

NMAI059 Probability and Statistics
Lecture overview: Definitions, Theorems, Examples, Problems.
February 2, 2017

1. AXIOMS; PROBABILITY SPACE; EVENTS

Definition 1 (σ -algebra). Let Ω be a nonempty set. System \mathcal{F} of subsets of Ω is called σ -algebra if

- (1) $\emptyset \in \mathcal{F}, \Omega \in \mathcal{F}$.
- (2) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$.
- (3) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

For any Ω the set \emptyset, Ω is *trivial* σ -algebra. Another simple example is $\emptyset, A, \Omega \setminus A, \Omega$ for any $A \subset \Omega$. Also the system 2^Ω of *all* subsets of Ω is σ -algebra. However, this system is useful mostly for *finite or countable* Ω .

For any finite or countable system \mathbf{A} of subsets of Ω there exists unique σ -algebra $\mathcal{F}_{\mathbf{A}}$ which is the smallest σ -algebra containing \mathbf{A} .

Problem 1. Intersection of (countably many) σ -algebras is again σ -algebra. Union of σ -algebras need not to be σ -algebra. Find proof or examples.

Definition 2. Consider nonempty set Ω and some σ -algebra \mathcal{F} on it. Let $P : \mathcal{F} \rightarrow [0, 1]$ be mappings such that

- (1) $P(\Omega) = 1$
- (2) For any pairwise disjoint $A_1, A_2, \dots \in \mathcal{F}$ it holds

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Then P is called *probability measure* on \mathcal{F} . The triple (Ω, \mathcal{F}, P) is called *probability space*.

We say that probability measure is *countably additive*. If Ω is finite or countable then it is usually possible to use $\mathcal{F} = 2^\Omega$. For uncountable Ω it may be impossible to define probability measure on 2^Ω and strictly smaller σ -algebra is usually needed.

Theorem 1. Let (Ω, \mathcal{F}, P) be a probability space. Then

- (1) $P(\emptyset) = 0$.
- (2) For any $A \in \mathcal{F}$ it holds $P(\Omega \setminus A) = 1 - P(A)$.
- (3) For any $A, B \in \mathcal{F}$ it holds $A \subset B \Rightarrow P(A) \leq P(B)$ and $P(B \setminus A) = P(B) - P(A)$.

Definition 3 (Classical probability space). Let Ω be a nonempty finite set, $\mathcal{F} = 2^\Omega$ and for any $A \subset \Omega$ set

$$P(A) = \frac{|A|}{|\Omega|},$$

where $|\cdot|$ is the cardinality of set. (Ω, \mathcal{F}, P) is called *classical* or *elementary probability space*.

Each $\omega \in \Omega$ is called *elementary event*. Each $A \in \mathcal{F}$ is called *random event*.

Theorem 2 (Continuity of probability measure). Let P be a probability measure on (Ω, \mathcal{F}) .

- (1) If $\{A_i\} \subset \mathcal{F}$ is a sequence of random events such that $A_n \nearrow A$ then $\lim_{n \rightarrow \infty} P(A_n) = P(A)$

- (2) If $\{A_i\} \subset \mathcal{F}$ is a sequence of random events such that $A_n \searrow A$ then $\lim_{n \rightarrow \infty} P(A_n) = P(A)$
 (3) If $\{A_i\} \subset \mathcal{F}$ is a sequence of random events such that $A_n \searrow \emptyset$ then $\lim_{n \rightarrow \infty} P(A_n) = 0$

In fact, (1)–(3) are equivalent.

Problem 2. Let P be positive finitely additive function defined on (Ω, \mathcal{F}) . If (3) of Theorem 2 holds then P is probability measure (i.e., P is also countably measurable).

Problem 3. Show that classical probability space is probability space. Consider three dices. There are more possible classical probability spaces describing this random experiment.

Problem 4. Throw two dices and take sum of the results. Find **classical** probability space for the outcome sum.

Definition 4 (Discrete probability space). Let Ω be nonempty finite or countable set. Let $\mathcal{F} = 2^\Omega$ and consider for each $\omega \in \Omega$ value p_ω such that

$$\forall \omega p_\omega \geq 0, \text{ and } \sum_{\omega \in \Omega} p_\omega = 1.$$

Define

$$P(A) = \sum_{\omega \in A} p_\omega \text{ for any } A \in \mathcal{F}.$$

Then (Ω, \mathcal{F}, P) is *discrete probability space*.

Problem 5. Throw two dices and take sum of the results. Find **discrete** probability space for the outcome sum.

Definition 5 (Conditional probability I). Consider (Ω, \mathcal{F}, P) . Let $B \in \mathcal{F}$ be random event such that $P(B) > 0$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

is called *conditional probability of A given B*.

Problem 6. Show that $P(\cdot|B)$ is probability measure on Ω and \mathcal{F} . Conditional probability is probability of a random event given that the random event B occurred.

Definition 6 (Independence I). Random events A, B are *independent* if $P(A \cap B) = P(A)P(B)$.

Problem 7. Show that random events A and B are independent if

- (1) $P(A)$ or $P(B)$ is either 0 or 1.
- (2) $P(A) = P(A|B)$.

Show that disjoint events cannot be independent unless (1) holds.

Definition 7 (Independence II). Random events A_1, A_2, \dots are (mutually) independent if for any $n \in \mathbb{N}$ and any finite subset of indices $\Lambda = \{i_1, \dots, i_n\}$ it holds

$$P\left(\bigcap_{j=1}^n A_{i_j}\right) = \prod_{j=1}^n P(A_{i_j}).$$

Remark: Independence may be defined for *any* number of random event, even uncountably many random events, in the same way as in Definition 7.

Problem 8. Random events A_1, A_2, \dots are pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for any pair $i \neq j$. Show that if A_1, A_2, \dots are mutually independent then the random events are pairwise independent but the opposite implication is not true.

Theorem 3 (Gradual conditioning). Let E_1, E_2, \dots, E_n be random events such that $P(E_1 \cap \dots \cap E_{n-1}) > 0$. Then

$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_n | E_1 \cap \dots \cap E_{n-1})P(E_{n-1} | E_1 \cap \dots \cap E_{n-2}) \dots P(E_2 | E_1)P(E_1).$$

Theorem 4 (Inclusion and exclusion). Let A_1, A_2, \dots, A_n be random events. Then

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P\left(\bigcap_{i=1}^n A_i\right).$$

Definition 8 (Disjoint decomposition). System (finite or countable) of random events E_1, E_2, \dots is called disjoint decomposition of Ω if $P(E_i) > 0$ for all i , $E_i \cap E_j = \emptyset$ for any $i \neq j$ and $\Omega = \bigcup_i E_i$.

By \bigcup_i we mean union (sum, product, ...) over all indices i , both finite or countable.

Theorem 5 (Total probability). Let A_1, A_2, \dots be disjoint decomposition of Ω . Then for any random event B

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Theorem 6 (Bayes theorem). Let A_1, A_2, \dots be disjoint decomposition of Ω . Then for any random event B , $P(B) > 0$ and for any i

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}.$$

Theorems 2—4 are essential for classical probability problems.

Problem 9 (Pólya urn scheme). Consider an urn with balls of different colours. Each time we draw exactly one ball from the urn randomly, i.e. each ball has the same chance to be drawn (classical probability space). When we see the colour of the ball we return the ball to the urn together with Δ additional balls of the same colour. Special cases are

- $\Delta = -1$ corresponds to the case that the ball is not returned to the urn. We call this *sampling without replacement*.
- $\Delta = 0$ corresponds to the situation that only the drawn ball is replaced back to the urn. This is called *sampling with replacement*.

Consider classical urn scheme starting with n black balls and m white balls in the urn. Calculate the probability, that white ball is drawn in the *second* round if

- (1) $\Delta = 0$

$$(2) \Delta = -1$$

$$(3) \Delta = 2$$

Calculate this probability also for further rounds of sampling.

Theorem 7 (Bonferroni inequality). *Let A_1, \dots, A_n be random events. Then*

$$P\left(\bigcap_{i=1}^n A_i\right) \geq 1 - \sum_{i=1}^n (1 - P(A_i)).$$

2. DISKRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Definition 9 (Random variable). Let (Ω, \mathcal{F}, P) be a probability space. Function

$$X : \Omega \rightarrow \mathbb{R} \text{ such that } X^{-1}(-\infty, a] = \{\omega \in \Omega; X(\omega) \leq a\} \in \mathcal{F} \text{ for all } a \in \mathbb{R}$$

is called *random variable*.

If X takes at most countably many values we say that X is *discrete* random variable. In such case it is usually possible to consider only set \mathbb{Z} (whole numbers) as values of X .

Definition 10 (Distribution of r.v.). Probability measure P_X defined on \mathbb{R} by

$$P_X((-\infty, a]) = P(\{\omega; X(\omega) \leq a\})$$

is called *distribution of random variable X* .

Notation remark: For simpler notation we adopt the convention

$$[X \in A] = \{\omega \in \Omega; X(\omega) \in A\} \text{ for } A \in \mathcal{B}.$$

Here \mathcal{B} is the smallest σ -algebra (see Definition 1) of subsets of \mathbb{R} containing *all open and closed sets*. Hence, $(-\infty, a] \in \mathcal{B}$ for any $a \in \mathbb{R}$. Definition of the distribution may be written using this notation as

$$P_X(A) = P[X \in A].$$

Clearly, $[X \in A]$ is a *random event* for any $A \in \mathcal{B}$.

Definition 11. Let X be a random variable defined on probability space (Ω, \mathcal{F}, P) . The σ -algebra

$$\mathcal{F}_X = \{X^{-1}(A), A \in \mathcal{B}\}$$

is the σ -algebra of random events generated by the r.v. X

Problem 10. Show that \mathcal{F}_X is indeed a σ -algebra, and $\mathcal{F}_X \subset \mathcal{F}$.

If the random variable X is clear from the context we may write simply P for its distribution. We try to distinguish the probability measure P on probability space Ω and the distribution P as measure on \mathbb{R} .

Example 1. Consider fair six-sided dice. Throw the dice two times independently. Then we may choose the probability space as $\Omega = \{(i, j); i = 1, \dots, 6, j = 1, \dots, 6\}$, $\mathcal{F} = 2^\Omega$ and $P(i, j) = 1/36$ for all $(i, j) \in \Omega$. Random variable $X(i, j) = i + j$ describes the sum of the two results. The distribution P_X of X is a probability measure on set $\{2, 3, \dots, 12\}$.

Problem 11. Find the distribution P_X from the last Example.

Definition 12 (Distribution function and density I). Let P_X be a distribution of discrete random variable X (integer valued). Function

$$F_X(x) = P_X((-\infty, x]) = P[X \leq x]$$

is called *distribution function* of r.v. X . Function

$$p(x) = P_X(\{x\}) = P[X = x]$$

is called *density with respect to the arithmetic measure of discrete* r.v. X .

Note: Clearly, $p(x) = 0$ for any $x \notin \mathbb{N}$ if X is integer-valued (here $\mathbb{N} = \{0, 1, 2, \dots\}$). For $i = 0, 1, \dots$ we often denote $p_i = p(i)$ and the set of probabilities $\{p_i, i = 0, 1, \dots\}$ uniquely determines the distribution P_x and distribution function

$$F_X(x) = \sum_{i \leq x} p_i.$$

Both distribution function and density uniquely determine the distribution of random variable.

Bernoulli (alternative) distribution Let $p \in (0, 1)$. Distribution given by

$$p_0 = 1 - p, \quad p_1 = p$$

is called *alternative*, or *Bernoulli* distribution. Corresponding random variable is two-valued. $X = 1$ denotes *success* and $X = 0$ denotes failure in the experiment. Abbreviation is usually $\text{Alt}(p)$. Such experiment (with success or failure outcome with success probability p) is called *Bernoulli trial*.

Binomial distribution Let $n > 0$, $n \in \mathbb{N}$, and $p \in (0, 1)$. Distribution on the set $\{0, 1, \dots, n\}$ given by

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}$$

is called *binomial with parameters* n and p , abbreviated as $\text{Bi}(n, p)$. Binomial distribution is the distribution of *number of successes* in n independent Bernoulli trials $\text{Alt}(p)$.

Geometric distribution Let $p \in (0, 1)$. Distribution on the set $\{0, 1, \dots\}$ given by

$$p_i = p(1-p)^i$$

is called *geometric distribution* with parameter p . Geometric distribution is the distribution of *number of failures* preceding the first success in a series of independent Bernoulli trials. Abbreviation for this distribution is $\text{Geom}(p)$.

Negative binomial distribution Let $n > 0$, $n \in \mathbb{N}$, and $p \in (0, 1)$. Distribution on the set $\{0, 1, \dots\}$ given by

$$p_i = \binom{n+i-1}{i} p^n (1-p)^i$$

is called *negative binomial with parameters* n and p abbreviated as $\text{NBi}(n, p)$. It is the distribution of number of failures preceding the n -th success in a series of independent Bernoulli trials.

Hypergeometric distribution Let n, M, N be integers such that $0 < n < N$, $0 < M < N$. Distribution

$$p_m = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}}, \quad m = \max\{0, n+M-N, \dots, \min\{n, M\}\}$$

is called *hypergeometric distribution*. Consider two sets of object (white and black balls), M is the number of white balls, N is the total number of balls. Exactly n balls are drawn *without replacement* from the set of balls. Then p_m is the probability that exactly m white balls are drawn in the sample.

Poisson distribution Let $\lambda > 0$. Distribution on the set $\{0, 1, \dots\}$ given by

$$p_i = e^{-\lambda} \frac{\lambda^i}{i!}$$

is called *Poisson distribution*. Poisson distribution is a limit case of $\text{Bi}(n, \pi_n)$ distribution if $n\pi_n \rightarrow \lambda \in (0, \infty)$ as $n \rightarrow \infty$.

For arbitrary integer-valued (or diskrete) random variable X it is possible to define its *canonical* probability space

$$\Omega = \mathbb{N} \text{ (or } \mathbb{S}), \quad \mathcal{F} = 2^\Omega, \quad \mathbb{P} = P_X,$$

where \mathbb{S} is the set of all possible values of r.v. X . \mathbb{S} is often called *sample space*.

Definition 13 (Independence). Random variables X_1, X_2, \dots are mutually independent if for any $n \in \mathbb{N}$ and any finite subset of indices $\Lambda = \{i_1, \dots, i_n\}$ and any $x_1, \dots, x_n \in \mathbb{R}$ it holds

$$\mathbb{P} \left(\bigcap_{j=1}^n [X_{i_j} \leq x_j] \right) = \prod_{j=1}^n \mathbb{P}[X_{i_j} \leq x_j].$$

In other words, $X_i, i = 1, 2, \dots$ are mutually independent if and only if any random events A_1 generated by X_1, A_2 generated by X_2 etc. are mutually independent.

In particular, integer-valued random variables are independent iff for any $n \in \mathbb{N}$ and any finite subset of indices $\Lambda = \{i_1, \dots, i_n\}$ and any $x_1, \dots, x_n \in \mathbb{R}$ it holds

$$\mathbb{P} \left(\bigcap_{j=1}^n [X_{i_j} = x_j] \right) = \prod_{j=1}^n \mathbb{P}[X_{i_j} = x_j].$$

Theorem 8 (Properties of distribution function). *Let X be a random variable and F_X its distribution function. Then*

- (1) F_X is right continuous and non-decreasing.
- (2) $\lim_{x \rightarrow \infty} F_X(x) = 1$.
- (3) $\lim_{x \rightarrow -\infty} F_X(x) = 0$.

Theorem 9 (Characterisation of distribution function). *Let $F : \mathbb{R} \rightarrow \mathbb{R}$ be a function satisfying (1)–(3) of Theorem 8. Then there exists probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variable X defined on it such that F is the distribution function of X .*

3. EXPECTATION AND HIGHER MOMENTS

Distribution function fully describes the distribution of random variable and its random behaviour. However, simpler numerical characteristics are often needed.

Definition 14 (Mean value of random variable (general)). Let X be random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *mean value* (or *expectation*) of X is defined as

$$EX = \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$$

if the integral on the right hand side exists. If the mean value of X exists and it is finite, then we say that the random variable X has *finite expectation* (or *finite mean value*).

Notation: In what follows we fix (at most countable) discrete set \mathbb{S} to be the sample space, i.e. the set of all possible values of random variable X . If needed, we will use \mathbb{S}_X to specify the random variable.

Theorem 10 (Mean value of discrete random variable). *Let X be an \mathbb{S} -valued random variable with finite mean value and distribution P_X . Then*

$$EX = \sum_{s \in \mathbb{S}} sP[X = s] = \sum_{s \in \mathbb{S}} sP_X(s) = \sum_{s \in \mathbb{S}} sp_s.$$

The mean value (expectation) characterises the *location* of random variable. It is also called *first moment* of X . Theorem 10 gives simple computation rule for expectation of random variable with given distribution (or simply expectation of given distribution).

Problem 12. Calculate the expectation of random variables with distributions Bernoulli, binomial, geometric, negative binomial, hypergeometric and Poisson.

Remark: Let X be non-negative random variable, i.e. $P[X \geq 0] = 1$. Then its expectation always exists and may be either finite or $+\infty$.

If $E|X|$ (defined naturally as $\int_{\Omega} |X(\omega)|dP(\omega)$) is finite, then EX exists and is finite.

Problem 13. It is not difficult to find distribution P on \mathbb{Z} such that expectation of P does not exist. (*Hint: look for distribution such that both $\sum_{s=1}^{\infty} sp_s = \infty$ and $\sum_{s=-1}^{-\infty} sp_s = -\infty$ hold.*)

Definition 15 (Other moments). Let X be random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$ function such that $g(X)$ is again random variable. Then we define

$$Eg(X) = \int_{\Omega} g(X(\omega))dP(\omega)$$

if the integral exists (finite or infinite). If X is a discrete \mathbb{S} -valued random variable, then

$$Eg(X) = \sum_{s \in \mathbb{S}} g(s)P_X(s)$$

if the sum exists.

Definition 16 (Higher moments and moment generating function). Let X be a random variable. Then

- (1) EX^r is called *r-th moment* of X (if it exists).
- (2) $E(X - EX)^r$ is called *r-th central moment* of X (if it exists).
- (3) $\text{var}(X) := E(X - EX)^2$ is called *variance* (or dispersion) of X (if it exists).
- (4) $\mu_3(X) = E(X - EX)^3 / (\text{var}(X))^{3/2}$ is called *skewness* of X (if it exists).
- (5) $\psi_X(t) = Ee^{tX}$ is called *moment generating function* of X (needs not exist for all t).

Theorem 11 (Moments from moment generating function). *Let X be a random variable and ψ_X its moment generating function. If ψ_X does exist on an open neighbourhood of 0 then for any $r \in \mathbb{N}$*

$$EX^r = \frac{d^r \psi_X(t)}{dt^r}(0),$$

i.e. the r-th moment of X is the r-th derivative of ψ_X at 0.

Problem 14. Calculate the moments of known discrete distributions using the moment generating function.

Theorem 12 (MGF and characterisation of distribution). *Let X and Y be random variables and ψ_X and ψ_Y their moment generating functions. Assume that both ψ_X and ψ_Y exist and are finite on some open neighbourhood of zero, and that for some $\delta > 0$*

$$\psi_X(t) = \psi_Y(t) \text{ for all } |t| < \delta.$$

Then $F_X = F_Y$, i.e. the two random variables are identically distributed.

Theorem 13 (Jenssen's inequality). *Let X be a random variable and f be a convex function. If both EX and $Ef(X)$ exist then*

$$f(EX) \leq Ef(X)$$

Theorem 14 (Expectation of non-negative r.v.). *Let X be an \mathbb{N}_0 -valued random variable. Then*

$$EX = \sum_{n=0}^{\infty} (1 - F_X(n)).$$

4. DISCRETE RANDOM VECTORS AND THEIR DISTRIBUTIONS

Definition 17 (Random vector). Let Ω, \mathcal{F}, P be a probability space. Function

$\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ such that $\mathbf{X}^{-1} \left(\prod_{i=1}^d (-\infty, a_i] \right) = \{\omega \in \Omega; X_i(\omega) \leq a_i\} \in \mathcal{F}$ for all $\mathbf{a} = (a_1, \dots, a_d) \in \mathbb{R}^d$

is called *d-dimensional random vector*.

If \mathbf{X} takes at most countably many values we say that \mathbf{X} is *discrete* random vector (**d.r.v.**). In such case it is usually possible to consider only set \mathbb{N}_0^d (non-negative integers) as values of \mathbf{X} . **From now on we shall always consider d.r.v. \mathbf{X} to be \mathbb{N}_0^d -valued unless we specify another sample space!**

Definition 18 (Distribution of random vector). Probability measure $P_{\mathbf{X}}$ defined on \mathbb{R}^d by

$$P_{\mathbf{X}} \left(\prod_{i=1}^d (-\infty, a_i] \right) = P \left(\bigcap_{i=1}^d \{\omega; X_i(\omega) \leq a_i\} \right)$$

is called (*joint*) *distribution of random vector \mathbf{X}* .

Special case is again a discrete random vector.

Theorem 15 (Probabilities of d.r.v.). *Let \mathbf{X} be a d-dimensional d.r.v. with distribution $P_{\mathbf{X}}$. Then there exist non-negative function $p : \mathbb{N}_0^d \rightarrow [0, 1]$ such that*

$$P_{\mathbf{X}} \left(\prod_{i=1}^d (-\infty, a_i] \right) = \sum_{\mathbf{z} \leq \mathbf{a}} p(\mathbf{z}),$$

where $\mathbf{z} \leq \mathbf{a}$ if $z_i \leq a_i, i = 1, \dots, d$. It holds $P[\mathbf{X} = \mathbf{z}] = p(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{N}_0^d$.

Therefore $p(\mathbf{z})$ fully characterise the distribution $P_{\mathbf{X}}$ and we also may call the set $\{p(\mathbf{z}), \mathbf{z} \in \mathbb{N}_0^d\}$ the distribution of \mathbf{X} . Sometimes we call it also *density of d.r.v. with respect to arithmetic measure* (see Definition 12).

Also for random vectors we may define its distribution function which fully characterises the distribution.

Definition 19 (Distribution function of random vector). Let $P_{\mathbf{X}}$ be a distribution of random vector \mathbf{X} . Function $F_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, 1]$ defined as

$$F_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}} \left(\prod_{i=1}^d (-\infty, x_i] \right) = \mathbb{P}[\mathbf{X} \leq \mathbf{x}]$$

is called (*joint*) *distribution function* of d.r.v. \mathbf{X} .

Definition 20 (Marginal distribution). Let \mathbf{X} be random vector with distribution $P_{\mathbf{X}}$ and distribution function $F_{\mathbf{X}}$. Then

$$P_{X_i}(-\infty, a] = \lim_{a_j \rightarrow \infty, j \neq i} P_{\mathbf{X}} \left(\prod_{j=1}^d (-\infty, a_j] \right)$$

is the *marginal distribution* of X_i , and

$$F_{X_i}(x) = \lim_{x_j \rightarrow \infty, j \neq i} F_{\mathbf{X}}(\mathbf{x})$$

is called *marginal distribution function* of X_i .

Clearly, $\mathbf{X} = (X_1, \dots, X_d)$, and X_i is a discrete random variable and P_{x_i} is its distribution, and F_{X_i} its distribution function. In a similar way it is possible to define random *subvector* of \mathbf{X} and its distribution and distribution function.

Theorem 16 (Marginal distribution determined). *Let \mathbf{X} be d.r.v. with distribution $P_{\mathbf{X}}$ then all its marginal distributions P_{X_i} , $i = 1, \dots, d$ are uniquely determined by $P_{\mathbf{X}}$ (and the same holds for the distribution function and “density”).*

Reverse implication does not hold at all!

Problem 15 (Infinitely many possibilities). Consider two discrete marginal distribution $p_1(n)$ and $p_2(n)$, $n \in \mathbb{N}_0$. Find **at least** two d.r.v. (X_1, X_2) and (Y_1, Y_2) such that the marginal distributions of both d.r.v are p_1 and p_2 , but the joint distributions differ!

Notation: $\mathbf{a} < \mathbf{b} \in \mathbb{R}^d$ iff $a_i < b_i$ for all $i = 1, \dots, d$. For $1 \leq k \leq d$ denote $\Delta_k(\mathbf{a}, \mathbf{b})$ the set of all $\mathbf{c} \in \mathbb{R}^d$ such that there is exist $1 \leq i_1 < \dots < i_k \leq d$ and

$$c_i = \begin{cases} b_i & \text{if } i = i_j \text{ for some } j = 1, \dots, k \\ a_i & \text{otherwise.} \end{cases}$$

Clearly $\Delta_0(\mathbf{a}, \mathbf{b}) = \mathbf{a}$. In other words $\Delta_k(\mathbf{a}, \mathbf{b})$ denotes those vertices of a d -dimensional cube $[\mathbf{a}, \mathbf{b}]$ for which exactly k coordinates belong to \mathbf{b} .

Theorem 17 (Properties of distribution function). *Let $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector and $F_{\mathbf{X}}$ its distribution function. Then*

- (1) For any $i \in \{1, \dots, d\}$ it holds $\lim_{x_i \rightarrow -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0$.
- (2) $\lim_{\forall i \ x_i \rightarrow \infty} F_{\mathbf{X}}(\mathbf{x}) = 1$.
- (3) $F_{\mathbf{X}}$ is nondecreasing and right continuous in all variables.
- (4) For any $\mathbf{a} < \mathbf{b}$ it holds

$$\sum_{k=0}^d (-1)^k \sum_{\mathbf{c} \in \Delta_k(\mathbf{a}, \mathbf{b})} F_{\mathbf{X}}(\mathbf{c}) \geq 0.$$

Theorem 18 (Characterisation of distribution function). *Let $F : \mathbb{R}^d \rightarrow [0, 1]$ satisfy (1)–(4) of Theorem 17. Then there exist probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random vector \mathbf{X} such that F is the distribution function of \mathbf{X} .*

Condition (4) cannot be removed since condition (3) itself is not sufficient to assure that any nondegenerated d -dimensional cube $[\mathbf{a}, \mathbf{b}]$ has non-negative probability.

Theorem 19 (Characterisation of independence). *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector. The random variables X_1, \dots, X_n are independent iff for any $\mathbf{x} = (x_1, \dots, x_n)$*

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n F_{X_i}(x_i).$$

If \mathbf{X} is a discrete random vector then this is equivalent to

$$p_{\mathbf{X}}(\mathbf{z}) = \prod_{i=1}^n p_{X_i}(z_i) \quad \forall \mathbf{z} = (z_1, \dots, z_n) \in \mathbb{N}_0^n.$$

Note: The joint distribution is crucial! It fully describes the relation between the elements of the random vector, their dependence (or independence), the form of the dependence. This important information is not available from the marginal distributions.

From this point of view a random vector \mathbf{X} is *more than just mere set of random variables* X_1, X_2, \dots, X_d . It is also the model describing the mutual relations between the random variables X_1, \dots, X_d (and also subvectors of \mathbf{X}).

5. (DISCRETE) RANDOM VECTORS: MOMENTS AND TRANSFORMATIONS

Definition 21 (Expectation of random vector). Let \mathbf{X} be a d.r.v. and $g : \mathbb{R}^d \rightarrow \mathbb{R}$. Then

- (1) $E\mathbf{X} = (EX_1, EX_2, \dots, EX_d)$ if all expectations on the right hand side exist.
- (2) $Eg(\mathbf{X}) = \sum_{\mathbf{z} \in \mathbb{Z}^d} g(\mathbf{z})p_{\mathbf{X}}(\mathbf{z})$ if the sum exists.

If \mathbf{X} is a *discrete* random vector then the expectation may be calculated as

$$Eg(\mathbf{X}) = \sum_{\mathbf{z} \in \mathbb{N}^d} g(\mathbf{z})P[\mathbf{X} = \mathbf{z}] = \sum_{\mathbf{z} \in \mathbb{N}^d} g(\mathbf{z})p(\mathbf{z})$$

Theorem 20 (Linearity of expectation). *Let \mathbf{X} be a d.r.v. with finite $E\mathbf{X}$, $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^d$. Then*

$$Ea + \sum_{i=1}^d b_i X_i = a + \sum_{i=1}^d b_i EX_i.$$

Theorem 21 (Independence and expectation). *Let X_i , $i = 1, \dots, d$ be independent random variables and let EX_i exists finite for all i . Then*

$$E \prod_{i=1}^d X_i = \prod_{i=1}^d EX_i.$$

Definition 22. Let X and Y be random variables (defined on the same probability space) such that both $EX^2 < \infty$ and $EY^2 < \infty$. Then define their *covariance* as

$$\text{cov}(X, Y) = E(X - EX)(Y - EY).$$

Theorem 22 (Variance of sum). *Let X_i , $i = 1, \dots, d$ be random variables with finite second moment. Then*

$$\begin{aligned} \text{var} \left(\sum_{i=1}^d X_i \right) &= \sum_{i=1}^d \sum_{j=1}^d \text{cov}(X_i, X_j) = \sum_{i=1}^d \text{var}(X_i) + \sum_{\substack{i,j \\ i \neq j}} \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^d \text{var}(X_i) + 2 \sum_{1 \leq i < j \leq d} \text{cov}(X_i, X_j). \end{aligned}$$

Definition 23 (Variance and correlation matrix). Let \mathbf{X} be a d -dimensional random vector such that $\text{E}X_i^2 < \infty$ for all $i = 1, \dots, d$. Denote

$$\text{cov}(X_i, X_j) = \text{E}((X_i - \text{E}X_i)(X_j - \text{E}X_j))$$

the *covariance* of random variables X_i and X_j . Denote

$$\varrho_{i,j} = \text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var } X_i \text{ var } X_j}}$$

the *correlation* of random variables X_i and X_j . The matrix

$$\text{Var } \mathbf{X} = \{\text{cov}(X_i, X_j)\}_{i,j=1}^d$$

is called the *variance matrix* of random vector \mathbf{X} and the matrix

$$\text{Corr } \mathbf{X} = \{\text{corr}(X_i, X_j)\}_{i,j=1}^d$$

is called the *correlation matrix* of random vector \mathbf{X} .

Notice that $\text{var}(X) = \text{cov}(X, X)$.

Theorem 23 (Properties of covariance and correlation). Let \mathbf{X} be a random vector $\text{Var } \mathbf{X}$ its variance matrix and $\text{Corr } \mathbf{X}$ its correlation matrix. Let X and Y be random variables with finite second moment. Then

- (1) $-1 \leq \text{corr}(X, Y) \leq 1$, $\text{corr}(X, X) = 1$.
- (2) $|\text{corr}(X, Y)| = 1$ iff there exist $a \neq 0$ and b such that $\text{P}[X = aY + b] = 1$.
- (3) $\text{cov}(aX + b, cY + d) = ac \text{cov}(X, Y)$.
- (4) $\text{corr}(aX + b, cY + d) = \text{sign}(ac) \text{corr}(X, Y)$.
- (5) If X and Y are independent then $\text{cov}(X, Y) = \text{corr}(X, Y) = 0$.
- (6) $\text{Var } \mathbf{X}$ and $\text{Corr } \mathbf{X}$ are positively semidefinite.
- (7) $\text{Var}(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A} \text{Var}(\mathbf{X})\mathbf{A}^T$ for any $l \times d$ matrix \mathbf{A} and l -dimensional vector \mathbf{b} .

Definition 24 (Uncorrelated random variables). Random variables X and Y with finite second moments are called *uncorrelated* if $\text{corr}(X, Y) = 0$.

Problem 16. Find uncorrelated random variables which are not independent.

Theorem 24 (Transformation of random vector). Let \mathbf{X} be a d -dimensional random vector and $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^l$ (measurable¹) function. Then $\mathbf{Y} = \phi(\mathbf{X})$ is a l -dimensional random vector and its distribution is given by

$$P_{\mathbf{Y}}(B) = \text{P}[\mathbf{Y} \in B] = \text{P}[\mathbf{X} \in \phi^{-1}(B)] = P_{\mathbf{X}}(\phi^{-1}(B))$$

for any $B \in \mathcal{B}(\mathbb{R}^l)$.

If \mathbf{X} is a discrete random vector, then \mathbf{Y} is also discrete random vector (not necessary \mathbb{N}^l -valued) and

$$\text{P}[\mathbf{Y} = \mathbf{y}] = p_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}; \phi(\mathbf{x})=\mathbf{y}} p_{\mathbf{X}}(\mathbf{x}).$$

¹By measurable we mean here that the pre-image of Borel set is again Borel, i.e. $\{\mathbf{x}; \phi(\mathbf{x}) \in B\} \in \mathcal{B}(\mathbb{R}^d)$ for any $B \in \mathcal{B}(\mathbb{R}^l)$.

Theorem 25 (Distribution of sum and product I). *Let $\mathbf{X} = (X_1, X_2)$ be discrete (\mathbb{N}^2 -valued) random vector. Let $Y = X_1 + X_2$ and $Z = X_1 X_2$. Then*

$$\begin{aligned} P[Y = y] &= \sum_{n \in \mathbb{N}} P[X_1 = n, X_2 = y - n], \\ P[Z = z] &= \sum_{n \in \mathbb{N}} P[X_1 = n, X_2 = z/n]. \end{aligned}$$

The first distribution (or formula) is called convolution of distribution of X_1 and X_2 .

Theorem 26 (Moment generating function of sum). *Let X and Y be independent random variables and ψ_X and ψ_Y their moment generating functions. Then*

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t).$$

Problem 17. It is possible to extend the formula for the distribution of sum and product also to more general discrete random vectors. Consider random vector $\mathbf{X} = (X_1, X_2)$. Find formula for

- (1) distribution of $X_1 + X_2$ if \mathbf{X} is \mathbb{Z}^2 -valued random vector.
- (2) distribution of $X_1 X_2$ if \mathbf{X} is \mathbb{Z}^2 -valued random vector (work carefully with zero and the fact that $z/n = (-z)/(-n)$).
- (3) distribution of $X_1 - X_2$ and X_1/X_2 if \mathbf{X} is \mathbb{Z}^2 -valued random vector.

Problem 18. Throw two dices and denote (X_1, X_2) the result. Find the distribution of transformed random vector $(X_1 + X_2, X_2)$ and show that the first marginal distribution of the transformed random vector is exactly the same as the one you get from the convolution formula. Try also the transformation to random vector $(X_1 + X_2, X_1 - X_2)$ and its marginal distribution. Repeat the exercise for the product $X_1 X_2$.

6. ABSOLUTELY CONTINUOUS RANDOM VECTORS

Definitions 17, 18, 19 and 20 hold also for continuous random vectors.

Definition 25 (Absolutely continuous random vector). Random vector $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ is called absolutely continuous (AC) if there exists a non-negative function $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$F_{\mathbf{X}}(\mathbf{a}) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_d} f_{\mathbf{X}}(x_1, \dots, x_d) dx_d \dots dx_1 \text{ for any } \mathbf{a} \in \mathbb{R}^d.$$

Function f is called (joint) probability density function, or just joint density, of random vector \mathbf{X} .

Theorem 27 (Density of absolutely continuous random vector). *Let $\mathbf{X} = (X_1, \dots, X_d)$ be an AC random vector with density $f_{\mathbf{X}}$. Then for its marginal distribution holds*

$$F_{X_i}(a) = \int_{-\infty}^a f_{X_i}(x) dx,$$

where

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_d) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_d$$

is the marginal density of X_i .

Theorem 28 (Characterisation of independence II). *Let $\mathbf{X} = (X_1, \dots, X_d)$ be an AC random vector with density $f_{\mathbf{X}}$. Then the random variables X_1, \dots, X_d are independent iff for any $\mathbf{a} = (a_1, \dots, a_d)$*

$$f_{\mathbf{X}}(\mathbf{a}) = \prod_{i=1}^d f_{X_i}(a_i)$$

Theorem 29 (Moments of AC random vector). *Let $\mathbf{X} = (X_1, \dots, X_d)$ be an AC random vector with density $f_{\mathbf{X}}$. Then for any suitable function $g: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\text{E}g(\mathbf{X})$ exists it holds*

$$\text{E}g(\mathbf{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_d) f_{\mathbf{X}}(x_1, \dots, x_d) dx_1 \dots dx_d.$$

The distribution of a transformation of AC random vector may be always determined by Theorem 24. There are also specific formulas for particular cases.

Theorem 30 (Convolution formula for density). *Let $\mathbf{X} = (Y, Z)$ be absolutely continuous random vector with joint density $f(y, z)$. Then the sum $V = Y + Z$ is an absolutely continuous random variable with density*

$$f_V(v) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(y, v - y) dy.$$

Theorem 31 (Density of ratio and product). *Let $\mathbf{X} = (Y, Z)$ be absolutely continuous random vector with joint density $f(y, z)$ and such that $\text{P}[Z > 0] = 1$. Then the ratio $V = Y/Z$ is an absolutely continuous random variable with density*

$$f_V(v) = \int_0^{\infty} f_{\mathbf{X}}(vz, z) z dz,$$

and the product $U = Y \cdot Z$ is an absolutely continuous random variable with density

$$f_U(u) = \int_0^{\infty} f_{\mathbf{X}}\left(\frac{v}{z}, z\right) \frac{1}{z} dz.$$

Uniform distribution Let $-\infty < a < b < \infty$. Random variable X has uniform distribution on the interval (a, b) if the density f_X is constant on (a, b) and zero elsewhere. Clearly,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in (a, b), \\ 0 & \text{elsewhere.} \end{cases}$$

Let $B \in \mathbb{R}^d$ be a connected set with nonzero volume $\lambda(B)$. Random vector \mathbf{X} has uniform distribution on B if the density $f_{\mathbf{X}}$ is constant on B and zero elsewhere. Clearly

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{\lambda(B)} & \text{for } \mathbf{x} \in B, \\ 0 & \text{elsewhere.} \end{cases}$$

Typical examples are uniform distribution on rectangles, triangles, circles, etc.

Normal (Gaussian) distribution Let $\mu \in \mathbb{R}$, and $\sigma^2 > 0$. Random variable X has normal distribution (also called Gaussian distribution) if its density has form

$$f_X(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

For X it holds $\text{E}X = \mu$ and $\text{var } X = \sigma^2$.

Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and Σ be $d \times d$ symmetric positive definite matrix. Random vector \mathbf{X} has d -variate normal distribution if its density is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi \det(\Sigma))^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d.$$

For \mathbf{X} it holds $E\mathbf{X} = \boldsymbol{\mu}$ and $\text{Var } \mathbf{X} = \Sigma$.

Random variable (vector) with normal distribution is also called *Gaussian r.v.* Note that marginal distribution of Gaussian distribution is again Gaussian.

Theorem 32 (Independence in Gaussian distribution). *Let (X, Y) is normally distributed random vector. Then X and Y are independent iff $\text{cov}(X, Y) = 0$. Let $\mathbf{X} = (X_1, \dots, X_d)$ be a d -dimensional Gaussian r.v. Then X_1, \dots, X_d are independent iff the variance matrix Σ is diagonal.*

Exponential distribution Let $\mu > 0$. Random variable X has exponential distribution with parameter μ if it has density

$$f_X(x) = \begin{cases} \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) & \text{for } x > 0, \\ 0 & \text{elsewhere.} \end{cases}$$

7. CONDITIONAL DISTRIBUTION AND CONDITIONAL EXPECTATION

We restrict the full definition *only to discrete random vectors*.

Definition 26 (Conditional distribution of d.r.v.). Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ be a d -dimensional discrete random vector, where $\mathbf{Y} = (X_1, \dots, X_p)$ and $\mathbf{Z} = (X_{p+1}, \dots, X_d)$. For any $\mathbf{z} \in \mathbb{N}_0^{d-p}$ such that $P[\mathbf{Z} = \mathbf{z}] > 0$ define the *conditional distribution of \mathbf{Y} given $\mathbf{Z} = \mathbf{z}$* (or simply “given \mathbf{z} ”) by

$$p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = P[\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z}] = \frac{p_{(\mathbf{Y}, \mathbf{Z})}(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} = \frac{P[(\mathbf{Y}, \mathbf{Z}) = (\mathbf{y}, \mathbf{z})]}{P[\mathbf{Z} = \mathbf{z}]}, \quad \mathbf{y} \in \mathbb{N}_0^d.$$

The conditional distribution is probability distribution. Hence, we may define its moments.

Definition 27 (Conditional expectation of d.r.v.). Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ be a d -dimensional discrete random vector, where $\mathbf{Y} = (X_1, \dots, X_p)$ and $\mathbf{Z} = (X_{p+1}, \dots, X_d)$. For any $\mathbf{z} \in \mathbb{N}_0^{d-p}$ such that $P[\mathbf{Z} = \mathbf{z}] > 0$ and random variable $S = g(\mathbf{Y}, \mathbf{Z})$ define the *conditional expectation of S given $\mathbf{Z} = \mathbf{z}$* (or simply “given \mathbf{z} ”) by

$$E(S|\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{y} \in \mathbb{N}_0^{d-p}} g(\mathbf{y}, \mathbf{z}) p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z})$$

if the sum on the right hand side is defined. In particular, if $p = d - 1$, i.e. $Y = X_1$ is one-dimensional random variable we define conditional expectation and conditional variance as

$$E(Y|\mathbf{Z} = \mathbf{z}) = \sum_{y \in \mathbb{N}_0} y p_{Y|\mathbf{Z}}(y|\mathbf{z}),$$

$$\text{var}(Y|\mathbf{Z} = \mathbf{z}) = E((Y - E(Y|\mathbf{Z} = \mathbf{z}))^2|\mathbf{Z} = \mathbf{z}).$$

Theorem 33 (“Total expectation”). *Let $\mathbf{X} = (Y, \mathbf{Z})$ be a discrete random vector such that Y is a random variable and $E|Y|$ is finite. Then*

$$EY = \sum_{\substack{\mathbf{z} \in \mathbb{N}_0^{d-1} \\ P[\mathbf{Z} = \mathbf{z}] > 0}} E(Y|\mathbf{z})P[\mathbf{Z} = \mathbf{z}]$$

Compare this theorem with the “law of total probability”. Conditional expectation may be defined also as a random variable.

Definition 28 (Conditional expectation as random variable). Let (X, Y) be a bivariate discrete random vector such that $E|X|$ is finite. Then random variable $E(X|Y) : \Omega \rightarrow \mathbb{R}$ defined as

$$E(X|Y)(\omega) = E(X|Y = Y(\omega))$$

is well defined random variable. (The definition may be simply generalised to more-dimensional random vectors).

Theorem 34 (Expectation of conditional expectation). Let (Y, \mathbf{Z}) be a discrete random vector such that $E|Y|$ is finite. Then

$$E(E(Y|\mathbf{Z})) = EY.$$

Problem 19. Throw two dices, denote X and Y the results, respectively. Find

- (1) Joint and marginal distributions of random vector $(X, X + Y, X - Y)$.
- (2) $\text{Var}(X, X + Y, X - Y)$, $\text{Corr}(X, X + Y, X - Y)$.
- (3) Are these variables independent? Is there any pair of independent r.v.'s?
- (4) $E(X|X + Y = z)$, $E(X|X + Y = z, X - Y = w)$.
- (5) $E(X + Y|X = x)$, $E(X + Y|X - Y = w)$.
- (6) Values and distribution of random variable $E(X|X + Y)$.
- (7) $E(X + Y)$ using Theorem 33.

For absolutely continuous random vectors we may correctly define *conditional density*. The main problem is that we need to condition by random event $[\mathbf{Z} = \mathbf{z}]$ which probability is zero.

Definition 29 (Conditional density of AC r.v.). Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ be an absolutely continuous random vector with density function $f_{\mathbf{X}}$. Denote $f_{\mathbf{Z}}$ the (marginal) density of the absolutely continuous random vector \mathbf{Z} . The conditional density of \mathbf{Y} given $\mathbf{Z} = \mathbf{z}$ is defined as

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = \begin{cases} \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{z})}{f_{\mathbf{Z}}(\mathbf{z})} & \text{if } f_{\mathbf{Z}}(\mathbf{z}) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 30 (Conditional expectation of AC r.v.). Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ be an absolutely continuous random vector and let $S = g(\mathbf{Y}, \mathbf{Z})$ be random variable. The conditional expectation of S given $\mathbf{Z} = \mathbf{z}$ is defined as

$$E(S|\mathbf{Z} = \mathbf{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{y}, \mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) dy_1 \cdots dy_p$$

if the integral exists.

8. INEQUALITIES AND BOUNDS

Theorem 35 (Markov inequality I). Let X be a non-negative random variable. Then for any $a > 0$

$$P[X \geq a] \leq \frac{EX}{a}.$$

Theorem 36 (Markov inequality II). Let X be a non-negative random variable. Then for any $a > 0$ and for $k > 0$

$$P[X \geq a] \leq \frac{EX^k}{a^k}.$$

Theorem 37 (Chebyshev inequality). *Let X be random variable with finite expectation EX . Then for any $\varepsilon > 0$*

$$P[|X - EX| \geq \varepsilon] \leq \frac{\text{var } X}{\varepsilon^2}.$$

Theorem 38 (Kolmogorov inequality). *Let X_1, X_2, \dots, X_n be independent random variables with finite expectations EX_i and finite variances $\text{var } X_i$. Then for any $\varepsilon > 0$*

$$P\left[\max_{1 \leq k \leq n} \left| \sum_{i=1}^k X_i - EX_i \right| \geq \varepsilon\right] \leq \frac{\sum_{i=1}^n \text{var } X_i}{\varepsilon^2}.$$

Problem 20. Suppose we toss a fair dice 100 times. Denote X_1, \dots, X_{100} the results, denote $S_k = \sum_{i=1}^k X_i$ and use Chebyshev and Kolmogorov inequalities to find upper bounds for

$$P[|S_{100} - 350| \geq l], \quad P\left[\max_{1 \leq k \leq n} |S_k - 3.5k| \geq m\right]$$

for different values of l and m .

Problem 21. Given positive integer k find a non-negative random variable X such that

$$P[X \geq k] = \frac{EX}{k}.$$

(Looking for non-negative random variable with given properties is equivalent to problem of finding probability measure on \mathbb{R}^+ satisfying these conditions.)

Theorem 39 (Chernoff bounds). *Let X be a random variable and ψ_X its moment generating function. Then*

$$P[X \geq a] \leq \min_{t>0} \frac{\psi_X(t)}{\exp(ta)}$$

$$P[X \leq a] \leq \min_{t<0} \frac{\psi_X(t)}{\exp(ta)}$$

Definition 31 (Poisson trials). *Let X_1, X_2, \dots be independent random variables such that*

$$P[X_i = 1] = p_i = 1 - P[X_i = 0], \quad p_i \in (0, 1).$$

Such random variables are called *Poisson trials*.

Note: Bernoulli trials are special case of Poisson trials for which $p_i = p$ for all $i = 1, 2, \dots$

Theorem 40 (Upper Chernoff bounds for Poisson trials). *Let X_1, \dots, X_n be independent Poisson trial with $P[X_i = 1] = p_i$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu_s = \sum_{i=1}^n p_i$. Then*

(1) For $\delta > 0$

$$P[S_n \geq (1 + \delta)\mu_s] \leq \left(\frac{e^\delta}{(1 + \delta)^{1 + \delta}}\right)^{\mu_s}.$$

(2) For $0 < \delta \leq 1$

$$P[S_n \geq (1 + \delta)\mu_s] \leq e^{-\mu_s \delta^2 / 3}.$$

(3) For $\delta > 6\mu_s$

$$P[S_n \geq \delta] \leq 2^{-\delta}.$$

Theorem 41 (Lower Chernoff bounds for Poisson trials). *Let X_1, \dots, X_n be independent Poisson trial with $P[X_i = 1] = p_i$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu_s = \sum_{i=1}^n p_i$. Then*

(1) For $0 < \delta < 1$

$$P[S_n \leq (1 - \delta)\mu_s] \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^{\mu_s}.$$

(2) For $0 < \delta \leq 1$

$$P[S_n \leq (1 - \delta)\mu_s] \leq e^{-\mu_s \delta^2 / 2}.$$

Problem 22. Consider n independent Bernoulli trials X_1, \dots, X_n with (unknown) probability of success $p \in (0, 1)$. Let

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

be your guess about p . For given $\alpha \in (0, 1)$ use Chebyshev and Markov inequalities and Chernoff bounds to find smallest possible δ such that

$$P[p \in (\hat{p} - \delta, \hat{p} + \delta)] \geq 1 - \alpha.$$

9. RANDOM SAMPLES AND LIMIT THEOREMS

Definition 32 (Set limsup and liminf). Let $A_n, n = 1, 2, \dots$ be sets (random events). Define

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i, \quad \liminf_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i.$$

Clearly, if $a \in \limsup_{n \rightarrow \infty} A_n$ then there exists infinitely many sets A_i such that $a \in A_i$. If $b \in \liminf_{n \rightarrow \infty} A_n$ then there is at most finite number of sets A_i such that $b \notin A_i$.

Theorem 42 (Borel-Cantelli 0-1 law). (1) *Let $A_n, n = 1, 2, \dots$ be random events. Then*

$$\sum_{i=1}^{\infty} P(A_i) < \infty \Rightarrow P\left(\limsup_{n \rightarrow \infty} A_n\right) = 0.$$

(2) *Let $A_n, n = 1, 2, \dots$ be independent random events. Then*

$$\sum_{i=1}^{\infty} P(A_i) = \infty \Rightarrow P\left(\limsup_{n \rightarrow \infty} A_n\right) = 1.$$

Note that $\liminf A_n = (\limsup A_n^C)^C$. Hence, Theorem 42 may be used both for $\limsup A_n$ and for $\liminf A_n$

Definition 33 (Random sample). A sequence X_1, X_2, \dots, X_n of independent and identically distributed (iid) random variables (vectors) is called random sample of (sample) size n (from distribution P_X).

Definition 34 (Sample moments). Let X_1, \dots, X_n be a random sample. Then

- (1) $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is called *sample mean*.
- (2) $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is called *sample variance*.
- (3) $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \chi(X_i \leq x)$, where $\chi(\cdot)$ is the indicator function, is called *empirical distribution function*.

Theorem 43 ((weak) Law of large numbers). *Let X_1, X_2, \dots be independent and identically distributed random variables with finite mean $EX_1 = \mu$ and finite variance $\text{var } X_1 = \sigma^2$. Then*

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \varepsilon \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

This property is called (weak) *consistency of sample mean*. The assumption $\sigma^2 < \infty$ may be relaxed but the proof is then much more difficult. This is denoted as

$$\overline{X}_n - \mu \xrightarrow{\mathbb{P}} 0, \text{ or } \overline{X}_n \xrightarrow{\mathbb{P}} \mu$$

Theorem 44 (Consistency of empirical d.f.). *Let X_1, X_2, \dots be independent and identically distributed random variables with c.d.f. F_X . Then for any x*

$$\mathbb{P} \left[\left| \widehat{F}_n(x) - F_X(x) \right| > \varepsilon \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

The previous Theorem holds even uniformly, i.e. $\mathbb{P} \left[\sup_x \left| \widehat{F}_n(x) - F_X(x) \right| > \varepsilon \right] \rightarrow 0$

Theorem 45 (Central limit theorem). *Let X_1, X_2, \dots be independent and identically distributed random variables with finite mean $EX_1 = \mu$ and finite positive variance $0 < \text{var } X_1 = \sigma^2$. Then*

$$\mathbb{P} \left[\sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} \leq x \right] \rightarrow \Phi(x),$$

where $\Phi \cdot$ is the distribution function of standard normal distribution (with zero mean and unit variance). Equivalently we write

$$\sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} N(0, 1),$$

or

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0, 1),$$

or

$$\sqrt{n} (\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

This is really very important theorem!!! It is also written that

Theorem 46 (Delta theorem). *Let Y_1, Y_2, \dots be a sequence of random variables such that*

$$\sqrt{n} (Y_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

and let g be a differentiable function. Then

$$\sqrt{n} (g(Y_n) - g(\mu)) \xrightarrow{d} N(0, (g'(\mu))^2 \sigma^2)$$

There exist multivariate and more general versions of both CLT and Delta theorem.

Theorem 47 (Cramér–Slutskij theorem). *Let $X_n \xrightarrow{d} N(\mu, \sigma^2)$, $U_n \xrightarrow{\mathbb{P}} a$, and $Z_n \xrightarrow{\mathbb{P}} s > 0$. Then*

$$Z_n X_n + U_n \xrightarrow{d} N(s\mu + a, s^2 \sigma^2)$$

The Cramér–Slutskij and Delta theorems are essential tools for asymptotic estimation techniques based on Central limit theorem. From C–S theorem follows that the convergence \xrightarrow{d} is weaker than the convergence $\xrightarrow{\mathbb{P}}$.

10. POINT AND INTERVAL ESTIMATES, HYPOTHESES

Definition 35. Parametric family of distributions Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be a family of distributions such that

- (1) All P_θ are distributions on the same sample space (here \mathbb{R} or \mathbb{R}^d).
- (2) P_θ are explicitly defined with *unknown* parameter θ .
- (3) Θ is the set of all possible parameters.

Example 2. Examples of parametric families Parameric families are usually given by the density function or by the distribution function.

- Bernoulli distribution: The parameter is $p \in (0, 1)$, the unknown probability of success.
- Univariate normal distribution: The parameter is two-dimensional: (μ, σ^2) , where μ is the expectation and σ^2 is the variance.

Note that there may be more equivalent parametrisations of given family of distribution.

Definition 36. Point estimator Let X_1, \dots, X_n be independent and identically distributed random variables (vectors) following distribution P_θ from some parametric family \mathcal{P} . This is called *random sample from the distribution P_θ* . The point estimation problem is the problem to find *statistics* $T : \mathbb{R}^n \rightarrow \Theta$, $T(X_1, \dots, X_n)$, where T does not depend on the unknown parameter θ such that T is “good” approximation of θ .

- (1) The point estimator T is called *unbiased* if for all $\theta \in \Theta$ holds $ET = \theta$ provided θ is the true value of the parameter.
- (2) The point estimator T is called *consistent* if for all $\theta \in \Theta$ holds $T \xrightarrow{P} \theta$ provided θ is the true value of the parameter.

Consistency and unbiasedness are usually considered as the desirable properties for any estimator. Having two consistent or unbiased estimators the one with smaller variance is preferred.

There are several methods how to find estimator.

Definition 37. Method of moments Consider X_1, \dots, X_n random sample from distribution P_θ , θ being unknown parameter. Suppose the moments of X_1 are functions of θ (which is usually the case), say $EX_1^j = \tau_j(\theta)$. Define sample moments

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

The moment estimator is then defined as solution to equations

$$\tau_j(\theta) = \hat{\mu}_j, j = 1, \dots, l,$$

where l is (usually) the dimension of θ .

Law of large numbers, central limit theorem, delta theorem, Cramér–Slutsky theorem are often useful to determine the properties of point estimators.

Problem 23. Consider following parametric families, and random samples form these distributoins. Use method of moments to find the point estimation of the parameter. Are these estimates unbiased and consistent?

- (1) $P[X = 1] = p \in (0, 1), P[X = 0] = 1 - p$.
- (2) $P[X \leq y] = 1 - \exp(-\lambda y)$ for $y > 0$, where $\lambda > 0$ is unknown parameter.

(3) X has density

$$f(x) = \begin{cases} \lambda \exp\{-\lambda(x-a)\} & x > a, \\ 0 & x < a, \end{cases}$$

where $a \in \mathbb{R}$ and $\lambda > 0$ are unknown parameters.

(4) X has normal distribution with $\sigma^2 = 1$ and $\mu \in \mathbb{R}$ is parameter.

(5) X has normal distribution with unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Definition 38. Interval estimator Let X_1, \dots, X_n be independent and identically distributed random variables (vectors) following distribution P_θ from some parametric family \mathcal{P} , where $\Theta \subset \mathbb{R}$. The interval estimation problem is the problem to find two *statistics* $L, U : \mathbb{R}^n \rightarrow \mathbb{R}$, where L and U do not depend on the unknown parameter θ , such that

$$\mathbb{P}[L \leq \theta \leq U] \geq 1 - \alpha$$

if θ is the true value of the parameter. We call the interval $[L, U]$ interval estimate (or *confidence interval*) with *significance level* $1 - \alpha$.

There is no general algorithm how to find interval estimate. However, this procedure may be often successful:

- (1) Find function $H : \Theta \times \mathbb{R}^n \rightarrow \mathbb{R}$ such that $H(\theta; X_1, \dots, X_n)$ is random variable whose distribution P_H doesn't depend on θ (it is sufficient in asymptotic sense).
- (2) Find appropriate quantiles of P_H , i.e. values q_l and q_u such that

$$\mathbb{P}[q_l \leq H \leq q_u] \geq 1 - \alpha.$$

Usually $q_{\alpha/2}$ and $q_{1-\alpha/2}$ defined as

$$q_{\alpha/2} = \inf\{q : \mathbb{P}[H \leq q] > \alpha/2\}, \quad q_{1-\alpha/2} = \inf\{q : \mathbb{P}[H \leq q] > 1 - \alpha/2\}$$

are used as q_l and q_u , respectively. (Show that indeed $\mathbb{P}[q_{\alpha/2} \leq H \leq q_{1-\alpha/2}] \geq 1 - \alpha$.)

- (3) Find $L(X_1, \dots, X_n; q_{\alpha/2}, q_{1-\alpha/2})$ and $U(X_1, \dots, X_n; q_{\alpha/2}, q_{1-\alpha/2})$ such that the inequalities

$$q_{\alpha/2} \leq H(\theta; X_1, \dots, X_n) \leq q_{1-\alpha/2}$$

hold if and only if

$$L(X_1, \dots, X_n; q_{\alpha/2}, q_{1-\alpha/2}) \leq \theta \leq U(X_1, \dots, X_n; q_{\alpha/2}, q_{1-\alpha/2})$$

- (4) Then, since

$$\mathbb{P}[L \leq \theta \leq U] = \mathbb{P}[q_{\alpha/2} \leq H \leq q_{1-\alpha/2}] \geq 1 - \alpha$$

we have found the confidence interval.

Interval estimates may be used for *test of statistical hypothesis*. Assume we have two claims about the parameter:

- (1) $H_0 : \theta = t$ is the hypothesis.
- (2) $H_1 : \theta \neq t$ is the alternative.

Such H_0 is called *simple hypothesis*, H_1 is called *both-sided alternative*. The value t is given before the experiment and it is hypothetical value of the unknown parameter θ .

Having interval estimator $[L, U]$ of θ with significance level $1 - \alpha$ we may do this decision:

- (1) We *reject* the hypothesis H_0 if $t \notin [L, U]$. The hypothesis is *rejected at statistical significance level* α .
- (2) We *do not reject* the hypothesis H_0 if $t \in [L, U]$. The hypothesis is *not rejected at statistical significance level* α .

Note that **we do not accept** the hypothesis. It is either rejected or not rejected. The decision depends on the significance level.