Applied Probability for Computer Science

Isadora Antoniano-Villalobos
isadora.antoniano@unive.it

Master in Computer Science
Ca' Foscari University of Venice

Academic year 2023/2024

Central Limit Theorem and Law of Large Numbers

Central Limit Theorem

Let us consider a sequence of random variables, $X_1, X_2, X_3, \ldots$. The sum of the first n elements of the series is, itself, a random variable,

$$S_n = X_1 + X_2 + \ldots + X_n$$

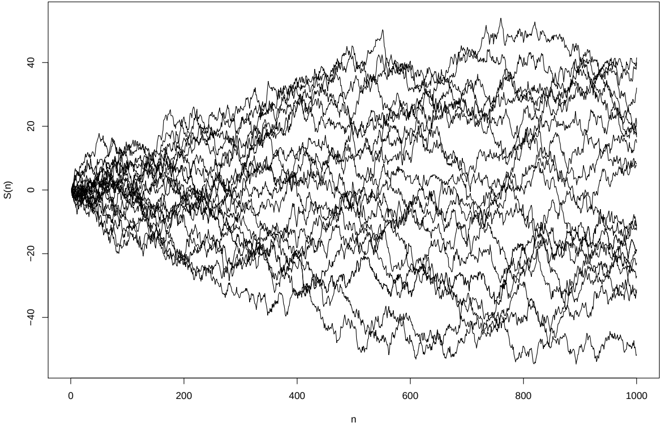
- If the X_i have a common mean, $\mu = \mathbb{E}[X_i]$, then $\mathbb{E}[S_n] = n\mu$
- If the X_i are independent and have common variance $\sigma^2 = \mathrm{Var}[X_i]$, then $\mathrm{Var}[S_n] = n\sigma^2$

☞ But what can we say about the sequence of sums, $S_1, S_2, S_3, \ldots$ as n grows?

We can try to learn something by simulating various realizations of such series using **R**

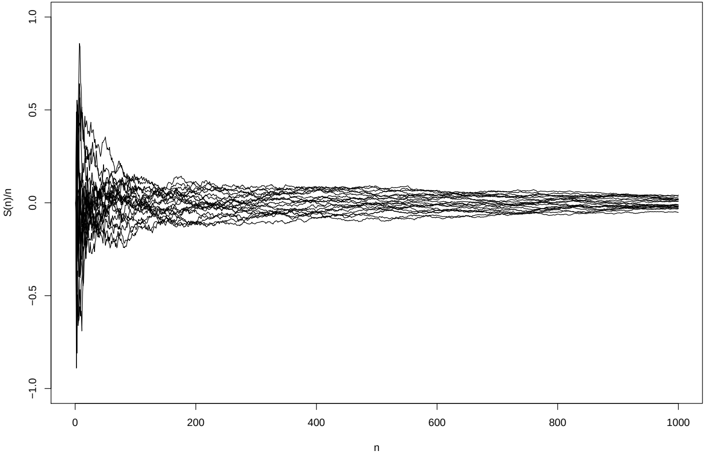
Example 1: $X_i \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$

- Plot of 20 realizations of S_n for $n = 1, \ldots, 1000$



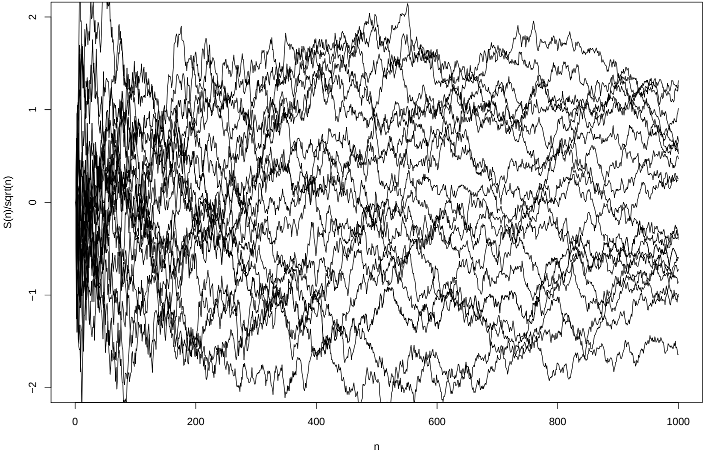
Central Limit Theorem

- Plot of 20 realizations of S_n/n for $n = 1, \ldots, 1000$



Central Limit Theorem

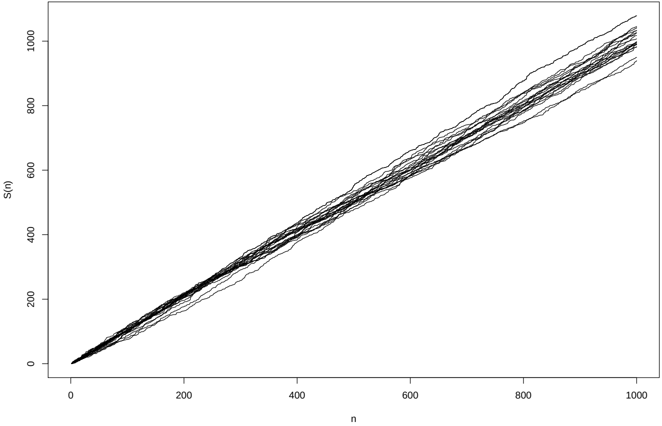
- Plot of 20 realizations of $S_n/\sqrt{n}$ for $n = 1, \ldots, 1000$



Central Limit Theorem

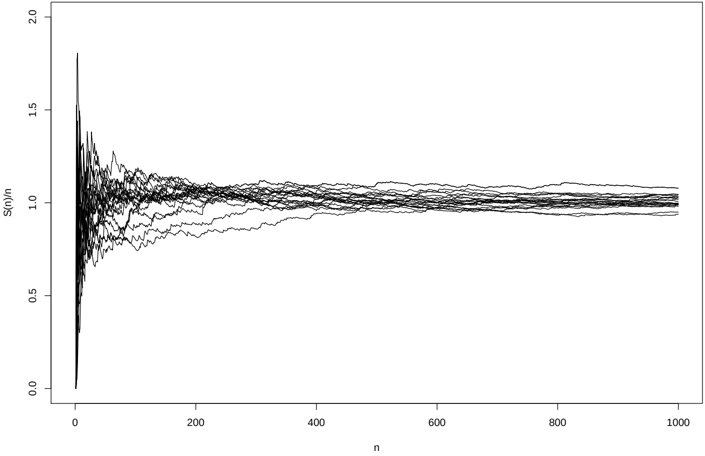
Example 2: $X_i \overset{i.i.d.}{\sim} \text{Exp}\,(1)$

- Plot of 20 realizations of S_n for $n = 1, \ldots, 1000$



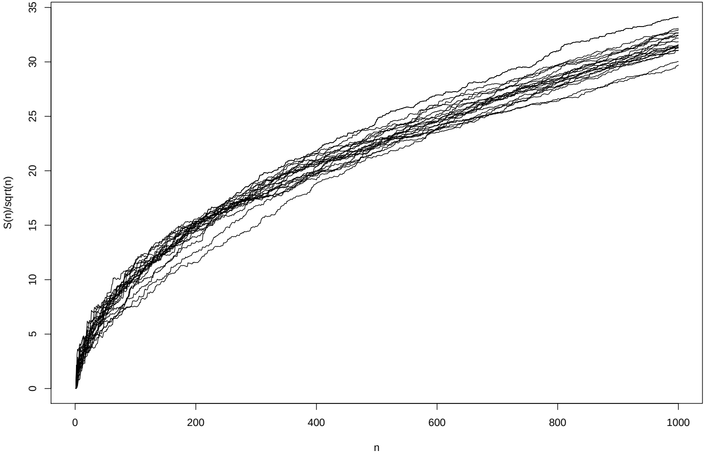
Central Limit Theorem

- Plot of 20 realizations of S_n/n for $n = 1, \ldots, 1000$



Central Limit Theorem

- Plot of 20 realizations of $S_n/\sqrt{n}$ for $n = 1, \ldots, 1000$



Central Limit Theorem

We can summarize our observations as follows:

- The **pure sum** S_n diverges. This is because the variability of S_n grows unboundedly as n goes to infinity,

$$\mathsf{Var}\left[S_n\right] = n\sigma^2 \xrightarrow[n\to\infty]{} \infty$$

- The **average** S_n/n converges, because its variability vanishes as n grows

$$\mathsf{Var}\left[S_n/n\right] = \mathsf{Var}\left[S_n\right]/n^2 = \sigma^2/n \xrightarrow[n\to\infty]{} 0$$

- When we use the normalization factor $1/\sqrt{n}$, we see that $S_n/\sqrt{n}$ has an interesting behavior. When $\mu = 0$ (example 1), it takes values around 0, behaving like some random variable!

Central Limit Theorem

Theorem 1 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be independent random variables with the same expectation $\mu = \mathbf{E}(X_i)$ and the same standard deviation $\sigma = \mathrm{Std}(X_i)$, and let*

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n.$$

As $n \to \infty$, the standardized sum

$$Z_n = \frac{S_n - \mathbf{E}(S_n)}{\mathrm{Std}(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a Standard Normal random variable, that is,

$$F_{Z_n}(z) = \boldsymbol{P}\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right\} \to \Phi(z) \qquad (4.18)$$

for all z.

Central Limit Theorem

Among the random variables discussed in Chapters 3 and 4, at least three have a form of S_n:

Binomial variable	=	sum of independent Bernoulli variables
Negative Binomial variable	=	sum of independent Geometric variables
Gamma variable	=	sum of independent Exponential variables

Hence, the Central Limit Theorem applies to all these distributions with sufficiently large n in the case of Binomial, k for Negative Binomial, and α for Gamma variables.

Example: Normal approximation to the Binomial Distribution

If $X_i \overset{i.i.d.}{\sim}$ Bern (p), then $S_n \sim$ Bin (n, p), so for sufficiently large n and moderate values of p, the Binomial (the distribution of S_n) can be approximated by a Normal with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$

Law of Large Numbers

☞ Carlton-Devore textbook **Section 4.5.4**

- A collection $X_1, X_2, \ldots, X_n$ of independent and identically distributed (i.i.d) random variables is called a **random sample**
- Their average is also called the **sample mean** and denoted $\bar{X} = S_n/n$

LAW OF LARGE NUMBERS

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ, then $\overline{X}$ converges to μ

1. In mean square: $E\left[\left(\overline{X} - \mu\right)^2\right] \to 0$ as $n \to \infty$
2. In probability: $P\left(|\,\overline{X} - \mu\,| \geq \varepsilon\right) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$

Introduction to Stochastic Processes

Stochastic Process: Two (Equivalent) Definitions

Baron (**B**) textbook **Section 6.1**

DEFINITION 6.1

A **stochastic process** is a random variable that also depends on time. It is therefore a function of two arguments, $X(t, \omega)$, where:

- $t \in \mathcal{T}$ is time, with $\mathcal{T}$ being a set of possible times, usually $[0, \infty)$, $(-\infty, \infty)$, $\{0, 1, 2, \ldots\}$, or $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$;
- $\omega \in \Omega$, as before, is an outcome of an experiment, with Ω being the whole sample space.

Values of $X(t, \omega)$ are called *states*.

Stochastic Process: Two (Equivalent) Definitions

At any fixed time t, we have a random variable $X_t(\omega)$, a function of a random outcome. On the other hand, if we fix the outcome ω, we obtain a function of time $X_\omega(t)$. This function is called a **realization**, a **sample path**, or a **trajectory** of the process $X = \{X(t) : t \in \mathcal{T}\}$

☞ Carlton-Devore (**CD**) textbook **Section 7.1**

DEFINITION

For a given sample space $\mathcal{S}$ of some experiment, a **random process** is any rule that associates a time-dependent function with each outcome in $\mathcal{S}$. Any such function that may result is a **sample function** of the random process. The collection of all possible sample functions is called the **ensemble** of the random process.

Types of Stochastic Processes

DEFINITION 6.2

Stochastic process $X(t, \omega)$ is **discrete-state** if variable $X_t(\omega)$ is discrete for each time t, and it is a **continuous-state** if $X_t(\omega)$ is continuous.

DEFINITION 6.3

Stochastic process $X(t, \omega)$ is a **discrete-time process** if the set of times $\mathcal{T}$ is discrete, that is, it consists of separate, isolated points. It is a **continuous-time process** if $\mathcal{T}$ is a connected, possibly unbounded interval.

Types of Stochastic Processes

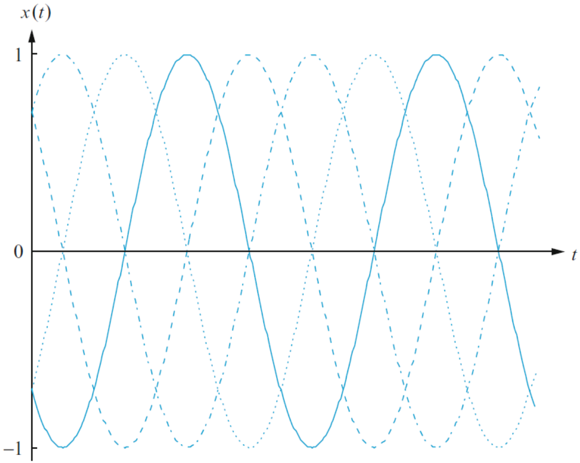
- **CD Example 7.1:** Some communication systems use phase-shift keying to transmit information. A quaternary phase-shift keying (QPSK) system can transmit four distinct symbols (often used to encode two bits at a time: 00, 01, 10, 11). The four symbols are distinguished by varying the phase at which they are transmitted; specifically, for $k \in \{1, 2, 3, 4\}$, the k-th symbol is transmitted for T seconds with the wave

$$x_k(t) = \cos(2\pi f_0 t + \pi/4 + k\pi/2), \quad 0 \leq t \leq T$$

for some predetermined frequency f_0.
Consider the transmission of a single randomly selected symbol, and let $X(t)$ denote the corresponding transmitted wave.
Each function $x_k(t)$ is a sample function and the set of these four functions is the ensemble of $X(t)$

Types of Stochastic Processes

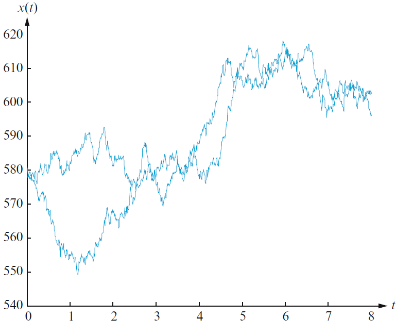
👉 This is a continuous-time, continuous-space process



Types of Stochastic Processes

- **CD Example 7.2:** Let $X(t)$ be the fluctuation in the value of Apple Inc. stock (AAPL) during tan 8-hour trading day, measuring time from the opening bell on Wall Street. The ensemble of $X(t)$ is subject to the constraint $X(0) =$yesterday's closing value. If, for example, if the closing value yesterday was \$580, the ensemble of $X(t)$ consists of all possible paths that the price of Apple stock could hypothetically take tomorrow, starting at \$580 per share.

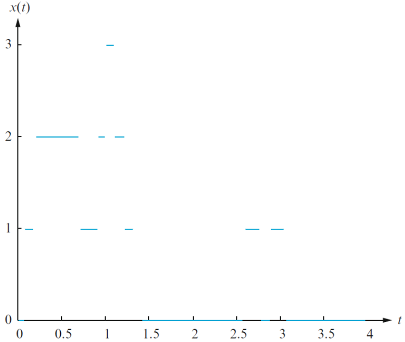
☞ This is a continuous-time, continuous-space process



Types of Stochastic Processes

- **CD Example 7.3:** Consider modeling the number of people $N(t)$ logged in to a specific server at time t (perhaps measured from midnight).
  The ensemble of $N(t)$ consists of all nonnegative integer-valued functions $n(t)$ that might hypothetically arise from successive logins and logouts.

☞ This is a continuous-time, discrete-space process



Types of Stochastic Processes

- **B Example 6.4** In a printer shop, let $X(n)$ be the amount of time required to print the n-th job. ☞ This is a discrete-time, continuous-state stochastic process, because $n = 1, 2, 3, \ldots$, and $X \in (0, \infty)$

  Let $Y(n)$ be the number of pages of the n-th printing job. Now, $Y = 1, 2, 3, \ldots$ is discrete ☞ This process is discrete-time and discrete-state.
- The listing $X_1, X_2, \ldots$, or more simply X_n, is a discrete-time random process, also called a **random sequence**

Note: The difference between discrete- and continuous-space processes is less important than distinguishing how we model time.

Random Processes as Collections of RVs

- At any fixed time point t_0, the ensemble of a random process $X(t)$ forms a probability distribution ☞ $X(t_0)$ is a random variable with support determined by the ensamble

- A random process is characterized by its simultaneous behavior at all time points ☞ To be precise, a random process $X(t)$ is characterized only if we know the **joint distribution** of $X(t_1), \ldots, X(t_r)$ for all finite sets of time points $t_1 < \ldots < t_r$ and $r \in \{1, 2, 3, \ldots\}$. The collection of all such joint distributions constitutes the **finite dimensional distributions** of the process.

In this course

From now on all stochastic processes considered will be in continuous time, unless otherwise stated

Mean and Variance Functions

Main concepts from **CD Subsection 7.2.1**

Recall: Let $X = \{X(t) : t \in \mathcal{T}\}$ be a stocastic process. For each fixed $t \in \mathcal{T}$, $X(t) := X_t$ is a random variable. In particular, it has a mean and a variance which may depend on t

DEFINITION

The **mean function** of a random process $X(t)$ is given by

$$\mu_X(t) = E[X(t)],$$

where $E[X(t)]$ is the expected value of the random variable $X(t)$ for the fixed time point t.

Similarly, we define the **variance function** of $X(t)$ by

$$\sigma_X^2(t) = \mathrm{Var}(X(t)) = E\left[(X(t) - \mu_X(t))^2\right] = E\left[X^2(t)\right] - [\mu_X(t)]^2$$

and the **standard deviation function** of $X(t)$ by $\sigma_X(t) = \sqrt{\mathrm{Var}(X(t))}$.

Mean and Variance Functions

Notice: $\mu_X(t), \sigma^2(t)$, and $\sigma(t)$ are deterministic (nonrandom) functions of t, just as the mean, variance, and standard deviation of a random variable are numbers and not random quantities.

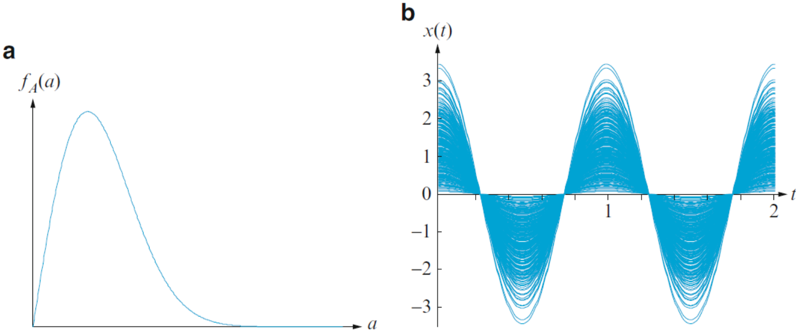
CD Example 7.8:
An intended signal may have the form $\nu_0 + a\cos(\omega_0 t + \theta_0)$, but amplitude variation may occur (for instance, due to natural current or voltage variation), so we can model this as a random process

$$X(t) = \nu_0 + A\cos(\omega_0 t + \theta_0),$$

where A is a random variable whose distribution describes the amplitude variation. Engineers frequently model amplitude variation A with a Rayleigh distribution, $A \sim$Raleigh(σ) with mean and variance

$$\mathbb{E}[A] = \sigma\sqrt{\pi/2}, \quad \text{Var}[A] = \frac{4-\pi}{2}\sigma^2$$

Mean and Variance Functions



(a) Rayleigh pdf for $\sigma = 1$; **(b)** Ensemble of $X(t) = A\cos(2\pi t)$ (for the specifications $\nu_0 = 0, \omega_0 = 2\pi$, and $\theta 0 = 0$)

Mean and Variance Functions

➔ Notice that for each fixed t, so we can apply the properties of expected value and variance to find the mean and variance functions of the process $X\ cos(\omega_0 t + \theta_0)$ is a constant, so

$$\mu_X(t) = \mathbb{E}\left[X(t)\right] = \mathbb{E}\left[\nu_0 + A\cos(\omega_0 t + \theta_0)\right] = \nu_0 + \mathbb{E}\left[A\right]\cos(\omega_0 t + \theta_0)$$

$$\sigma_X^2(t) = \mathsf{Var}\left[X(t)\right] = \mathsf{Var}\left[\nu_0 + A\cos(\omega_0 t + \theta_0)\right] = \mathsf{Var}\left[A\right]\cos^2(\omega_0 t + \theta_0)$$

For example, for the specifications $\nu_0 = 0, \omega_0 = 2\pi$, and $\theta_0 = 0$ with $\sigma = 1$, we get $\mu_X(t) \approx 1.253\cos(t)$, which is again a sinusoid. Notice that in this case $0 \leq \sigma_X^2(t) \approx 0.429\cos^2(t) \leq 0.429$ and the variance is 0 whenever $t = \{1/4, 3/4, 5/4, \ldots\}$, which we can clearly see on the right side (b) of the figure above.

Autocovariance Function

Main concepts from **CD Subsection 7.2.2**

Notice: The mean and variance functions contain information about the behavior of the ensemble at each single point in time. For two different times t and s, the random variables $X(t)$ and $X(s)$ will typically be related ➔ A complete statistical analysis of a random process should also include an exploration of that relationship.

DEFINITION

The **autocovariance function** of a random process $X(t)$ is defined by

$$C_{XX}(t,s) = \mathrm{Cov}(X(t), X(s)) = E[(X(t) - \mu_X(t))(X(s) - \mu_X(s))]$$

Notice that the autocovariance function is a nonrandom function of *two* time points, t and s.

Autocovariance Function

- The autocovariance function is sometimes also denoted $\sigma_X(t,s)$ and, when $t=s$, we recover the variance function $\sigma_X(t,t)=\sigma^2(t)$

Properties of the autocovariance function follow directly from the properties previously derived for covariance. In particular,

PROPOSITION

Let $C_{XX}(t,s)$ denote the autocovariance function of a random process $X(t)$.

1. $C_{XX}(t,s)=C_{XX}(s,t)$
2. $C_{XX}(t,s)=E[X(t)X(s)]-\mu_X(t)\mu_X(s)$
3. $\sigma_X^2(t)=\mathrm{Var}(X(t))=\mathrm{Cov}(X(t),X(t))=C_{XX}(t,t)=E\left[X^2(t)\right]-\mu_X^2(t)$

Autocovariance Function

The autocovariance function of $X(t)$ has the same interpretation as the covariance between two variables:

- If $C_{XX}(t,s) > 0$, when the process X is above its mean function at time t, it also tends to be above its mean function at time s (and vice versa)
- If $C_{XX}(t,s) < 0$, then above-average values of the random process at time t are associated with below-average values at time s (and vice versa).
- $C_{XX}(t,s) = 0$ does not necessarily imply independence

Autocorrelation Function

Warning! The name **autocorrelation function** is used for two different things, depending on the context!

- In the context of time series analysis, the **autocorrelation function**, denoted $\rho_X t, s$ is defined as

$$\rho_X t, s = \frac{\sigma_X(t, s)}{\sigma_X(t)\sigma_X(s)}$$

and its interpretation is analogous to that of the correlation between random variables. In particular, $0 \leq \rho_X t, s \leq 1$ indicates the magnitude and direction of the association between $X(t)$ and $X(s)$

Autocorrelation Function

- However, in the context of signal processing and in the engineering literature, the **autocorrelation function**, denoted $R_{XX}t, s$ is defined as

$$R_{XX}t, s = \mathbb{E}\left[X(t)X(s)\right]$$

and is equivalent to $\rho_X t, s$ only when the mean and variance functions are constant and equal to 0 and 1, respectively. In general, the sign of $R_{XX}t, s$ does not indicate the direction of the association between $X(t)$ and $X(s)$, and its magnitude is not bounded by 1

  - Our textbook **CD** uses this definition!

In this course

The autocorrelation function will not play an important role, as we will focus on other types of properties

Stationarity and Weak Stationarity

- Main concepts from **CD Section 7.3** (subsections 7.3.1 and 7.3.2 excluded)

Informally: We say that a stochastic process is **stationary** if its behavior remains stable over time. But what do we mean by stable?

> **DEFINITION**
> A random process $X(t)$ is **(strict-sense) stationary** if all of its statistical properties are invariant with respect to time. More precisely, $X(t)$ is stationary if, for any time points $t_1, \ldots, t_r$ and any value τ, the joint distribution of $X(t_1), \ldots, X(t_r)$ is the same as the joint distribution of $X(t_1+\tau), \ldots, X(t_r+\tau)$.

This definition requires that the statistical properties of $X(t)$ remain stable over time

Stationarity and Weak Stationarity

- In particular $X(t)$ and $X(t+\tau$ must have the same distribution for all t and all τ ➔ it follows that $X(t)$ must have the same mean, variance, standard deviation, etc. at all times t
- However, the definition requires more. Since the joint distribution of $X(t_1)$ and $X(t_2)$ must be translation-invariant, the autocovariance function of $X(t)$ must be translation-invariant as well
- And this is true for the joint distribution the process at any number of points in time!

☞ It is rarely practical to determine whether a particular random process model is strict-sense stationary. Fortunately, a weaker version of stationarity suffices for the purposes of many analyses.

Stationarity and Weak Stationarity

DEFINITION

A random process $X(t)$ is **wide-sense stationary (WSS)** if the following two conditions hold:

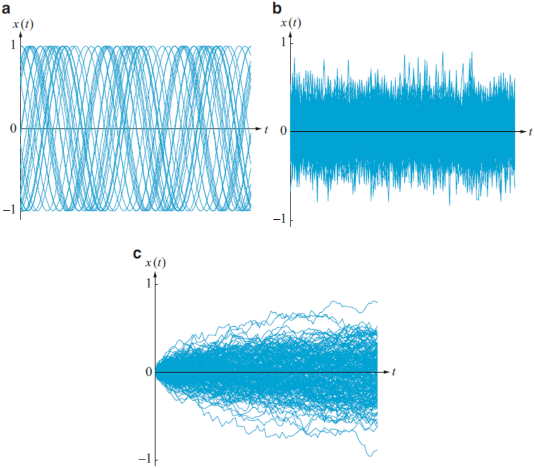
1. The mean function of $X(t)$, $\mu_X(t)$, is a constant.
2. The autocovariance function of $X(t)$, $C_{XX}(t, s)$, depends only on $s - t$.

- A wide-sense stationary process is also called **weakly stationary**, as opposed to a strict-sense stationary process, which is also called **strongly stationary**

Condition 2 states that the degree of association between $X(t)$ and $X(s)$, measured by the covariance, depends only on the distance between the times s and t, but not on the position of those times on an absolute scale.

- For a **weakly stationary**, X both the mean function $\mu_X(t) = \mu_X$ and the covariance function $C_{XX}(t, t+\tau) = C_{XX}(\tau)$ are independent of t

Stationarity and Weak Stationarity



(a) and **(b)** seem weakly stationary; **(c)** clearly is not

Markov Processes

The Markov Property and Markov Chains

☞ Initial definitions from **B Section 6.2** and **CD Section 7.7**

DEFINITION 6.4

Stochastic process $X(t)$ is **Markov** if for any $t_1 < ... < t_n < t$ and any sets $A; A_1, \ldots, A_n$

$$\begin{aligned} \boldsymbol{P}\{X(t) \in A \mid X(t_1) \in A_1, \ldots, X(t_n) \in A_n\} \\ = \boldsymbol{P}\{X(t) \in A \mid X(t_n) \in A_n\}. \end{aligned} \tag{6.1}$$

- For a **Markov process**, the conditional distribution of $X(t)$ is the same under two different conditions:
  1. given observations of the process X at several moments in the past;
  2. given only the latest observation of X

The Markov Property and Markov Chains

- In other words, knowing the present, there is no additional information from the past that can be used to better predict the future,

$$\mathbb{P}\left[\text{future}\middle|\text{past, present}\right] = \mathbb{P}\left[\text{future}\middle|\text{present}\right]$$

- For the future development of a Markov process, only its present state is important, and it does not matter how the process arrived to this state.

☞ Some processes satisfy the Markov property, and some don't

The Markov Property and Markov Chains

(B) Example 6.5: Internet connections

Let $X(t)$ be the total number of internet connections registered by some internet service provider by the time t

- Typically, people connect to the internet at random times, regardless of how many connections have already been made.
- Therefore, the number of connections in a minute will only depend on the current number.

☞ This process is Markov.

(B) Example 6.6 Stock prices

Let $Y(t)$ the value of some stock or some market index at time t

- If we know $Y(t)$ and we want to predict $Y(t+1)$, is it useful to also know $Y(t-1)$?
- One may argue that if $Y(t-1) < Y(t)$, then the market is rising, therefore, $Y(t+1)$ is more likely to exceed $Y(t)$. On the other hand, if $Y(t-1) > Y(t)$, we may conclude that the market is falling and may expect $Y(t+1) < Y(t)$
- It looks like knowing the past in addition to the present does help us to predict the future.

☞ This process is NOT Markov.

The Markov Property and Markov Chains

DEFINITION 6.5

A **Markov chain** is a discrete-time, discrete-state Markov stochastic process.

- More generally, the term **Markov Chain** is used to refer to any discrete-space stochastic process with the Markov property
  ➔ Over the next lessons, we will focus on the study of **Continuous Time Markov Chains (CTMC)**