Statistics/Econometrics: Fundamentals

· **2026**

Fatih Kansoy

University of Oxford

What econometrics is

Econometrics: measuring relationships with data

- Economic theory tells us the **direction** of many relationships, but rarely the **number**. Higher cigarette prices cut smoking, but by how much? Another year of education raises pay, but by how much?
- **Econometrics** is the use of economic theory and statistical methods to put numbers on these relationships using real-world data.
- Two distinct jobs run through the course:
  - measuring a **causal effect**: what would change if we intervened,
  - making a **prediction**: a good forecast of an outcome we do not yet observe.

The questions look like these

- **A causal question.** What is the effect on cigarette consumption of a 1% rise in the price? That number is the **price elasticity of demand**, and tax policy needs it.
- **A causal question.** Holding the borrower's ability to repay fixed, does an applicant's race change the chance a mortgage is approved? The Boston Fed found 28% of black applicants but only 9% of white applicants were denied in the early 1990s.
- **A prediction question.** By how much will GDP grow next year? A forecaster does not need to know *why*, only to predict well.

The ideal: a randomised controlled experiment

The benchmark for a causal effect · Split subjects at random into a **treatment group** (gets the treatment) and a **control group** (does not). Because assignment is random, the groups differ *only* by the treatment, so the difference in average outcomes is the **causal effect**.

- A horticulturalist gives 100 grams of fertiliser to half the plots, chosen by computer, and none to the rest. The gap in average yield is the effect of the fertiliser.
- Randomisation is what breaks the link between the treatment and every other factor (how sunny each plot is, how good the soil). That is the whole point.

We define a causal effect as the result of an ideal experiment, even when we cannot run one. It is the yardstick against which all other estimates are judged.

Why we usually cannot run the experiment

- Real experiments on people are often too expensive, too slow, or unethical. We cannot randomly hand teenagers cheap cigarettes, and we cannot randomly assign firms a credit rating.
- So in economics and finance we almost always work with **observational data**: numbers recorded by watching the world, not by intervening in it.
- In observational data the "treatment" (a firm's leverage, an investor's strategy, a country's policy) is *not* assigned at random. Other factors move with it and muddy the comparison.

Correlation is not causation. Two things can move together because one causes the other, because a third factor drives both, or by chance. Sorting this out is most of the course.

Three kinds of data

How the observations are arranged ·

- **Cross-section**: many entities (firms, funds, countries) at one point in time.
- **Time series**: one entity tracked over many periods. We write T for the number of periods.
- **Panel**: many entities, each followed over several periods.

Type	Finance example	Dimension
Cross-section	returns of 500 stocks on one day	n entities
Time series	the S&P 500 monthly, 1990 to 2025	T periods
Panel	48 funds, each over 10 years	$n \times T$

Here n is the number of entities and T the number of time periods. A panel of 48 funds over 10 years holds $48 \times 10 = 480$ observations.

Random variables and distributions

Random variables: the language of uncertainty

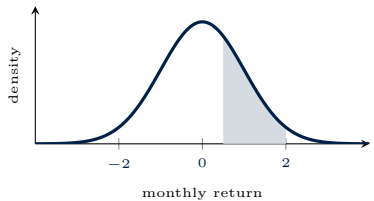
The basic objects ·

- An **outcome** is one possible result of a random process; its **probability** is the share of the time it occurs in the long run.
- A **random variable** Y is a number attached to a random outcome, for example next month's return on a stock.

- A **discrete** random variable takes separate values (the number of defaults in a bond portfolio: $0, 1, 2, \dots$). A **continuous** one takes any value in a range (a return, which can be 1.3% or 1.34% or anything between).
- Before the month ends, Y is random. Once the month is over and the return is recorded, it is just a number. This shift from random to realised is the heart of everything today.

Describing a random variable: the distribution

- For a **discrete** variable, the **probability distribution** lists each value and the probability it occurs; the probabilities sum to 1.
- For a **continuous** variable we use the **probability density function** (the **p.d.f.**). The variable cannot equal one exact value with positive probability, so instead the *area* under the density between two points is the probability of landing between them.
- The **cumulative distribution function** (the **c.d.f.**) gives $\mathbb{P}(Y \leq y)$, the probability of being at or below a value. It runs from 0 up to 1.



The shaded area is $\mathbb{P}(0.5 \leq Y \leq 2)$.

Read it as a shape over possible returns. Most of the mass sits near the centre; the tails hold the rare large moves, up or down, that matter most for risk.

The mean: the expected value

Expected value, the long-run average · For a discrete Y taking values $y_1, \dots, y_k$ with probabilities $p_1, \dots, p_k$,

$$\mu_Y = \mathbb{E}(Y) = y_1 p_1 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i \, p_i .$$

- The **mean** μ_Y (also called the **expectation** of Y) is the probability-weighted average of the possible values: the average you would get over very many repetitions.
- **Example.** You lend a friend \$100 at 10%. With probability 0.99 you are repaid \$110; with probability 0.01 they default and you get \$0. The expected repayment is $110 \times 0.99 + 0 \times 0.01 = \mathbf{\$108.90}$.

In finance μ_Y is the expected return: the centre of gravity of the return distribution, the number a long-run investor earns on average per period.

The spread: variance and standard deviation

How far Y typically falls from its mean ·

$$\sigma_Y^2 = \operatorname{var}(Y) = \mathbb{E}\big((Y - \mu_Y)^2\big), \qquad \sigma_Y = \sqrt{\sigma_Y^2}.$$

- The **variance** σ_Y^2 is the expected squared distance of Y from its mean. Squaring keeps positive and negative deviations from cancelling, and it punishes large deviations heavily.
- Because the variance is in squared units (squared percent), we usually quote the **standard deviation** σ_Y, which is back in the original units (percent).

In finance the standard deviation of returns is the standard measure of **risk**, and it goes by a special name: **volatility**. A higher σ_Y means a wider, riskier spread of outcomes around the same average.

A linear rescaling moves the mean and the spread

- Suppose a new variable is a straight-line function of Y: $W = a + bY$, with fixed numbers a and b. Then

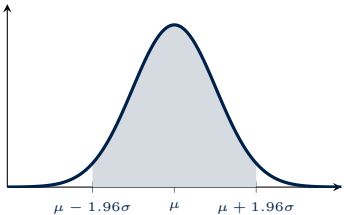
$$\mathbb{E}(W) = a + b\,\mu_Y, \qquad \operatorname{var}(W) = b^2\,\sigma_Y^2.$$

- Adding a constant a shifts the mean but not the spread. Multiplying by b scales the mean by b and the standard deviation by $|b|$.

So what: this is exactly how leverage works. If you hold a position at $2\times$ exposure, so $b = 2$, your expected return doubles and your volatility also doubles. Risk and reward scale together under a linear rule.

The normal distribution

The normal distribution



95% of the area lies within 1.96σ of μ.

- The **normal distribution** is the familiar bell curve. It is fixed by just two numbers, its mean μ and variance σ^2, and we write it $N(\mu, \sigma^2)$.
- It is symmetric around μ, and 95% of its probability sits between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$. That 1.96 will reappear in every confidence interval in this course.

The **standard normal** is the special case $N(0, 1)$, with mean 0 and variance 1. We write a standard normal variable as Z, and $\Phi(c) = \mathbb{P}(Z \leq c)$ for its cumulative distribution function.

Standardising: putting any normal on the same ruler

Subtract the mean, divide by the standard deviation · If $Y \sim N(\mu, \sigma^2)$, then $Z = \dfrac{Y-\mu}{\sigma} \sim N(0,1)$.

- Standardising answers "how many standard deviations from the mean is this value?". It turns any normal question into one about the standard normal, which is tabulated.
- **Worked check.** Let $Y \sim N(1,4)$, so $\mu = 1$ and $\sigma = \sqrt{4} = 2$. Then
$$\mathbb{P}(Y \le 2) = \mathbb{P}\big(Z \le \tfrac{2-1}{2}\big) = \Phi(0.5) = \mathbf{0.691}.$$

So what: a return that is 3 standard deviations below its mean is a -3 in standardised units, a once-in-a-few-years move *if* returns were normal. Whether they really are normal is the next slide.

A bad day on Wall Street: returns are not normal

Black Monday, 19 October 1987 · The Dow Jones fell 22.6% in one day. Over 1980 to 2012 its average daily change was about 0.04% with a standard deviation of about 1.12%. So in standardised units the drop was

$$\frac{-22.6 - 0.04}{1.12} \approx \textbf{-20} \text{ standard deviations.}$$

- If daily returns were exactly normal, a move of 20 standard deviations would have probability about 5.5×10^{-89}: it should essentially never happen in the life of the universe. Yet it did, and days nearly this extreme recur.
- The lesson: real return distributions have **fat tails**. Extreme moves are far more likely than the normal curve allows. The normal is a useful first approximation, not the truth, and risk models that assume it understate the danger.

This is why Nassim Taleb's term "black swan" entered finance, and why we treat the normal with respect but not faith.

Two random variables

Two variables together: joint and marginal

The joint distribution · The **joint distribution** of two random variables X and Y gives the probability that they take particular values *at the same time*, $\mathbb{P}(X = x, Y = y)$.

- The **marginal distribution** of Y is just its own distribution, recovered from the joint one by summing over all values of X. "Marginal" simply means "on its own, ignoring the other variable".
- **Example.** Let $X = 1$ if a month is a recession, 0 otherwise, and Y a stock's return. The joint distribution records how returns and recessions occur together; the marginal of Y is the return distribution across all months, recession or not.

Conditional distribution and the conditional mean

The distribution of Y given that we know X ·

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}.$$

- The **conditional distribution** is the distribution of Y once we are told the value of X. Its mean is the **conditional mean** $\mathbb{E}(Y \mid X = x)$: the average of Y within the group that has $X = x$.
- The conditional mean is just the familiar idea of a *group average*, given a name. The mean return of stocks *in recession months* is $\mathbb{E}(Y \mid X = 1)$; the mean return *in normal months* is $\mathbb{E}(Y \mid X = 0)$.

The conditional mean is the best prediction

- Suppose you must predict Y knowing X, and you are judged by the **mean squared prediction error**: the average of $(Y - \text{guess})^2$ over many cases. Squaring penalises big misses.
- Of all predictions that can depend on X, the one with the smallest mean squared error is the **conditional mean** $\mathbb{E}(Y \mid X)$.

Why prediction and causation are different jobs · To *predict* Y you only need the conditional mean, a statistical pattern. To know the *causal effect* of changing X you need much more. A good forecast need not be a causal story.

Umbrellas predict rain well, but opening one does not cause it. A signal can forecast returns without being a lever you can pull to earn them.

Independence

When one variable tells you nothing about the other · X and Y are **independent** if knowing X leaves the distribution of Y unchanged: $\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$ for all values. Equivalently, the joint factorises: $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$.

- Under independence the conditional mean of Y does not move with X: the group averages are all the same.

Covariance and correlation

Covariance: do they move together?

The direction of co-movement ·

$$\sigma_{XY} = \operatorname{cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

- When X is above its mean and Y tends to be above its mean too, the product is positive and the **covariance** is positive: the two move in the same direction. If they move in opposite directions, the covariance is negative.
- If X and Y are independent, their covariance is 0. The converse does not hold: a covariance of 0 does not by itself prove independence.
- The covariance of a variable with itself is its variance: $\operatorname{cov}(X, X) = \sigma_X^2$.

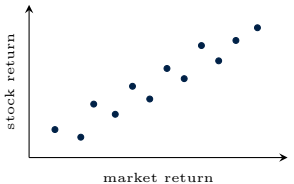
The drawback: covariance has awkward units (units of X times units of Y), so its size is hard to read. The fix is correlation.

Correlation: a unit-free number in $[-1, 1]$

Covariance, rescaled ·

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \, \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1, 1].$$

- Dividing by the two standard deviations cancels the units, giving a pure number between -1 and $+1$: $+1$ a perfect upward line, -1 a perfect downward line, 0 no *linear* link.
- **It measures only *linear* association.** A strong curved pattern can still have a correlation near 0, so always look at the scatterplot too.



Positive correlation, about $+0.85$.

So what: a stock's correlation with the market, scaled by their volatilities, becomes its **beta** in the CAPM, the single number that says how much market risk it carries.

Variance of a sum: where diversification comes from

The variance of a combination of two variables ·

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}.$$

- Risk does not simply add. The cross term $2\operatorname{cov}(X,Y)$ means the risk of a combined position depends on how the two pieces move together.
- **Worked example.** Hold half your money in each of two stocks, each with volatility 20%. If their correlation is 1, the portfolio volatility stays 20%, no benefit. If their correlation is only 0.2, it falls to about 15.5% for the same average return.

Combining assets that do not move in lockstep lowers risk without lowering the expected return. That free lunch is **diversification**, and the covariance term is its source.

Diversification: spreading risk across many assets

Do not put all your eggs in one basket · Split \$1 equally across n assets, each with the same volatility σ and the same pairwise correlation ρ. As n grows, the portfolio's risk falls toward $\sqrt{\rho}\,\sigma$, not to zero.

Number of assets n	1	5	20	100
Portfolio volatility, $\rho = 0$	20.0%	8.9%	4.5%	2.0%
Portfolio volatility, $\rho = 0.3$	20.0%	13.3%	11.6%	11.1%

- If assets were uncorrelated ($\rho = 0$), enough of them would drive risk to almost nothing.
- Because real assets are positively correlated ($\rho \approx 0.3$ here), risk stops falling at a floor of about 11%. That floor is **market risk**, the part you cannot diversify away.

The sample average

Simple random sampling and i.i.d. data

How the data are drawn · In **simple random sampling** we draw n items at random from a population, each equally likely. Write the draws $Y_1, \ldots, Y_n$. They are then **i.i.d.**:

- *independent*, one draw tells you nothing about another, and
- *identically distributed*, all from the same population.

- Before sampling, each Y_i is random. After sampling, each is a recorded number. The subscript i just indexes the observations, from 1 to n.

The sample average is itself random

The estimator we will use again and again · The **sample average** of the n observations is

$$\overline{Y} = \frac{1}{n}\left(Y_1 + \cdots + Y_n\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

- Because the sample is random, $\overline{Y}$ is a **random variable**: a different sample would have given a different value. Its distribution across all possible samples is the **sampling distribution**.
- This is the key conceptual step. We use the one $\overline{Y}$ we computed to learn about the unknown population mean μ_Y, and to do that we need to know how $\overline{Y}$ behaves from sample to sample.

Example: the average of 60 monthly returns estimates a fund's true mean return, but it is only an estimate, and a different 60 months would have produced a different average.

Mean and variance of the sample average

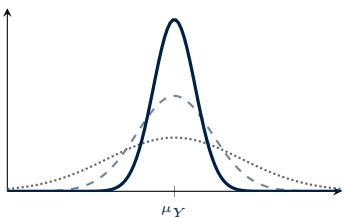
Centre and spread of $\overline{Y}$ · If $Y_1, \ldots, Y_n$ are i.i.d. with mean μ_Y and variance σ_Y^2, then

$$\mathbb{E}(\overline{Y}) = \mu_Y, \qquad \operatorname{var}(\overline{Y}) = \frac{\sigma_Y^2}{n}, \qquad \sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}}.$$

- The sample average is centred on the truth: $\mathbb{E}(\overline{Y}) = \mu_Y$, so on average it is right. This is the property of being **unbiased**.
- Its spread shrinks as the sample grows, but only like $1/\sqrt{n}$. To *halve* the uncertainty in $\overline{Y}$ you need *four times* as much data.

So what: this square-root law is costly in finance. To pin down a fund's true average return you need many years of data, and by then the strategy or the market may have changed.

The law of large numbers



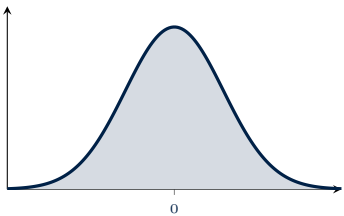
As n rises the distribution of $\overline{Y}$ tightens onto μ_Y.

Consistency · If $Y_1, \ldots, Y_n$ are i.i.d. and large outliers are unlikely (the variance is finite), then as n grows the sample average $\overline{Y}$ gets arbitrarily close to the true mean μ_Y with probability approaching 1.

- We say $\overline{Y}$ is a **consistent** estimator of μ_Y: more data brings it ever closer to the truth.
- The large values balance the small values, and the average settles near their common mean.

This is the formal version of "the average of many draws is reliable". It is why long return histories are worth more than short ones.

The central limit theorem



For large n the standardised $\overline{Y}$ is close to standard normal.

The shape becomes normal · If $Y_1, \ldots, Y_n$ are i.i.d. with finite variance, then as n grows the standardised sample average

$$\frac{\overline{Y} - \mu_Y}{\sigma_{\overline{Y}}} \longrightarrow N(0,1),$$

whatever the shape of the underlying Y.

- Standardising means subtracting the mean and dividing by the standard deviation of $\overline{Y}$.
- The remarkable part: the population Y can be skewed or fat-tailed, yet the *average* is still approximately normal once n is large.

So what: even though single returns are not normal, an average of many is close to normal. This is what lets us use 1.96 for tests and intervals in the next lecture.

Summary

Recap

- **Econometrics** measures relationships from data, separating the job of finding a **causal effect** from the job of making a **prediction**.
- The cleanest causal evidence is a **randomised experiment**; in finance we almost always have **observational** data instead, so **correlation is not causation**.
- A **random variable** is summarised by its **mean** μ_Y (expected return) and its **variance** σ_Y^2, whose square root is the **volatility**.
- **Covariance** and **correlation** measure co-movement; the variance of a sum carries a covariance term, which is the source of **diversification** and of **beta**.
- For i.i.d. data the **sample average** has mean μ_Y and standard deviation $\sigma_Y/\sqrt{n}$; the **law of large numbers** makes it consistent and the **central limit theorem** makes it approximately normal.

> We learn about an unknown population from a single random sample. The sample average is centred on the truth, tightens like $1/\sqrt{n}$, and is approximately normal for large n.

From population to sample

The problem: we see a sample, we want the population

- A trading strategy returned 0.6% a month on average over the past few years, about 7% **a year**. That looks good. But is its **true** long-run edge really positive, or was the 0.6% just **luck** in a short, noisy run?

- We can never see the whole **population** (every return the strategy could ever produce). We see only a **sample**, the months that happened, and a few noisy years cannot settle the question by eye.

- Three tools take us from the sample back to the population:

  - **estimation**: a best guess for an unknown population number,

  - **hypothesis testing**: is a claim about the population consistent with the data,

  - **confidence intervals**: a plausible range for the unknown number.

The words we will use

Population objects: fixed and unknown ·

- **Random variable** Y: a numerical outcome, for example next month's return.
- **Population mean** $\mu_Y = \mathbb{E}(Y)$: the true long-run average of Y. A fixed number we do not know.
- **Population variance** $\sigma_Y^2 = \mathbb{E}\big((Y - \mu_Y)^2\big)$: how spread out Y is; σ_Y is the standard deviation.

- A **random sample** $Y_1, \ldots, Y_n$ is drawn by **simple random sampling**, so the observations are **i.i.d.**:
  - *independent*, one draw tells you nothing about another, and
  - *identically distributed*, all from the same population.
- Greek letters (μ, σ) are unknown population truths; Latin letters $(\overline{Y}, s)$ are things we compute from data.

Estimation

Estimator versus estimate

Three things to keep apart ·

- **Estimand**: the population number you want, here μ_Y. Fixed, unknown.
- **Estimator**: the *rule* you apply to a sample, for example the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. A **random variable**: a new sample gives a new value.
- **Estimate**: the one *number* the rule returns for your sample, for example 0.6%.

- The estimator is random because the sample is random. Before you draw the data it could come out many ways; after you draw it, you hold one number.
- There are **many** rules for μ_Y: the sample mean, the first observation Y_1, the median, a weighted average. **So which rule should we trust?**

Three things we want from an estimator

Unbiased, consistent, efficient ·

- **Unbiased**: right *on average* across samples, $\mathbb{E}(\text{estimator}) = \mu_Y$.
- **Consistent**: it gets closer to the truth as the sample grows, estimator $\to \mu_Y$ as $n \to \infty$.
- **Efficient**: among unbiased rules, it has the **smallest variance**, so it varies least from sample to sample.

- **Bias** means the rule is pointed at the wrong target: $\mathbb{E}(\text{estimator}) \neq \mu_Y$. More observations reduce noise, but they do not fix a systematically wrong sample.
- **Survivorship bias** is the finance example. The population is *all funds ever launched*; the database contains only funds *still alive*. Since dead funds usually had poor returns, the average return of surviving funds is too high.
- Likewise a poll of only phone owners (1936, "Landon beats Roosevelt") was biased: the sample was not drawn at random from the population.

Why the sample mean? It uses the data efficiently

- Y_1 is unbiased ($\mathbb{E}(Y_1) = \mu_Y$), but it puts weights $(1, 0, \ldots, 0)$ on the sample: the other $n-1$ observations do no work, so its precision never improves as n grows.
- The sample mean uses all observations equally:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \operatorname{var}\left(\overline{Y}\right) = \frac{\sigma_Y^2}{n} \quad \text{instead of} \quad \operatorname{var}(Y_1) = \sigma_Y^2.$$

  As n grows, $\operatorname{var}\left(\overline{Y}\right)$ shrinks towards zero while $\operatorname{var}(Y_1)$ stays fixed at σ_Y^2.
- Therefore $\overline{Y}$ is unbiased, consistent, and among all *linear unbiased* rules has the smallest variance: it is **BLUE**.

Least squares view · If one number m must summarise the sample, choose the value that makes squared errors as small as possible:

$$\min_m \sum_{i=1}^{n}(Y_i - m)^2, \qquad -2\sum_{i=1}^{n}(Y_i - m) = 0 \quad \Rightarrow \quad m = \overline{Y}.$$

The mean is the best constant fitted value. Next lecture, OLS chooses the best fitted line by the same principle.

How precise is one estimate? The standard error

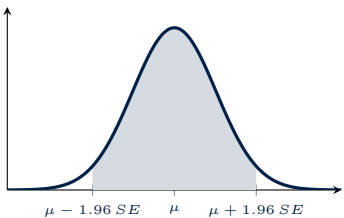
The standard error of the mean · $\overline{Y}$ has variance σ_Y^2/n. We do not know σ_Y, so we use the **sample standard deviation** s_Y, where

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2, \qquad SE(\overline{Y}) = \frac{s_Y}{\sqrt{n}}.$$

- The standard error is the estimated standard deviation of $\overline{Y}$: roughly how far your one estimate typically lands from the truth.
- The divisor $n-1$ (the **degrees of freedom**) makes s_Y^2 unbiased: computing $\overline{Y}$ first uses up one piece of information.
- Precision improves like $1/\sqrt{n}$: to *halve* the standard error you need *four times* the data.

So what: in finance the square-root law is costly. To tell a real edge from noise you need many years of data, and by then the strategy may have changed.

The keystone: the sampling distribution



Distribution of $\overline{Y}$ over many hypothetical samples.

- Each possible sample gives a slightly different $\overline{Y}$. The distribution of $\overline{Y}$ across samples is the **sampling distribution**.
- It is centred at μ_Y (unbiased) with spread SE. As n grows it tightens around μ_Y (**law of large numbers**) and becomes **normal** (**central limit theorem**), whatever the shape of Y.
- So about 95% of samples land within $1.96\,SE$ of μ_Y.

You see *one* draw from this curve. Every test and interval today works in reverse: from your single $\overline{Y}$ to the unknown μ_Y at the centre.

Hypothesis testing

The logic: a yes/no question about the population

Two competing claims ·

- **Null** H_0: the status quo or "no effect", for example $\mu_Y = 0$ (the fund has no skill).
- **Alternative** H_1: what you suspect instead, for example $\mu_Y \neq 0$ (a *two-sided* alternative).

- Treat it like a courtroom: the null is presumed true. You do not try to *prove* it; you ask whether the data are strong enough to **reject** it.
- A **one-sided** alternative ($\mu_Y > 0$) is used when only one direction is of interest, for example "does education *raise* earnings". We use two-sided by default.

You never "accept H_0". You **reject** it or **fail to reject** it. Failing to reject is an acquittal for lack of evidence, not a verdict of innocence.

The t-statistic

Distance from the null, in standard errors ·

$$t = \frac{\text{estimate} - \text{null}}{SE} = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}.$$

- t counts how many standard errors the estimate sits from the hypothesised value $\mu_{Y,0}$.
- If H_0 is true then, by the central limit theorem, t is approximately $N(0,1)$ in large samples. So a value beyond ± 1.96 happens only 5% of the time by chance.
- This is why $|t| > 1.96$ is the working definition of "surprising" at the 5% level, and $|t| > 2$ is the quick rule you will see in every empirical finance paper.
- More generally, any number computed from a sample to decide a test is a **test statistic**; the t-statistic is the one we use here.

The p-value, and what it is not

The probability of a result this extreme, if H_0 were true ·

$$p = \mathbb{P}(|Z| > |t|) = 2\,\Phi(-|t|), \qquad Z \sim N(0,1).$$

Here Φ is the standard normal CDF. Small p means the data would be unlikely under H_0, so it is evidence *against* H_0. Convention: reject when $p < 0.05$.

A p-value is *not* the probability that H_0 is true. It is the probability of the data given H_0, not the probability of H_0 given the data.

- Read it plainly: $p = 0.03$ means "if the true mean return were zero, a result this large would appear about 3 times in 100 by luck alone".
- The p-value says more than "reject or do not reject", so report it.

Fix the rule before you look: level, errors, critical values

- Choose a **significance level** α (usually 5%) *in advance*: the false-alarm rate you will tolerate. Reject when $|t|$ exceeds the **critical value**.
- Two ways to be wrong:
  - **Type I error**: reject a *true* null (a false discovery). Its probability is α, the **size**.
  - **Type II error**: fail to reject a *false* null (a missed effect). The **power** is the chance of catching a true effect.
- The t-values that reject H_0 form the **rejection region** ($|t| > 1.96$); the rest is the **acceptance region**.

Level α	Two-sided critical value	One-sided
10%	1.64	1.28
5%	1.96	1.64
1%	2.58	2.33

Lowering α (being stricter) cuts false discoveries but lowers power: the two errors trade off.

Worked example: is the fund's edge real?

A long-short strategy, $n = 100$ months · Mean monthly excess return $\overline{Y} = 0.60\%$, sample SD $s = 2.50\%$. Test $H_0 : \mu = 0$ (no edge), two-sided, at 5%.

1. $SE = \dfrac{2.50}{\sqrt{100}} = \dfrac{2.50}{10} = \mathbf{0.25}$.
2. $t = \dfrac{0.60 - 0}{0.25} = \mathbf{2.40}$.
3. $2.40 > 1.96$, so **reject** H_0 at 5%.
4. $p = 2\,\Phi(-2.40) \approx \mathbf{0.016}$: about 1 chance in 60.

The edge passes the test: a 0.6% average month over 100 months is too large to explain by luck alone. But 0.6% per month is also *economically* large (roughly 7% a year before costs), so here statistics and economics agree. They do not always.

Statistical significance is not economic importance

- Because $SE = s/\sqrt{n}$ shrinks with n, a *large enough* sample makes even a trivial effect "significant". A signal earning $+0.01\%$ per trade over 200,000 trades can have $t = 8$ and still be **worthless** once fees and slippage are paid.
- The reverse also happens: a genuinely large effect measured on a short, noisy sample can fail to reach significance.

> **Two different questions** · **Statistically significant**: unlikely to be exactly zero. **Economically significant**: large enough to matter after costs.

Always report the **magnitude** and its units, not only the stars. A big t or a big n is not a big effect.

When "significant" goes wrong: multiple testing

- At the 5% level a true null is falsely rejected 1 time in 20. Screen 20 useless signals and the chance that *at least one* looks significant by luck is

$$1 - 0.95^{20} = \mathbf{0.64},$$

  not 5%. Test enough useless signals and one will look significant by chance.
- This is **p-hacking** (or data mining): trying many variables, samples or rules and reporting only the winner.

> **The factor zoo** · Hundreds of "return-predicting factors" have been published. Harvey, Liu and Zhu (2016) argue that after accounting for all the testing, the honest hurdle is about $t > 3.0$, not 2.0. Many celebrated anomalies do not clear it, and stop working on new data.

The fix is set in advance: hold out data, pre-register the hypothesis, and demand **out-of-sample** evidence.

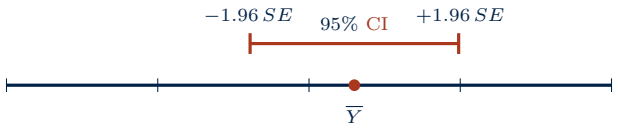
Confidence intervals

The confidence interval

A plausible range for the truth ·

$$95\%\ \text{CI for } \mu_Y: \qquad \overline{Y} \pm 1.96\, SE(\overline{Y}).$$

It is exactly the set of null values $\mu_{Y,0}$ that a two-sided 5% test would *not* reject.



The half-width $1.96\, SE$ is the **margin of error**; the full width shrinks like $1/\sqrt{n}$.

What a confidence interval means, and what it does not

Coverage: a statement about the method · Over many random samples, 95% of the intervals built this way contain the true μ_Y. The **coverage probability** is a property of the *procedure*.

The wrong reading: "there is a 95% probability that μ_Y lies in *this* interval".

- Once computed, the interval either contains μ_Y or it does not. μ_Y is a fixed number, not random; the *interval* is what changes from sample to sample.
- The 95% describes how often the recipe works, not your one interval.

Tests and intervals are two views of one thing

Duality ·

$$\mu_{Y,0} \notin 95\%\ \text{CI} \quad \Longleftrightarrow \quad \text{reject } H_0 : \mu_Y = \mu_{Y,0} \text{ at } 5\%.$$

- To test whether an effect is zero, just check whether 0 lies in the interval. For the fund, the 95% CI is $0.60 \pm 1.96(0.25) = \mathbf{[0.11\%, 1.09\%]}$, which *excludes* 0, the same "reject" we got from $t = 2.40$.
- A confidence interval is usually more useful than a test: it shows the whole plausible range, not just a yes or no.

Why we still argue about the equity premium · US stocks have beaten bonds by roughly 6% a year, with annual SD near 20%. Even with 100 years, $SE = 20/\sqrt{100} = 2\%$, so the 95% CI is about $\mathbf{[2\%, 10\%]}$. A century of data still leaves the single most important number in asset allocation badly imprecise.

Comparing two means

Is the gap between two groups real?

Take two groups (value vs growth stocks; treated vs control; men vs women) and estimate the gap:

$$\widehat{d} = \overline{Y}_1 - \overline{Y}_2, \qquad SE(\widehat{d}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \qquad t = \frac{\widehat{d}}{SE(\widehat{d})}.$$

- Reject $H_0 : \mu_1 = \mu_2$ when $|t| > 1.96$; the 95% CI for the gap is $\widehat{d} \pm 1.96\, SE(\widehat{d})$.
- **The common slip:** $SE(\widehat{d})$ is the **square root of the sum of variances**, *not* $SE_1 + SE_2$. You add the s_i^2/n_i first, then take the root.

This form does not assume the two groups share a variance, so it is the safe default.

Worked example: does value beat growth?

Value vs growth portfolios, $n = 50$ months each · Value: mean 0.90%, SD 3.0%. Growth: mean 0.30%, SD 3.0%. Test whether the average returns differ, two-sided, at 5%.

1. $SE(\widehat{d}) = \sqrt{\frac{3.0^2}{50} + \frac{3.0^2}{50}} = \sqrt{0.18 + 0.18} = \mathbf{0.60}$.
2. $\widehat{d} = 0.90 - 0.30 = \mathbf{0.60}$.
3. $t = \dfrac{0.60}{0.60} = \mathbf{1.00} < 1.96$.
4. So **fail to reject**; $p \approx 0.32$, about 1 in 3.

Value beat growth by 0.6% a month in this sample, almost 8% a year, yet with this much noise a gap that size turns up about one time in three even when the two are truly equal.

"Fail to reject" does not mean the two are equal. It means 50 months cannot tell them apart. Detecting return premiums is really hard.

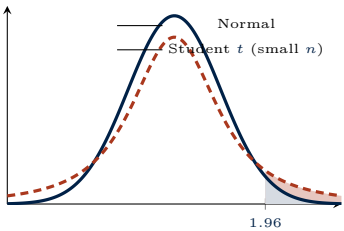
From correlation to causation

- When the two groups come from a **randomised experiment** (a treatment group and a control group assigned at random), the difference in means estimates a **causal effect**, also called the **treatment effect**.
- With **observational** data, which is almost all finance data, the same difference is just an **association**. The groups may differ in other ways (size, sector, risk) that drive the gap.

Correlation is not causation. A difference-in-means test answers "is the gap real?", never "what caused it?". Separating cause from association is the job of regression, and of the whole rest of the course, starting next lecture.

Two loose ends: small samples and correlation

Small samples: the Student t



Lower peak, and more mass in the tails (the red sliver beyond 1.96), so the cutoff grows.

- With few observations the central limit theorem applies less well and s is a noisy estimate of σ. The **Student** t distribution accounts for this: same shape as the normal but **fatter tails**, so the critical value is a little larger.
- Two-sided 5% critical values, with $n - 1$ degrees of freedom:

n	10	20	50	large
crit.	2.26	2.09	2.01	1.96

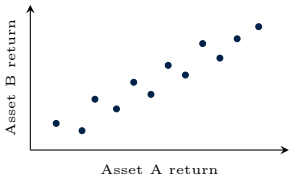
So what: the t matters only for small n, and only if Y is truly normal, which returns rarely are. In finance, samples are usually large, so 1.96 serves. Keep the t for a short track record, say a fund with 24 months.

Covariance and correlation

Two variables moving together ·

$$s_{XZ} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z}), \qquad r_{XZ} = \frac{s_{XZ}}{s_X\, s_Z} \in [-1, 1].$$

- **Sample covariance** s_{XZ} gives the *direction* of the relationship, but its size depends on units. **Sample correlation** r_{XZ} rescales it to a unit-free number in $[-1, 1]$: $+1$ a perfect upward line, -1 a perfect downward line, 0 no linear link.
- Plot the pairs on a **scatterplot** to see it. **Correlation measures only *linear* association:** a strong curved pattern can still have $r \approx 0$.



Positive correlation, $r \approx 0.85$.

So what: correlation is the engine of **diversification**. Weakly correlated assets lower portfolio risk, and an asset's correlation with the market becomes its **beta** in the CAPM. The conditional mean of one variable given another becomes the **regression** of Lecture 3.

Summary

Recap

- **Estimate** a population number with the sample mean; report its **standard error** $SE = s/\sqrt{n}$, the measure of precision.
- **Test** a claim with $t = \dfrac{\text{estimate} - \text{null}}{SE}$; small p leads you to reject. A p-value is not the probability that H_0 is true, and you never "accept H_0".
- **Report** a confidence interval, estimate $\pm 1.96\,SE$: the plausible range, and the set of nulls a test would not reject.
- **Compare** two groups with the difference in means and its root-sum-of-squares standard error.
- **Remember** the three traps: significance is not importance; multiple testing inflates false discoveries; correlation is not causation.

An estimate is one number with a margin of error around it; a test asks only whether chance alone could have produced a gap this big. "Significant" means "unlikely to be luck", not "large" and not "true".

The problem: does class size matter?

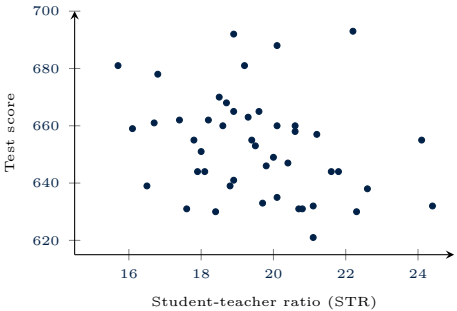
A decision a school district has to make

- A superintendent has some money and a choice: hire more teachers so that classes are **smaller**. Smaller classes mean fewer students per teacher, the **student-teacher ratio** (STR).
- The hope is that with fewer students each, teachers help more and **test scores rise**. But teachers are expensive. Before spending the money, she wants an answer to a simple question:

> **The question** · If a district lowers its student-teacher ratio by one student, how much do its test scores change on average, and **how sure** can we be of that number?

- We have data on **420 California school districts**: for each one, its average test score and its student-teacher ratio. One dot per district.
- The lecture answers her in five steps: draw the best **line**; ask how well it **fits**; ask when its slope can be **trusted**; measure its **uncertainty**; then **test** it and give a **range**. By the end we can hand her a number with an honest margin around it.

Look at the data first

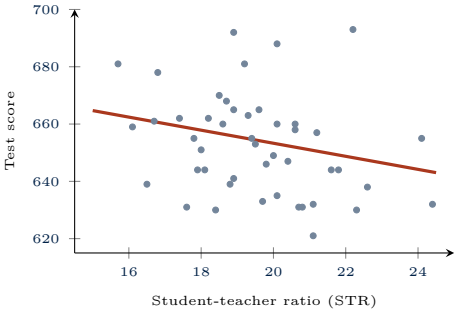


A representative slice of the 420 districts, one dot per district.

- Before any formula, look. Is there a pattern? Which way does it lean?
- The cloud tilts gently **downwards**: districts with larger classes (higher STR) tend to have *lower* scores, hinting the effect is negative.
- But the cloud is **wide**. Two districts with the same class size can score very differently, so class size is not the only thing that matters.
- We want one number: the **typical** change in score per one-student change in class size.

A line and its error

A straight line as a summary



One line through the cloud: score $\approx a + b \times$STR.

- The simplest summary of the tilt is a **straight line**: score $\approx a + b \times$ STR.
- The **slope** b is the number we want: how much the score changes when STR goes up by one student. A downward line means $b < 0$.
- The **intercept** a is where the line sits: the score the line gives when STR is zero.

Now we can name the parts. Write the outcome as Y (test score) and the thing we explain it with as X (STR).

The line is $\beta_0 + \beta_1 X$, with intercept β_0 and slope β_1.

The part the line cannot explain

- No district sits exactly on the line. Each real score is the line *plus* a leftover: the district scored a little above or below what its class size alone would predict.
- That leftover holds **everything else** about the district: family incomes, how well the school is run, luck on the test day. We give it a name, the **error term** u_i, and write the model:

Key result 1: the single-regressor model ·

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \ldots, n.$$

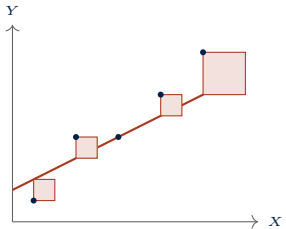
The line $\beta_0 + \beta_1 X$ is the **average** value of Y at each X. The error u_i is the vertical gap between district i and that line.

- The subscript i labels the districts, $i = 1$ up to $i = n = 420$.
- Here is the catch: we do **not** see the true line. We see only the 420 dots. Our job is to use the dots to draw our best estimate of the hidden line.

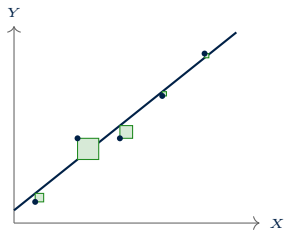
Estimating the line: ordinary least squares

Why we square the mistakes

A ruler and a good eye give everyone a different line; we need a **rule**. For a candidate line, call its miss at district i the **mistake** $e_i = Y_i - (b_0 + b_1 X_i)$.



A poor line: total area = 7



The OLS line: total area = 1.6

- If we simply **added** the mistakes, positive and negative ones would cancel: many bad lines sum to exactly zero, so the plain sum cannot choose.
- So we **square** each mistake and add those. A squared mistake is the area of a square whose side is the mistake: big misses count much more, and nothing cancels.
- **Ordinary least squares (OLS)** is the line that makes the total area smallest: $\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$.

Solving it: the two conditions the best line obeys

- Minimise $S(b_0, b_1) = \sum_i (Y_i - b_0 - b_1 X_i)^2$. Set each partial derivative to zero. Two conditions drop out (the **normal equations**). Write $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$ for the winning line's miss at district i, its **residual**:

$$\frac{\partial S}{\partial b_0} = 0 : \ \sum_i (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \qquad \Rightarrow \qquad \sum_i \widehat{u}_i = 0,$$

$$\frac{\partial S}{\partial b_1} = 0 : \ \sum_i X_i (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \qquad \Rightarrow \qquad \sum_i X_i \, \widehat{u}_i = 0.$$

- The first says the residuals **sum to zero**, so the line passes through the point of averages $(\overline{X}, \overline{Y})$. This gives the intercept at once: $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$.
- The second says the residuals do not move with X. Putting $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$ into it and tidying gives the slope. We read off both on the next slide.

The OLS slope and intercept

Key result 2: the estimators ·

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

- A "hat" means **estimated from the sample**. The slope is the **sample covariance** of X and Y divided by the **sample variance** of X.
- Read it plainly: **how much X and Y move together, per unit of how much X varies on its own**. If they move together (covariance positive) the slope is positive; against each other, negative.
- An equivalent form links it to correlation: $\widehat{\beta}_1 = r_{XY}\,\dfrac{s_Y}{s_X}$, where r_{XY} is the correlation of X and Y. Zero correlation means a zero slope.

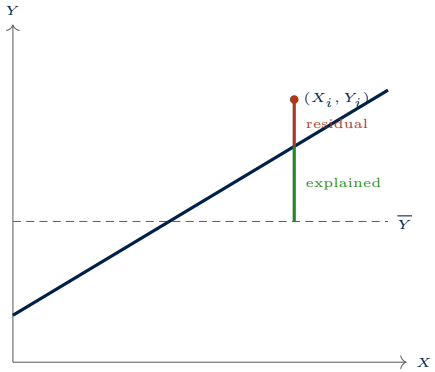
Back to the districts: reading the fitted line

The estimated line on the California data · OLS on all 420 districts gives $\widehat{\text{Score}} = 698.9 - 2.28 \times \text{STR}$.

- First, two names. The **fitted value** $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ is the line's prediction for district i; the **residual** $\widehat{u}_i = Y_i - \widehat{Y}_i$ is the miss we can compute (unlike the true error u_i, which we never see). By the first normal equation, residuals average to zero.
- **Slope** -2.28: a district with one more student per teacher scores, on average, 2.28 points lower.
- **Intercept** 698.9: the line's value at STR $= 0$. No district is there, so this is only where the line meets the axis, **not** a prediction. Reading a line outside your data is extrapolation.
- **Predict**: at STR $= 20$, the line gives $698.9 - 2.28(20) = 653.3$. A district there that scored 660 has a residual of $660 - 653.3 = +6.7$: better than its class size alone would suggest.
- **Same tool, another question**: in a standard US wage sample, $\widehat{\text{wage}} = -0.90 + 0.54 \times (\text{years of schooling})$: each extra year adds about 54 cents an hour.

How well does the line fit?

R^2, the idea: splitting the variation



Each point's distance from $\overline{Y}$ splits at the line.

- The cloud was wide, so the line is not the whole story. Two numbers judge the fit: R^2 (what share it explains) and the **SER** (how big a typical miss is). Build R^2 first.
- Scores vary around their average $\overline{Y}$. For each district, split the distance from the average:

$$\underbrace{Y_i - \overline{Y}}_{\text{total}} = \underbrace{(\widehat{Y}_i - \overline{Y})}_{\text{explained by the line}} + \underbrace{(Y_i - \widehat{Y}_i)}_{\text{residual}}.$$

- The **explained** piece is how far the line moved from the average because X changed; the **residual** piece is what it still missed. A good fit: big explained pieces, small residuals.

R^2, the number: the share of variation explained

Key result 3: the coefficient of determination · Square the split and add over all districts. The pieces add up:

$$\underbrace{\sum_i (Y_i - \overline{Y})^2}_{\text{TSS: total}} = \underbrace{\sum_i (\widehat{Y}_i - \overline{Y})^2}_{\text{ESS: explained}} + \underbrace{\sum_i \widehat{u}_i^2}_{\text{SSR: residual}}, \qquad R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}.$$

- R^2 is the **share** of the total variation in scores that the line explains. $R^2 = 1$ means every point is on the line; $R^2 = 0$ means the line explains nothing. With one regressor, $R^2 = r_{XY}^2$.
- On the California data, $R^2 = 0.051$: class size explains only about **5%** of why districts differ in scores. The other 95% is everything else.
- **A low R^2 is normal, not a failure.** It says other things matter too. It does *not* say the slope -2.28 is wrong or useless.

The standard error of the regression

Key result 4: the typical size of a miss ·

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}.$$

- Where R^2 is a share, the SER is measured in the **units of** Y. It is roughly the average size of a residual: how far, in points, a typical district sits from the line.
- On the California data the SER is about **18.6 points**. So even knowing a district's class size, a typical prediction of its score is off by around 18.6 points. That is the wide cloud, put in numbers.
- We divide by $n-2$, not n, because two numbers (the slope and the intercept) were already estimated from the same data. When n is large, the choice barely matters.

When can we trust the slope?

The real worry: is -2.28 the effect of class size?

- OLS will always draw a line. The hard question is whether the slope is the **effect** of class size, or whether something else is hiding in it.
- Think about **which districts have small classes**. They tend to be richer districts, which can afford more teachers. Richer districts also have more educated parents, better resources, and so on, all of which raise scores on their own.
- So a small class size travels together with many advantages. The slope -2.28 may be crediting class size with gains that really came from income. This is **confounding**, and it is the central danger in all observational data.

The next three conditions, together the **least squares assumptions**, say when this danger is absent and the OLS slope can be trusted. The first is the one that does the real work.

Assumption 1: the error averages to zero at each X

Key result 5a ·

$$\mathbb{E}(u \mid X = x) = 0 \quad \text{for every } x.$$

- Whatever the class size, the other factors in u average out to zero. They do not lean one way as X changes.
- If they do lean (income falls as STR rises), OLS blames X for their effect and the slope is **biased**.
- This holds by design in a **randomised experiment**. In observational data like ours, it must be argued, and here it is doubtful.

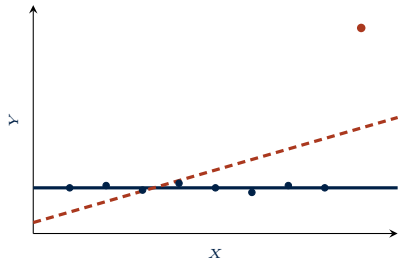
Assumptions 2 and 3

Key result 5b and 5c ·

2. (X_i, Y_i) are a **random sample**: independent and identically distributed across districts.
3. **Large outliers are unlikely**: X and Y have finite fourth moments.

- **Assumption 2** holds when each district is a fresh, comparable draw from one population. It is what lets us describe, below, how the estimate would vary across samples.
  - **Where it fails: time series**, such as one asset's return recorded month after month. Today's value is tied to yesterday's, so the draws are not independent. We handle that later.
- **Assumption 3** keeps a few extreme points from dominating. Because OLS squares the mistakes, a single wrong or freak value can swing the whole line, as the next slide shows. Always plot and check.

One point can move the whole line



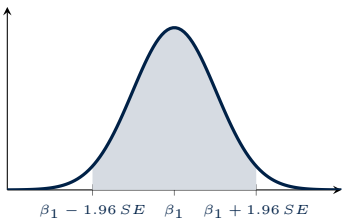
Solid: no outlier. Dashed: the one red point drags the line up.

- The blue points show no relationship: the true line is flat.
- Add one extreme point (red) and OLS, which squares every miss, bends the line sharply towards it.
- One bad observation can manufacture a relationship that is not there, or hide a real one.

Outliers are often data errors: a misplaced decimal, a wrong unit. Look at your data before you trust the slope.

How uncertain is the estimate?

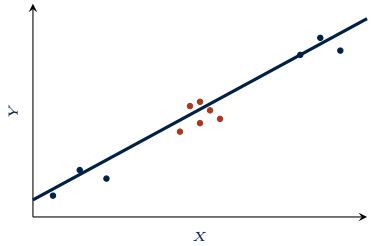
The slope is itself random



The values $\widehat{\beta}_1$ would take across many samples.

- Our -2.28 came from these 420 districts. A different sample would have given a slightly different number. So $\widehat{\beta}_1$ is a **random variable**.
- The distribution of its values across all possible samples is its **sampling distribution**. It is the source of our uncertainty about the true β_1.
- We want its **centre** (right on average?), its **spread** (how much does it bounce?), and its **shape** (what curve?). These three answers are the bridge from an estimate to an inference.

Centred at the truth, and more precise with better data



Blue: X spread wide. Red: X bunched. Same noise, very different certainty about the slope.

Key result 6a: unbiased and consistent ·

$$\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1, \quad \text{and } \widehat{\beta}_1 \to \beta_1 \text{ as } n \text{ grows.}$$

- Under the assumptions, OLS is right **on average**, and it gets closer to the truth with more data.
- The estimate is more **precise** when: there are more districts; X is more **spread out**; and the errors are smaller.
- Through the wide (blue) points you can pin the slope; through the bunched (red) points the same noise allows many slopes. To learn a slope, X must actually *move*.

In large samples the slope is normal

Key result 6b: the large-sample shape · If the least squares assumptions hold, then in large samples

$$\widehat{\beta}_1 \text{ is approximately } N\left(\beta_1,\ \sigma^2_{\widehat{\beta}_1}\right).$$

- By the **central limit theorem**, the sampling distribution of $\widehat{\beta}_1$ is close to a **normal** (bell) curve when n is large, whatever the shape of the data.
- It is centred at the true β_1, with a spread given by the drivers on the previous slide. We write that spread $\sigma^2_{\widehat{\beta}_1}$; its square root, the **standard error**, is next.
- Everything that follows rests on this bell shape: because we know the curve, we know how *surprising* any estimate would be if the truth were some particular value. That is exactly what a test needs, and we build one now.

The standard error of the slope

Key result 7: the slope's own standard error · $SE(\widehat{\beta}_1)$ is the estimated standard deviation of the slope's sampling distribution: roughly how far your one estimate typically lands from the true β_1.

- It plays exactly the role $s/\sqrt{n}$ played for the sample mean. Small SE: repeated samples would give nearly the same slope, so yours is trustworthy. Large SE: the slope is poorly pinned down.
- The formula (software computes it) is heavy but plain in spirit:

$$SE(\widehat{\beta}_1) = \sqrt{\frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_i (X_i - \overline{X})^2 \, \widehat{u}_i^2}{\left[\frac{1}{n}\sum_i (X_i - \overline{X})^2\right]^2}} \qquad \leftarrow \text{error noise on top, spread of } X \text{ below.}$$

  More noise around the line widens it; more spread in X and more data shrink it: the same three drivers as before, now in one number.
- On the California data, $SE(\widehat{\beta}_1) = \mathbf{0.52}$. Our slope is -2.28, give or take about half a point.

Testing a claim about the slope

The sceptic's claim, and how a test treats it

Two competing claims about the true slope ·

- **Null hypothesis** $H_0: \beta_1 = 0$. Class size does not matter: the true line is flat, and our -2.28 is only sampling luck.
- **Alternative** $H_1: \beta_1 \neq 0$. There is a real relationship, in one direction or the other.

- Treat it as a courtroom. The null is **presumed true**; we ask only whether the data are strong enough to **reject** it. We never set out to "prove" the null, and failing to reject it is not proving it.
- The hypothesised value need not be zero: you can test $H_0 : \beta_1 = \beta_{1,0}$ for any number $\beta_{1,0}$ a theory or a claim supplies. Zero is simply the most common sceptic.
- We use the **two-sided** alternative by default. Fixing one direction in advance is justified only when the other direction is genuinely impossible, which is rare.

The t-statistic

Key result 8: distance from the null, in standard errors ·

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error}} = \frac{\widehat{\beta}_1 - \beta_{1,0}}{SE(\widehat{\beta}_1)}.$$

- The logic: if the null were true, the bell of $\widehat{\beta}_1$ would be centred at $\beta_{1,0}$, and estimates more than about two standard errors away would be rare. So count how far away we landed.
- Under H_0, t is approximately standard normal in large samples, so a value beyond ± 1.96 happens only 5% of the time by chance: $|t| > 1.96$ means "surprising at the 5% level".

The class-size slope against the sceptic ·

$$t = \frac{-2.28 - 0}{0.52} = \textbf{-4.38}.$$

More than four standard errors below zero: if class size truly did not matter, a slope this far from flat would almost never appear. **Reject** the null at the 5% level.

The p-value, and what it is not

The chance of a result this extreme, if the null were true ·

$$p = \mathbb{P}(|Z| > |t|) = 2\,\Phi(-|t|), \qquad Z \sim N(0,1),$$

where Φ is the standard normal cumulative distribution function. Reject at 5% when $p < 0.05$.

- For the class-size slope, $t = -4.38$ gives $p \approx \mathbf{0.00001}$: if class size truly did not matter, a tilt this strong would appear about once in a hundred thousand samples. That is why we reject.
- Read p plainly: it measures how *surprising* the data are under the null, so a small p is evidence against the null. It also says more than a bare verdict, so report the number itself.

A p-value is not the probability that the null is true. It is the probability of data this extreme *given* the null, not the probability of the null given the data.

How results are reported, and how to read them

A regression is reported with each standard error in parentheses under its coefficient:

$$\widehat{\text{Score}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times \text{STR}, \qquad R^2 = 0.051, \quad SER = 18.6.$$

- This one line carries everything. The estimate answers **how big**; the number in parentheses answers **how sure**; R^2 and the SER report the fit.
- You can redo the inference from the printed line alone: $t = -2.28/0.52 \approx -4.4$, beyond 1.96, so the slope is significantly different from zero at 5%. No software needed.
- This is how results appear in papers and reports. Reading such a line at a glance (estimate, precision, fit) is the practical skill you will use in every empirical paper.

A range for the truth: confidence intervals

The confidence interval

Key result 9: a plausible range for the true slope ·

$$95\%\ \text{CI for } \beta_1: \qquad \widehat{\beta}_1 \pm 1.96\, SE(\widehat{\beta}_1).$$

It is exactly the set of null values that a two-sided 5% test would *not* reject.

- The reasoning runs the test backwards. Instead of asking "is zero plausible?", ask "**which** values are plausible?": every $\beta_{1,0}$ within 1.96 standard errors of the estimate survives the test, and every value further away is rejected.
- For the class-size slope:

$$-2.28 \ \pm \ 1.96 \times 0.52 \ = \ \textbf{[-3.30, -1.26]}.$$

- Zero is far outside, so the interval and the test give the same verdict. But the interval says more: it shows *how big* the effect plausibly is, not just that it is not zero.

What the interval means, and what it does not

Coverage: a statement about the method · Over many random samples, 95% of the intervals built this way contain the true β_1. The **coverage** is a property of the *procedure*, not of any one interval.

The wrong reading: "there is a 95% probability that β_1 lies in *this* interval".

- Once computed, an interval either contains the true β_1 or it does not. The truth is a fixed number; it is the *interval* that changes from sample to sample.
- The 95% describes how often the recipe works across samples, not the odds for your single interval.
- The half-width, $1.96\,SE$, is the **margin of error**. It shrinks like $1/\sqrt{n}$: to halve your uncertainty you need four times the data.

The answer the superintendent can use

- She never asked about one student. Her plan is to cut the ratio by **two** students. The predicted gain is $\beta_1 \times (-2)$, so multiply both ends of the interval by -2 (the signs flip):

From the slope interval to the policy interval ·

slope interval $[-3.30, -1.26]$ $\xrightarrow{\times(-2)}$ predicted gain between **2.5** and **6.6** points.

- This is the honest deliverable: not "class size is significant", but "cutting two students is associated with a gain of roughly 2.5 to 6.6 points, and we can now weigh that against the cost of the teachers".
- An interval answers the question a decision-maker actually asks: **how big**, and **how sure**. A bare "significant" answers neither.

Summary

The whole argument on one slide

- **Question**: what does one more student per teacher do to test scores?
- **Line (OLS)**: $\widehat{\text{Score}} = 698.9 - 2.28 \times \text{STR}$; the slope is the covariance of X and Y per unit of variance of X.
- **Fit**: $R^2 = 0.051$, about 5% explained; SER $= 18.6$ points, the typical miss. A low R^2 is normal and does not undermine the slope.
- **Trust**: the slope is the *effect* of class size only if the error averages to zero at every X. Doubtful here: income travels with class size. The fix is **multiple regression**, next topic.
- **Uncertainty**: $SE(\widehat{\beta}_1) = 0.52$, so $t = -4.38$ and $p \approx 0.00001$: not sampling luck. The 95% interval is $[-3.30, -1.26]$.
- **Answer**: cutting two students is associated with a gain of about 2.5 to 6.6 points, a range to weigh against the cost, remembering the causal caveat.

> An estimate answers "how big"; a standard error answers "how sure". A defensible empirical answer always carries both, and in real data the standard error must be computed robustly.