Study Note

AIS 31 version 3.0

RASE 류지은

Content

1. Probability, Stochastics, and Random Variables [Ch.4.2 in AIS31]

1.1 Definitions and Basic Concepts

488

In the following, Ω denotes a non-empty set.

489

In this document, Ω usually represents the admissible 의도하여 허용되는 values of random numbers, random experiments, or measurements. Usually, Ω is finite (typically, $\Omega = \{0,1\}^k$ or $\Omega = \mathbb{Z}_n$) or it equals $\mathbb{R}^m$ or a subset of $\mathbb{R}^m$ ($m \geq 1$).

Note 1: Experiments with finite Ω are, for example, coin tosses and dice rolls. In the context of RNGs, random numbers are important examples that assume values in a finite or in a countable set Ω, e.g., $\Omega = \{0,1\}$ and $\Omega = \mathbb{N}_0$.

Note 2: Examples for $\Omega \subseteq \mathbb{R}^m$ are timing measurements and voltage measurements.

490

$\mathcal{P}(\Omega)$ denotes the power set of Ω. The power set contains all subsets of Ω. If Ω is finite then $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$.

491

Paragraphs 492 to 501 contain basic definitions and facts from probability and measure theory, which will be needed below for proper definitions of independence or stationary stochastic processes, for example. However, these concepts are rather "technical". Paragraphs 503 to 506 provide a "light version" thereof, which should suffice to understand the subsequent definitions and concepts.

492

A σ-algebra $\mathcal{A}$ structual over Ω is a set of subsets of Ω, i.e., $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, that fulfills the following conditions: axiom

a $\Omega \in \mathcal{A}$

b If $A \in \mathcal{A}$, then also its complement $A^c := \Omega \setminus A \in \mathcal{A}$

c If $A_1, A_2, \cdots \in \mathcal{A}$ then $\bigcup_{n \geq 1} A_n \in \mathcal{A}$

493

Note: Par. 492, Condition (c), includes finite sequences $A_1, A_2, \ldots, A_k$. Note that such a finite sequence can formally be extended by $A_{k+1} = A_{k+2} = \cdots = \{\}$ to an infinite sequence with the same union set.

> It has completeness in itself. $\bigcap_{n \geq 1} A_n \in \mathcal{A}$ with De Morgan's laws by it.

494

Example:

i $\mathcal{P}(\Omega)$ is a σ-algebra over Ω. trivial

ii The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ over $\mathbb{R}$ is the smallest σ-algebra that contains the open intervals (equivalently, that contains the open subsets of $\mathbb{R}$).

iii More generally, for $m \geq 1$, the Borel σ-algebra $\mathcal{B}(\mathbb{R}^m)$ over $\mathbb{R}^m$ is the smallest σ-algebra that contains the open subsets of $\mathbb{R}^m$.

> Most of the sets known at the undergraduate level are the Borel σ-algebra.

495

A **probability measure** 측도: 확률, 길이, 부피 등의 기준에 따른 정도를 부여하는 것 ν on a σ-algebra $\mathcal{A}$ is a mapping $\nu : \mathcal{A} \to [0, 1]$ with the following properties: axiom

1. $\nu(\Omega) = 1$
2. If the sets $A_1, A_2, \cdots \in \mathcal{A}$ are mutually disjoint, $\nu\left(\bigcup_{n \geq 1} A_n\right) = \sum_{n \geq 1} \nu(A_n)$.
   (The sequence $A_1, A_2, \ldots$ may be finite or countable.)

> Since $\mathcal{A}$ is a σ-algebra, the countable(∞) union $\bigcup_{n \geq 1} A_n$ belongs to $\mathcal{A}$, making the term $\nu(\bigcup_{n \geq 1} A_n)$ well-defined.

496

More generally, if a mapping $\nu : \mathcal{A} \to [0, \infty]$ fulfills Condition (b) from par. 495, and if $\nu(\Omega) < \infty$, we refer to ν as a finite measure; otherwise, ν is an infinite measure. If there exists a countable sequence $C_1 \subseteq C_2 \subseteq C_3 \cdots \in \mathcal{A}$ such that $\nu(C_n) < \infty$ for all $n \in \mathbb{N}$ and $\bigcup_{n\geq 1} C_n = \Omega$, then ν is a σ-finite measure.

497

Any $A \in \mathcal{A}$ is said to be an event or a measurable 확률 공간에서는 event만 measure 가능함. 즉, measurable=사건으로 볼 수 있는 set. A pair $(\Omega, \mathcal{A})$ is denoted as a measurable space 아직 measure가 할당되지 않은 space, while the triple $(\Omega, \mathcal{A}, \nu)$ is called a **measure space**. If ν is a probability measure, the triple $(\Omega, \mathcal{A}, \nu)$ is a **probability space**.

Probability space $(\Omega, \mathcal{F}, \Pr)$: (sample space, event set, probability function)

498

Example:

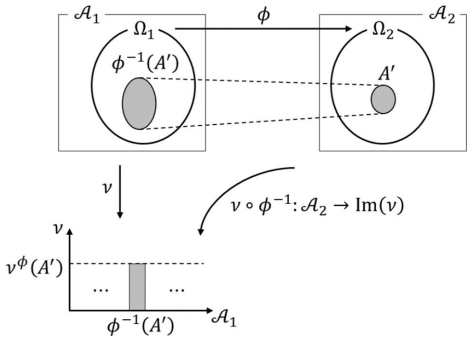
1. Let $B(n, p)$ denote a binomial distribution with parameters n and p. Then $B(n, p)$ is a probability measure on $\mathcal{P}(\{0, \ldots, n\})$.
2. The Lebesgue measure λ is a σ-finite measure on $\mathcal{B}(\mathbb{R})$.
   Note: The Lebesgue measure corresponds to the "geometric" measure Euclid space의 subset에 길이/부피/너비 등을 할당하는 map on $\mathbb{R}$, i.e., $\lambda([a, b)) = b - a$ if $a \leq b$.
3. The standard normal distribution (standard Gaussian distribution) $N(0, 1)$ is a probability measure on $\mathcal{B}(\mathbb{R})$.
4. The Lebesgue measure λ_m on $\mathbb{R}^m$ is a σ-finite measure.

499

If there is no ambiguity 상황에 따라 판단 about the σ-algebra $\mathcal{A}$, we often loosely speak of "measures on Ω". Unless otherwise stated, in this document $\mathcal{A} = \mathcal{P}(\Omega)$ for countable Ω (finite or infinite), and for $\mathbb{R}$, $\mathbb{R}^m$, and measurable subsets $\Omega \subseteq \mathbb{R}^m$, we use the Borel σ-algebras $\mathcal{A} = \mathcal{B}(\mathbb{R})$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^m)$, or $\mathcal{A} = \mathcal{B}(\Omega)$, respectively.

500

Assume that $(\Omega_1, \mathcal{A}_1, \nu)$ is a probability space and that $(\Omega_2, \mathcal{A}_2)$ is a measurable space. Furthermore, let $\phi : \Omega_1 \to \Omega_2$ be a mapping. We call ϕ measurable (or more precisely, $(\mathcal{A}_1, \mathcal{A}_2)$-measurable) if for each $A' \in \mathcal{A}_2$ the pre-image $\phi^{-1}(A') \in \mathcal{A}_1$. If ν is a measure on $\mathcal{A}_1$, then $\nu^\phi(A') := \nu(\phi^{-1}(A'))$ for all $A' \in \mathcal{A}_2$ defines a measure on $\mathcal{A}_2$. We denote ν^ϕ the image measure (or: transformed measure) of ν under ϕ.



In the figure, $\mathrm{Im}(\nu)$ is an image of ν

The function $\nu^\phi(A')$ gives a measure on $\mathcal{A}_\in$.

Furthermore, where X is a random variable onto a mathematical space and Pr is a probability measure of an event set of noise sources, some probability measure $\Pr^X$ gives a probability measure on Mathematical space with regard to the RNG.

501

Assume that $\mathcal{A}_1$ and $\mathcal{A}_2$ are σ-algebras over Ω_1 and Ω_2. A **random variable** X is a measurable mapping $X : \Omega_1 \to \Omega_2$. In our context, usually Ω_2 is finite, countable, or a subset of $\mathbb{R}^m$.

502

Outside of mathematical proofs, the probability space of a random variable is usually not explicitly stated. We point out that a random variable $X : \Omega_1 \to \Omega_2$ with probability space $(\Omega_1, \mathcal{A}_1, \nu)$ can also be interpreted as a random variable on the measure space $(\Omega_2, \mathcal{A}_2, \nu^X)$. Here, ν^X denotes the image measure (or: transformed measure) of X, i.e., $\nu^X(A_2) = \nu(X^{-1}(A_2))$ for all $A_2 \in \mathcal{A}_2$.

The last sentence of Par. 502 "Furthermore, $\mathrm{Prob}(X \in A_1) = \nu(A_1)$ quantifies the probability that the random variable X assumes a value in A_1." seems to need revision for the equation to be "$\mathrm{Prob}(X \in A_2) = \mathrm{Prob}(X^{-1} \in A_1) = \nu(A_1)$."

503

["light version" of pars. 492 to 501] As already mentioned above, these definitions and concepts are needed for mathematically precise definitions in the following. Fortunately, in the context of RNG evaluations, problems concerning measurability hardly occur. The paragraphs 504 to 506 thus provide a "light version". This light version should suffice for at least an intuitive understanding of the following definitions and concepts and to apply them correctly. This, in particular, refers to the material collected in Subsection 4.2.2.

504

["light version" of pars. 492 to 501 ctd.] Some of the following definitions and conditions refer to "measurable subsets" of some space Ω (equivalently, to elements of a σ-algebra on Ω). If Ω is finite or countable, all subsets of Ω are measurable (to be precise: with regard to the σ-algebra $\mathcal{P}(\Omega)$). If $\Omega \subseteq \mathbb{R}^m$ one may think of "regular" subsets as (depending on the dimension m) intervals, rectangles, circles, cuboids, balls, etc. and countable unions thereof. (There exist further measurable and non-measurable subsets, but this should be of little importance for RNG evaluations.)

505

["light version" of pars. 492 to 501 ctd.] In this document and, more generally, in the context of the evaluation of RNGs, random variables usually assume values in finite or countable sets or in subsets of $\mathbb{R}$ or $\mathbb{R}^m$. We may speak of random variables on finite or countable set Ω (e.g., $\Omega = \{0, 1\}$), or random variables on $\mathbb{R}$ (also: "real-valued random variables"), random variables on $\mathbb{R}^m$, or random variables on Ω.

506

["light version" of pars. 492 to 501 ctd.] The expression $\text{Prob}(X \in A)$ quantifies the probability that the random variable X assumes a value in the set $A \subseteq \Omega$.

507

$X \sim \nu$ means that the random variable X has distribution ν, i.e., that $\text{Prob}(X \in A) = \nu(A)$. The term $\text{Prob}(X \in A)$ quantifies the probability that X assumes a value in the set A. Values that are assumed (or: taken on) by a random variable X are called realizations 시행, 측정(evaluation)을 통해 얻은 결과 of X.

508

[Notation] In this document, we denote random variables by capital letters and their realizations usually by the corresponding small letters.

509

Example: Assume that the random variable X models the tossing of a fair coin. Then $\text{Prob}(X = 0) = \text{Prob}(X = 1) = 0.5$ if we identify "head" and "tail" with 1 and 0. These probabilities quantify the knowledge on the outcome of a future coin toss (and on a past experiment to a person who does not know its outcome). Possible realizations of X are 0 and 1.

510

In this document, we model non-deterministic phenomena by random variables. Their realizations are observable as random numbers, voltage, or timing, for example.

511

Definition: The term $B(n, p)$ denotes the **binomial distribution** with parameters n and p, which is given by

$$\text{Prob}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, \ldots, n. \tag{4.1}$$

Definition: The Poisson distribution with parameter $\tau > 0$ is given by

$$\text{Prob}(X = k) = \frac{\tau^k}{k!} e^{-\tau} \quad \text{for } k \in \mathbb{N}_0. \tag{4.2}$$

Note: The parameter $\tau > 0$ equals the mean number of events per time interval of length 1.

> The Poisson distribution is a special case of the binomial distribution with $p = \frac{\tau}{n}$ where $n \to \infty$, for the number of repetitions k, i.e., $\lim_{n\to\infty} \binom{n}{k} \left(\frac{\tau}{n}\right)^k \left(1 - \frac{\tau}{n}\right)^{n-k}$.

513

Definition: The geometric대수학의 geometric mean 등과 관련됨 distribution $\mathcal{G}_p$ with parameter $p \in (0, 1]$ is a discrete distribution on $\mathbb{N}$. More precisely,

$$\mathcal{G}_p(k) := p(1-p)^{k-1} \quad \text{for } k \in \mathbb{N}. \tag{4.3}$$

The term $\mathcal{G}_p(k)$ equals the probability that a sequence of iid Bernoulli trials with individual success probability p is successful for the first time in the k^{th} trial.

Note: In the literature, there also exists an alternative definition of the geometric distribution that only counts the number of failures, i.e., $k-1$ in place of k.

514

Definition: The letters λ and λ_m denote the Lebesgue measures on $\mathbb{R}$ or $\mathbb{R}^m$, respectively. It is $\lambda([a,b)) = b-a$ if $a \leq b$. Accordingly, $\lambda_m\left(\prod_{j=1}^m [a_j, b_j)\right) = \prod_{j=1}^m (b_j - a_j)$ if $a_j \leq b_j$ for $1 \leq j \leq m$.

Note: The Lebesgue measure λ corresponds to the "geometricEuclid geometry와 관련됨" measure on $\mathbb{R}$ (not to be mixed up with the geometric distribution defined in par. 513).

515

Definition: The term $N(\mu, \sigma^2)$ denotes the **normal (Gaussian) distribution** with expectation μ and variance σ^2. It has the density

$$\phi(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.4}$$

Induction of the density

- Preliminary 1

  $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ by Stirling's approximation for large n.

- Preliminary 2

$$\begin{aligned} \ln \frac{n}{2}\left(1 + \frac{2\delta}{n}\right) &= \ln \frac{n}{2} + \ln\left(1 + \frac{2\delta}{n}\right) \\ &\approx \ln \frac{n}{2} + \left(\frac{2\delta}{n} - \frac{1}{2}\left(\frac{2\delta}{n}\right)^2\right) \qquad \text{by Taylor Series approx.} \end{aligned}$$

Similarly, $\ln \frac{n}{2}\left(1 + \frac{2\delta}{n}\right) \approx \ln \frac{n}{2} - \left(\frac{2\delta}{n} + \frac{1}{2}\left(\frac{2\delta}{n}\right)^2\right)$.

- Preliminary 3

  For given $R = \frac{n^n}{\left(\frac{n}{2}+\delta\right)^{n/2+\delta}\left(\frac{n}{2}-\delta\right)^{n/2-\delta}2^n}$,

$$
\begin{aligned}
\ln R &= n\ln n - \left(\frac{n}{2}+\delta\right)\ln\left(\frac{n}{2}+\delta\right) \\
&\quad - \left(\frac{n}{2}-\delta\right)\ln\left(\frac{n}{2}-\delta\right) - n\ln 2 \\
&\approx n\ln n - \left(\frac{n}{2}+\delta\right)\left(\ln\frac{n}{2} + \frac{2\delta}{n} - \frac{1}{2}\left(\frac{2\delta}{n}\right)^2\right) \\
&\quad - \left(\frac{n}{2}-\delta\right)\left(\ln\frac{n}{2} - \frac{2\delta}{n} - \frac{1}{2}\left(\frac{2\delta}{n}\right)^2\right) - n\ln 2 \qquad \text{by Pre.2} \\
&= n\ln\frac{n}{2} - \frac{n}{2}\ln\frac{n}{2} + \frac{n}{4}\left(\frac{2\delta}{n}\right)^2 - \delta\left(\frac{2\delta}{n}\right) \\
&\quad - \frac{n}{2}\ln\frac{n}{2} + \frac{n}{4}\left(\frac{2\delta}{n}\right)^2 - \delta\left(\frac{2\delta}{n}\right) \\
&= \frac{n}{2}\left(\frac{2\delta}{n}\right)^2 - \frac{(2\delta)^2}{n} \\
&= -\frac{(2\delta)^2}{2n}. \\
\therefore R &\approx \exp(-2\delta^2/n).
\end{aligned}
$$

For the fair coin tossing, $\Pr[X=x] = \binom{n}{x}\left(\frac{1}{2}\right)^n = \frac{n!}{x!(n-x)!}\left(\frac{1}{2}\right)^n$.
Let $\mu = \mathbb{E}X = \frac{n}{2}$ and $\delta := x - \frac{n}{2}$.
Then $x = \frac{n}{2}+\delta$, $n - x = \frac{n}{2} - \delta$.

$$
\begin{aligned}
\Pr[X=x] &= \frac{n!}{x!(n-x)!}\left(\frac{1}{2}\right)^n \\
&= \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi x}\left(\frac{x}{e}\right)^x\sqrt{2\pi(n-x)}\left(\frac{n-x}{e}\right)^{(n-x)}}\left(\frac{1}{2}\right)^n \qquad \text{by Pre.1} \\
&= \frac{\sqrt{n}\cdot n^n}{\sqrt{2\pi x}\cdot x^x\sqrt{n-x}\cdot(n-x)^{(n-x)}}\left(\frac{1}{2}\right)^n \\
&= \frac{\sqrt{n}}{\sqrt{2\pi x(n-x)}}\cdot\frac{n^n}{x^x\cdot(n-x)^{(n-x)}\cdot 2^n} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi\left(\frac{n}{2}+\delta\right)\left(\frac{n}{2}-\delta\right)}}\cdot\frac{n^n}{\left(\frac{n}{2}+\delta\right)^{n/2+\delta}\cdot\left(\frac{n}{2}-\delta\right)^{n/2-\delta}\cdot 2^n} \\
&= \frac{\sqrt{n}}{\sqrt{\frac{n^2}{2}\pi\left(1+\frac{2\delta}{n}\right)\left(1-\frac{2\delta}{n}\right)}}\cdot\frac{n^n}{\left(\frac{n}{2}+\delta\right)^{n/2+\delta}\cdot\left(\frac{n}{2}-\delta\right)^{n/2-\delta}\cdot 2^n}
\end{aligned}
$$

$$\approx \frac{1}{\sqrt{\frac{n\pi}{2}\left(1-\left(\frac{2\delta}{n}\right)^2\right)}} \cdot \exp(-2\delta^2/n) \qquad \text{by Pre.3}$$

$$\approx \sqrt{\frac{2}{n\pi}} \cdot \exp(-2\delta^2/n) \qquad \text{for } \delta \ll n$$

Let $\sigma^2 = \frac{n}{4}$.
Since $\delta = x - \frac{n}{2} = x - \mu$,

$$\begin{aligned}\Pr[X = x] &= \sqrt{\frac{2}{n\pi}} \cdot \exp(-2\delta^2/n) \\ &= \sqrt{\frac{2}{4 \cdot \frac{n}{4}\pi}} \cdot \exp(-2(x-\mu)^2/(4 \cdot n/4)) \\ &= \sqrt{\frac{1}{2\sigma^2\pi}} \cdot \exp(-(x-\mu)^2/(2\sigma^2))\end{aligned}$$

In particular, $N(0,1)$ is called **standard normal distribution**. Its cumulative distribution function $\Phi(\cdot)$ is given by

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \qquad (4.5)$$

Note 1: To be precise, "density" means "Lebesgue density". In this document, densities with respect to other measures than the (one-dimensional or multidimensional) Lebesgue measure are not relevant. For this reason, we briefly speak of "density" in place of "Lebesgue density" in the following.

Note 2: Normal distributions exist in $\mathbb{R}^k$ for each $k \geq 1$ (multivariate normal distributions); cf. par. 542.

516

Definition: The **Gamma distribution** with the shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has the density

$$\gamma_{\alpha,\beta}(x) := \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0. \tag{4.6}$$

> Gamma function $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} dx$ $(\alpha > 0)$ is defined analytic continuation2!과 3!의 사이인 2.5! 등의 of this integral function. It offers a useful property from its integral, which takes the form of (polynomial) * (exponential).
> The gamma distribution is a Conjugate prior distribution.

Note: Occasionally, the Gamma distribution is not characterized by a shape parameter and a rate parameter but by a shape parameter and a scale parameter. Thus, caution is advised when results from different books and papers are applied. The scale parameter is the reciprocal value of the shape parameter.

517

The random variable X is called discrete if Ω is countable (finite or infinite). If Ω is finite, we also call X a finite random variable. Examples are binomially distributed random variables and Poisson-distributed random variables. Section 4.4 treats given random mappings. There, the realizations of the random variables are mappings between sets.

518

Let X be a random variable that assumes values in a finite set Ω. We say that X is **uniformly distributed** (or equivalently: unbiased, equidistributed) if it assumes all $\omega \in \Omega$ with the same probability, namely $\text{Prob}(X = \omega) = |\Omega|^{-1}$. Otherwise, X is said to be biased.

Note: Precisely formulated, it should actually read $\text{Prob}(X = \{\omega\})$ instead of $\text{Prob}(X = \omega)$. However, the shorter expression "$\text{Prob}(X = \omega)$" is common for finite and countable Ω.

519

A random variable X has density $f : \Omega \to [0, \infty]$ with respect to a measure τ if $\text{Prob}(X \in A) = \int_A f(\omega)\, d\tau(\omega)$ for all measurable sets A. Equivalently, a measure ν has density $f : \Omega \to [0, \infty]$ with respect to a measure τ if $\nu(A) = \int_A f(\omega)\, d\tau(\omega)$ for all measurable sets A.

Note: Densities do not exist for each pair of measures (ν, τ). v는 point 기준, t는 length 기준일 때 $p_v(x) = 1 \neq p_t(x) = 0$

520

In our context, usually $\Omega \subseteq \mathbb{R}^m$ with $m \geq 1$, and $\tau = \lambda_m$. Then lambda(dx) = lambda(x-x')=—x-x'—=dx 이므로

$$\text{Prob}(X \in A) = \int_A f(x)\, \lambda_m(dx) = \int_A f(x)\, dx. \qquad (4.7)$$

521

Let X denote a random variable that assumes values in $\mathbb{R}^m$ and has distribution ν. If the integral

$$E(X) := \int_\Omega x\, \nu(dx) \qquad (4.8)$$

exists (i.e., if $\int_\Omega |x|\, \nu(dx) < \infty$), then $E(X)$ is called the expectation of X.

Note: The expectation $E(X)$ does not exist for every random variable. Counterexamples are, for example, Cauchy-distributed 꼬리가 두꺼운 Bell-shape 분포로 적분 결과가 발산함 random variables.

For discrete random variables X with values in $\Omega \subseteq \mathbb{R}$ (e.g., $\Omega = \{0,1\}$, $\mathbb{N}$, or $\mathbb{Z}$), formula (4.8) simplifies to

$$E(X) := \sum_{x \in \Omega} x \operatorname{Prob}(X = x). \tag{4.9}$$

For any random variables X and Y regardless of independence,

$$\begin{aligned}
E(X+Y) &= \sum_x \sum_y (x+y) P(X=x, Y=y) \\
&= \sum_x \sum_y x P(X=x, Y=y) + \sum_x \sum_y y P(X=x, Y=y) \\
&= \sum_x x \left(\sum_y P(X=x, Y=y) \right) + \sum_y y \left(\sum_x P(X=x, Y=y) \right) \\
&= \sum_x x P(X=x) + \sum_y y P(Y=y) \\
&= E(X) + E(Y)
\end{aligned}$$

It can be generalized to $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$.
If X and Y are independent, then $P(X=x, Y=y) = P(X=x)P(Y=y)$. Thus,

$$\begin{aligned}
E(XY) &= \sum_x \sum_y xy P(X=x, Y=y) \\
&= \sum_x \sum_y xy P(X=x) P(Y=y) \\
&= \left(\sum_x x P(X=x) \right) \left(\sum_y y P(Y=y) \right) \\
&= E(X) E(Y)
\end{aligned}$$

If X assumes values in $\mathbb{R}^m$ and has Lebesgue density f then (4.8) reads

$$E(X) := \int_{\mathbb{R}^m} x f(x)\, dx. \tag{4.10}$$

In the context of **PTRNG** evaluations, we are usually faced with these two special cases.

523

Note: For random variables with values in $\{0,1\}^n$, no meaningful definition for the mean is evident.

524

The variance of a real-valued random variable X is defined by

$$\mathrm{Var}(X) := E(E(X) - X)^2. \qquad (4.11)$$

provided that both expectations exist. This is not always the case. Cauchy-dist.는 EX가 없으므로 Var도 없음

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= E\left(E\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i\right)^2 \\
&= E\left(\mu - \sum_{i=1}^{n} X_i\right)^2 \\
&= E\left[\mu^2 - 2\mu\sum_{i=1}^{n} X_i + \left(\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \mu^2 - 2\mu E\sum_{i=1}^{n} X_i + E\left(\sum_{i=1}^{n} X_i\right)^2 \\
&= (EX)^2 - 2\,(EX)^2 + E\left(\sum_{i=1}^{n} X_i\right)^2 \\
&= -\,(EX)^2 + EX^2
\end{aligned}$$

525

Assume that $\mathrm{Var}(X)$ exists. Then

$$\sigma_X := \sqrt{\mathrm{Var}(X)} \qquad (4.12)$$

is the standard deviation of X.

526

[sum of normal distributions] If X_1 and X_2 denote **independent** normally distributed random variables with expectations μ_1, μ_2 and variances σ_1^2, σ_2^2, then $X_1 + X_2$ is normally distributed with expectation $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. More generally, if the random variables $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$-distributed then the sum $X_1 + \cdots + X_n$ is $N(n\mu, n\sigma^2)$-distributed.

$$E\left(\sum X_i\right) = \sum E(X_i) = n\mu$$

$$E\left(\sum_{i=1}^{n} X_i\right)^2 = E\left(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right) = \sum_{i=1}^{n} EX_i^2 + \sum_{i \neq j} E(X_i X_j)$$

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= E\left(\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i\right)^2 \\
&= E\left(\sum_{i=1}^{n} (X_i - EX_i)\right)^2 \\
&= E\left(\sum_{i=1}^{n} (X_i - \mu_i)\right)^2 \\
&= E\left(\sum_{i=1}^{n} (X_i - \mu_i)^2 + \sum_{i \neq j} (X_i - \mu_i)(X_j - \mu_j)\right) \\
&= \sum_{i=1}^{n} E(X_i - \mu_i)^2 + \sum_{i \neq j} E[(X_i - \mu_i)(X_j - \mu_j)] \\
&= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)
\end{aligned}$$

If $X_1, ..., X_n$ are independent,

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

because $\operatorname{Cov}(X_i, X_j) = 0$ for all i, j s.t. $i \neq j$.

527

[Gamma distribution] The Gamma distribution with the shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has the density $\gamma_{\alpha,\beta}(\cdot)$, cf. par. 516. A random variable that is Gamma distributed with parameters α and β has mean $\mu = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$.

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx\Gamma(\alpha+1) = \alpha\Gamma(\alpha)f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}, \quad x > 0$$

$$\begin{aligned}
\mu = E(X) &= \frac{\alpha}{\beta}\int_0^\infty xf(x;\alpha,\beta)\ dx \\
&= \int_0^\infty x\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}dx \\
&= \int_0^\infty \frac{\beta^\alpha x^\alpha e^{-\beta x}}{\Gamma(\alpha)}dx \\
&= \frac{\alpha}{\beta}\int_0^\infty \frac{\beta^{\alpha+1}x^\alpha e^{-\beta x}}{\Gamma(\alpha+1)}dx \quad (\because \Gamma(\alpha+1) = \alpha\Gamma(\alpha)) \\
&= \frac{\alpha}{\beta}\int_0^\infty f(x;\alpha+1,\beta)\ dx \\
&= \frac{\alpha}{\beta}
\end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_0^\infty x^2\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}dx \\
&= \frac{\alpha(\alpha+1)}{\beta^2}\int_0^\infty \frac{\beta^{\alpha+2}x^{\alpha+1}e^{-\beta x}}{\Gamma(\alpha+2)}dx \\
&= \frac{\alpha(\alpha+1)}{\beta^2}\int_0^\infty f(x;\alpha+2,\beta)\ dx \\
&= \frac{\alpha(\alpha+1)}{\beta^2} \\
&= \frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2}
\end{aligned}$$

$$\sigma^2 = \mathrm{Var}(X) = EX^2 - (EX)^2 = \left(\frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2}\right) - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}$$

528

[sum of Gamma distributions] If X and Y are **independent** random variables with densities $\gamma_{\alpha_1,\beta}(\cdot)$ and $\gamma_{\alpha_2,\beta}(\cdot)$, respectively, then $X + Y$ is Gamma-distributed with density $\gamma_{\alpha_1+\alpha_2,\beta}(\cdot)$. Consequently, if the random variables $X_1, \ldots, X_n$ are iid Gamma distributed with parameters α and β then the sum $X_1 + \cdots + X_n$ is Gamma distributed with parameters $n\alpha$ and β.

529

The random variables $X_1, X_2, \ldots, X_k$ are said to be **independent** if for each k-tuple $(A_1, \ldots, A_k)$ of measurable sets the equality

$$\text{Prob}(X_1 \in A_1, \ldots, X_k \in A_k) = \prod_{j=1}^{k} \text{Prob}(X_j \in A_j). \qquad (4.13)$$

holds.

530

More generally, the (infinite) sequence $X_1, X_2, \ldots$ of random variables is said to be independent if for each integer $k' \geq 1$ and for each k'-tuple $(A_1, \ldots, A_{k'})$ not distinct of measurable sets, condition (4.13) is valid (with k' in place of k).

Note: Independence can be generalized to uncountable index sets.

531

For discrete random variables $X_1, X_2, \ldots$ with values in Ω, condition (4.13) simplifies to

$$\text{Prob}(X_1 = x_1, \ldots, X_k = x_k) = \prod_{j=1}^{k} \text{Prob}(X_j = x_j) \qquad (4.14)$$

for each k-tuple $(x_1, \ldots, x_k) \in \Omega^k$. xi={xi} in Omega

532

In the context of random variables $X_1, X_2, \ldots$, the abbreviation **iid** stands for "independent and identically distributed."

533

Mathematically, a sequence of iid uniformly distributed random variables $X_1, X_2, \ldots$ on a finite set Ω (e.g., $\Omega = \{0, 1\}$) describes an ideal RNG.

534

Assume that the random variables $X_1, X_2, \ldots, X_n$, resp. $X_1, X_2, \ldots$ are independent. If $X_j \sim \nu_j$, the joint distribution of $(X_1, X_2, \ldots, X_n)$, resp. of the sequence $X_1, X_2, \ldots$ is given by the product measure $\otimes_{j=1}^{n} \nu_j$, resp. by $\otimes_{j=1}^{\infty} \nu_j$. These product measures are characterized by the conditions from pars. 529 and 530. If the random variables $X_1, X_2, \ldots$ are identically distributed, i.e., if $\nu_1 = \nu_2 = \cdots = \nu_n$, we alternatively also use the notation ν^n and $\nu^{\mathbb{N}}$.

For the joint distribution

1. $\sum_x \sum_y \Pr(X = x, Y = y) = 1$

$$\begin{aligned} \Pr[X = x \text{ and } Y = y] &= \Pr[X = x | Y = y] \Pr[Y = y] \\ &= \Pr[Y = y | X = x] \Pr[X = x] \end{aligned}$$

2. $\int_x \int_y f_{x,y} \, dydx = 1$

$$\begin{aligned} f_{x,y}(X = x \text{ and } Y = y) &= f_{x,y}(X = x | Y = y) f_y(Y = y) \\ &= f_{x,y}(Y = y | X = x) f_x(X = x) \end{aligned}$$

535

Assume that for the real-valued random variables X and Y, expectations and variances exist. Then the right-hand sides of (4.15) and (4.16) exist

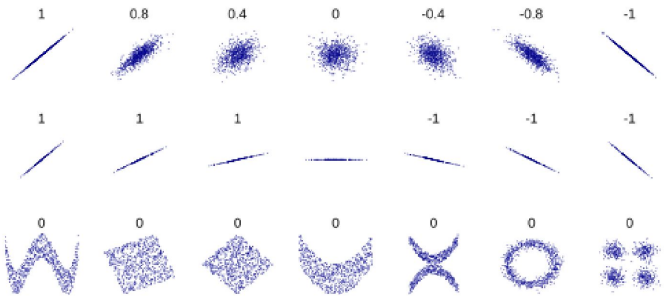
$$\mathrm{Cov}(X,Y) := E(XY) - E(X)E(Y) \quad \text{(covariance)} \tag{4.15}$$

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[(X - E(X))(Y - E(Y))] \\
&= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

$$\mathrm{corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \quad \text{(correlation coefficient)} \tag{4.16}$$

Cov¡0일 때가 있어서 일반적으로 sqrt(Cov)는 정의하지 않음

If $\mathrm{Cov}(X,Y) = 0$, we say that X and Y are uncorrelated.



The sign of the covariance shows the tendency in the linear relationship between the variables.

536

Independence implies uncorrelatedness but, in general, the converse is not true (cf. pars. 537 and 543).

If X, Y are independent, $\mathrm{Cov}(X,Y) = 0$ because $E(XY) = E(X)E(Y)$.

537

Counterexample ([Geor15], Beispiel (4.26)): Assume that X and Y are random variables that assume values in $\Omega_1 = \{-1, 0, 1\}$ and in $\Omega_2 = \{0, 1\}$, respectively. Assume further that $\text{Prob}(X = 1, Y = 0) = \text{Prob}(X = 0, Y = 1) = \text{Prob}(X = -1, Y = 0) = 1/3$. Hence, $\text{Prob}(X = 0) = \text{Prob}(X = 1) = \text{Prob}(X = -1) = 1/3$ and thus $E(X) = 0$. Similarly, $\text{Prob}(Y = 0) = 2/3$, $\text{Prob}(Y = 1) = 1/3$ and thus $E(Y) = 1/3$. Finally,

$$\begin{aligned}\text{Cov}(X, Y) = E(XY) - 0 \cdot \frac{1}{3} &= \sum_{x \in \Omega_1, y \in \Omega_2} xy \,\text{Prob}(X = x, Y = y) - 0 \\ &= \left(1 \cdot 0 \cdot \frac{1}{3} + 0 \cdot 1 \cdot \frac{1}{3} - 1 \cdot 0 \cdot \frac{1}{3}\right) = 0.\end{aligned}$$

Thus, X and Y are uncorrelated, but since $\text{Prob}(X = 1, Y = 1) = 0 \neq 1/9 = \text{Prob}(X = 1) \cdot \text{Prob}(Y = 1)$, the random variables X and Y are not independent.

538

[**Strong law of large numbers**] Assume that the random variables $X_1, X_2, \ldots$ have expectations $E(X_1), E(X_2), \ldots$. We say that the sequence $X_1, X_2, \ldots$ satisfies the strong law of large numbers if

$$\text{Prob}\left(\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} (X_j - E(X_j)) = 0\right) = 1. \qquad (4.17)$$

If the random variables $X_1, X_2, \ldots$ are identically distributed with expectation $E(X_j) = \mu$, (4.17) simplifies to

$$\text{Prob}\left(\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} X_j = \mu\right) = 1. \qquad (4.18)$$

The proving process of the Weak Law of Large Numbers

[Theorem. Markov Inequality] If $X \geq 0$ is a random variable and $a > 0$, then

$$P(X \geq a) \leq \frac{1}{a} EX.$$

Proof. Let the indicator function IX≥a be defined as follows.

$$I := \begin{cases} 1 & \text{if } X \geq a, \\ 0 & \text{if } X < a. \end{cases}$$

Then, here is the crucial observation:

$$I_{\{X \geq a\}} \leq \frac{1}{a} X.$$

Indeed, if $X < a$, the left-hand side is 0 and the right-hand side is nonnegative; if $X \geq a$, the left-hand side is 1, and the right-hand side is at least 1. Taking the expectation of both sides, we get

$$P(X \geq a) = E(I_{\{X \geq a\}}) \leq \frac{1}{a} EX. \qquad \square$$

[Theorem. Chebyshev inequality] If $EX = \mu$ and $\mathrm{Var}(X) = \sigma^2$ are both finite and $k > 0$, then

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

Proof. By the Markov inequality,

$$\begin{aligned} P(|X - \mu| \geq k) &= P((X - \mu)^2 \geq k^2) \\ &\leq \frac{1}{k^2} E(X - \mu)^2 \\ &= \frac{1}{k^2} \mathrm{Var}(X). \qquad \square \end{aligned}$$

The Chebyshev inequality is useful if either σ is small or k is large.

[Theorem. Weak Law of Large Numbers] If $X, X_1, X_2, \ldots$ are iid R.V.s with finite expectation and variance, then $\frac{X_1 + \cdots + X_n}{n}$ converges to EX in the sense that, for any fixed $\epsilon > 0$,

$$P\left(\left| \frac{X_1 + \cdots + X_n}{n} - EX \right| \geq \epsilon \right) \to 0,$$

as $n \to \infty$.

Proof. Denote $\mu = EX$, $\sigma^2 = \mathrm{Var}(X)$, and let $S_n = X_1 + \cdots + X_n$. Then,

$$ES_n = EX_1 + \cdots + EX_n = n\mu \quad \text{and} \quad \mathrm{Var}(S_n) = n\sigma^2.$$

Therefore, by the Chebyshev inequality,

$$\begin{aligned} P\left(\left| \frac{S_n}{n} - EX \right| > \epsilon \right) &= P(|S_n - n\mu| \geq n\epsilon) \\ &\leq \frac{n\sigma^2}{n^2 \epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0, \qquad \text{as } n \to \infty. \quad \square \end{aligned}$$

In particular, if S_n is the number of successes in n independent trials, each of which is a success with probability p, then, as we have observed before, $S_n = I_1 + \cdots + I_n$, where $I_i = I_{\{\text{success at trial } i\}}$. So, for every $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \to 0, \quad \text{as } n \to \infty.$$

Thus, the proportion of successes converges to p in this sense.

1. $X_n \to X$ in **probability**, if for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.
2. $X_n \to X$ in distribution, if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ as $n \to \infty$, for all $x \in \mathbb{R}$ at which $x \mapsto \mathbb{P}(X \leq x)$ is continuous.
3. $X_n \to X$ in L^1, if $\mathbb{E}(|X_n|) < \infty$ for all $n \geq 1$ and $\mathbb{E}(|X_n - X|) \to 0$ as $n \to \infty$.
4. $X_n \to X$ **almost surely** (a.s.), if $\mathbb{P}(X_n \to X \text{ as } n \to \infty) = 1$.

$$\begin{array}{ccc} X_n \to X \text{ almost surely (Strong)} & & \\ \Downarrow & & \\ X_n \to X \text{ in probability (Weak)} & \Rightarrow & X_n \to X \text{ in distribution} \\ \Uparrow & & \\ X_n \to X \text{ in } L^1 & \Rightarrow & \mathbb{E}(X_n) \to \mathbb{E}(X) \end{array}$$

539

[Strong law of large numbers] If the random variables $X_1, X_2, \ldots$ fulfill one of the following conditions, they satisfy the strong law of large numbers.

i The random variables $X_1, X_2, \ldots$ are iid with expectation $E(X_j) = \mu$.

ii The random variables $X_1, X_2, \ldots$ are identically distributed with expectation $E(X_j) = \mu$ and mutually independent.

Note 1: Condition (ii) generalizes condition (i).

Note 2: Under suitable conditions, the strong law of large numbers also applies to dependent random variables.

540

[Convolution of probability measures] Let X and Y denote independent real-valued random variables that are η-distributed and τ-distributed, respectively. The distribution of the sum $X+Y$ is the **convolution** product of η and τ, denoted by $\eta * \tau$. If X has density $g(\cdot)$ and Y has density $h(\cdot)$, then $X+Y$ has density

$$f(x) := \int_{-\infty}^{\infty} g(x-y)h(y)\,dy. \qquad (4.19)$$

Z=X+Y=z가 되는 모든 (x,y)에 대한 joint distribution density의 누적합

The CDF of $Z = X + Y$ is

$$\begin{aligned}\Pr(X+Y \le z) = \int f * g(z)dz &= \int\int_{x+y\le z} f(x)g(y)dxdy \\ &= \int_{-\infty}^{\infty} g(y) \int_{-\infty}^{z-y} f(x)dx\ dy\end{aligned}$$

$$\begin{aligned}\therefore\ f * g(z)dz &= \int_{-\infty}^{\infty} g(y)\left(\frac{d}{dz}\int_{-\infty}^{z-y} f(x)dx\right)dy \\ &= \int_{-\infty}^{\infty} g(y)f(z-y)dy \qquad \because f(-\infty)\to 0\end{aligned}$$

Replace z with x and denote the convolution $f * g$ as h then, 본문은 각 분포에 대한 density가 h-¿f, f-¿g, g-¿h로 표기됨

$$h(x)dx = \int_{-\infty}^{\infty} f(x-y)g(y)dy$$

Note 1: The convolution of special distributions is discussed in pars. 1202 to 1206.

$$\begin{aligned}\Pr(Z=z) &= \Pr[(X,Y)\in A] \qquad \text{Let } A := \{(x,y) : x+y=z\} \\ &= \sum_y P(X=z-y|Y=y)P(Y=y) \\ &= \sum_y P(X=z-y)P(Y=y) \qquad \because X, Y \text{ are independent} \\ &= \sum_y f(z-y)g(y)\end{aligned}$$

Note 2: The term τ^{*k} denotes the k-fold convolution product $\tau * \cdots * \tau$.

541

[Convolution of probability measures on groups] The convolution of probability measures is not only defined on $\mathbb{R}$ but also on groups. In our context, finite groups are relevant. Assume that the independent random variables X and Y take on values on a finite (not necessarily commutative) group G with group operation "$\circ$", and that X and Y are τ-distributed and η-distributed, respectively. In particular,

$$\tau * \eta(g) = \sum_{h \in G} \eta(g \circ h^{-1})\tau(h) \quad \text{for each } g \in G. \tag{4.20}$$

$$\begin{aligned}
\tau * \eta(z) &= \Pr(Z = z) \\
&= \Pr((X, Y) \in A) \qquad \text{Let } A := \{(x, y) \in G \times G \mid y \circ x = z\} \\
&= \sum_{(x,y) \in A} \Pr(X = x, Y = y) \\
&= \sum_{x \in G} \Pr(X = x, Y = z \circ x^{-1}) \\
&= \sum_{x \in G} P(Y = z \circ x^{-1}) P(X = x) \qquad \because X, Y \text{ are independent} \\
&= \sum_{x \in G} \eta(z \circ x^{-1})\tau(x)
\end{aligned}$$

Replace z and x with g and h, respectively, then,

$$\tau * \eta(g) = \sum_{h \in G} \eta(g \circ h^{-1})\tau(h)$$

Example:

1. $(G, \circ) = (\mathbb{Z}_n, +(\text{mod } n))$.
2. $(G, \circ) = (\{0, 1\}^\ell, \oplus)$, where "$\oplus$" denotes the bitwise addition mod 2.
3. $(G, \circ) = \text{GL}(n, \text{GF}(2))$, the group of invertible $(n \times n)$-matrices over GF(2). Here, the group operation is the matrix multiplication. The group $\text{GL}(n, \text{GF}(2))$ is not commutative for $n \geq 2$.

Note 1: In an additive group, $X \circ Y$ and $g \circ h^{-1}$ read $X + Y$ and $g - h$, respectively.

Note 2: Pars. 596 to 603 provide useful theorems.

다변량 RNG의 Stochastic model 구성 시

Definition: Let C is a positive definite $(k \times k)$-matrix. The term $N(\vec{\mu}, C)$ denotes the k-dimensional normal (Gaussian) distribution with expectation $\vec{\mu}$ and covariance matrix C. It has the k-dimensional density

$$f_C(x) := \frac{1}{\sqrt{(2\pi)^k}\sqrt{\det C}} e^{-0.5(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})}. \tag{4.21}$$

Let $\mathbf{z} \sim N(\mathbf{0}, I)$. (It means $E[\mathbf{z}] = \vec{0}$, $\mathrm{Cov}(\mathbf{z}) = I$.)
Note: Covariance of a 1-dim Normal distribution is $\mathrm{Cov}(z_1, z_1) = \sigma^2$
The joint PDF of $\mathbf{z}$ is

$$\begin{aligned} f_{\mathbf{z}}(\mathbf{z}) &= \prod_{i=1}^{k} f(z_i) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \\ &= \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum z_i^2} \\ &= \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} \end{aligned}$$

Now we want to replace $\mathbf{x}$ with $\mathbf{z}$. The $\mathbf{x}$ has two conditions in this proof.
Let A be the $k \times k$ invertible matrix, and the covariance matrix of $\mathbf{x}$ is $C := AA^T$ (Cholesky decomposition).
1. $\mathrm{Cov}(\mathbf{x}) = C$.

$$\begin{aligned} \mathrm{Cov}(\mathbf{x}) &= C \\ &= AA^T \\ &= AIA^T \\ &= A\mathrm{Cov}(\mathbf{z})A^T \qquad \because \mathrm{Cov}(\mathbf{z}) = I \\ &= \mathrm{Cov}(A\mathbf{z}) \\ &= \mathrm{Cov}(A\mathbf{z} + \mathbf{c}) \qquad \text{for some vector } \mathbf{c} \end{aligned}$$

2. $E[\mathbf{x}] = \boldsymbol{\mu}$.

$$\begin{aligned} E[\mathbf{x}] &= \boldsymbol{\mu} \\ &= B \cdot \mathbf{0} + \boldsymbol{\mu} \qquad \text{for some matrix } B \\ &= B \cdot E[\mathbf{z}] + \boldsymbol{\mu} \qquad \because E[\mathbf{z}] = \mathbf{0} \\ &= E[B\mathbf{z} + \boldsymbol{\mu}] \end{aligned}$$

Since the $\mathbf{x}$ need to satisfy both 1. and 2.,

$$\mathbf{x} = A\mathbf{z} + \boldsymbol{\mu}$$

and consequently,

$$\therefore\ \mathbf{z} = A^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Next, we find the PDF of $\mathbf{x}$.

$$\begin{aligned}
\int_{\Omega_1} f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} &= \int_{\Omega_2} f_{\mathbf{z}}(\mathbf{z})d\mathbf{z} && \because P(\mathbf{x} \in \Omega_1) = P(\mathbf{z} \in \Omega_2) \\
&= \int_{\Omega_1} f_{\mathbf{z}}(\mathbf{z}) \left|\det \frac{d\mathbf{z}}{d\mathbf{x}}\right| d\mathbf{x} && \because \frac{d\mathbf{z}}{d\mathbf{x}} = \mathrm{Vol}\left(\frac{d\mathbf{z}}{d\mathbf{x}}\right) = \left|\det \frac{d\mathbf{z}}{d\mathbf{x}}\right| \\
\therefore\ f_{\mathbf{x}}(\mathbf{x}) &= f_{\mathbf{z}}(\mathbf{z}) \left|\det \frac{d\mathbf{z}}{d\mathbf{x}}\right|
\end{aligned}$$

where

$$\begin{aligned}
\left|\det\left(\frac{d\mathbf{z}}{d\mathbf{x}}\right)\right| &= |\det A^{-1}| && \because \frac{d\mathbf{z}}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}(A^{-1}(\mathbf{x} - \boldsymbol{\mu})) = A^{-1} \\
&= \frac{1}{|\det A|} \\
&= \frac{1}{\sqrt{|\det C|}} && \because \det C = \det(AA^T) = (\det A)^2
\end{aligned}$$

Therefore,

$$\begin{aligned}
f_{\mathbf{x}}(\mathbf{x}) &= f_{\mathbf{z}}(\mathbf{z}) \left|\det\left(\frac{d\mathbf{z}}{d\mathbf{x}}\right)\right| \\
&= \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} \cdot \frac{1}{\sqrt{|\det C|}} \\
&= \frac{1}{\sqrt{(2\pi)^k |\det C|}} e^{-\frac{1}{2}(A^{-1}(\mathbf{x}-\boldsymbol{\mu}))^T (A^{-1}(\mathbf{x}-\boldsymbol{\mu}))} \\
&= \frac{1}{\sqrt{(2\pi)^k |\det C|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T (A^{-1})^T A^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{\sqrt{(2\pi)^k |\det C|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu})} && \because (A^{-1})^T = (A^T)^{-1}
\end{aligned}$$

Note 1: If $\vec{X} = (X_1, \ldots, X_k)$ is $N(\vec{\mu}, C)$-distributed, its expectation $E(\vec{X}) = \vec{\mu} \in \mathbb{R}^k$, and $C = (c_{ij})_{1 \le i,j \le k} = (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le k}$ is the covariance matrix of $\vec{X}$.

Note 2: Definition (4.21) generalizes (4.4). In the one-dimensional case, $\vec{\mu} = \mu$, and the covariance matrix C equals the variance σ^2.

Multivariate normal distribution $\Rightarrow$ each is normally distributed (the converse is not true)

Note 3: If the random variables $X_1, \ldots, X_k$ are iid $N(0,1)$-distributed, the covariance matrix is given by the k-dimensional identity mapping.

Note 4: Above, we assume that the covariance matrix C is regular. It should be noted that normal distributions with singular covariance matrices 역행렬이 없는 정사각 행렬 C also exist. These distributions do not have k-dimensional densities.

543

Assume that the random variables X and Y are bivariate normally distributed. If X and Y are uncorrelated, X and Y are independent.

Note: In general, this conclusion is not true; cf. pars. 536 and 537.

544

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Formally, a **stochastic process**process에 time의 의미가 내포되는 경우가 많음 $(X_t)_{t\in T}$ with state space Ω is a collection of real-valued random variables $\{X_t \mid t \in T\}$, where the index t is usually interpreted as "time."

545

If $T \subseteq \mathbb{R}$ is an interval (e.g., $T = (a,b)$, $T = [0,\infty)$ or $T = \mathbb{R}$), we speak of (time-)continuous stochastic processes. If $T \subseteq \Delta\mathbb{Z}$ for some $\Delta > 0$, e.g., $T = \mathbb{Z}, T = \mathbb{N}$ or $T = \mathbb{N}_0$, the stochastic process is called (time-)discrete.

546

Example: Markov chains (time-discrete stochastic process; cf. par. 578), Gaussian process (time-continuous stochastic process; cf. par. 547), Wiener process (time-continuous stochastic process).

[Wiener process] For the R.V.s $W_{t_1}, W_{t_2}, W_{t_3}, W_{t_4}$, a Markov chain is established by $(W_{t_2} - W_{t_1}), (W_{t_4} - W_{t_3})$ with Cov $= 0$. And $W_{t_j} - W_{t_i} \sim N(0, t_j - t_i)$

547

A time-continuous real-valued stochastic process $(X_t)_{t\in T}$ is called a Gaussian process if the random vector $(X_{t_1}, \ldots, X_{t_k})$ is (multivariate) normally (Gaussian) distributed for each finite index set $\{t_1, \ldots, t_k\} \subseteq T$; cf. par. 542.

> For any $T' \subseteq T$, $\{X_t \mid t \in T'\}$ consist multivariate normal distribution.
> Each X_i satisfy $X_i \sim N(\mu, \sigma^2)$.

548

A stochastic process $(X_t)_{t\in T}$ is called **stationary** (or: stationary in a strict sense) if

$$\text{Prob}(X_{t_1} \in A_1, X_{t_2} \in A_2, \ldots, X_{t_k} \in A_k) = \text{Prob}(X_{t_1+\tau} \in A_1, X_{t_2+\tau} \in A_2, \ldots, X_{t_k+\tau} \in A_k) \tag{4.22}$$

for each $k \in \mathbb{N}$, $\tau > 0$, all $t_1 < \cdots < t_k$ with $t_j, t_j + \tau \in T$ $(j \leq k)$, and all measurable sets $A_1, \ldots, A_k$.

If the random variables X_j are discrete, (4.22) simplifies to

$$\text{Prob}(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_k} = x_k) = \text{Prob}(X_{t_1+\tau} = x_1, X_{t_2+\tau} = x_2, \ldots, X_{t_k+\tau} = x_k) \tag{4.23}$$

for each $k \in \mathbb{N}$, $\tau > 0$, all $t_1 < \cdots < t_k$ with $t_j, t_j + \tau \in T$ $(j \leq k)$, and all $x_1, \ldots, x_k \in \Omega$.

> $(t_1, ..., t_k)$ and $(t_{\tau+1}, ..., t_{\tau+k})$ can overlap. This means the distributions at t_i and $t_j = \tau + t_i$ $(t_i < t_j)$ are identical for any $\tau \in T$. For example, $\Pr[(1, 0, 1) \text{ in } t_1] = \Pr[(1, 0, 1) \text{ in } t_3 = t_1 + \tau]$. This property is called time-invariance.

549

Stationarity means that the distribution of the stochastic process is time-invariant. In other words: For admissible shifts τ (that is, $T + \tau \subseteq T$), the stochastic processes $(X_t)_{t\in T}$ and $(X_{t+\tau})_{t\in T}$ are identically distributed. If $T = \mathbb{R}$ or $T = [0, \infty)$, for example, any $\tau > 0$ is admissible. For $T = \mathbb{Z}$ or $T = \mathbb{N}$ (time-discrete stochastic processes), the shift parameter τ must be a positive integer.

550

A stochastic process $(X_t)_{t\in T}$ is stationary in a weak sense (or: stationary in a wide sense) if

expectation

$$E(X_t) = E(X_{t+\tau}) \tag{4.24}$$

covariance

$$E((X_{t_1} - \mu)(X_{t_2} - \mu)) = E((X_{t_1+\tau} - \mu)(X_{t_2+\tau} - \mu)) \tag{4.25}$$

for all $t, t+\tau \in T, \tau > 0$.

551

Assume that $(X_t)_{t\in T}$ is a stochastic process on the state space $\mathbb{R}$ for which $E(X_t)$ and $\mathrm{Var}(X_t)$ exist for all $t \in T$. Then

$$\overline{K}_X(t_1, t_2) := E((X_{t_1} - E(X_{t_1}))(X_{t_2} - E(X_{t_2}))) \quad (t_1, t_2 \in T) \tag{4.26}$$

is the autocovariance 하나의 X로부터 Xt1, Xt2를 관측함 function of $(X_t)_{t\in T}$. Furthermore,

$$\overline{\gamma}_X(t_1, t_2) := \frac{E((X_{t_1} - E(X_{t_1}))(X_{t_2} - E(X_{t_2})))}{\sqrt{\mathrm{Var}(X_{t_1})\mathrm{Var}(X_{t_2})}} \quad (t_1, t_2 \in T) \tag{4.27}$$

defines the autocorrelation function of $(X_t)_{t\in T}$.

552

Assume that $(X_t)_{t\in T}$ is a stochastic process on the state space $\mathbb{R}$ for which $E(X_t)$ and $\mathrm{Var}(X_t)$ exist for all $t \in T$. If $(X_t)_{t\in T}$ is stationary or stationary in a weak sense, (4.26) and (4.27) only depend on $|t_1 - t_2|$.

Analogously to par. 551, we define the autocovariance function

$$K_X(h) := E((X_t - E(X_t))(X_{t+h} - E(X_{t+h}))) \quad (h \geq 0; t, t+h \in T) \tag{4.28}$$

and the autocorrelation function (ACF)

$$\gamma_X(h) := \frac{E((X_t - E(X_t))(X_{t+h} - E(X_{t+h})))}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t+h})}} \quad (h \geq 0; t, t+h \in T). \tag{4.29}$$

Note: If $T = [0, \infty)$, for example, $h \in [0, \infty)$ while $h \in \mathbb{N}_0$ if $T \in \{\mathbb{Z}, \mathbb{N}_0, \mathbb{N}\}$.

553

Pars. 585 and 587 to 590 collect useful facts on stationary stochastic processes. Stationarity plays an important role in stochastic models for PTRNGs; cf. Section 4.5. It captures the desired feature that if a PTRNG is analyzed in a certain period in time, its stochastic behavior should be the same at different times.

Note: For stochastic models of physical noise sources, the requirement is relaxed to time-local stationarity; cf. pars. 788 to 790.

554

If a (time-continuous or time-discrete) stationary stochastic process is ergodic, statistical properties of this stochastic process can be deduced from a single, sufficiently long realization of this stochastic process with probability 1.

> [Ergodic] The process $X(t)$ is said to be (mean-)ergodic if the time average estimate converges in squared mean to the ensemble calculated over all possible sample functions X, not time averages average as time goes to infinity.
>
> $$\text{time-dependent} \quad \lim_{T\to\infty}\left(\hat{\mu}_X = \frac{1}{T}\int_0^T X(t)\,dt\right) = \mu_X \quad \text{time-independent}$$
>
> or for the discret-time r.v.
>
> $$\lim_{N\to\infty}\left(\hat{\mu}_X = \frac{1}{N}\sum_{n=1}^{N} X[n]\right) = EX$$
>
> Ergodicity allows a single RNG to function as multiple RNGs. For example, we can generate a random bit sequence with a fair coin by tossing it 100 times instead of using 100 separate fair coins. An RNG can demonstrate ergodicity by restarting itself.

Note 1: In the context of the evaluation of PTRNGs, this feature is exploited for the estimation of parameters, by online tests and by statistical tests applied by the evaluator, for example.

Note 2: There exist several equivalent formal definitions for ergodicity, e.g., that the invariant events are attained with probability 0 or 1. We refer the interested reader to the relevant literature, e.g., to [KaTa75], Chapter 9.

Note 3: Pars. 555 and 556 provide an example of an ergodic process and a counterexample. Loosely speaking, to ensure ergodicity, long-term dependencies of the stochastic process need to decrease sufficiently fast.

555

Example: Assume that the random variables $X_1, X_2, \ldots$ are iid $B(1,p)$-distributed. If we observe a realization sequence $x_1, x_2, \ldots$, the empirical mean $n^{-1}\sum_{j=1}^{n} x_j$ converges to p with probability 1 (Strong law of large numbers; cf. pars. 538 and 539). If the random variables model the repeated tossing of a particular coin (cf. Subsec. 4.5.2), a sequence of realizations can easily be obtained by tossing this coin several times, which allows the estimation of the (unknown) parameter p. The random variables $X_1, X_2, \ldots$ define a stationary ergodic process (cf. par. 556).

556

Counterexample: Assume that the random variables $X_1, X_2, \ldots$ are identically $B(1,p)$. Unlike in par. 555, these random variables are not independent but fully dependent, namely $X_1 = X_2 = \ldots$. Then the realization of X_1 determines the whole realization sequence. In this case, one can only observe the realization sequences $1, 1, \ldots$ (with probability p) or $0, 0, \ldots$ (with probability $1-p$). Hence, it is not possible to estimate p on the basis of a single realization sequence. This stochastic process is stationary but not ergodic.

557

[empirical mean and empirical variance] Assume that $x_1, x_2, \ldots, x_m$ are realizations of the iid random variables $X_1, X_2, \ldots, X_m$. Assume further that the expectation $\mu = E(X_j)$ and the variance $\sigma^2 = \mathrm{Var}(X_j)$ exist. The arithmetic mean $\overline{x}$ and the empirical variance $\overline{s}^2$ of $x_1, x_2, \ldots, x_m$ are given by

$$\overline{x} := \frac{x_1 + x_2 + \cdots + x_m}{m}, \qquad (4.30)$$

$$\overline{s}^2 := \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \overline{x})^2. \qquad (4.31)$$

$\overline{x}$ and $\overline{s}^2$ are *unbiased* estimators of μ and σ^2. In this context, unbiased means, that if the sample values x_j in the right-hand sides of (4.30) and (4.31) are replaced by random variables X_j, the expectation of these terms is μ and σ^2, respectively. Unbiased Estimators

$$\begin{aligned} E[\bar{x}] &= E\left[\frac{x_1 + \cdots + x_m}{m}\right] \\ &= \frac{1}{m}\left[Ex_1 + \cdots + Ex_m\right] \\ &= \frac{1}{m}\left[\mu + \cdots + \mu\right] = \frac{1}{m} m\mu \\ &= \mu = EX \end{aligned}$$

$$
\begin{aligned}
E\left[\overline{s}^2\right] &= E\left[\frac{1}{m-1}\sum_{j=1}^{m}(x_j-\overline{x})^2\right] \\
&= \frac{1}{m-1}\sum_{j=1}^{m}E(x_j-\overline{x})^2 \\
&= \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathrm{Var}(x_j-\overline{x})+\left[E(x_j-\overline{x})\right]^2\right) \\
&= \frac{1}{m-1}\sum_{j=1}^{m}\mathrm{Var}(x_j-\overline{x}) \quad \because E[\overline{x}]=\mu \text{ and } Ex_j=\mu \implies E(x_j-\overline{x})=0 \\
&= \frac{1}{m-1}\sum_{j=1}^{m}\mathrm{Var}\left(\frac{m-1}{m}x_j-\frac{1}{m}\sum_{i\neq j}x_i\right) \quad \because \overline{x}=\frac{x_1+\cdots+x_m}{m} \\
&= \frac{1}{m-1}\sum_{j=1}^{m}\left[\left(\frac{m-1}{m}\right)^2\mathrm{Var}\left(x_j\right)+\left(-\frac{1}{m}\right)^2\sum_{i\neq j}\mathrm{Var}\left(x_i\right)\right] \\
&= \frac{1}{m-1}\sum_{j=1}^{m}\left[\left(\frac{m-1}{m}\right)^2\sigma^2+\left(\frac{1}{m}\right)^2\sum_{i\neq j}\sigma^2\right] \\
&= \frac{m}{m-1}\left[\frac{(m-1)^2}{m^2}\sigma^2+\frac{m-1}{m^2}\sigma^2\right] \\
&= \frac{m-1}{m}\sigma^2+\frac{1}{m}\sigma^2 \\
&= \sigma^2 \\
&= \mathrm{Var}(X)
\end{aligned}
$$

Note: Occasionally, formula (4.31) is used with factor $1/m$ in place of $1/(m-1)$. In this case, the estimator is biased (but asymptotically unbiased).

558

[empirical mean and empirical variance] Assume that the random variables $X_1, X_2, \ldots, X_m$ are iid $N(\mu, \sigma^2)$-distributed. Then

$$\frac{X_1 + X_2 + \cdots + X_m}{m} \sim N\left(\mu, \frac{\sigma^2}{m}\right) \qquad (4.32)$$

and

$$\frac{m-1}{\sigma^2} \cdot \frac{1}{m-1} \sum_{j=1}^{m} (X_j - \overline{X})^2 \sim \chi^2_{m-1} \qquad (4.33)$$

where χ^2_{m-1} denotes the χ^2-distribution with $m-1$ degrees of freedom. Formula (4.33) is a well-known corollary from Cochrane's Theorem. 정규모집단에서 표본평균과 표본분산의 분포

$$E\left[\frac{X_1 + X_2 + \cdots + X_m}{m}\right] = \frac{1}{m}E[X_1] + \cdots + \frac{1}{m}E[X_m] = \mu$$

$$\mathrm{Var}\left[\frac{X_1 + X_2 + \cdots + X_m}{m}\right] = \frac{1}{m^2}\mathrm{Var}[X_1] + \cdots + \frac{1}{m^2}\mathrm{Var}[X_m] = \frac{1}{m}\sigma^2$$

For the iid R.V.s s.t. $(X_i - \mu)/\sigma \sim N(0,1)$ $(i = 1, 2, \cdots, m)$,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Since, $n\bar{X} = \sum_{i=1}^{n} X_i$, and $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$,

$$\begin{aligned}
\sum_{i=1}^{m} \left(\frac{X_i - \mu}{\sigma}\right)^2 &= \sum_{i=1}^{m} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + m\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + \frac{2}{\sigma^2}(\bar{X} - \mu)\sum_{i=1}^{m}(X_i - \bar{X}) \\
&= \sum_{i=1}^{m} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + m\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 \quad \because \sum_{i=1}^{m}(X_i - \bar{X}) = \sum_{i=1}^{m} X_i - m\bar{X} = 0 \\
&= \left(\frac{(m-1)S^2}{\sigma^2}\right) + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{m}}\right)^2
\end{aligned}$$

[Cochrane's Theorem] Let $U_1, ..., U_N$ be i.i.d. standard normally distributed random variables, and $U = [U_1, ..., U_N]^T$. Let $B^{(1)}, B^{(2)}, \ldots, B^{(k)}$ be symmetric matrices. Define r_i to be the rank of $B^{(i)}$. Define $Q_i = U^T B^{(i)} U$, so that the Q_i are quadratic forms. Further assume $\sum_i Q_i = U^T U$.

Cochran's theorem states that the following are equivalent:

- $r_1 + \cdots + r_k = N$
- the Q_i are independent
- each Q_i has a chi-squared distribution with r_i degrees of freedom.

The rank $\sum_{i=1}^{m}\left(\frac{X_i-\mu}{\sigma}\right)^2 = m$ is trivial. Informaly, the rank of $\sum_{i=1}^{m}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ is $m-1$ because the r.v.s are independent except only one r.v. (e.g. X_m), and the rank of $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{m}}\right)^2$ is 1 because $\bar{X} \sim N(\mu, \sigma^2/m)$ is the single normal standard dist. So, rank $\sum_{i=1}^{m}\left(\frac{X_i-\mu}{\sigma}\right)^2 =$ rank $\sum_{i=1}^{m}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 +$ rank $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{m}}\right)^2$, and it means $\sum_{i=1}^{m}\left(\frac{X_i-\mu}{\sigma}\right)^2$ and $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{m}}\right)^2$ are independent by the Cochran's thm.

Furthermore,

$$\sum_{i=1}^{m}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(m), \quad \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{m}}\right)^2 \sim \chi^2(1).$$

Therefore, by the additivity of the Chi-square dist. for the independent terms $\sum_{i=1}^{m}\left(\frac{X_i-\mu}{\sigma}\right)^2, \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{m}}\right)^2$

$$\frac{(m-1)S^2}{\sigma^2} \sim \chi^2(m-1).$$

559

[empirical mean and empirical variance] If the random variables $X_1, X_2, \ldots, X_m$ are iid (but not necessarily normally distributed), then

$$E\left(\frac{1}{m-1}\sum_{j=1}^{m}(X_j-\overline{X})^2\right) = \sigma^2 \tag{4.34}$$

$$\mathrm{Var}\left(\frac{1}{m-1}\sum_{j=1}^{m}(X_j-\overline{X})^2\right) = \frac{1}{m}\left(E((X-\mu)^4) - \frac{m-3}{m-1}\sigma^4\right) \tag{4.35}$$

$$\bar{s}^2 = \frac{1}{m-1}\sum(X_i-\bar{X})^2$$

Let $Y_i = X_i - \bar{X}$ be a r.v. for all i. Then, we can see the

$$\bar{Y} := EY_i = \mu = 0$$

and

$$\begin{aligned}
\mathrm{Var}(Y_i) &= E(Y_i^2) - (EY_i)^2 \\
&= E(Y_i^2) \\
&= E(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= EX_i^2 - 2\bar{X}EX_i + \bar{X}^2 \\
&= EX_i^2 - (EX_i)^2 \\
&= \sigma^2
\end{aligned}$$

That is, $\mathrm{Var}(Y_i) = E(Y_i^2) = \sigma^2$.

Thus,

$$\bar{s}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2 = \frac{1}{m-1}(\sum_{i=1}^{m} Y_i^2 - m\bar{Y}^2)$$

Next, we want to take the variance of $\bar{s}^2$, i.e.,

$$\mathrm{Var}(\bar{s}^2) = \frac{1}{(m-1)^2}\mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2 - m\overline{Y}^2\right)$$

So, we need to find the $\mathrm{Var}\left(\sum Y_i^2 - m\overline{Y}^2\right)$.

$$\mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2 - m\overline{Y}^2\right) = \mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2\right) + m^2\mathrm{Var}\left(\overline{Y}^2\right) - 2\,\mathrm{Cov}\left(\sum_{i=1}^{m} Y_i^2,\ m\overline{Y}^2\right)$$

First,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2\right) &= \sum_{i=1}^{m}\mathrm{Var}(Y_i^2) \\
&= \sum_{i=1}^{m}\left[EY^4 - \left(EY^2\right)^2\right] \\
&= m\left[\mu_4 - \sigma^4\right] \qquad \because E(Y_i^2) = \sigma^2 \text{ and define } EY^4 = \mu_4
\end{aligned}$$

and

$$\begin{aligned}
m^2\mathrm{Var}\left(\overline{Y}^2\right) &= m^2\left[E\overline{Y}^4 - (E\overline{Y}^2)^2\right] \\
&= m^2\left[E\overline{Y}^4 - \left(\frac{\sigma^2}{m}\right)^2\right] \quad \because E(\overline{Y}^2) = \mathrm{Var}(\bar{Y}) + (E\bar{Y})^2 = \frac{\sigma^2}{m} - 0 = \frac{\sigma^2}{m} \\
&= m^2\left[\frac{\mu_4}{m^3} + \frac{3(m-1)\sigma^4}{m^3} - \frac{\sigma^4}{m^2}\right] \\
&= m^2\left[\frac{\mu_4}{m^3} + \frac{2m-3}{m^3}\sigma^4\right] \\
&= \frac{\mu_4}{m} + \frac{2m-3}{m}\sigma^4
\end{aligned}$$

where

$$E\left(\sum Y_i\right)^4 = m\mu_4 + 3m(m-1)\sigma^4 \quad \because E[Y_iY_jY_kY_l] = E[Y_iY_j^3] = 0$$

$$\begin{aligned}
\Longrightarrow\ E\left(\overline{Y}^4\right) &= \frac{1}{m^4}\left\{m\mu_4 + 3m(m-1)\sigma^4\right\} \quad \because \overline{Y} = \frac{\sum Y_i}{m} \implies \overline{Y}^4 = \frac{(\sum Y_i)^4}{m^4} \\
&= \frac{\mu_4}{m^3} + \frac{3(m-1)\sigma^4}{m^3}
\end{aligned}$$

Last, because

$$2\,\mathrm{Cov}\left(\sum_{i=1}^{m} Y_i^2,\ m\overline{Y}^2\right) = 2m\left(E\left[\sum_{i=1}^{m} Y_i^2 \cdot \overline{Y}^2\right] - E\sum_{i=1}^{m} Y_i^2 \cdot E\overline{Y}^2\right),$$

$$\begin{aligned}
E\left[\sum_{i=1}^{m} Y_i^2 \cdot \overline{Y}^2\right] &= E\left[(\sum Y_i^2)\cdot\left(\frac{1}{m}\sum Y_j\right)^2\right] \\
&= \frac{1}{m^2}E\left[\sum_i Y_i^2 \cdot \sum_j\sum_k Y_jY_k\right] \\
&= \frac{1}{m^2}[m\mu_4 + m(m-1)\sigma^4] \qquad \because E[Y_iY_jY_kY_l] = E[Y_iY_j^3] = 0 \\
&= \frac{\mu_4}{m} + \frac{m-1}{m}\sigma^4
\end{aligned}$$

and

$$E\sum_{i=1}^{m} Y_i^2 \cdot E\overline{Y}^2 = (m\sigma^2)\left(\frac{\sigma^2}{m}\right) = \sigma^4$$

Thus,

$$2\,\mathrm{Cov}\left(\sum_{i=1}^{m} Y_i^2,\ m\overline{Y}^2\right) = 2m\left[\left(\frac{\mu_4}{m} + \frac{m-1}{m}\sigma^4\right) - \sigma^4\right] = 2(\mu_4 - \sigma^4)$$

Consequently,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2 - m\overline{Y}^2\right) &= \mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2\right) + m^2\mathrm{Var}\left(\overline{Y}^2\right) - 2\,\mathrm{Cov}\left(\sum_{i=1}^{m} Y_i^2,\ m\overline{Y}^2\right) \\
&= m\left[\mu_4 - \sigma^4\right] + \left(\frac{\mu_4}{m} + \frac{2m-3}{m}\sigma^4\right) - 2(\mu_4 - \sigma^4) \\
&= \frac{(m-1)^2}{m}\mu_4 - \frac{(m-1)(m-3)}{m}\sigma^4
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{m-1}\sum(X_i-\bar{X})^2\right) &= \mathrm{Var}(\bar{s}^2) \\
&= \frac{1}{(m-1)^2}\mathrm{Var}\left(\sum_{i=1}^{m} Y_i^2 - m\overline{Y}^2\right) \\
&= \frac{1}{(m-1)^2}\left(\frac{(m-1)^2}{m}\mu_4 - \frac{(m-1)(m-3)}{m}\sigma^4\right) \\
&= \frac{1}{m}\mu_4 - \frac{m-3}{m(m-1)}\sigma^4 \\
&= \frac{1}{m}E(X-\mu)^4 - \frac{m-3}{m(m-1)}\sigma^4 \qquad \because EY^4 = \mu_4 \ \text{ and } \ Y_i = X_i - \bar{X} \quad \square
\end{aligned}$$

560

[Allan variance] When estimating the jitter of digital clock signals, for example, the empirical variance may overestimate the jitter if low frequency noise such as flicker noise is present. In such scenarios, often the (empirical) Allan variance is used instead; cf., e.g., [ASPB+18]. Assume that the measurement values $x_1, x_2, \ldots, x_m$ are taken at times $\tau, 2\tau, \ldots, m\tau$ (in practice, the x_j often are fractional frequencies that have been averaged over an interval of length τ). The (empirical) Allan variance of $x_1, x_2, \ldots, x_m$ is defined by

$$\overline{\mathrm{AVar}} = \frac{1}{2(m-1)} \sum_{j=1}^{m-1} (x_j - x_{j+1})^2. \qquad (4.36)$$

> Allan variance analyzes the difference between consecutive measurements rather than comparing each measurement to the overall mean. Typically, implementations use this approach because measured values tend to change monotonically over time in the real world. For example, in the case of thermal noise, the noise level can differ during the day, when temperatures rise, and at night, when temperatures fall. Calculating the variance with this in mind helps prevent the variance from increasing unnecessarily.

Note 1: By construction, the Allan variance is only a little sensitive to slow drifts of the distributions of the corresponding random variables $X_1, X_2, \ldots, X_m$.

Note 2: The definition of the Allan variance is not uniform in the literature.

561

[Allan variance] Assume that the measurement values $x_1, x_2, \ldots$ are realizations of the random variables $X_1, X_2, \ldots$. If the random variables are stationarily distributed and uncorrelated (i.e., $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$), the Allan variance coincides with the "usual" variance [ASPB+18], Theorem 1.

> The stationary distribution, such as IID, also conforms to the Allan variance.

1.2 Useful Theorems and Facts

562

Subsection 4.2.2 provides facts and theorems that can be useful in the context of this document.

563

[Stirling's approximation]

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} \quad \text{(Stirling's approximation)} \quad (4.37)$$

Stirling's approx. (general)

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad \text{for large } n$$

proof. We use Laplace's method

$$\int_a^b e^{Mf(x)}dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}\, e^{Mf(x_0)}$$

as $M \to \infty$ for x_0 is a maximum point. Then,

$$\begin{aligned}
n! &= \Gamma(n+1) \\
&= \int_0^\infty x^n e^{-x}dx = \int_0^\infty e^{n\ln x}e^{-x}dx = \int_0^\infty e^{n\ln x - x}dx \\
&= \int_0^\infty e^{n\ln ny - ny} n dy \qquad \text{(changing variables to } x = ny \text{ for applying the Laplace's method)} \\
&= \int_0^\infty n e^{n\ln n} e^{n(\ln y - y)}dy = n e^{n\ln n}\int_0^\infty e^{n(\ln y - y)}dy = n\cdot n^n \int_0^\infty e^{n(\ln y - y)}dy \\
&\approx n\cdot n^n \sqrt{\frac{2\pi}{n}}\, e^{-n} \\
&= \sqrt{2\pi n}\left(\frac{n}{e}\right)^n
\end{aligned}$$

[Euler-Maclaurin formula] for B_{2k}: Bernoulli numbers

$$\sum_{i=m}^{n} f(i) \approx \int_m^n f(x)\,dx + \frac{f(m)+f(n)}{2} + \sum_{k=1}^{p}\frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(m)\right)$$

[Robbins Inequality]

$$\ln n! = \sum_{k=1}^{n}\ln k = (n+\frac{1}{2})\ln n - n + \ln\sqrt{2\pi} + \mu(n)$$

For $\mu(n)$,

$$\mu(n) - \mu(n+1) = \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \frac{1}{7(2n+1)^6} + \dots$$

where $x = \frac{1}{2n+1}$, because

$$\mu(n) - \mu(n+1) = (n + \frac{1}{2}) \ln(1 + \frac{1}{n}) - 1.$$

Then, the upper bound is

$$\mu(n) - \mu(n+1) > \frac{1}{3(2n+1)^2} > \frac{1}{12n(n+1)} = \frac{1}{12n} - \frac{1}{12(n+1)} \implies \mu(n) < \frac{1}{12n}$$

and the lower bound is

$$\mu(n) - \mu(n+1) < \frac{1}{12n+1} - \frac{1}{12(n+1)+1} \implies \mu(n) > \frac{1}{12n+1}$$

Therefore,

$$\ln \sqrt{2\pi} n^{(n+1/2)}/e^n + \ln e^{\frac{1}{12n+1}} < \ln n! < \ln \sqrt{2\pi} n^{(n+1/2)}/e^n + \ln e^{\frac{1}{12n}}$$

$$\therefore \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} \qquad \square$$

564

[Stirling's approximation] If n, k, and $n-k$ are large, applying the lower bound in Stirling's formula (4.37) to the factorials of $\binom{n}{k}$ yields the approximation:

$$\binom{n}{k} \approx \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}} \qquad (4.38)$$

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
&\approx \frac{\sqrt{2\pi n} \cdot \frac{n^n}{e^n}}{\left(\sqrt{2\pi k} \cdot \frac{k^k}{e^k}\right)\left(\sqrt{2\pi(n-k)} \cdot \frac{(n-k)^{n-k}}{e^{n-k}}\right)} \qquad \text{(by Stirling's approx.)} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}}
\end{aligned}$$

565

[Expectation: computation rules] Assume that for the (not necessarily independent nor identically distributed) random variables $X_1, \ldots, X_k$, the expectations $E(X_j)$ exist. Let $Y = a_1X_1 + \cdots + a_kX_k$ with $a_1, \ldots, a_k \in \mathbb{R}$. Then the expectation of Y exists. More precisely,

$$E(Y) = E(a_1X_1 + \cdots + a_kX_k) = \sum_{j=1}^{k} a_jE(X_j). \tag{4.39}$$

If the random variables are iid and $a_j = 1/k$ for each $j \leq k$, then $E(Y) = E(X_1) = \cdots = E(X_k)$.

$$E(aX) = \sum_i (ax_i)P(X = x_i) = a\sum_i x_iP(X = x_i) = aE(X) \therefore\ E(ax) = aE(X)$$

Next, for $P(X_1 = x_i, X_2 = y_j)$,

$$\begin{aligned}
E(X_1 + X_2) &= \sum_i \sum_j (x_i + y_j)P(X_1 = x_i, X_2 = y_j) \\
&= \sum_i \sum_j x_iP(X_1 = x_i, X_2 = y_j) \\
&\quad + \sum_i \sum_j y_jP(X_1 = x_i, X_2 = y_j) \\
&= \sum_i x_i \left[\sum_j P(X_1 = x_i, X_2 = y_j)\right] \\
&\quad + \sum_j y_j \left[\sum_i P(X_1 = x_i, X_2 = y_j)\right] \\
&\quad \because \sum_j P(X_1 = x_i, X_2 = y_j) = P(X_1 = x_i) \\
&= \sum_i x_iP(X_1 = x_i) + \sum_j y_jP(X_2 = y_j) \\
&= E(X_1) + E(X_2)
\end{aligned}$$

$$\therefore\ E(X_1 + X_2) = E(X_1) + E(X_2)$$

Moreover,

$$E(Y) = \sum_{j=1}^{k} \frac{1}{k}E(X_j) = \frac{1}{k}\sum_{j=1}^{k} \mu = \frac{1}{k} \cdot k\mu = \mu$$

566

[Variance: computation rules] Assume that for the independent (but not necessarily identically distributed) random variables $X_1, \ldots, X_k$, the variances $\mathrm{Var}(X_j)$ exist. Let $Y = a_1X_1 + \cdots + a_kX_k$ for $a_1, \ldots, a_k \in \mathbb{R}$. Then the expectation of Y exists. More precisely,

$$\mathrm{Var}(Y) = \mathrm{Var}(a_1X_1 + \cdots + a_kX_k) = \sum_{j=1}^{k} a_j^2 \mathrm{Var}(X_j). \tag{4.40}$$

If we drop the assumption that the random variables $X_1, \ldots, X_k$ are independent, then (4.40) becomes more complicated:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(a_1X_1 + \cdots + a_kX_k) \\ &= \sum_{j=1}^{k} a_j^2 \mathrm{Var}(X_j) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j). \end{aligned} \tag{4.41}$$

$$\begin{aligned} Var(Y) &= E\left[\left(\sum_{j=1}^{k} a_jX_j - \sum_{j=1}^{k} a_j\mu_j\right)^2\right] \qquad \text{by (4.11)} \\ &= E\left[\left(\sum_{j=1}^{k} a_j(X_j - \mu_j)\right)^2\right] \\ &= E\left[\sum_{j=1}^{k} a_j^2(X_j - \mu_j)^2 + \sum_{i \neq j} a_ia_j(X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{j=1}^{k} a_j^2 E[(X_j - \mu_j)^2] + \sum_{i \neq j} a_ia_j E[(X_i - \mu_i)(X_j - \mu_j)] \\ \therefore\ Var(Y) &= \sum_{j=1}^{k} a_j^2 Var(X_j) + \sum_{i \neq j} a_ia_j Cov(X_i, X_j) \end{aligned}$$

If $X_1, \ldots, X_k$ are Independent, $Cov(X_i, X_j) = 0$ for all $i \neq j$. Thus

$$Var(Y) = \sum_{j=1}^{k} a_j^2 Var(X_j)$$

567

Example: Expectation and variance of $B(n, p)$-distributed random variables.

The random variable $Y = Y_1 + \cdots + Y_n \sim B(n, p)$ if $Y_1, \ldots, Y_n$ are iid $B(1, p)$-distributed. By (4.39) and (4.40) we conclude that $E(Y) = E(Y_1) + \cdots + E(Y_n) = np$ and $\mathrm{Var}(Y) = \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_n) = np(1-p)$.

568

[Central Limit Theorem (CLT)] Assume that the real-valued random variables $X_1, X_2, \ldots$ are iid with expectation μ and variance σ^2. For $n = 1, 2, \ldots,$

$$S_n^* := \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \qquad (4.42)$$

define normalized partial sums. The Central Limit Theorem (CLT) applies to the sequence $X_1, X_2, \ldots$. More precisely,

$$\lim_{n\to\infty} \text{Prob}(S_n^* \le x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad \text{for all } x \in \mathbb{R}. \qquad (4.43)$$

First, we need a few facts about moment generating functions(mgf). Recall that the mgf of a r.v. X is $M_X(t) = E(e^{tX})$. We'll write this as

$$\int_{-\infty}^{\infty} e^{tx} f(x) dx$$

and think of f as a continuous random variable, although the result holds for discrete r.v.s as well. If we differentiate the mgf r times, we get

$$M^{(r)}(t) = \frac{d^r}{dt^r} \int_{-\infty}^{\infty} e^{tx} f(x)\ dx = \int_{-\infty}^{\infty} \frac{d^r}{dt^r} \left[e^{tx} f(x)\right] dx = \int_{-\infty}^{\infty} x^r e^{tx} f(x)\ dx.$$

The third equality requires some justification, in order to interchange differentiation and integration, but this should seem believable.

Therefore, we get

$$M^{(r)}(0) = \int_{-\infty}^{\infty} x^r e^{tx} f(x) dx = E(X^r).$$

Thm.1 *Let F_n be a sequence of cumulative distribution functions(cdf) with the correspond mgfs M_n. Let F be a cdf with the mgf M. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ for all x at which F is continuous.*

This says that if the moment generating functions of a sequence of distributions approach some limit, then that limiting mgf is the mgf of the limit of that sequence of distributions. So in order to prove the CLT, it will be enough to show that the mgf of a standardized sum of n iid r.v.s approaches th mgf of a standard normal as $n \to \infty$.

Thm.2 (CLT) *Let $X_1, X_2, \ldots$ be a sequence of independent r.v.s having mean μ and variance σ^2. Let each X_i have the cdf $P(X_i \le x) = F(x)$ and the mgf $M(t) = E(e^{tX_i})$. Let $S_n = \sum_{i=1}^{n} X_i$. Then*

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu n}{\sigma\sqrt{n}} \le x\right) = \Phi(x)$$

for $-\infty < x < \infty$

Proof. It suffices to do the proof in the case $\mu = 0$. If $\mu \neq 0$, let $Y_i = X_i - \mu$ for each i. Let $T_n = Y_1 + \cdots + Y_n$. Then we have

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu n}{\sigma\sqrt{n}} \leq x\right) = \lim_{n\to\infty} P\left(\frac{T_n}{\sigma\sqrt{n}} \leq x\right)$$

and so it suffices to prove the central limit thm. in the case $\mu = 0$. Let $Z_n = S_n/\sigma\sqrt{n}$. We'll show that the mgf of Z_n tends to the mgf $M_Z(t) = \exp(t^2/2)$ of the standard normal dist. Since S_n is a sum of independent r.v.s,

$$M_{S_n}(t) = [M(t)]^n$$

and we have

$$M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

It will suffice to show that $\lim_{n\to\infty} n \log M(t/\sigma\sqrt{n}) = t^2/2$; then we can take exponentials and use Thm.1 to get the result. Call this limit L. Let $x = 1/\sqrt{n}$ in this limit; then we need to find

$$L = \lim_{n\to\infty} \frac{\ln M\left(\frac{t}{\sigma}\frac{1}{\sqrt{n}}\right)}{\left(\frac{1}{\sqrt{n}}\right)^2} = \lim_{x\to 0} \frac{\ln M(tx/\sigma)}{x^2}.$$

This is of indeterminate form $0/0$, since $M(0) = 1$. Differentiating (L'Hopital's rule) gives

$$L = \lim_{x\to 0} \frac{\frac{M'(tx/\sigma)}{M(tx/\sigma)}\frac{t}{\sigma}}{2x}$$

and we can pull out a constant to get

$$L = \frac{t}{2\sigma} \lim_{x\to 0} \frac{M'(tx/\sigma)}{xM(tx/\sigma)}.$$

This is again indeterminate of form 0 over 0. Differentiating again gives

$$L = \frac{t}{2\sigma} \lim_{x\to 0} \frac{M''(tx/\sigma)\cdot\frac{t}{\sigma}}{M(tx/\sigma) + xM'(tx/\sigma)\cdot\frac{t}{\sigma}}.$$

Pulling out a constant and rearranging limits gives

$$L = \frac{t^2}{2\sigma^2} \frac{\lim_{x\to 0} M''(tx/\sigma)}{\lim_{x\to 0} M(tx/\sigma) + \frac{t}{\sigma}\lim_{x\to 0} xM'(tx/\sigma)} = \frac{t^2}{2\sigma^2}\frac{M''(0)}{M(0)}.$$

Now we recall $M^{(r)}(0) = E(X^r)$; thus $M(0) = E(1), M'(0) = E(X) = 0, M''(0) = E(X^2) = E(X)^2 + \mathrm{Var}(X) = 0 + \sigma^2 = \sigma^2$. This finally gives

$$L = \frac{t^2}{2\sigma^2}\frac{\sigma^2}{1} = \frac{t^2}{2},$$

which is what we wanted. □

569

[tail of the standard normal distribution] ([GaSt77], Lemma 1.19.2) For $x > 0$, it is

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \le 1 - \Phi(x) = \Phi(-x) \le \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}. \qquad (4.44)$$

For the $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$,

$$\Phi(x) = 1 - \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \frac{1}{u}\cdot(ue^{-u^2/2})du \qquad \text{for } x > 0.$$

and then we can calculate

$$\int_x^\infty \frac{1}{u}\cdot(ue^{-u^2/2})du = \left[-\frac{1}{u}e^{-u^2/2}\right]_x^\infty - \int_x^\infty \frac{1}{u^2}e^{-u^2/2}du \;=\; \frac{1}{x}e^{-x^2/2} - \int_x^\infty \frac{1}{u^2}e^{-u^2/2}du$$

using the integration by parts. Since the term $\int_x^\infty \frac{1}{u^2}e^{-u^2/2}du$ is positive,

$$1 - \Phi(x) < \frac{1}{\sqrt{2\pi}\,x}e^{-x^2/2}.$$

Similarly, for the remain term $\int_x^\infty \frac{1}{u^2}e^{-u^2/2}du$,

$$\int_x^\infty \frac{1}{u^2}e^{-u^2/2}du = \left[-\frac{1}{u^3}e^{-u^2/2}\right]_x^\infty - \int_x^\infty \frac{3}{u^4}e^{-u^2/2}du = \frac{1}{x^3}e^{-x^2/2} - \int_x^\infty \frac{3}{u^4}e^{-u^2/2}du.$$

Consequently,

$$\begin{aligned} 1 - \Phi(x) &= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{x}e^{-x^2/2} - \left(\frac{1}{x^3}e^{-x^2/2} - \int_x^\infty \frac{3}{u^4}e^{-u^2/2}du\right)\right) \\ &= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{x} - \frac{1}{x^3}\right)e^{-x^2/2} + \frac{3}{\sqrt{2\pi}}\int_x^\infty \frac{1}{u^4}e^{-u^2/2}du. \end{aligned}$$

Since the term $\frac{3}{\sqrt{2\pi}}\int_x^\infty \frac{1}{u^4}e^{-u^2/2}du$ is positive,

$$1 - \Phi(x) \ge \left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$ □

570

[CLT, parameter estimation] Assume that $X_1, X_2, \ldots$ are iid $B(1,p)$-distributed. Then the CLT implies

$$\begin{aligned}\text{Prob}\left(\left|\frac{1}{N}\sum_{j=1}^{N} X_j - p\right| > \epsilon\right) &= \text{Prob}\left(\left|\frac{\sum_{j=1}^{N} X_j - Np}{N}\right| > \epsilon\right) \\ &= \text{Prob}\left(\left|\frac{\sum_{j=1}^{N} X_j - Np}{\sqrt{N}\sqrt{p(1-p)}}\right| > \frac{\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \\ &= 2\Phi\left(\frac{-\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \leq 2\Phi(-2\epsilon\sqrt{N}). \qquad (4.45)\end{aligned}$$

By the assumption, $E[X_j] = p, Var(X_j) = p(1-p)$ for all j and $E\bar{X} = E\left[\frac{1}{N}\sum_{j=1}^{N} X_j\right] = p$, $\text{Var}(\bar{X})\frac{p(1-p)}{N}$ for N independent samples.

$$\begin{aligned}\text{Prob}(|\bar{X} - p| > \epsilon) &= \text{Prob}\left(\left|\frac{\sum_{j=1}^{N} X_j}{N} - p\right| > \epsilon\right) \\ &= \text{Prob}\left(\left|\sum_{j=1}^{N} X_j - Np\right| > N\epsilon\right) \\ &= \text{Prob}\left(\frac{|\sum X_j - Np|}{\sqrt{Np(1-p)}} > \frac{\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \\ &= 2 \cdot \text{Prob}\left(\frac{\sum X_j - Np}{\sqrt{Np(1-p)}} < -\frac{\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \\ &\qquad \text{(take the negative part to modify into the form } \Phi\text{)}\end{aligned}$$

Moreover, the function $f(p) = p(1-p)$ has maximum $1/4$ where $p = 1/2$. Therefore, $\frac{-1}{\sqrt{p(1-p)}} \leq -2$ and $\epsilon\sqrt{N} > 0$

$$\begin{aligned}\text{Prob}(|\bar{X} - p| > \epsilon) &= 2\ \text{Prob}\left(\frac{\sum X_j - Np}{\sqrt{Np(1-p)}} < -\frac{\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \\ &= 2\,\Phi\left(\frac{-\epsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \leq 2\Phi(-2\epsilon\sqrt{N})\end{aligned}$$

571

[CLT] Par. 568 formulates the Central Limit Theorem (CLT) for iid random variables. The CLT is very robust and holds under weak conditions. Under suitable conditions, the iid assumption and even the independence property can be dropped. Some special cases are covered in paragraphs 572, 583, and 584.

Background information: If the CLT applies, the random variables $S_1^*, S_2^*, \ldots$ converge to $N(0,1)$ in distribution. We do not go deeper but refer the interested reader to ([Geor15], Subsection 5.3).

572

[CLT] Assume that the real-valued random variables $X_1, X_2, \ldots$ are independent (but not necessarily iid) with expectations $E(X_j) = \mu_j$ and variances $\mathrm{Var}(X_j) = \sigma_j^2$ for $j \in \mathbb{N}$. For $n = 1, 2, \ldots,$

$$S_n^* := \frac{\sum_{j=1}^n (X_j - \mu_j)}{\sqrt{s_n^2}} \quad \text{with } s_n^2 := \sum_{j=1}^n \sigma_j^2 \tag{4.46}$$

defines normalized partial sums.

$$E[S_n^*] = E\left[\frac{\sum_{j=1}^n (X_j - \mu_j)}{s_n}\right] = \frac{1}{s_n}\sum_{j=1}^n (E[X_j] - \mu_j) = \frac{1}{s_n}\sum_{j=1}^n (\mu_j - \mu_j) = 0$$

$$\mathrm{Var}(S_n^*) = \mathrm{Var}\left[\frac{\sum (X_j - \mu_j)}{s_n}\right] = \frac{1}{s_n^2}\sum_{j=1}^n \mathrm{Var}(X_j - \mu_j) = \frac{1}{s_n^2}\sum_{j=1}^n \sigma_j^2 = \frac{s_n^2}{s_n^2} = 1$$

Assume further that the Lindeberg condition holds

$$\lim_{n\to\infty} L_n(\delta) = 0 \quad \text{for all } \delta > 0$$
$$\text{where} \quad L_n(\delta) := \tfrac{1}{s_n^2}\textstyle\sum_{j=1}^n E((X_j - \mu_j)^2 1_{\{|X_j - \mu_j| \geq \delta s_n\}}). \tag{4.47}$$

The equation

$$L_n(\delta) := \frac{1}{s_n^2}\sum_{j=1}^n E\left[(X_j - \mu_j)^2 \cdot 1_{\{|X_j - \mu_j| \geq \delta s_n\}}\right] \to 0$$

implies that the sum of the $(X_j - \mu_j)^2$ exceeding a certain value δs_n converges to 0 as $n \to \infty$. In other words, there is no r.v. indicating overwhelming deviation.

Then the Central Limit Theorem (CLT) applies to the sequence $X_1, X_2, \ldots$. In particular,

$$\lim_{n\to\infty} \mathrm{Prob}(S_n^* \leq x) = \Phi(x) \quad \text{for all } x \in \mathbb{R}. \tag{4.48}$$

Let $Y_j = \frac{X_j - \mu_j}{s_n}$ is a characteristic function. Then, $S_n^* = \sum Y_j$. The characteristic function of S_n^* is

$$\varphi_{S_n^*}(t) = \prod_{j=1}^{n} \varphi_{Y_j}(t) = \prod_{j=1}^{n} E[e^{itY_j}] \qquad (\varphi_X(t) := E[e^{itX}])$$

with the independent property.

Applying Taylor series on $E[e^{itY_j}]$ at $t = 0$,

$$\begin{aligned} e^{itY_j} &= 1 + itY_j + \frac{(itY_j)^2}{2} + R(itY_j) \\ &= 1 + itY_j - \frac{t^2 Y_j^2}{2} + R(itY_j) \\ \varphi_{Y_j}(t) = E[e^{itY_j}] &= E[1] + itE[Y_j] - \frac{t^2 E[Y_j^2]}{2} + E[R(itY_j)] \\ &= 1 - \frac{t^2 \sigma_j^2}{2s_n^2} + E[R(itY_j)] \qquad (\because E[Y_j] = 0, E[Y_j^2] = Var(Y_j) = \frac{\sigma_j^2}{s_n^2}) \end{aligned}$$

where $R_{nj}(t)$ is remainder.

Next, use the property $\ln(1 + x) \approx x$.

$$\begin{aligned} \ln \varphi_{S_n^*}(t) &= \sum_{j=1}^{n} \ln \left(1 - \frac{t^2 \sigma_j^2}{2s_n^2} + R_{nj}(t) \right) \\ &\approx \sum_{j=1}^{n} \left(-\frac{t^2 \sigma_j^2}{2s_n^2} + R_{nj}(t) \right) \\ &= -\frac{t^2}{2s_n^2} \sum \sigma_j^2 + \sum R_{nj}(t) \\ &= -\frac{t^2}{2} + \sum R_{nj}(t). \end{aligned}$$

If the condition of (4.47) satisfies, $\sum R_{nj}(t) \to 0$ Higher-order terms converge faster than the first and second-order terms as $n \to \infty$. Consequently

$$\lim_{n\to\infty} \ln \varphi_{S_n^*}(t) = -\frac{t^2}{2} \implies \lim_{n\to\infty} \varphi_{S_n^*}(t) = e^{-t^2/2}$$

By the Levy's Continuity Theorem, the characteristic function converges to that of normal distribution. Thus, r.v. S_n^* converges in normal distribution.

$$\lim_{n\to\infty} \text{Prob}(S_n^* \le x) = \Phi(x)$$

573

[CLT] Assume that the random variables $X_1, X_2, \ldots$ are iid and that besides $E(X_1)$ and $E(X_1^2)$, also the third moment $E(X_1^3)$ exist. Then the well-known Berry-Esséen-Theorem provides an upper bound for the maximal difference between the exact cumulative distribution function of S_n^* and $\Phi(\cdot)$. It is

$$\text{(Berry-Esséen-Theorem)} \quad |\text{Prob}(S_n^* \le x) - \Phi(x)| \le C \frac{E(|X_1 - E(X_1)|^3)}{(\text{Var}(X_1))^{1.5}} \frac{1}{\sqrt{n}} \quad \text{for each } x \in \mathbb{R} \quad (4.49)$$

for a suitable constant C (cf. [Geor15], Bemerkung (5.31), with $C = 0.8$). In [Shev11], it is proved that $C < 0.4748$. In particular, (4.49) says that the rate of convergence is $O(n^{-0.5})$.

First, we assume that $X_1, X_2, \ldots, X_n$ are iid and satisfy the follow conditions.

$$E(X_1) = 0, \qquad Var(X_1) = \sigma^2, \qquad E|X_1|^3 = \rho < \infty, \qquad S_n^* = \frac{\sum_{j=1}^n X_j}{\sigma\sqrt{n}}$$

The most powerful tool for proving Esseen's Smoothing Lemma is the inequality below, which transforms a gap between CDFs into a gap between characteristic functions. The following conditions hold for all $T > 0$.

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le \frac{k}{T} + \frac{1}{\pi} \int_{-T}^{T} \left| \frac{\varphi_n(t) - e^{-t^2/2}}{t} \right| dt$$

where $\varphi_n(t) = E[e^{itS_n^*}]$ is the characteristic function of S_n^* and k is a constant.
We expand the characteristic function $\varphi(t)$ of X_j to third-order at $t = 0$.

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + \theta \frac{\rho |t|^3}{6}, \quad (|\theta| \le 1)$$

The characteristic function $\varphi_n(t)$ of the standardized sum S_n^* is shown as follows due to its independence.

$$\varphi_n(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}} \right) \right]^n = \left[1 - \frac{t^2}{2n} + \theta \frac{\rho |t|^3}{6\sigma^3 n^{3/2}} \right]^n$$

Next, we estimate the error using the log approximation $\ln(1 + z) \approx z$ as $n \to \infty$.

$$\ln \varphi_n(t) \approx n \left(-\frac{t^2}{2n} + \frac{\rho |t|^3}{6\sigma^3 n^{3/2}} \right) = -\frac{t^2}{2} + \frac{\rho |t|^3}{6\sigma^3 \sqrt{n}}$$

Let $T = \frac{\sigma^3 \sqrt{n}}{4\rho}$. Then

$$|\varphi_n(t) - e^{-t^2/2}| \le \frac{\rho |t|^3}{\sigma^3 \sqrt{n}} e^{-t^2/4}.$$

Now we substitute it into the integration of Esseen's Smoothing Lemma.

$$\int_{-T}^{T} \frac{|\varphi_n(t) - e^{-t^2/2}|}{|t|} dt \leq \frac{\rho}{\sigma^3\sqrt{n}} \int_{-\infty}^{\infty} t^2 e^{-t^2/4} dt$$

Because the integral term on the right side converges to a constant, the full term is proportional to $\frac{\rho}{\sigma^3\sqrt{n}}$.
The first term k/T is also proportional to $T \propto \sqrt{n}$, thus it is of the same order as $1/\sqrt{n}$.
Eq.(4.49) is derived by combining two terms as follows.

$$|Prob(S_n^* \leq x) - \Phi(x)| \leq C \cdot \frac{E|X_1 - E(X_1)|^3}{(Var(X_1))^{1.5}} \cdot \frac{1}{\sqrt{n}}.$$

As a result, the error size decreases at a rate of $1/\sqrt{n}$; $O(n^{-0.5})$. Additionally, the case where $C < 0.4748$ has been examined.

574

A sequence $X_1, X_2, \ldots$ of random variables is called q-dependent if the random vectors $(X_1, \ldots, X_u)$ and $(X_v, \ldots, X_n)$ are independent for all $1 \leq u < v \leq n$ with $v - u > q$.

Note: The components of each vector need not be independent.

Simply, it means "if variables are separated by a distance greater than q, they have no relationship (they are independent)."
The formula $v - u > q$ implies the following:

- First Group: The set from X_1 to X_u (past data)
- Second Group: The set from X_v to X_n (future data)
- Condition: If the distance between the start of the second group (v) and the end of the first group (u) is greater than q, these two groups are mathematically independent.

The 'Note' at the bottom of the image is a crucial distinction. This means that while groups separated by a large distance are independent, items within the same group or close neighbors (like X_1 and X_2) do not need to be independent (i.e., they can have dependence).

575

CLT의 R.V.들이 IID라는 가정을 거리가 q보다 멀 때 independent라는 가정으로 완화함

[CLT for q-dependent random variables, [HoRo48]] Let $X_1, X_2, \dots$ be a q-dependent (not necessarily stationary) sequence of random variables such that $E\left[|X_i|^3\right]$ is uniformly bounded for all $i \in \mathbb{N}$.

$$A_i := \mathrm{Var}(X_{i+q}) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_{i+q-j}, X_{i+q}) \quad \text{for } i \in \mathbb{N}. \tag{4.50}$$

Since,

$$\mathrm{Var}(S_n) = \sum_{j=1}^{n} \mathrm{Var}(X_j) + 2 \sum_{1 \le j < k \le n} \mathrm{Cov}(X_j, X_k)$$

for the sum of r.v.s $S_n = \sum_{j=1}^{n} X_j$.

By the definition of the q-dependent, $\mathrm{Cov}(X_i, X_k) = 0$ if $|k - j| > q$. Thus, the series of the covariance term can be truncated at the q-th term.

$$\mathrm{Var}(S_n) = \sum_{j=1}^{n} \mathrm{Var}(X_j) + 2 \sum_{j=q+1}^{n} \sum_{k=1}^{q} \mathrm{Cov}(X_j, X_{j-k}) + C$$

where C indicates truncated terms.

For any (time) index i, A_i is the sum of the variance at $i+q$ and the covariance of q terms.

$$A_i = \mathrm{Var}(X_{i+q}) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_{i+q-j}, X_{i+q}).$$

Then, $\sum A_i$ forms S_n excluding the boundaries of each A_i. As a result, each A_i contributes as 'Incremental Variance' to show the average rate at which variance increases as n increases.

If the limit $A := \lim_{u\to\infty} u^{-1} \sum_{h=1}^{u} A_{i+h}$ exists uniformly for all $i \in \mathbb{N}$, then

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{\sum_{j=1}^{n} (X_j - E(X_j))}{\sqrt{An}} \le x \right) = \Phi(x)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad \text{for all } x \in \mathbb{R}. \tag{4.51}$$

It shows that the convergence of the distribution of the total sum as n increases.

First, let A, the average rate at which variance increases, as follow.

$$A := \lim_{u\to\infty} u^{-1} \sum_{h=1}^{u} A_{i+h}$$

If it exists, $\mathrm{Var}(S_n)$ increases linearly with A_n as $n \to \infty$; $Var(S_n) \sim An$.

For using Bernstein's Blocking Method, we split the number of r.v.s n into p and q; $n = k(p+q)+r$ $(0 \le r < p+q)$. It serves block sizes; p-size large and q-size small blocks. Let

$$U_j = \sum_{i=(j-1)(p+q)+1}^{(j-1)(p+q)+p} (X_i - E[X_i]), \qquad j = 1, \ldots, k$$

$$V_j = \sum_{i=(j-1)(p+q)+p+1}^{j(p+q)} (X_i - E[X_i]), \quad j = 1, \ldots, k$$

$$R_n = \sum_{i=k(p+q)+1}^{n} (X_i - E[X_i]).$$

U_j represents the sum of p-size blocks and V_j is the sum of q-size blocks (R_n is remainder). Then,

$$S_n = \sum_{j=1}^{k} U_j + \sum_{j=1}^{k} V_j + R_n.$$

According to the definition of the q-dependence, X_i and X_j are independent if $|i-j| > q$ because the last index of U_i is $(j-1)(p+q)+p$ and the first index of U_{j+1} is $j(p+q)+1$; $\{j(p+q)+1\} - \{(j-1)(p+q)+p\} = q+1 > q$. It means that U_j and U_m are independent for any $j \neq m$.

Next, we adjust the parameters so that $p \to \infty, q \to \infty$, and $\frac{q}{p} \to 0$ as $n \to \infty$. Based on the assumption, $\mathrm{Var}(S_n) \sim An$ and $E[V_j^2] \approx Aq$, where the block size of V_j is q. Since $k \approx n/p$,

$$\frac{\mathrm{Var}(\sum V_j)}{\mathrm{Var}(S_n)} \approx \frac{k \cdot Aq}{An} \approx \frac{(n/p)Aq}{An} = \frac{q}{p} \to 0.$$

Moreover, the number of R_n is less than $p+q$,

$$\frac{\mathrm{Var}(R_n)}{An} \to 0.$$

By the Slutsky's Theorem, the limiting distribution of $S_n/\sqrt{An}$ depends only on the $(\sum U_j)/\sqrt{An}$ $\sum V_j \to 0, R_n \to 0$, which are the sum of the independent terms.

Assume that the following conditions hold for the two r.v.s X_n and Y_n:

$$X_n \xrightarrow{d} X \quad \text{and} \quad Y_n \xrightarrow{p} c.$$

Then, these conditions apply.

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} cX$
3. $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$

Now we check Lyapunov's condition to show that the CLT holds for the independent U_j.

$$\lim_{n\to\infty} \frac{1}{s_n^3} \sum_{i=1}^{n} E[|X_i - \mu_i|^3] = 0$$

Since $E|X_i|^3$ is uniformly bounded,

$$\frac{\sum_{i=1}^{k} E|U_j|^3}{s_k^3} = \frac{\sum_{j=1}^{k} E|U_j|^3}{(\mathrm{Var}(\sum U_j))^{3/2}} \leq \frac{k \cdot C \cdot p^{3/2}}{(Akp)^{3/2}} = \frac{k \cdot C}{(Ak)^{3/2}} = \frac{C}{A^{3/2} \cdot \sqrt{k}} \approx \frac{1}{\sqrt{k}} \to 0$$

where C is a proportional constant of the third moment of the X_i. Therefore standardized $\sum U_j$ converges to the standard normal distribution because this condition holds for the given U_i.

Finally, First condition of the Slutsky's Theorem

$$\frac{S_n}{\sqrt{An}} = \underbrace{\frac{\sum U_j}{\sqrt{An}}}_{X_n} + \underbrace{\frac{\sum V_j + R_n}{\sqrt{An}}}_{Y_n} \xrightarrow{d} N(0,1) + 0$$

and then

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{\sum_{j=1}^{n} (X_j - E(X_j))}{\sqrt{An}} \leq x \right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

for all $x \in \mathbb{R}$. □

576

[CLT for q-dependent random variables] If the random variables $X_1, X_2, \ldots$ in par. 575 are **stationary**, the necessary conditions simplify considerably: It suffices that $E(|X_i|^3)$ and

$$A = \sigma^2 := \mathrm{Var}(X_1) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_1, X_{1+j}). \qquad (4.52)$$

exist.

Regarding eq.(4.50)

$$A_i = \mathrm{Var}(X_{i+q}) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_{i+q-j}, X_{i+q}),$$

the stationary property provides the following conditions.

$$\mathrm{Var}(X_{i+q}) = \mathrm{Var}(X_1), \qquad \mathrm{Cov}(X_{i+q-j}, X_{i+q}) = \mathrm{Cov}(X_1, X_{1+j})$$

Therefore, A_i is constant as follows for all $i \in \mathbb{N}$,

$$A_i = \mathrm{Var}(X_1) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_1, X_{1+j})$$

and we then redefined it as A or σ^2, that is,

$$A = \sigma^2 := \mathrm{Var}(X_1) + 2\sum_{j=1}^{q} \mathrm{Cov}(X_1, X_{1+j}).$$

In particular, for $\mu := E(X_1)$, we obtain the equivalent to (4.51)

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{\sum_{j=1}^{n}(X_j - \mu)}{\sigma\sqrt{n}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt \quad \text{for all } x \in \mathbb{R}. \qquad (4.53)$$

Eq.(4.51) is

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{\sum_{j=1}^{n}(X_j - E(X_j))}{\sqrt{An}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

$E[X_j] = \mu$ is trivial for the stationary r.v.s and we obtain $A = \sigma^2$ in eq.(4.52). Therefore,

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{\sum_{j=1}^{n}(X_j - \mu)}{\sigma\sqrt{n}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt.$$

577

[CLT, dependent random variables] The CLT may even hold if $X_1, X_2, \ldots$ has no finite memory, provided that the dependencies decrease sufficiently fast. If the sequence $X_1, X_2, \ldots$ is stationary and, e.g., strongly mixing, the CLT holds if some further conditions are fulfilled. If needed, the reader is referred to [Jone04], Section 4, for details.