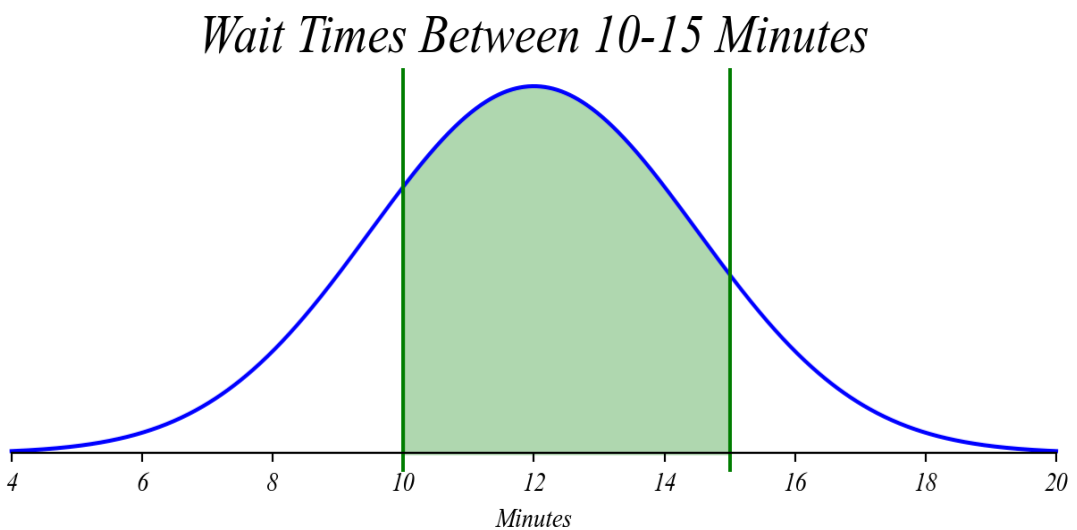
> *in general, we only see a sample*

> *if we only see a sample, what* can *we say about the population?*

Known Parameters

Example: Normally distributed wait times

- *Probability wait time < 10:*
  - $P(X < 10) = 0.21$
- *Probability wait time > 15:*
  - $P(X > 15) = 0.11$
- *Probability between 10 - 15:*
  - $P(10 < X < 15) = 0.59$

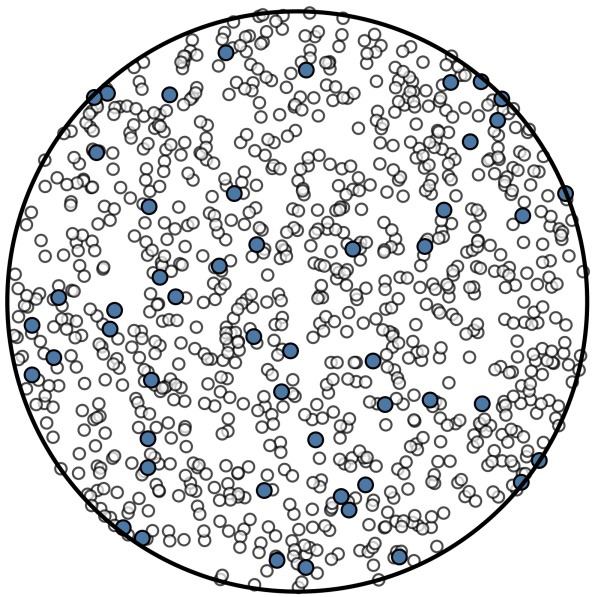


> with known parameters, we can calculate everything exactly

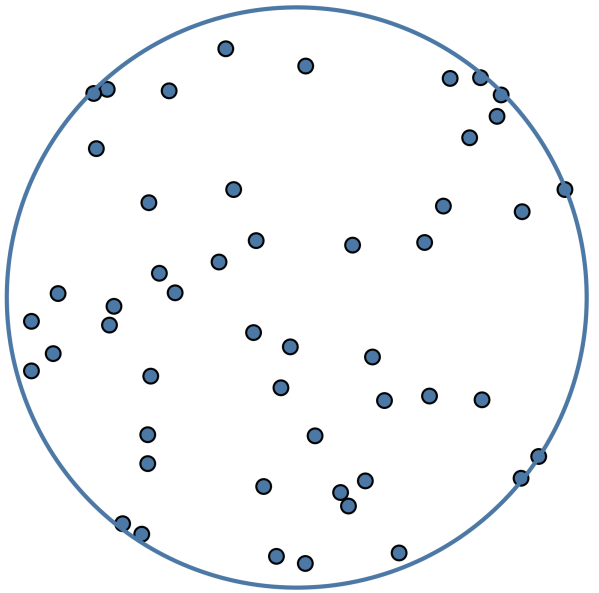
Sampling Error

But if all we see is the sample, how do we learn about a population?

Population (μ=?; σ=?)



Sample ($n = 50$; $\bar{x}$; S)



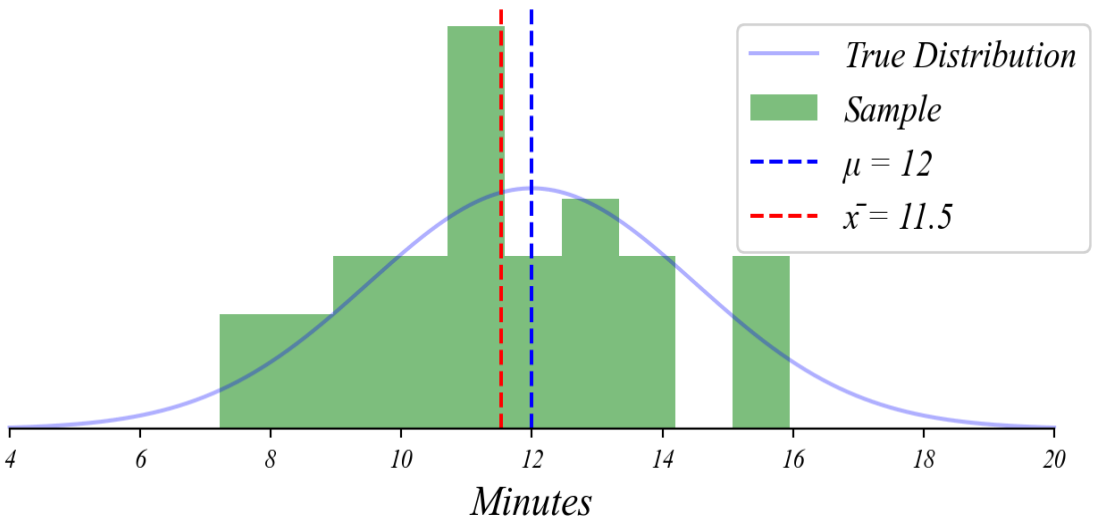
> *how do we learn about* μ *if all we have is* n, $\bar{x}$, *and* S?

Population vs Sample

Question: How close are $\bar{x}$ and s to μ and σ?

Instead of μ and σ, we know:

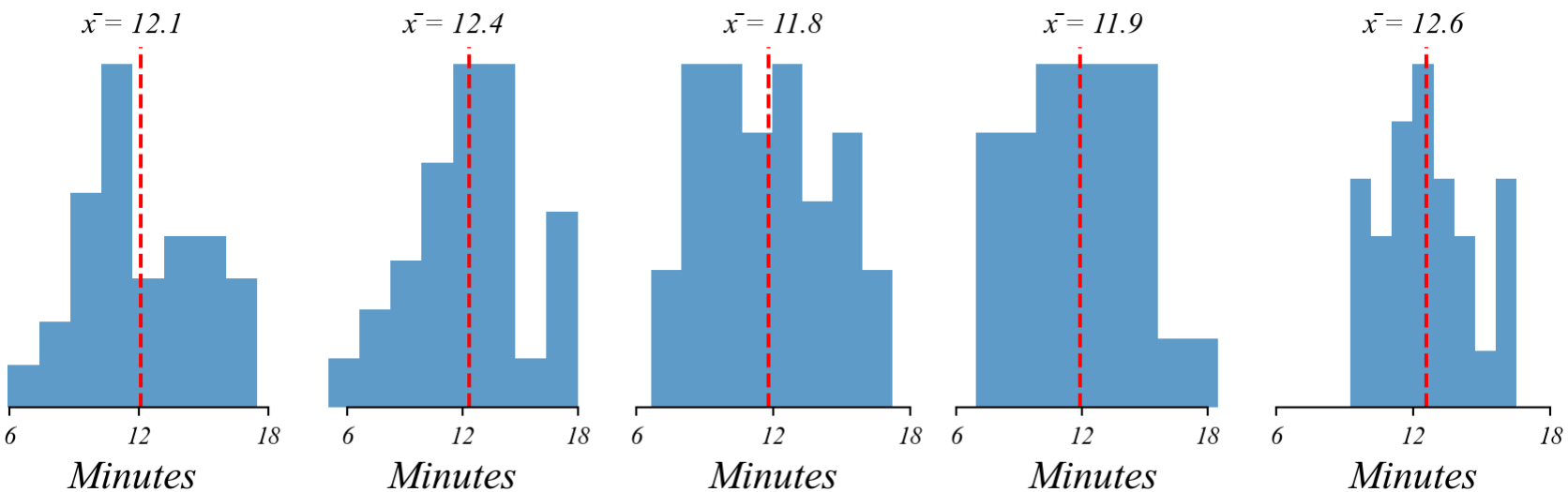
- *A sample of observations*
- *Sample statistics ($\bar{x}$, s)*



> we want to systematically measure the "closeness" of the sample/population

Sampling Error

If we take multiple samples, we get different sample means.

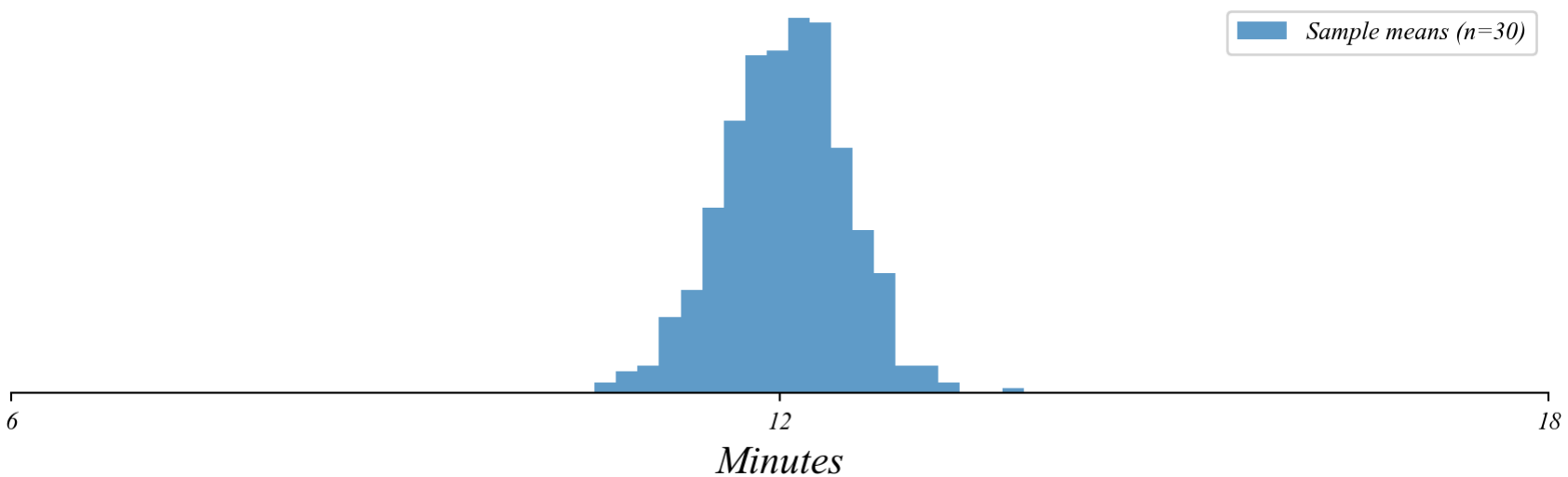


> *each sample gives us a different estimate of the population mean*

> *lets plot these sample means ($\bar{x}$)*

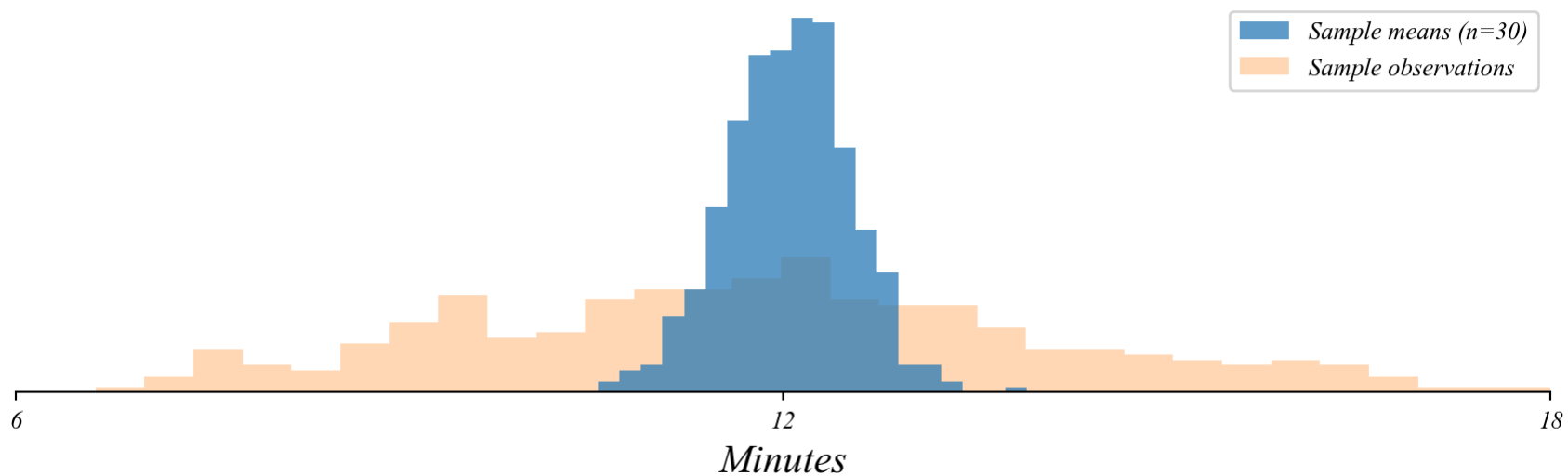
Sampling Error

If we take multiple samples, their means will vary.



Sampling Error

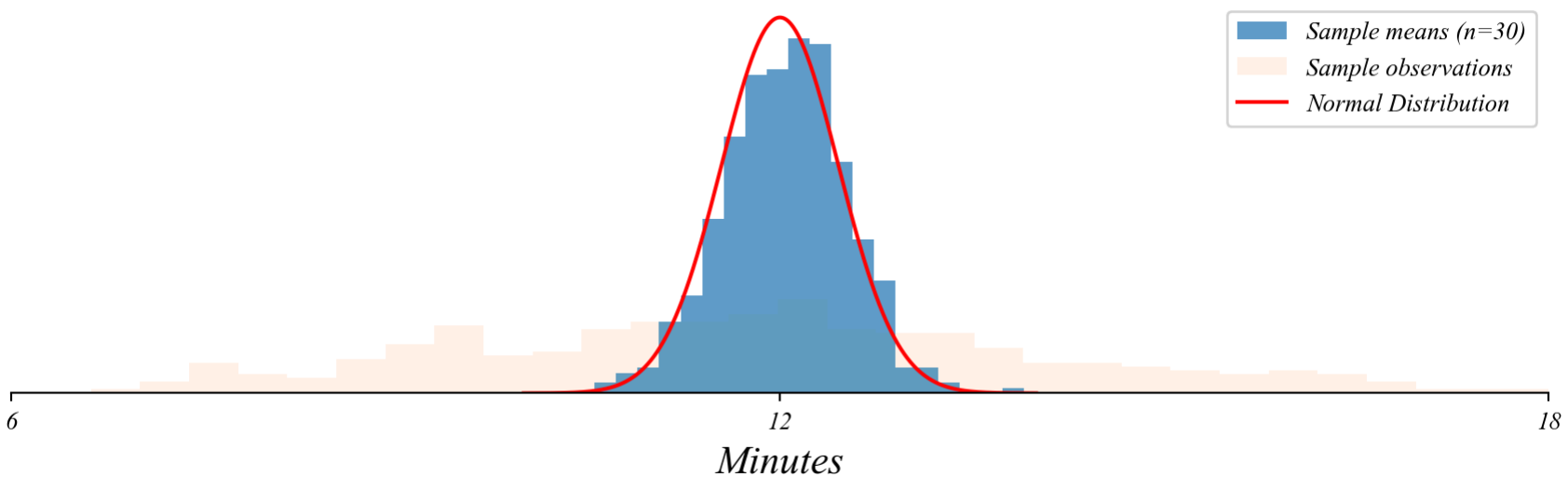
If we take multiple samples, their means will vary, and by much less than the original distribution.



> why? think about rolling two dice… it's much less likely to get a 1 and a 7

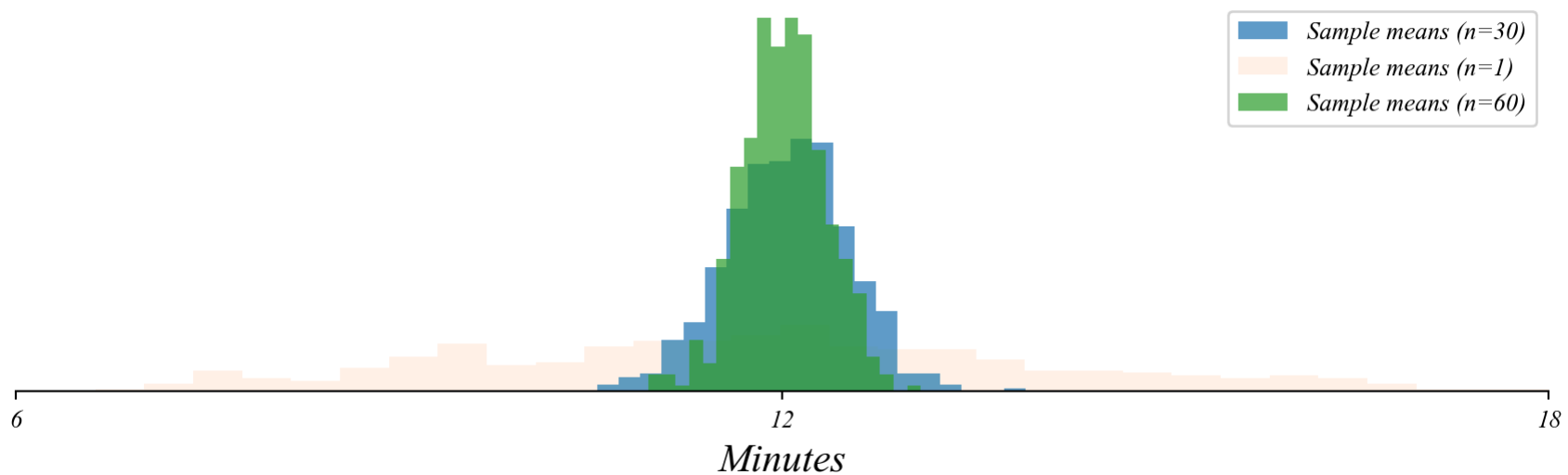
Sampling Error

If we take multiple samples, their means will follow a normal distribution.



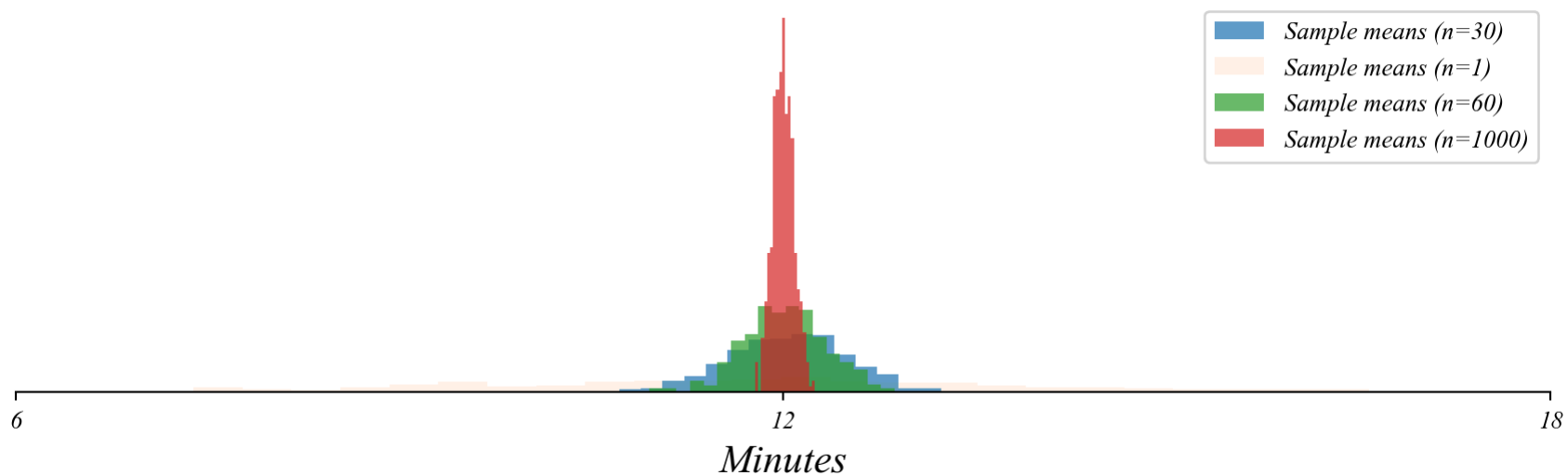
Sampling Error

If we take multiple samples, their distribution is narrower with larger sample sizes.



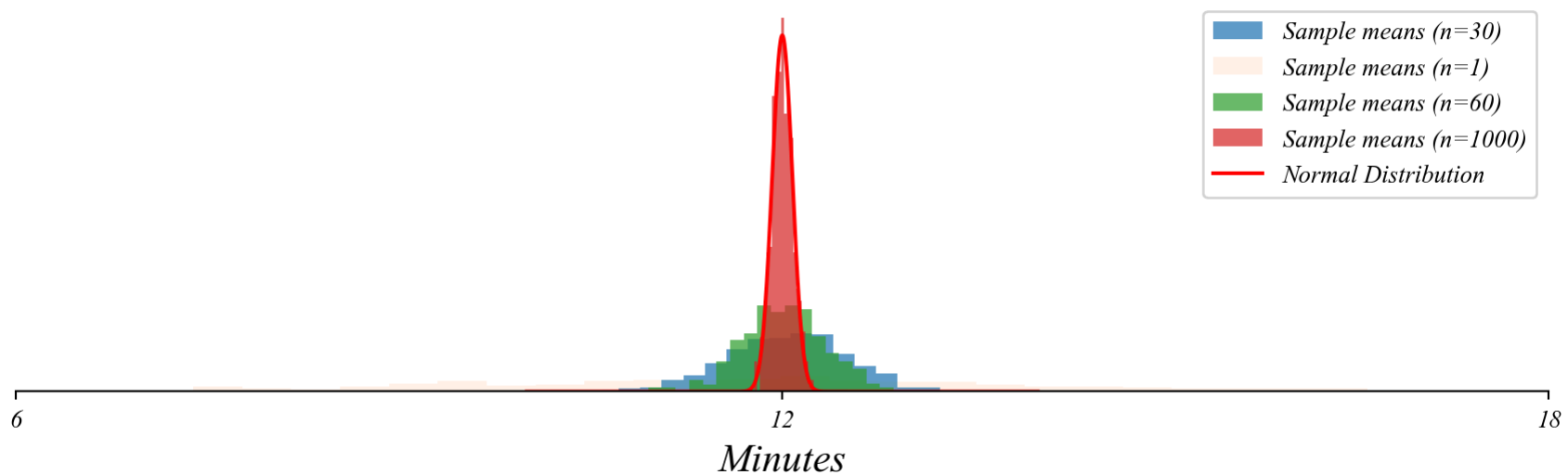
Sampling Error

If we take multiple samples, their distribution is narrower with larger sample sizes.



Sampling Error

If we take multiple samples, their distribution is narrower with larger sample sizes.



The Central Limit Theorem

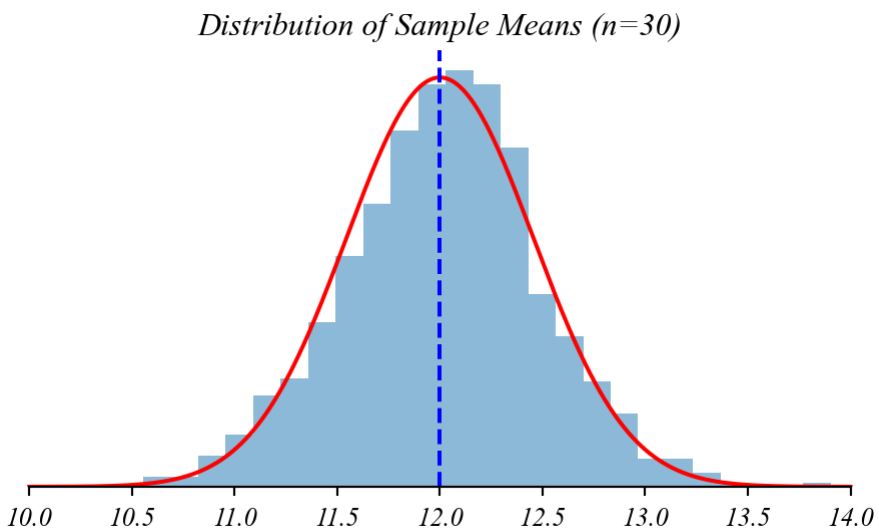
The distribution of sample means approximates a normal distribution as sample size increases, regardless of the population's distribution.

Key insights:

- *Sample means cluster around μ*
- *Standard error = $\sigma/\sqrt{n}$*
- *Normal shape emerges*

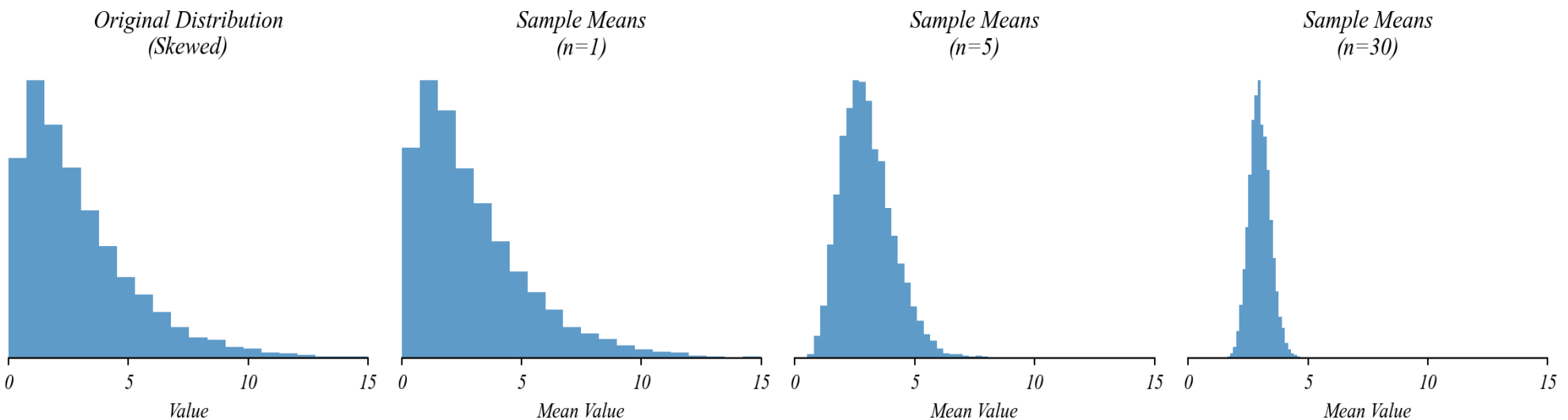
Implications:

- *We can predict the behavior of $\bar{x}$*
- *This works for ANY distribution*



CLT: The Magic of Large Samples

The CLT works for any distribution shape as sample size increases.



> *as n increases, the sampling distribution becomes more normal!*

> *and **THIS** is something we know and can work with*