Mathematische Methoden der Unsicherheitsquantifizierung

Oliver Ernst

Professur Numerische Mathematik

Sommersemester 2016

Mathematik!
TU Chemnitz

Contents

Contents

Contents

We denote an abstract probability space by $(\Omega, \mathfrak{A}, \mathbf{P})$, in which

- Ω is an abstract set of elementary events,
- $\mathfrak{A}$ is a σ-algebra of subsets of Ω containing the measurable events and
- $\mathbf{P}$ is a probability measure on $\mathfrak{A}$.

Definition A.1

A measure $\mathbf{P}$ on a measurable space $(\Omega, \mathfrak{A})$ is called a probability measure if $\mathbf{P}(\Omega) = 1$.

Definition A.2

An event $A \in \mathfrak{A}$ is said to occur almost surely with respect to the measure $\mathbf{P}$ ($\mathbf{P}$-a.s.) if $\mathbf{P}(A) = 1$.

Proposition A.3 (Boole's inequality)

For events $\{A_n\}_{n\in\mathbb{N}}$ there holds

$$\mathbf{P}\left(\cup_{n=1}^{\infty} A_n\right) \leqslant \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

Definition A.4

The set of all $\omega \in \Omega$ such that $\omega \in A_n$ for infinitely many values of n is defined as

$$\{A_n, \text{ i.o. }\} := \limsup_{n\in\mathbb{N}} A_n := \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n$$

Theorem A.5 (Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$, then $\mathbf{P}\{A_n, \text{i.o.}\} = 0$. For independent events $\{A_n\}_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ there holds $\mathbf{P}\{A_n, \text{i.o.}\} = 1$.

Definition A.6

Let $(\Omega, \mathfrak{A}, \mathbf{P})$ be a probability space and $(E, \mathfrak{E})$ a measurable space. A measurable function $X : \Omega \to E$ is called an (E-valued) random variable. Individual values $X(\omega)$ for $\omega \in \Omega$ are called realisations of the random variable.

Remark: If E is a topological space then the σ-algebra generated by the open subsets of E is called the Borel σ-algebra $\mathfrak{B}(E)$.

Definition A.7

Let X be an E-valued random variable where $(E, \mathfrak{E})$ is a measurable space and $(\Omega, \mathfrak{A}, \mathbf{P})$ is the underlying probability space. The probability distribution $\mathbf{P}_X$ of X (also called the law of X) is the probability measure on $(E, \mathfrak{E})$ defined by $\mathbf{P}_X(A) := \mathbf{P}(X^{-1}(A))$ for pre-images $X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\}$ of sets $A \in \mathfrak{E}$.

Remark: This construction is sometimes called the push-forward measure defined by $(\Omega, \mathfrak{A}, \mathbf{P})$, $(E, \mathfrak{E})$ and X.

Theorem A.8 (Doob-Dynkin lemma)

Let $f : \Omega \to E$ and $g : \Omega \to F$ be two measurable functions from a measurable space $(\Omega, \mathfrak{A})$ to two measurable spaces $(E, \mathfrak{E})$ and $(F, \mathfrak{F})$ of which the first is a separable and complete metric space. Then f is g-measurable if and only if there exists some measurable mapping $h : F \to E$ with $f = h \circ g$.

See [Kallenberg, 1997], Lemma 1.13 for a proof.

Definition A.9

The expectation of a Banach space-valued random variable X is defined as the integral

$$\mathbf{E}[X] := \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega).$$

Definition A.10

The k-th moment ($k \in \mathbb{N}$) of a real-valued random variable X is $\mathbf{E}\left[X^k\right]$.
The first moment $\mu := \mathbf{E}[X]$ is also called the mean or mean value.
The central moments $\mathbf{E}\left[(X-\mu)^k\right]$ measure the deviation of X from its mean.
The second central moment

$$\mathbf{Var}\,X := \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\left[X^2\right] - \mu^2$$

of a random variable X is called its variance.

Remark: The quantity $\sigma := \sqrt{\mathbf{Var}\,X}$ is called the standard deviation of X.

Moments of a random variable are sometimes more easily computed by integrating over the image variable.

Consider a real-valued random variable X from $(\Omega, \mathfrak{A})$ to $(\Gamma, \mathfrak{B}(\Gamma))$ where $\Gamma \subset \mathbb{R}$. For $B \in \mathfrak{B}(\Gamma)$, set $A := X^{-1}(B)$. Then by the definition of the probability distribution $\mathbf{P}_X$

$$\int_\Omega \mathbb{1}_A(\omega) \, \mathrm{d}\mathbf{P}(\omega) = \mathbf{P}(A) = \mathbf{P}_X(B) = \int_\Gamma \mathbb{1}_B(x) \, \mathrm{d}\mathbf{P}_X(x).$$

For measurable functions $f : \Gamma \to \mathbb{R}$ we have

$$\int_\Omega f(X(\omega)) \, \mathrm{d}\mathbf{P}(\omega) = \int_\Gamma f(x) \, \mathrm{d}\mathbf{P}_X(x)$$

and, in particular,

$$\mathbf{E}[X] = \int_\Omega X(\omega) \, \mathrm{d}\mathbf{P}(\omega) = \int_\Gamma x \, \mathrm{d}\mathbf{P}_X(x).$$

Definition A.11

Let $\mathbf{P}$ be a probability measure on $(\Gamma, \mathfrak{B}(\Gamma))$ for some $\Gamma \subset \mathbb{R}$. If there exists a function $p : \Gamma \to [0, \infty)$ such that $\mathbf{P}(B) = \int_B p(x)\,\mathrm{d}x$ for any $B \in \mathfrak{B}(\Gamma)$ we say that $\mathbf{P}$ has a density p with respect to Lebesgue measure and we call p its probability density function (pdf). If X is a Γ-valued random variable on $(\Omega, \mathfrak{A}, \mathbf{P})$, the pdf p_X of X (if it exists) is the pdf of the probability distribution $\mathbf{P}_X$.

For real-valued random variables X from $(\Omega, \mathfrak{A}, \mathbf{P})$ to $(\Gamma, \mathfrak{B}(\Gamma))$ we then have[3]

$$\mathbf{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_\Gamma x\,\mathrm{d}\mathbf{P}_X(x) = \int_\Gamma x p(x)\,\mathrm{d}x. \tag{A.1}$$

Event probabilities are then easily calculated as

$$\mathbf{P}(X \in (a,b)) = \mathbf{P}(\{\omega \in \Omega : a < X(\omega) < b\}) = \mathbf{P}_X((a,b)) = \int_a^b p(x)\,\mathrm{d}x.$$

[3](where we have omitted the subscript X)

A random variable X is uniformly distributed on $D = [a, b] \subset \mathbb{R}$, $(a < b)$, denoted

$$X \sim \mathrm{U}(a, b),$$

if its pdf is

$$p(x) = \frac{1}{b - a}, \qquad x \in [a, b].$$

Using (A.1), we easily obtain

$$\mathbf{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{a+b}{2}, \qquad \mathbf{E}\left[X^2\right] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{b^3 - a^3}{3(b-a)},$$

so that $\mathbf{Var}\, X = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \frac{(b-a)^2}{12}$.

A random variable X is said to follow the Gaussian or normal distribution on $\Gamma = \mathbb{R}$ if its pdf is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R},$$

with two real parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, denoted $X \sim \mathsf{N}(\mu, \sigma^2)$.
As is easily verified,

$$\mathbf{E}[X] = \mu, \qquad \mathbf{Var}\, X = \sigma^2.$$

The probability that X is within α of its mean is given by

$$\mathbf{P}(|X - \mu| \leqslant \alpha) = \operatorname{erf}\left(\frac{\alpha}{\sqrt{2\sigma^2}}\right),$$

with the error function erf defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t.$$

The cumulative distribution function (cdf) of the standard normal distribution $\mathsf{N}(0,1)$ is denoted by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

Any (finite) linear combination of (jointly) random variables is normally distributed.

Lemma A.12 (Change of variables)

Suppose $Y : \Omega \to \mathbb{R}$ is a real-valued random variable and $f : (a, b) \to \mathbb{R}$ is continuously differentiable with inverse function f^{-1}. If p_Y is the pdf of Y, the pdf of the random variable $X : \Omega \to (a, b)$ defined via $X = f^{-1}(Y)$ is

$$p_X(x) = p_Y(f(x))\,|f'(x)| \quad \text{for } a < x < b.$$

If $Y \sim \mathsf{N}(\mu, \sigma^2)$, then the random variable

$$X := \exp(Y)$$

is said to follow a lognormal distribution. With $f(x) = \log x$, Lemma A.12 yields the pdf of X as

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2 x^2}} \exp\left(-\frac{[\log(x) - \mu]^2}{2\sigma^2}\right).$$

Moreover, there holds

$$\mathbf{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \qquad \mathbf{Var}\, X = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}.$$

Definition A.13

The covariance between two real-valued random variables is defined as

$$\mathbf{Cov}(X, Y) = \mathbf{E}\left[(X - \mu_X)(Y - \mu_Y)\right],$$

where $\mu_X := \mathbf{E}[X]$ and $\mu_Y := \mathbf{E}[Y]$. In particular, $\mathbf{Cov}(X, X) = \mathbf{Var}\, X$.

Note: An equivalent expression is $\mathbf{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$.

Calculation of the covariance requires evaluating the integral

$$\mathbf{E}[XY] = \int_\Omega X(\omega) Y(\omega)\, \mathrm{d}\mathbf{P}(\omega) = \int_{X(\Omega)\times Y(\Omega)} xy\, \mathrm{d}\mathbf{P}_{X,Y}(x, y),$$

in which $\mathbf{P}_{X,Y}$ is the joint probability distribution of X and Y.
Sometimes it is useful to scale the covariance to lie in $[-1, 1]$. The resulting quantity is known as the correlation coefficient

$$\rho(X, Y) := \frac{\mathbf{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Definition A.14

The joint probability distribution of two random variables X and Y is the distribution of the bivariate random variable $\boldsymbol{X} = (X, Y)$, i.e., for all $B \in \mathfrak{B}(X(\Omega) \times Y(\Omega))$

$$\mathbf{P}_{X,Y}(B) = \mathbf{P}(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}).$$

If it exists, the density $p_{X,Y}$ of $\mathbf{P}_{X,Y}$ is known as the joint pdf and

$$\mathbf{P}_{X,Y} = \int_B p_{X,Y}(x, y) \,\mathrm{d}x \,\mathrm{d}y.$$

Definition A.15

If $\mathbf{Cov}(X, Y) = 0$ the random variables X and Y are said to be uncorrelated. A family $\{X_\alpha\}_\alpha$ is said to be pairwise uncorrelated if X_α and X_β are uncorrelated for all $\alpha \neq \beta$.

Note: Uncorrelated random variables may still be strongly related. As an example,

$$X \sim \mathrm{N}(0, 1), \qquad \text{and} \qquad Y := \cos X$$

satisfy $\mu_X = 0$ and hence

$$\begin{aligned}\mathbf{Cov}(X, Y) = \mathbf{E}\left[X \cos X\right] &= \int_{\mathbb{R}} x \cos(x) \, \mathrm{d}\mathbf{P}_X(x) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x \cos(x) \exp\left(\frac{-x^2}{2}\right) \, \mathrm{d}x = 0.\end{aligned}$$

A stronger notion is that of independent random variables.

Definition A.16

A σ-algebra $\mathfrak{B}$ is a sub σ-algebra of $\mathfrak{A}$ if $\mathfrak{B} \subset \mathfrak{A}$, i.e., if $A \in \mathfrak{B}$ implies $A \in \mathfrak{A}$.

Definition A.17

Let X be an E-valued random variable on $(\Omega, \mathfrak{A}, \mathbf{P})$ for a measurable space $(E, \mathfrak{E})$. The σ-algebra generated by X, denoted $\sigma(X)$, is defined as

$$\sigma(X) := \{X^{-1}(A) : A \in \mathfrak{E}\} \subset \mathfrak{A}.$$

Remark: $\sigma(X)$ is the smallest σ-algebra such that X is measurable. It may be considerably smaller than $\mathfrak{A}$.

Definition A.18

Two events $A, B \in \mathfrak{A}$ are independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.
Two σ-algebras $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are independent if all pairs of events A_1 and A_2 with $A_1 \in \mathfrak{A}_1$ and $A_2 \in \mathfrak{A}_2$ are independent.

Definition A.19

Two random variables X, Y on a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$ are said to be independent if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.
A family $\{X_\alpha\}_\alpha$ of random variables is said to be pairwise independent if X_α and X_β are independent for all $\alpha \neq \beta$.

Independence of random variables X and Y can be conveniently determined using their joint distribution $\mathbf{P}_{X,Y}$: X and Y are independent if and only if $\mathbf{P}_{X,Y}$ equals the product measure $\mathbf{P}_X \times \mathbf{P}_Y$. If X and Y are real-valued with densities p_X and p_Y, they are independent if and only if their joint pdf is

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Lemma A.20

If X and Y are independent real-valued random variables and $\mathbf{E}[|X|], \mathbf{E}[|Y|] < \infty$, then X and Y are uncorrelated.

Note: The converse is generally false.

Theorem A.21 (Jensen's inequality)

If X is a real-valued random variable with $\mathbf{E}[|X|] < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function, then

$$\phi(\mathbf{E}[X]) \leqslant \mathbf{E}[\phi(X)]. \tag{A.2}$$

Definition A.22

Let $(\Omega, \mathfrak{A}, \mathbf{P})$ be a probability space and let W be a separable Banach space with norm $\|\cdot\|$. We denote by $L^p(\Omega; W)$, $1 \leqslant p < \infty$, the space of W-valued $\mathfrak{A}$-measurable random variables $X : \Omega \to W$ with $\mathbf{E}\left[\|X\|^p\right] < \infty$. The resulting space is a Banach space with the norm

$$\|X\|_{L^p(\Omega;W)} := \left(\int_\Omega \|X(\omega)\|^p \, \mathrm{d}\mathbf{P}(\omega)\right)^{1/p} = \mathbf{E}\left[\|X\|^p\right]^{1/p}.$$

Similarly, $L^\infty(\Omega; W)$ is the Banach space of W-valued random variables $X : \Omega \to W$ for which

$$\|X\|_{L^\infty(\Omega;W)} = \operatorname*{ess\,sup}_{\omega\in\Omega} \|X(\omega)\| < \infty.$$

The case $p = 2$ when W is a Hilbert space $W = H$ with inner product $(\cdot, \cdot)$ occurs frequently. In this case $L^2(\Omega; H)$ is a Hilbert space with inner product

$$(X, Y)_{L^2(\Omega;H)} := \mathbf{E}\left[(X, Y)\right] = \int_\Omega (X(\omega), Y(\omega)) \, \mathrm{d}\mathbf{P}(\omega).$$

Random variables in $L^2(\Omega; H)$ are called mean-square integrable random variables.

For random variables $X, Y \in L^2(\Omega; H)$ the Cauchy-Schwarz inequality takes on the form

$$|(X, Y)_{L^2(\Omega;H)}| \leqslant \|X\|_{L^2(\Omega;H)} \|Y\|_{L^2(\Omega;H)}$$

or

$$\mathbf{E}\left[(X, Y)\right] \leqslant \mathbf{E}\left[\|X\|^2\right]^{1/2} \mathbf{E}\left[\|Y\|^2\right]^{1/2}.$$

Definition A.23

Let H be a separable Hilbert space. A linear operator $C : H \to H$ is the covariance of two H-valued random variables X and Y if

$$(C\phi, \psi) = \mathbf{Cov}((\phi, X), (\psi, Y)) \qquad \forall \phi, \psi \in H.$$

X and Y are said to be uncorrelated if C is the zero operator. If $Y = X$ then C is called the covariance of X.

More generally, the covariance of two random variables X and Y with values in a separable Banach space W may be defined as a bilinear map $c : W' \times W' \to \mathbb{R}$ on the dual space W' of W such that

$$c(\phi, \psi) = \mathbf{Cov}(\langle \phi, X \rangle_{W' \times W}, \langle \psi, Y \rangle_{W' \times W}) \qquad \forall \phi, \psi \in W'.$$

Here $\langle \cdot, \cdot \rangle_{W' \times W}$ denotes the duality bracket between W' and W. The bilinear map c may be identified with a linear operator from $C : W' \to W''$ via the identity

$$\langle C\phi, \psi \rangle_{W'' \times W'} = c(\phi, \psi).$$

Definition A.24

Let W be a Banach space with norm $\|\cdot\|$ and $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of W-valued random variables. We say X_n converges to $X \in W$

almost surely if $X_n(\omega) \to X(\omega)$ for almost all $\omega \in \Omega$, i.e., if

$$\mathbf{P}\left(\|X_n - X\| \to 0 \text{ for } n \to \infty\right) = 1.$$

in probability if $\mathbf{P}\left(\|X_n - X\| > \epsilon\right) \to 0$ for $n \to \infty$ for any $\epsilon > 0$.

in p-th mean or in $L^p(\Omega; W)$ if $\mathbf{E}\left[\|X_n - X\|^p\right] \to 0$ as $n \to \infty$. When $p = 2$ this is known as convergence in mean square.

in distribution if $\mathbf{E}\left[\phi(X_n)\right] \to \mathbf{E}\left[\phi(X)\right]$ as $n \to \infty$ for any bounded and continuous function $\phi : W \to \mathbb{R}$.

Theorem A.25

Let $X_k \to X$ in p-th mean and, for $r > 0$ and a constant $K = K(p)$, assume that

$$\|X_k - X\|_{L^p(\Omega;W)} := \mathbf{E}\left[\|X_k - X\|^p\right]^{1/p} \leqslant \frac{K(p)}{k^r}. \tag{A.3}$$

Then the following convergence properties apply:

(a) $X_k \to X$ in probability and, for any $\epsilon > 0$,

$$\mathbf{P}\left(\|X_k - X\| \geqslant k^{-r+\epsilon}\right) \leqslant \frac{K(p)^p}{k^{p\epsilon}}. \tag{A.4}$$

(b) $\mathbf{E}\left[\phi(X_k)\right] \to \mathbf{E}\left[\phi(X)\right]$ for all Lipschitz continuous functions on W and, if L denotes a Lipschitz constant of ϕ,

$$\left|\mathbf{E}\left[\phi(X_k)\right] - \mathbf{E}\left[\phi(X)\right]\right| \leqslant L\frac{K(p)}{k^r}.$$

(c) If (A.3) holds for all p sufficiently large, then $X_k \to X$ a.s. Furthermore, for each $\epsilon > 0$ there exists a nonnegative random variable K such that $\|X_k(\omega) - X(\omega)\| \leqslant K(\omega)k^{-r+\epsilon}$ for almost all ω.

Contents

Random variables $\boldsymbol{X} = (X_1, \dots, X_n)^\mathsf{T}$ from $(\Omega, \mathfrak{A}, \mathbf{P})$ to $(\Gamma, \mathfrak{B}(\Gamma))$ with $\Gamma \subset \mathbb{R}^n$ are known as random vectors or multivariate random variables (bivariate for $n = 2$).

Their expected value

$$\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{X}] = \int_\Omega \boldsymbol{X}(\omega)\,\mathrm{d}\mathbf{P}(\omega) = [\mathbf{E}[X_1], \dots, \mathbf{E}[X_n]]^\mathsf{T}$$

is a vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has a pdf p, then for $B \in \mathfrak{B}(\Gamma)$

$$\mathbf{P}(\boldsymbol{X} \in B) = \mathbf{P}(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}) = \mathbf{P}_{\boldsymbol{X}}(B) = \int_B p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

The components $\{X_j\}_{j=1}^n$ of $\boldsymbol{X}$ are (pairwise) independent if and only if $\mathbf{P}_{\boldsymbol{X}}$ is the product measure $\mathbf{P}_{X_1} \times \cdots \times \mathbf{P}_{X_n}$. In terms of the pdf, this is equivalent to

$$p(\boldsymbol{x}) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdots p_{X_n}(x_n).$$

A random vector $\boldsymbol{X} : \Omega \to \Gamma$ with values in a set $\Gamma \subset \mathbb{R}^n$ with finite Lebesgue measure $|\Gamma|$ follows a multivariate uniform distribution on Γ, denoted by

$$\boldsymbol{X} \sim \mathrm{U}(\Gamma)$$

if it has the pdf

$$p(\boldsymbol{x}) \equiv \frac{1}{|\Gamma|}, \qquad \boldsymbol{x} \in \Gamma.$$

Definition A.26

The covariance of two real-valued random vectors $\boldsymbol{X} = [X_1, \dots, X_m]^\mathsf{T}$ and $\boldsymbol{Y} = [Y_1, \dots, Y_n]^\mathsf{T}$ is given by the $m \times n$ matrix

$$\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbf{E}\left[(\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbf{E}[\boldsymbol{Y}])^\mathsf{T}\right].$$

$\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be uncorrelated if $\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{O}$ (the $m \times n$ zero matrix). The matrix $\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$ is called the covariance matrix of $\boldsymbol{X}$.

Proposition A.27

Let $\boldsymbol{X}$ be an $\mathbb{R}^n$-valued random variable with mean vector $\boldsymbol{\mu}$ and covariance matric $\boldsymbol{C}$. Then $\boldsymbol{C}$ ist symmetric positive semi-definite and its trace is given by $\mathbf{E}\left[\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2\right]$.

A random vector with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{C}$ is said to follow an n-variate Gaussian distribution if it has the pdf

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{C}}} \exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2} \right). \tag{A.5}$$

To cover the case that $\boldsymbol{C}$ is singular we introduce the characteristic function.

Definition A.28

The characteristic function of an $\mathbb{R}^n$-valued random vector $\boldsymbol{X}$ is $\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\top \boldsymbol{X})\right]$, for $\boldsymbol{\lambda} \in \mathbb{R}^n$. If $\boldsymbol{X}$ has the pdf p, then its characteristic function is

$$\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\top \boldsymbol{X})\right] = (2\pi)^{n/2}\hat{p}(-\boldsymbol{\lambda}),$$

where $\hat{p}$ is the Fourier transform of p. (The minus sign is a convention in probability theory.)

Proposition A.29

A random vector $\boldsymbol{X}$ has the density (A.5) for a given vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and symmetric positive definite matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ if and only if its characteristic function is

$$\mathbf{E}\left[\exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{X})\right] = \exp(i\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{C}\boldsymbol{\lambda}). \tag{A.6}$$

Definition A.30

An $\mathbb{R}^n$-valued random vector $\boldsymbol{X}$ follows a multivariate normal (or Gaussian) distribution, denoted

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{C}),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite, if its characteristic function is (A.6).

If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{C})$ is a multivariate normal random vector, then for any $\boldsymbol{a} \in \mathbb{R}^n$ the linear combination

$$Y = \boldsymbol{a}^\top \boldsymbol{X} = \sum_{k=1}^{n} a_k X_k$$

follows the normal distribution $Y \sim \mathrm{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{C} \boldsymbol{a})$.

Contents

Definition A.31

A sequence $\{X_j\}_{j\in\mathbb{N}}$ of random variables is said to be independent and identically distributed (i.i.d.) if they all follow the same probability distribution and, in addition, are pairwise independent.

The classical limit theorems of probability theory concern sums of iid random variables. For an iid sequence $\{X_j\}_{j\in\mathbb{N}}$, we introduce the notation

$$S_n := X_1 + \cdots + X_n, \qquad n \in \mathbb{N}.$$

Theorem A.32 (Chebyshev inequality)

A random variable X with finite mean μ and finite variance σ^2 satisfies

$$c^2 \mathbf{P}(|X-\mu| \geqslant c) \leqslant \sigma^2.$$

Theorem A.33 (WLLN)

Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of i.i.d. random variables on a given probability space $(\Omega, \mathfrak{A}, \mathbf{P})$ with mean μ and finite variance. Then

$$\frac{S_n}{n} \to \mu \quad \text{in probability, i.e.}$$

for ever fixed $\epsilon > 0$ there holds

$$\mathbf{P}\left(|S_n/n - \mu| > \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Theorem A.34 (SLLN)

Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of i.i.d. real-valued random variables on a given probability space $(\Omega, \mathfrak{A}, \mathbf{P})$. Then S_n/n has a finite limit if and only if $\mathbf{E}[|X_1|] < \infty$, in which case

$$\frac{S_n}{n} \to \mathbf{E}[X_1] \qquad \text{a.s.}$$

If $\mathbf{E}[|X_1|] = \infty$, then $\limsup_{n\to\infty} |S_n|/n \to \infty$ a.s.

Lemma A.35 (Kronecker's Lemma)

If the series $\sum_{k=1}^{\infty} x_k/k$ converges (not necessarily absolutely) for a sequence $\{x_k\}_{k\in\mathbb{N}}$ of real numbers, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = 0.$$

Lemma A.36

The sequence $\{X_k\}_{k\in\mathbb{N}}$ converges a.s. if and only if

$$\lim_{n\to\infty} \mathbf{P}\{\sup_{k\in\mathbb{N}} |X_{n+k} - X_n| > \epsilon\} = 0 \qquad \forall \epsilon > 0.$$

Theorem A.37 (Kolmogorov Inequality)

Let $X_1, \dots, X_n$ be independent real-valued random variables with $\mathbf{E}[X_j] = 0$ and $0 < \sigma_j^2 = \mathbf{Var}\, X_j < \infty$ for all j. Then for each $\epsilon > 0$

$$\mathbf{P}\left\{ \max_{1 \leqslant k \leqslant n} |S_k| > \epsilon \right\} \leqslant \frac{1}{\epsilon^2} \sum_{j=1}^{n} \sigma_j^2. \tag{A.7}$$

Conversely, if there exists c such that $\mathbf{P}\{|X_k| < \epsilon\} = 1$ for each k, then for each ϵ

$$\mathbf{P}\left\{ \max_{1 \leqslant k \leqslant n} |S_k| > \epsilon \right\} \geqslant 1 - \frac{(c+\epsilon)^2}{\sum_{j=1}^{n} \sigma_j^2}. \tag{A.8}$$

Theorem A.38

Let $\{X_k\}_{k \in \mathbb{N}}$ be independent real-valued random variables with $\mathbf{E}[X_k] = 0$ for all k. If

$$\sum_{k=1}^{\infty} \mathbf{E}\left[X_k^2\right] = \sum_{k=1}^{\infty} \mathbf{Var}\, X_k < \infty$$

then $\sum_{k=1}^{\infty} X_k$ converges a.s.

Definition A.39

For a real-valued random variable X and $c > 0$ we denote the truncation of X at c by

$$X^c := X\mathbb{1}_{\{|X|\leqslant c\}} = \begin{cases} X & \text{if } |X| \leqslant c, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem A.40 (Three-series theorem)

Let $\{X_k\}_{k\in\mathbb{N}}$ be independent. If, for some $c > 0$,

$$\sum_{k=1}^{\infty} \mathbf{P}\{|X_k| > c\} < \infty, \tag{A.9a}$$

$$\sum_{k=1}^{\infty} |\mathbf{E}[X_k^c]| < \infty, \tag{A.9b}$$

$$\sum_{k=1}^{\infty} \mathbf{Var}\, X_k^c < \infty, \tag{A.9c}$$

then $\sum_{k=1}^{\infty} X_k$ converges a.s.
Conversely, if $\sum_{k=1}^{\infty} X_k$ converges a.s., then (A.9a)–(A.9c) hold for every $c > 0$.

Let the sequence $\{X_j\}_{j\in\mathbb{N}}$ of real-valued random variables be independent, but not necessarily identically distributed. In addition, let $\mathbf{E}[X_j] = 0$ and $\mathbf{E}\left[X_j^2\right] < \infty$ for all j.

Besides $S_n = \sum_{j=1}^n X_j$, introduce the quantities

$$\sigma_j^2 := \mathbf{Var}\, X_j,$$
$$\Sigma_n^2 := \sum_{j=1}^n \sigma_j^2 = \mathbf{Var}\, S_n.$$

The central limit theorem (CLT) is the statement that

$$\lim_{n\to\infty} \frac{S_n}{\Sigma_n} = \lim_{n\to\infty} \frac{S_n - \mathbf{E}[S_n]}{\sqrt{\mathbf{Var}\, S_n}} \sim \mathrm{N}(0,1) \quad \text{in distribution.}$$

Definition A.41 (Lyapunov condition)

The sequence $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lyapunov condition if $\mathbf{E}\left[|X_k|^3\right] < \infty$ for each k and

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^{n} \mathbf{E}\left[|X_k|^3\right] = 0.$$

Theorem A.42 (CLT)

If $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lyapunov condition, then $S_n/\Sigma_n \to \mathrm{N}(0,1)$ in distribution.

Definition A.43 (Lindeberg condition)

The sequence $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lindeberg condition if for every $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^{n} \mathbf{E}\left[X_k^2 \cdot \mathbb{1}_{\{|X_k|>\epsilon\Sigma_n\}}\right] = 0.$$

Proposition A.44

The Lyapunov condition implies the Lindeberg condition.

Example A.45

(1) If $\mathbf{P}\{|X_k| \leqslant c\} = 1$ for some constant c and if $\Sigma_n^2 \to \infty$, then the Lindeberg condition is satisfied.

(2) If $\{X_k\}_{k\in\mathbb{N}}$ are i.i.d. with variance $\sigma^2 \in (0,\infty)$, then the Lindeberg condition is satisfied.

Theorem A.46 (Lindeberg-Feller CLT)

If $\{X_k\}_{k\in\mathbb{N}}$ satisfies the Lindeberg condition, then $S_n/\Sigma_n \to \mathrm{N}(0,1)$ in distribution.

Theorem A.47 (Berry, 1941; Esseen 1942)

Let $\{X_k\}_{k\in\mathbb{N}}$ be i.i.d. random variables with (common)

$$\mu := \mathbf{E}[X_1], \quad \sigma^2 := \mathbf{Var}\, X_1 > 0, \quad \rho := \mathbf{E}\left[|X_1 - \mu|^3\right] < \infty.$$

If F_n denotes the distribution function of $(S_n - n\mu)/(\sigma\sqrt{n})$ and Φ that of the standard normal distribution $\mathsf{N}(0,1)$, then, with a universal constant C,

$$\sup_{x\in\mathbb{R}} |\Phi(x) - F_n(x)| \leqslant C \cdot \frac{\rho}{\sigma^3\sqrt{n}}.$$

Note: the constant C is known to satisfy $0.4097 \leqslant C \leqslant 0.7056$ [Shevtsova, 2007].

Contents

Statistical Estimation

- Estimation theory is concerned with determining an unknown quantity θ associated with the probability distribution of a random variable X given n i.i.d. samples $\{X_k\}_{k=1}^n$ of X.
- Typical examples of such quantities θ are moments of X's distribution such as the mean and the variance. Another common situation is the estimation of one or more parameters which determine the distribution of X.
- An estimator for a scalar quantity θ is a function

$$\phi : \mathbb{R}^n \to \mathbb{R}, \qquad \hat{\theta} = \phi(X_1, \dots, X_n)$$

  mapping n i.i.d. realizations of X to the estimate $\hat{\theta}$ of θ.
- Note that, since each of the n random samples X_k are random variables, the same is true of

$$\hat{\theta} = \hat{\theta}(\omega) = \phi(X_1(\omega), \dots, X_n(\omega)).$$

  Once the samples have been drawn/realized, the estimate $\hat{\theta}$ is a real number.

Statistical Estimation

Sample average, unbiased estimator

- The sample average
$$\hat{\mu}_n := \frac{X_1 + \cdots + X_n}{n}$$
is an estimate for the mean $\mu = \mathbf{E}[X]$.
- Since the X_k are i.i.d., we conclude from the linearity of expectation that
$$\mathbf{E}[\hat{\mu}_n] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[X_k] = \frac{1}{n} \cdot n\mu = \mu.$$
- If $\mathbf{E}[|X|] < \infty$ the SLLN tells us that also $\hat{\mu}_n \to \mu = \mathbf{E}[X]$ a.s. as $n \to \infty$.
- Since $\mathbf{Var}\,\hat{\mu}_n = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathbf{Var}\,X$, we note that the variance $\hat{\mu}_n$ decreases like $1/n$ with growing sample size.

Definition A.48

An estimator for which $\mathbf{E}\left[\hat{\theta}\right] = \theta$ is called unbiased.

The sample variance

$$\hat{\sigma}_n^2 := \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \hat{\mu}_n)^2$$

is an unbiased estimator for $\sigma^2 = \mathbf{Var}\, X$.

In addition, there holds $\hat{\sigma}_n^2 \to \sigma^2$ a.s. as $n \to \infty$.

An estimator $\hat{\theta}$ is, in general, only close to the estimated quantity θ in a probabilistic sense, i.e., it will fluctuate around the true value θ from realization to realization.

For a probability distribution depending on a real-valued parameter θ, we denote by

$$\mathbf{P}(A \,|\, \theta)$$

the probability of event A if the true value of the parameter is θ.

Definition A.49

Given n i.i.d. random variables $\{X_k(\omega)\}_{k=1}^n$ and a number $\gamma \in [0, 1]$, a confidence interval of level γ for a quantity θ is determined by two functions $\tau_-, \tau^+ : \mathbb{R}^n \to \mathbb{R}$ such that, for all possible values of θ,

$$\mathbf{P}\left(\tau_-(X_1, \dots, X_n) \leqslant \theta \leqslant \tau_+(X_1, \dots, X_n) \,|\, \theta\right) = \gamma.$$

As an example, take the random variables

$$X_k = \mu + \epsilon_k, \qquad \mu \in \mathbb{R}, \quad \epsilon_k \sim \mathrm{N}(0,1) \text{ i.i.d.}, \quad k = 1, \dots, n.$$

Then $\mu = \mathbf{E}[X]$ and for the estimation error we obtain

$$\hat{\mu}_n - \mu = \frac{1}{n} \sum_{k=1}^{n} \epsilon_k \sim \mathrm{N}\left(0, \tfrac{1}{n}\right)$$

and therefore $\sqrt{n}(\hat{\mu}_n - \mu) \sim \mathrm{N}(0,1)$.

Given $\gamma \in [0,1]$ we choose $a \geqslant 0$ such that $\Phi(a) - \Phi(-a) = \gamma$ and obtain

$$\gamma = \mathbf{P}(-a \leqslant \sqrt{n}(\hat{\mu}_n - \mu) \leqslant a \,|\, \mu) = \mathbf{P}\left(\hat{\mu}_n - \frac{a}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu}_n + \frac{a}{\sqrt{n}} \,|\, \mu\right),$$

so that $\tau_\pm = \hat{\mu}_n \pm \frac{a}{\sqrt{n}}$ yield a confidence interval of level γ for μ.

Beyond the mathematical definition of confidence intervals, there is substantial variation as to their interpretation.

- In Frequentist statistics, the unknown parameter θ is a real number. If the experiment is performed many (independent) times, the γ-confidence interval would contain θ in a proportion of γ of the number of such trials. For a single experiment with observation vector $\boldsymbol{x}$, this is meant by $[\tau_-(\boldsymbol{x}), \tau_+(\boldsymbol{x})]$ being a γ-confidence interval.
- In Bayesian statistics, the parameter θ is a random variable with a given prior probability distribution. Here, a γ-confidence interval I is one for which the posterior density of θ conditioned on the observation integrated over I equals γ.

More on Frequentist vs. Bayesian confidence intervals can be found in the excellent book [Williams, 2001; Chapter 6].

An investigation of common fallacies in the (scientific) interpretation of confidence intervals is given in [Morey & al., 2016].