# A SHORT SUMMARY ON 'A FIRST COURSE IN PROBABILITY'

LONG CHEN

ABSTRACT. This is an outline of the book 'A First Course in Probability', which can serve as a minimal introduction to the probability theory.

## CONTENTS

1. Combinatorial Analysis 2
2. Axioms of Probability 2
3. Conditional Probability and Bayes' Formulae 3
Bayes's Formulae 3
4. Random Variables 4
Important Discrete Random Variables 5
5. Continuous Random Variables 7
Important Continuous Random Variables 8
6. Jointly Distributed Random Variables 10
6.1. Jointly Distribution 10
6.2. Summation of Independent Random Variables 11
7. Properties of Expectation and Variance 11
7.1. Expectation of Sums of Random Variables 11
7.2. Moments of the Number of Events that Occur 12
7.3. Covariance, Variance of Sums, and Correlations 13
7.4. Moment Generating Functions 14
7.5. The Sample Mean and Variance 14
8. Limit Theorems 15
8.1. Tail bound 15
8.2. The Central Limit Theorem 16
8.3. Law of Large Numbers 17
Appendix 19

*Date*: November 12, 2015.

## 1. Combinatorial Analysis

The following sampling table presents the number of possible samples of size $k$ out of a population of size $n$, under various assumptions about how the sample is collected.

Table 1. Sampling Table

| | Order | No Order |
|---|---|---|
| With Rep | $n^k$ | $\binom{n+k-1}{k}$ |
| Without Rep | $k!\binom{n}{k}$ | $\binom{n}{k}$ |

Here 'Rep' stands for 'Replacement' or 'Repetition' meaning that in the sampling the output can have duplication items.

## 2. Axioms of Probability

Let $S$ denote the set of all possible outcomes of an experiment. $S$ is called the sample space of the experiment. An event is a subset of $S$.

An intuitive way of defining the probability is as follows. For each event $E$ of the sample space $S$, we define $n(E)$ to be the number of outcomes favorable to $E$ in the first $n$ experiments. Then the naive definition of the probability of the event $E$ is

$$\Pr(E) = \lim_{n\to\infty} \frac{n(E)}{n}.$$

That is $\Pr(E)$ can be interpreted as a long-run relative frequency. It possesses a serious drawback: How do we know the limit exists? We have to believe (assume) the limit exists!

Instead, we can accept the following axioms of probability as the definition of a probability:

(A1)

$$0 \leq \Pr(E) \leq 1$$

(A2)

$$\Pr(S) = 1$$

(A3) For mutually exclusive events

$$\Pr\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$$

The inclusive-exclusive formulate for two events is

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F)$$

which can be generalized to $n$-events and will be proved later

$$\begin{aligned}\Pr\left(\cup_{i=1}^{n} E_i\right) = &\sum_{i=1}^{n} \Pr(E_i) - \sum_{i<j} \Pr(E_i \cap E_j) \\ &+ \sum_{i<j<k} \Pr(E_i \cap E_j \cap E_k) + \cdots + (-1)^{n+1} \Pr(\cap_{i=1}^{n} E_i).\end{aligned}$$

More rigorously, the probability can be defined as a normalized measure. A $\sigma$-field $\mathcal{F}$ (on $S$) is a collection of subsets of $S$ satisfying the following conditions:

(1) It is not empty: $S \in \mathcal{F}$;
(2) If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$;
(3) If $E_1, E_2, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} E_i \in \mathcal{F}$.

From these properties, it follows that the $\sigma$-algebra is also closed under countable intersections (by applying De Morgan's laws).

Then the probability is a normalized measure defined on the $\sigma$-algebra satisfying the three axioms. Usually we denote as a triple $(S, \mathcal{F}, \Pr)$.

## 3. Conditional Probability and Bayes' Formulae

For events $E$ and $F$, the conditional probability of $E$ given that $F$ has occurred is denoted by $\Pr(E|F)$ and is defined by

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

It can be rewritten as a multiplication rule of probability

$$\Pr(E \cap F) = \Pr(F)\Pr(E|F).$$

If $\Pr(E \cap F) = \Pr(E)\Pr(F)$, then we say that the events $E$ and $F$ are independent. This condition is equivalent to $\Pr(E|F) = \Pr(E)$ and to $\Pr(F|E) = \Pr(F)$. Thus, the events $E$ and $F$ are independent if knowledge of the occurrence of one of them does not affect the probability of the other.

The conditional independence of $E$ and $F$ given $A$ is defined as

$$\Pr(E \cap F|A) = \Pr(E|A)\Pr(F|A).$$

Conditional independence does not imply independence, and independence does not imply conditional independence.

A valuable identity is

$$\Pr(E) = \Pr(E|F)\Pr(F) + \Pr(E|F^c)\Pr(F^c)$$

which can be used to compute $\Pr(E)$ by "conditioning" on whether $F$ occurs. The formulae can be generalized to a set of mutually exclusive and exhaustive events (a partition of the sample space $S$) $F_1, \ldots, F_n$ which is known as the law of total probability

$$\Pr(E) = \sum_{i=1}^{n} \Pr(E|F_i)\Pr(F_i). \tag{1}$$

**Bayes's Formulae.** Let $H$ denote the Hypothesis and $E$ as the Evidence. The Bayes' formulae

$$\Pr(H|E) = \frac{\Pr(H \cap E)}{\Pr(E)} = \frac{\Pr(E|H)\Pr(H)}{\Pr(E|H)\Pr(H) + \Pr(E|H^c)\Pr(H^c)}.$$

The ordering in the conditional probability is switched since in practice the information is known for $\Pr(E|H)$ not $\Pr(H|E)$.

For a set of mutually exclusive and exhaustive events (a partition of the sample space $S$) $F_1, \ldots, F_n$, the formula is

$$\Pr(F_j|E) = \frac{\Pr(E \cap F_j)}{\Pr(E)} = \frac{\Pr(E|F_j)\Pr(F_j)}{\sum_{i=1}^{n} \Pr(E|F_i)\Pr(F_i)}.$$

If the events $F_i, i = 1, \ldots, n$, are competing hypotheses, then Bayes's formula shows how to compute the conditional probabilities of these hypotheses when additional evidence $E$ becomes available.

**Example 3.1.** The color of a person's eyes is determined by a single pair of genes. If they are both blue-eyed genes, then the person will have blue eyes; if they are both brown-eyed genes, then the person will have brown eyes; and if one of them is a blue-eyed gene and the other a brown-eyed gene, then the person will have brown eyes. (Because of the latter fact, we say that the brown-eyed gene is dominant over the blue-eyed one.) A newborn child independently receives one eye gene from each of its parents, and the gene it receives from a parent is equally likely to be either of the two eye genes of that parent. Suppose that Smith and both of his parents have brown eyes, but Smith's sister has blue eyes.

(1) Suppose that Smith's wife has blue eyes. What is the probability that their first child will have brown eyes?
(2) If their first child has brown eyes, what is the probability that their next child will also have brown eyes?

The answers are $2/3$ and $3/4$. Let the three events be $F_1 = (br, br), F_2 = (br, bl), F_3 = (bl, br)$. The second question is different with the first one due to the Bayes's formulae. After the evidence $E =$'their first child has brown eyes', the hypothesis $\Pr(F_j|E)$ is changed. The conditional probability of $A =$ 'their next child will also have brown eyes' given $E$ can be computed by the law of total probability

$$\Pr(A|E) = \sum_{i=1}^{3} \Pr(A|(F_i|E)) \Pr(F_i|E).$$

Note that $\Pr(A) = \sum_{i=1}^{3} \Pr(A|F_i) \Pr(F_i) = 2/3$. The new evidence $E$ increases the probability of having brown eyes since $E$ is in favor of that fact.

The *odds* of an event $A$ are defined by

$$\frac{\Pr(A)}{\Pr(A^c)} = \frac{\Pr(A)}{1 - \Pr(A)}.$$

If the odds are equal to $\alpha$, then it is common to say that the odds are '$\alpha$ to 1' in favor of the hypothesis. For example, if $\Pr(A) = 2/3$, then the odds are 2.

Consider now a hypothesis $H$ and a new evidence $E$ is introduced. The new odds after the evidence $E$ has been introduced are

$$\frac{\Pr(H|E)}{\Pr(H^c|E)} = \frac{\Pr(H)}{\Pr(H^c)} \frac{\Pr(E|H)}{\Pr(E|H^c)}. \tag{2}$$

The *posterior odds* of $H$ are the likelihood ratio times the *prior odds*.

## 4. Random Variables

A real-valued function defined on the sample space is called a random variable (RV), i.e., $X : S \to \mathbb{R}$. The event $\{X \le x\}$ is a subset of $S$.

Two functions in the classical senses are associated to a random variable. The *cumulative distribution function* (CDF) $F : \mathbb{R} \to [0, 1]$ is defined as

$$F(x) = \Pr\{X \le x\},$$

which is an increasing and right continuous function, and the *density function* (for a continuous RV) is

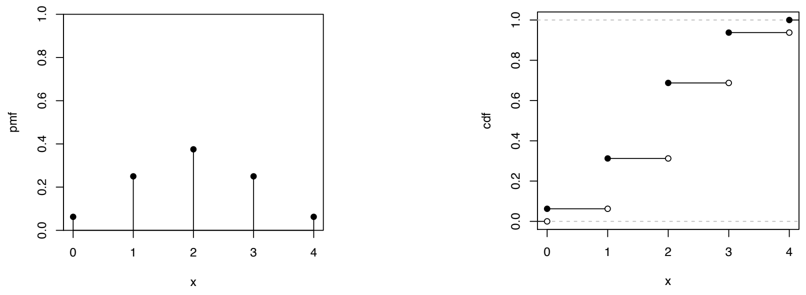
$$f(x) = F'(x).$$

All probabilities concerning $X$ can be stated in terms of $F$.

Random variables can be classified into: discrete RV and continuous RV. A random variable whose set of possible values is either finite or countably infinite is called discrete. If $X$ is a discrete random variable, then the function

$$p(x_i) = \Pr\{X = x_i\}$$

is called the *probability mass function* (PMF) of $X$ and $p(x) = F'(x)$ is understood in the distribution sense; see Fig 4.



(a) A PMF

(b) A CDF

FIGURE 1. A PMF and CDF for a discrete random variable.

The *expectation*, commonly called the *mean*, $\mathbb{E}[X]$ is

$$\mathbb{E}[X] = \sum_i x_i p(x_i).$$

The *variance* of a random variable $X$, denoted by $\mathrm{Var}(X)$, is defined by

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X].$$

It is a measure of the spread of the possible values of $X$. The quantity $\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of $X$.

An important property of the expected value is the linearity of the expectation

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

It can be proved using another equivalent definition of expectation

$$\mathbb{E}[X] = \sum_{s \in S} X(s) p(s).$$

**Important Discrete Random Variables.** We list several discrete random variables corresponding to the sampling schemes.

TABLE 2. Distribution for Sampling Schemes

| | Replace | No Replace |
|---|---|---|
| Fixed $n$ trials | Binomial $(n, p)$ (Bern if $n = 1$) | Hypergeometric |
| Draw until $r$ success | Negative Binomial (Geom if $r = 1$) | Negative Hypergeometric |

Consider in total $n$ trials with $r$ successful trials. Each trial results in a success with probability $p$. Denoted by the triple $(n, r, p)$. If $n$ is fixed, $r$ is the random variable called binomial random variable. When $n = 1$, it is Bernoulli random variable. If $r$ is fixed, $n$ is the random variable called negative binomial random variable. A special case is $r = 1$ called geometric random variable.

The case without replacement is better understood as drawing balls from a bin. Consider in total $N$ balls of which $m$ are white and $N - m$ are black. If drawing white is considered as successful, then the probability of success is $p = m/N$ when only one ball is chosen. If the selected ball is put back for each sampling, then it is the case considered before. Now without replacement, a sample of size $n$ is chosen and ask for $r$ are white. Denoted by $(n, r, N, m)$. If $n$ is fixed, $r$ is the random variable called hypergeometric random variable. If $r$ is fixed, $n$ is the random variable called negative hypergeometric random variable.

We explain these distributions in detail below.

- $\mathrm{Bin}(n, p)$. Binomial random variable with parameters $(n, p)$
$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad i = 0, \dots, n.$$
Such a random variable can be interpreted as being the number of success that occur when $n$ independent trials, each of which results in a success with probability $p$, are performed. Its mean and variance are given by
$$\mathbb{E}[X] = np, \quad \mathrm{Var}(X) = np(1-p).$$
The special case $n = 1$ is called Bernoulli random variable and denoted by $\mathrm{Bern}(p)$.
- $\mathrm{NBin}(r, p)$. Negative binomial random variable with parameters $(r, p)$
$$p\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r} \quad n = r, r+1, \dots$$
Suppose that independent trials, each having probability $p, 0 < p < 1$, of being a success are preformed until a total of $r$ successes is accumulated. The random variable $X$ equals the number of trials required. In order for the $r$th success to occur at the $n$th trial, there must be $r - 1$ successes in the first $n - 1$ trials and the $n$th trial must be a success.
- $\mathrm{Geom}(p)$. Geometric random variable with parameter $p$
$$p(i) = p(1-p)^{i-1} \quad i = 1, 2, \dots$$
Such a random variable represents the trial number of the first success when each trial is independently a success with probability $p$. Its mean and variance are
$$\mathbb{E}[X] = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$
- $\mathrm{HGeom}(n, N, m)$. Hypergeometric random variable with parameters $(n, N, m)$
$$\Pr\{X = i\} = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad i = 0, 1, \dots, n.$$
An urn contains $N$ balls, of which $m$ are white and $N - m$ are black. A sample of size $n$ is chosen randomly from this urn. $X$ is the number of white balls selected.

- $\mathrm{NHGeom}(N, m, r)$. Negative hypergeometric random variables with integer parameters $(N, m, r)$

  $$\Pr\{X = n\} = \frac{\binom{m}{r-1}\binom{N-m}{n-r}}{\binom{N}{n-1}} \frac{m-(r-1)}{N-(n-1)} \quad n = 0, 1, \ldots, m.$$

  An urn contains $N$ balls, of which $m$ are special and $N - m$ are ordinary. These balls are removed one at a time randomly. The random variable $X$ is equal to the number of balls that need be withdrawn until a total of $r$ special balls have been removed.

  To obtain the probability mass function of a negative hypergeometric random variable $X$, not that $X$ will equal $n$ if both
  (1) the first $n - 1$ withdrawals consist of $r - 1$ special and $n - r$ ordinary balls
  (2) the $k$th ball withdrawn is special

When $p \ll 1$ and $n \gg 1$, the $\mathrm{Bin}(n, p) \approx \mathrm{Pois}(\lambda)$ defined below.
$\mathrm{Pois}(\lambda)$. Poisson random variable with parameter $\lambda$

$$p(i) = \frac{e^{-\lambda}\lambda^i}{i!} \quad i \geq 0.$$

If a large number of (approximately) independent trials are performed, each having a small probability of being successful, then the number of successful trials that result will have a distribution which is approximately that a Poisson random variable. The mean and variance of a Poisson random variable are both equal to its parameter $\lambda$.

## 5. CONTINUOUS RANDOM VARIABLES

A random variable $X$ is continuous if there is a nonnegative function $f$, called the probability density function of $X$, such that, for any set $B \subset \mathbb{R}$,

$$\Pr\{X \in B\} = \int_B f(x)\,\mathrm{d}x.$$

If $X$ is continuous, then its distribution function $F$ will be differentiable and

$$\frac{d}{dx}F(x) = f(x).$$

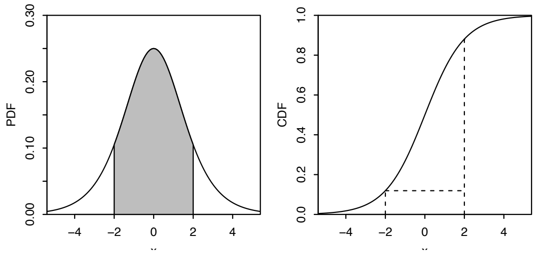


FIGURE 2. PDF and CDF functions of a continuous random variable.

The expected value of a continuous random variable $X$ is defined by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

For a nonnegative random variable $X$, the following formulae can be proved by interchanging the order of integration

$$\mathbb{E}[X] = \int_0^{\infty} \Pr\{X > x\}\,\mathrm{d}x = \int_0^{\infty}\int_x^{\infty} f(y)\,\mathrm{d}y\,\mathrm{d}x. \tag{3}$$

The Law of the Unconscious Statistician (LOTUS) is that, for any function $g$,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x,$$

which can be proved using (3) and again interchanging the order of integration.

**Important Continuous Random Variables.**

- $\mathrm{Unif}(a,b)$. Uniform random variable over the interval $(a,b)$

  $$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

  Its expected value and variance are

  $$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

  Universality of Uniform (UoU). For any continuous random variable, we can transform it into a uniform random variable and vice versa. Let $X$ be a continuous random variable with CDF $F_X(x)$. Then

  $$F_X(X) \sim \mathrm{Unif}(0,1). \tag{4}$$

  Indeed for any increasing function $g$ and for any $a$ in the range of $g$,

  $$\Pr\{g(X) < a\} = \Pr\{X < g^{-1}(a)\} = F_X(g^{-1}(a)),$$

  which implies (4) by choosing $g = F_X$.

  Similarly if $U \sim \mathrm{Unif}(0,1)$, then

  $$F_X^{-1}(U) \sim X.$$

- $N(\mu,\sigma)$. Normal random variable with parameters $(\mu,\sigma^2)$

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

  It can be shown that

  $$\mu = \mathbb{E}[X], \quad \sigma^2 = \mathrm{Var}(X).$$

  Then transformed random variable

  $$Z = \frac{X-\mu}{\sigma}$$

  is normal with mean $0$ and variance $1$. Such a random variable is said to be a standard normal random variable. Probabilities about $X$ can be expressed in terms of probabilities about the standard normal variable $Z$.

The identity $\int_{\mathbb{R}} f = 1$ for the normal distribution can be proved as follows:

$$I^2 = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\,\mathrm{d}x \int_{\mathbb{R}} e^{-\frac{1}{2}y^2}\,\mathrm{d}y = \int_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y.$$

The last integral can be easily evaluated in the polar coordinate.

- $\mathrm{Expo}(\lambda)$. Exponential random variable with parameter $\lambda$

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

Its expected value and variance are, respectively,

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

A key property possessed only by exponential random variables is that they are memoryless, in the sense that, for positive $s$ and $t$,

$$\Pr\{X > s + t | X > t\} = \Pr\{X > s\}.$$

If $X$ represents the life of an item, then the memoryless property states that, for any $t$, the remaining life of a $t$-year-old item has the same probability distribution as the life of a new item. Thus, one need not remember the age of an item to know its distribution of remaining life. In summary, a product with an $\mathrm{Expro}(\lambda)$ lifetime is always ‘as good as new’.

- $\Gamma(\alpha, \lambda)$. Gamma distribution with parameters $(\alpha, \lambda)$

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

where the gamma function is defined by

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

The gamma distribution often arises, in practice as the distribution of the amount of time one has to wait until a total of $n$ events has occurred. When $\alpha = 1$, it reduces to the exponential distribution.

- $\beta(a, b)$. Beta distribution with parameters $(a, b)$

$$f(x) = \begin{cases} \dfrac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where the constant $B(a, b)$ is given by

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x.$$

- $\chi^2(n)$ Chi-square distribution. Let $Z_i \sim N(0, 1)$ for $i = 1, \ldots, n$. Then the square sum of $Z_i$, i.e., $X = \sum_{i=1}^n Z_i^2$ is called the $\chi^2(n)$ distribution and

$$X \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

- Weibull distribution with parameters $(\nu, \alpha, \beta)$ The Weibull distribution function
$$F(x) = \begin{cases} 0 & x \leq \nu \\ 1 - \exp\left\{ -\left( \frac{x-\nu}{\alpha} \right)^{\beta} \right\} & x > \nu. \end{cases}$$
and the density function can be obtained by differentiation.

  It is widely used in the field of life phenomena as the distribution of the lifetime of some object, especially when the "weakest link" model is appropriate for the object.
- Cauchy distribution with parameter $\theta$
$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2}, \quad -\infty < x < \infty.$$

## 6. Jointly Distributed Random Variables

A random variable is like a single variable function $S \to \mathbb{R}$. Given two random variables, we can the obtain a multivariable function $S^2 \to \mathbb{R}^2$.

### 6.1. **Jointly Distribution.**

The *joint cumulative probability distribution function* of the pair of random variables $X$ and $Y$ is defined by
$$F(x, y) = \Pr\{X \leq x, Y \leq y\} \quad -\infty < x, y < \infty.$$
All probabilities regarding the pair can be obtained from $F$. To find the individual probability distribution functions of $X$ and $Y$, use
$$F_X(x) = \lim_{y\to\infty} F(x, y), \quad F_Y(y) = \lim_{x\to\infty} F(x, y).$$

If $X$ and $Y$ are both discrete random variables, then their *joint probability mass function* is defined by
$$p(i, j) = \Pr\{X = i, Y = j\}.$$
The individual mass functions are
$$\Pr\{X = i\} = \sum_j p(i, j), \quad \Pr\{Y = j\} = \sum_i p(i, j).$$
If we list $p(i, j)$ as a table, then the above mass functions are called marginal PMF.

The random variables $X$ and $Y$ are said to be *jointly continuous* if there is a function $f(x, y)$, called the *joint probability density function*, such that for any two-dimensional set $C$,
$$\Pr\{(X, Y) \in C\} = \iint_C f(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$
If $X$ and $Y$ are jointly continuous, then they are individually continuous with density functions
$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x.$$

6.2. **Summation of Independent Random Variables.** The random variables $X$ and $Y$ are *independent* if, for all sets $A$ and $B$,

$$\Pr\{X \in A, Y \in B\} = \Pr\{X \in A\} \Pr\{Y \in B\}.$$

In terms of the joint distribution or the joint density function, for independent random variables, they can be factorized as

$$\begin{aligned} F(x,y) &= F_X(x)F_Y(y) \\ p(x,y) &= p_X(x)p_Y(y) \\ f(x,y) &= f_X(x)f_Y(y). \end{aligned}$$

If $X$ and $Y$ are independent continuous random variables, then the distribution function of their sum can be obtained from the identity

$$F_{X+Y}(a) = \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy.$$

The density function is the convolution

$$f_{X+Y} = f_X * f_Y.$$

Examples of sums of independent random variables. We assume $X_1, X_2, \ldots, X_n$ are independent random variables of the same type.

- $X_i = N(\mu_i, \sigma_i^2)$:
$$\sum_{i=1}^n X_i = N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$
- $X_i$ is Poisson RV with parameter $\lambda_i$:
$$\sum_{i=1}^n X_i \text{ is Poisson with parameter } \sum_{i=1}^n \lambda_i.$$
- $X_i$ is binomial RV with parameter $(n_i, p)$:
$$\sum_{i=1}^n X_i \text{ is binomial with parameter } \left(\sum_{i=1}^n n_i, p\right).$$
- $X_i$ is continuous Gamma distribution with parameter $(t_i, \lambda)$:
$$\sum_{i=1}^n X_i \text{ is Gamma distribution with parameter } \left(\sum_{i=1}^n t_i, \lambda\right).$$

## 7. Properties of Expectation and Variance

7.1. **Expectation of Sums of Random Variables.** If $X$ and $Y$ have a joint probability mass function $p(x,y)$ or a joint density function $f(x,y)$, then.

$$\begin{aligned} \mathbb{E}[g(X,Y)] &= \sum_y \sum_x g(x,y)p(x,y) \\ \mathbb{E}[g(X,Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) \, \mathrm{d}x \, \mathrm{d}y. \end{aligned}$$

It can be proved by simply switching the order of integration. A consequence of the preceding equations is that

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y],$$

which generalizes to

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]. \tag{5}$$

By definition $\mathbb{E}[cX] = c\mathbb{E}[X]$. Therefore $\mathbb{E}[\cdot]$ is linear.
**Question**: How about the expectation of the product of random variables? Do we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]? \tag{6}$$

The answer is NO in general and YES if $X$ and $Y$ are independent. In general, if $X$ and $Y$ are independent, then, for any functions $h$ and $g$,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]. \tag{7}$$

This fact can be easily proved by the separation of variables in the corresponding integral as the joint density function of independent variables are separable.

Using (5), we can decompose a random variable into summation of simple random variables. For example, the binomial random variable can be decomposed into sum of Bernoulli random variables.

Two important and interesting examples:

- a random walk in the plane;
- complexity of the quick-sort algorithm.

7.2. **Moments of the Number of Events that Occur.** Express a random variable as a combination of indicator random variables. For an event $A$, the indicator random variable $I_A = 1$ if $A$ occurs and 0 otherwise. Then $I_A \sim \mathrm{Bern}(p)$ and $\mathbb{E}[I_A] = \Pr(A)$. For two events $A$ and $B$, we have the properties

- $I_A I_B = I_{A\cap B}$;
- $I_{A\cup B} = I_A + I_B - I_A I_B$.

For given events $A_1, \ldots, A_n$, let $X$ be the number of these events that occur. If we introduce an indicator variable $I_i$ for even $A_i$, then

$$X = \sum_{i=1}^{n} I_i,$$

and consequently

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_i] = \sum_{i=1}^{n} \Pr(A_i).$$

Now suppose we are interested in the number of pairs of events that occur. Then

$$\binom{X}{2} = \sum_{i<j} I_i I_j.$$

Taking expectations yields

$$\mathbb{E}\left[\binom{X}{2}\right] = \sum_{i<j} \mathbb{E}[I_i I_j] = \sum_{i<j} \Pr(A_i A_j)$$

giving that

$$\mathbb{E}[X^2] - \mathbb{E}[X] = 2\sum_{i<j} \Pr(A_i A_j).$$

Moreover, by considering the number of distinct subsets of $k$ events that all occur, we have

$$\mathbb{E}\left[\binom{X}{k}\right] = \sum_{i_1<i_2<\ldots<i_k} \mathbb{E}[I_{i_1} I_{i_1} \cdots I_{i_k}] = \sum_{i_1<i_2<\ldots<i_k} \Pr(A_{i_1} A_{i_1} \cdots A_{i_k}).$$

7.3. **Covariance, Variance of Sums, and Correlations.** The *covariance* between $X$ and $Y$, denoted by $\mathrm{Cov}(X,Y)$, is defined by

$$\mathrm{Cov}(X,Y)=\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]=\mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y].$$

$\mathrm{Cov}(\cdot,\cdot)$ is a symmetric bilinear functional and $\mathrm{Cov}(X,X)=\mathrm{Var}(X)>0$. That is $\mathrm{Cov}(\cdot,\cdot)$ defines an inner product on a quotient subspace of the space of random variables. Let us identify this subspace. First of all, we restrict to the subspace of random variables with finite second moment. Second, we identify two random variables if they differ by a constant. The obtained quotient space is isomorphic to the subspace of random variables with finite second moment and mean zero; on that subspace, the covariance is exactly the $L^2$ inner product of real-valued functions.

The *correlation* of two random variables $X$ and $Y$, denoted by $\rho(X,Y)$, is defined by

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

The correlation coefficient is a measure of the degree of linearity between $X$ and $Y$. A value of $\rho(X,Y)$ near $+1$ or $-1$ indicates a high degree of linearity between $X$ and $Y$, whereas a value near $0$ indicates that such linearity is absent. If $\rho(X,Y)=0$, then $X$ and $Y$ are said to be *uncorrelated*. From the inner product point of view, $\rho(X,Y)=\cos\theta(X,Y)$ contains the angle information between $X$ and $Y$ and uncorrelation is equivalent to the orthogonality.

By the definition, we have a precise characterization of the question (6):

$$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]\qquad\Longleftrightarrow\qquad X,Y\text{ are uncorrelated.}$$

If $X$ and $Y$ are independent, then $\mathrm{Cov}(X,Y)=0$. However, the converse is not true. That is two random variables could be dependent but uncorrelated. An example is: $X\sim N(0,1)$ and $Y=X^2$, then $X$ and $Y$ are dependent but uncorrelated. When both $X$ and $Y$ are normal, then the converse is true. Namely two normal random variables are uncorrelated iff they are independent.

Writing

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right)=\mathrm{Cov}\left(\sum_{i=1}^n X_i,\sum_{j=1}^n X_j\right),$$

we get a formula on the variance of summation of random variables

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right)=\sum_{i=1}^n\mathrm{Var}(X_i)+2\sum_{i<j}\mathrm{Cov}(X_i,X_j).$$

In particular, if $X_i$ are independent, we can exchange the sum and Var, i.e.,

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right)=\sum_{i=1}^n\mathrm{Var}(X_i),\ \text{when } X_i \text{ are independent.}$$

Note that $\mathrm{Var}(\cdot)$ is not linear but quadratic to the scaling. By definition

$$\mathrm{Var}(cX)=\mathrm{Cov}(cX,cX)=c^2\mathrm{Cov}(X,X)=c^2\,\mathrm{Var}(X).$$

7.4. **Moment Generating Functions.** The moment generating function $M(t)$ of the random variable $X$ is defined for $t$ in some open interval containing 0 as

$$M(t) = \mathbb{E}[e^{tX}].$$

We call $M(t)$ the moment generating function because all of the moments of $X$ can be obtained by successively differentiating $M(t)$ and then evaluating the result at $t = 0$. Specifically, we have

$$M^{(n)}(0) = \mathbb{E}[X^n] \quad n \geq 1.$$

When $X$ and $Y$ are independent, using (7), we know

$$M_{X+Y}(t) = M_X(t) M_Y(t). \tag{8}$$

In particular,

$$M_{aX+b} = \mathbb{E}\left[e^{taX}e^{tb}\right] = e^{tb}\mathbb{E}\left[e^{taX}\right] = e^{bt}M_X(at).$$

For the standardized version $Z = (X - \mu)/\sigma$, we have the relation

$$M_X(t) = e^{\mu t} M_Z(\sigma t).$$

An important result is that the moment generating function uniquely determines the distribution. This result leads to a simple proof that the sum of independent normal (Poisson, gamma) random variables remains a normal (Poisson, gamma) random variable by using (8) to compute the corresponding moment generating function.

Moments describes the shape of a distribution. Recall that $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$ which implies $f(x) \sim o(1/x)$ as $|x| \to \infty$. If the even moments $(2k)$ is finite, it implies $f(x)$ decay faster than $1/x^{2k+1}$.

7.5. **The Sample Mean and Variance.** Let $X_1, \ldots, X_n$ be i.i.d (independent and identically distributed) random variables having distribution function $F$ and expected value $\mu$. Such a sequence of random variables is said to constitute a sample from the distribution $F$. The quantity

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the *sample mean*. By the linearity of expectation, $\mathbb{E}\left[\overline{X}\right] = \mu$. When the distribution mean $\mu$ is unknown, the sample mean is often used in statistics to estimate it.

The quantities $X_i - \overline{X}_i, i = 1, \ldots, n$, are called *deviations*: the differences between the individual data and the sample mean. The random variable

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

is called the *sample variance*. Then

$$\operatorname{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[S^2\right] = \sigma^2.$$

The sample mean and the sample variance are independent. The sample mean and a deviation from the sample mean are uncorrelated, i.e.,

$$\operatorname{Cov}(X_i - \overline{X}, \overline{X}) = 0.$$

Although $\overline{X}$ and the deviation $X_i - \overline{X}$ are uncorrelated, they are not, in general independent. A special exception is $X_i$ are normal random variables.

The sample mean $\overline{X}$ is a normal random variable with mean $\mu$ and variance $\sigma^2/n$; the random variable $(n-1)S^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ is a Chi-squared random variable with

$n-1$ degrees of freedom which explains the denominator is $n-1$ not $n$ in the definition of $S$.

## 8. Limit Theorems

The most important theoretical results in the probability theory are limit theorems. Of these, the most important are:

- *Laws of large numbers*. The average of a sequence of random variables converges (in certain topology) to the expected average.
- *Central limit theorems*. The sum of a large number random variables has a probability distribution that is approximately normal.

8.1. **Tail bound.** Taking expectation of the inequality

$$\chi(\{X \geq a\}) \leq X/a,$$

we obtain the Markov's inequality.

**Markov's inequality**. Let $X$ be a *non-negative* random variable, i.e., $X \geq 0$. Then, for any value $a > 0$,

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}.$$

**Chebyshev's inequality**. If $X$ is a random variable with finite mean $\mu$ and variance $\sigma^2$, then, for any value $a > 0$,

$$\Pr\{|X-\mu| \geq a\} \leq \frac{\sigma^2}{a^2}.$$

Just apply Markov's inequality to the non-negative RV $(X-\mu)^2$.

The importance of Markovs and Chebyshevs inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known.

If we know more moment of $X$, we can obtain more effective bounds. For example, if $r$ is a nonnegative even integer, then applying Markov's inequality to $X^r$

$$\Pr\{|X| \geq a\} = \Pr\{X^r \geq a^r\} \leq \frac{\mathbb{E}[X^r]}{a^r}, \tag{9}$$

a bound that falls off as $1/a^r$. The larger the $r$, the greater the rate is, given a bound on $\mathbb{E}[X^r]$ is available. If we write the probability $\bar{F}(a) := \Pr\{X > a\} = 1 - \Pr\{X \leq a\} = 1 - F(a)$, then the bound (9) tells how fast the function $\bar{F}$ decays. The moments $\mathbb{E}[X^r]$ is finite implies the PDF $f$ decays faster than $1/x^{r+1}$ and $\bar{F}$ decays like $1/x^r$.

On the other hand, as Markov and Chebyshevs inequalities are valid for all distributions, we cannot expect the bound on the probability to be very close to the actual probability in most cases.

Indeed the Markov's inequality is useless when $a \leq \mu$ and Chebyshev's inequality is useless when $a \leq \sigma$ since the upper bound will be greater than or equal to one. We can improve the bound to be strictly less than one.

**One-sided Chebyshev inequality**. If $\mathbb{E}[X] = \mu$ and $\operatorname{Var}(X) = \sigma^2$, then, for any $a > 0$,

$$\Pr\{X \geq \mu + a\} \leq \frac{\sigma^2}{\sigma^2 + a^2},$$

$$\Pr\{X \leq \mu - a\} \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

*Proof.* Let $b > 0$ and note that $X \geq a$ is equivalent to $X + b \geq a + b$. Hence

$$\Pr\{X \geq a\} = \Pr\{X + b \geq a + b\} \leq \Pr\{(X + b)^2 \geq (a + b)^2\}.$$

Upon applying Markov's inequality, the preceding yields that

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}[(X + b)^2]}{(a + b)^2} = \frac{\sigma^2 + b^2}{(a + b)^2}.$$

Letting $b = \sigma^2/a$, which minimizes the upper bound, gives the desired result. □

When the moment generating function is available (all moments are finite), we have the Chernoff bound which usually implies exponential decay of the tail.

**Chernoff bounds**.

$$\Pr\{X \geq a\} \leq e^{-ta}M(t) \quad \text{for all } t > 0,$$
$$\Pr\{X \leq a\} \leq e^{-ta}M(t) \quad \text{for all } t < 0.$$

We can obtain the best bound by using the $t$ that minimizes the upper bound $e^{-ta}M(t)$.

**Example 8.1** (Chernoff bounds for the standard normal distribution)**.** Let $X \sim N(0, 1)$ be the standard normal distribution. Then $M(t) = e^{t^2/2}$. So the Chernoff bound is given by

$$\Pr\{X \geq a\} \leq e^{-ta}e^{t^2/2} \quad \text{for all } t > 0.$$

The minimum is achieved at $t = a$ which gives the exponential decay tail bound

$$\Pr\{X \geq a\} \leq e^{-a^2/2} \quad \text{for all } a > 0. \tag{10}$$

Similarly, we get the tail bound in the left

$$\Pr\{X \leq a\} \leq e^{-a^2/2} \quad \text{for all } a < 0.$$

For $X \sim N(0, \sigma)$, the tail bound becomes

$$\Pr\{X \geq a\} \leq e^{-\frac{1}{2}\frac{a^2}{\sigma^2}} \quad \text{for all } a > 0. \tag{11}$$

The smaller $\sigma$ is, the faster the decay is.

8.2. **The Central Limit Theorem.** The central limit theorem is one of the most remarkable results in the probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal.

**Theorem 8.2** (The central limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having finite mean $\mu$ and variance $\sigma^2$. Then the distribution of*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

*tends to the standard normal as $n \to \infty$.*

Equivalently $\overline{X} \to N(\mu, \sigma^2/n)$. Here recall that the sample mean $\overline{X} = \sum_{i=1}^n X_i/n$. The variance of $\overline{X}$ scales like $\mathcal{O}(1/n)$ and will approach to zero which implies the law of large numbers by Chernoff bounds.

The central limit result can be extended to independent but may not identical random variables.

**Theorem 8.3** (Central limit theorem for independent random variables)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables having finite mean $\mu_i$ and variance $\sigma_i^2$. If*

*(1) the $X_i$ are uniformly bounded;*
*(2) $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$,*

*then the distribution of*

$$\Pr\left\{\sum_{i=1}^{n}(X_i - \mu_i)\Big/\sqrt{\sum_{i=1}^{n}\sigma_i^2} \le a\right\} \to \Phi(a) \quad \textit{as } n \to \infty.$$

The application of the central limit theorem to show that measurement errors are approximately normally distributed is regarded as an important contribution to science. Indeed, in the 17th and 18th centuries the central limit theorem was often called the *law of frequency of errors*. The central limit theorem was originally stated and proven by the French mathematician Pierre-Simon, Marquis de Laplace.

8.3. **Law of Large Numbers.**

**Theorem 8.4** (The weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having finite mean $\mathbb{E}[X_i] = \mu$. Then, for any $\varepsilon > 0$,*

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \varepsilon\right\} \to 0 \quad \textit{as} \quad n \to \infty.$$

*Proof.* Let $\overline{X} = \sum_{i=1}^{n} X_i/n$. Then by the linearity of expectation, $\mathbb{E}\left[\overline{X}\right] = \mu$. Since $X_i$ are i.i.d., the variance is additive to independent variables and scale quadratically to the constant scaling, we have $\operatorname{Var}(\overline{X}) = \sigma^2/n$. Apply Chebyshev’s inequality to get

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \varepsilon\right\} \le \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{as} \quad n \to \infty.$$

□

**Theorem 8.5** (The strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having finite mean $\mathbb{E}[X_i] = \mu$. Then, with probability $1$, $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$ as $n \to \infty$, i.e.*

$$\Pr\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right\} = 1.$$

What is the difference between the weak and the strong laws of large numbers? In the weak version, $n$ depends on $\varepsilon$ while the strong version is not.

As an application of the strong law of large numbers, suppose that a sequence of independent trials of some experiment is performed. Let $E$ be a fixed event of the experiment, and denote by $\Pr(E)$ the probability that $E$ occurs on any particular trial. Let $X_i$ be the indicator random variable of $E$ on the $i$th trial. We have, by the strong law of large numbers, that with probability $1$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X] = \Pr(E).$$

Therefore, if we accept the interpretation that “with probability 1” means “with certainty,” we obtain the theoretical justification for the long-run relative frequency interpretation of probabilities.

The weak law of large numbers was originally proven by James Bernoulli for the special case where the $X_i$ are Bernoulli random variables. The general form of the weak law of large numbers was proved by the Russian mathematician Khintchine.

The strong law of large numbers was originally proven, in the special case of Bernoulli random variables, by the French mathematician Borel. The general form of the strong law was proven by the Russian mathematician A. N. Kolmogorov.

include a sketch of proofs

## APPENDIX

Formulae

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n,$$

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x}\, \mathrm{d}x,$$

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, \mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \ln n + 0.577$$

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

TABLE 3. Table of Distributions

| Distribution | PMF/PDF | Mean | Variance | MGF |
|---|---|---|---|---|
| Binomial | $\Pr\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$ | $np$ | $np(1-p)$ | $(1 - p + pe^t)^n$ |
| Poisson | $\Pr\{X = k\} = \frac{k!}{e^{-\lambda}\lambda^k}$ | $\lambda$ | $\lambda$ | $e^{\lambda(e^t - 1)}$ |
| Uniform | $f(x) = 1/(b-a)$ | $(a+b)/2$ | $(b-a)^2/12$ | $\frac{e^{tb} - e^{ta}}{t(b-a)}$ |
| Normal | $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ | $e^{t\mu + \sigma^2 t^2/2}$ |
| Gamma | $f(x) = \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}/\Gamma(\alpha), x \geq 0$ | $\alpha/\lambda$ | $\alpha/\lambda^2$ | $\left(\frac{\lambda}{\lambda - t}\right)^{\alpha}, t < \lambda$ |
| Chi-square | $x^{n/2-1}e^{-x/2}/(2^{n/2}\Gamma(n/2)), x > 0$ | $n$ | $2n$ | $(1-2t)^{-n/2}, t < 1/2$ |