STA447/STA2006 Stochastic Processes

Gun Ho Jang

Lecture on January 9, 2014

Note

This lecture note is prepared for the course STA447/STA2006 Stochastic Processes. This lecture note may contain flaws. Please consult text book or reference books for confidence.

- Gun Ho Jang

* indicates graduate level. So you may skip those parts.

1 Review of Probability Theory

1.1 Probability

Probability is well defined when the number of expected outcomes is finite.

Example 1 (Coin toss)**.** The number of head in 3 consecutive tosses can be $0, 1, 2$ and 3. The probability of having only one head is defined by the number of cases having only one head to the number of all possible outcomes. All possible outcomes are HHH, HHT, HTH, HTT, THH, THT, TTH and TTT. Among them HTT, THT and TTH have only one head. Hence, the probability having only one head is $3/8 = 0.375$

Definition 1. A non-empty space Ω is called a *sample space*. An *event* is a subset of the sample space Ω.

Example. In Example 1, the sample space is $\Omega = \{$ HHH, HHT, HTH, HTT, THH, THT, TTH, TTT $\}$ and the event of interest is $E = \{$ HTT, THT, TTH $\}$.

Note. The set of all event is 2^Ω (the *power set* of Ω), that is, $2^\Omega = \{A \subset \Omega\}$.

Definition 2. When the sample space Ω is finite or at most countable, a probability measure P is a function $2^\Omega \mapsto [0, 1]$ such that

(a) (non-negativity) For any $E \subset \Omega$, $P(E) \geq 0$.

(b) (countable additivity) For any disjoint events $E_1, E_2, \ldots \subset \Omega$, $P(\cup_{k=1}^\infty E_k) = \sum_{k=1}^\infty P(E_k) = P(E_1) + P(E_2) + \cdots$.

(c) (totality) $P(\Omega) = 1$.

Example. In Example 1, the probability measure was taken as a uniform measure, that is, $P(\{x\}) = 1/|\Omega|$ for all $x \in \Omega$.

A function $P_2 : 2^\Omega \to [0, 1]$ defined as $P(E) = |E \cap \{HHH\}|/2 + |E \cap \{HHH\}^c|/14$ for any $E \in 2^\Omega$ is also a probability measure.

Note. Let Ω be the sample space. Probability is well-defined for most of subsets $A \subset \Omega$. But there exists some sets $N \subset \Omega$ on which probability cannot be defined. Such sets are called *non-measurable sets*. Not all subsets $A \subset \Omega$ have probability.

Let $\mathcal{F}$ be the collection of subsets on which probability is well-defined. Such collections $\mathcal{F}$ is called a *σ-field.*

Definition* 3. An *σ-field* $\mathcal{F} \subset 2^{\Omega}$ is a collection of subsets satisfying

(a) (empty set) $\emptyset \in \mathcal{F}$,

(b) (closure under complement) For any $A \in \mathcal{F}$, $A^c \in \mathcal{F}$,

(c) (closure under countable union) For any $A_1, A_2, \ldots \in \mathcal{F}$, $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

A *Borel σ-field* $\mathcal{B} = \mathcal{B}(\Omega)$ is the smallest σ-field containing all open subsets of Ω.

Example 2. $\mathcal{F} = \{\emptyset, \Omega\}$ is a σ-field which is called the *trivial σ-field.*

Exercise 1. Show that 2^{Ω} is a σ-field if Ω is finite or countably infinite.

Exercise* 2. Show that there exists the smallest field containing a finite number of events $E_1, \ldots, E_n$.

Exercise* 3. Show that there exists the smallest σ-field containing a collection $\mathcal{F}_0$ of measurable sets.

Definition* 4. A pair $(\Omega, \mathcal{F})$ is called a *measurable space* if $\mathcal{F}$ is a σ-field of Ω. A measure μ on a measure space $(\Omega, \mathcal{F})$ is a mapping $\mathcal{F} \to \mathbb{R}_+ := \{r \in \mathbb{R} : r \geq 0\}$ satisfying that

(a) (non-negativity) $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$,

(b) (countable additivity) if $A_n \in \mathcal{F}$ is a countable sequence of disjoint sets, then

$$\mu(\bigcup_n A_n) = \sum_n \mu(A_n).$$

If $\mu(\Omega) = 1$, then μ is called a *probability measure.* Elements in $\mathcal{F}$ are called *measurable sets.* A probability measure P always comes with a *probability triple* or *probability space* $(\Omega, \mathcal{F}, P)$.

Theorem 1 (Properties of probability)**.** Let P be a measure on Ω.

(a) (monotonicity). If $A \subset B$, then $P(A) \leq P(B)$.

(b) (subadditivity). If $A \subset \cup_{n=1}^{\infty} A_n$, then $P(A) \leq \sum_{n=1}^{\infty} P(A_n)$.

(c) (continuity from below). If $A_n \nearrow A$ ($A_1 \subset A_2 \subset \cdots$ and $\cup_n A_n = A$), then $P(A_n) \nearrow P(A)$.

(d) (continuity from above). If $A_n \searrow A$ ($A_1 \supset A_2 \supset \cdots$ and $\cap_n A_n = A$), then $P(A_n) \searrow P(A)$.

Theorem* 2 (Properties of measure)**.** Let μ be a measure on $(\Omega, \mathcal{F})$.

(a) (monotonicity). If $A \subset B$, then $\mu(A) \leq \mu(B)$.

(b) (subadditivity). If $A \subset \cup_{n=1}^{\infty} A_n$, then $\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)$.

(c) (continuity from below). If $A_n \nearrow A$ ($A_1 \subset A_2 \subset \cdots$ and $\cup_n A_n = A$), then $\mu(A_n) \nearrow \mu(A)$.

(d) (continuity from above). If $A_n \searrow A$ ($A_1 \supset A_2 \supset \cdots$ and $\cap_n A_n = A$), then $\mu(A_n) \searrow \mu(A)$.

Proof. Part (a). Since $B = A \cup (B - A)$,

$$\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A).$$

Part (b). Let $B_1 = A_1 \cap A$, $B_n = \cup_{i \leq n}(A_i \cap A) - \cup_{i<n} B_i$ for $n \geq 2$. It is easy to see that $A = \cup_n B_n$ and $B_1 \subset A_1, B_n \subset A_n - \cup_{i<n} B_i$. Hence,

$$\mu(A) = \mu(\cup_n B_n) = \sum_n \mu(B_n) \leq \sum_n \mu(A_n)$$

Part (c). Let $B_1 = A_1$, $B_n = A_n - A_{n-1}$ for $n \geq 2$. Then $A_n = \cup_{i \leq n} B_i$ and $A = \cup_n B_n$. Hence $\mu(A_n) = \mu(A_{n-1}) + \mu(B_n)$ is non-decreasing and

$$\mu(A_n) = \mu(\cup_{i \leq n} B_i) = \sum_{i \leq n} \mu(B_i) \to \sum_n \mu(B_n) = \mu(\cup_n B_n) = \mu(A).$$

Part (d). If $A \subset B$, then $\mu(B - A) = \mu(B) - \mu(A)$. Note $A_1 - A_n \uparrow A_1 - A$. (c) implies $\mu(A_1) - \mu(A_n) = \mu(A_1 - A_n) \uparrow \mu(A_1 - A) = \mu(A_1) - \mu(A)$. Hence $\mu(A_n) \downarrow \mu(A)$. $\square$

1.2 Random Variable and Distribution Functions

Definition 5. Let P be a probability on a sample space Ω. A function $X : \Omega \to \mathbb{R}$ is said to be a *random variable* if $P(X > x) = P(\{\omega \in \Omega : X(\omega) > x\})$ are well-defined for all $x \in \mathbb{R}$, that is, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable.

When $k > 1$, a multivariate function $X = (X_1, \ldots, X_k)$ is called a *random vector* if $X_1, \ldots, X_k$ are random variable.

Definition* 6. Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is said to be *measurable* if $f^{-1}(E_2) \in \mathcal{F}_1$ for all $E_2 \in \mathcal{F}_2$ where f^{-1} is the preimage, that is, $f^{-1}(E_2) = \{x \in \Omega_1 : f(x) \in E_2\}$. If both $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are Borel spaces, then a measurable function $f : \Omega_1 \to \Omega_2$ is called a *Borel function*.

Example 3. If A is a subset of Ω, the indicator function 1_A is measurable if A is a measurable set and non-measurable if A is non-measurable.

Theorem 3. Let $Y, X, X_1, X_2, \ldots$ be random variables.
(a) $P(X > r)$ is well-defined for any $r \in \mathbb{R}$.
(b) If $g : \mathbb{R} \to \mathbb{R}$ is continuous (or measurable), then $g(X)$ is also a random variable.
(c) $X + Y$ and XY are random variables.
(d) In general $X_1 + X_2 + \cdots + X_n$, $\inf_n X_n$, $\sup_n X_n$, $\limsup_n X_n$, $\liminf_n X_n$ are all random variables.

Proof. A map $(X_1, \ldots, X_n) \mapsto X_1 + \cdots + X_n$ is continuous. Hence, $X_1 + \cdots + X_n$ is a measurable mapping, that is, a random variable. Note that $\{\inf_n X_n < a\} = \cup_n \{X_n < a\} \in \mathcal{F}$, $\{\sup_n X_n > a\} = \cup_n \{X_n > a\} \in \mathcal{F}$, and $\liminf_n X_n = \sup_n \inf_{m \geq n} X_m$, $\limsup_n X_n = \inf_n \sup_{m \geq n} X_m$ are random variables. $\square$

Definition 7. A function $F : \mathbb{R} \to [0, 1]$ is said to be a *(cumulative) distribution function of* X if $F(x) = P(X \leq x) = P(X^{-1}((-\infty, x]))$ for all $x \in \mathbb{R}$.

Theorem 4 (Properties of distribution functions)**.** Let F be a distribution function. Then (a) F is nondecreasing,
(b) $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$,
(c) F is right continuous, that is, $\lim_{y \searrow x} F(y) = F(x)$,
(d) $F(x-) \coloneqq \lim_{y \nearrow x} F(y) = P(X < x)$,
(e) $P(X = x) = F(x) - F(x-)$.

Proof. (a) Since $\{X \leq x\} \subset \{X \leq y\}$ for all $x \leq y$, $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$.
(b) $\lim_{x \to \infty} F(x) = \lim_{x \to \infty} P(X \leq x) = P(\Omega) = 1$ and $\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} P(X \leq x) = P(\emptyset) = 0$.
(c) For $y > x$, $F(y) = P(X \leq y) = P(X \leq x) + P(x < X \leq y) = F(x) + P(X \in (x, y]) \to F(x)$ as $y \searrow x$.
(d) $F(x-) = \lim_{y \nearrow x} F(y) = \lim_{y \nearrow x} P(X \in (-\infty, y]) = P(X \in (-\infty, x)) = P(X < x)$.
(e) $P(X = x) = P(X \leq x) - P(X < x) = F(x) - F(x-)$. $\square$

Theorem* 5. If a real function F satisfies (a)-(c) of Theorem 4, then it is a distribution function of a random variable.

Proof. We build a random variable of which distribution function is $F : \mathcal{T} \to [0, 1]$ where $\mathcal{T} = \mathbb{R}$. Let $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}(\Omega)$ and P be the Lebesgue measure on Ω. Define $W : \Omega \to \mathcal{T}$ so that $W(x) = \sup\{t : F(t) \leq x\}$. Obviously $\{W \leq t\} = \{x \in \Omega : W(x) \leq t\} = \{x \in \Omega : x \leq F(t)\}$. Then, the distribution function of W is $F_W(t) = P(W \leq t) = P(\{x : x \leq F(t)\} = P([0, F(t)]) = F(t)$. $\square$

Definition 8. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, P is a probability measure on the same space and X is a random variable. Suppose that P is absolutely continuous with respect to μ. When Ω is countable and μ is counting measure, the Radon-Nikodym derivative $p(x) = \frac{dP}{d\mu}(x)$ is called a *probability mass function* of X. When μ is a σ-finite measure, the Radon-Nikodym derivative $f(x) = \frac{dP}{d\mu}(x)$ is called a *probability density function* of X.

Note. When an abstract probability space can be constructed from random variables, we will not specify the probability space unless it is necessary.

Example 4. Some useful distributions.

$X \sim \text{Poisson}(\mu)$ for $\mu > 0$. Then, $\text{pmf}_X(x) = P(X = x) = I(X \in \mathbb{N}_+)e^{-\mu}\mu^x/x!$ for all $x \in \mathbb{N}_+$.

$U \sim \text{Uniform}([0,1])$. Then $\text{pdf}_U(u) = 1_{[0,1]}(x)$ or $\text{cdf}_U(u) = uI(0 \le u \le 1) + I(u > 1) = \min(1, \max(0,u))$.

$Y \sim \text{Exp}(\mu)$ for $\mu > 0$. Then $\text{pdf}_Y(y) = I(y > 0)e^{-y/\mu}/\mu$ or $\text{cdf}_Y(y) = 1 - e^{-y/\mu}$ for $y > 0$.

$Z \sim N(0,1)$, that is, the standard normal. Then $\text{pdf}_Z(z) = \exp(-z^2/2)/(2\pi)^{1/2}$.

Note (Notation)**.** Throughout this course ϕ and Φ are assumed as the density and distribution function of the standard normal.

1.2.1 Background on Sequences

A sequence is an iteration of numbers like $1, 1, 2, 3, 5, \ldots$.

Definition 9. A mapping $a : \mathbb{N} \to \mathbb{R}$ is called a *sequence*.

Note. Operations are well-defined on sequences like $a_n \pm b_n$.

Definition 10 (Limits)**.** Let a_n be a sequence.

A sequence a_n converges to x if for any $\epsilon > 0$, there exists $N > 0$ such that $|a_n - x| < \epsilon$ for all $n \ge N$. It is denoted by $\lim_n a_n = \lim_{n\to\infty} a_n = x$ and x is called the limit of the sequence a_n.

The supremum of a sequenced is denoted by $\sup_n a_n =$ the smallest number x satisfying $x \ge a_n$ for all n, that is, $a_n \le x$ for all n and there exists a subsequence a_{n_k} such that $\lim_{k\to\infty} a_{n_k} = x$.

The infimum of a sequence is denoted by $\inf_n a_n =$ the biggest number x satisfying $x \le a_n$ for all n, that is, $a_n \ge x$ for all n and there exists a subsequence a_{n_k} such that $\lim_{k\to\infty} a_{n_k} = x$.

The limit supremum is defined by $\limsup_n a_n = \limsup_{n\to\infty} a_n = \lim_{m\to\infty} \sup_{n\ge m} a_n$.

The limit infimum is defined by $\liminf_n a_n = \liminf_{n\to\infty} a_n = \lim_{m\to\infty} \inf_{n\ge m} a_n$.

Exercise 4. Show that $\inf_n a_n \le \liminf_n a_n \le \limsup_n a_n \le \sup_n a_n$.

Exercise 5. Show that the limit of a_n does not exist if $\liminf_n a_n \ne \limsup_n a_n$.

1.3 Expectation

Definition 11. The expectation of X is defined by

$$\mathbb{E}(X) = \int_0^\infty P(X > x)dx - \int_{-\infty}^0 P(X < x)dx$$

when at least one integral is finite. Otherwise, the expectation of X is not defined.

Example 5 (Discrete Random Variables)**.** Suppose X is a discrete random variable such that $P(X = x_n) = p_n$ for some x_n and p_n satisfying $p_n \ge 0$ and $\sum_n p_n = 1$.

Assume further $P(X \ge 0) = 1$ and $0 = x_0 \le x_1 < x_2 < \cdots$. Then, $\mathbb{E}(X) = \int_0^\infty P(X > x)dx = \sum_{n=1}^\infty P(X \ge x_n)(x_n - x_{n-1}) = \sum_{n=1}^\infty \left(\sum_{k=n}^\infty p_k\right)(x_n - x_{n-1}) = \sum_{k=1}^\infty \sum_{n=1}^k p_k(x_n - x_{n-1}) = \sum_{k=1}^\infty p_k x_k$.

In general, for discrete X, $\mathbb{E}(X) = \sum_x xP(X = x)$.

Example 6 (Continuous Random Variables)**.** Suppose X is a continuous random variable having density f. Then, $\int_0^\infty P(X > x)dx = \int_0^\infty \int_x^\infty f(y)dydx = \int_0^\infty \int_0^x f(y)dxdy = \int_0^\infty yf(y)dy$ and similarly $\int_{-\infty}^0 P(X < 0)dx = \int_{-\infty}^0 |y|f(y)\,dy$. Hence, $\mathbb{E}(X) = \int xf(x)\,dx$.

Theorem 6 (Properties of Expectation)**.** (a) if $P(X \geq 0) = 1$, then $\mathbb{E}(X) \geq 0$.
(b) $\mathbb{E}(rX) = r\mathbb{E}(X)$.
(c) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
(d) if $P(X \leq Y) = 1$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$,
(e) $|\mathbb{E}(X)| \leq \mathbb{E}|X|$.

Proof. (a) If $P(X \geq 0) = 1$, then $P(X < 0) = 0$ and $\mathbb{E}(X) = \int_0^\infty P(X > x)\, dx \geq 0$.
(b) If $r = 0$, then obvious provided by $\mathbb{E}|X| < \infty$. If $r > 0$, then $\mathbb{E}(rX) = \int_0^\infty P(rX > x)\, dx - \int_{-\infty}^0 P(rX < x)\, dx = \int_0^\infty P(X > y)\, r dy - \int_{-\infty}^0 P(X < y)\, r dy = r\mathbb{E}(X)$. Similarly, for $r < 0$, $\mathbb{E}(rX) = r\mathbb{E}(X)$.
(c) Note that $P(X+Y > t) = P(X > t, X+Y > t) + P(X \leq t < X+Y) = P(X > t) - P(X+Y \leq t < X) + P(X \leq t < X+Y)$. Then, $\mathbb{E}(X+Y) = \int_0^\infty P(X+Y > t)\, dt - \int_{-\infty}^0 P(X+Y < t)\, dt = \int_0^\infty P(X > t)\, dt - \int_{-\infty}^0 P(X < t)\, dt + \int_{-\infty}^\infty P(X < t < X+Y)\, dt - P(X+Y < t < X)\, dt = \mathbb{E}(X) + \int_0^\infty P(Y > t)\, dt - \int_{-\infty}^0 P(Y < t)\, dt = \mathbb{E}(X) + \mathbb{E}(Y)$. Note that the equality sign inside of the probability in the term $\int_0^\infty P(X \leq t < X+Y)\, dt$ can be dropped because $\int_0^\infty P(X = t)\, dt = 0$. Similarly equality was dropped in some terms. Also note that $\int_{-\infty}^\infty P(X < t < X+Y)\, dt = \int_{-\infty}^\infty \int 1((t,\omega) : X(\omega) < t < X(\omega)+Y(\omega))\, dP\, dt = \int \int_{-\infty}^\infty 1((t,\omega) : X(\omega) < t < X(\omega)+Y(\omega))\, dt\, dP = \int \max(0, Y(\omega))\, dP = \int_0^\infty \int 1(0 < t < Y(\omega))\, dP\, dt = \int_0^\infty P(Y > t)\, dt$ and $\int_{-\infty}^\infty P(X+Y < t < X)\, dt = \int_{-\infty}^0 P(Y < t)\, dt$.
(d) It can be obtained by combining (a)-(c), that is, $\mathbb{E}(Y) - \mathbb{E}(X) = \mathbb{E}(Y) + \mathbb{E}(-X) = \mathbb{E}(Y - X) \geq 0$.
(e) $|\mathbb{E}(X)| = |\int_0^\infty P(X > x)\, dx - \int_{-\infty}^0 P(X < x)\, dx| \leq\leq |\int_0^\infty P(X > x)\, dx| + |\int_{-\infty}^0 P(X < x)\, dx| = \int_0^\infty P(X > x)\, dx + \int_{-\infty}^0 P(X < x)\, dx = \int_0^\infty P(X > x)\, dx + \int_0^\infty P(-X > x)\, dx = \int_0^\infty P(|X| > x)\, dx = \mathbb{E}|X|$. ☐

Note. Let $L_+ = L_+(P)$ be the collection of non-negative random variables having finite expectation, $L^1 = L^1(P)$ be the collection of random variables having finite expectation and $L^p = L^p(P)$ be the collection of random variables having p-th absolute moment, that is $\mathbb{E}|X|^p < \infty$ for $p > 0$.

Exercise* 6. Show that if $X_n \in L_+$ with $X_n \nearrow X$ of which expectation is finite, i.e., $\lim_{n\to\infty} \mathbb{E}(X_n) < \infty$ then $X \in L_+$ and $\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n)$.

1.4 Inequalities

Theorem 7 (Chebyshev's inequality)**.** If $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$, then for any $\alpha > 0$,

$$P(|X - \mu| \geq \alpha\sigma) \leq 1/\alpha^2.$$

Proof. $P(|X-\mu| \geq \alpha\sigma) = \mathbb{E}(1(|X-\mu| \geq \alpha\sigma)) \leq \mathbb{E}((X-\mu)^2/(\alpha\sigma)^2) = 1/\alpha^2$. ☐

Example 7 (Weak law of large numbers)**.** Let $X_1, \ldots, X_n$ be an i.i.d. (independent and identically distributed) from F with mean μ and finite variance σ^2. Then the sample mean $\overline{X}_n = (X_1 + \cdots + X_n)/n$ has mean $\mathbb{E}(\overline{X}_n) = \mu$ and variance $\mathbb{V}\mathrm{ar}(\overline{X}_n) = \mathbb{V}\mathrm{ar}(X_1)/n = \sigma^2/n$. By apply Chebyshev's inequality, we get, for any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| > \epsilon) = P(|\overline{X}_n - \mu| > (\epsilon/\sigma)\sigma) \leq \mathbb{V}\mathrm{ar}(\overline{X}_n)/(\epsilon/\sigma)^2 = \sigma^2/(n\epsilon^2) \to 0.$$

In other words, $\overline{X}_n$ converges to the mean μ in probability as n increases.

Exercise 7. Show that Chebyshev's inequality holds the 'equality' for some cases.

Theorem 8 (Markov's inequality)**.** For $X \geq 0$ a.s. and any $\alpha > 0$,

$$P(X \geq \alpha) \leq \mathbb{E}(X)/\alpha.$$

Proof. $P(X \geq \alpha) = \mathbb{E}(1(X \geq \alpha)) \leq \mathbb{E}(X/\alpha)) = \mathbb{E}(X)/\alpha$. $\square$

Exercise 8 (Cantelli's inequality)**.** Show that for $\alpha > 0$, $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$,

$$P(|X - \mu| > \alpha) \leq \frac{2\sigma^2}{\sigma^2 + \alpha^2}.$$

When is the Cantelli's inequality more accurate than Chebyshev's inequality?

Definition 12. A function φ is *convex* if and only if

$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \geq \varphi(\lambda x + (1-\lambda)y)$$

for all $\lambda \in [0,1]$ and $x, y \in \mathbb{R}$. A function f is *concave* if $-f$ is convex.

Example 8. For $x \geq 0$, functions $\varphi_p(x) = x^p$ is convex if $p \geq 1$ and concave if $p \leq 1$. The logarithm $\log(x)$ is concave while $\exp(x)$ is convex. Trigonometric functions $\sin(x)$, $\cos(x)$ and $\tan(x)$ are neither convex nor concave.

Exercise 9. If a function φ is both convex and concave, then it is a line of the form $\varphi(x) = ax + b$ for some $a, b \in \mathbb{R}$.

Exercise 10. If f, g are convex and g is nondecreasing, then $g \circ f$ is also convex.

Exercise 11. If f is twice differentiable and $f''(x) \geq 0$, then f is convex.

Exercise 12. Let f be a convex function. Show that there exist $a, b \in \mathbb{R}$ such that $f(x) \geq ax + b$ for all $x \in \mathbb{R}$.

Let $S = \{(a,b) \in \mathbb{R}^2 : f(x) \geq ax + b \text{ for all } x \in \mathbb{R}\}$. Show that $f(x) = \sup_{(a,b)\in S}(ax + b)$.

Note. Roughly speaking, a convex function has the property that the function value of weighted average is not greater than the same weighted average of function values. This property can be generalized to random variables through the expectation.

Theorem 9 (Jensen's inequality)**.** Let φ be a convex function and X be a random variable. Then,

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$$

Proof. Let $S = \{(a,b) \in \mathbb{R}^2 : \varphi(x) \geq ax + b \text{ for all } x \in \mathbb{R}\}$. For any $(a,b) \in S$, $\mathbb{E}(\varphi(X)) \geq \mathbb{E}(aX + b) = a\mathbb{E}(X) + b$. Hence, $\mathbb{E}(\varphi(X)) \geq \sup_{(a,b)\in S}(a\mathbb{E}(X) + b) = \varphi(\mathbb{E}(X))$. $\square$

Example 9 (Lyapounov's inequality)**.** If $\mathbb{E}(|X|^p) < \infty$ for $p > 0$, then $\mathbb{E}(|X|^q) \leq \{\mathbb{E}(|X|^p)\}^{q/p}$ for all $0 < q \leq p$. Note $\varphi(x) = x^{p/q}$ is a convex function. Hence,

$$\mathbb{E}(|X|^p) = \mathbb{E}(\varphi(|X|^q)) \geq \varphi(\mathbb{E}(|X|^q)) = \{\mathbb{E}(|X|^q)\}^{p/q}.$$

Exercise 13. Show that the Young's inequality, that is, for $p, q > 1$ with $1/p + 1/q = 1$, two nonnegative real numbers $x, y \geq 0$ satisfies

$$xy \leq x^p/p + y^q/q.$$

Theorem 10 (Hölder's inequality)**.** Define $||f||_p = (\mathbb{E}(|f|^p))^{1/p}$ for $p \geq 1$. For $p, q > 1$ with $1/p + 1/q = 1$,

$$\mathbb{E}(|fg|) \leq ||f||_p ||g||_q.$$

Proof. If $||f||_p = 0$ or $||g||_q = 0$, then $|fg| = 0$ a.e. So the theorem holds. If both $||f||_p$, $||g||_q$ are positive and finite, then two can be normalized, i.e., $\tilde{f} = f/||f||_p$ and $\tilde{g} = g/||g||_q$. Using Young's inequality, we get $\frac{1}{||f||_p||g||_q}\mathbb{E}(|fg|) = \mathbb{E}(|\tilde{f}\tilde{g}|) \leq \mathbb{E}\Big(\frac{|\tilde{f}|^p}{p} + \frac{|\tilde{g}|^q}{q}\Big) = \frac{1}{p} + \frac{1}{q} = 1.$ $\square$

Note. The special case $p = q = 2$ is called the Cauchy-Schwartz inequality.

Exercise 14. Prove the following Minkowski's inequality.

Theorem 11 (Minkowski's inequality)**.** For $p \geq 1$, $||f + g||_p \leq ||f||_p + ||g||_p$.

1.5 Mode of Convergence

Definition 13. Let X, X_n be random variables. A sequence X_n converges in *distribution* to X if $P(X_n < t) \to P(X < t)$ for all $t \in \mathbb{R}$ with $P(X = t) = 0$. A sequence X_n converges in *probability* to X if for any $\epsilon > 0$ $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. A sequence X_n converges in *almost surely* to X if $P(\limsup_n |X_n - X| = 0) = 1$. A sequence X_n converges in L^p to X if $\mathbb{E}|X_n - X|^p \to 0$ as $n \to \infty$.

Example 10. Let $U \sim \text{Uniform}(0,1)$. Let $X_n = 1(0 < U < 1/n)$. Then $X_n \to 0$ in probability, a.s. and in L^p for $p > 0$. Let $Y_n = n1(0 < U < 1/n)$. Then $Y_n \to 0$ in probability, a.s. but not in L^p for $p > 0$. Let $Z_n = 1(a_n < U < b_n)$ where $n = 2^k + m$ with $0 \leq m < 2^k$, $a_n = m/2^k$ and $b_n = (m+1)/2^k$. Then $Z_n \to 0$ in probability and in L^p for $p > 1$ but not a.s. because $\limsup_{n\to\infty} Z_n = 1$.

Theorem 12. (a) $X_n \to X$ a.s. $\Longrightarrow X_n \to X$ in measure.
(b) $X_n \to X$ in $L^p \Longrightarrow X_n \to X$ in probability.

Proof. (a) Fix $\epsilon > 0$. Let $B_n^\epsilon = \{|X_n - X| > \epsilon\}$ and $A_n^\epsilon = \cup_{m \geq n} B_m = \{\sup_{m\geq n} |X_m - X| > \epsilon\}$. Note that $A^\epsilon = \{|\lim_n X_n - X| > \epsilon\} = \cap_{n=1}^\infty A_n^\epsilon$. It is easy to see that $A_n^\epsilon \searrow A^\epsilon$. By the almost sure convergence, $P(A^\epsilon) = 0$. Then the continuity from above implies $P(B_n^\epsilon) \leq P(A_n^\epsilon) \to \mu(A^\epsilon) = 0$.

(b) For $\epsilon > 0$. Let $A_n = \{|X_n - X| > \epsilon\}$. Then $P(A_n) = \mathbb{E}(1_{A_n}) \leq \mathbb{E}[(|X_n - X|/\epsilon)^p] = \epsilon^{-p}\mathbb{E}(|X_n - X|^p) \to 0$. $\square$

Theorem 13. A sequence x_n of real numbers converges to x if and only if for any subsequence n_k there exists a further subsequence n_{k_l} such that $x_{n_{k_l}}$ converges to x.

Proof. Sufficiency ($\Longrightarrow$) is obvious. Necessity ($\Longleftarrow$). If x_n does not converge to x, then the sequence $|x_n - x|$ does not converge to 0. Then there exists a $\delta > 0$ and a subsequence n_k such that $|x_{n_k} - x| > \delta$. However, from the assumption, there exists a further sequence n_{k_l} such that $x_{n_{k_l}} \to x$, i.e., $|x_{n_{k_l}} - x| \to 0$. Two statements contradicts. Thus x_n converges to x. $\square$

Theorem 14. A sequence of random variables X_n converges to X in probability if and only if for any subsequence n_k there exists a further subsequence n_{k_l} such that $X_{n_{k_l}}$ converges to X a.s.

Proof. Necessity part ($\Longleftarrow$) is direct from Theorem 12. Sufficiency ($\Longrightarrow$). We prove if $f_n \to f$ in probability then there exists a subsequence n_k such that $X_{n_k} \to X$ a.s. Let $n_0 = 0$. Sequentially take $n_k > n_{k-1}$ such that $P(|X_n - X| > 2^{-k}\}) < 2^{-k}$ for all $n \geq n_k$. Then $\{\lim_{k\to\infty} X_{n_k}(x) \neq X(x)\} \subset \cap_{m=1}^\infty \cup_{k\geq m} B_k$ where $B_k = \{|X_{n_k} - X| > 2^{-k}\}$. So we get

$$P(\{\lim_{k\to\infty} X_{n_k} \neq X\}) \leq \lim_{m\to\infty} P(\cup_{k\geq m} B_k) \leq \lim_{m\to\infty} \sum_{k\geq m} P(B_k) \leq \lim_{m\to\infty} \sum_{k\geq m} 2^{-k} = \lim_{m\to\infty} 2^{1-m} = 0.$$

Hence the theorem follows. $\square$

Theorem 15 (Fatou's lemma)**.** Let X_n be non-negative random variables.

$$\mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

Proof. If $\liminf_{n\to\infty} \mathbb{E}(X_n) = \infty$, then the theorem is obvious. So assume that $\liminf_{n\to\infty} \mathbb{E}(X_n) < \infty$. Let $Y_n = \inf_{m\geq n} X_m$ and $Y = \liminf_{n\to\infty} X_n$. Then $Y_n \nearrow Y$. Hence $\mathbb{E}(Y) = \lim_{n\to\infty} \mathbb{E}(Y_n)$. Since $Y_n \leq X_m$ for any $m \geq n$, $\mathbb{E}(Y_n) \leq \inf_{m\geq n} \mathbb{E}(X_m)$. Therefore

$$\mathbb{E}(\liminf_{n\to\infty} X_n) = \mathbb{E}(Y) = \lim_{n\to\infty} \mathbb{E}(Y_n) \leq \lim_{n\to\infty} \inf_{m\geq n} \mathbb{E}(X_m) = \liminf_{n\to\infty} \mathbb{E}(X_n).$$

$\square$

Theorem 16 (Monotone Convergence Theorem)**.** If $X_n \geq 0$ and $X_n \nearrow X$, then $\mathbb{E}(X_n) \nearrow \mathbb{E}(X)$.

Proof. See Exercise 6. $\square$

Theorem 17 (Dominated Convergence Theorem)**.** If $X_n \to X$ a.s., $|X_n| \le Y$ for all n and Y has finite expectation, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

Proof. Note that $X_n + Y \ge 0$. Fatou's lemma implies

$$\liminf_{n\to\infty} \mathbb{E}(X_n + Y) \ge \mathbb{E}(\liminf_{n\to\infty}(X_n + Y)) = \mathbb{E}(X + Y).$$

By subtracting $\mathbb{E}(Y)$, we get

$$\liminf_{n\to\infty} \mathbb{E}(X_n) \ge \mathbb{E}(X).$$

Applying the last result to $-X_n$, we obtain

$$\limsup_{n\to\infty} \mathbb{E}(X_n) \le \mathbb{E}(X).$$

Hence the theorem applies. $\square$

Exercise 15. Show the next theorem.

Theorem 18 (Generalized Dominated Convergence Theorem)**.** If $X_n \to X$ a.s., $|X_n| \le Y_n$ for all n, $Y_n \to Y$ a.e. and $\mathbb{E}(Y_n) \to \mathbb{E}(Y) < \infty$, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

Exercise 16. Prove the dominated convergence theorem and the generalized dominated convergence theorem with $X_n \to X$ in probability.

1.5.1 Expectation

Definition 14. The variance and covariance are defined as

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$$
$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Two random variables X and Y are uncorrelated if $\mathrm{Cov}(X, Y) = 0$.

Note. Let X be a random variable, F_X be the distribution function of X and $P^X = P \circ X^{-1}$. The expectation of X can be denoted in several ways like

$$\mathbb{E}(X) = \int X\, dP = \int X\, dP^X = \int X(x)\, P(dx) = \int X(x)\, dF_X(x) = \int X\, dF_X = \int X(x)\, F_X(dx).$$

Theorem 19. Let $X, X_1, X_2, \ldots$ be random variables.
(a) $X_n \to X$ in probability if $X_n \to X$ a.s.
(b) $X_n \to X$ in probability if $X_n \to X$ in L^p
(c) if $X_n \to X$ in probability, then there exists a subsequence n_k such that $X_{n_k} \to X$ a.s.
(d) $X_n \to X$ in probability if and only if for any subsequence n_k there exists a further subsequence n_{k_l} such that $X_{n_{k_l}} \to X$ a.s.

Exercise 17. Show that $X_n \to X$ in L^p if and only if for any subsequence n_k there exists a further subsequence n_{k_l} such that $X_{n_{k_l}} \to X$ in L^p and a.s. Note. L^p is vector space equipped with a topology.

Exercise 18. (a) Show that $X_n + Y_n \to X + Y$, $X_n Y_n \to XY$ a.s. if $X_n \to X$, $Y_n \to Y$ a.s.
(b) Show that $X_n + Y_n \to X + Y$, $X_n Y_n \to XY$ in probability if $X_n \to X_n$, $Y_n \to Y$ in probability.

1.6 Independence

Definition 15. Two events $A, B \in \mathcal{F}$ are independent if $P(A \cap B) = P(A) \cdot P(B)$.
Two random variables X and Y are independent if $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ for all $A, B \in \mathcal{B}$.

Note. Two events $A, B \in \mathcal{F}$ are independent if and only if two indicator functions 1_A and 1_B are uncorrelated.

Example 11. Find events A, B, C are pairwise independent but are not (mutually) independent.
Let $\Omega = \{1, 2, 3, 4\}$, P is uniform on Ω, $A = \{1, 2\}, B = \{1, 3\}, C = \{1, 4\}$. Then $P(A \cap B \cap C) = P(\{1\}) = \frac{1}{4} \neq \frac{1}{8}$.
Let $X_i \sim^{i.i.d.}$ *Bernoulli*$(\frac{1}{2})$, $A = (X = Y), B = (X = Z), C = (Y = Z)$

The distribution function of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is defined as $F_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_{})$. Hence $X_1, \ldots, X_n$ are independent if and only if $F_{\mathbf{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n)$.

Theorem 20. Let X and Y be two random variables. Two random variables X and Y are independent if and only if, for any measurable functions $f, g : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}|f(X)|, \mathbb{E}|g(Y)| < \infty$,

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

Proof. Necessity ($\Longleftarrow$). For any $A, B \in \mathcal{B}(\mathbb{R})$, $1_A, 1_B$ are measurable and $\mathbb{E}|1_A(X)|, \mathbb{E}|1_B(Y)| \leq 1 < \infty$. $P(X \in A, Y \in B) = \mathbb{E}[1_A(X)1_B(Y)] = \mathbb{E}[1_A(X)]\mathbb{E}[1_B(Y)] = P(X \in A)P(Y \in B)$. Hence, X and Y are independent.
Sufficiency ($\Longrightarrow$). Using Fubini's theorem,

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \int\int f(x)g(y)\ P^X(dx)\ P^Y(dy) = \int\int f(x)\ P^X(dx)\ g(y)\ P^Y(dy) = \int \mathbb{E}[f(X)]\ g(y)\ P^Y(dy) \\ &= \mathbb{E}[f(X)] \int g(y)\ P^Y(dy) = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].\end{aligned}$$

□

Example 12. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. But the converse is not true. A simple counterexample is $X \sim N(0, 1)$ and $Y = |X|$. Since X is symmetric and $\mathbb{E}|X| < \infty$, $\mathbb{E}(XY) = 0$ and $\mathbb{E}(X)\mathbb{E}(Y) = 0$ but X and Y are not independent obviously.

Example 13. We may restrict f, g as bounded measurable functions, for example, $f = 1_{(-\infty, x]}$ and $g = 1_{(-\infty, y]}$. Still X and Y are independent if and only if

$$F_{(X,Y)}(x, y) = \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)] = F_X(x)F_Y(y).$$

1.7 Law of Large Numbers

Theorem 21 (Weak Law of Large Numbers)**.** Let $X_1, \ldots, X_n$ be pair-wise uncorrelated random variables having the same mean μ and finite variance, that is, $\mathbb{E}(X_i) = \mu$, $\mathbb{V}\mathrm{ar}(X_i) \leq M$ for all i, for some $M > 0$. Then the sample mean $\overline{X}_n = (X_1 + \cdots + X_n)/n \to \mu$ in L^2 as well as in probability.

Proof. The mean of $\overline{X}_n$ is $\mathbb{E}(\overline{X}_n) = (\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n))/n = \mu$ and variance becomes

$$\mathbb{E}(|\overline{X}_n - \mu|)^2 = \mathbb{V}\mathrm{ar}(\overline{X}_n) = \frac{1}{n^2}(\mathbb{V}\mathrm{ar}(X_1) + \cdots + \mathbb{V}\mathrm{ar}(X_n)) \leq \frac{M}{n} \to 0.$$

Hence $\overline{X}_n \to \mu$ in L_2 and in probability. □

Lemma 22 (Cesàro's Sum)**.** Let x_n be a sequence of real numbers that converges to x and v_n be a nondecreasing sequence diverging to infinity with $v_0 = 0$, that is, $x_n \to x$ and $v_n \nearrow \infty$. Then, $v_n^{-1} \sum_{k=1}^n (v_k - v_{k-1}) x_k \to x$.

Proof. For $\epsilon > 0$, there exists a number $N > 0$ such that $|x_n - x| < \epsilon$ for all $n \geq N$.

$$\begin{aligned}\Big|\frac{1}{v_n}\sum_{k=1}^n (v_k - v_{k-1})x_k - x\Big| = \Big|\frac{1}{v_n}\sum_{k=1}^n (v_k - v_{k-1})(x_k - x)\Big| &\leq \frac{1}{v_n}\sum_{k=1}^n (v_k - v_{k-1})|x_k - x| \\ &\leq \frac{1}{v_n}\Big[\sum_{k=1}^N (v_k - v_{k-1})|x_k - x|\Big] + \frac{1}{v_n}(v_n - v_N)\epsilon \to \epsilon.\end{aligned}$$

The lemma follows because ϵ is arbitrary. □

Example 14. If $x_n \to x$, then $(x_1 + \cdots + x_n)/n \to x$ as $n \to \infty$ by applying the above lemma with $v_n = n$.

Lemma 23 (Kronecker's lemma)**.** If $a_n \nearrow \infty$ and $\sum_n x_n/a_n$ converges, then $(x_1 + \cdots + x_n)/a_n \to 0$.

Proof. Let $a_0 = b_0 = 0$ and $b_n = x_1/a_1 + \cdots + x_n/a_n$. Then $x_n = a_n(b_n - b_{n-1})$ and

$$\frac{1}{a_n}\sum_{m=1}^n x_m = \frac{1}{a_n}\sum_{m=1}^n a_m(b_m - b_{m-1}) = b_n - \frac{1}{a_n}\sum_{m=1}^n (a_m - a_{m-1})b_{m-1} \to 0$$

In the last convergence, Cesàro's sum is applied. □

Theorem 24 (Weak Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be i.i.d. with finite mean μ. Then, $\overline{X}_n = (X_1 + \cdots + X_n)/n \to \mu$ in probability.

Proof. Let $T_n = \sum_{i=1}^n X_i 1(|X_i| \leq n)$ and $\mu_n = \mathbb{E}(T_n)/n$. Fix $\epsilon > 0$. Then

$$\begin{aligned}&P(|S_n/n - \mu| > \epsilon) \leq P(S_n \neq T_n) + P(|T_n/n - \mu| > \epsilon) \\ &\leq P(S_n \neq T_n) + P(|\mu_n - \mu| > \epsilon/2) + P(|T_n/n - \mu_n| > \epsilon/2) = I_1 + I_2 + I_3. \qquad (1)\end{aligned}$$

The first term I_1 in (1) is bounded by

$$I_1 = P(S_n \neq T_n) \leq P(\cup_{i=1}^n \{|X_i| > n\}) \leq nP(|X_1| > n) \leq \mathbb{E}[|X_1| 1_{|X_1|>n}] \to 0.$$

In the last convergence, dominated convergence theorem was used. Note $\mathbb{E}(X_n 1(|X_n| < n)) \to \mu$ by the dominated convergence theorem. Using Cesàro's sum, $\mu_n \to \mu$. Hence $I_2 = P(|\mu_n - \mu| > \epsilon/2) = 0$ for sufficiently large n. For the last part, we will apply Chebyshev's inequality. So the variance of T_n/n is required. Note $\mathbb{E}(T_n/n) = \mu_n$ and

$$\mathbb{E}\Big(\frac{T_n}{n} - \mu_n\Big)^2 = \frac{1}{n}\mathbb{V}\mathrm{ar}(X_1 1(|X_1| \leq n)) \leq \frac{1}{n}\mathbb{E}[X_1^2 1(|X_1| \leq n)] = \mathbb{E}[|X_1| \cdot |X_1/n| 1(|X_1| \leq n)] \to 0.$$

In the last convergence, the dominated convergence theorem was used, that is, $|X_1| \cdot |X_1/n| 1(|X_1| \leq n) \to 0$, $|X_1| \cdot |X_1/n| 1(|X_1| \leq n) \leq |X_1|$ and $\mathbb{E}|X_1| < \infty$. Hence,

$$I_3 = P(|T_n/n - \mu_n| > \epsilon/2) \leq \mathbb{V}\mathrm{ar}(T_n/n)/(\epsilon/2)^2 \to 0.$$

It completes the proof. □

Theorem* 25 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be pairwise independent and identically distributed random variables with $\mathbb{E}|X_n| < \infty$. Let $\mathbb{E}X_n = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.

Example 15 (Empirical Distribution)**.** Let $X_1, X_2, \ldots$ be i.i.d from F. Then the sample empirical distribution function $F_n(x) = n^{-1}\sum_{i=1}^n 1(X_i \leq x)$ converges to $F(x)$ almost surely.

Exercise 19. Let $X_1, X_2, \ldots$ be i.i.d with $\mathbb{E}X_n = +\infty$. Let $S_n = X_1 + \cdots + X_n$. Show that $S_n/n \to \infty$.

1.8 Conditional Expectation

Example 16. If $P(A) > 0$, the conditional probability given A is $P(B \mid A) = P(B \cap A)/P(A)$. If $P(A) = 0$, it is difficult to define a conditional probability given A.

Definition 16. Let X, Y be two random variables having a joint density. The conditional probability $P(X \in A \mid Y = y) = \int_A \mathrm{pdf}_{X,Y}(x, y)\, dx / \mathrm{pdf}_Y(y)$.

Theorem 26 (Properties of conditional expectation)**.** Let X_n, X, Y, Z, W be random variables.
(a) $Y = \mathbb{E}(X \mid W)$ has finite expectation if $\mathbb{E}|X| < \infty$.
(b) (Uniqueness) If both Y and Z are versions of $\mathbb{E}(X \mid W)$, then $Y = Z$ a.s.
(c) $\mathbb{E}(aX + bY \mid W) = a\mathbb{E}(X \mid W) + b\mathbb{E}(Y \mid W)$.
(d) $X \le Y$ implies $\mathbb{E}(X \mid W) \le \mathbb{E}(Y \mid W)$.
(e) If $X_n \ge 0$ and $X_n \nearrow X$ with $\mathbb{E}X < \infty$, then $\mathbb{E}(X_n \mid W) \nearrow \mathbb{E}(X \mid W)$.
(f) If X is a function of W, then $\mathbb{E}(X \mid W) = X$.
(g) If $V = g(W)$ and $\mathbb{E}(X \mid W)$ is a function of V, then $\mathbb{E}(X \mid W) = \mathbb{E}(X \mid V)$.
(h) If $V = g(W)$, then $\mathbb{E}(\mathbb{E}(X \mid V) \mid W) = \mathbb{E}(X \mid V)$ and $\mathbb{E}(\mathbb{E}(X \mid W) \mid V) = \mathbb{E}(X \mid V)$.
(i) If X is a function of W, $\mathbb{E}|Y|, \mathbb{E}|XY| < \infty$, then $\mathbb{E}(XY \mid W) = X\mathbb{E}(Y \mid W)$.

Exercise 20. Show Markov's inequality, that is, for any $a > 0$, $P(|X| > a \mid W) \le a^{-1}\mathbb{E}(|X| \mid W)$.

Exercise 21. Let φ be a convex function. Show that if $\mathbb{E}|X|, \mathbb{E}|\varphi(X)| < \infty$, then $\varphi(E(X \mid W)) \le \mathbb{E}(\varphi(X) \mid W)$.

Exercise 22. If $\mathbb{E}X^2 < \infty$, then $\mathbb{E}(X \mid W)$ is the W-measurable r.v. Y which minimizes $\mathbb{E}(X - Y)^2$.

Exercise 23. Show that $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(\mathbb{V}\mathrm{ar}(X \mid W)) + \mathbb{V}\mathrm{ar}(\mathbb{E}(X \mid W))$ where $\mathbb{V}\mathrm{ar}(X \mid W) = \mathbb{E}(X^2 \mid W) - [\mathbb{E}(X \mid W)]^2$.