Lecture 1 — Foundations of Univariate Time Series

Course map, dependence, stationarity, ergodicity, mixing, and Wold decomposition

Jiajing Sun†

School of Economics and Management, University of Chinese Academy of Sciences†

March 2026

Lecture 1 in the 20-lecture sequence

- **Today:** the probabilistic foundations of univariate time series.
- **Lecture 2:** AR, MA, ARMA, lag polynomials, causality, and invertibility.
- **Lecture 3:** ACF, PACF, model identification, estimation, and forecasting.
- **Later blocks:** nonstationarity, VAR / VECM / SVAR, volatility, nonparametrics, robust inference, filtering, and continuous-time finance.

Why start here?

Without stationarity, ergodicity, mixing, and the Wold theorem, most later econometric machinery has no clean probabilistic foundation.

Learning goals for this 3-hour lecture

By the end of this lecture, students should be able to:

1. explain why time series data are different from i.i.d. samples;
2. distinguish strict stationarity from weak stationarity;
3. interpret autocovariance and autocorrelation functions;
4. explain why ergodicity and mixing matter for LLN / CLT arguments;
5. state the Wold decomposition theorem and explain why it motivates ARMA modeling.

A practical plan for the 3 hours

- **Hour 1:** motivation, notation, dependence, stationarity.
- **Hour 2:** examples, nonstationarity, ergodicity, mixing, and asymptotic implications.
- **Hour 3:** Wold decomposition, innovations, representation, and preparation for Lecture 2.

Teaching suggestion

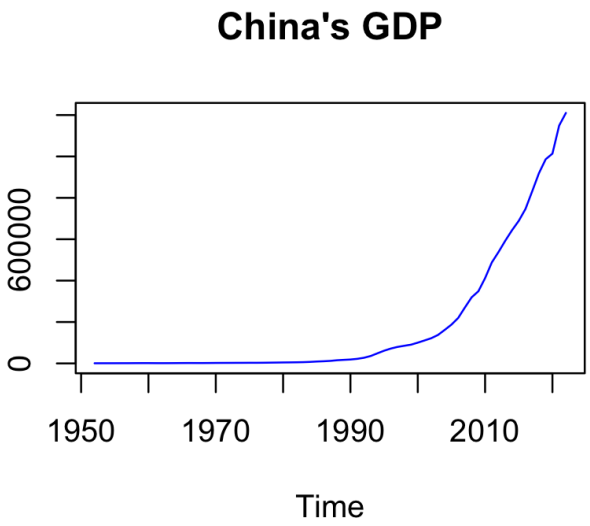
Do not rush the definitions. In this lecture, the vocabulary is the theory.

Lecture Roadmap

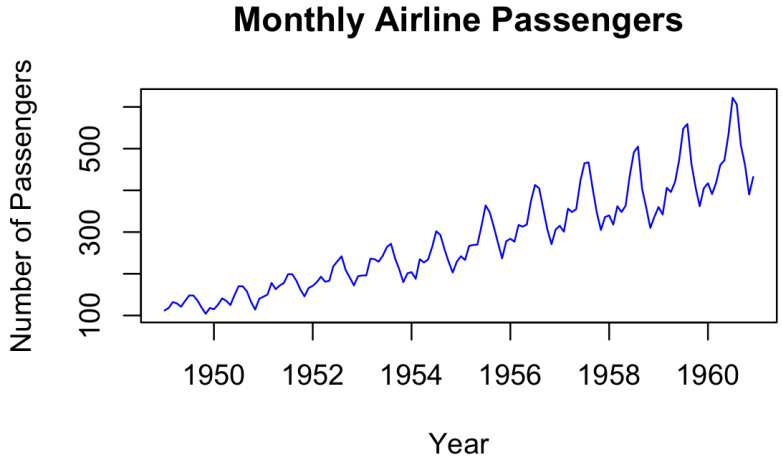
Why do we study time series?

- Many economic and financial variables are observed sequentially: GDP, inflation, interest rates, returns, volatility, exchange rates, consumption, and output.
- The objective is not only description, but also:
  - dynamic interpretation,
  - forecasting,
  - policy evaluation,
  - risk measurement.
- Time ordering creates dependence, and dependence changes both estimation and inference.

Examples of economic and business time series



Real macroeconomic series often contain trend and persistence.



Business and transport series often combine trend, seasonality, and noise.

Typical questions in time-series econometrics

- Is the series stable over time or drifting?
- How persistent are shocks?
- How much of today's value is predictable from the past?
- Are the data mean-reverting, trend-dominated, or shock-accumulating?
- Can we forecast reliably, and how uncertain are those forecasts?

Static data versus dynamic data

Cross-sectional work

The main issue is heterogeneity across units at one point in time.

Time-series work

The main issue is dependence across dates for the same process.

- Yesterday's value may help predict today's value.
- Shocks may propagate, decay, or persist permanently.

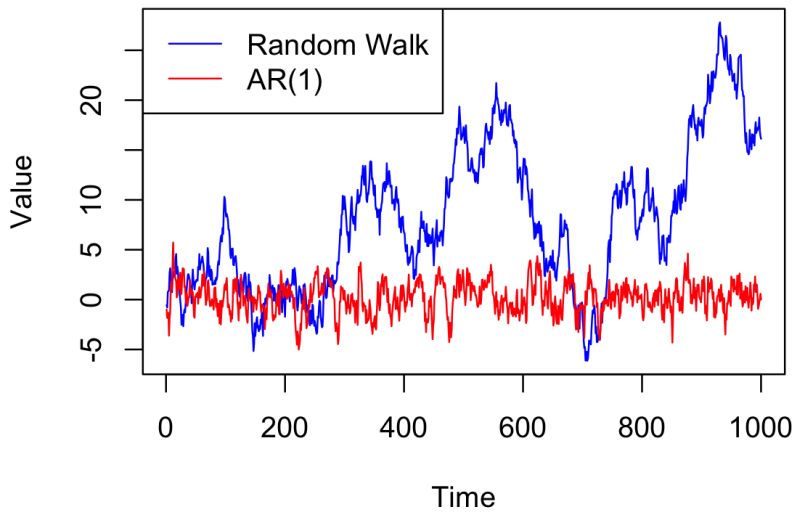
Why dependence changes inference

- In i.i.d. data, the sample mean has variance σ^2 / T.
- In dependent data, the sample mean depends on the whole autocovariance structure.
- Hence standard errors, confidence intervals, and hypothesis tests must be rethought.

Big picture

Dependence is not a nuisance detail; it is the central object of the course.

A first visual contrast: stable versus unstable dynamics



A stable AR(1) fluctuates around a fixed mean; a random walk accumulates shocks over time.

Lecture 1 as the foundation for the rest of Chapter 2

1. Today: foundational probability and representation.
2. Next: AR, MA, ARMA models and lag operators.
3. Then: ACF / PACF, model selection, estimation, and forecasting.

Interpretation

Lecture 1 explains why later parametric models are meaningful at all.

What students should keep in mind from the start

- A time series is one realized path from a stochastic process.
- We usually observe one history, not repeated experimental replications.
- Therefore, assumptions that connect one path to population properties are crucial.

Hour 1 begins here

Motivation, notation, dependence, and stationarity

- First objective: understand what kind of probabilistic object a time series is.
- Second objective: understand what it means for its probabilistic environment to be stable.

Lecture Roadmap

What is a time series?

- A univariate time series is a stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$.
- We observe one realization over time:

$$Y_1, Y_2, \ldots, Y_T.$$

- The same notation may represent levels, growth rates, returns, spreads, or volatility proxies.

Common transformations before modeling

A model is built for the stochastic object we choose to analyze, not automatically for the raw data as recorded.

- **Level:** Y_t, useful for spreads, rates, and already stable series.
- **First difference:** $\Delta Y_t = Y_t - Y_{t-1}$, often used when levels look nonstationary.
- **Growth rate:** $\Delta \log Y_t \approx (Y_t - Y_{t-1}) / Y_{t-1}$.
- **Log return:** $r_t = \log P_t - \log P_{t-1}$, standard in finance.
- **De-meaning / detrending / deseasonalizing:** often necessary before applying stationary tools.

Teaching message

Always ask first: what transformation is most likely to produce a stable probabilistic environment?

One process, one realization, many questions

- The process is the underlying probabilistic mechanism.
- The sample path is the realized history we actually see.
- Econometrics asks: what can one sample path tell us about the underlying process?

Notation for past information

- Let $\mathcal{F}_{t-1}$ denote the information available up to time $t-1$.
- This can be thought of as the sigma-field generated by $Y_{t-1}, Y_{t-2}, \ldots$.
- It formalizes the notion of "the past."

Prediction from the past

- The best mean-squared predictor from past information is

$$\mathrm{E}(Y_t \mid \mathcal{F}_{t-1}).$$

- The unforecastable component is

$$\varepsilon_t = Y_t - \mathrm{E}(Y_t \mid \mathcal{F}_{t-1}).$$

Important

This innovation language reappears later in ARMA models, Kalman filtering, and the Wold decomposition.

Dependence is the central difficulty

- If observations were i.i.d., standard probability theory would do most of the work.
- In time series, serial dependence changes convergence rates and asymptotic variances.
- A mean based on dependent data is not "just" an average.

Variance of the sample mean under dependence

Starting from

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t, \qquad \mathrm{Var}(\bar{Y}_T) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{Cov}(Y_t, Y_s),$$

weak stationarity implies

$$\mathrm{Var}(\bar{Y}_T) = \frac{1}{T}\gamma_0 + \frac{2}{T}\sum_{h=1}^{T-1}\left(1 - \frac{h}{T}\right)\gamma_h, \qquad \gamma_h = \mathrm{Cov}(Y_t, Y_{t-h}).$$

- The double sum collapses into a weighted sum of autocovariances because lag h appears $T - h$ times.
- Under i.i.d. sampling, $\gamma_h = 0$ for $h \geq 1$, so we recover the familiar σ^2 / T.
- Positive serial correlation inflates uncertainty; negative serial correlation can reduce it.
- This is the first place where the *entire* dependence structure enters econometric inference.

A preview of long-run variance

If the autocovariances are absolutely summable, then the *long-run variance*

$$\Omega = \sum_{h=-\infty}^{\infty} \gamma_h = \gamma_0 + 2\sum_{h=1}^{\infty} \gamma_h$$

is well defined, and

$$T\,\mathrm{Var}(\bar{Y}_T) \to \Omega.$$

Under suitable moment and dependence conditions, this leads to

$$\sqrt{T}\,(\bar{Y}_T - \mu) \Rightarrow N(0, \Omega).$$

- When data are dependent, Ω replaces σ^2 in asymptotic uncertainty calculations.
- This object reappears later in HAC estimation, robust standard errors, and long-run covariance estimation.

Benchmark concept 1: i.i.d. noise

$$\varepsilon_t \sim i.i.d.(0, \sigma^2)$$

means the shocks are independent across time and share the same distribution.

- **Independent**: no temporal dependence in any order.
- **Identically distributed**: the entire marginal law is stable over time.
- This is the strongest benchmark and often the first one used in probability theory.
- Empirically it is often too strong: financial returns may be nearly unpredictable in the mean but clearly dependent in volatility.
- Still, i.i.d. innovations remain a useful starting point for building linear time-series models.

Benchmark concept 2: white noise

$$\mathrm{E}(\varepsilon_t) = 0, \qquad \mathrm{Var}(\varepsilon_t) = \sigma^2, \qquad \mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0 \ (t \neq s).$$

- White noise is a *second-order* concept: it only restricts means, variances, and autocovariances.
- It rules out linear predictability at nonzero lags, but it does *not* rule out all nonlinear dependence.
- Therefore white noise is weaker than i.i.d.
- In practice, a sequence can be white noise in levels and still exhibit dependence in squares or other transforms.

Key empirical example

ARCH/GARCH innovations are often uncorrelated in levels but dependent in volatility.

Benchmark concept 3: martingale difference sequence

$$\mathrm{E}(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0.$$

- There is no predictable component in the conditional mean.
- This is stronger than saying the unconditional mean is zero, because it conditions on all past information.
- However, the conditional variance $\mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1})$ may still vary over time.
- Hence a martingale difference sequence can capture mean unpredictability without imposing homoskedasticity or independence.

Financial interpretation

Asset returns are often modeled as approximately martingale differences in the mean, even when their volatility is strongly time varying.

How the benchmark concepts are related

Useful implications

$$i.i.d. \Longrightarrow \text{martingale difference}, \qquad i.i.d. \Longrightarrow \text{white noise}.$$

If a martingale difference sequence also has constant finite variance, then it is white noise.

Implications that generally fail

- White noise $\not\Rightarrow$ i.i.d.
- White noise $\not\Rightarrow$ martingale difference.
- Martingale difference $\not\Rightarrow$ constant conditional variance.

- In applied econometrics, residual diagnostics often aim for "close to white noise" rather than true independence.

A compact comparison table

Concept	Main content
i.i.d.	Independent, identically distributed, strongest benchmark.
White noise	Zero mean, constant variance, zero autocovariances.
MDS	Zero conditional mean given the past.
Stationary process	Distribution or moments stable over time.

An important econometric example

Daily asset returns are often written as

$$r_t = \mu_t + \varepsilon_t, \qquad \mathrm{E}(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0,$$

while

$$\mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = h_t$$

is allowed to vary over time.

- In the mean, the series may behave approximately like a martingale difference.
- In the variance, it may display strong serial dependence through volatility clustering.
- This is exactly why "uncorrelated", "white noise", and "independent" should not be treated as synonyms.

Quick checkpoint

Question

Can a series be white noise but not i.i.d.?

- Yes. Zero autocovariance does not rule out nonlinear dependence or conditional heteroskedasticity.
- This distinction matters later for volatility modeling.

Mini-summary before stationarity

- Time series econometrics studies one realization of a dependent process.
- Past information is formalized through $\mathcal{F}_{t-1}$.
- Innovations, white noise, and dependence structure will appear throughout the course.

Lecture Roadmap

Why stationarity comes first

- Time-series analysis asks whether the future is probabilistically comparable to the past.
- Stationarity is the formal way to capture that idea.
- If the probabilistic environment drifts over time, simple learning from history becomes much harder.

Informal idea of stationarity

Informal statement

A stationary process behaves in the same probabilistic way tomorrow as it does today.

- Means do not drift.
- Variability does not explode or collapse.
- Dependence patterns depend on lag, not calendar time.

Strict stationarity: formal definition

Definition

The process $\{Y_t\}$ is **strictly stationary** if, for every $k \geq 1$ and every shift h,

$$(Y_t, Y_{t-1}, \ldots, Y_{t-k+1}) \stackrel{d}{=} (Y_{t+h}, Y_{t+h-1}, \ldots, Y_{t+h-k+1}).$$

- Equivalently, *all* finite-dimensional distributions are invariant to time shifts.
- This means not only the mean and variance, but also skewness, tails, and nonlinear dependence are unchanged over time.
- It is conceptually elegant, but difficult to verify directly from one finite sample path.

How to interpret strict stationarity

- Every finite-dimensional distribution is invariant to time shifts.
- This is a distribution-level concept, not just a moment-level concept.
- It is mathematically elegant but empirically difficult to verify directly.

Weak stationarity: formal definition

Definition

The process $\{Y_t\}$ is **weakly stationary** if

$$\mathrm{E}(Y_t) = \mu < \infty, \qquad \mathrm{Var}(Y_t) = \gamma_0 < \infty,$$

and

$$\mathrm{Cov}(Y_t, Y_{t-k}) = \gamma_k$$

depends only on the lag k, not on the date t.

- Weak stationarity is a *second-order* concept.
- The autocorrelation function is $\rho_k = \gamma_k / \gamma_0$.
- It is the workhorse notion for linear time-series methods, including ACF/PACF analysis, Wold decomposition, and ARMA theory.

Why strict stationarity is hard to verify in practice

- Strict stationarity concerns the full distribution of every finite block of observations.
- With one finite sample path, we almost never have enough evidence to check that property directly.
- Weak stationarity keeps exactly the part of the probabilistic structure needed by linear tools: mean, variance, and covariance.
- This is why most applied time-series work begins with weak stationarity and only adds stronger assumptions later if the method requires them.

Practical rule

When applied papers say "stationary" without qualification, they often mean weakly stationary.

How to interpret weak stationarity

- The mean is time-invariant.
- The variance is time-invariant and finite.
- The covariance between two observations depends only on how far apart they are.

Why this matters

Weak stationarity is the workhorse for linear prediction, ARMA modeling, and second-order inference.

Autocovariance function

For a weakly stationary process,

$$\gamma_k = \mathrm{Cov}(Y_t, Y_{t-k}).$$

- γ_0 is the variance.
- γ_k summarizes linear dependence at lag k.
- The full sequence $\{\gamma_k\}$ describes second-order temporal structure.

Autocorrelation function

$$\rho_k = \frac{\gamma_k}{\gamma_0}.$$

- ρ_k normalizes dependence to the interval $[-1, 1]$.
- It is scale-free and easy to interpret.
- Later, ACF patterns will help us identify ARMA models.

Basic properties of the autocovariance function

- Symmetry: $\gamma_{-k} = \gamma_k$.
- Positive semidefiniteness: covariance matrices built from γ_k must be positive semidefinite.
- If autocovariances decay slowly, shocks may be highly persistent.

Strict and weak stationarity are not the same

Strict does not imply weak

If moments do not exist, strict stationarity does not guarantee finite mean or variance.

Weak does not imply strict

Time-invariant second moments do not guarantee that the full distribution is invariant over time.

Important special case

For a Gaussian process, weak stationarity *does* imply strict stationarity, because the finite-dimensional distributions are fully determined by means and covariances.

Example: strict but not weak stationarity

Let Y_t be i.i.d. Cauchy.

- The sequence is strictly stationary because every finite block has the same joint distribution after any time shift.
- But the Cauchy distribution has no finite mean and no finite variance.
- Therefore the weak-stationarity requirements fail at the moment-existence stage.

Why this matters

Many tools in econometrics are second-order tools, so strict stationarity alone is not enough when moments are undefined.

Example: weak but not strict stationarity

Consider $X_{2t-1} = \varepsilon$ and $X_{2t} = \eta$, where $\varepsilon \sim N(0,1)$, $\eta \sim U(-\sqrt{3}, \sqrt{3})$, and the two are independent.

$$\mathrm{E}(X_t) = 0, \qquad \mathrm{Var}(X_t) = 1,$$

and

$$\gamma_{t,t+s} = \begin{cases} 1, & \text{if } s \text{ is even,} \\ 0, & \text{if } s \text{ is odd.} \end{cases}$$

- The covariance pattern depends only on the lag parity, so second moments are time-invariant.
- But odd and even observations have different marginal distributions.
- Hence the series is weakly stationary but not strictly stationary.

Stationary benchmark 1: i.i.d. noise

- i.i.d. noise is both strictly and weakly stationary, provided moments exist.
- It is the cleanest theoretical benchmark.
- But most economic time series are more dependent than this.

Stationary benchmark 2: white noise

- White noise is weakly stationary with $\gamma_k = 0$ for $k \neq 0$.
- It may still exhibit nonlinear dependence.
- Therefore "no autocorrelation" does not mean "no dependence."

Stationary benchmark 3: AR(1)

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \qquad |\phi| < 1.$$

- This is the basic example of structured stationary dependence.
- The past matters, but shocks decay geometrically.

R illustration: simulating a stationary AR(1)

```
set.seed(1)
n <- 500
phi <- 0.9
eps <- rnorm(n)
y <- numeric(n)
for (t in 2:n) y[t] <- phi * y[t - 1] + eps[t]

plot(y, type = "l", main = "Stationary AR(1)")
acf(y, main = "Sample ACF")
```

When $|\phi| < 1$, the process fluctuates around a stable mean and its dependence decays geometrically.

Mean and variance of a stationary AR(1)

Assume $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$, and $|\phi| < 1$.

$$\mathrm{E}(Y_t) = 0, \qquad \mathrm{Var}(Y_t) = \frac{\sigma^2}{1-\phi^2}.$$

- Finite variance requires $|\phi| < 1$.
- This is why the unit-root boundary is special.

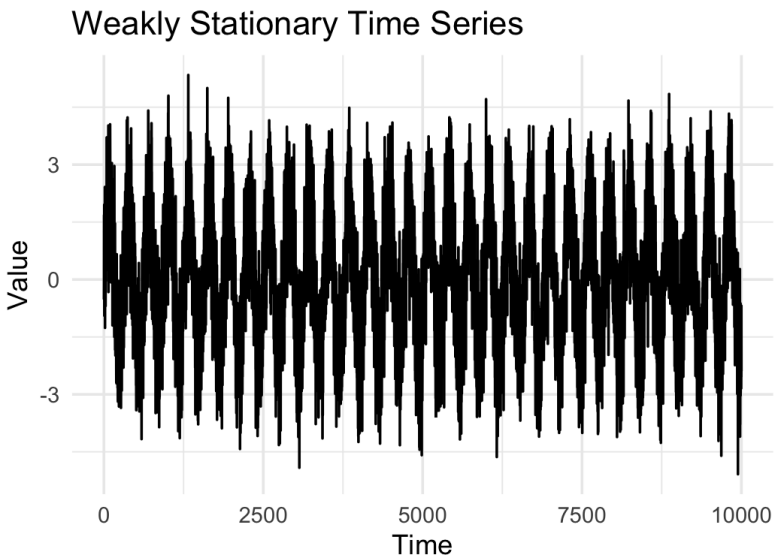
Autocorrelation of a stationary AR(1)

For $Y_t = \phi Y_{t-1} + \varepsilon_t$ with $|\phi| < 1$,

$$\rho_k = \phi^k, \qquad k = 0, 1, 2, \ldots$$

- Dependence decays geometrically.
- Larger $|\phi|$ means stronger persistence.

A stationary-looking series



What to notice:

- fluctuations remain around a stable center;
- the spread looks broadly constant over time;
- shocks seem to die out rather than cumulate permanently;
- visual inspection is useful, but it is never a proof of stationarity.

A nonstationary benchmark: random walk

$$Y_t = Y_{t-1} + \varepsilon_t.$$

- The shock ε_t is added permanently to the level.
- $\mathrm{Var}(Y_t)$ grows with t, so weak stationarity fails.

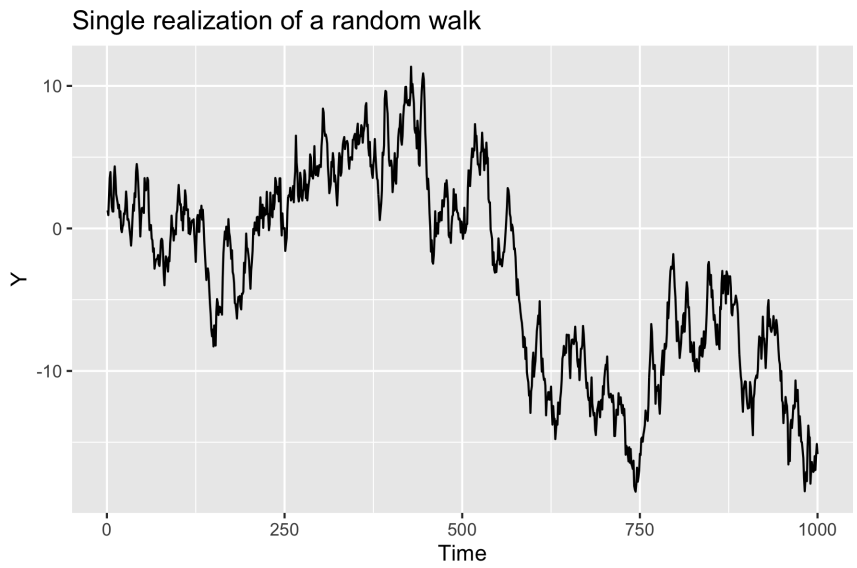
Why a random walk is not weakly stationary

If $Y_t = Y_{t-1} + \varepsilon_t$ and ε_t has variance σ^2, then

$$Y_t = Y_0 + \sum_{j=1}^{t} \varepsilon_j, \qquad \mathrm{Var}(Y_t) = \mathrm{Var}(Y_0) + t\sigma^2.$$

- The variance depends on calendar time.
- So the process cannot be weakly stationary.

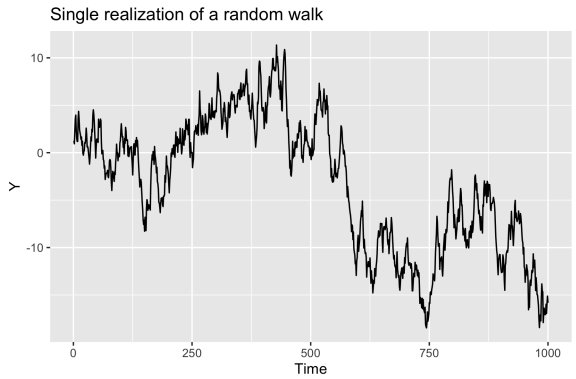
A random walk in pictures



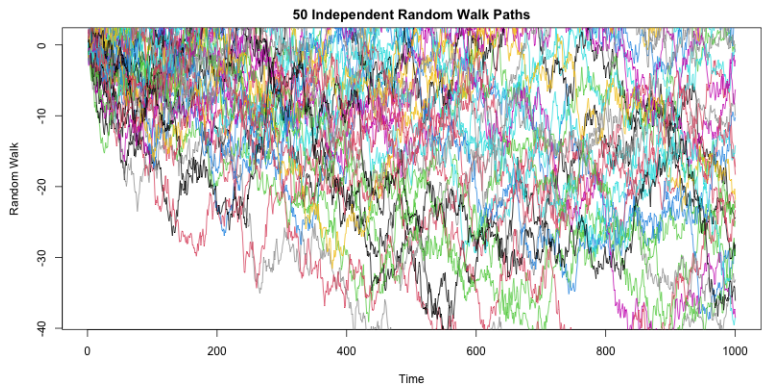
What to notice:

- there is no stable long-run mean around which the series fluctuates;
- the effect of old shocks remains in the current level;
- the dispersion of the process grows with the sample length;
- this is why random walks are nonstationary even if their increments are i.i.d.

One path is deceptive: single realization versus ensemble



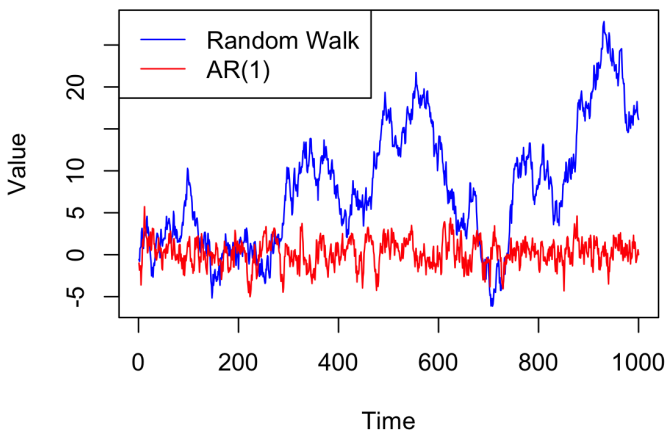
A single path can temporarily look trend-like, flat, or even mean-reverting.



Across many paths, the dispersion widens as shocks accumulate.

This contrast is useful for understanding both nonstationarity and the distinction between a single realization and an ensemble of possible realizations.

Stationary AR(1) versus random walk



- A stable AR(1) forgets shocks geometrically when $|\phi| < 1$.
- A random walk keeps every past shock in its current level.
- Mean reversion and permanent shock accumulation therefore create very different sample paths.

Trend-stationary versus difference-stationary

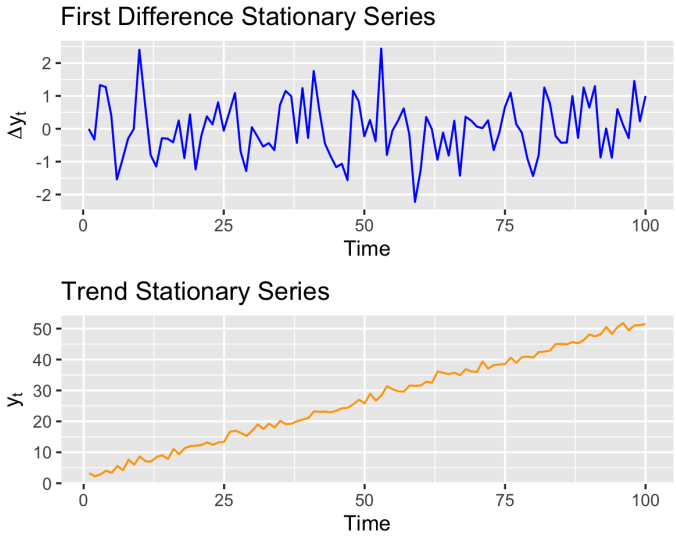
Trend-stationary

$Y_t = f(t) + u_t$, with stationary u_t.
After removing the deterministic trend, the remainder is stationary.

Difference-stationary

ΔY_t is stationary, but Y_t itself is not.
Shocks have permanent effects on levels.

A visual preview of the distinction



- Trend-stationary series become stable after removing a deterministic trend.
- Difference-stationary series become stable only after differencing.
- These are not cosmetic distinctions: they lead to different forecasts and different responses to shocks.

Structural breaks also threaten stationarity

- A one-time shift in mean or variance can destroy stationarity.
- The same is true for changes in persistence or volatility regime.
- In practice, a series may be locally stable but globally unstable.

Useful visual diagnostics

- plot the level series;
- compare early and late subsamples;
- inspect whether volatility is roughly stable;
- ask whether the series appears to revert to a stable center.

These are not formal tests, but they are good economic diagnostics before formal modeling.

R illustration: first inspection of a time series

```
plot(y, type = "l", main = "Series in levels")
plot(diff(y), type = "l", main = "First difference")
acf(y, main = "ACF in levels")
acf(diff(y), main = "ACF of first difference")
```

- In Lecture 1, R is mainly a diagnostic companion.
- Plot the series, inspect transformations, and compare dependence patterns before choosing a formal model.

Algebra of weak stationarity

- If Y_t is weakly stationary, then $a + bY_t$ is also weakly stationary.
- If Y_t and X_t are jointly weakly stationary, then linear combinations remain weakly stationary.
- Linear filtering preserves weak stationarity under mild summability conditions.

Why weak stationarity is the workhorse

- ARMA theory is second-order theory.
- Forecasting formulas rely on covariance structure.
- Spectral analysis is built from autocovariances.
- Robust inference later depends on long-run variance, which is also a covariance object.

Checkpoint 1

Question

Suppose Y_t has constant mean and variance, but $\mathrm{Cov}(Y_t, Y_{t-1})$ changes with calendar time. Is the process weakly stationary?

No. Weak stationarity requires the covariance structure to depend on lag only.

Checkpoint 2

Question

Is every white-noise process i.i.d.?

No. White noise only requires zero autocovariance, not independence.

End of the stationarity block

- Stationarity stabilizes the probabilistic environment.
- Weak stationarity is the central practical concept.
- The next question is: when do time averages from one path reveal population moments?

Hour 2 begins here

Ergodicity, mixing, and why asymptotics work

- We now move from "stability of the process" to "learnability from one history."

Lecture Roadmap

Why stationarity is not enough

- Stationarity says the distribution is stable over time.
- But we observe only one sample path.
- To estimate population quantities from one path, we need time averages to be informative.

The core identification problem

Question

Why should the sample mean of one realized path estimate the population mean of the process?

- This is the role of ergodicity.
- Without it, one path may be misleading even if the process is stationary.

Ergodicity: intuition

In words

A process is ergodic if long-run time averages equal the corresponding population averages.

- History becomes informative.
- One realization can be used for learning.

Ergodicity and the law of large numbers

If $\{Y_t\}$ is stationary and ergodic with finite mean μ, then

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu.$$

- This is the time-series analogue of the law of large numbers.
- It explains why one long realization can be informative about a population moment.
- Without ergodicity, a sample mean may converge to the wrong random object, not to μ.

Common sufficient condition for ergodicity in the mean

A standard sufficient condition is absolute summability:

$$\sum_{k=0}^{\infty} |\gamma_k| < \infty \qquad \text{or} \qquad \sum_{k=0}^{\infty} |\rho_k| < \infty.$$

A sufficient condition for ergodicity in the mean

Starting from weak stationarity,

$$\mathrm{Var}(\bar{Y}_T) = \frac{\gamma_0}{T} \sum_{k=-(T-1)}^{T-1} \left(1 - \frac{|k|}{T}\right) \rho_k.$$

A common sufficient condition for $\mathrm{Var}(\bar{Y}_T) \to 0$ is

$$\sum_{k=0}^{\infty} |\gamma_k| < \infty \qquad \text{or equivalently} \qquad \sum_{k=0}^{\infty} |\rho_k| < \infty.$$

- Absolute summability prevents very persistent dependence from overwhelming averaging.
- Intuitively, the process "forgets" enough of its distant past for the LLN to work.

Ergodicity is about one path being representative

- Imagine many possible histories generated by the same process.
- Ergodicity says a long enough single history behaves like the ensemble average.
- That is why empirical macro and finance can learn from one historical record.

A non-ergodic thought experiment

Consider

$$Y_t = v + \varepsilon_t,$$

where v is drawn once and then fixed forever, while ε_t is white noise.

- Conditional on v, the process looks stable.
- But the sample mean converges to v, not to a unique constant.

Why the previous process is problematic

- Different realizations have different long-run averages because they carry different values of v.
- So one path cannot identify one common population mean.
- This shows stationarity alone is not enough for learning.

What ergodicity buys us empirically

- consistency of sample means and autocovariances;
- meaningful long-run averages;
- a bridge from theory to observed history.

Mixing: another key idea

- Ergodicity tells us that time averages converge.
- Mixing explains how the process forgets its remote past.
- Distant blocks become approximately independent as the lag grows.

Mixing: intuition in one sentence

Intuition

Strong short-run dependence is allowed; what matters is that sufficiently distant observations are only weakly related.

A standard strong-mixing coefficient

Let $\mathcal{F}_a^b = \sigma(Y_a, \ldots, Y_b)$. The strong-mixing coefficient is

$$\alpha(k) = \sup_{t \in \mathbb{Z}} \sup_{A \in \mathcal{F}_{-\infty}^t,\, B \in \mathcal{F}_{t+k}^{\infty}} \left| \mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B) \right|.$$

- A is an event determined by the distant past; B is an event determined by the distant future.
- $\alpha(k)$ measures the largest departure from independence at gap k.
- Strong mixing means $\alpha(k) \to 0$: as the gap grows, the future becomes nearly independent of the past.
- This is stronger than plain ergodicity because it gives quantitative control over dependence decay.

Other mixing notions used in the book

- **Beta mixing** (*absolute regularity*): controls the average gap between conditional and unconditional probabilities.
- **Phi mixing** (*uniform mixing*): requires uniform convergence of $P(B \mid A)$ to $P(B)$; it is quite strong.
- **Rho mixing**: controls maximal correlations between square-integrable functions of the distant past and future.

Useful hierarchy

Roughly speaking, ϕ-mixing is strongest, while β-mixing and ρ-mixing both imply α-mixing; converse implications usually fail.

- In applications, we impose the weakest condition that still delivers the LLN or CLT we need.

How to read the strong-mixing definition

- A depends only on the distant past.
- B depends only on the distant future.
- If the joint probability is close to the product of probabilities, then remote blocks are almost independent.

Why mixing matters for asymptotics

- Mixing conditions help prove LLNs and CLTs for dependent sequences.
- They support consistency and asymptotic normality of estimators.
- They justify HAC estimators, bootstrap methods, and other robust tools.

From mixing to central limit theory

- Under suitable moment and mixing conditions,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (Y_t - \mu)$$

  can converge to a normal limit.
- The asymptotic variance is typically the long-run variance, not just γ_0.

Examples of mixing processes

- Many stable ARMA processes are mixing under mild conditions.
- Many GARCH processes are also mixing, though proofs are more delicate.
- Short-memory stationary processes are often good candidates.

Non-examples or difficult cases

- A random walk is not stationary, so the usual stationary mixing framework does not apply.
- Processes with permanent common components can fail ergodicity.
- Long-memory processes need special care because dependence decays slowly.

Practical message for econometrics

Key message

Dependence is not the problem; uncontrolled or badly behaved dependence is the problem.

- Good asymptotic theory requires structure.
- Ergodicity and mixing are part of that structure.

Checkpoint 3

Question

Can a stationary process fail to be ergodic?

Yes. The process $Y_t = v + \varepsilon_t$ is the standard warning example.

Checkpoint 4

Question

Why do econometricians care about mixing rather than only stationarity?

Because stationarity stabilizes the environment, while mixing helps justify asymptotic inference under dependence.

End of the ergodicity / mixing block

- Ergodicity explains why one long history can be informative.
- Mixing helps explain why LLN / CLT arguments still work under dependence.
- We are now ready for the general representation theorem for stationary processes.

Hour 3 begins here

Wold decomposition and the representation view of time series

- The next theorem explains why ARMA-type models are natural approximations to stationary data.

Lecture Roadmap

Why representation theorems matter

- Definitions tell us what a stationary process is.
- Representation theorems tell us how to write such a process in a useful way.
- Wold is the key representation theorem for covariance-stationary time series.

The question Wold answers

Question

Once we know that a process is covariance stationary, what general form can it take?

Wold's theorem says: a very general one-sided linear innovation representation exists.

Wold decomposition theorem: statement

Wold decomposition

If $\{Y_t\}$ is covariance stationary with finite variance, then

$$Y_t = v_t + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}, \qquad \sum_{j=0}^{\infty} \theta_j^2 < \infty,$$

where v_t is linearly deterministic and $\{\varepsilon_t\}$ is a white-noise innovation sequence.

- No mixing assumption is required for the representation itself.
- The stochastic component is a one-sided linear filter of current and past innovations.
- The deterministic component is perfectly predictable from the infinite past.
- This is the conceptual reason ARMA models are natural rather than ad hoc.

What is the deterministic component?

- v_t is perfectly predictable from the infinite past.
- It contains the linearly deterministic part of the process.
- In many standard stochastic models, $v_t = 0$.

What is the nondeterministic component?

- $\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$ accumulates current and past innovations.
- It is a one-sided moving-average representation.
- The coefficients are square summable, so the variance is finite.

Why the theorem is powerful

- It is completely general for covariance-stationary processes.
- It is not just about ARMA models; ARMA models are special low-dimensional cases.
- It tells us that innovations are the basic building blocks of stationary dynamics.

Innovations are forecast errors

Define the innovation as

$$\varepsilon_t = Y_t - \mathrm{E}(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots).$$

- It is the part of Y_t not linearly forecastable from the infinite past.
- This makes Wold fundamentally a theorem about prediction.

Orthogonality of innovations

Define the innovation by

$$\varepsilon_t = Y_t - \mathrm{E}(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots).$$

Then

$$\mathrm{E}(\varepsilon_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$$

and, for every $j \geq 1$,

$$\mathrm{Cov}(\varepsilon_t, Y_{t-j}) = 0.$$

- The innovation is the one-step-ahead linear forecast error.
- Orthogonality is what makes ε_t "new information".
- Forecasting and representation are therefore two sides of the same theorem.

A recursive projection intuition

- Project Y_t onto its past.
- The residual is ε_t.
- Then project again recursively to express Y_t as accumulated innovations plus any deterministic remainder.

AR(1) as an MA(∞) process

If

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \qquad |\phi| < 1,$$

then repeated substitution gives

$$Y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots .$$

What the AR(1) example teaches us

For $Y_t = \phi Y_{t-1} + \varepsilon_t$ with $|\phi| < 1$,

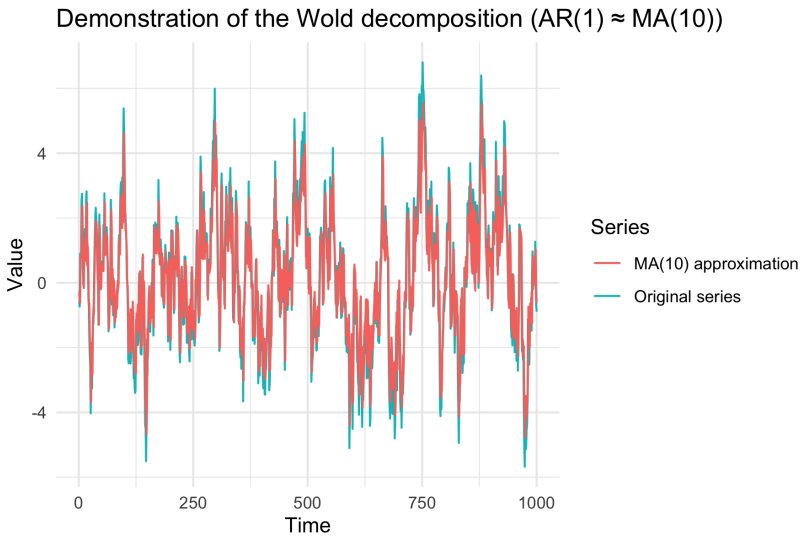
$$Y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}.$$

- A stable autoregression can always be rewritten as an infinite moving average.
- The coefficient ϕ^j tells us how much a shock from j periods ago still matters today.
- The closer $|\phi|$ is to one, the more slowly shocks die out.
- The ACF $\rho_k = \phi^k$ mirrors the same geometric persistence.

An MA(∞) interpretation of persistence

- The coefficients $1, \phi, \phi^2, \ldots$ tell us how quickly shocks decay.
- If $|\phi|$ is close to 1, the process is highly persistent.
- If $|\phi|$ is small, shocks die out rapidly.

A visual illustration of Wold approximation



- The original AR(1) series and the finite MA approximation track each other closely.
- Over a finite sample, the infinite Wold representation can often be approximated well by a truncated moving average.
- This is the practical bridge from theorem to model building.

R illustration: approximating Wold with a finite MA model

```
set.seed(123)
n <- 1000
phi <- 0.9
epsilon <- rnorm(n)
Y <- numeric(n)
Y[1] <- epsilon[1]
for (t in 2:n) Y[t] <- phi * Y[t - 1] + epsilon[t]

fit <- forecast::Arima(Y, order = c(0, 0, 10), include.mean = FALSE)
fitted_values <- fitted(fit)
```

A stable AR(1) is theoretically an MA(∞) process; a finite-order MA model gives a workable approximation over a finite sample.

From Wold to ARMA

- Wold gives an MA(∞) representation.
- Empirical work seeks parsimonious low-dimensional approximations.
- Finite-order AR, MA, and ARMA models provide exactly that.

Why ARMA is not an arbitrary model class

- Wold says every covariance-stationary process has a one-sided innovation representation.
- ARMA models impose a parsimonious parametric structure on that representation through finite autoregressive and moving-average polynomials.
- The point is not that every real-world process is exactly ARMA, but that ARMA gives a tractable approximation to the underlying dependence structure.
- This is why Lecture 2 can move naturally from Wold to AR, MA, and ARMA dynamics.

Autocovariances from the Wold representation

If $v_t = 0$ and

$$Y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}, \qquad \mathrm{Var}(\varepsilon_t) = \sigma^2,$$

then

$$\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} \theta_j^2, \qquad \gamma_k = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+k}, \quad k \geq 1.$$

- The entire second-order structure is encoded in the innovation weights $\{\theta_j\}$.
- Fast decay of θ_j typically produces short memory; slow decay can produce strong persistence.

Long-run variance from Wold

If the coefficients are absolutely summable, then

$$\Omega = \sum_{k=-\infty}^{\infty} \gamma_k = \sigma^2 \left(\sum_{j=0}^{\infty} \theta_j \right)^2 .$$

- The long-run variance depends on the *sum* of impulse weights, not only on the innovation variance.
- If impulse weights accumulate strongly, shocks have a large cumulative effect on sample averages.
- This is another place where Wold connects directly to robust inference for dependent data.

A note on long memory

- If the weights θ_j decay slowly, dependence can remain strong even at long lags.
- Then autocovariances may fail to be absolutely summable.
- Such processes require special tools beyond short-memory ARMA intuition.

Wold and linear prediction

- The deterministic component is perfectly forecastable.
- The innovation component is the irreducible uncertainty.
- Thus prediction decomposes naturally into a predictable part and a new-shock part.

Wold as a bridge between probability and econometrics

- Probability side: a general theorem about stationary processes.
- Econometric side: the conceptual reason we fit linear dynamic models.
- Forecasting side: a theorem about innovations and projection errors.

What Lecture 2 will build on

Lecture 2 will turn the Wold intuition into concrete model classes:

- AR(p), MA(q), and ARMA(p, q),
- lag polynomials,
- causality and invertibility,
- dynamic interpretation of shocks.

Checkpoint 5

Question

Why is Wold's theorem a natural justification for ARMA modeling?

Because it says every covariance-stationary process has a one-sided innovation representation, and ARMA models are parsimonious approximations to that representation.

Checkpoint 6

Question

What is an innovation in the Wold representation?

It is the part of the current observation that cannot be linearly forecast from the infinite past.

End of the Wold block

- Wold is the master representation theorem for covariance-stationary processes.
- It explains why innovations and filters are the right language for time-series analysis.
- It is the conceptual doorway to ARMA theory.

Lecture Roadmap

Main takeaways from Lecture 1

1. Time series are about dependent data, not just data indexed by time.
2. Stationarity stabilizes the probabilistic environment.
3. Ergodicity lets one long realization reveal population moments.
4. Mixing controls dependence decay and supports asymptotic inference.
5. Wold provides the general representation behind ARMA modeling.

A fuller conceptual map before we close

1. Start with a stochastic process and one observed realization.
2. Decide whether the probabilistic environment is stable enough to justify stationary tools.
3. Ask whether one long realization is representative enough for averages and estimators to work.
4. Use dependence-decay conditions to justify LLN and CLT arguments.
5. Represent the stationary component through innovations, then move to ARMA modeling.

This is the logic of the whole first module

Definitions $\rightarrow$ probabilistic foundations $\rightarrow$ representation $\rightarrow$ parametric modeling.

A compact map of the logic

$$\text{Dependence} \to \text{stationarity} \to \text{ergodicity / mixing} \to \text{Wold} \to \text{ARMA}$$

- This is the conceptual spine of the first part of the course.

Suggested board plan for the instructor

Write the following items carefully and slowly on the board:

- strict stationarity,
- weak stationarity,
- autocovariance and autocorrelation,
- ergodicity,
- strong mixing,
- the Wold representation.

Three questions students should be able to answer now

1. Why is a random walk not weakly stationary?
2. Why is white noise weaker than i.i.d.?
3. Why is Wold a theorem about prediction as well as representation?

Reading for Lecture 2

- Review today's definitions until they can be stated without notes.
- Read the ARMA section of Chapter 2.
- Be ready to explain the AR(1) as MA(∞) derivation on the board.

Final slide: the lecture in one sentence

A time series becomes econometrically useful only when we understand how dependence behaves, how stability is defined, why one history can be informative, and how stationary dynamics can be represented through innovations.