

STA452

Lecture Notes

Yuchen Wang

April 16, 2021

Contents

1	Preface	3
1.1	Preliminary	3
2	Virtual Dice	4
2.1	The Law of Large Numbers	4
2.2	Some examples: “virtual dice”	5
2.2.1	A higher level of virtuality: ‘continuous dice’ and random ‘real’ numbers	5
2.2.2	Equality-in-distribution	6
2.3	Nature makes them, so can you	6
2.3.1	Exponential distribution	6
2.3.2	Consider the generalization	7
2.4	Expected Value	8
2.4.1	Expected Value for an Arbitrary Finite Discrete Distribution	9
2.4.2	Full generality: lebesgue-stieltjes	9
2.4.3	Examples	10
2.4.4	Expected Value for Continuous Functions	10
2.4.5	Expected Value for C^1 functions	10
2.5	Exponential Distribution	10
2.6	Gamma Distribution	10
2.7	Continuity Revisited	11
2.7.1	Sequential Continuity of Probability	11
2.7.2	Right Continuity of Cumulative Distribution Function	12
2.8	Back to the Uniform	13
2.9	Back to the Uniform II	13
2.9.1	Percentiles	13
2.9.2	Medians	16
3	Reduction to an Axiomatic System	16
3.1	The Kolmogorov Axioms	16
3.1.1	Reducing the Reduction	17
3.1.2	Recovering the Expectation Operator	18
3.1.3	Classical Random Variables	19
3.1.4	Expectation Operator	19
3.2	Recovering the Physics - Kolmogorov Synthesis	19

4	Geometry of Data	20
4.1	The Natural Geometry of \mathbb{R}^n	20
4.2	The Natural Geometry of Random Variables	21
4.2.1	Markov & Chebyshev	21
4.2.2	The Geometry of L_2	22
4.3	Covariance & Correlation	22
4.4	Simple Linear Model	23
4.5	General Linear Model	23

1 Preface

In our course, you will not see very much data. Our job is to express the ideas that drive the logic (or the logic that drives the ideas). As a consequence, the most of the examples in elementary book appear as pure theory. But they are not defined, they are consequences of abstract mathematical ideas. For example, conditional density is a consequence of conditional expectation, which is an orthogonal projection in a vector space, a pure euclidean geometric idea.

1.1 Preliminary

Definition 1.1. A sequence of sets $A_n \rightarrow A$ iff $I(A_n) \rightarrow I(A)$.

Proposition 1.1. A limit exists when the limsup is equal to the liminf:

$$\lim = \overline{\lim} = \underline{\lim} \quad (1.1)$$

Proof. For $w \in \Omega$,

$$\sup_{t \in T} I(A_t)(w) = I(\cup_{t \in T} A_t)(w) \quad (1.2)$$

$$= 1 \text{ or } 0 \quad (1.3)$$

$$\inf_{t \in T} I(A_t)(w) = I(\cap_{t \in T} A_t)(w) \quad (1.4)$$

Therefore,

$$\lim_{n \rightarrow \infty} I(A_n) = \overline{\lim}_{n \rightarrow \infty} I(A_n) \quad (1.5)$$

$$= \inf_{n=1}^{\infty} \sup_{k=n}^{\infty} I(A_k) \quad (1.6)$$

$$= I(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k) \quad (1.7)$$

$$= \underline{\lim}_{n \rightarrow \infty} I(A_n) \quad (1.8)$$

$$= \sup_{n=1}^{\infty} \inf_{k=n}^{\infty} I(A_k) \quad (1.9)$$

$$= I(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k) \quad (1.10)$$

■

Property 1.1. Therefore it is clear that

1.

$$A_n \rightarrow A \iff A = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \quad (1.11)$$

2.

$$A_n \uparrow \implies A_n \uparrow \cup_{n=1}^{\infty} A_n \quad (1.12)$$

3.

$$A_n \downarrow \implies A_n \downarrow \cap_{n=1}^{\infty} A_n \quad (1.13)$$

2 Virtual Dice

2.1 The Law of Large Numbers

Consider a *mechanism/process/system*, W , which generates *outcomes*, w , in a sample space Ω :

$$W : w_1, w_2, \dots, w_n, \dots$$

The outcomes are often referred to as *trials* of the *process* W . w_n is called the n th trial, and the finite sequence (w_1, w_2, \dots, w_n) is the first n trials.

Consider any real-valued function $g : \Omega \rightarrow \mathbb{R}$ defined on the *sample space* Ω . Let $X = g(W)$ denote the *extended process* that applies the function g to the outcome w from W to produce the outcome $x = g(w)$. This new process has its own sequence of trial outcomes:

$$g(W) : g(w_1), g(w_2), \dots, g(w_n), \dots \quad (2.1)$$

$$\text{or } X : x_1, x_2, \dots, x_n, \dots \quad (2.2)$$

These transformed outcomes are all real values, with which we can do lots of easy arithmetic, while the abstract sample space Ω may not have this property.

Definition 2.1 (sample mean). For each $n \in \mathbb{N}$, the *sample mean* over the first n trials is the *arithmetic average* of the function values over those n trials:

$$\hat{E}_n X := \frac{g(w_1) + \dots + g(w_n)}{n} = \bar{x}_n \quad (2.3)$$

Definition 2.2 (random variable). A given process W is said to be a *random process / random variable* iff it satisfies the *empirical law of large numbers*, in that, for any real-valued $X = g(W)$, we have

1. *stability*: the sequence of *sample averages* $(\hat{E}_n g(W), n \in \mathbb{N})$ converges;
2. *invariance*: the limit is independent of any particular realization $(w_n, n \in \mathbb{N})$.

Definition 2.3 (expected value). For each real-valued $X = g(W)$, we obtain a *expected value* in the above limit:

$$EX := \lim_{n \rightarrow \infty} \hat{E}_n g(W) = \lim_{n \rightarrow \infty} \hat{x}_n \quad (2.4)$$

Definition 2.4 (indicator function). The indicator function of a subset A of a set X is a function $I_A : X \rightarrow \{0, 1\}$ defined as

$$I_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \quad (2.5)$$

Definition 2.5 (probability). Now *probability* itself is a special case of an *expected value*: for any *indicator function* $g = I_A$ with $A \subset \Omega$ we will get the usual sequence of averages, but now to be referred to as *empirical relative frequencies*. These averages give the proportion of times that A occurs in the first n trials.

$$\hat{P}_n(W \in A) := \hat{E}_n I_A(W) = \frac{I_A(w_1) + \dots + I_A(w_n)}{n} \quad \forall n \in \mathbb{N} \quad (2.6)$$

As $n \rightarrow \infty$, the above equation gives the *long-run frequency*, or *probability*:

$$P_W(A) = P(W \in A) := \lim_{n \rightarrow \infty} \hat{E}_n I_A(W) = \lim_{n \rightarrow \infty} \hat{P}_n(W \in A) \quad (2.7)$$

Notation 2.1. Given a random variable W and a *probability distribution* P_W , we can use the following notation:

$$W \sim P_W \quad \text{on } \Omega$$

to be read as “ W is distributed as P_W on Ω ” or “ W is distributed as P_W ”.

2.2 Some examples: “virtual dice”

Definition 2.6. For any specific $n \in \mathbb{N}$, the random variable X is said to have a (*finite discrete*) *uniform distribution* on the sample space $\Omega = \{1, \dots, n\}$ (denoted $X \sim \text{unif}\{1, \dots, n\}$) iff

$$P(X = k) = \frac{1}{n}, \quad k = 1, \dots, n \quad (2.8)$$

Example 2.1. A ten-sided die: $Y \sim \text{unif}\{0, \dots, 9\}$ Let Y be a 2-stage procedure: Divide the ten digits $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$ into two batches

$$A = \{1, 2, 3, 4, 5\} \quad \& \quad B = \{6, 7, 8, 9, 0\} \quad (2.9)$$

and then toss a standard six-sided die twice. On the first toss, if the die shows 1, 2 or 3 then we go to A , if the die shows 4, 5 or 6 then we go to B . Thus each batch is selected half the time. On the second toss, ignoring the digit 6, and if the die shows k we take the k th digit in the batch and report the result. It should be clear then we will arrive at each of the ten digits with identical frequency $\frac{1}{10}$.

2.2.1 A higher level of virtuality: ‘continuous dice’ and random ‘real’ numbers

Let U denote the hypothetical possibility of generating an infinite decimal expansion of a number between 0 and 1, by performing the physical algorithm outlined in Example (2.1) an infinite number of times. So an outcome u for U entails an infinite number of repetitions $Y_i, i = 1, 2, \dots$ of the finite procedure Y :

$$U = \sum_{i=1}^{\infty} \frac{Y_i}{10^i} = 0.Y_1Y_2Y_3\dots \quad (2.10)$$

Example 2.2. If we generate U explicitly to four places $.y_1y_2y_3y_4$, then there are 10,000 equally likely possibilities, and our ‘actual’ U is known to be somewhere between $.y_1y_2y_3y_4$ and 0.0001 higher. In other words, the outcome is in one particular of 10,000 equally likely subintervals of $[0, 1]$:

$$P(.y_1y_2y_3y_4 \leq U \leq .y_1y_2y_3y_4 + 0.0001) = 1/10,000 \quad \forall y_1, y_2, y_3, y_4 \in \Omega \quad (2.11)$$

Thus we can deduce that

$$P(0 \leq U \leq .a_1a_2a_3a_4) = .a_1a_2a_3a_4/10,000 = .a_1a_2a_3a_4 \quad \forall a_1, a_2, a_3, a_4 \in \Omega \quad (2.12)$$

More generally, if u is an n -place finite decimal in the interval $[0, 1]$ for any $n \in \mathbb{Z}$, then $P(U \leq u) = u$, and for any pair of n -place finite decimals $a, b \in [0, 1]$ with $a \leq b$, we will have the *uniformity condition*

$$P(a \leq U \leq b) = b - a \quad (2.13)$$

Corollary 2.1. The probability of U obtaining any specific value u is zero.

$$P(U = u) = P(u \leq U \leq u) = u - u = 0 \quad (2.14)$$

Definition 2.7 (uniform distribution). The random variable U is said to have a (*continuous*) *uniform distribution* on the unit interval $[0, 1]$ (denoted $U \sim \text{unif}[0, 1]$) iff

$$P(U \leq u) = u \quad \forall 0 \leq u \leq 1 \quad (2.15)$$

Remark 2.1. This is a mathematical statement, which is different from physical existence as in Definition (2.6).

Corollary 2.2. If $X \sim \text{unif}[a, b]$ and $U \sim \text{unif}[0, 1]$, then

$$x = a + (b - a) \cdot u \quad (2.16)$$

Example 2.3. Let $V = 1 - U$, then

$$P(V \leq u) = P(1 - U \leq u) = P(U \geq 1 - u) \quad (2.17)$$

$$= 1 - P(U \leq 1 - u) \quad (2.18)$$

$$= 1 - (1 - u) = u = P(U \leq u) \quad (2.19)$$

As random variables, U and V behave exactly the same way. They have the same *stochastic behavior*. Accordingly, they are said to be *equal-in-distribution*: $V \stackrel{d}{=} U$.

2.2.2 Equality-in-distribution

Definition 2.8 (equality-in-distribution). Two random variables W_1, W_2 on the same sample space Ω are said to be *identically distributed* / *stochastically identical* (denoted $W_1 \stackrel{d}{=} W_2$) iff

$$Eg(W_1) = Eg(W_2) \quad \forall g : \Omega \rightarrow \mathbb{R} \quad (2.20)$$

iff

$$P(W_1 \in A) = P(W_2 \in A) \quad \forall A \subset \Omega \quad (2.21)$$

Proposition 2.1 (invariance 1). For any function $\phi : \Omega \rightarrow \chi$

$$W_1 \stackrel{d}{=} W_2 \implies \phi(W_1) \stackrel{d}{=} \phi(W_2) \quad (2.22)$$

Proof.

$$Eh(\phi(W_1)) = Eh(\phi(W_2)) \quad \forall h : \chi \rightarrow \mathbb{R} \quad (2.23)$$

■

Proposition 2.2 (invariance 2).

$$W_1 \stackrel{d}{=} W_2 \iff g(W_1) \stackrel{d}{=} g(W_2) \quad \forall g : \Omega \rightarrow \mathbb{R} \quad (2.24)$$

2.3 Nature makes them, so can you

2.3.1 Exponential distribution

Let $Z = -\ln U$ with $U \sim \text{unif}[0, 1]$. Then it is straightforward to compute that, for any non-negative $0 \leq s \leq t \leq \infty$:

$$P(s \leq Z \leq t) = e^{-s} - e^{-t} \quad (2.25)$$

Proof.

$$s \leq Z \leq t \iff s \leq -\ln U \leq t \quad (2.26)$$

$$\iff -t \leq \ln U \leq -s \quad (2.27)$$

$$\iff e^{-t} \leq U \leq e^{-s} \quad (2.28)$$

Therefore

$$P(s \leq Z \leq t) = P(e^{-t} \leq U \leq e^{-s}) \quad (2.29)$$

$$= e^{-s} - e^{-t} \quad (2.30)$$

■

Definition 2.9 (standard exponential distribution). The random variable Z is said to have a *standard exponential distribution* on $[0, \infty)$ (denoted $Z \sim \exp(1)$)
iff

$$P(Z \leq z) = 1 - e^{-z} \quad \forall z \geq 0 \quad (2.31)$$

Definition 2.10 (scaled exponential distribution). The random variable X is said to have a *scaled exponential distribution*, with *scale parameter* $\theta > 0$ on $[0, \infty)$ (denoted $X \sim \exp(\theta)$)
iff

$$X \stackrel{d}{=} \theta Z, \quad \text{where } Z \sim \exp(1) \quad (2.32)$$

2.3.2 Consider the generalization

Consider any strictly monotone and C^1 function, g on the interval $[0, 1]$, and let $X \stackrel{d}{=} g(U)$, where $U \sim \text{unif}[0, 1]$. Then

$$P(s < X \leq t) = \begin{cases} g^{-1}(t) - g^{-1}(s), & g \uparrow\uparrow \\ g^{-1}(s) - g^{-1}(t), & g \downarrow\downarrow \end{cases} \quad (2.33)$$

Corollary 2.3. Suppose $F : \mathbb{R} \rightarrow [0, 1]$ $x \mapsto P(X \leq x)$. Then F is certainly *non-decreasing*, and for any $s \leq t$,

$$P(s < X \leq t) = F(t) - F(s) \quad (2.34)$$

Definition 2.11 (distribution function). For any real-valued random variable, X , the *distribution function* of X is given by

$$F(x) \stackrel{\text{or}}{=} F_X(x) := P(X \leq x) \quad \forall x \in \mathbb{R} \quad (2.35)$$

Remark 2.2. Let $f(x) = F'(x)$, then we immediately have

$$P(s < X \leq t) = F(t) - F(s) = \int_s^t f(x) dx \quad \forall s, t \quad (2.36)$$

At each $x \in g[0, 1]$,

$$\lim_{s \uparrow x, t \downarrow x} \frac{P(s < X \leq t)}{t - s} = \lim_{s \uparrow x, t \downarrow x} \frac{F(t) - F(s)}{t - s} = f(x) \quad (2.37)$$

Remark 2.3. $f(x)$ can be interpreted as “amount of probability per unit length at the point x ”.

Definition 2.12 (probability density function). A real-valued random variable X is said to be *absolutely continuous* (wrt length measure)
iff

$$\exists f : \mathbb{R} \rightarrow [0, \infty), P(s < X \leq t) = \int_s^t f(x) dx \quad \forall s \leq t \quad (2.38)$$

in which case, the function f (**not necessarily unique**) is referred to as the *probability density function* of X .

Remark 2.4. For any abs. cont. X ,

$$P(X = x) = \int_x^x f(x) dx = 0 \quad \forall x \quad (2.39)$$

so there is no discrete contribution to the distribution at any $x \in \mathbb{R}$. Thus,

$$P(s \leq X \leq t) = P(s < X < t) = P(s < X \leq t) = P(s \leq X < t)$$

Proposition 2.3. $F : [a, b] \rightarrow [0, 1]$ is C^1 , iff

$$F(x) = \int_a^x f(s) ds \quad \text{with } f = F' > 0 \text{ cont. on } [a, b]$$

Proposition 2.4. If $g = F^{-1}$ and $g \in C^1$, then $F(X) \stackrel{d}{=} U$

Proof.

$$P(F(X) \leq u) = P(X \leq g(u)) \quad (2.40)$$

$$= P(X \leq g(u)) \quad (2.41)$$

$$= F(g(u)) \quad (2.42)$$

$$= u \quad (2.43)$$

$$= P(U \leq u) \quad (2.44)$$

■

Definition 2.13 (quantile). For any $0 \leq p \leq 1$, the value $x_p = g(p) = F^{-1}(p)$ is called the $100 \times p$ th *quantile* (or *percentile*) of X . The function g is called the *quantile function*.

$$P(X \leq x_p) = p \quad (2.45)$$

2.4 Expected Value

Property 2.1 (finite additivity of probability). If two sets A and B are mutually disjoint, then

$$I(A + B) = I(A) + I(B) \quad (2.46)$$

Therefore

$$P(A + B) = EI(A + B)(W) = E(I(A)(W) + I(B)(W)) \quad (2.47)$$

$$= EI(A)(W) + EI(B)(W) \quad (2.48)$$

$$= P(A) + P(B) \quad (2.49)$$

We can prove by induction that

$$P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) \quad (2.50)$$

Property 2.2 (E is normed on constant random variables). E is *normed* on $g(W) = c$ where $c \in \mathbb{R}$

$$Ec = c \quad \forall c \in \mathbb{R} \quad (2.51)$$

Property 2.3. The indicator function of the whole sample space Ω is 1

$$I_\Omega(W) = 1 \implies P(\Omega) = E1 = 1 \quad (2.52)$$

Property 2.4 (non-negativity of probability).

$$0 \leq I(A) \leq 1 \implies 0 \leq P(A) = EI(A) \leq 1 \quad (2.53)$$

2.4.1 Expected Value for an Arbitrary Finite Discrete Distribution

Definition 2.14 (finite scheme). For any finite discrete distribution, we can write a *finite scheme*

$$W \sim \begin{pmatrix} w_1 & \cdots & w_N \\ p_1 & \cdots & p_N \end{pmatrix} \quad (2.54)$$

to symbolize the *probability mass function*

$$P(W = w_i) = p_i, \quad i \in \{1, 2, \dots, N\}$$

where $\sum_{i=1}^N p_i = 1$.

Corollary 2.4. For any real-valued function $g(W)$, $W \in \Omega$, the expected value is

$$Eg(W) = \sum_{i=1}^N g(w_i)P(W = w_i) = \sum_{i=1}^N g(w_i)p_i \quad (2.55)$$

Proof. $g(W)$ can be explicitly represented as a finite linear combination of simple indicator functions

$$g(W) = \sum_{i=1}^N g(w_i)I(W = w_i)$$

So that applying E to both sides gives us the result. ■

Remark 2.5. For $U \sim \text{unif}\{1, \dots, n\}$, we have $EU = \frac{1+\dots+n}{n}$ and generally $EU^k = \frac{1+\dots+n^k}{n}$.

Note that $EU^k - E(U-1)^k = n^{k-1}$ for $k \in \mathbb{N}$. This provides an iterative basis for the computation of EU^k :

$$EU^2 - E(U-1)^2 = n \implies EU = \frac{n+1}{n} \quad (2.56)$$

$$EU^3 - E(U-1)^3 = n^2 \implies EU^2 = \frac{2n+1}{3}EU \quad (2.57)$$

$$EU^4 - E(U-1)^4 = n^3 \implies EU^3 = n(EU)^2 \quad (2.58)$$

$$(2.59)$$

2.4.2 Full generality: lebesgue-stieltjes

Suppose we are given a distribution function $F(x) = P(X \leq x)$, $x \in \mathbb{R}$, for a real-valued random variable $X = g(W)$, with $W \sim P$ on sample space Ω . Then consider some discrete approximation to X , for example,

$$X_n = \sum_{i=-n}^n \frac{i-1}{\sqrt{n}} I\left(\frac{i-1}{\sqrt{n}} < X < \frac{i}{\sqrt{n}}\right) \quad (2.60)$$

For this particular approximation,

$$|X - X_n| \leq \frac{1}{\sqrt{n}} + |X|I(|X| > \sqrt{n}) \quad (2.61)$$

Thus $X_n \rightarrow X$ as $n \rightarrow \infty$. Then any continuous real-valued function $h(X_n) \rightarrow h(X)$ as $n \rightarrow \infty$. If $h(X)$ is bounded, then

$$Eh(X) = \lim_{n \rightarrow \infty} \sum_{i=-n}^n h\left(\frac{i-1}{\sqrt{n}}\right) \left(F\left(\frac{i}{\sqrt{n}}\right) - F\left(\frac{i-1}{\sqrt{n}}\right)\right) \quad (2.62)$$

which is called the *lebesgue-stieltjes integral* of the function $h(x)$. It may be denoted

$$Eh(X) := \int_{-\infty}^{\infty} h(x) dF(x) \quad (2.63)$$

2.4.3 Examples

Definition 2.15 (bernoulli trial). The random variable Z is said to be a *bernoulli trial* (denoted $Z \sim \text{bern}(p), 0 \leq p \leq 1$) iff

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

2.4.4 Expected Value for Continuous Functions

2.4.5 Expected Value for C^1 functions

Consider the special case where function g is strictly monotone and C^1 .

Proposition 2.5. If $X = g(U), g : [0, 1] \rightarrow [a, b]$ is strictly monotone and C^1 , then for any continuous function $h : \mathbb{R} \rightarrow \mathbb{R}$,

$$Eh(X) = \int_a^b h(x)f(x) dx \quad (2.64)$$

where

$$F = \begin{cases} g^{-1} & , g \uparrow\uparrow \\ 1 - g^{-1} & , g \downarrow\downarrow \end{cases} \quad \text{and} \quad f(x) = F'(x)$$

Proof. For any $0 \leq t < 1$, when $g \uparrow\uparrow$ and C^1 , we have

$$\int_0^t h(g(u)) du = ??? \quad (2.65)$$

■

complete it later

2.5 Exponential Distribution

notes

2.6 Gamma Distribution

Property 2.5. As follows.

- $\Gamma(\alpha) = (\alpha - 1)!$ for positive integer α
- $\Gamma(p + 1) = p\Gamma(p)$
- $\Gamma(1/2) = \pi^{1/2}$
- If $Z \sim \text{Gamma}(p, 1)$, then $EZ^s = \Gamma(p + s)/\Gamma(p) \quad \forall s \in \mathbb{R}$
- $\text{Gamma}(v/2, 1/2) \stackrel{d}{=} \chi_{(v)}^2$
- $\text{Gamma}(1, \lambda) \stackrel{d}{=} \text{Exp}(\lambda)$
- If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ for $i = 1, 2, \dots, N$, then $\sum_{i=1}^N X_i \sim \text{Gamma}\left(\sum_{i=1}^N \alpha_i, \beta\right)$
- If $X \sim \text{Gamma}(\alpha, \theta), Y \sim \text{Gamma}(\beta, \theta)$ are independently distributed, then $X/(X + Y) \sim \text{Beta}(\alpha, \beta)$ is independent of $X + Y$.
- If $X \sim \text{Gamma}(\alpha, \beta)$, then $cX \sim \text{Gamma}\left(\alpha, \frac{\beta}{c}\right)$
- If $X_i \sim \text{Gamma}(\alpha_i, 1)$ are independently distributed, then the vector $(X_1/S, \dots, X_n/S)$, where $S = X_1 + \dots + X_n$ follows a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_n$.

2.7 Continuity Revisited

2.7.1 Sequential Continuity of Probability

Definition 2.16 (σ -additivity). P is said to be σ -additive / countably additive iff for any mutually disjoint sequence of events A_n ($n \in \mathbb{N}$)

$$P\left(\sum_1^\infty A_n\right) = \sum_1^\infty P(A_n) \quad (2.66)$$

Remark 2.6. Equation (2.66) is equivalent to the following pair of equations:

$$\text{finite-additivity: } P\left(\sum_1^n A_i\right) = \sum_1^n P(A_i) \quad (2.67)$$

$$\text{continuity: } A_n \rightarrow A \implies P(A_n) \rightarrow P(A) \quad (2.68)$$

Proposition 2.6. If $A_n \uparrow A$ or $A_n \downarrow A$, then

$$P(A_n) \rightarrow P(A)$$

Proof. If $A_n \uparrow A$ then we have that

$$A = \cup_{n=1}^\infty A_n = \sum_{n=1}^\infty (A_n - A_{n-1})$$

where, for convenience, we have $A_0 = \emptyset$.

Then

$$P(A) = \sum_{n=1}^\infty (P(A_n) - P(A_{n-1})) \quad (2.69)$$

$$= \lim_{n \rightarrow \infty} \sum_{i=1}^\infty (P(A_i) - P(A_{i-1})) \quad (2.70)$$

$$= \lim_{n \rightarrow \infty} P(A_n) \quad (2.71)$$

On the other hand, $A_n \downarrow A$ is equivalent to $A_n^c \uparrow A^c$. ■

Corollary 2.5 (sequential continuity).

$$A_n \rightarrow A \implies P(A_n) \rightarrow P(A) \quad (2.72)$$

Proof. Suppose $A_n \rightarrow A$, then

$$\cup_{n=1}^\infty \cap_{k \geq n} A_k = A = \cap_{n=1}^\infty \cup_{k \geq n} A_k \quad (2.73)$$

$$\cap_{k \geq n} A_k \leq A_n, A \leq \cup_{k \geq n} A_k \quad (2.74)$$

$$P(\cap_{k \geq n} A_k) \leq P(A_n), P(A) \leq P(\cup_{k \geq n} A_k) \quad (2.75)$$

$$|P(A_n) - P(A)| \leq P(\cup_{k \geq n} A_k) - P(\cap_{k \geq n} A_k) \quad (2.76)$$

$$\rightarrow P(A) - P(A) \quad (2.77)$$

$$= 0 \quad (2.78)$$

Therefore,

$$|P(A_n) - P(A)| \rightarrow 0 \quad (2.79)$$

$$P(A_n) \rightarrow P(A) \quad (2.80)$$

■

2.7.2 Right Continuity of Cumulative Distribution Function

For any $x_n \downarrow x$, simply let $A_n = (-\infty, x_n]$ and $A = (-\infty, x]$.

Then $A_n \downarrow A$, so

$$F(x_n) = P(X \in A_n) \downarrow P(X \in A) = F(x)$$

Denoting the right-limit of F at x by $F(x+) := \lim_{y \downarrow x} F(y)$, and the left-limit $F(x-) := \lim_{y \uparrow x} F(y)$, we get the property of *right-continuity* for CDF

$$F(x+) = F(x) \quad \forall x \in \mathbb{R} \quad (2.81)$$

Remark 2.7. Any distribution function $F(x)$ can actually be discontinuous at no more than a **countable** number of points, which corresponds to all the jumps on the discrete part of the distribution.

Definition 2.17 (probability mass function). For any real-valued random variable X , the *probability mass function* of X is given by

$$p(x) = p_X(x) = P(X = x) \quad \forall x \in \mathbb{R}$$

Proposition 2.7. Probability mass function

$$p(x) = F(x) - F(x-) \quad \forall x \in \mathbb{R} \quad (2.82)$$

Proof. For any $x_n \uparrow x$, simply let $A_n = (-\infty, x_n]$ and $A = (-\infty, x)$.

Then $A_n \uparrow A$, so

$$F(x-) := \lim_{n \rightarrow \infty} P(X \in A_n) = P(X \in A) = P(X < x)$$

Therefore

$$p(x) = P(X \leq x) - P(X < x) = F(x) - F(x-)$$

■

Remark 2.8. The points of continuity C_F of any distribution function correspond perfectly to the points where pmf is zero.

$$C_F = \{x \in \mathbb{R} | F(x-) = F(x+)\} \quad (2.83)$$

$$= \{x \in \mathbb{R} | F(x-) = F(x)\} \quad (2.84)$$

$$= \{x \in \mathbb{R} | p(x) = 0\} = p^{-1}(0) \quad (2.85)$$

The complementary region being the discrete part of the distribution

$$D_F = \{x \in \mathbb{R} | p(x) > 0\} = p^{-1}(0)^c \quad (2.86)$$

Proposition 2.8. D_F is at most countable.

$$\#D_F \leq \#\mathbb{N}$$

Proof. Note that

$$\{x \in \mathbb{R} | p(x) > 0\} = \cup_{n=1}^{\infty} \{x \in \mathbb{R} | p(x) > 1/n\}$$

It is clear that for every $n \in \mathbb{N}$, $\{x \in \mathbb{R} | p(x) > 1/n\}$ has less than n point in it. Otherwise

$$\exists A_n = \{a_1, \dots, a_n\} \subset \{x \in \mathbb{R} | p(x) > 1/n\} \text{ with } P(A_n) > 1$$

which is a contradiction.

Since a countable union of countable sets is still countable, we have D_F is at most countable.

■

2.8 Back to the Uniform

Definition 2.18 (p-adic series). For any $p \in \mathbb{N}$ with $p \geq 2$, any real number $U \in [0, 1)$ can be written as a base p expansion in the form

$$U = \sum_{i=1}^{\infty} Z_i p^{-i}$$

where $Z_i \in \{0, 1, 2, \dots, p-1\}$.

Notation 2.2. Let \dot{p}^∞ denote the collection of all the infinite p-sequences which do not end in $p-1$ repeated forever.

Lemma 2.1 (p-adic coding of the unit interval). $u = \sum_{i=1}^{\infty} z_i p^{-i}$ defines a correspondence $\Phi : \dot{p}^\infty \mapsto [0, 1)$.

Lemma 2.2 (p-adic partitioning). If $u = \sum_{i=1}^{\infty} z_i p^{-i}$ with $\mathbf{z} \in \dot{p}^\infty$ then

$$z_1 = b_1, \dots, z_n = b_n \iff \sum_{i=1}^n z_i p^{-i} \leq u < \sum_{i=1}^n z_i p^{-i} + p^{-n}$$

Theorem 2.1 (digital coding of the uniform). For $U = \sum_{i=1}^{\infty} z_i p^{-i}$ with $p \geq 2$ and $\mathbf{Z} \in \dot{p}^\infty$,

$$U \sim \text{unif}[0, 1] \iff Z_i \stackrel{i.i.d.}{\sim} \text{unif}\{0, \dots, p-1\}$$

Proof. Omitted here because it is very long and nuanced. ■

Remark 2.9. It is regarded as the *Fundamental Theorem of Applied Probability*.

2.9 Back to the Uniform II

2.9.1 Percentiles

For any given $X \sim F$ and any $0 < p < 1$

Definition 2.19 (percentile/quantile). A p-th *percentile/quantile* of X is any value, $\theta = \theta_p$, such that $F(\theta-) \leq p \leq F(\theta)$

Remark 2.10. Not necessarily unique, could be a closed interval on the real line.

Definition 2.20 (lower and upper quantile functions). We define the *lower quantile function* of the distribution function, F , to be the real-valued function $g : (0, 1) \rightarrow \mathbb{R}$ with

$$g(u) = \inf F^{-1}[u, 1] \tag{2.87}$$

and the *upper quantile function* to be $h : (0, 1) \rightarrow \mathbb{R}$ with

$$h(u) = \sup F^{-1}[0, u] \tag{2.88}$$

Remark 2.11. Both of these functions are non-decreasing. But even when $F(x)$ is strictly increasing, either of $g(u)$ or $h(u)$ may actually be constant over various intervals.

Proposition 2.9. For every $0 < u < 1$ and $x \in \mathbb{R}$ we have both

$$u \leq F(g(u)) \quad \text{and} \quad g(F(x)) \leq x \tag{2.89}$$

Proof. (1) By the definition of infimum, we may choose $x_n \downarrow g(u)$ with $F(x_n) \geq u \forall n$. Since F is right-continuous, then $F(x_n) \downarrow F(g(u))$. So $\lim F(x_n) = F(g(u)) \geq u$.
(2)

$$x \in F^{-1}(F(x)) \subset F^{-1}[F(x), 1]$$

Since $g(F(x)) = \inf F^{-1}[F(x), 1]$, then $g(F(x)) \leq x$. ■

Proposition 2.10. For every $0 < u < 1$ and $x \in \mathbb{R}$ we have both

$$u \leq F(h(u)) \quad \text{and} \quad x \leq h(F(x)) \quad (2.90)$$

Proof. (1) (2)

$$x \in F^{-1}(F(x)) \subset F^{-1}[0, F(x)]$$

Since $h(F(x)) = \sup F^{-1}[0, F(x)]$, then $x \leq h(F(x))$. ■

Corollary 2.6.

$$g(u) \leq x \iff u \leq F(x) \quad (2.91)$$

Proof.

$$g(u) \leq x \xrightarrow{F} u \leq F(g(u)) \leq F(x) \xrightarrow{g} g(u) \leq gF(x) \leq x$$

Corollary 2.7.

$$F(g(p)-) \leq p \leq F(g(p)) \quad (2.92)$$

Proof. From $g(u) \leq x \iff u \leq F(x)$, we can conclude $x < g(u) \iff F(x) < u$. Then $F(x) < p \quad \forall x < g(p)$, so $F(g(p)-) \leq p$. ■

Remark 2.12. Indeed, the set of all p th percentiles is the simple compact interval $[g(p), h(p)]$.

Corollary 2.8. $F(x-) \leq pF(x) \iff g(p) \leq x \leq h(p)$

Corollary 2.9.

$$FgF = F \quad \text{and} \quad gFg = g$$

Proof. Since $u \leq Fg(u)$, then $F(x) \leq F(g(F(x)))$.

Since $g(F(x)) \leq x$ and F is non-decreasing, then $F(g(F(x))) \leq F(x)$

Therefore, $F(x) \leq F(g(F(x))) \leq F(x) \implies FgF = F$.

Similarly, Since $u \leq F(g(u))$ and g is non-decreasing, then $g(u) \leq g(F(g(u)))$.

Since $g(F(x)) \leq x$, then $g(F(g(u))) \leq g(u)$.

Therefore, $g(u) \leq g(F(g(u))) \leq g(u) \implies gFg = g$. ■

Corollary 2.10. $g(u)$ is left-continuous.

$$u_n \uparrow u \implies g(u_n) \uparrow g(u) \quad (2.93)$$

Proof. $u_n \uparrow u \implies g(u_n) \uparrow c \leq g(u)$ for some upper limit c .

But $g(u_n) \leq c \quad \forall n$. Then $u_n \leq Fg(u_n) \leq F(c) \quad \forall n$

Then $u \leq F(c)$, then $g(u) \leq g(F(c)) \leq c$. Therefore $c \leq g(u) \leq c \implies g(u) = c$, then $g(u_n) \uparrow g(u)$. ■

Corollary 2.11. F is continuous iff $u = Fg(u) \quad \forall u$.

Proof. (\Rightarrow) Obvious, since $F(g(u)) = F(g(u)-) \leq u \leq F(g(u)) \implies u = Fg(u)$

(\Leftarrow) If $u = Fg(u)$ for every $0 < u < 1$, then we only need to show that F is left-continuous.

$x_n \uparrow x \implies F(x_n) \uparrow p = Fg(p) \leq F(x)$ for some p .

So if $g(p) = x$, we are done.

If $g(p) < x$, then $g(p) < x_n$ for n sufficient large.

So $p = Fg(p) \leq F(x_n) \leq p$ for n sufficient large, so $F(x_n) = p$ and

Proposition 2.11 (the quantile transform).

$$U \sim \text{unif}[0, 1] \implies g(U) \stackrel{d}{=} X \quad (2.94)$$

Proof. We know from Corollary (2.6) that $g(U) \leq x \iff U \leq F(x)$.

Therefore,

$$P(g(U) \leq x) = P(U \leq F(x)) \quad (2.95)$$

$$= F(x) \quad (\text{by the property of uniform distribution})$$

$$= P(X \leq x) \quad (2.96)$$

Corollary 2.12.

$$gF(X) \stackrel{d}{=} X \quad (2.97)$$

Proof. $gF(X) \stackrel{d}{=} gFg(U) \stackrel{d}{=} g(U) \stackrel{d}{=} X$.

Corollary 2.13.

$$P(F(X) \leq F(x)) = P(X \leq x) \quad \forall x \in \mathbb{R} \quad (2.98)$$

Proof. Let $x \in \mathbb{R}$

$$F(X) \leq F(x) \implies gF(X) \leq \underbrace{g(F(x))}_X \leq x \implies F(X) \leq F(x)$$

Therefore, $F(X) \leq F(x)$ iff $X \leq x$

Then $P(F(X) \leq F(x)) = P(X \leq x)$.

Proposition 2.12 (probability integral transform). F is continuous iff $F(X) \stackrel{d}{=} U$

Proof. (\Rightarrow): Assume F is continuous. From Proposition (2.11), we know $g(U) \stackrel{d}{=} X$, so $F(X) \stackrel{d}{=} Fg(U)$.

But since F is continuous, $Fg(U) = U$. Therefore, $F(X) \stackrel{d}{=} U$. (\Leftarrow): Assume $F(X) \stackrel{d}{=} U$.

$X = x$ implies $F(X) = F(x)$. This means $F(X) = F(x)$ may have a higher probability than $X = x$.

Therefore,

$$P(X = x) \leq P(F(X) = F(x)) = P(U = F(x)) = 0$$

Hence $P(X = x) = 0$ for all $x \in \mathbb{R}$ so F is continuous.

Proposition 2.13. Both g and F are continuous iff $g = h = F^{-1}$ on $(0, 1)$.

Proof. (\Rightarrow): Assume g and F are continuous.

Then from Corollary 2.11, $u = Fg(u)$. Also we can easily conclude that g is onto.

Property 2.6. Given any $f : \mathbb{R} \rightarrow (0, \infty) \in C$ s.t. $\int_{-\infty}^{\infty} f(x) dx = 1$, the function defined by $F(x) = \int_{-\infty}^x f(s) ds, x \in \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\} = [-\infty, \infty]$ determines a homeomorphism $F : \bar{\mathbb{R}} \xrightarrow{\cong} [0, 1]$ with quantile function $g = h = F^{-1}$.

2.9.2 Medians

Definition 2.21. A *median* for a random variable X is any $\theta = \theta_{1/2}$ s.t. $F(\theta-) \leq \frac{1}{2} \leq F(\theta)$ (denoted $\theta = \text{median}(X)$).

Remark 2.13. A median is simply a 50th percentile.

Proposition 2.14. Assuming $E|X| < \infty$ (the mean of X exists):

$$\theta = \text{median}(X) \iff E|X - \theta| = \inf_{t \in \mathbb{R}} E|X - t| \quad (2.99)$$

Remark 2.14. A median of a r.v. X is the closest constant to X in L_1 metric, a specific way of measuring the distance between two random objects:

$$d_1(x, y) = E|x - y|$$

Proposition 2.15. Assuming $E|X| < \infty$ (the mean of X exists):

$$\mu = EX \iff \sqrt{E(X - \mu)^2} = \inf_{t \in \mathbb{R}} \sqrt{E(X - t)^2} \quad (2.100)$$

Remark 2.15. A mean of a r.v. X is the closest constant to X in L_2 metric:

$$d_2(x, y) = \sqrt{E(x - y)^2}$$

3 Reduction to an Axiomatic System

3.1 The Kolmogorov Axioms

Definition 3.1 (probability space). A *probability space (distribution)* is a triple of objects (Ω, L, E)

1. Ω : any set, called the *sample space*
2. L : any vector space of real-valued functions on Ω that contains the constants, and is closed under taking absolute values ($X \in L \implies |X| \in L$), the elements of which are referred to as *random variables*
3. $E : L \rightarrow \mathbb{R}$, any functional that is
 - *normed*: $Ec = c$
 - *non-negative*: $X \geq 0 \implies EX \geq 0$
 - *linear*: $E \sum_1^n a_i X_i = \sum_1^n a_i EX_i$
 - *continuous*: $0 \leq X_n \uparrow X \implies 0 \leq EX_n \uparrow EX$

referred to as an *expectation operator*, while its value EX at any $X \in L$ is called the *expected value* of that X .

Property 3.1 (continuity). A useful variant of E 's *continuous* property is stated as:

If $Z_n \geq 0, n = 1, 2, \dots$, then

$$E \sum_{i=1}^{\infty} Z_n = \sum_{i=1}^{\infty} EZ_n \quad (3.1)$$

3.1.1 Reducing the Reduction

As understood, probability is a very special case of expected values. Thus we can reduce the definition of a probability space as follows

Definition 3.2 (probability space). A *probability space (distribution)* is a triple of objects (Ω, \mathcal{F}, P)

1. Ω : any set, called the *sample space*
2. \mathcal{F} : any σ -algebra of subsets of Ω , which is a non-empty collection closed under countable unions and complements. The elements of \mathcal{F} are referred to as *events*
3. $E : \mathcal{F} \rightarrow \mathbb{R}$, any functional that is
 - *normed*: $Ec = c$
 - *non-negative*: $X \geq 0 \implies EX \geq 0$
 - *σ -additive*: $P(\sum_1^\infty A_i) = \sum_1^\infty P(A_i)$

referred to as *probability measure*, while its value $P(A)$ at any $A \in \mathcal{F}$, is called the *probability* of that A .

Remark 3.1. σ -algebra is identical to σ -field.

Proposition 3.1 (nullity).

$$P(\emptyset) = 0$$

Proof. First we show that $\Omega \in \mathcal{F}$.

If $F \neq \emptyset$, then $\exists A \in \mathcal{F}$ s.t. $A^c \in \mathcal{F}$.

So let $A_1 = A, A_n = A^c \forall n \geq 2$.

Then

$$\cup_1^\infty A_n = A \cup A^c \cup A^c \cup \dots \quad (3.2)$$

$$= A \cup A^c \quad (3.3)$$

$$= \Omega \in \mathcal{F} \quad (3.4)$$

Define the sequence of mutually disjoint events

$$A_1 = \Omega \quad \& \quad A_n = \emptyset, \quad n \geq 2$$

Then we have $\Omega = \sum_{n=1}^\infty A_n$ and thus

$$1 = 1 + \lim_{n \rightarrow \infty} nP(\emptyset)$$

which forces the result. ■

Proposition 3.2 (finite-additivity).

$$P(A + B) = P(A) + P(B)$$

Corollary 3.1 (complementarity).

$$P(A^c) = 1 - P(A)$$

Corollary 3.2 (negative additivity).

$$P(A - B) = P(A) - P(A \cap B)$$

Proof. Since $A = AB + AB^c = AB + (A - B)$, then $P(A) = P(AB) + P(A - B)$, hence the result. ■

Corollary 3.3 (monotonicity).

$$A \subset B \implies P(A) \leq P(B)$$

Proof. Since $B = (B - A) \cup A$, then $P(B) - P(A) = P(B - A) \geq 0$, hence the result. ■

Proposition 3.3. Assuming *normed*, *non-negative* and *σ -additive*. If $A_n \uparrow A$ or $A_n \downarrow A$, then

$$P(A_n) \rightarrow P(A)$$

3.1.2 Recovering the Expectation Operator

The space of bernoulli trials (indicator functions) is

$$\mathcal{J} = \{I_A | A \in \mathcal{F}\}$$

On this collection, we have to define the expectation operator to be $E : \mathcal{J} \rightarrow \mathbb{R}$

$$E(I_A) = P(A) \quad \forall A \in \mathcal{F}$$

Starting from \mathcal{J} , we can create a vector space that contains [finite linear combinations of indicator functions](#). They are all the finite discrete random variables

$$\mathcal{S} = \left\{ S \mid S = \sum_{i=1}^m a_i I(A_i), a_i \in \mathbb{R}, A_i \in \mathcal{F}, i = 1, \dots, m, m \in \mathbb{N} \right\}$$

In this case, E is required to be linear, so we define it as $E : \mathcal{S} \rightarrow \mathbb{R}$

$$E(S) = \sum_{i=1}^m a_i P(A_i) \quad \forall S = \sum_{i=1}^m a_i I(A_i)$$

Lemma 3.1 (invariance at zero).

$$\sum_{i=1}^m a_i I(A_i) = 0 \implies \sum_{i=1}^m a_i P(A_i) = 0 \tag{3.5}$$

Corollary 3.4 (invariance).

$$\sum_{i=1}^m a_i I(A_i) = \sum_{j=1}^m b_j I(B_j) \implies \sum_{i=1}^m a_i P(A_i) = \sum_{j=1}^m b_j P(B_j) \tag{3.6}$$

Corollary 3.5 (linearity).

$$E : \mathcal{S} \xrightarrow{\text{linear}} \mathbb{R}$$

by

$$ES = \sum_{i=1}^m a_i P(A_i)$$

for any

$$S = \sum_{i=1}^m I(A_i)$$

3.1.3 Classical Random Variables

Consider the event $(X \leq x)$ for any $x \in \mathbb{R}$. This is a subset of the original sample space Ω :

$$(X \leq x) = \{w \in \Omega | X(w) \leq x\} = X^{-1}(-\infty, x]$$

Definition 3.3 (classical real-valued random variables). The classical real-valued random variables, X , wrt to a given distribution, (Ω, \mathcal{F}, P) , consist in the collection, $\mathcal{R} = \langle \mathcal{F} \rangle$ (generated by \mathcal{F}):

$$\mathcal{R} = \langle \mathcal{F} \rangle = \{X : \Omega \rightarrow \mathbb{R} | (X \leq x) \in \mathcal{F} \forall x \in \mathbb{R}\} \quad (3.7)$$

Proposition 3.4 (\mathcal{R} is closed wrt countable maxima and minima). For any given sequence $X_n, n \in \mathbb{N}$ in \mathcal{R} , provided they are real-valued, both $\inf_{n=1}^{\infty} X_n \in \mathcal{R}$ and $\sup_{n=1}^{\infty} X_n \in \mathcal{R}$

Corollary 3.6 (\mathcal{R} is sequentially closed). For a given sequence $X_n, n \in \mathbb{N}$ in \mathcal{R} ,

$$X_n \rightarrow X \implies X \in \mathcal{R}$$

Proposition 3.5 (fundamental representation). For any $Z \geq 0$ in \mathcal{R} , there are $S_n \geq 0$ in \mathcal{S} s.t.

$$0 \leq S_n \uparrow Z$$

Proposition 3.6. \mathcal{R} is an algebra (a vector space with multiplication), closed wrt sequential limits and absolute values.

3.1.4 Expectation Operator

Theorem 3.1 (monotone convergence theorem (MCT)).

$$0 \leq Z_n \uparrow Z \implies 0 \leq EZ_n \uparrow EZ \quad (3.8)$$

Corollary 3.7 (dominated convergence theorem (DCT)). Assume $X_n, Y \in \mathcal{R}$,

$$X_n \rightarrow X \text{ w. } |X_n| \leq Y, EY < \infty \implies E|X_n - X| \implies 0 \quad (3.9)$$

Property 3.2 (decomposition). $X \in \mathcal{R}$ can be decomposed into its positive and negative parts:

$$\begin{aligned} X^+ &= \max(X, 0) \\ X^- &= -\min(X, 0) \\ X &= X^+ - X^- \\ EX &= \begin{cases} EX^+ - EX^-, & EX^+ < \infty \text{ or } EX^- < \infty \\ \text{undefined}, & EX^+ = EX^- = \infty \end{cases} \end{aligned} \quad (3.10)$$

3.2 Recovering the Physics - Kolmogorov Synthesis

Theorem 3.2 (empirical law of large numbers (ELLN)). Suppose $X_i, i \in \{1, 2, \dots, n\}$ i.i.d. and the sample mean $\overline{X}_n = \frac{X_1 + \dots + X_n}{n}$

$$EX = \lim_{n \rightarrow \infty} \overline{X}_n \quad (3.11)$$

Theorem 3.3 (strong law of large numbers (SLLN)). Suppose $X_i, i \in \{1, 2, \dots, n\}$ i.i.d. and the sample mean $\overline{X}_n = \frac{X_1 + \dots + X_n}{n}$

$$P(\overline{X}_n \rightarrow EX) = 1 \quad (3.12)$$

4 Geometry of Data

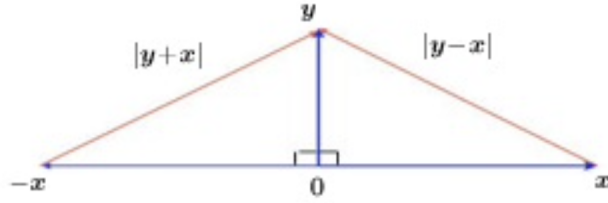
4.1 The Natural Geometry of \mathbb{R}^n

On \mathbb{R}^n we define three standard geometric devices

1. **inner product** $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}'\mathbf{y} = \sum_{i=1}^n x_i y_i$
2. **length (norm)** $|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$
3. **distance (metric)** $d(\mathbf{x}, \mathbf{y}) = |\mathbf{y} - \mathbf{x}|$

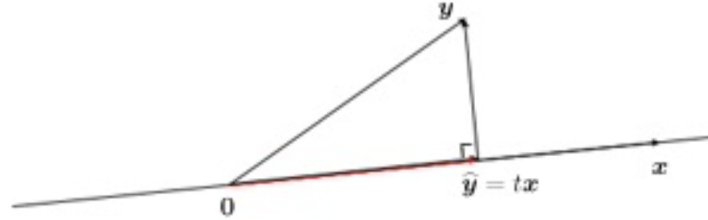
Definition 4.1 (orthogonality). For non-zero vectors \mathbf{x} and \mathbf{y} ,

$$\mathbf{x} \perp \mathbf{y} \iff |\mathbf{y} - \mathbf{x}| = |\mathbf{y} + \mathbf{x}| \iff \mathbf{x} \cdot \mathbf{y} = 0 \quad (4.1)$$



Definition 4.2 (orthogonal projection). The *orthogonal projection* $\hat{\mathbf{y}}$ of \mathbf{y} on any non-zero \mathbf{x} is a scalar multiple of \mathbf{x} and the *residual vector* $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to \mathbf{x} :

$$\begin{cases} \hat{\mathbf{y}} = t\mathbf{x} & \text{for some } t \in \mathbb{R} \\ \mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x} \end{cases} \quad (4.2)$$



Remark 4.1. As follows.

1. Then $\mathbf{x} \cdot (\mathbf{y} - t\mathbf{x}) = 0$ and $t = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2}$.
2. Let θ be the angle described by \mathbf{x} and \mathbf{y} . Then $\cos \theta = \text{sign}(\mathbf{x} \cdot \mathbf{y}) \frac{|\hat{\mathbf{y}}|}{|\mathbf{y}|} = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$

Property 4.1 (Pythagorean Theorem).

$$|\mathbf{y}|^2 = |\hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})|^2 = |\hat{\mathbf{y}}|^2 + |\mathbf{y} - \hat{\mathbf{y}}|^2 \quad (4.3)$$

Property 4.2 (Cauchy-Schwarz inequalities). As follows.

1. $|\hat{\mathbf{y}}| \leq |\mathbf{y}|$ w.eq. iff $\mathbf{y} = \hat{\mathbf{y}}$
2. $|\mathbf{x} \cdot \mathbf{y}| \leq |\mathbf{x}||\mathbf{y}|$ w.eq. iff $\mathbf{y} = t\mathbf{x}$ for $t \in \mathbb{R}$
3. $|\cos \theta(\mathbf{x}, \mathbf{y})| \leq 1$ w.eq. iff $\mathbf{y} = t\mathbf{x}$ for $t \in \mathbb{R}$

Property 4.3 (Triangle Inequality).

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}| \quad \text{w.eq. iff } \mathbf{y} = t\mathbf{x} \text{ for } t > 0 \quad (4.4)$$

4.2 The Natural Geometry of Random Variables

Definition 4.3 (L2 space). By definition, L is the vector space of real-valued random variables. We move to the particular **sub-space of L** where the natural geometry of \mathbb{R}^n has been reinvested in the random variables.

$$L2 = \{X \in L | EX^2 < \infty\} \quad (4.5)$$

In $L2$ we define

1. **inner product** $\langle X, Y \rangle = EXY$
2. **length (norm)** $\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{EX^2}$
3. **distance (metric)** $d(X, Y) = \|Y - X\| = \sqrt{E(Y - X)^2}$

4.2.1 Markov & Chebyshev

Theorem 4.1 (Markov's Inequality). For any $Z \geq 0, t \geq 0$ and non-decreasing $g : [0, \infty) \rightarrow [0, \infty)$, we have

$$P(Z \geq t) \leq \frac{Eg(Z)}{g(t)} \quad \forall t \text{ s.t. } g(t) > 0 \quad (4.6)$$

Proof.

$$g(t)I(Z \geq t) \leq g(Z) \quad (4.7)$$

$$g(t)P(Z \geq t) \leq Eg(Z) \quad (\text{applying } E \text{ to both sides})$$

$$P(Z \geq t) \leq \frac{Eg(Z)}{g(t)} \quad (4.8)$$

■

Consider an arbitrary random variable X in $L2$ with $\mu = EX$ and $\sigma^2 = E(X - \mu)^2$ and let $Z = |X - \mu|$ and $g(t) = t^2$ to get

Corollary 4.1 (Chebyshev I). For any $X \in L2, \epsilon > 0$, we have

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \quad (4.9)$$

Or take $Z = \frac{|X - \mu|}{\sigma}$ and $g(t) = k^2$ to find

Corollary 4.2 (Chebyshev II). For any $X \in L2, k > 0$, we have

$$P\left(\frac{|X - \mu|}{\sigma} > k\right) \leq \frac{1}{k^2} \quad (4.10)$$

Corollary 4.3 (Markov's Equality).

$$E|X| = 0 \iff X \stackrel{wP1}{=} 0 \quad (4.11)$$

4.2.2 The Geometry of L_2

Definition 4.4 (orthogonality).

$$X \perp Y \iff \|Y - X\| = \|Y + X\| \iff \langle X, Y \rangle = EXY = 0 \quad (4.12)$$

Definition 4.5 (orthogonal projection).

$$\begin{cases} \hat{Y} = tX & \text{for some } t \in \mathbb{R} \\ Y - \hat{Y} \perp X \end{cases} \quad (4.13)$$

Remark 4.2. 1. Then $E(Y - tX)X = 0$ and $t = \frac{EXY}{EX^2}$ provided $EX^2 > 0$.

$$2. \cos \theta(X, Y) = \text{sign}(EXY) \frac{\|\hat{Y}\|}{\|Y\|} = \frac{EXY}{\sqrt{EX^2 EY^2}}$$

Property 4.4 (Pythagorean Theorem).

$$\|Y\|^2 = \|\hat{Y} + (Y - \hat{Y})\|^2 = \|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2 \quad (4.14)$$

Property 4.5 (Cauchy-Schwarz inequalities). As follows.

1. $\|\hat{Y}\| \leq \|Y\|$ w.eq. iff $Y \stackrel{wP1}{=} \hat{Y}$
2. $(EXY)^2 \leq EX^2 Y^2$ w.eq. iff $Y \stackrel{wP1}{=} tX$ for $t \in \mathbb{R}$
3. $|\cos \theta(X, Y)| \leq 1$ w.eq. iff $Y = tX$ for $t \in \mathbb{R}$

Property 4.6 (Triangle Inequality).

$$\|X + Y\| \leq \|X\| + \|Y\| \quad \text{w.eq. iff } Y \stackrel{wP1}{=} tX \text{ for } t > 0 \quad (4.15)$$

4.3 Covariance & Correlation

Definition 4.6 (centred random variables). We define *centred* version of random variable X as

$$\dot{X} = X - EX$$

The expected value of any centred random variables is zero:

$$E\dot{X} = 0$$

Remark 4.3. • Variance is a quadratic operation instead of a linear one

- It is the inner product of the two centred variables:

$$\text{cov}(X, Y) := E\dot{X}\dot{Y} = E(X - EX)(Y - EY) \quad (4.16)$$

Definition 4.7 (correlation coefficient). We define the *correlation coefficient* of X and Y to be the *cosine of the angle* between the centred X and Y

$$\rho(X, Y) := \cos \theta(\dot{X}, \dot{Y}) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} \quad (4.17)$$

Property 4.7 (Cauchy-Schwarz inequalities). As follows.

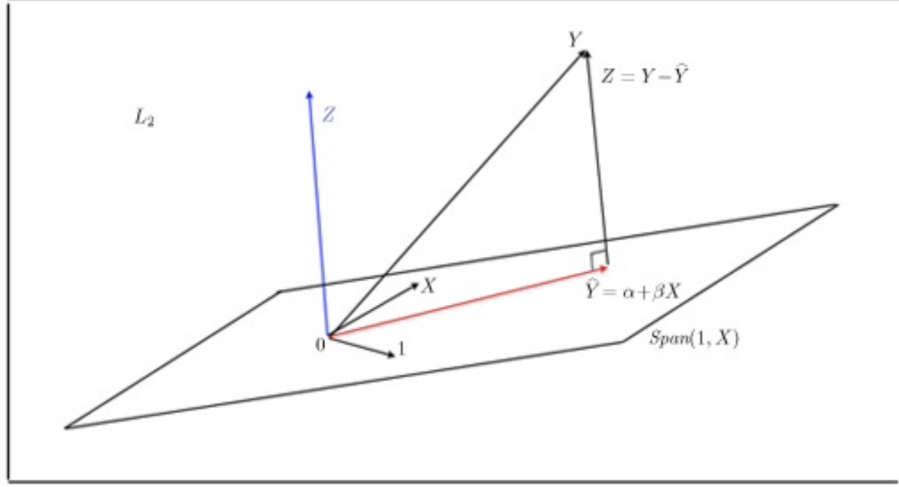
1. $|\text{cov}(X, Y)| \leq \sqrt{\text{var} X \text{var} Y}$ w.eq. iff $Y \stackrel{wP1}{=} \alpha + \beta X$
2. $|\rho(X, Y)| \leq 1$ w.eq. iff $Y \stackrel{wP1}{=} \alpha + \beta X$

4.4 Simple Linear Model

Proposition 4.1. Given any two real-valued r.v.s X and Y , in L_2 there exist unique scalars α and β and a unique r.v. Z s.t.

$$Y = \alpha + \beta X + Z \quad \text{w. } EZ = 0 = \rho(Z, X) \quad (4.18)$$

Remark 4.4. We call $Z = Y - \alpha - \beta X$ a *residual* variable.



Corollary 4.4.

$$\|Y - \alpha - \beta X\| \leq \|Y - s - tX\| \quad \text{w. eq. iff } s = \alpha, t = \beta \quad (4.19)$$

Property 4.8 (Pythagorean expression).

$$\|Y\|^2 = \|\alpha + \beta X\|^2 + \|Z\|^2 \quad (4.20)$$

$$\Rightarrow \text{var}Y = \text{var}(\alpha + \beta X) + \text{var}Z \quad (4.21)$$

$$= \underbrace{\beta^2 \text{var}X}_{\text{cov}(X,Y)=\beta \text{var}X} + \text{var}Z \quad (4.22)$$

$$= \rho(X, Y)^2 \text{var}Y + \text{var}Z \quad (4.23)$$

$$\Rightarrow \|Z\|^2 = \text{var}Z = (1 - \rho(X, Y)^2) \text{var}Y \quad (4.24)$$

$$\Rightarrow \frac{\|Y - \alpha - \beta X\|}{\|Y - EY\|} = \frac{\sigma(Z)}{\sigma(Y)} = \sqrt{1 - \rho(X, Y)^2} \leq 1 \quad (4.25)$$

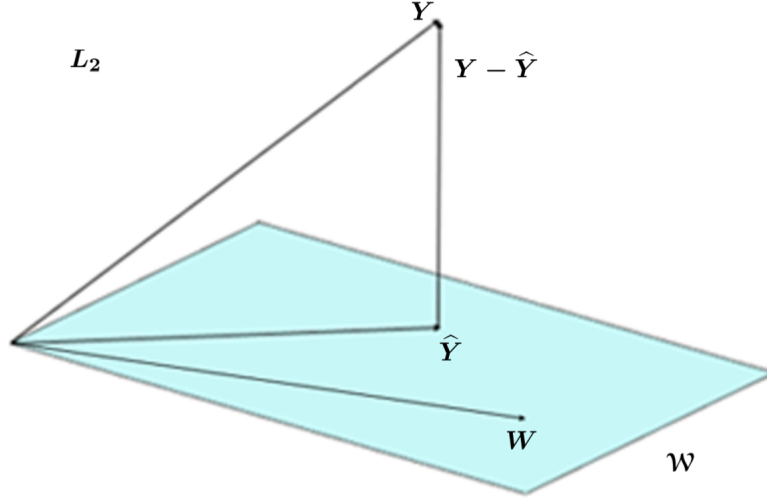
Remark 4.5. (Equation 4.25) tells us the ratio of two physical distances: from Y to $\hat{Y} = \alpha + \beta X$ and from Y to its own mean value. It will be strictly less than 1 if $\rho(X, Y)$ is not trivial, which means \hat{Y} , a linear transformation of a random variable, can easily do better than a single real value.

4.5 General Linear Model

A L_2 *prediction space* is any closed vector subspace $\mathcal{W} \subseteq L_2$, of r.v.s that contains the constants:

$$1 \in \mathcal{W} \stackrel{\text{closed}}{\subseteq} L_2$$

The best predictor of Y in \mathcal{W} is the unique element \hat{Y} that is closest to Y .



The *orthogonal complement* of \mathcal{W} is the vector subspace \mathcal{W}^\perp , of all r.v.s that are orthogonal to everything in \mathcal{W} itself:

$$\mathcal{W}^\perp = \{V \in L_2 \mid EVW = 0 \quad \forall W \in \mathcal{W}\} \quad (4.26)$$

It is clear that $\mathcal{W} \cap \mathcal{W}^\perp = \{0\}$.

Definition 4.8 (orthogonal projection). The *orthogonal projection* of Y on \mathcal{W} , denoted $\hat{Y} = op(Y|\mathcal{W})$, is the random variable \hat{Y} s.t.

$$\hat{Y} \in \mathcal{W} \quad \text{and} \quad Y - \hat{Y} \in \mathcal{W}^\perp$$

Proposition 4.2 (orthogonal projection and minimum distance).

$$\hat{Y} = op(Y|\mathcal{W}) \iff \|Y - \hat{Y}\| = \inf_{W \in \mathcal{W}} \|Y - W\| \quad (4.27)$$

Proposition 4.3 (general linear model). For any $Y \in L_2$ and $1 \in \mathcal{W} \subseteq L_2$, there are unique $\hat{Y} \in \mathcal{W}$ and Z s.t.

$$Y = \hat{Y} + Z \quad w. \quad EZ = 0 = \rho(Z, W) \quad \forall W \in \mathcal{W}$$

Proposition 4.4 (maximum correlated estimator). $Y = \hat{Y} + Z$ with $\hat{Y} \in \mathcal{W}, Z \in \mathcal{W}^\perp$ iff

$$\rho(\hat{Y}, Y) = \sup_{W \in \mathcal{W}} \rho(W, Y) \quad w. \quad EZ = 0 = \rho(Z, \hat{Y})$$