Random Sampling and Sampling Distributions

31 Aug 2026 | Sampling and the reliability of averages

ECO 351 – Introduction to Econometrics

Shi Qiu – Fall 2026

Recall: one outcome and its expected value

Let M count computer crashes while a student completes a paper. Last lecture used this distribution:

m	0	1	2	3	4
$P(M = m)$	.80	.10	.06	.03	.01

The calculation

$$E(M) = \sum_m mP(M = m) = 0.35$$

Check the interpretation

Can one paper have 0.35 crashes?
What does the value 0.35 describe?

Distinguish the outcome on one paper from the average described by the distribution.

A small commuting model we can work out completely

One randomly selected day

Let Y be commuting time in minutes. In this simplified model:

Time	Probability
10	1/2
30	1/2

The population quantities

$$\mu_Y = E(Y) = 20$$

$$\sigma_Y^2 = \tfrac{1}{2}(10-20)^2 + \tfrac{1}{2}(30-20)^2 = 100$$

$$\sigma_Y = 10 \text{ minutes}$$

We want to learn how reliably an average of sampled days recovers the population mean of 20 minutes.

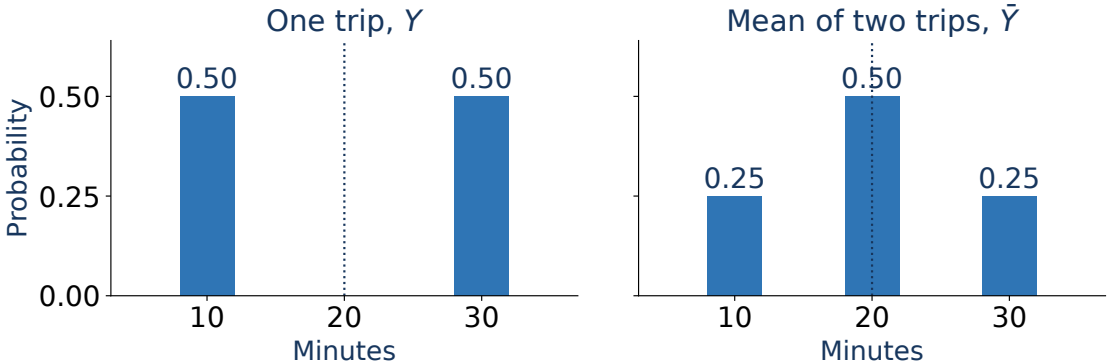
Draw two days and list every possible average

Let Y_1 and Y_2 be independent draws and define $\bar{Y} = (Y_1 + Y_2)/2$.

First day Y_1	Second day Y_2	Probability	Average $\bar{Y}$
10	10	$(1/2)(1/2) = 1/4$	10
10	30	$1/4$	20
30	10	$1/4$	20
30	30	$1/4$	30

The two mixed samples are different samples, but they produce the same average. Thus $P(\bar{Y} = 20) = 1/4 + 1/4 = 1/2$.

The sampling distribution describes the averages



Both distributions are centered at 20. Averaging moves probability toward the center: $P(Y = 20) = 0$, but $P(\bar{Y} = 20) = 1/2$.

Averaging preserves the center and reduces the spread

Center of the averages

$$E(\bar{Y}) = \tfrac{1}{4}(10) + \tfrac{1}{2}(20) + \tfrac{1}{4}(30) = 20$$

The population mean and the expected sample mean are the same.

Spread of the averages

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \tfrac{1}{4}(10-20)^2 \\ &\quad + \tfrac{1}{2}(20-20)^2 \\ &\quad + \tfrac{1}{4}(30-20)^2 \\ &= 50.\end{aligned}$$

One trip has SD 10 minutes. The average of two trips has SD $\sqrt{50} \approx 7.07$ minutes.

Try it: an average of three independent days

Same commuting model

Each day is 10 or 30 minutes, independently, with probability $1/2$ each.

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3}{3}$$

Work with a partner

1. Which samples give an average **above 20** minutes?
2. Find $P(\bar{Y} > 20)$.
3. Is an average of exactly 20 possible?

You may list the eight ordered samples, or count how many 30-minute days each sample contains.

The event determines which samples to count

30-minute days	Number of samples	Average	Probability
0	1	10	1/8
1	3	50/3	3/8
2	3	70/3	3/8
3	1	30	1/8

$P(\bar{Y} > 20) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$

None of the four possible averages equals 20. Yet $E(\bar{Y}) = 20$.

The center of a sampling distribution need not be a value attained by one particular sample.

The same reasoning extends to a sample of size n

Population and sampling

Y_i: outcome on sampled observation i.
n: number of observations.
μ_Y: population mean.
σ^2_Y: population variance.

Average the observations

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} Y_i$$

The summation sign means *add the n observations*.

Assume the observations are i.i.d. and have finite variance. The population need not have only two possible outcomes.

Why the expected sample mean equals the population mean

$$
\begin{aligned}
E(\bar{Y}) &= \frac{E(Y_1) + \cdots + E(Y_n)}{n} \\
&= \frac{\mu_Y + \cdots + \mu_Y}{n} \\
&= \frac{n\mu_Y}{n} = \mu_Y,
\end{aligned}
$$

Read each step

Take the expected value of each observation.

Each has the same expected value, μ_Y.

Adding n copies and dividing by n returns μ_Y.

Correct on average does not mean correct in every sample. An observed $\bar{Y}$ can differ from μ_Y.

Why the variance falls as the sample grows

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \frac{1}{n^2}\mathrm{Var}(Y_1 + \cdots + Y_n) \\ &= \frac{1}{n^2}(\sigma_Y^2 + \cdots + \sigma_Y^2) \\ &= \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n},\end{aligned}$$

Two rules do the work

Dividing a variable by n divides its variance by n^2. Independent variables have additive variances.

Taking the square root gives $SD(\bar{Y}) = \sigma_Y/\sqrt{n}$. For two days: $10/\sqrt{2} \approx 7.07$ minutes.

Standard deviation and standard error describe different spreads

Individual observations

$$SD(Y_i) = \sigma_Y = 10$$

The spread of individual commuting times does not shrink when we observe more days.

Sample averages

n	$SD(\bar{Y}) = 10/\sqrt{n}$
1	10
4	5
25	2
100	1

The **standard error** describes the sampling spread of an estimator. For the sample mean, the theoretical spread is $\sigma_Y/\sqrt{n}$.

The law of large numbers describes increasing reliability

What becomes unlikely

A sample average that misses the fixed population mean by more than a chosen tolerance.

$$P\left(|\bar{Y} - \mu_Y| > \varepsilon\right) \longrightarrow 0$$

Here $\varepsilon > 0$ is the tolerated error.

For commuting times

Choose a tolerance of 2 minutes.

$$P(\bar{Y} < 18 \text{ or } \bar{Y} > 22) \longrightarrow 0$$

As n grows, averages outside that band become less likely.

Under i.i.d. sampling with finite variance, $\bar{Y}$ converges in probability to μ_Y. Individual trips remain variable.

Standardization measures distance in sampling units

Subtract the center

$$\bar{Y} - \mu_Y = 22 - 20 = 2$$

The average is 2 minutes above the population mean.

Divide by the relevant spread

$$z = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} = \frac{22 - 20}{1} = 2$$

The average is 2 sampling standard deviations above the population mean.

Average in minutes	18	20	22
Standardized average z	-2	0	2

For an average, divide by $SD(\bar{Y}) = 1$, not by the individual-trip SD of 10.

The central limit theorem supplies the approximation

For i.i.d. observations with a finite, positive variance, the standardized sample mean becomes approximately standard normal as n grows:

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \overset{\text{approx}}{\sim} N(0,1).$$

The original outcomes

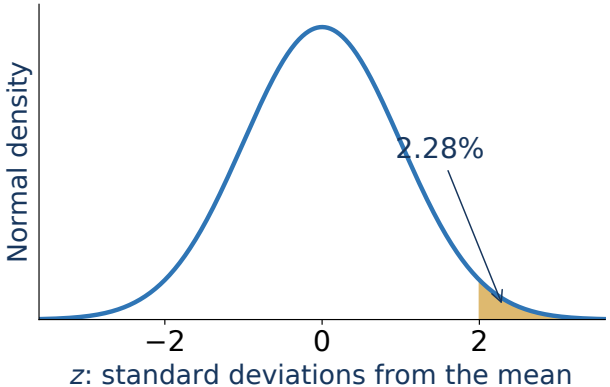
They may be skewed or discrete. They do not have to be normal.

The sample average

For sufficiently large n, its distribution is approximately $N(\mu_Y, \sigma_Y^2/n)$.

If the observations themselves are normal, their average is exactly normal for every n.

An average above 22 minutes is an upper-tail event



100 commuting days

$$SD(\bar{Y}) = 10/\sqrt{100} = 1$$

$$\begin{aligned} P(\bar{Y} > 22) &\approx P(Z > 2) \\ &= 1 - \Phi(2) \\ &= 1 - .9772 \\ &= .0228. \end{aligned}$$

$\Phi(z) = P(Z \leq z)$ is the standard normal CDF. About 2.3% of these sample averages exceed 22 minutes under the approximation.

Stata: generate 1,000 samples of 100 commuting days

```
clear
set seed 35103
set obs 100000
generate study = ceil(_n/100)
generate commute = rgamma(4,5)
```

What the commands create

`_n` is the row number.
`ceil(_n/100)` assigns a study number to each block of 100 rows.

The commuting population

`rgamma(4,5)` generates positive, skewed times with mean 20 and SD 10.

The seed makes this simulation reproducible. At this point, each row still represents one commuting day.

Stata: turn the trips into study averages

```
collapse (mean) commute, by(study)
histogram commute, normal
```

What collapse changes

The 100 trips in each study are replaced by their average.
The dataset now contains 1,000 observations.

What the histogram should show

$$\text{Center} \approx 20$$

$$\text{SD} \approx 10/\sqrt{100} = 1$$

Each plotted observation is one study average.

The normal curve is an overlay fitted to the simulated averages. A similar shape is evidence from a demonstration, not a proof of the CLT.

Stata: compare the observed tail with the approximation

```
summarize commute
count if commute > 22
display r(N)/_N
display 1-normal(2)
```

Quantity	Model / approximation	Simulation
Mean of study averages	20	20.003
SD of study averages	1	0.994
Fraction above 22	0.0228	0.026

26 of 1,000 study averages exceed 22. Simulation and approximation both introduce error.

From a probability model to an unknown population mean

What we can now explain

An average has a sampling distribution.
Its center is μ_Y and its SD is $\sigma_Y/\sqrt{n}$ under i.i.d. sampling.
The CLT provides an approximate normal shape.

What changes next lecture

We observe one sample, not the population.
The sample mean is known, but μ_Y and σ_Y are not.
We use sampling uncertainty to judge what the data tell us about μ_Y.

Today: a known population implies probabilities for sample averages. Next: an observed sample helps us estimate an unknown population mean.