STA347: Probability

Tianyu Du

October 28, 2019

Contents

1 Preliminaries

1.1 Random Processes

Definition 1.1. A **process**[1] W is a mechanism generating **outcomes** w from a sample space Ω. Any realized trail of process W can be denoted as a potentially infinite sequence in Ω:

$$W : w_1, w_2, \cdots, w_n, \cdots \tag{1.1}$$

Definition 1.2. A **random variable** (extended process), $X := g(W)$, can be constructed from a process W and a real-valued function $g : \Omega \to \mathbb{R}$.

Definition 1.3. Given a random variable $X = g(W)$, the **sample mean** (i.e. empirical expectation) of the first n trials from a sequence of realizations, $g(w_1), \cdots, g(w_n), \cdots$, is defined to be

$$\hat{\mathbb{E}}_n g(W) := \frac{\sum_{i=1}^n g(w_i)}{n} \tag{1.2}$$

Definition 1.4. A process W is said to be a **random process/variable** if it satisfies the *empirical law of large numbers*, in that, $\forall g \in \mathbb{R}^\Omega$:

(i) *stability*: $(\hat{\mathbb{E}}_n g(W))_{n\in\mathbb{N}}$ converges;

(ii) *Invariance*: $\forall\ (w_n)_{n\in\mathbb{N}} \subseteq \Omega$, the limits of $(\hat{\mathbb{E}}_n g(W))_{n\in\mathbb{N}}$ are the same.

Definition 1.5. Let W be a random process and $g \in \mathbb{R}^\Omega$, the **expected value** of $g(W)$ is defined as

$$\mathbb{E}g(W) := \lim_{n\to\infty} \hat{\mathbb{E}}_n g(W) \tag{1.3}$$

the limit is well-defined given ELLN.

Definition 1.6. Let W be a random process. For every $A \subseteq \Omega$, take $g := I_A \in \mathbb{R}^\Omega$, the **empirical relative frequencies** (i.e. empirical probability) is defined as

$$\hat{P}(W \in A) := \hat{\mathbb{E}}_n I_A(W) \tag{1.4}$$

Given ELLN, the limit is well-defined, then the **probability** is defined to be the limit:

$$P(W \in A) := \lim_{n\to\infty} \hat{P}(W \in A) \tag{1.5}$$

Remark 1.1. The notation of expected values and probabilities on W is well-defined only when W satisfies the empirical law of large numbers, that is, W is a random process.
Given W defined on Ω satisfies ELLN, the behaviour of W can be fully characterized by its **probability distribution**.

$$W \sim P_W \text{ on } \Omega \tag{1.6}$$

1.2 Indicator Functions and Set Operations

Proposition 1.1. The **indicator map** $I : \mathcal{P}(\Omega) \to \{0,1\}^\Omega$ is bijective.

[1] This is just a process, not necessarily a random process.

Proof. Injective: Let $f, g \in \{0,1\}^\Omega$, show the corresponding sets A_f, A_g are the same.
Let $w \in A_f$, then $f(w) = 1 = g(w)$, which implies $w \in A_g$. Therefore, $A_f \subseteq A_g$.
The other direction is similar.
Surjective: Let $f \in \{0,1\}^\Omega$, then one can construct $A_f := \{w \in \Omega : f(w) = 1\}$ so that $I(A_f) = f$. ■

Proposition 1.2. Let W be a random variable defined on outcome space Ω, and let (A_n) be a collection[2], then

$$I \bigcap_{n=1}^{\infty} A_n = \inf_{n=1}^{\infty} IA_n \tag{1.7}$$

$$I \bigcup_{n=1}^{\infty} A_n = \sup_{n=1}^{\infty} IA_n \tag{1.8}$$

1.3 Real Analysis

Definition 1.7. A sequence (x_n) is **Cauchy** if $\sup_{i,j\geq n} |x_i - x_j| \to 0$ as $n \to \infty$.

Proposition 1.3.

$$\sup_{i,j\geq n} |x_i - x_j| = \sup_{i\geq n} x_i - \inf_{j\geq n} x_j \tag{1.9}$$

Theorem 1.1. Let (x_n) be a sequence in a Hilbert space, then

$$(x_n) \to x \iff \liminf x_n = \limsup x_n \tag{1.10}$$

Proposition 1.4 (Corollary of Order Limit Theorem)**.** It is evident that

$$\forall n \in \mathbb{N}, \inf_{i\geq n} x_i \leq \sup_{i\geq n} x_i \tag{1.11}$$

therefore,

$$\liminf x_n \leq \limsup x_n \tag{1.12}$$

2 Distributions

2.1 Construction of Uniform Distributions

Definition 2.1. For any $n \in \mathbb{N}$, a random variable X is said to have a **(finite discrete) uniform distribution** on the sample space $\Omega = \{1, \cdots, n\}$ if

$$P(X = k) = \frac{1}{n} \ \forall k \in \{1, \cdots, n\} \tag{2.1}$$

Denoted as $X \sim unif\{1, \cdots, n\}$.

Definition 2.2. A random variable U wit $\Omega = [0,1]$ is said to follow a **(continuous) standard uniform** if

$$P(U \leq u) = u \ \forall u \in [0,1] \tag{2.2}$$

[2]The lecture note presents it as a countable set, indeed, it can be an arbitrary set.

Denoted as $U \sim unif[0,1]$.

Construction of Continuous Uniform Let Y be a random variable on $\{0, \cdots, 9\}$. For each sequence of realizations of Y, $(Y_i)_{i\in\mathbb{N}}$, one can construct $U \in [0,1]$ using the following decimal expansion:

$$U := \sum_{i=1}^{\infty} \frac{Y_i}{10^i} \tag{2.3}$$

Proposition 2.1. It is evident that each finite decimal of Y, $(Y_1, \cdots, Y_n)$, partitions the unitary into 10^n small intervals

$$P(0.Y_1 \cdots Y_n \leq U \leq 0.Y_1 \cdots Y_n + 10^{-n}) = \frac{1}{10^n} \tag{2.4}$$

and

$$P(0 \leq U \leq 0.Y_1 \cdots Y_n) = \frac{Y_1 Y_2 \cdots Y_n}{10^n} = 0.Y_1 \cdots Y_n \tag{2.5}$$

Proposition 2.2. $P(a \leq U \leq b) = b - a$.

Corollary 2.1. $P(U = u) = P(u \leq U \leq u) = u - u = 0$.

Proposition 2.3. Let $U \sim unif[0,1]$ and $V := 1 - U$, then $V \overset{d}{=} U$.

Proof.

$$P(V \leq u) = P(1 - U \leq u) = P(U \geq 1 - u) \tag{2.6}$$

$$= 1 - P(U \leq 1 - u) = 1 - 1 + u = u \tag{2.7}$$

■

Theorem 2.1 (The Fundamental Theorem Applied Probability)**.** If $U = \sum_{n=1}^{\infty} Z_i p^{-i}$, then the following are equivalent:

(i) $U \sim unif[0,1]$;

(ii) $Z_i \overset{i.i.d.}{\sim} Z \overset{d}{=} unif\{0, \cdots, p-1\}$ for some $p \geq 2$.

Proof. The result should be evident by construction of uniform distribution, (ii) is simply the base p expansion of real numbers, which is equivalent to the decimal expansion. ■

Definition 2.3. Two random processes W_1, W_2 defined on Ω are **identically distributed**, $W_1 \overset{d}{=} W_1$ if

$$\mathbb{E}g(W_1) = \mathbb{E}g(W_2) \quad \forall g : \Omega \to \mathbb{R} \tag{2.8}$$

Theorem 2.2 (Invariance I)**.** If $X \overset{d}{=} Y$, then

$$\varphi(X) \overset{d}{=} \varphi(Y) \quad \forall \varphi : \Omega \to \mathcal{X} \tag{2.9}$$

Note that $\varphi(X)$ and $\varepsilon(Y)$ are random variable defined on sample space $\mathcal{X}$.

Proof. Let $h : \mathcal{X} \to \mathbb{R}$, note that $h \circ \varepsilon : \Omega \to \mathbb{R}$. It is evident that $\mathbb{E}h \circ \varphi(X) = \mathbb{E}h \circ \varphi(Y)$. ■

Theorem 2.3 (Invariance II)**.**

$$W_1 \overset{d}{=} W_2 \iff g(W_1) \overset{d}{=} g(W_2) \quad \forall g : \Omega \to \mathbb{R} \tag{2.10}$$

Proof. ($\implies$) The sufficient direction is direct by previous theorem.
($\impliedby$) Suppose $g(W_1) \overset{d}{=} g(W_2) \quad \forall g : \Omega \to \mathbb{R}$. The proof is immediate by applying identity mapping in the definition of $g(W_1) \overset{d}{=} g(W_2)$, which implies $\mathbb{E}g(W_1) = \mathbb{E}g(W_2)$ for any function $g : \Omega \to \mathbb{R}$. ■

Corollary 2.2. Let $A \overset{d}{=} B$, for each $A \subseteq \Omega$, take $g = I_A$. Then,

$$P(X \in A) = \mathbb{E}I_A(X) = \mathbb{E}I_A(Y) = P(Y \in A) \tag{2.11}$$

2.2 Constructing Other Distributions

Definition 2.4. Let $Z = -\ln U$ with $u \sim unif[0, 1]$. Then Z is said to follow **exponential distribution** such that

$$P(s \leq Z \leq t) = e^{-s} - e^{-t} \tag{2.12}$$

The derivation of distribution is immediate from the monotone property of $-\ln$.

Definition 2.5. The random variable Z is said to have a **standard exponential distribution** on $\mathbb{R}_+$, denoted as $Z \sim exp(1)$ if

$$P(Z \leq z) = 1 - \exp(-z) \tag{2.13}$$

Definition 2.6. The random variable X is said to have a **scaled exponential distribution** with some $\theta > 0$, denoted as $X \sim exp(\theta)$ if

$$X \overset{d}{=} \theta Z \quad Z \sim exp(1) \tag{2.14}$$

Definition 2.7. For any real-valued random variable, X[3], the **distribution function** of X is defined as

$$F_X(x) := P(X \leq x) \ \forall x \in \mathbb{R} \tag{2.15}$$

Definition 2.8. A real-valued random variable, X, is said to be **absolutely continuous** with respective to length measure if there exists $f : \mathbb{R} \to \mathbb{R}_+$ such that

$$P(s < X \leq t) = \int_s^t f_X(x)\ dx \quad \forall s \leq t \tag{2.16}$$

Where f_X is defined as the **probability density function** of X.
Remark 1: A random variable is absolutely continuous if there exists an integrable density function f.
Remark 2: The density function f_X is not necessarily unique.

Definition 2.9. The **percentile/quantile function**, $x_p = g(p) : [0, 1] \to \mathbb{R}$, is defined so that

$$F(X \leq x_p) = p \tag{2.17}$$

[3]A real-valued random variable can be deemed as the composite of an arbitrary random variable W on Ω and a function $g : \Omega \to \mathbb{R}$

Remark 2.1. In any event, it is (or certainly should be) perfectly clear that each of the three methods outlined above will lead to the same basic result, each with its own probability density function, f, its own distribution function, F, and its own percentile function, g.

2.3 More on Expectation Operators

Remark 2.2. The **expectation** operator

$$\mathbb{E} : \mathcal{R} \to \mathbb{R} \cup \{\pm\infty\} \cup \{\text{DNE}\} \tag{2.18}$$

where $\mathcal{R}$ is the space of *real-valued* random processes.

Theorem 2.4. By the algebraic limit theorem, it is evident that for any pair of real-valued random process X, Y and scalar a, the expectation operator is linear:

(i) $\mathbb{E}(X+Y) + \mathbb{E}X + \mathbb{E}Y$;

(ii) $\mathbb{E}(aX) = a\mathbb{E}X$.

Proposition 2.4 (Expectation is Normed)**.** $\mathbb{E}c = c \ \forall c \in \mathbb{R}$.

Proposition 2.5. $X \geq 0 \implies \mathbb{E}X \geq 0$ by order limit theorem.

Theorem 2.5. Let W be a real-valued random variable. Then for any finite mutually disjoint set $(A_1, \cdots, A_n)$,

$$P(W \in \sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(W \in A_i) \tag{2.19}$$

That is, the probability measure is linear given mutual disjoining.

Proposition 2.6. Properties of probability measure:

(i) $I_\Omega(W) = 1 \implies P(W \in \Omega) = \mathbb{E}1 = 1$;

(ii) $0 \leq I_A(W) \leq 1 \implies 0 \leq P(W \in A) = \mathbb{E}I_A(W) \leq 1$.

2.4 Expected Value for an Arbitrary Discrete Distribution

Definition 2.10. A **finite scheme** or a **finite discrete distribution** can be written as

$$W \sim \begin{pmatrix} \omega_1 & \cdots & \omega_N \\ p_1 & \cdots & p_N \end{pmatrix} \tag{2.20}$$

with **probability mass function** (pmf)

$$P\left(W = \omega_i\right) = p_i \ s.t. \ \sum_{i=1}^{N} p_i = 1 \tag{2.21}$$

With vector notation

$$W \sim \begin{pmatrix} \omega \\ p \end{pmatrix} \ s.t. \ \omega \subseteq \Omega, \ \langle p, 1 \rangle = 1 \tag{2.22}$$

Proposition 2.7. The expected value of a finite discrete random variable is more or less obvious:

$$Eg(W) = \sum_{i=1}^{N} g(\omega_i) P(W = \omega_i) = \sum_{i=1}^{N} g(\omega_i) p_i \tag{2.23}$$

Proposition 2.8. A finite discrete random variable, $g(W)$, can be explicitly represented as a *finite linear combination of simple indicator functions*:

$$g(W) = \sum_{i=1}^{N} g(\omega_i) I(W = \omega_i) = \sum_{i=1}^{N} g(\omega_i) I_{\{\omega_i\}}(W) \tag{2.24}$$

2.5 Discrete Uniform Distributions

Proposition 2.9. Let $W \sim unif\{1, \cdots, n\}$, then

$$n + 1 - W \stackrel{d}{=} W \tag{2.25}$$
$$\implies (n + 1 - W)^2 \stackrel{d}{=} W^2 \tag{2.26}$$
$$\implies (n+1)^2 - 2(n+1)W + W^2 \stackrel{d}{=} W^2 \tag{2.27}$$
$$\implies \mathbb{E}[(n+1)^2 - 2(n+1)W + W^2] = \mathbb{E}[W^2] \tag{2.28}$$
$$\implies \mathbb{E}[W] = \frac{n+1}{2} \tag{2.29}$$

Proposition 2.10.

$$(n + 1 - W)^3 \stackrel{d}{=} W^3 \tag{2.30}$$
$$\implies 2\mathbb{E}[W^3] = (n+1)^3 - 3(n+1)^2\mathbb{E}[W] + 3(n+1)\mathbb{E}[W^2] \tag{2.31}$$
$$\implies 2\mathbb{E}[W^3] = (n+1)^3 - 3(n+1)^2\frac{n+1}{2} + 3(n+1)\mathbb{E}[W^2] \tag{2.32}$$
$$\implies 2\mathbb{E}[W^3] = -\frac{(n+1)^2}{2} + 3(n+1)\mathbb{E}[W^2] \tag{2.33}$$
$$\implies \mathbb{E}[W^3] = n(\mathbb{E}[W])^2 \tag{2.34}$$

Proposition 2.11. $\mathbb{E}[W^4]$. TODO

Definition 2.11. $W \sim unif\{1, \cdots, n\}$, then the *distance between* W^2 and $\mathbb{E}[W^2]$ is defined as

$$d(W^2, \mathbb{E}[W^2]) := \sqrt{\mathbb{E}[W^2 - \mathbb{E}[W])^2]^2} = \sqrt{\mathbb{V}[W^2]} = \sigma_{W^2} \tag{2.35}$$

Corollary 2.3 (Corollary of Jensen's Inequality)**.**

$$\mathbb{E}[W^2] \geq (\mathbb{E}[W])^2 \tag{2.36}$$

and equality holds if and only if

$$\mathbb{E}[(W - \mathbb{E}[W])^2] = 0 \tag{2.37}$$

which is equivalent to

$$P(W = \mathbb{E}[W]) = 1 \tag{2.38}$$

Proof.

$$\mathbb{V}[W] = \mathbb{E}[(W - \mathbb{E}[W])^2] \geq 0 \tag{2.39}$$

■

Theorem 2.6.

$$EU^k - E(U-1)^k = n^{k-1} \quad k \in \mathbb{N} \tag{2.40}$$

Proof. WLOG, let $U \sim unif\{1, \cdots, n\}$, it is evident that $U - 1 \sim unif\{0, \cdots, n-1\}$

$$EU^k - E(U-1)^k = \frac{\sum_{i=1}^{n} i^k - \sum_{i=0}^{n-1} i^k}{n} \tag{2.41}$$

$$= \frac{\sum_{i=1}^{n} i^k - \sum_{i=1}^{n-1} i^k}{n} \tag{2.42}$$

$$= \frac{n^k}{n} = n^{k-1} \tag{2.43}$$

■

2.6 Bernoulli Trials

Definition 2.12. A random variable Z is said to be a **Bernoulli trial** if

$$Z \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix} \tag{2.44}$$

with $p \in [0, 1]$, denoted as $Z \sim bern(p)$.

Remark 2.3. Let $A \subseteq \Omega$, the **Bernoulli trial** associated with A is defined to be the finite scheme distribution

$$Z = I(A) \sim \begin{pmatrix} 0 & 1 \\ 1 - P(A) & P(A) \end{pmatrix} \tag{2.45}$$

Proposition 2.12 (Invariance)**.** Given the outcome space of Bernoulli trials to be $\{0, 1\}$, $Z^s = Z$ for every $s \in \mathbb{N}$.
Remark: Z^2 and Z are indeed equal, which is a stronger statement than equal in distribution.

Proposition 2.13 (Negation)**.** $Z \sim bern(p) \quad \Leftrightarrow \quad 1 - Z \sim bern(q)$ where $p + q = 1$.

2.7 Binomial Distribution

Definition 2.13. Given $n \in \mathbb{N}$ and $0 \leq p \leq 1$, a random variable X is said to follow **binomial distribution** with n trails and chance p if

$$X \overset{d}{=} Z_1 + \cdots + Z_n \quad Z_i \overset{i.i.d.}{\sim} \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix} \tag{2.46}$$

Denoted as $X \sim bin(n, p)$.

Proposition 2.14. $X^2 \overset{d}{=} \left(\sum_{i=1}^{n} Z_i\right)^2$.

Proof. The statement follows immediately from the definition of Binomial distribution. ■

Corollary 2.4.

$$\mathbb{E}(X^2) = \mathbb{E}\left(\sum_{i=1}^{n} Z_i\right)^2 \tag{2.47}$$

$$Var(X) = Var\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n} Var(Z_i) \tag{2.48}$$

$$= \sum_{i=1}^{n} pq = npq \tag{2.49}$$

More generally, for any $f : \mathbb{R} \to \mathbb{R}$, $f(X) \overset{d}{=} f\left(\sum_{i=1}^{n} Z_i\right)^2$.

Proposition 2.15. Let $X \sim bin(m, p)$, $X \sim bin(n, p)$ and $X \perp Y$, then $X + Y \sim bin(m + n, p)$.

Proof. Note that the joint distribution of (X, Y) follows

$$\begin{pmatrix} X \\ Y \end{pmatrix} \overset{d}{=} \begin{pmatrix} Z_1 + \cdots + Z_m \\ Z_{m+1} + \cdots + Z_{m+n} \end{pmatrix} \tag{2.50}$$

then the result is immediate. ■

Example 2.1. Define $g(Z_1, \cdots, Z_n) := \sum_{i=1}^{n} a_i Z_i^i$, then

$$g(Z_1, \cdots, Z_n) \overset{d}{=} \sum_{i=1}^{n} a_i Z_i \tag{2.51}$$

$$\implies \mathbb{E}g(Z_1, \cdots, Z_n) = \sum_{i=1}^{n} a_i p \tag{2.52}$$

$$Var(g(Z_1, \cdots, Z_n)) = \sum_{i=1}^{n} a_i pq \tag{2.53}$$

2.8 Probability Mass Function

Definition 2.14. Let $X \sim bin(n, p)$, then for each $k \in \{0, \cdots, n\}$, the **probability mass function (pmf)** is defined as

$$P(X = k) := P(\sum_{i=1}^{n} Z_i = k) \tag{2.54}$$

Conversely, for every $k \notin \{0, \cdots, n\}$, $P(X = k) := 0$.

Theorem 2.7 (pmf for Binomial)**.** Note that $P(Z = z) = p^z q^{1-z}$ for $z \in \{0, 1\}$, then for every $(z_i)_{i=1}^n \in$

$\{0,1\}^n$, it is evident that

$$P((Z_i)_{i=1}^n = (z_i)_{i=1}^n) = P\left[\bigcap_{i=1}^n \{Z_i = z_i\}\right] \tag{2.55}$$

$$= \prod_{i=1}^n P(Z_i = z_i) \tag{2.56}$$

$$= \prod_{i=1}^n p^{z_i} q^{1-z_i} \tag{2.57}$$

$$= p^{\sum_{z_i}} q^{n-\sum z_i} \tag{2.58}$$

Define

$$C_k^n := \{(z_1, \cdots, z_n) \in \{0,1\}^n : \sum_{i=1}^k z_i = k\} \tag{2.59}$$

Then

$$P(X = k) = \sum_{(z_1,\cdots,z_n)\in C_k^n} P[(Z_1, \cdots, Z_n) = (z_1, \cdots, z_n)] \tag{2.60}$$

$$= \sum_{(z_1,\cdots,z_n)\in C_k^n} p^k q^{n-k} \tag{2.61}$$

$$= |C_k^n| p^k q^{n-k} \tag{2.62}$$

Define $\binom{n}{k} := |C_k^n|$.

Definition 2.15. Given $k, n \in \mathbb{N}$ such that $k < n$, the **descending operator** $n^{(k)}$ is defined as

$$n^{(k)} := n(n-1)(n-2)\cdots(n-(r-1)) \tag{2.63}$$

Proposition 2.16. Let $X \sim bin(n,p)$, then

$$\mathbb{E}X^{(r)} = \begin{cases} n^{(r)} p^r & 0 \le r \le w \\ 0 & 0 \ge n+1 \end{cases} \tag{2.64}$$

2.9 Standard Uniform Distribution

Theorem 2.8. Let $U \sim unif[0,1]$, then

$$EU^k = \frac{1}{k+1}, \quad k \ge 0 \tag{2.65}$$

Proof. Given the outcome space of standard uniform, $P(U \leq t) = F(t) = t$ for every $t \in [0,1]$.

$$EU^k = \int_0^1 F'(t)t^k \, dt \tag{2.66}$$

$$= \int_0^1 t^k \, dt \tag{2.67}$$

$$= \frac{1}{k+1} t^{k+1} \Big|_0^1 \tag{2.68}$$

$$= \frac{1}{k+1} \tag{2.69}$$

■

2.10 Scaled Exponential Distribution

2.11 Gamma Distribution

3 Distribution Functions in General

3.1 Probability

Definition 3.1. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ consisting of **sample space**, **event space**, and a **probability** satisfying the following properties:
Properties of $\mathcal{F}$:

(i) $\Omega \in \mathcal{F}$;

(ii) Closed under complement;

(iii) Closed under countable union.

Properties of $P : \mathcal{F} \to [0,1]$:

(i) σ-additivity;

(ii) Non-negativity: $P(A) \geq 0$ for every $A \in \mathcal{F}$;

(iii) Normality: $P(\Omega) = 1$.

Definition 3.2. A probability measure $P : \mathcal{F} \to [0,1]$ is said to be **σ-additive** if for any countable mutually disjoint events (A_n),

$$P(\sum_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} P(A_n) \tag{3.1}$$

Definition 3.3. Let (A_n) be a sequence of subsets of Ω, let $A \subseteq \Omega$. Then $(A_n) \to A$ if and only if $(I(A_n))$ point-wise converges to $I(A)$.

Definition 3.4. A probability measure P is **continuous** if for any convergent sequence $(A_n) \to A$ in $\mathcal{F}$, $P(A_n) \to P(A)$.
Remark: the continuity of P is only defined through the sequential definition, there is δ-ε definition for it.

Remark 3.1. The *constructor of indicator function* is a bijection between $\mathcal{P}(\Omega)$ and function space $\{0,1\}^\Omega$.

Proposition 3.1 (Set Theoretic Limits)**.** A sequence of sets (A_n) converges if and only if that $I(A_n)$ converges, which is equivalent to $\lim_{n\to\infty} I(A_n) = \limsup I(A_n) = \liminf I(A_n)$:

$$\lim_{n\to\infty} I(A_n) = \lim_{n=1}^{\infty} \sup_{k\geq n} I(A_k) \tag{3.2}$$

$$= \inf_{n=1}^{\infty} \sup_{k\geq n} I(A_k) \tag{3.3}$$

$$= \inf_{n=1}^{\infty} I(\bigcup_{k\geq n} A_k) \tag{3.4}$$

$$= I(\bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k) \tag{3.5}$$

Similarly,

$$\lim_{n\to\infty} I(A_n) = \lim_{n=1}^{\infty} \inf_{k\geq n} I(A_k) \tag{3.6}$$

$$= \sup_{n=1}^{\infty} \inf_{k\geq n} I(A_k) \tag{3.7}$$

$$= I(\bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k) \tag{3.8}$$

Therefore,

$$(A_n) \to A \iff A = \bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k \tag{3.9}$$

Corollary 3.1. Let $A_n \uparrow$ be an increasing sequence, that is, $A_n \subseteq A_{n+1}$. Then

$$A_n \uparrow \bigcup_{n=1}^{\infty} A_n \tag{3.10}$$

Proof. TODO: *Prove this.* ■

Corollary 3.2. Let $A_n \downarrow$ be a decreasing sequence, that is, $A_n \supseteq A_{n+1}$. Then

$$A_n \downarrow \bigcap_{n=1}^{\infty} A_n \tag{3.11}$$

Proof. TODO: *Prove this.* ■

Proposition 3.2. If $A_n \uparrow A$ or $A_n \downarrow A$, then $P(A_n) \to P(A)$.

Proof. Let $A_0 = \varnothing$. Suppose $A_n \uparrow A$, then by previous corollary

$$A = \bigcup_{n=1}^{\infty} A_n = \sum_{i=1}^{\infty} (A_i - A_{i-1}) \tag{3.12}$$

Then

$$P(A) = \sum_{i=1}^{\infty} P(A_i - A_{i-1}) \tag{3.13}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i - A_{i-1}) \tag{3.14}$$

$$= \lim_{n\to\infty} P(A_n) - P(A_0) \tag{3.15}$$

$$= \lim_{n\to\infty} P(A_n) \tag{3.16}$$

TODO: *Prove the other case.* ■

Proposition 3.3 (Sequential Continuity)**.**

$$A_n \to A \implies P(A_n) \to P(A) \tag{3.17}$$

Theorem 3.1. Any distribution function $F_X(x)$ is right-continuous. That is, $F(x+) = F(x)$.

Proof. Let $x \in \mathbb{R}$, let $x_n \downarrow x$. Construct sequence of sets as $A_n := (-\infty, x_n]$.
By construction, $A_n \downarrow A$.
Therefore, $F_X(x_n) = P_X(A_n) \downarrow P_X(A) = F_X(x)$. ■

Definition 3.5. Let X be a real-valued random variable, then the **probability mass function** of X is given by

$$p_X(x) := P(X = x) \tag{3.18}$$

Proposition 3.4.

$$p(x) = F(x) - F(x-) \tag{3.19}$$

Proof. Let $x_n \uparrow\uparrow x$, let $A_n = (-\infty, x_n]$ and $A = (-\infty, x)$. Clearly, $A_n \uparrow A$. Then,

$$F(x-) = \lim_{n\to\infty} P(A_n) \tag{3.20}$$

$$= P(A) = P(X < x) \tag{3.21}$$

Hence $p(x) = P(X \le x) - P(X < x) = F(x) - F(x-)$. ■

Proposition 3.5. It is evident that $p(x) = 0$ wherever $F(x)$ is continuous.

Definition 3.6. The **points of continuity** of a distribution function F is defined as

$$C_F := p^{-1}(0) \tag{3.22}$$

And the **points of discontinuity** is simply $D_F := \mathbb{R} - C_F$.

Theorem 3.2. A distribution function F can have at most countably many discontinuous point. That is, $|D_F| \le \aleph_0$.

Proof. Note that

$$D_F = \bigcup_{n=1}^{\infty} \left\{ x \in \mathbb{R} : p(x) > \frac{1}{n} \right\} \tag{3.23}$$

Note that $\left|\left\{x \in \mathbb{R} : p(x) > \frac{1}{n}\right\}\right| < n$, therefore D_F is a countable union of countable sets, so $|D_F| \leq \aleph_0$. ■

3.2 Covariance as an Inner Product

Motivation Let W be a random variable defined on Ω, let $X := f(W)$ and $Y := g(W)$. Let $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_n)$ denote the sequence of outcomes.
Recall that the expectation of X is defined as

$$\mathbb{E}X := \lim_{n\to\infty} \hat{\mathbb{E}}_n X \tag{3.24}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=1}^n x_n}{n} \tag{3.25}$$

Such limit is well defined given X satisfies the empirical law of large numbers.

Definition 3.7. Define the **norm** (i.e. variance) of a random variable X as

$$||X|| := \sqrt{\mathbb{E}X^2} \tag{3.26}$$

$$= \lim_{n\to\infty} \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n}} \tag{3.27}$$

Definition 3.8. Define the **inner product** between two random variables X, Y as

$$\langle X, Y \rangle := \lim_{n\to\infty} \frac{\sum_{i=1}^n x_i y_i}{n} \tag{3.28}$$

Definition 3.9. The **cosine similarity** between two random variable X, Y can e written as

$$\cos\angle(X, Y) = \frac{\langle X, Y \rangle}{||X||||Y||} \tag{3.29}$$

$$= \lim_{n\to\infty} \frac{\frac{\sum_{i=1}^n x_i y_i}{n}}{\frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n}} \frac{\sqrt{\sum_{i=1}^n y_i^2}}{\sqrt{n}}} \tag{3.30}$$

$$\overset{n\to\infty}{\to} \frac{\mathbb{E}XY}{\sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2}} \tag{3.31}$$

Definition 3.10. The **covariance** between two random variables X, Y is defined to be the decentralized cosine similarity:

$$Cov(X, Y) := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) \tag{3.32}$$

In particular, for single random variable X, the **variance** is defined as

$$Var(X) := Cov(X, X) \tag{3.33}$$

The **coefficient of correlation** is defined to be the normalized covariance:

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \tag{3.34}$$

Proposition 3.6. $\mathbb{E}X$ is the closest constant (in terms of norm defined on random variables) near X. That is, $||X - \mathbb{E}X|| = \inf_{t\in\mathbb{R}} ||X - t||$.

Proof. Define $g(t) := \mathbb{E}(X-t)^2$, solving the first order condition gives $t^* = \mathbb{E}X$. ■

Proposition 3.7. Random variable space is a vector space.

Definition 3.11. The **centralization operator** for random variable is defined as

$$\dot{X} := X - \mathbb{E}X \tag{3.35}$$

Proposition 3.8. The centralization operator is linear, that is,

$$(X \dot{+} Y) = \dot{X} + \dot{Y} \tag{3.36}$$

$$(\dot{cX}) = c\dot{X} \tag{3.37}$$

Proof. The proof follows immediately from the linearity of expectation operator. ■

Theorem 3.3. Covariance operator is an inner product on the space of random variables.

Proof. Symmetry: obviously, $Cov(X,Y) = Cov(Y,X)$.
Bi-linearity:

$$Cov(X+Y,Z) = \mathbb{E}(X \dot{+} Y\dot{Z}) \tag{3.38}$$

$$= \mathbb{E}(\dot{X} + \dot{Y})\dot{Z} \tag{3.39}$$

$$= \mathbb{E}\dot{X}\dot{Z} + \mathbb{E}\dot{Y}\dot{Z} \tag{3.40}$$

$$= Cov(X,Z) + Cov(Y,Z) \tag{3.41}$$

Non-negativity

$$Cov(X,X) = \mathbb{E}\dot{X}\dot{X} \tag{3.42}$$

$$= \mathbb{E}(\dot{X})^2 \geq 0 \tag{3.43}$$

and the equality holds if and only if $X = \mathbb{E}X$ with probability 1 (i.e. X is deterministic). ■

3.3 Markov Inequality

Theorem 3.4 (Markov Inequality)**.** Let $Z \geq 0$ be a random variable, let $t \in \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function, then

$$P(Z \geq t) \leq \frac{\mathbb{E}g(Z)}{g(t)} \tag{3.44}$$

Proof. Note that

$$g(Z) \geq g(t)1(Z \geq t) \tag{3.45}$$
$$\implies \mathbb{E}g(Z) \geq g(t)P(Z \geq t) \tag{3.46}$$
$$\implies P(Z \geq t) \leq \frac{\mathbb{E}g(Z)}{g(t)} \tag{3.47}$$

■

Corollary 3.3 (Chebyshev Inequality)**.**

$$P(|X - \mathbb{E}X| \geq k) \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{k^2} \tag{3.48}$$

Proof. Let $Z = |X - \mathbb{E}X|$, $g(t) = t^2$, and apply the Markov inequality. ■

Corollary 3.4.

$$P(|X - \mathbb{E}X| \geq k\sigma) \leq \frac{1}{k^2} \tag{3.49}$$

where $\sigma^2 = \mathbb{E}(X - \mathbb{E}X)^2$.

Proof. Applying Chebyshev inequality gives

$$P(|X - \mathbb{E}X| \geq k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2} \tag{3.50}$$

■

Corollary 3.5. Let X be a random variable, $Var(X) = 0$ if and only if $X = \mathbb{E}X$ with probability 1.

Proof. ($\implies$) Suppose $Var(X) = 0$, then for every $n \in \mathbb{N}$,

$$0 \leq P(|X - \mathbb{E}X| \geq \frac{1}{n}) \leq Var(X)n^2 = 0 \tag{3.51}$$

Because the above inequality holds for arbitrarily large n, the only case is that $P(|X - \mathbb{E}X| > 0) = 0$.
Given $|X - \mathbb{E}X| \geq 0$, it must be $P(|X - \mathbb{E}X| = 0) = 1$.
($\impliedby$) Suppose $P(|X - \mathbb{E}X| = 0) = 1$.
Let $Y := (X - \mathbb{E}X)^2$, obviously $Var(X) = \mathbb{E}Y$.
It's been given that $P(Y = 0) = 1$, which means $\mathbb{E}Y = 0$. Therefore, $Var(X) = 0$. ■

Theorem 3.5 (An Application)**.**

$$U \sim unif[0, 1] \iff [nU] \sim unif\{0, \cdots, n - 1\} \tag{3.52}$$

where $[\,\cdot\,]$ denotes the floor function.

Proof. ($\implies$) Note that for every $0 \leq k \leq n - 1$,

$$P([nU] = k) = P(k \leq nU < k + 1) \tag{3.53}$$
$$= P(\frac{k}{n} \leq U < \frac{k + 1}{n}) \tag{3.54}$$
$$= \frac{k + 1}{n} - \frac{k}{n} = \frac{1}{n} \tag{3.55}$$

($\Longleftarrow$) Suppose $[nU] \sim unif\{0, \cdots, n-1\}$, we are going to show

$$\forall \alpha \in [0,1], P(U \leq \alpha) = \alpha \tag{3.56}$$

Considering $r \in [0,1) \cap \mathbb{Q}$ such that $r = \frac{k}{n}$ for some $\mathbb{Z} \ni k < n$.
Then

$$\begin{aligned} P(U < r) &= P(nU < k) && (3.57)\\ &= \sum_{i=0}^{n-1} P(nU < k \wedge [nU] = i) && (3.58)\\ &= \sum_{i=0}^{n-1} P(0 \leq nU < k \wedge i \leq nU < i+1) && (3.59)\\ &= \sum_{i=0}^{k-1} P(0 \leq nU < k \wedge i \leq nU < i+1) && (3.60)\\ &= \sum_{i=0}^{k-1} P(i \leq nU < i+1) && (3.61)\\ &= \sum_{i=0}^{k-1} P([nu] = i) && (3.62)\\ &= \frac{k}{n} = r && (3.63) \end{aligned}$$

Therefore, $P(U < r) = r$ for every rational $r \in [0,1)$. Because $\mathbb{Q}$ is dense in $\mathbb{R}$, every $x \in [0,1]$ can be approximated using a rational number r arbitrarily precisely. Hence, the result generalizes to $[0,1]$. ■

4 Conditional Probability

Definition 4.1. Given a random variable W on Ω with sequence of realizations (w_n), let $A, B \subseteq \Omega$. Then the **conditional empirical relative frequency** for A conditioned on B is defined to be

$$\begin{aligned} \hat{P}_n(W \in A | W \in B) &= \frac{I_A(w_1) I_B(w_1) + \cdots + I_A(w_n) I_B(w_n)}{I_B(w_1) + \cdots + I_B(w_n)} && (4.1)\\ &= \frac{\hat{P}_n(W \in A \cap B)}{\hat{P}_n(W \in B)} && (4.2) \end{aligned}$$

whenever $\hat{P}_n(W \in B) \neq 0$.

Definition 4.2. Given random variable W defined on sample space Ω and $A, B \subseteq \Omega$, the **conditional probability** of A given B is defined as

$$\begin{aligned} P(W \in A | W \in B) &:= \lim_{n \to \infty} \frac{\hat{P}_n(W \in A \cap B)}{\hat{P}_n(W \in B)} && (4.3)\\ &= \frac{P(W \in A \cap B)}{P(W \in B)} && (4.4) \end{aligned}$$

provided that $P(W \in B) \neq 0$.