ECON 0150 | Economic Data Analysis

The economist's data analysis skillset.

Part 3.2 | Sampling and the Central Limit Theorem

A Big Question

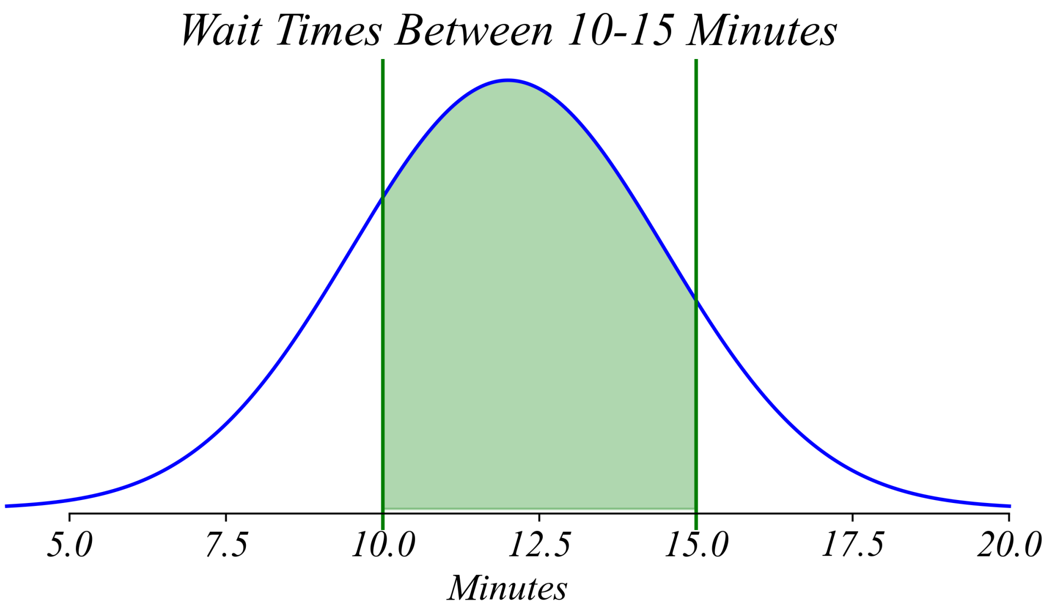
If all we see is the sample, how do we learn about a population?

- *In general, a population's random variables will be unobservable.*
- *If we only see a sample, what can we say about the population?*

Random Variables: Known

If we know the random variable, we can learn many things about the population.

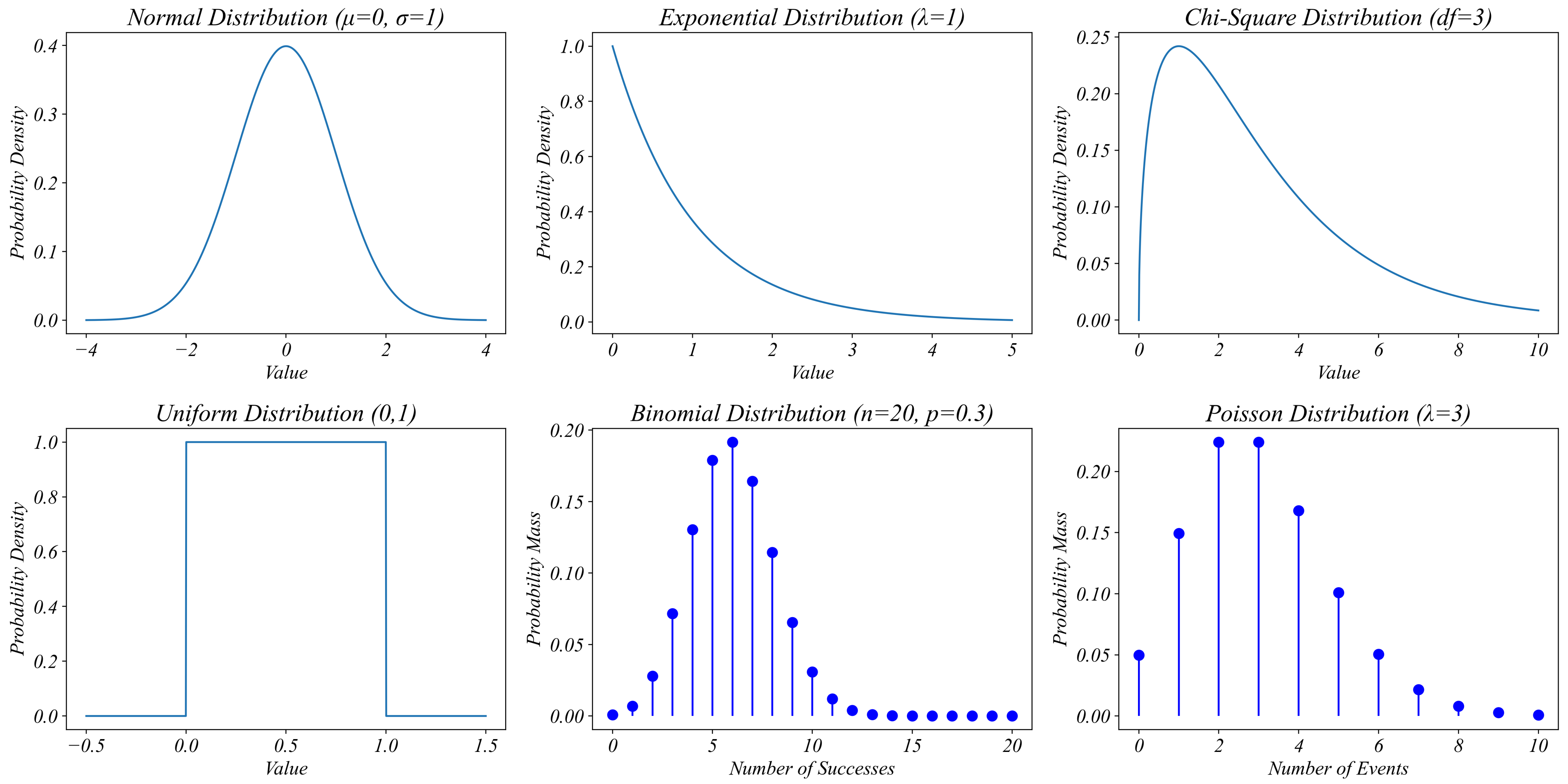
- *Probability wait time < 10:*
  - $P(X < 10) = 0.21$
- *Probability wait time > 15:*
  - $P(X > 15) = 0.11$
- *Probability between 10 - 15:*
  - $P(10 < X < 15) = 0.59$



> when we know the probability function, we can calculate everything exactly

Random Variables: Known

If we know the random variable, we can learn many things about the population.

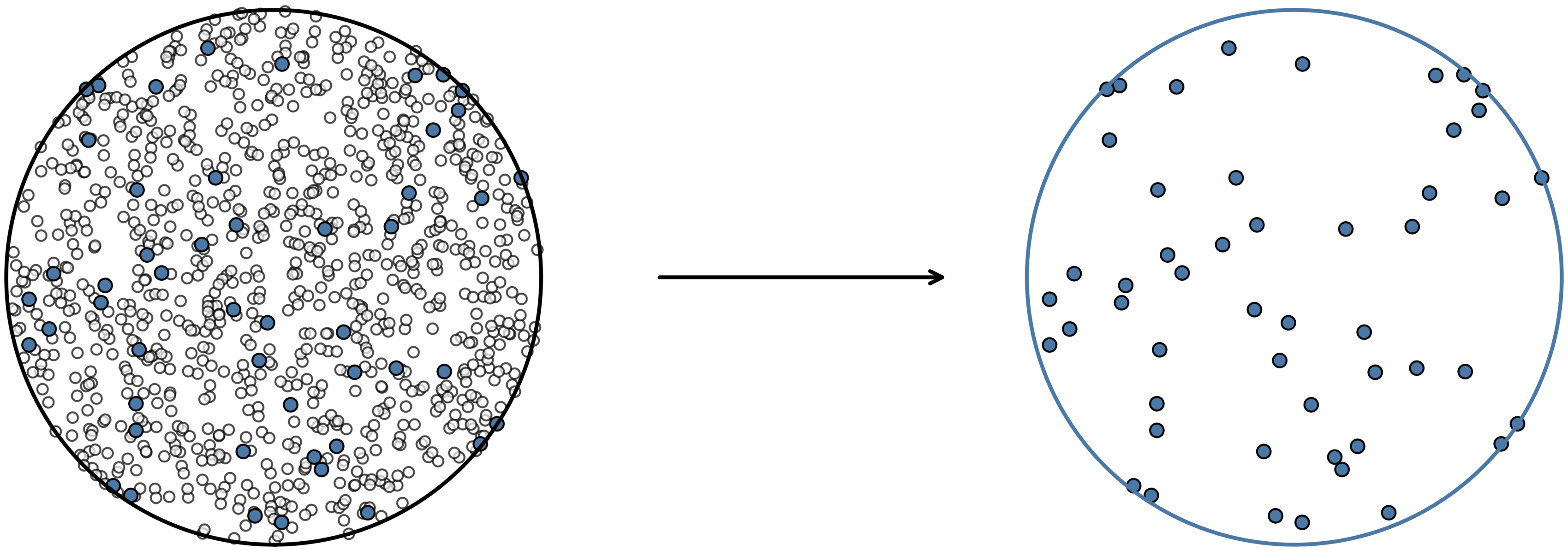


> but what can we know about the population if we only see the sample?

Random Variables: Unknown

But if all we see is the sample, what can we know 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?*

Exercise 3.2 | Sampling Dice (n=1)

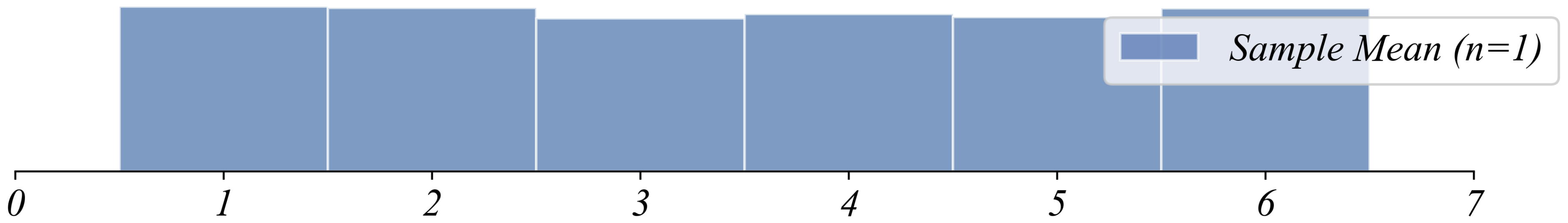
Let's pretend we don't know the probability function for dice.

Let's start with something simple.

1. Roll a die once (sample size: n=1).
2. We'll plot the distribution of our samples.

Exercise 3.2 | Results (n=1)

Your samples have a lot of variability!



> this variability perfectly matches what we would expect from a fair die

Exercise 3.2 | Sampling Dice (n=2)

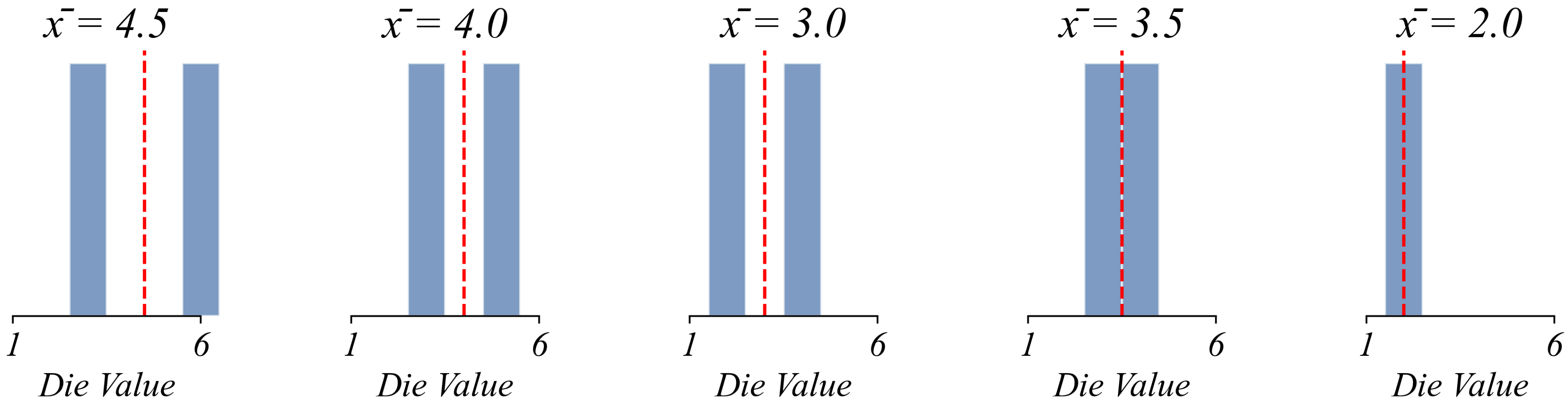
Now take a sample of two rolls and compute the mean.

Next is something slightly less boring.

1. Roll a die twice (sample size: n=2).
2. Calculate the mean of your two rolls.
3. We'll plot the distribution of your sample means.

Exercise 3.2 | Results (n=2)

Each sample has a slightly different sample mean.

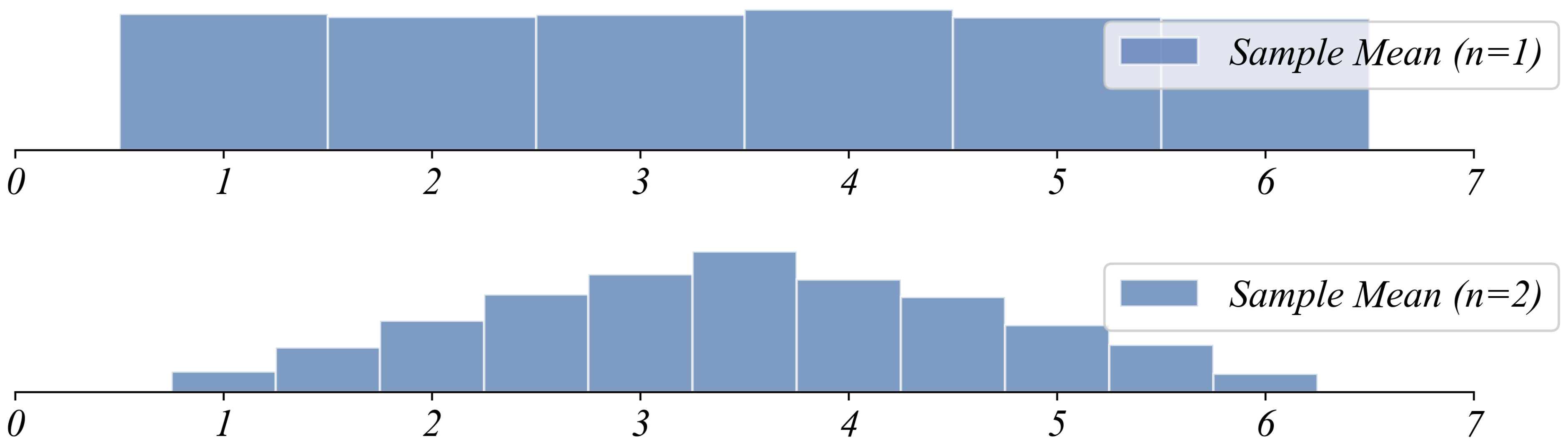


> *there's a lot of variability in your sample means!*

> *what do you expect to see when we plot these sample means ($\bar{x}$)?*

Exercise 3.2 | Results (n=2)

The distribution of sample means bunches in the middle.



> our sample means are more bunched (like a pyramid) in the middle! why?

> there are more ways to get 7/2 than 2/2!

Exercise 3.2 | Sampling Dice (n=3)

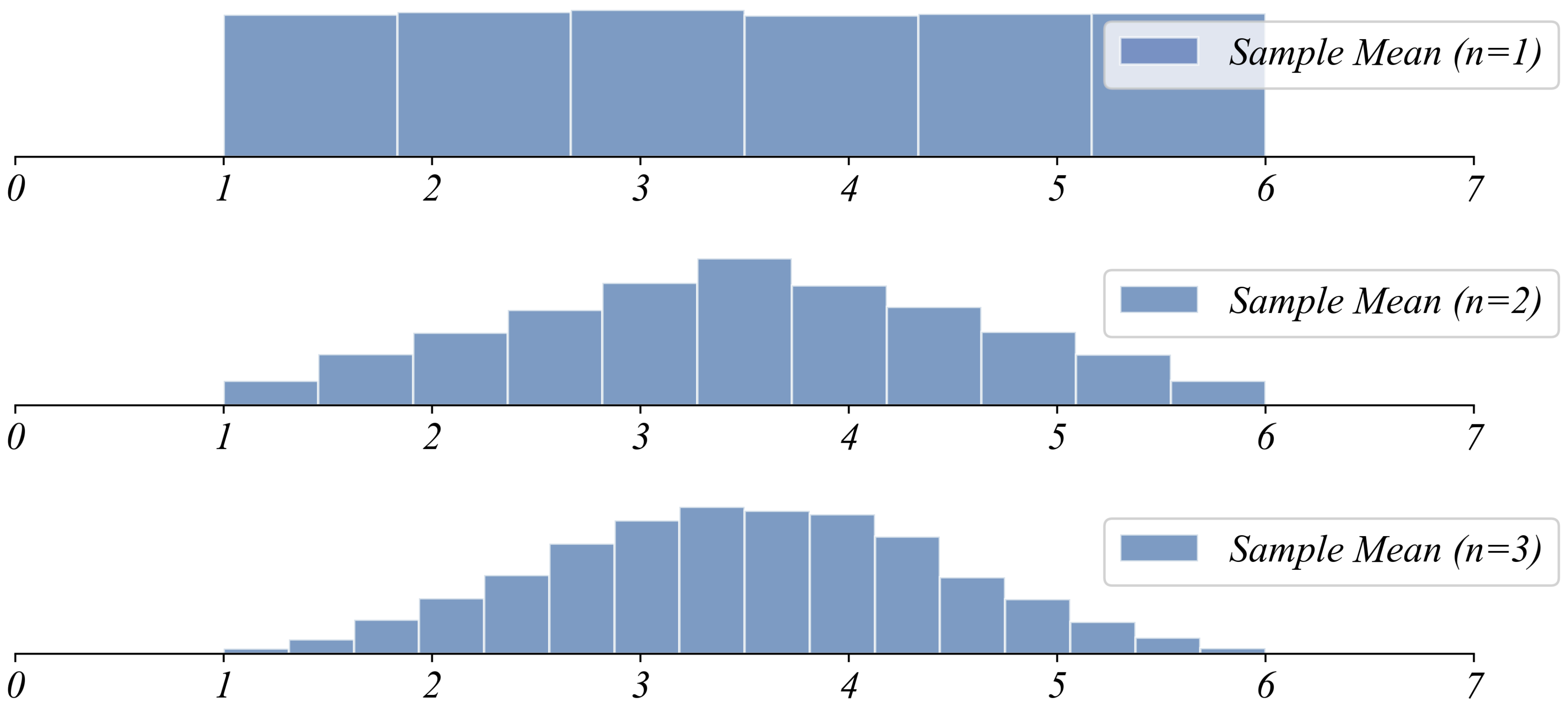
Now take a sample of three rolls and compute the mean.

Next is something even less boring.

1. *Roll a die three times (sample size: n=3).*
2. *Calculate the mean of your three rolls.*
3. *We'll plot the distribution of your sample means.*

Exercise 3.2 | Results (n=3)

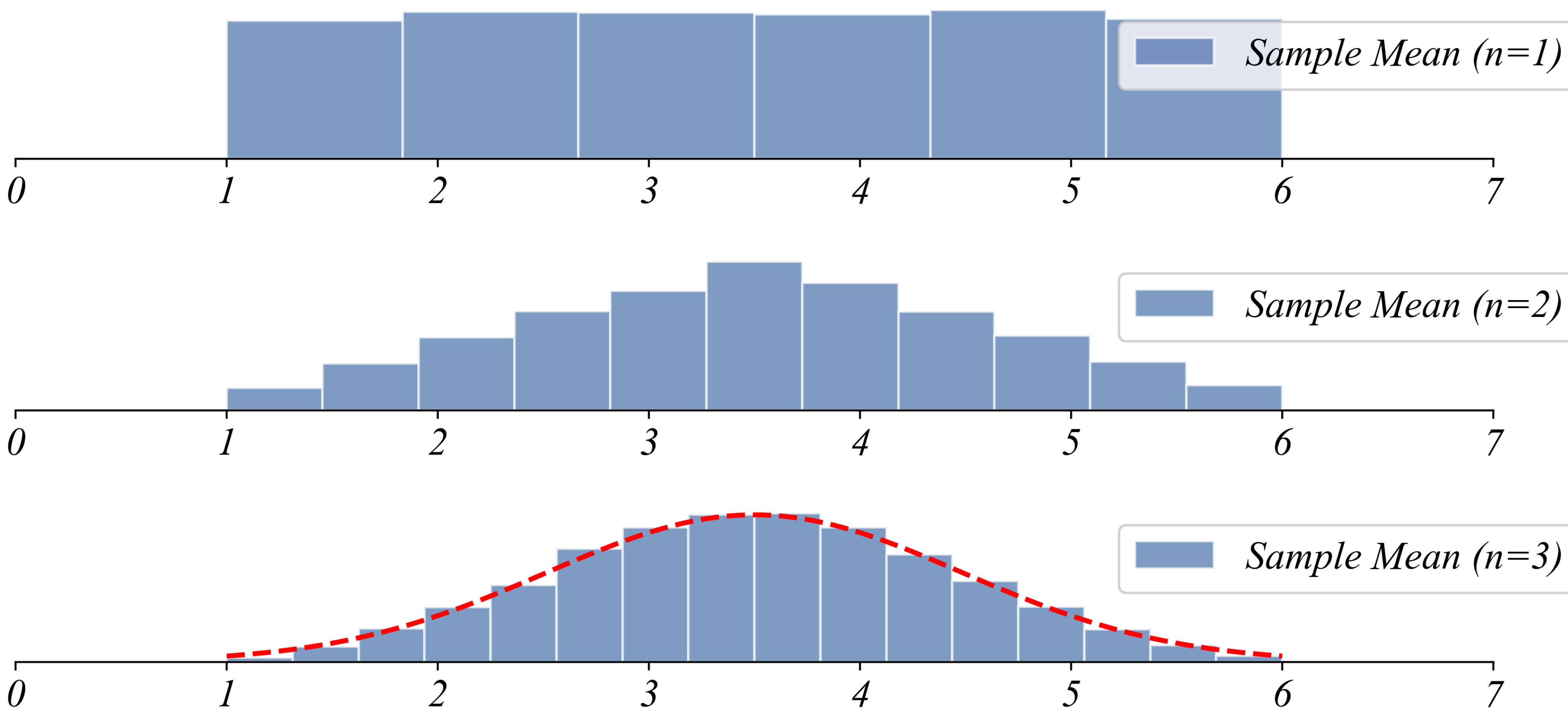
The distribution of sample means with n=3.



> *what do you notice about the shape with n=3?*

Exercise 3.2 | Results (n=3)

The distribution of sample means with n=3.



> there's some curvature to the shape — the edges are rounding into a curve

Exercise 3.2 | Sampling Dice (n=30)

Now let's really increase the sample size.

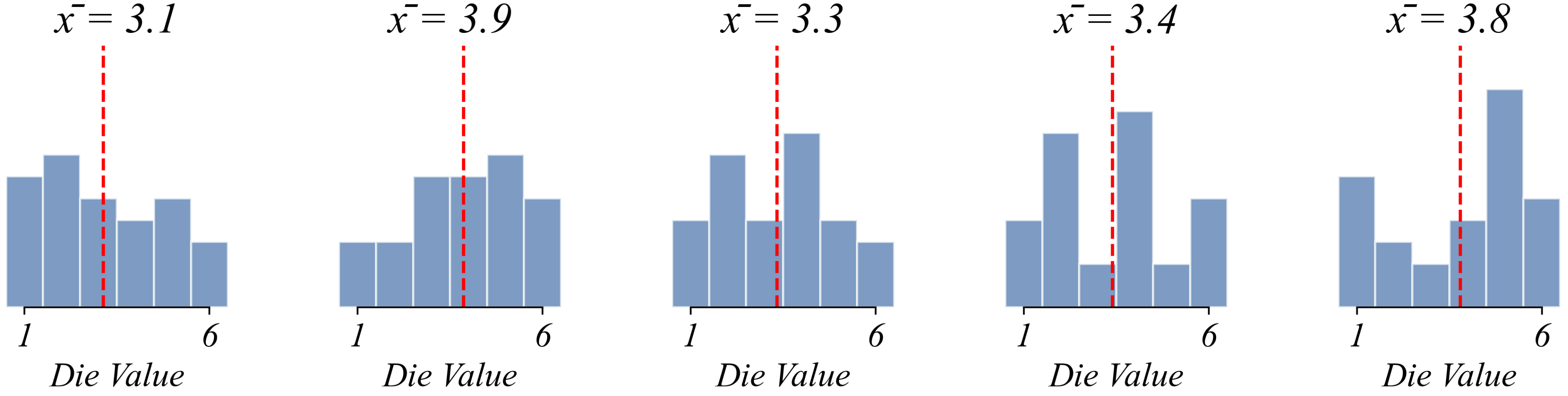
Next is something very un-boring.

1. Roll a die thirty times (sample size: n=30).

2. We'll simulate this 1,000 times and plot the distribution of sample means.

Exercise 3.2 | Results (n=30)

Your individual samples each look different.

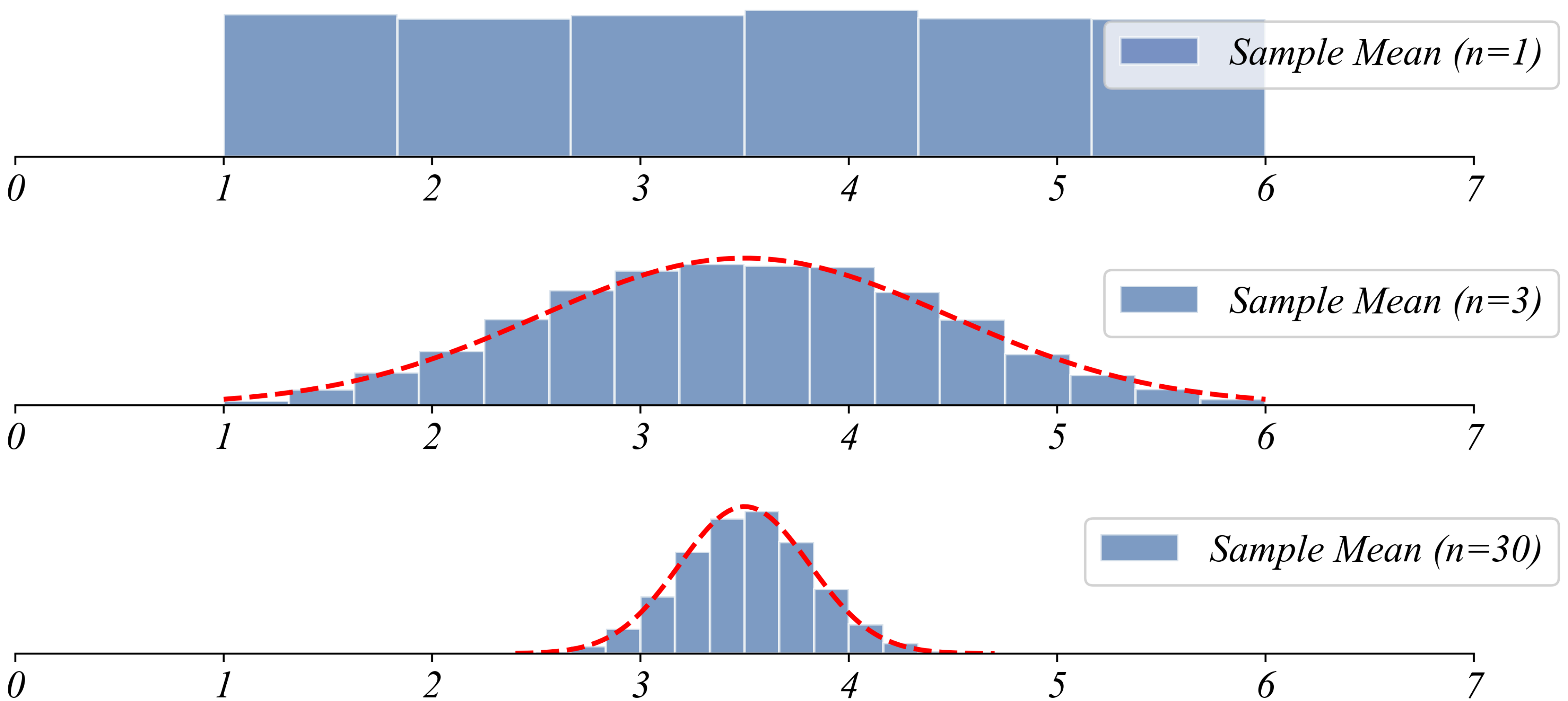


> *there are even more ways your sample could look!*

> *what do you expect to see when we plot these sample means ($\bar{x}$)?*

Exercise 3.2 | Results (n=30)

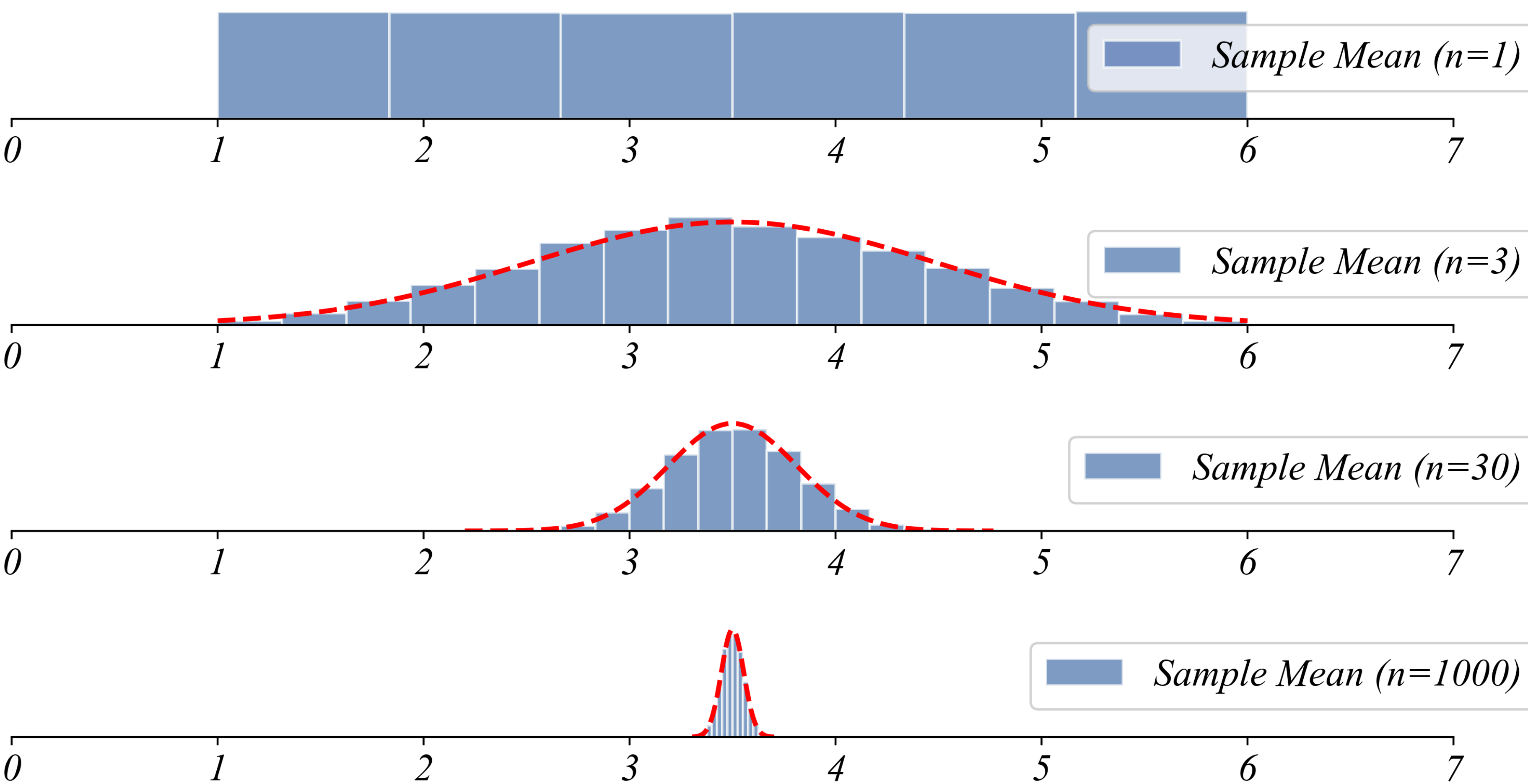
What happens when we really increase the sample size?



> the distribution of sample means gets tighter and more bell-shaped

Exercise 3.2 | Results (n=30)

What happens when we really increase the sample size?

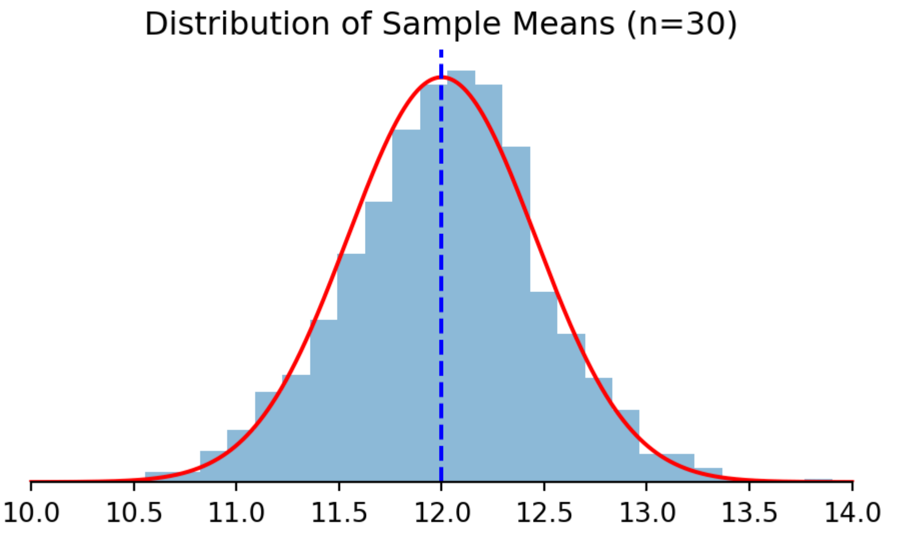


> what is this probability function in red?

The Central Limit Theorem

The distribution of sample means approximates a normal distribution as sample size increases.

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



1. ***Shape***: *the sampling distribution is normal*
2. ***Center***: *it's centered on the population mean* μ
3. ***Spread***: *the standard error* $\sigma/\sqrt{n}$ *shrinks with larger* n

The Standard Error

Where does $\sigma/\sqrt{n}$ come from?

Each observation x_i is drawn independently with variance σ^2, so:

$$Var(x_1 + x_2 + \cdots + x_n) = n\sigma^2$$

Dividing by n divides the variance by n^2:

$$Var\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Take the square root:

$$SD(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Skewed Distributions

Does the CLT work for distributions that aren't as nice?

Question: *Does the CLT still work when the population looks asymmetric?*

Exercise 3.2 | Skewed Population

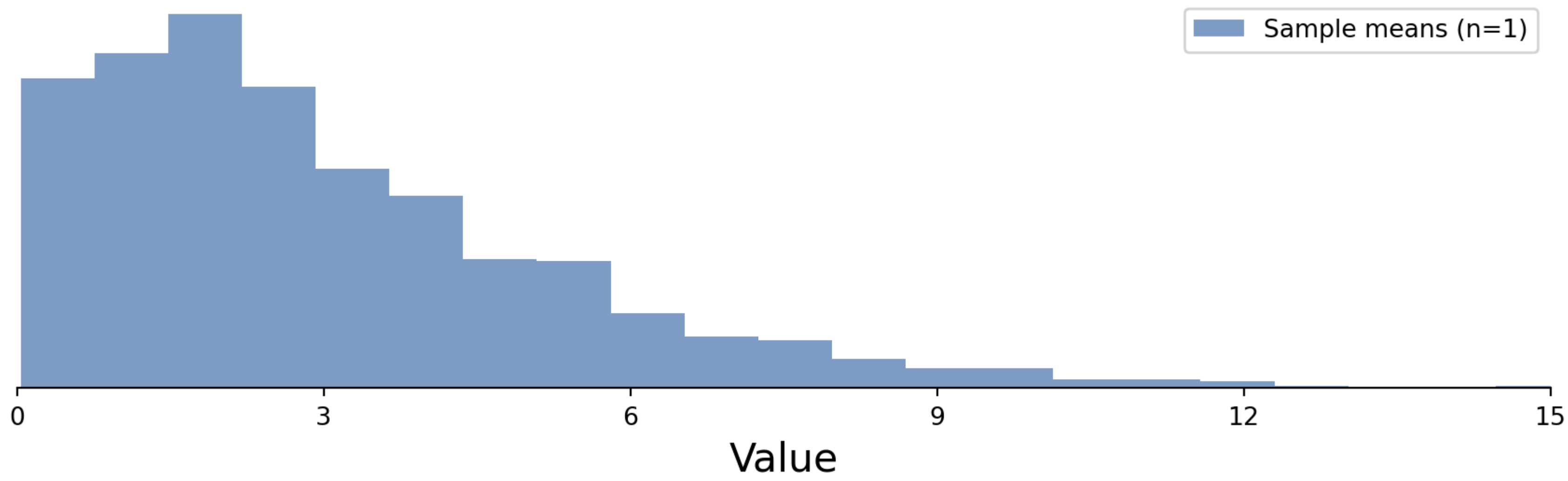
Simulate 1,000 sample means from a chi-squared population with n=1.

```python
# Simulate 1000 sample means from a skewed population
samples = stats.chi2.rvs(df=3, size=(1000, 1))
sample_means = samples.mean(axis=1)
sns.histplot(sample_means, bins=30)
```

> *with n=1, the sample means are just the raw observations*

Exercise 3.2 | Skewed Population (n=1)

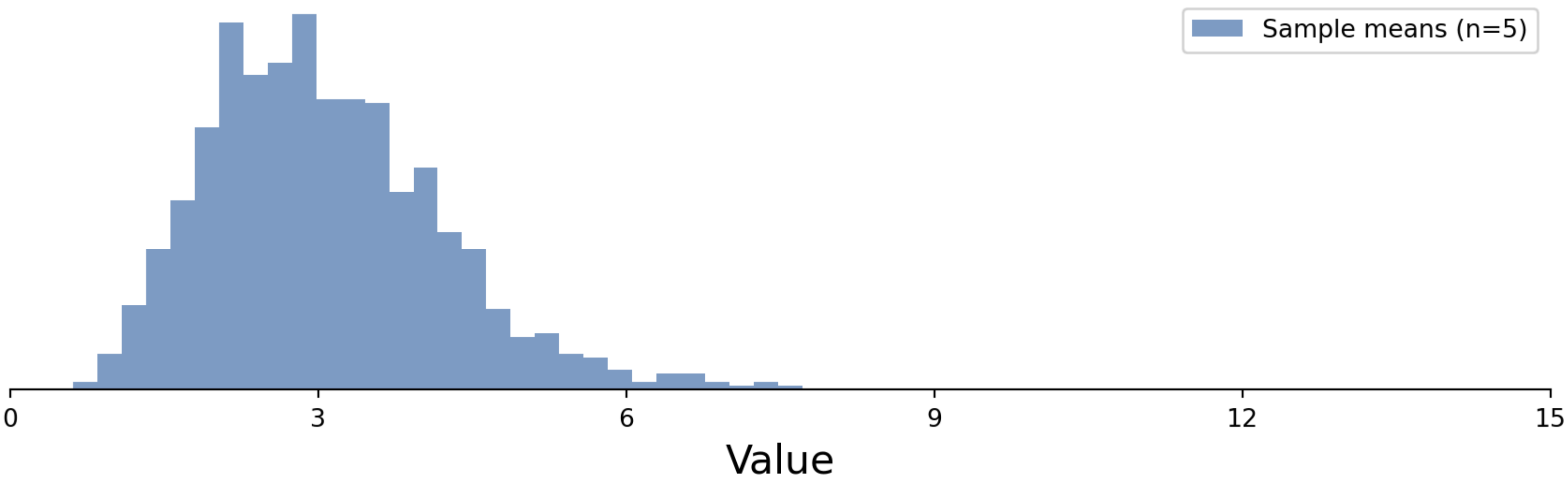
The distribution of sample means looks just like the population — very skewed.



> now increase the sample size to n=5

Exercise 3.2 | Skewed Population (n=5)

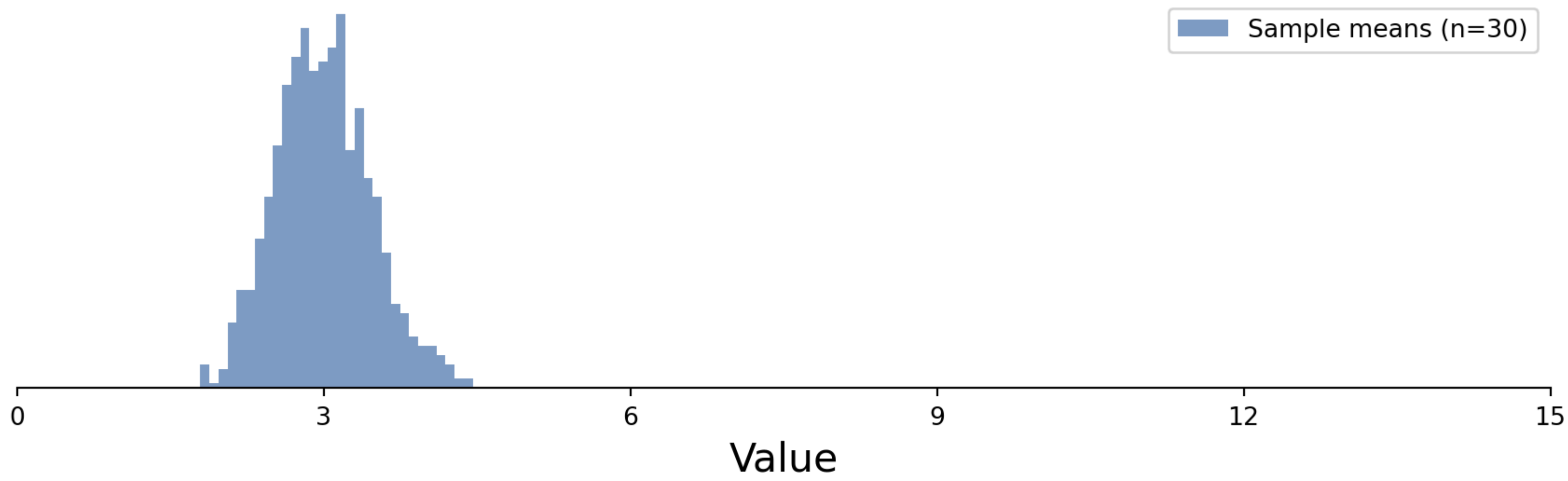
The skew is already diminishing.



> now increase to n=30

Exercise 3.2 | Skewed Population (n=30)

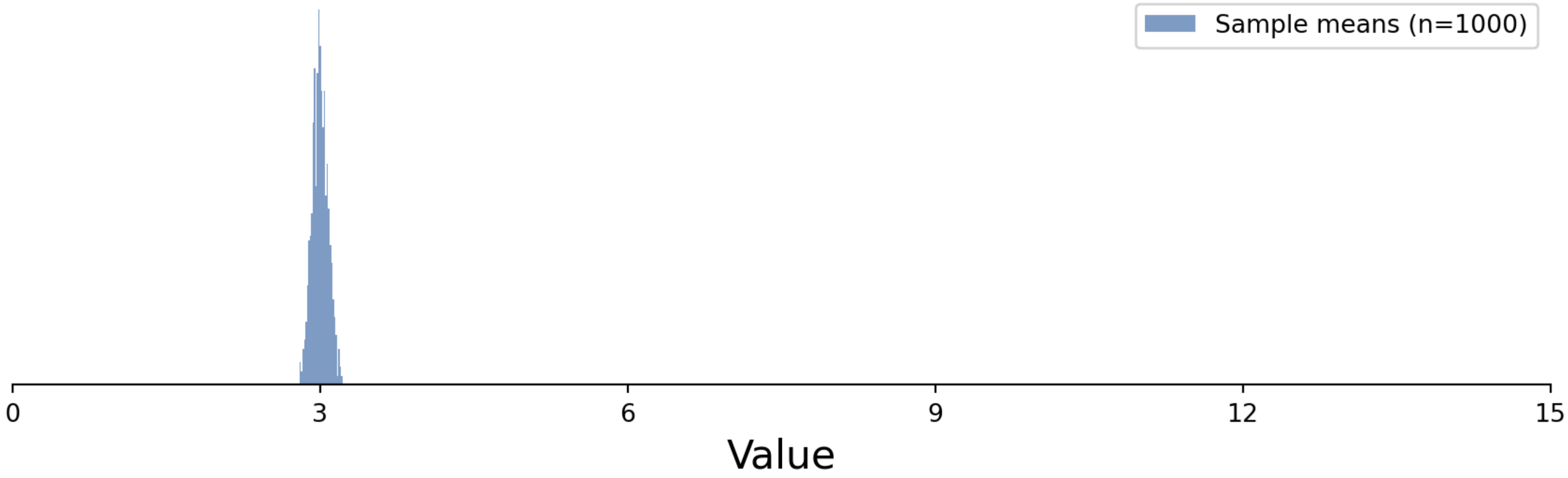
It looks normal — despite the skewed population.



> now increase to n=1000

Exercise 3.2 | Skewed Population (n=1000)

Very tight, very normal.



> *the skew has completely disappeared*

Exercise 3.2 | Skewed Population

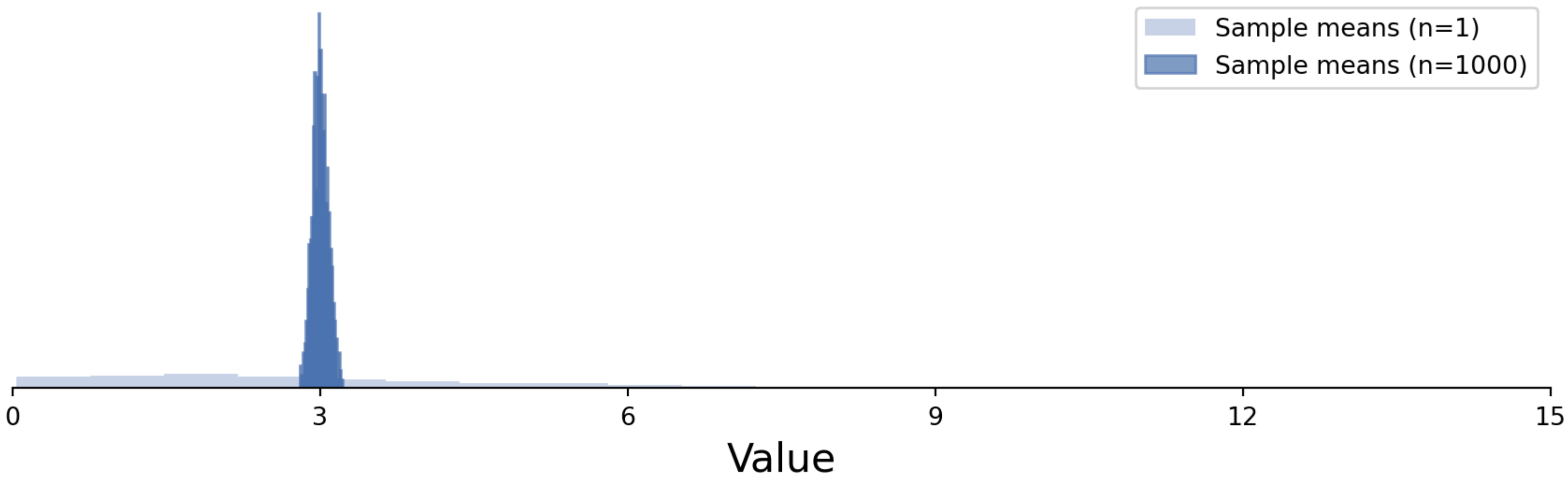
Overlay the n=1 distribution behind the n=1000 distribution.

```python
# Overlay n=1 (raw population) behind n=1000 (tight, normal)
means_1 = stats.chi2.rvs(df=3, size=(1000, 1)).mean(axis=1)
means_1000 = stats.chi2.rvs(df=3, size=(1000, 1000)).mean(axis=1)

sns.histplot(means_1, bins=30, alpha=0.3, stat='density', label='Sample means (n=1)')
sns.histplot(means_1000, bins=30, alpha=0.7, stat='density', label='Sample means (n=1000)')
```

Exercise 3.2 | Skewed Population

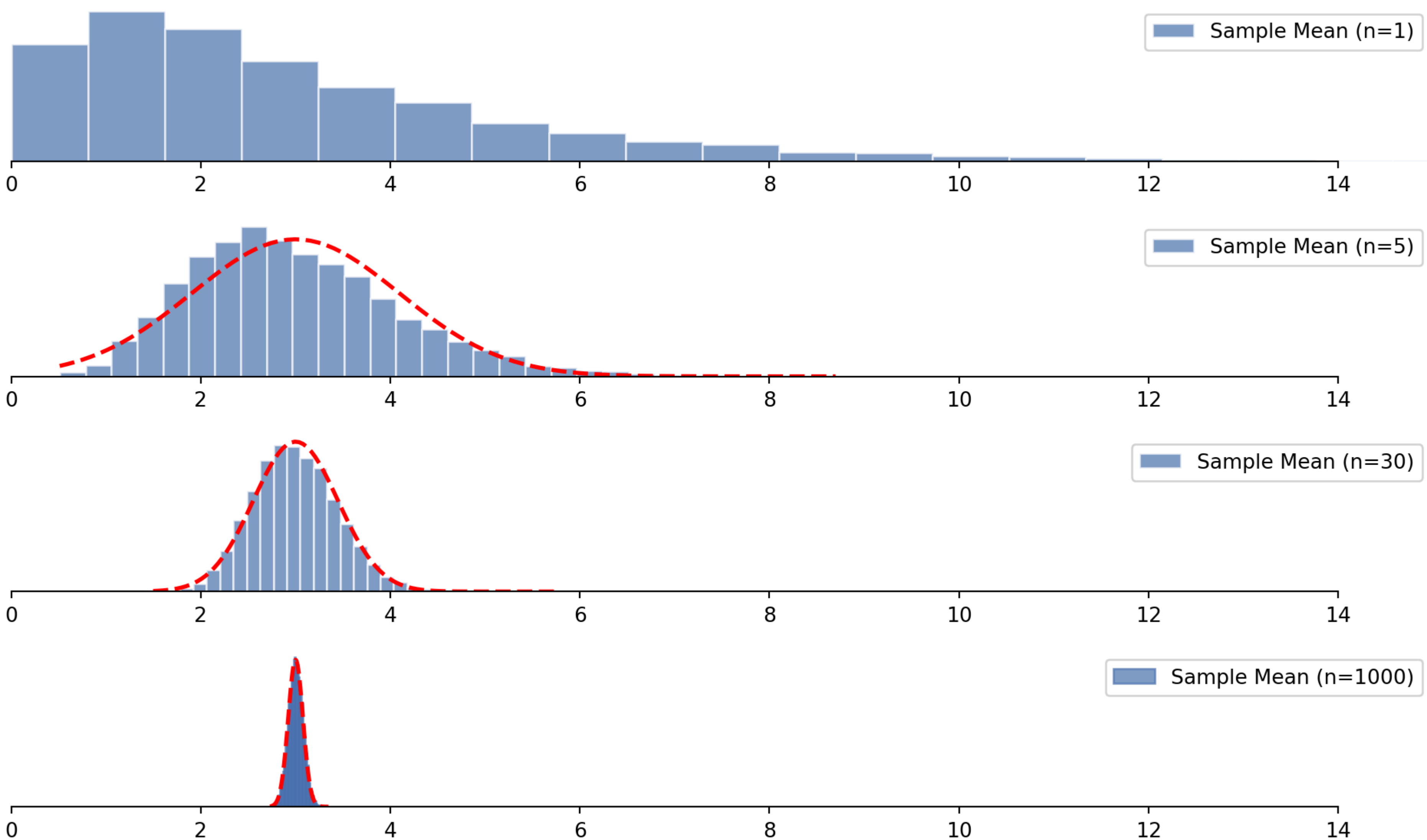
From skewed population to normal sampling distribution.



> the CLT works for (nearly) any distribution shape

Exercise 3.2 | Skewed Population

The full picture — sample means converge to normal as n increases.



Key Properties

Three things to notice about the sampling distribution of $\bar{x}$.

1. ***Unbiased***: *the sampling distribution is centered on the population mean* μ.
2. ***Precise***: *the standard error* $\sigma/\sqrt{n}$ *shrinks as sample size increases.*
3. ***Universal***: *the shape approaches normal regardless of the population.*

Assumptions

The CLT isn't magic. There are a few conditions.

1. ***Independence****: observations don't influence each other.*
2. ***Identical distribution****: observations come from the same population.*
3. ***Sample size****:* $n \geq 30$ *is usually sufficient.*

What We've Achieved

From an unobservable population to a knowable sampling distribution.

1. ***Problem***: *the population distribution is unobservable.*
2. ***Insight***: *the distribution of* $\bar{x}$ *is knowable even when the population isn't.*
3. ***Implication***: *that distribution is centered on* μ, *linking sample to population.*

Looking Ahead

We know the sampling distribution. Now what do we do with it?

- *Part 3.3 | **Confidence Intervals** - how close is $\bar{x}$ to the true μ?*
- *Part 3.4 | **Hypothesis Testing** - can we test whether μ equals a specific value?*

> *the CLT gives us the distribution — Parts 3.3 and 3.4 show us how to use it*