MATH1312: Lecture Note on Probability Theory and Mathematical Statistics

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2024 年 11 月 11 日

Contents

Chapter 1

Review of Probability Theory

Randomness should be taken into account in data, model, equations (PDEs), etc. This can be realized by allowing the model to be probabilistic in nature, which is referred to as a probability model. The reference book is [2].

1.1 Probability Space

https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/02%3A_Probability_Spaces/2.03%3A_Probability_Measures

https://www.stat.berkeley.edu/∼wfithian/courses/stat210a/measure-theory-basics.html. Also one can see in kejianbeifen_latex for Stochastic_Chap5_C2.pdf

A probability theory is made up of three part, $(\Omega, \mathcal{F}, P)$. Suppose that we have a random experiment with **sample space** $(\Omega, \mathcal{F})$, so that Ω is the set of outcomes of the experiment and $\mathcal{F}$) is the collection of events. When we run the experiment, a given event A either occurs or does not occur, depending on whether the outcome of the experiment is in A or not. Intuitively, the probability of an event is a measure of how likely the event is to occur when we run the experiment. Mathematically, probability is a function on the collection of events that satisfies certain axioms.

1.1.1 Sample space

Ω is a sample space.

Example 1.1.1 $\Omega = \{H(ead), T(ail)\}$ *for a coin flipping.*

Example 1.1.2 $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ *for flipping two coins.*

Example 1.1.3 $\Omega = \{1, 2, \ldots, 6\}$ *for rolling of a die.*

Example 1.1.4 $\Omega = \begin{Bmatrix} (1,1) & \cdots & (1,6) \\ \vdots & & \vdots \\ (6,1) & \cdots & (6,6) \end{Bmatrix}$ *for rolling of two dice.*

1.1.2 Events

Subset E of Ω is known as an event.

Example 1.1.5 $E = \{H\}$

Example 1.1.6 $E = \{2, 4, 6\}$. *Even number appears.*

- union of events. Given $E_1 = \{1, 3, 5\}$ and $E_2 = \{1, 2, 3\}$, then $E_1 \cup E_2 = \{1, 2, 3, 5\}$.
- intersection of events. Given above, then $E_1E_2 := E_1 \cap E_2 = \{1, 3\}$.
- complement. $E_1^c = \{2, 4, 6\}$.
- mutually exclusive. E, F, G are called mutually exclusive if $EF = \varnothing, EG = \varnothing, FG = \varnothing$. For example, $E = \{1, 2\}, F = \{3, 4\}, G = \{5, 6\}$ are mutually exclusive.

Definition 1.1.7 *$\mathcal{F}$ is a family of subsets of Ω satisfying:*
(1) $\Omega \in \mathcal{F}$.
(2) $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$.
(3) $E_j \in \mathcal{F} \Rightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{F}$.
Then $\mathcal{F}$ is a σ-albegra of Ω. $\Omega, \mathcal{F}$ is called a measurable space.

1.1.3 Probability space

Definition 1.1.8 *P is a function defined on satisfying:*
(1) non-negative. $0 \leq P(E) \leq 1, \forall E \in \mathcal{F}$.
(2) completeness. $P(\Omega) = 1$.
(3) countable additivity. For any countable mutually exclusive sets in $\mathcal{F}$, $P\big(\bigcup_{j=1}^{\infty} E_j\big) = \sum_{j=1}^{\infty} P(E_j)$.
Then $P(E)$ is the probability (or probability measure or probability distribution) of $E \in \mathcal{F}$. That is, $P : \mathcal{F} \to [0, 1], E \mapsto P(E)$. The three axioms are known as the **Kolmogorov axioms**, *in honor of Andrei Kolmogorov who was the first to formalize probability theory in an axiomatic way. Moreover, $(\Omega, \mathcal{F}, P)$ is the triple elements of a probability space.*

Example 1.1.9 $P(\{H\}) = P(\{T\}) = \frac{1}{2}$.

Example 1.1.10 $P(\{1\}) = \cdots = P(\{6\}) = \frac{1}{6}$.
$P(\{1, 3, 5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{2}$.

Example 1.1.11 $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1E_2) - P(E_1E_3) - P(E_2E_3) + P(E_1E_2E_3)$.

1.1.4 The Distribution of a Random Variable

Suppose now that X is a random variable for the experiment, taking values in a set T. Recall that mathematically, X is a function from Ω into T, and $\{X \in B\}$ denotes the event $\{\omega \in \Omega : X(\omega) \in B\}$ for $B \subset T$. Intuitively, X is a variable of interest for the experiment, and every meaningful statement about X defines an event.

Definition 1.1.12 *The function $B \mapsto P(X \in B)$ for $B \subset T$ defines a probability measure on T.*

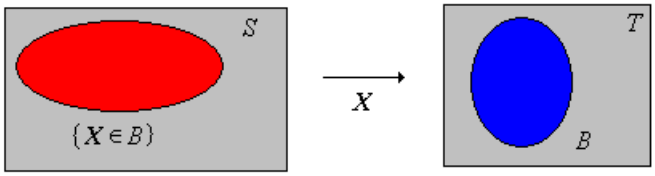

Figure 1.1: Here S should be changed to Ω. A set $B \in \mathcal{T}$ corresponds to the event $\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$.

The probability measure above is called the probability distribution of X, so we have all of the ingredients for a new probability space.

Definition 1.1.13 *A random variable X with values in T defines a new probability space:*
1. T is the set of outcomes.
2. Subsets of T are the events.
3. The probability distribution of X is the probability measure on T.

This probability space corresponds to the new random experiment in which the outcome is X. Moreover, recall that the outcome of the experiment itself can be thought of as a random variable. Specifically, if we let $T = \Omega$ we let X be the identity function on Ω, so that $X(\omega) = \omega$ for $\omega \in \Omega$. Then X is a random variable with values in Ω and $P(X \in A) = P(A)$ for each event A. Thus, every probability measure can be thought of as the distribution of a random variable.

1.2 Constructions of Probability Meaures

1.2.1 Measures

How can we construct probability measures? As noted briefly above, there are other measures of the size of sets; in many cases, these can be converted into probability measures. First, a **positive measure** μ on the sample space $(\Omega, \mathcal{F})$ is a real-valued function defined on Ω that satisfies axioms (1) and (3) in Definition 1.1.8, and then $(\Omega, \mathcal{F}, \mu)$ is a measure space. In general, $\mu(A)$ is allowed to be infinite. However, if $\mu(\Omega)$ is positive and finite (so that μ is a finite positive measure), then μ can easily be re-scaled into a probability measure.

Definition 1.2.1 *If μ is a positive measure on Ω with $0 < \mu(\Omega) < \infty$ then P defined below is a probability measure,*

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}, \qquad A \in \mathcal{F}.$$

In this context, $\mu(\Omega)$ is called the normalizing constant. In the next two subsections, we consider some very important special cases.

1.2.2 Discrete Distributions

In this discussion, we assume that the sample space $(\Omega, \mathcal{F})$ is discrete. Recall that this means that the set of outcomes Ω is countable and that $\mathcal{F} = \mathcal{P}(\Omega)$ is the collection of all subsets of Ω, so that every subset

is an event. The standard **uniform** measure on a discrete space is counting measure #, so that $\#(A)$ is the number of elements in A for $A \subset \Omega$. When Ω is finite, the probability measure corresponding to counting measure as constructed in above is particularly important in combinatorial and sampling experiments.

Definition 1.2.2 *Suppose that Ω is a finite, nonempty set. The discrete uniform distribution on Ω is given by*

$$P(A) = \frac{\#(A)}{\#(\Omega)}, \quad A \subset \Omega.$$

The underlying model is refereed to as the classical probability model, because historically the very first problems in probability (involving coins and dice) fit this model.

In the general discrete case, if P is a probability measure on Ω, then since Ω is countable, it follows from countable additivity that P is completely determined by its values on the singleton events. Specifically, if we define $f(x) = P(\{x\})$ for $x \in \Omega$, then $P(A) = \sum_{x \in A} f(x)$ for every $A \subset \Omega$. By axiom (1), $f(x) \geq 0$ for $x \in \Omega$ and by axiom (2), $\sum_{x \in \Omega} f(x) = 1$. Conversely, we can give a general construction for defining a probability measure on a discrete space.

Definition 1.2.3 *Suppose that $g : S \to [0, \infty)$. Then μ defined by $\mu(A) = \sum_{x \in A} g(x)$ for $A \subset \Omega$ is a positive measure on Ω. If $0 < \mu(\Omega) < \infty$ then P defined as follows is a probability measure on Ω,*

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{\sum_{x \in A} g(x)}{\sum_{x \in \Omega} g(x)} = \sum_{x \in A} \frac{g(x)}{\sum_{x \in \Omega} g(x)} = \sum_{x \in A} f(x), \quad A \subset \Omega.$$

In the context of our previous remarks, $f(x) = g(x)/\mu(\Omega) = g(x)/\sum_{y \in \Omega} g(y)$ for $x \in \Omega$. Distributions of this type are said to be discrete. Discrete distributions will be reviewed in detail in the following sections.

Proposition 1.2.4 *If Ω is finite and g is a constant function, then the probability measure P associated with g is the discrete uniform distribution on Ω.*

1.2.3 Continuous Distributions

The probability distributions that we will construct next are continuous distributions on $\mathbb{R}^n$ for $n \in \mathbb{N}_+$ and require some calculus.

Definition 1.2.5 *For $n \in \mathbb{N}_+$, the standard measure λ_n on $\mathbb{R}^n$ is given by*

$$\lambda_n(A) = \int_A 1 dx, \quad A \subset \mathbb{R}^n.$$

In particular, $\lambda_1(A)$ is the length of $A \subseteq \mathbb{R}^1$, $\lambda_2(A)$ is the area of $A \subseteq \mathbb{R}^2$, and $\lambda_3(A)$ is the volume of $A \subseteq \mathbb{R}^3$.

When $n > 3$, $\lambda_n(A)$ is sometimes called the n-dimensional volume of $A \subset \mathbb{R}^n$. The probability measure associated with λ_n on a set with positive, finite n-dimensional volume is particularly important.

Definition 1.2.6 *Suppose that $\Omega \subset \mathbb{R}^n$ with $0 < \lambda_n(\Omega) < \infty$. The continuous uniform distribution on Ω is defined by*

$$P(A) = \frac{\lambda_n(A)}{\lambda_n(\Omega)}, \quad A \subset \Omega.$$

Note that the continuous uniform distribution is analogous to the discrete uniform distribution, but with **Lebesgue** measure λ_n replacing counting measure $\#$. We can generalize this construction to produce many other distributions.

Definition 1.2.7 *Suppose again that $\Omega \subset \mathbb{R}^n$ and that $g : \Omega \to [0, \infty)$. Then μ defined by $\mu(A) = \int_A g(x)dx$ for $A \subset \Omega$ is a positive measure on Ω. If $0 < \mu(\Omega) < \infty$, then P defined as follows is a probability measure on Ω.*

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{\int_A g(x)dx}{\int_\Omega g(x)dx}, \quad A \in \mathcal{F}.$$

Distributions of this type are said to be continuous. Continuous distributions will also be reviewed in detail in the following sections. Note that the continuous distribution above is analogous to the discrete distribution, but with integrals replacing sums. The general theory of integration allows us to unify these two special cases, and many others besides.

1.3 Conditional Probability

Definition 1.3.1

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Example 1.3.2 *Choose one number from 1-10. The number is at least five, then what is the cond. probability that it is ten?*
Sol: Let $E = \{10\}$ and $F = \{\geq 5\}$. Then

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{\frac{1}{10}}{\frac{6}{10}} = \frac{1}{6}.$$

Example 1.3.3 *An urn contains 7 black balls and 5 white balls. Draw two balls without replacement. Each ball is equally drawn. What is probability that both drawn balls are black?*
Sol: F={first ball is black}, E={2nd ball is black}. Since $P(E|F) = \frac{6}{11}, P(F) = \frac{7}{12}$, then

$$P(EF) = P(F)P(E|F) = \frac{7 * 6}{12 * 11}.$$

Another solution is given directly by

$$P(\text{both black}) = \frac{C_7^2}{C_{12}^2} = \frac{7 * 6}{12 * 11}.$$

1.4 Independent events

$$\begin{aligned} E, F \text{ areindependent} &\Leftrightarrow P(EF) = P(E)P(F) \\ &\Leftrightarrow P(E|F) = P(E), \quad P(F) \neq 0 \\ &\Leftrightarrow P(F|E) = P(F), \quad P(E) \neq 0 \end{aligned}$$

Example 1.4.1 *Let a ball be drawn from an urn containing 4 balls {1,2,3,4}. Let* $E = \{1, 2\}$, $F = \{1, 3\}$, $G = \{1, 4\}$. *Then*

$$P(EF) = P(E)P(F) = \frac{1}{4},$$
$$P(EG) = P(E)P(G) = \frac{1}{4},$$
$$P(FG) = P(F)P(G) = \frac{1}{4}.$$

However,

$$\frac{1}{4} = P(EFG) \neq P(E)P(F)P(G) = \frac{1}{8}.$$

E, F, G *are not jointly independ.*

1.5 Discrete random variables

Definition 1.5.1 *If* X *is discrete with probability mass function* $p(x)$, *then for any real-valued function* g, *the expectation is defined as*

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

1.5.1 the Bernoulli random variable

An experiment, whose outcome is either a success or a failure. $X = 1$ is a success and $X = 0$ is a failure. Then X is denoted as $X \sim B(1, p)$ and the pmf is

$$\begin{aligned} p(0) &= P(X = 0) = 1 - p, \\ p(1) &= P(X = 1) = p. \end{aligned}$$

Its expect and var is

$$\begin{aligned} EX &= 1 \cdot p + 0 \cdot q = p, \\ Var(X) &= EX^2 - (EX)^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p(1 - p). \end{aligned}$$

1.5.2 the Binomial random variable

Suppose there are n trials of Bernoulli experiments. That is, If $X_1, \ldots, X_n$ are samples from $B(1, p)$, then
(1) $Y = X_1 + \cdots + X_n \sim B(n, p)$.
(2) pdf is given by $p(i) = C_n^i p^i (1 - p)^{n-i}$, $i = 0, \ldots, n$.
(3) $EY = \sum_{i=0}^n ip(i) = \sum_{i=0}^n iC_n^i p^i q^{n-i} = \sum_{i=1}^n i\frac{n!}{i!(n-i)!}p^i q^{n-i} = np\sum_{i=1}^n \frac{(n-1)!}{(i-1)!(n-i)!}p^{i-1}q^{n-i} = np$.
$Var(Y) = npq$.

Example 1.5.2 *Suppose each independent engine of an airplane will fail, when in flight, with probability* $1 - p$. *Suppose that the airplane will make a successful flight if at least 50 percent of its engines remain*

operative. For what values of p is a four-engine plane preferable to a two-engine plane?
Sol: A four-engine plane makes a successful flight with probability

$$\begin{aligned} & C_4^2 p^2(1-p)^2 + C_4^3 p^3(1-p)^1 + C_4^4 p^4(1-p)^0 \\ = \quad & 6p^2(1-p)^2 + 4p^3(1-p) + p^4. \end{aligned}$$

The probability for a two-engine plane is

$$C_2^1 p^1(1-p)^1 + C_2^2 p^2(1-p)^0 = 2p(1-p) + p^2.$$

Hence a four-engine plane is safer if

$$\begin{aligned} 6p^2(1-p)^2 + 4p^3(1-p) + p^4 &\geq 2p(1-p) + p^2 \\ 6p^3 - 12p^2 + 6p + 4p^2 - 4p^3 + p^3 &\geq 2 - p \\ 3p^3 - 8p^2 + 7p - 2 &\geq 0 \\ (p-1)^2(3p-2) &\geq 0. \\ p &\geq \frac{2}{3}. \end{aligned}$$

1.5.3 the geometric random variable

Suppose that independent trials, each having probability p of being a success, are performed until a success occurs. Let X be the number of trials required until the first success. The pmf is given by

$$p(n) = P(X = n) = (1-p)^{n-1}p, \quad n = 1, 2, \ldots$$

To check it is a pmf

$$\sum_{n=1}^{\infty} p(n) = p\sum_{n=1}^{\infty}(1-p)^{n-1} = \frac{p}{1-(1-p)} = 1.$$

The expect. and var is

$$\begin{aligned} EX &= \sum_{i=1}^{\infty} iq^{i-1}p = p\sum_{i=1}^{\infty}\frac{dq^i}{dq} = p\left(\frac{q}{1-q}\right)' = \frac{p}{(1-q)^2} = \frac{1}{p}. \\ Var(X) &= \frac{1-p}{p^2}. \end{aligned}$$

1.5.4 the Poisson random variable

X is said to be a Poisson random variable with parameter λ, denoted by $X \sim P(\lambda)$,

$$p(i) = P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}, i = 0, 1,$$

Check it is pmf

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1.$$

(1) $EX = \lambda, Var(X) = \lambda$.
(2) If $X_1, \ldots, X_n$ are indepen, $X_i \sim P(\lambda_i)$, then

$$X_1 + \cdots + X_n \sim P(\lambda_1 + \cdots + \lambda_n).$$

Pf. By induction.

$$\begin{aligned} P(X_1+X_2 &= i) = \sum_{k=0}^{i} P(X_1=k)P(X_2=i-k) \\ &= \sum_{k=0}^{i} e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{i-k}}{(i-k)!} = e^{-(\lambda_1+\lambda_2)}\frac{1}{i!}\sum_{k=0}^{i}\frac{i!}{k!(i-k)!}\lambda_1^k\lambda_2^{i-k} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{1}{i!}(\lambda_1+\lambda_2)^i \sim P(\lambda_1+\lambda_2). \end{aligned}$$

Example 1.5.3 *Suppose that the number of typo errors on a single page of a book has a Poisson distr. with parameter $\lambda = 1$. Calculate the probability that there is at least one error on this page.*
Sol: $P(X \geq 1) = 1 - P(X=0) = 1 - e^{-1} = 0.633$.

1.6 Continuous Random variable

Definition 1.6.1 *The cumulative distribution function (cdf) (or sometimes just distribution function) $F(\cdot)$ is defined by, $F(b) = P(X \leq b)$, satisfying (i) $F(b)$ is a nondecreasing function of b, (ii) $\lim_{b\to\infty} F(b) = F(\infty) = 1$, (iii) $\lim_{b\to-\infty} F(b) = F(-\infty) = 0$.*

One can see obviously that

$$P(a \leq X \leq b) = F(b) - F(a), \quad \text{for all } a < b.$$

Definition 1.6.2 *If there exists a nonnegative function $f(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B,*

$$P(X \in B) = \int_B f(x)dx.$$

The function $f(x)$ is called the probability density function (pdf) of X.

The relation bw the cdf F and the pdf f is

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx,$$

and

$$\frac{dF(a)}{da} = f(a).$$

Definition 1.6.3 *If X is continuous random variable with pdf $f(x)$, then for any real-valued function g, its expectation is defined by*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

1.6.1 Uniform random variable

Definition 1.6.4 *A random vari X is said to be uniformly distributed over $(0,1)$, if its pdf is given by*

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \textit{otherwise.} \end{cases}$$

Denote by $X \sim \mathcal{U}(0,1)$. Its expectation and variance are

$$EX = \int_0^1 xf(x)dx = \frac{1}{2}, \quad Var(X) = EX^2 - (EX)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

1.6.2 Exponential random variable

Definition 1.6.5 *A continuous random variable whose pdf is given, for some $\lambda > 0$, by*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

is said to be an exponential random variable with rate parameter λ. Denote by $X \sim \mathcal{E}(\lambda)$.

The cdf can be calculated by

$$\begin{aligned} F(a) &= \int_0^a \lambda e^{-\lambda x} dx = 1 - e^{-\lambda a}, \quad a \geq 0. \\ F(\infty) &= \int_0^\infty \lambda e^{-\lambda x} dx = 1. \end{aligned}$$

(1) $EX = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$.
(2) $P(X > t) = e^{-\lambda t}, t \geq 0$.
(3) Y is an exponential random variable if and only if $EY > 0$ and for $\forall s, t > 0$, such that

$$P(Y > s + t | Y > s) = P(Y > t),$$

where this condition is called **memoryless**. Denote $\overline{F}(t) = P(Y > t)$, then above Eq. is equivalent to

$$\overline{F}(t + s) = \overline{F}(t)\overline{F}(s).$$

Proof. One can easily check $\Rightarrow$ by noticing that $\overline{F}(t) = e^{-\lambda t}$. For the opposite direction, we want to prove $\overline{F}(t)$ is an exponential function. We first prove if $f(t + s) = f(t) + f(s)$, then f is linear function. For integers t and s, we have $f(n) = nf(1)$. For rational numbers t and s, $qf(p/q) = f(p) = pf(1)$, then $f(p/q) = p/qf(1)$. Since rational numbers are dense in real numbers, one can show $f(x) = xf(1)$ for all x real. Finally, $\overline{F}(t) = e^{f(t)} = e^{tf(1)}$, which is an expon. function. ■

Example 1.6.6 *Suppose a clock or a watch has a lifetime with exponential distribution with expectation 1 year. If it already works for 2 months, what's its remaining lifetime? (1 year since memoryless).*

Example 1.6.7 *Assume that the customer comes with interarrival time being exponential dist. If a cashier wants to go washroom, he/she goes right now or later on? (Right now since memoryless).*

1.6.3 Gamma ranodm variable

Definition 1.6.8 *A continuous random variable whose pdf is given, for some $\lambda > 0, \alpha > 0$, by*

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha - 1}}{\Gamma(\alpha)}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

is said to be a gamma random variable with rate parameters λ and α. Denote by $X \sim \Gamma(\alpha, \lambda)$. A Gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha - 1} dx.$$

The expectation and var of gamma vari is given by

$$EX = \frac{\alpha}{\lambda}, Var(X) = \frac{\alpha}{\lambda^2}.$$

1.6.4 Normal random variable

Definition 1.6.9 *X is normal random variable with parameters μ and σ^2 if the density of X is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty.$$

The density is bell-shaped curve that is symmetric around μ.

Definition 1.6.10 *multivariate normal distribution. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^T$. If $X = \mu + B\varepsilon$, then*

$$X \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma = BB^T$ is the covariance matrix of X.

(1) $X = (X_1, \ldots, X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ if and only if $\forall a_1, \ldots, a_n, \sum_{j=1}^n a_j X_j$ is normally distributed.
(2) Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then $X_1, \ldots, X_n$ are independent if and only if they are uncorrelated, i.e., $Cov(X_i, X_j) = 0$ for $i \neq j$. The proof can be found following.

Many real-world quantities tend to be normally distributed—for instance, human heights and other body measurements, cumulative hydrologic measures such as annual rainfall or monthly river discharge, errors in astronomical or physical observations, and diffusion of a substance in a liquid or gas. Some things are products of many independent variables (rather than sums), and in such cases the logarithm will be approximately normal since it is a sum of many independent variables—this is often the case for economic quantities such as stock market indices, due to the effect of compound interest.

1.6.5 Inverse Gamma Random Variable

If X is Gamma distributed then the distribution of $1/X$ is called the Inverse Gamma distribution. More precisely, if $X \sim \text{Gamma}(a, b)$ and $Y = 1/X$ then $Y \sim \text{InvGamma}(a, b)$, and the p.d.f. of Y is

$$\text{InvGamma}(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-b/y).$$

So, putting a $\text{Gamma}(a, b)$ prior on the precision λ is equivalent to putting an $\text{InvGamma}(a, b)$ prior on the variance $\sigma^2 = 1/\lambda$. The Inverse Gamma can be used to define a NormalInvGamma distribution for use as a prior on (μ, σ^2), which is sometimes more convenient than (but equivalent to) using a NormalGamma prior on (μ, λ).

1.6.6 History of Normal distribution

In 1809, Carl Friedrich Gauss (1777–1855) proposed the normal distribution as a model for the errors made in astronomical measurements, as a formal way of justifying the use of the sample mean, by showing it to be the most likely estimate—that is, the maximum likelihood estimate—of the true value (and more generally, to justify the method of least squares in linear regression). With astonishing speed, following Gauss' proposal, Laplace proved the central limit theorem in 1810. Laplace also calculated the normalization constant of the normal distribution, which is not a trivial task. James Clerk Maxwell (1831–1879) showed that the normal distribution arose naturally in physics, particularly in thermodynamics. Adolphe Quetelet (1796–1874) pioneered the use of the normal distribution in the social sciences. (See Fig. 1.2.)

Carl Friedrich Gauss

James Clerk Maxwell

Adolphe Quetelet

Figure 1.2: History of the normal distribution.

1.7 Jointly distributed random variables

1.7.1 independent random variables

Definition 1.7.1 *The random variables X and Y are said to be independent if, for all a, b,*

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b).$$

In terms of the joint distribution function F, we have that

$$F(a, b) = F_X(a)F_Y(b) \quad \text{for all } a, b.$$

Corollary 1.7.2 *When X and Y are discrete, the condition of indep. reduces to*

$$p(x, y) = p_X(x)p_Y(y).$$

If X and Y are jointly continuous, independence reduces to

$$f(x, y) = f_X(x)f_Y(y).$$

Pf.

$$\begin{aligned} P(X \leq a, Y \leq b) &= \sum_{y \leq b} \sum_{x \leq a} p(x, y) = \sum_{y \leq b} \sum_{x \leq a} p_X(x)p_Y(y) \\ &= \sum_{y \leq b} p_Y(y) \sum_{x \leq a} p_X(x) = P(Y \leq b)P(X \leq a). \end{aligned}$$

If X and Y are independ., then for any h and g

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Pf.

$$E[g(X)h(Y)] = \sum_y \sum_x g(x)h(y)p(x, y) = \sum_y \sum_x g(x)h(y)p_X(x)p_Y(y) = E[g(X)]E[h(Y)].$$

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy = E[g(X)]E[h(Y)]. \end{aligned}$$

Example 1.7.3 ***(Variance of a Binomial Random Variable)*** *Compute the Variance of a Binomial Random Variable. Sol. Binomial is the sum of n indep. Bernoulli.*

$$Var(X) = Var(X_1) + \cdots + Var(X_n) = npq,$$

since $Var(X_i) = pq$ for each Bernoulli distribution.

1.7.2 Covariance and Variance of Sums of Random Variables

Definition 1.7.4 *The covariance of any two random variables is*

$$Cov(X,Y) = E[(X - EX)(Y - EY)] = E[XY] - E[X]E[Y].$$

If X and Y are independent, then $Cov(X,Y) = 0$.

Corollary 1.7.5 *Property of Covariance*

(1) $Cov(X,X) = Var(X)$,

(2) $Cov(X,Y) = Cov(Y,X)$,

(3) $Cov(cX + dZ, Y) = cCov(X,Y) + dCov(Z,Y)$.

A useful expression for the variance can be found as follows:

$$\begin{aligned} Var(\sum_{i=1}^{n} X_i) &= Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Cov(X_i, X_i) + 2\sum_{i<j} Cov(X_i, X_j). \end{aligned}$$

Moreover, if X_i are indep. random variables, then above equation reduces to

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i).$$

Definition 1.7.6 *If $X_1, \ldots, X_n$ are i.i.d., then the random variable $\overline{X} = \sum_{i=1}^{n} X_i/n$ is called the* ***sample mean****.*

Proposition 1.7.7 *Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. Then*

(a) $E[\overline{X}] = \mu$.

(b) $Var(\overline{X}) = \sigma^2/n$.

(c) $Cov(\overline{X}, X_i - \overline{X}) = 0, i = 1, \ldots, n$.

Pf. Parts (a) and (b) are easy:

$$\begin{aligned} E[\overline{X}] &= \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu, \\ Var[\overline{X}] &= \left(\frac{1}{n}\right)^2 Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}. \end{aligned}$$

To prove (c), we follow

$$\begin{aligned} Cov(\overline{X}, X_i - \overline{X}) &= Cov(\overline{X}, X_i) - Cov(\overline{X}, \overline{X}) = \frac{1}{n} Cov(X_i + \sum_{j \neq i} X_j, X_i) - Var[\overline{X}] \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0. \end{aligned}$$

Proposition 1.7.8 *The **sample variance** is given by*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Then it is unbiased, that is,

$$ES^2 = \sigma^2.$$

Pf. Notice that

$$\begin{aligned} \sum_{i=1}^{n} (X_i - \overline{X})^2 &= \sum_{i=1}^{n} (X_i - \mu + \mu - \overline{X})^2 \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2(\mu - \overline{X}) \sum_{i=1}^{n} (X_i - \mu) \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\mu - \overline{X})^2. \end{aligned}$$

Then we obtain

$$\begin{aligned} E[(n-1)S^2] &= \sum_{i=1}^{n} E(X_i - \overline{X})^2 = \sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\mu - \overline{X})^2 \\ &= n\sigma^2 - nVar[\overline{X}] = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \end{aligned}$$

1.7.3 Sum of two independent variables

Let us derive the formula first. Suppose that X and Y are continuous and independent, X having pdf f and Y having pdf g. Letting $F_{X+Y}(a)$ be the cdf of $X + Y$, we have

$$\begin{aligned} F_{X+Y}(a) &= P(X + Y \leq a) = \iint_{x+y \leq a} f(x)g(y)dxdy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{a-y} f(x)dx \right) g(y)dy = \int_{-\infty}^{\infty} F_X(a-y)g(y)dy. \end{aligned}$$

By differentiating above, we obtain the pdf $f_{X+Y}(a)$ of $X + Y$ given by

$$f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)g(y)dy = \int_{-\infty}^{\infty} f(a-y)g(y)dy.$$

Thus f_{X+Y} is the **convolution** of functions f and g.

Example 1.7.9 *Two uniform random variables If X and Y are indepdt. both uniformly distributed on $(0, 1)$, then calculate the pdf of $X + Y$.*

Sol. The pdf's are

$$f(a) = g(a) = \begin{cases} 1, & 0 < a < 1, \\ 0, & otherwise. \end{cases}$$

we obtain

$$f_{X+Y}(a) = \int_0^1 f(a-y)g(y)dy.$$

For $0 \le a \le 1$, *this yields*

$$f_{X+Y}(a) = \int_0^a dy = a$$

since $0 \le a - y \le 1$ *and* $0 \le y \le 1 \Rightarrow 0 \le y \le a$. *And for* $1 \le a \le 2$, *this yields*

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

since $0 \le a - y \le 1$ *and* $0 \le y \le 1 \Rightarrow a - 1 \le y \le 1$. *Hence,*

$$f_{X+Y}(a) = \begin{cases} a, & 0 < a < 1, \\ 2-a & 1 < a < 2 \\ 0, & \textit{otherwise.} \end{cases}$$

1.8 Limit Theorems

Proposition 1.8.1 ***(Markov's Inequality)*** *If* X *is a random variable that takes only nonnegative values, then for any* $a > 0$

$$P\{X \ge a\} \le \frac{E[X]}{a}.$$

Pf. We give a proof for the case where X *is continuous with denfity* f,

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx = \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\ &\ge \int_a^\infty xf(x)dx \ge a\int_a^\infty f(x)dx = aP\{X \ge a\}. \end{aligned}$$

Proposition 1.8.2 ***(Chebyshev's Inequality)*** *If* X *is a random variable with mean* μ *and variance* σ^2, *then, for any* $k > 0$,

$$P\{|X - \mu| \ge k\} \le \frac{\sigma^2}{k^2}.$$

Pf. We apply Markov's inequality to the nonnegative $(X - \mu)^2$,

$$P\{(X-\mu)^2 \ge k^2\} \le \frac{E[(X-\mu)^2]}{k^2}.$$

Remark 1.8.3 *The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probs. when only the mean, or both the mean and the variance, are known. Of course, if the true distribution were known, then the desired probs. could be exactly computed, and we would not need to resort to bounds.*

Theorem 1.8.4 ***(Strong Law of Large Numbers)*** *Let* $X_1, X_2, \ldots$ *be a sequence of independent random variables have a commom distribution, and let* $E[X_t] = \mu$. *Then, with probability 1, or almost surely,*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty.$$

Definition 1.8.5 *If* $P(\lim_{n\to\infty} X_n = X) = 1$, *then we say* $X_n \to X$, *a.s. (almost surely) or* $X_n \to X$, *w.p.1 (with probability 1).*

Theorem 1.8.6 ***(Central Limit Theorem)*** *Let $X_1, X_2, \dots$ be a sequence of independent, identically distributed (i.i.d.) random variables, each with mean μ and variance σ^2. Then the distribution of*

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

goes to the standard normal as $n \to \infty$. That is,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx = \Phi(a),$$

as $n \to \infty$.

Proof. Note that the theorem holds for any distribution of the X_is; herein lies its power.

We now present a heuristic proof the CLT. Suppose first that the X_i have mena 0 and variance 1, and then the MGF can be computed,

$$E\left[\exp\left\{t\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right\}\right] = E[e^{tX_1/\sqrt{n}} \cdots e^{tX_n/\sqrt{n}}] = (Ee^{tX_i/\sqrt{n}})^n \text{ by independence.}$$

For large n, we obtain by Taylor expansion,

$$e^{tX_i/\sqrt{n}} = 1 + \frac{tX_i}{\sqrt{n}} + \frac{(tX_i)^2}{2n} + O(n^{-3/2}),$$

that is the reason for the **central** word. Taking expectations shows that when n is large,

$$E[e^{tX_i/\sqrt{n}}] = 1 + \frac{t^2}{2n} + O(n^{-3/2}), \text{ since } EX = 0 \text{ and } EX^2 = 1.$$

Therefore, we obtain

$$E\left[\exp\left\{t\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right\}\right] \approx \left(1 + \frac{t^2}{2n}\right)^n \to e^{t^2/2}.$$

Thus, the MGF of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the moment generating function of a standard normal random variable with mean 0 and variance 1. Notice that for $X \sim N(0,1)$, its MGF $\phi(t) = e^{t^2/2}$. Hence, it can be proven that the distribution function of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the distribution function of a standard normal Φ. When X_i have mean μ and variance σ^2, the random variables $\frac{X_i-\mu}{\sigma}$ have mean 0 and variance 1. Done. ■

Proposition 1.8.7 *The convergence has the following relations:*

$$\left.\begin{array}{c}\textit{conv. in moments or } L^p \textit{ converges} \\ \textit{a.s. or w.p.1}\end{array}\right\} \Rightarrow \textit{conv. in probability} \Rightarrow \textit{conv. in distribution.}$$

See https://zhuanlan.zhihu.com/p/70034585 for more details of theorems and counter-examples.

Lemma 1.8.8 ***(Lévy-Crammer)*** *$\{F_n\}$ is a set of distributions. If $\hat{F}_n \to \phi(t)$ conv. pointwisely, then $F_n \to F$ converges weakly, where ϕ is the character function of F and $\hat{F}_n$ is the character function of F_n.*

Theorem 1.8.9 ***(Linderberg-Lévy CLT)*** *check https://zhuanlan.zhihu.com/p/69862244. character function and Lindberg-Levy central limit theorem. See details of these two theorems in book Probability Theory and Mathematical Statistics by Marek Fisz [1].*

Example 1.8.10 *If X is binomially distributed with parameters n and p, then X is the sum of n independent Bernoulli random variables, each with parmeter p. Hence, the distribution of*

$$\frac{X - E[X]}{\sqrt{Var(X)}} = \frac{\sum X_i - n\mu}{\sqrt{n}\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches the standard normal distribution as n approaches ∞. The normal approximation will be quite good for $np(1-p) \geq 10$ or $\sqrt{Var(X)} \geq \sqrt{10}$.

Example 1.8.11 ***(Normal approximation to the Binomial)*** *Let X be the number of times that a fair coin, flipped 40 times, lands heads. Find the probability that $X = 20$.*
Sol.

$$\begin{aligned} P\{X = 20\} &= P\{19.5 < X < 20.5\} \\ &= P\left\{\frac{19.5 - 20}{\sqrt{10}} < \frac{X - 20}{\sqrt{10}} < \frac{20.5 - 20}{\sqrt{10}}\right\} \\ &= P\left\{-0.16 < \frac{X - 20}{\sqrt{10}} < 0.16\right\} = \Phi(0.16) - \Phi(-0.16) \\ &= 0.1272. \end{aligned}$$

The exact result is

$$P\{X = 20\} = C_{40}^{20}\left(\frac{1}{2}\right)^{20}\left(\frac{1}{2}\right)^{20} = 0.1268.$$

Example 1.8.12 *The lifetime of a battery is a random variable with mean 40 hours and standard deviation 20 hours. Assume a stockpile of 25 such batteries, approximate the probability that over 1100 hours of use can be obtained.*
Sol.

$$\begin{aligned} P\{X_1 + \cdots + X_{25} &> 1100\} = P\left\{\frac{X_1 + \cdots + X_{25} - 25 \times 40}{20\sqrt{25}} > \frac{1100 - 25 \times 40}{20\sqrt{25}}\right\} \\ &= P\{N(0,1) > 1\} = 1 - \Phi(1) = 0.1587. \end{aligned}$$

Bibliography

[1] M. Fisz. Probability theory and mathematical statistics. *(No Title)*, 1963.

[2] S. M. Ross. Introduction to probability models, ninth edition. In *Academic Press, Inc.*, 2006.