STA2004 Mathematical Statistics Lecture 1 - 2022-2023 Term 2

Lecture 1: Probability Review

Lecturer: Yunyi Zhang, Liyan Xie, Zhenxing Guo

Suggested Reading: Agresti and Kateri 2021, Chapter 2.

1 Probability Basics

1.1 Language of Probability

The basic terminologies in probability:

1. **Sample space**: The set containing all possible outcomes, usually denoted as Ω, e.g., $\Omega = \{\text{outcomes of flipping a coin}\} = \{\text{Head,Tail}\}$.
2. **Event**: a subset of Ω, e.g., $E = \{\text{Head}\}$.
3. **Empty event**: $E = \emptyset$; and **Non-empty event**: $E \neq \emptyset$.

Event Operations and Terminologies:

- Event Operation:
  1. **Union**: $A \cup B$ occurs if either A or B occurs.
  2. **Intersection**: $A \cap B$ occurs if both A and B occurs.
  3. **Complement**: A^c occurs if A does NOT occur.
- Event Terminologies:
  1. **Disjoint**: Two events A and B are called disjoint if $A \cap B = \emptyset$, i.e., they do not happen simultaneously.
  2. **Mutually exclusive**: A group of events $E_1, E_2, \ldots, E_n$ are mutually exclusive if any two events are disjoint, i.e., $E_i \cap E_j = \emptyset$ for any $i \neq j$.
  3. **Exhaustive**: A group of events $E_1, E_2, \ldots, E_n$ are exhaustive if the union of them is the entire sample space: $E_1 \cup E_2 \cup \cdots \cup E_n = \Omega$. Thus at least one of the events $E_1, E_2, \ldots, E_n$ will occur.
  4. **Partition**: A group of events $E_1, E_2, \ldots, E_n$ is called a partition of Ω if the events are mutually exclusive and exhaustive. Thus exactly one and ONLY one of the events $E_1, E_2, \ldots, E_n$ will occur.

Definition 1.1 (Probability). *A probability on the sample space Ω is an assignment of a value, $\mathbf{P}(E)$, to each event E such that:*

1. $\mathbf{P}(E) \geq 0$ *for any event* E.
2. $\mathbf{P}(\Omega) = 1$.
3. *For any sequence of mutually exclusive events* $E_1, E_2, \ldots$ *(i.e., events for which* $E_i \cap E_j = \emptyset$ *for any* $i \neq j$*),*

$$\mathbf{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbf{P}(E_i).$$

Definition 1.2 (Independence). *Two events A and B are called independent if and only if*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

1.2 Conditional Probability

When we have some prior information about the outcome, then the calculation of probability of its occurrence should be restricted to a smaller sample space based on the available information. Thus we introduce the concept of conditional probability.

Definition 1.3 (Conditional Probability). *For any two events A and B, the conditional probability of A given the occurrence of B is written as $\mathbf{P}(A|B)$ and is defined as*

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

provided that $\mathbf{P}(B) > 0$.

Theorem 1.1 (Law of Total Probability). *If* $0 < \mathbf{P}(B) < 1$, *then*

$$= P(A \cap B) + P(A \cap B^c)$$

$$\mathbf{P}(A) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|B^c)\mathbf{P}(B^c)$$

for any event A. Generally, if $B_1, B_2, \ldots, B_k$ are exclusive and exhaustive events (i.e., a partition of the sample space Ω), then for any event A,

$$= \sum_{j=1}^{k} P(A \cap B_j)$$

$$\mathbf{P}(A) = \sum_{j=1}^{k} \mathbf{P}(A|B_j)\mathbf{P}(B_j).$$

Theorem 1.2 (Bayes' Theorem or Bayes' Rule). *For any two events A and B with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$,*

$$\mathbf{P}(B|A) = \mathbf{P}(A|B)\frac{\mathbf{P}(B)}{\mathbf{P}(A)}.$$

$$= \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

2 Random Variables

2.1 Basic Definitions and Examples

A random variable is a mathematical formalization of a quantity or object which depends on random events.

e.g. $\Omega = \{H, T\}$ $\begin{cases} X(H)=1 \\ X(T)=0 \end{cases}$

Definition 2.1 (Random variables). *A random variable $X : \Omega \to \mathbf{R}$ is a numerical valued function defined on a sample space Ω. In other words, a number $X(\omega)$ is assigned to each outcome ω in the sample space.*

Conventionally, we use capital letters $X, Y, \ldots$ to denote the random variables and small letters $x, y, \ldots$ to denote the possible numerical values (or realizations) of these variables. We describe a random variable X via its distribution functions. Random variables usually have two categories: discrete random variable and continuous random variable.

A discrete random variable is one which may take on only a *countable* number of distinct values such as $0, 1, 2, 3, 4, \ldots$. The formal definition is as follows.

Definition 2.2 (Discrete random variables). *A random variable X defined on the sample space Ω is called a discrete random variable if $X(\Omega) := \{X(\omega) : \omega \in \Omega\}$ is countable (e.g., $X : \Omega \to \{0, 1, 2, \ldots\}$).*

The probability distribution of a discrete random variable is a list of probabilities associated with each possible values. It is also called the probability function or the probability mass function (pmf).

Definition 2.3 (Probability mass function). *The probability mass function (pmf) of a discrete random variable X is defined as*

$$p(x) = \mathbf{P}(X = x), \quad x \in X(\Omega),$$

where $X(\Omega)$ is the countable set of possible values of X. A valid pmf satisfies:

1. *$p(x) \geq 0$ for any $x \in X(\Omega)$;*
2. *$\sum_{x \in X(\Omega)} p(x) = 1$;*
3. *$\mathbf{P}(X \in A) = \sum_{x \in A} p(x)$ for any $A \subset X(\Omega)$.*

Examples of discrete random variables:

- Bernoulli distribution $X \sim Bernoulli(p)$: $p(x) = \mathbf{P}(X = x) = p^x(1-p)^{1-x}$, $x = 0, 1$. mgf: $1 - p + pe^t$
- Binomial distribution $X \sim B(n, p)$: $p(x) = \mathbf{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$. mgf: $(1-p+pe^t)^n$ (summation of n Bernoulli)
- Geometric distribution $X \sim Geo(p)$: $p(x) = \mathbf{P}(X = x) = p(1-p)^{x-1}$, $x = 1, 2, \ldots$. cdf: $1-(1-p)^x$ mgf: $\frac{pe^t}{1-(1-p)e^t}$

(summation of r Geometric)

mgf: $\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$

- Negative Binomial $X \sim NB(r,p)$: $p(x) = \mathbf{P}(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$, $x = r, r+1, \ldots$.
- Poisson distribution $X \sim Poi(\lambda)$: $p(x) = \mathbf{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, \ldots$. mgf: $e^{\lambda(e^t-1)}$
- Uniform distribution: $p(x_i) = \mathbf{P}(X = x_i) = \frac{1}{n}$, for $i = 1, 2, \ldots, n$.

Definition 2.4 (Cumulative distribution function). *The cumulative distribution function (cdf) of a discrete random variable X is defined as*

$$F(x) = \mathbf{P}(X \le x) = \sum_{t \le x} p(t), \quad -\infty < x < \infty.$$

The cdf $F(x)$ is non-decreasing and satisfies $F(-\infty) = 0$, $F(+\infty) = 1$. As can be seen from the definition, the cdf of a discrete random variable would be a *step-function* with $p(x)$ as the size of the jumps at the possible value x.

A continuous random variable is a random variable whose cumulative distribution function is continuous everywhere. The probability distribution of a continuous random variable is a function called the probability density function (pdf). The formal definition is as follows.

Definition 2.5 (Continuous random variable). *A random variable X is called a continuous random variable if its cumulative distribution function $F(x) = \mathbf{P}(X \le x)$ has the form*

$$F(x) = \int_{-\infty}^{x} f(t)dt, \quad -\infty < x < \infty,$$

for some function $f : \mathbf{R} \to [0, \infty)$. And the function f is called the probability density function (pdf) of X. The pdf satisfies $\int_{-\infty}^{\infty} f(t)dt = 1$.

Examples of continuous random variables:

$\Rightarrow X \sim Gamma(\alpha=1, \theta=\frac{1}{\lambda})$ cdf: $1-e^{-\lambda x}$

- Exponential distribution $X \sim Exp(\lambda)$: pdf $f(x) = \lambda e^{-\lambda x}$, $x > 0$. mgf: $\frac{\lambda}{\lambda - t}$
- Gamma distribution $X \sim Gamma(\alpha, \theta)$: $f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$, $0 \le x < \infty$, where $\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy$. $\Gamma(1)=1$; $\Gamma(\frac{1}{2})=\sqrt{\pi}$; $\Gamma(\alpha)=(\alpha-1)!\,\Gamma(\alpha-1)$, $\alpha \ge 2$; $\Gamma(n)=(n-1)!$, n is integer
- Normal (Gaussian) distribution $X \sim N(\mu, \sigma^2)$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, $x \in \mathbf{R}$.

$\Rightarrow X \sim Gamma(\alpha=\frac{r}{2}, \theta=2)$

- Chi-square distribution $X \sim \chi^2(r)$: $\frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}$, $0 < x < \infty$. $Z_1, Z_2, \ldots, Z_r \overset{iid}{\sim} N(0,1)$, $\sum_{i=1}^r Z_i \sim \chi^2(r)$
- Uniform distribution: $f(x) = \frac{1}{b-a}, a \le x \le b$.

2.2 Mean and Variance

We define the expectation (mean) as follow:

Definition 2.6 (Expectation). *(i). Suppose a* **discrete** *random variable X is defined on* $\mathbf{Z}$ *with* **probability mass function** *p, i.e., $p(x) = \mathbf{P}(X = x)$, then for any real value function $h(\cdot)$, define the mathematical expectation of $h(X)$ as:*

$$\mathbf{E}[h(X)] = \sum_{x \in X(\Omega)} [h(x)p(x)]. \tag{1}$$

(ii). Suppose a **continuous** *random variable X is defined on* $\mathbf{R}$ *with* **probability density function** *$f(x)$, then for any real value function $h(\cdot)$, define*

$$\mathbf{E}[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx \tag{2}$$

as the mathematical expectation of $h(X)$ if it exists.

In this class, we assume that all of the considered expectation exists. A special case of definition 1 gives the formal definition of expectation and variance of a random variable:

$h(X) = X$

Definition 2.7 (Mean and Variance). *For a given random variable X, $\mathbf{E}[X]$ is called the (population) mean of X, and is usually denoted by μ. Furthermore, $\mathbf{E}[(X-\mu)^2]$ is called the (population) variance of X and is usually denoted by σ^2 or $\mathrm{Var}(X)$. The positive square root of the variance, $\sigma = \sqrt{\mathrm{Var}(X)}$, is called the (population) standard deviation of X (of the distribution).*

$\mathrm{Var}(X) = E(X^2) - (E(X))^2$

Example 1 Bernoulli distribution: The pmf of Bernoulli distribution is:

$$p(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

(i). Choose $h(x) = x$, we have

$$\mathbf{E}[h(X)] = \mathbf{E}[X] = h(1) \times p + h(0) \times (1-p) = p.$$

(ii). Choose $h(x) = x^2$, we have

$$\mathbf{E}[h(X)] = \mathbf{E}[X^2] = h(1) \times p + h(0) \times (1-p) = p.$$

And thus the variance of the Bernoulli distribution is $\mathrm{Var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = p(1-p)$.

(iii). Choose $h(x) = e^x$, we have

$$\mathbf{E}[h(X)] = \mathbf{E}[e^X] = h(1) \times p + h(0) \times (1-p) = pe + 1 - p.$$

Example 2 **Uniform distribution:** The pdf of a continuous uniform distribution in $[0,1]$ is:

$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

similarly, choose $h(x) = x$,

$$\mathbf{E}[h(X)] = \mathbf{E}[X] = \int_{-\infty}^{\infty} h(x) \times f(x)dx = \int_0^1 xdx = 1/2.$$

Choose $h(x) = x^2$, then

$$\mathbf{E}[h(X)] = \mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 \times f(x)dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

Choose $h(x) = e^x$, then

$$\mathbf{E}[h(X)] = \mathbf{E}[e^X] = \int_0^1 e^x dx = e - 1.$$

2.3 Properties of Expectation

Theorem 2.1 (Linearity of expectation)**.** *(i). For any given random variables X, any function g, h and fixed number $a, b \in \mathbf{R}$, we have*

$$\mathbf{E}[ah(X) + bg(X)] = a\mathbf{E}[h(X)] + b\mathbf{E}[g(X)]. \tag{3}$$

(ii). More generally, for any given numbers $a_1, ..., a_n \in \mathbf{R}$, and function $h_i(\cdot)$, we have

$$\mathbf{E}\left(\sum_{i=1}^{n} a_i h_i(X)\right) = \sum_{i=1}^{n} a_i \mathbf{E}[h_i(X)]. \tag{4}$$

Proof Consider two situations: the discrete and continuous random variables. The proof is given separately as below:

Discrete: By definition

$$\begin{aligned} \mathbf{E}[ah(X) + bg(X)] &= \sum_{x \in X(\Omega)} (ah(x) + bg(x)) \times p(x) = \sum_{x \in X(\Omega)} ah(x) \times p(x) + \sum_{x \in X(\Omega)} bg(x) \times p(x) \\ &= a \sum_{x \in X(\Omega)} h(x) \times p(x) + b \sum_{x \in X(\Omega)} g(x) \times p(x) \\ &= a\mathbf{E}[h(X)] + b\mathbf{E}[g(X)]. \end{aligned}$$

Continuous: By definition

$$\begin{aligned}\mathbf{E}[ah(X)+bg(X)] &= \int_{-\infty}^{\infty}(ah(x)+bg(x))\cdot f(x)dx \\ &= \int_{-\infty}^{\infty} ah(x)f(x)dx + \int_{-\infty}^{\infty} bg(x)f(x)dx \\ &= a\int_{-\infty}^{\infty} h(x)f(x)dx + b\int_{-\infty}^{\infty} g(x)f(x)dx \\ &= a\mathbf{E}[h(X)] + b\mathbf{E}[g(X)].\end{aligned}$$

□

2.4 Moment Generating Function and Indicator Function

Definition 2.8 (Moment Generating Function). *The moment generating function (MGF) of random variable X is defined as*

$$M_X(t) = \mathbf{E}[e^{tX}].$$

And the cumulant generating function is defined as

$$R_X(t) = \ln M_X(t).$$

Theorem 2.2 (Properties of MGF). *The r-th order derivative of MGF at 0 equals to the r-th order moment of the random variable X:*

$$M_X^{(r)}(0) = \frac{d^r}{dt^r}M_X(t)|_{t=0} = \mathbf{E}[X^r].$$

Moreover, the cumulant generating function satisfies $\mu = \mathbf{E}[X] = R'_X(0)$ *and* $\sigma^2 = \mathrm{Var}[X] = R''_X(0)$.

$E[X] = M'_X(0)$ $\quad$ $E[X^2] = M''_X(0)$

Proof For discrete RV X, we have

$$M_X(t) = \mathbf{E}[e^{tX}] = \sum_{x\in X(\Omega)} e^{tx}p(x) \Rightarrow M'_X(0) = \sum_{x\in X(\Omega)} e^0 xp(x) = \mathbf{E}[X],$$

and

$$M'_X(t) = \sum_{x\in X(\Omega)} e^{tx}xp(x) \Rightarrow M''_X(0) = \sum_{x\in X(\Omega)} e^0 x^2 p(x)\mathbf{E}[X^2].$$

Moreover,

$$R'_X(t) = \frac{M'_X(t)}{M_X(t)} \Rightarrow R'_X(0) = M'_X(0) = \mu = \mathbf{E}[X],$$

and

$$R''_X(t) = \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{(M_X(t))^2} \Rightarrow R''_X(0) = M''_X(0) - (M'_X(0))^2 = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \mathrm{Var}[X] = \sigma^2.$$

Homework: prove for general order r and for continuous random variables. □

Definition 2.9 (Indicator function). *For a given set A, define the indicator function*

$$\mathbf{1}_A(x) = \begin{cases} 1, \text{ if } x \in A \\ 0, \text{ otherwise.} \end{cases} \tag{5}$$

Then, we define the probability

$$\mathbf{P}(X \in A) = \mathbf{E}[\mathbf{1}_A(X)]. \tag{6}$$

Example 3 Suppose X is drawn from a continuous uniform distribution in $[0, 1]$, i.e., $f(x) = 1$ if $x \in [0, 1]$ and 0 other wise.

(i). Choose $A = [0, 1/4] \cup [3/4, 1]$, then

$$\mathbf{P}(X \in A) = \int_0^1 \mathbf{1}_{[0,1/4]} dx + \int_0^1 \mathbf{1}_{[3/4,1]} dx = 1/2.$$

(ii). Consider $A = (-\infty, x]$, then the cdf of X:

$$F(x) = \mathbf{P}(X \in A) = \int_0^1 \mathbf{1}_{(-\infty,x]} dx = \begin{cases} 0, \text{ if } x \leq 0 \\ x, \text{ if } 0 < x \leq 1 \\ 1, \text{ if } x > 1 \end{cases}.$$

2.5 Concentration Inequalities

Theorem 2.3 (Markov Inequality). *Suppose random variable $X \geq 0$, i.e., the pmf $p(x) = 0$ or pdf $f(x) = 0$ for $x < 0$, then for any given $a > 0$, we have*

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}. \tag{7}$$

Proof **Discrete:** by definition

$$\mathbf{P}(X \geq a) = \mathbf{E}\mathbf{1}_{[a,\infty)}(X) = \frac{1}{a} \times \sum_{t \geq a} (a \times p(t))$$
$$\leq \frac{1}{a} \times \sum_{t \geq a} (t \times p(t)) \leq \frac{1}{a} \times \sum_{t \in X(\Omega)} (t \times p(t)) = \frac{\mathbf{E}[X]}{a}.$$

Continuous: Homework. □

Corollary 2.1. *Define $\mu = \mathbf{E}[X]$, then for any $a > 0$,*

$$\mathbf{P}\left(|X - \mu| > a\right) \leq \frac{\mathrm{Var}(X)}{a^2}. \tag{8}$$

Proof Homework. □

Theorem 2.4. *Suppose X is a continuous random variable. Besides, its cumulative distribution function F_X is strictly increasing in $\mathbf{R}$. Then $F_X(X)$ has uniform distribution on $[0,1]$, i.e.,*

$$\mathbf{P}\left(F_X(X) \leq x\right) = \begin{cases} 0, & \text{if } x \leq 0 \\ x, & \text{if } x \in [0,1] \\ 1, & \text{if } x > 1 \end{cases}$$

Proof Define $Y = F_X(X)$. By definition $Y \in [0,1]$, so $F_Y(y) = 0$ for $y \leq 0$ and 1 for $y > 1$. For any $y \in [0,1]$

$$\mathbf{P}\left(Y \leq y\right) = \mathbf{P}\left(F_X(X) \leq y\right) = \mathbf{P}\left(X \leq F_X^{-1}(y)\right) = F_X(F_X^{-1}(y)) = y.$$

Therefore, Y has uniform distribution on $[0,1]$. □

Definition 2.10 (Quantile and median). *Suppose X is a continuous random variable with cumulative distribution function F_X. If there is a number $q_{1-\alpha}$ such that $F_X(q_{1-\alpha}) = 1 - \alpha$, then we call $q_{1-\alpha}$ the $1-\alpha$ quantile of X. We define X's median as $q_{1/2}$.*

3 Multivariate Distributions

3.1 Joint and Marginal Distributions

When two or more random variables are observed simultaneously in an experiment, not only individual probabilistic behaviours but also the degree of relationship among the variables is investigated.

Definition 3.1. *The joint probability mass function (joint pmf) of the discrete random variables X and Y and is defined by*

$$p(x,y) = \mathbf{P}(X = x, Y = y), \quad x \in X(\Omega), y \in Y(\Omega).$$

Conditions for a joint pmf

- $0 \leq p(x,y) \leq 1$ for all $x \in X(\Omega)$ and $y \in Y(\Omega)$.
- $\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} p(x,y) = 1$.

- $\mathbf{P}((X,Y) \in A) = \sum_{(x,y)\in A} p(x,y)$ where $A \subset X(\Omega) \times Y(\Omega)$.

Definition 3.2. *Let X and Y be discrete random variables with joint pmf $p(x,y)$. The marginal pmfs of X and Y are respectively defined as*

$$\mathbf{P}_X(x) = \mathbf{P}(X = x) = \sum_{y\in Y(\Omega)} p(x,y),$$

and

$$\mathbf{P}_Y(y) = \mathbf{P}(Y = y) = \sum_{x\in X(\Omega)} p(x,y).$$

Definition 3.3. *We say that X and Y are jointly continuous if there exists a function $f(x,y)$ defined for all real x and y, having the property that for every (measurable) set C in the two-dimensional plane,*

$$\mathbf{P}\{(X,Y) \in C\} = \int\int_{(x,y)\in C} f(x,y)dxdy.$$

This function, if exists, is called the joint probability density function (joint pdf).

Joint pmf can uniquely determine the marginal pmfs, but the converse is not true.

Example 4 **Same marginal but different joint distributions**

- $\mathbf{P}_1(X,Y)$.

	Value of Y				Total
Value of X	0	1	2	3	
0	0.0454	0.1818	0.1364	0.0182	0.3818
1	0.1364	0.2727	0.0818	0	0.4909
2	0.0682	0.0545	0	0	0.1227
3	0.0045	0	0	0	0.0045
Total	0.2545	0.5091	0.2182	0.0182	1.0000

- $\mathbf{P}_2(X,Y)$

	Value of Y				Total
Value of X	0	1	2	3	
0	0.0972	0.1944	0.0833	0.0069	0.3818
1	0.1249	0.2499	0.1071	0.0089	0.4909
2	0.0312	0.0625	0.0268	0.0022	0.1227
3	0.0011	0.0023	0.0010	0.0001	0.0045
Total	0.2545	0.5091	0.2182	0.0182	1.0000

Properties of joint pdf

- $f(x,y) \geq 0$ for all $-\infty < x < \infty$ and $-\infty < y < \infty$.
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$.

- $\mathbf{P}((X,Y) \in A) = \int\int_A f(x,y)dxdy.$

- *Joint distribution function:*

$$F(x,y) = \mathbf{P}(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)dsdt.$$

- *Marginal pdf*

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, -\infty < y < \infty,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, -\infty < x < \infty.$$

3.2 Expectation of Function of Random Variables

Definition 3.4. *For random variables X_1, X_2, ..., X_n (not necessarily independent) with joint pmf $p(x_1, x_2, ..., x_n)$ or joint pdf $f(x_1, x_2, ..., x_n)$; if $u(X_1, X_2, ..., X_n)$ is a function of these random variables, then the expectation of $u(X_1, X_2, ..., X_n)$ is defined as*

- *Discrete*

$$\mathbf{E}[u(X_1, X_2, ..., X_n)] = \sum_{x_1}\sum_{x_2}...\sum_{x_n} u(x_1, x_2, ..., x_n)p(x_1, x_2, ..., x_n).$$

- *Continuous*

$$\mathbf{E}[u(X_1, X_2, ..., X_n)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} u(x_1, x_2, ..., x_n)f(x_1, x_2, ..., x_n)dx_1dx_2...dx_n.$$

Properties

- $\mathbf{E}(X+Y) = \mathbf{E}(X) + \mathbf{E}(Y)$. ($X$ and Y need not be independent.) In general, $\mathbf{E}[u_1(X) + u_2(Y)] = \mathbf{E}[u_1(X)] + \mathbf{E}[u_2(Y)]$.

- If X and Y are independent, then $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. In general, $\mathbf{E}[u_1(X)u_2(Y)] = \mathbf{E}[u_1(X)]\mathbf{E}[u_2(Y)]$. The converse is NOT necessarily true.

- If X and Y are independent, then the moment generating function of $X+Y$ is equal to the product of the moment generating functions of X and Y, i.e.,

$$M_{X+Y}(t) = M_X(t)M_Y(t),$$

and also for the linear combination $aX + bY$, the moment generating function is given by

$$M_{aX+bY}(t) = M_X(at)M_Y(bt).$$

- If $u(X_1, X_2, ..., X_n) = g(X_1)$ (i.e., u is a function of X_1 only), then the expectation of g can be obtained by the marginal pmf/pdf of X_1. That is, (for discrete version)

$$\begin{aligned}\mathbf{E}[g(X_1)] &= \sum_{x_1}\sum_{x_2}...\sum_{x_n} g(x_1)p(x_1, x_2, ..., x_n) \\ &= \sum_{x_1} g(x_1)\sum_{x_2}\sum_{x_3}...\sum_{x_n} p(x_1, x_2, ..., x_n) \\ &= \sum_{x_1} g(x_1)p_{X_1}(x_1)\end{aligned}$$

3.3 Independence, Population Covariance and Correlation

Definition 3.5 (Independence). *Two random variables X and Y are said to be independent if and only if their joint pmf (pdf) is equal to the product of their marginal pmfs (pdfs), i.e.,*

$$p(x, y) = p_X(x)p_Y(y), for\ \forall x, y,\ if\ X, Y\ are\ discrete;$$

or,

(独立的等价条件)

$$f(x, y) = f_X(x)f_Y(y),\ for\ \forall x, y,\ if\ X, Y\ are\ continuous.$$

Theorem 3.1. *Let X and Y be random variables with joint pdf (or pmf) $f(x, y)$. Then X and Y are independent if and only if*

- *the supports of X and Y do not depend on each other (i.e., the region of possible values is a rectangle); and*
- *$f(x, y)$ can be factorized as $g(x)h(y)$.* (normalize之后 $g(x) = f_X(x)$, $h(y) = f_Y(y)$)

The definitions of joint pdf (pmf), marginal pdf (pmf), and their properties can be generalized to multivariate case directly.

Definition 3.6 (Population Covariance). *Let X and Y be random variables with means μ_X and μ_Y, respectively. The population covariance between X and Y is defined as*

$$\sigma_{XY} = \text{Cov}(X, Y) = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbf{E}(XY) - \mu_X\mu_Y.$$

Note that a more correct notation should be $\sigma_{X,Y}$, but usually by convention σ_{XY} is used as long as it is not confused with the standard deviation of the product XY. *X and Y are called uncorrelated if* $\text{Corr}(X, Y) = 0$, *i.e.,* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Notably, independence implies un-correlation, but un-correlation does not imply independence. Examples (Homework).

Properties of the covariance

1. The magnitude of σ_{XY} depends on the scales of X and Y. Let $X' = aX + b$ and $Y' = cY + d$ where a and c are non-zero constants. Then

$$\mathrm{Cov}(X', Y') = \mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y).$$

$= E\{[aX+b-(aE(X)+b)]\cdot[cY+d-(cE(Y)+d)]\}$
$= acE[(X-E(X))\cdot(Y-E(Y))]$

2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
3. $\mathrm{Cov}(X, c) = 0$ for any constant c.
4. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
5. $\mathrm{Cov}(\sum_i a_i X_i, \sum_j b_j Y_j) = \sum_i \sum_j a_i b_j \mathrm{Cov}(X_i, Y_j)$.
6. $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, Y_j)$.

Definition 3.7 (Coefficient of Correlation). *Let X and Y be random variables with covariance σ_{XY}, standard deviations $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$. The population correlation coefficient between X and Y is defined as*

$$\rho = \rho_{XY} = \mathrm{Corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Similar to covariance, a more correct notation should be $\rho_{X,Y}$.

The sign and the magnitude of ρ_{XY} reveal the direction and the strength of the linear relationship between X and Y.

Theorem 3.2 (Cauchy-Schwarz Inequality). *Let X and Y be random variables with finite second moments. Then*

$$[\mathbf{E}(XY)]^2 \leq \mathbf{E}(X^2)\mathbf{E}(Y^2).$$

The equality holds if and only if either $\mathbf{P}(X = 0) = 1$ or $\mathbf{P}(Y = aX) = 1$ for some constant a, i.e., X and Y are proportional.

Proof For any $t \in R$ and random variables X and Y, consider

$$0 \leq \mathbf{E}[(tX + Y)^2] = t^2 * \mathbf{E}(X^2) + 2t * \mathbf{E}(XY) + \mathbf{E}(Y^2).$$

Prof can be finished by plugging in $t = -\frac{E(XY)}{E(X^2)}$. □

Properties of the correlation coefficient

1. $-1 \leqslant \rho \leqslant 1$
2. ρ is invariant under linear transformation of X and Y.

3. If X and Y are independent, then

$$\text{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(X)\mathbf{E}(Y) - \mathbf{E}(X)\mathbf{E}(Y) = 0,$$

and hence $\rho = 0$ if X and Y are not constants.

4. If X and Y are independent, then

$$\text{Var}(X+Y) = \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y).$$

Remarks

1. The converse of property 3 need not be true. That is, $\rho = 0$ does not imply X and Y are independent. Special case: X, Y 服从 joint Normal distribution: $\rho = 0 \Rightarrow$ X, Y independent

2. Correlation coefficient measures the strength of the linear relationship only. It may be possible that X and Y are strongly related but that the relation is curvilinear, and ρ would be nearly zero.

3. An observed correlation may be due to a third unknown casual variable.

3.4 Conditional Distributions

Definition 3.8. *Let (X,Y) be a discrete bivariate random vector with joint pmf $p(x,y)$ and marginal pmfs $p_X(x)$ and $p_Y(y)$. For any x such that $p_X(x) = \mathbf{P}(X = x) > 0$, the conditional pmf of Y given that $X = x$ is the function of y denoted by $p_{Y|X}(y|x)$ and is defined by*

$$p_{Y|X}(y|x) = \mathbf{P}(Y = y|X = x) = \frac{\mathbf{P}(Y = y, X = x)}{\mathbf{P}(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

On the other hand, for any y such that $p_Y(y) = \mathbf{P}(Y = y) > 0$, the conditional pmf of X given that $Y = y$ is the function of x denoted by $p_{X|Y}(x|y)$ and is defined by

$$p_{X|Y}(x|y) = \mathbf{P}(X = x|Y = y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Remarks

1. If X is independent of Y, then $\mathbf{P}_{X|Y}(x|y) = p_X(x)$ for $\forall y$. This holds vise versa.

2. For continuous variables (X, Y) with joint pdf $f(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \quad f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

3. conditional distribution function

$$F_{Y|X}(y|x) = \mathbf{P}(Y \le y|X = x) = \begin{cases} \sum_{i \le y} \mathbf{P}_{Y|X}(i|x), & \text{for discrete cases} \\ \int_{-\infty}^{y} f_{Y|X}(t|x)dt, & \text{for continuous cases} \end{cases}$$

4. conditional mean

$$\mathbf{E}[g(Y)|X = x] = \begin{cases} \sum_i g(i)\mathbf{P}_{Y|X}(i|x), & \text{for discrete cases} \\ \int_{-\infty}^{\infty} g(t) f_{Y|X}(t|x)dt, & \text{for continuous cases} \end{cases}$$

5. conditional variance

$$\begin{aligned} \mathrm{Var}(Y|X = x) &= \mathbf{E}\{[Y - \mathbf{E}(Y|X = x)]^2|X = x\} \\ &= \mathbf{E}(Y^2|X = x) - [\mathbf{E}(Y|X = x)]^2. \end{aligned}$$

Theorem 3.3 (Law of Total Expectation, or Adam's Law)**.** *If X and Y are two random variables, then for any function u,*

$$\mathbf{E}[u(X)] = \mathbf{E}\{\mathbf{E}[u(X)|Y]\}.$$

When u is the identity function, we have $\mathbf{E}(X) = \mathbf{E}[\mathbf{E}(X|Y)]$.

Theorem 3.4 (Law of Total Variance, or Eve's Law)**.** *If X and Y are two random variables, then*

$$\mathrm{Var}(X) = \mathbf{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathbf{E}(X|Y)].$$

Example 5 Let $X \sim Geo(p)$ and $Y \sim Geo(p)$ be two independent geometric random variables. Find the expected value of the proportion $\frac{X}{X+Y}$.

Solution:

Note that X, $Y \in 1, 2, ...$ Let $N = X + Y \in 2, 3, ...$ The possible values of X given $N = n$ should be from 1 to $n - 1$. By definition of Geometric distribution, we know that $X + Y \sim NB(2, p)$, the negative binomial distribution with $r = 2$ and success probability p, thus $\mathbf{P}(N = n) = \binom{n-1}{1} p^2(1-p)^{n-2} = (n-1)p^2(1-p)^{n-2}$.

Then for $k \in \{1, 2, ..., n-1\}$

$$\begin{aligned}
\mathbf{P}(X = k|N = n) &= \mathbf{P}(X = k|X + Y = n) \\
&= \frac{\mathbf{P}(X = k, X + Y = n)}{\mathbf{P}(X + Y = n)} \\
&= \frac{\mathbf{P}(X = k, Y = n - k)}{\mathbf{P}(X + Y = n)} \\
&= \frac{(1-p)^{k-1}p * (1-p)^{n-k-1}p}{(n-1)p^2(1-p)^{n-2}} \\
&= \frac{1}{n-1}
\end{aligned}$$

Then we have $X|(N = n) \sim DU\{1, 2, ..., n-1\}$ as a discrete uniform distribution. Then

$$E(X|N = n) = \frac{1 + 2 + ... + (n-1)}{n-1} = \frac{n}{2}$$

Thus

$$E[\frac{X}{X+Y}] = E[\frac{X}{N}] = E\{E[\frac{X}{N}|N]\} = E[\frac{1}{N} * \frac{N}{2}] = \frac{1}{2}.$$

3.5 Transformation of Multivariate Distributions

3.5.1 General form

Let $Y_i = g_i(X_1, X_2, ..., X_n), i = 1, 2, ..., n$ for some functions $g_i's$ such that the functions $g_i's$ satisfy the following conditions:

1. The equations $y_i = g_i(x_1, x_2, ..., x_n)$ can be uniquely solved for $x_1, x_2, ..., x_n$ in terms of $y_1, y_2, ..., y_n$ with solutions given by the inverse transformations, say, $x_i = h_i(y_1, y_2, ..., y_n)$, $i = 1, 2, ..., n$, i.e., the transformation from $X's$ to $Y's$ is one-to-one correspondence.

2. The functions $g_i's$ have continuous partial derivatives at all points $(x_1, x_2, ..., x_n)$, and are such that $n \times n$ Jacobian determinant is non-zero,

$$J_0(x_1, x_2, ..., x_n) = |(\frac{\partial g_i}{x_j})_{i,j}| \neq 0.$$

The inverse functions $h_i's$ should have continuous partial derivatives at all points $(y_1, y_2, ..., y_n)$ and the condition is again the Jacobian determinant being non-zero, i.e.,

$$J(y_1, y_2, ..., y_n) = |(\frac{\partial h_i}{y_j})_{i,j}| \neq 0.$$

Then, the joint pdf of $Y_1, Y_2, ..., Y_n$ is given by the following formula:

$$f_Y(y_1, y_2, ..., y_n) = f_X(x_1, x_2, ..., x_n) \times |J0(x_1, x_2, ..., x_n)|^{-1}$$

$$= f_X(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots y_n)) \cdot |J_0(x_1, \cdots, x_n)|^{-1}$$

or

$$f_Y(y_1, y_2, ..., y_n) = f_X(x_1, x_2, ..., x_n) \times |J(y_1, y_2, ..., y_n)|.$$

$$= f_X(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots y_n)) \cdot |J(x_1, \cdots, x_n)|$$

Example 6 Suppose that two random variables X_1, X_2 have a continuous joint distribution for which the joint pdf is as follows:

$$f_X(x_1, x_2) = \begin{cases} \frac{1}{2}(x_1 + x_2)e^{-x_1 - x_2}, & x_1 > 0, x_2 > 0; \\ 0, & \text{Otherwise} \end{cases}$$

Find the joint dist of $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$.

Solution:

Obviously the transformation $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$ is one-to-one correspondence with the inverse transformation $X_1 = (Y_1 + Y_2)/2, X_2 = (Y_1 - Y_2)/2$. The Jacobian determinant is given by

$$J(y_1, y_2) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{vmatrix} = -1/2$$

Then,

$$f_Y(y_1, y_2) = f_X(x_1, x_2) \times |J| = 1/2 y_1 e^{-y_1}| - 1/2| = 1/4 y_1 e^{-y_1},$$

with support $x_1 > 0$, $x_2 > 0$, or $-y_1 < y_2 < y_1, y_1 > 0$. Then

$$f_Y(y_1, y_2) = \begin{cases} \frac{1}{4} y_1 e^{-y_1}, & -y_1 < y_2 < y_1, y_1 > 0; \\ 0, & \text{Otherwise.} \end{cases}$$

3.5.2 Special transformation cases

- Sum of Two Random Variables: $Z = X + Y$.

  Discrete cases

$$p_Z(z) = \mathbf{P}(X + Y = z) = \sum_x p(x, z - x) = \sum_y p(z - y, y).$$

Continuous cases

$$
\begin{aligned}
F_Z(z) &= \mathbf{P}(Z \le z) = \mathbf{P}(X + Y \le z) \\
&= \mathbf{P}(Y \le z - X) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x,y)dxdy \\
&= \mathbf{P}(X \le z - Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x,y)dxdy \\
f_Z(z) &= F'(z) = \int_{-\infty}^{\infty} f(x, z-x)dx = \int_{-\infty}^{\infty} f(z-y, y)dy
\end{aligned}
$$

- Difference between Two Random Variables: $Z = X - Y$.

  Discrete cases

$$p_Z(z) = \mathbf{P}(X - Y = z) = \sum_x p(x, x-z) = \sum_y p(z+y, y).$$

  Continuous cases

$$
\begin{aligned}
F_Z(z) &= \mathbf{P}(Z \le z) = \mathbf{P}(X - Y \le z) \\
&= \mathbf{P}(Y \ge X - z) = \int_{-\infty}^{\infty} \int_{x-z}^{\infty} f(x,y)dxdy \\
&= \mathbf{P}(X \le z + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{z+y} f(x,y)dxdy \\
f_Z(z) &= F'(z) = \int_{-\infty}^{\infty} f(x, x-z)dx = \int_{-\infty}^{\infty} f(z+y, y)dy
\end{aligned}
$$

- Product: $Z = XY$.

  Discrete cases

$$p_Z(z) = \mathbf{P}(XY = z) = \sum_x p(x, z/x) = \sum_y p(z/y, y).$$

  Continuous cases

$$
\begin{aligned}
F_Z(z) &= \mathbf{P}(Z \le z) = \mathbf{P}(XY \le z) \\
&= \mathbf{P}(Y \le z/X, X > 0) + \mathbf{P}(Y \ge z/X, X < 0) \\
&= \int_0^{\infty} \int_{-\infty}^{z/x} f(x,y)dxdy + \int_{-\infty}^{0} \int_{z/x}^{\infty} f(x,y)dxdy \\
f_Z(z) &= F'(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x, z/x)dx = \int_{-\infty}^{\infty} \frac{1}{|y|} f(z/y, y)dy
\end{aligned}
$$

- Ratio: $Z = X/Y$. Discrete cases

$$p_Z(z) = \mathbf{P}(X/Y = z) = \sum_x p(x, x/z) = \sum_y p(zy, y).$$

Continuous cases

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \le z) = \mathbf{P}(X/Y \le z) \\ &= \mathbf{P}(X \le Yz, Y > 0) + \mathbf{P}(X \ge Yz, Y < 0) \\ &= \int_0^\infty \int_{-\infty}^{zy} f(x,y)dxdy + \int_{-\infty}^0 \int_{zy}^\infty f(x,y)dxdy \\ f_Z(z) = F'(z) &= \int_{-\infty}^\infty |y| f(zy,y)dy = \int_{-\infty}^\infty \frac{|x|}{z^2} f(x, x/z)dx \end{aligned}$$

4 Convergence in Probability and Convergence in Distribution

Definition 4.1 (Convergence in probability). *Suppose there is a sequence of random variables $X_n, n = 1, 2, \cdots$ and a random variable X. If for $\forall \epsilon > 0$, $\exists$ a constant $k > 0$ such that for any $n \ge k$,*

$P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$

$$\mathbf{P}\left(|X_n - X| > \epsilon\right) \le \epsilon, \tag{9}$$

then we say X_n converges to X in probability, i.e., $X_n \to_p X$.

In this class, we mainly consider the situation that X being a constant.

Theorem 4.1. *(i). Convergence in probability is preserved under summation and product, i.e., suppose $X_n \to_p X$ and $Y_n \to Y$, then*

$$X_n + Y_n \to_p X + Y \text{ and } X_n \times Y_n \to X \times Y, \tag{10}$$

and suppose $Y = c \neq 0$ is a constant, then

$$X_n / Y_n \to_p X/c. \tag{11}$$

(ii). Convergence in probability is preserved under function composition, i.e., suppose $X_n \to_p X$, then for any continuous function $h(\cdot)$,

$$h(X_n) \to_p h(X). \tag{12}$$

continuous mapping theorem

Proof (i). By definition for any given $\epsilon > 0$, there exists $M > 0$ such that for any $n \ge M$,

$$\mathbf{P}\left(|X_n - X| > \epsilon/2\right) \le \epsilon/2 \text{ and } \mathbf{P}\left(|Y_n - Y| > \epsilon/2\right) \le \epsilon/2$$

. Then

$$\begin{aligned} \mathbf{P}\left(|(X_n + Y_n) - (X + Y)| > \epsilon\right) &\le \mathbf{P}\left(\{|X_n - X| > \epsilon/2\} \cup \{|Y_n - Y| > \epsilon/2\}\right) \\ &\le \mathbf{P}\left(|X_n - X| > \epsilon/2\right) + \mathbf{P}\left(|Y_n - Y| > \epsilon/2\right) \le \epsilon. \end{aligned}$$

Others are left as homework. □

Remark Suppose we have data(random variable) $Z_1, Z_2, \cdots, Z_n, \cdots$, for each sample size define $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$, then $X_n, n = 1, 2, \cdots$ will be a sequence of random variables.

Definition 4.2 (Convergence in distribution). *Suppose we have a sequence of random variables $X_n, n = 1, \cdots$ and a random variable X. Suppose for any bounded continuous function $h(\cdot)$*

$$\mathbf{E}[h(X_n)] \to \mathbf{E}[h(X)] \quad \textit{i.e., numerically convergence} \tag{13}$$

then we say that X_n converges to X in distribution, i.e., $X_n \to_d X$.

Remark Notably, convergence in distribution is not preserved under summation and product. For example, suppose X is a normal random variable with mean 0 and variance 1, set $X_n = X$ and

$$Y_n = \begin{cases} X, & \text{if } n = 2k \\ -X, & \text{if } n = 2k+1 \end{cases}$$

we have $X_n \to_d X$ and $Y_n \to_d X$. However, we have

(i). $X_n + Y_n = 0$ for $n = 2k+1$ and $2X$ for $n = 2k$.

(ii). $X_n \times Y_n = -X^2$ for $n = 2k+1$ and X^2 for $n = 2k$.

Theorem 4.2 (Continuous mapping theorem). *Suppose $X_n \to_d X$ and $f(\cdot)$ is a continuous function. Then*

$$f(X_n) \to_d f(X). \tag{14}$$

Proof For any given continuous bounded function h, the function $h_1(x) = h(f(x))$ is also continuous and bounded. Therefore, by definition

$$\mathbf{E}[h(f(X_n))] \to \mathbf{E}[h(f(X))].$$

and $f(X_n) \to_d f(X)$. □

Corollary 4.1. *Suppose $X_n \to_d X$, then for any **fixed** $x, y \in \mathbf{R}$,*

$$\begin{aligned} \mathbf{P}(X_n \le x) &\to \mathbf{P}(X \le x); \\ \mathbf{P}(X_n < x) &\to \mathbf{P}(X < x); \\ \mathbf{P}(X_n \in [x, y]) &\to \mathbf{P}(X \in [x, y]). \end{aligned} \tag{15}$$

Proof Left as homework. □

Theorem 4.3 (Slutsky's theorem). *Suppose $X_n \to_d X$ and $Y_n \to_p c$, here c is a constant. Then*

(i). $X_n + Y_n \to_d X + c$.

(ii). $X_n Y_n \to_d cX_n$.

(iii). If $c \neq 0$, then $X_n/Y_n \to_d X_n/c$.

5 Law of Large Number & Central Limit Theorem

Theorem 5.1. *Suppose random variables $X_i, i = 1, \cdots$ are i.i.d. (i.e., independent and identically distributed) with mean $\mu = \mathbf{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$. Define $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$, then*

$$Z_n \to_p \mu. \tag{16}$$

Proof For any given $\epsilon > 0$, from Markov inequality

$$\mathbf{P}\left(|Z_n - \mu| > \epsilon\right) = \mathbf{P}\left((Z_n - \mu)^2 > \epsilon^2\right) \leq \frac{1}{\epsilon^2} \times \mathbf{E}(Z_n - \mu)^2$$

For

$$\mathbf{E}(Z_n - \mu)^2 = \frac{1}{n^2}\sum_{i_1=1}^n \sum_{i_2=1}^n \mathbf{E}(X_{i_1} - \mu)(X_{i_2} - \mu) = \frac{1}{n}\mathrm{Var}(X_1)$$

we have for sufficiently large n

$$\mathbf{P}\left(|Z_n - \mu| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} < \epsilon.$$

□

Theorem 5.2 (Central limit theorem, corollary 1.2 in Shao 2003). *Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with $\mu = \mathbf{E}[X_1]$ and $Var(X_1) = \sigma^2 < \infty$. Then*

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1) \qquad \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \to_d N(0, \sigma^2). \tag{17}$$

6 Properties of Normal Distributions

Recall that if a random variable X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{18}$$

then we call X a normal random variable. Especially, if $\mu = 0$ and $\sigma = 1$, then we call X *standard normal* random variable.

Definition 6.1 (χ^2 distribution). *Suppose $X_1, \cdots, X_n$ are independent standard normal random vari-*

ables, $Q = \sum_{i=1}^{n} X_i^2$. *Then the distribution of* Q *is called chi-squared* (χ^2) *distribution with degree of freedom* n*, denoted as* $Q \sim \chi_n^2$.

Definition 6.2 ($t-$distribution(student's distribution)). *Suppose* $Z \sim N(0,1)$ *follows the standard normal distribution,* $W \sim \chi_n^2$ *is the* χ^2 *distribution with degree of freedom* n*,* Z *and* W *are independent. Then the distribution of* $T = \frac{Z}{\sqrt{W/n}}$ *is called the* t *distribution with degree of freedom* n. $n \to \infty$, $T_n \to N(0,1)$

Definition 6.3 (F-distribution). *Suppose* $X \sim \chi_{n_1}^2$ *and* $Y \sim \chi_{n_2}^2$ *are independent* χ^2 *random variables. The distribution of the random variable* $F = \frac{X/n_1}{Y/n_2}$ *is called the* F *distribution with degree of freedom* n_1 *and* n_2.

Theorem 6.1 (Sampling distribution). *Suppose* $X_1, \cdots, X_n$ *are independent and identically distributed normal random variables with* $\mathbf{E}[X_1] = \mu$ *and* $\mathrm{Var}(X_1) = \mathbf{E}(X_1 - \mu)^2 = \sigma^2$*. Define the sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *and the sample variance* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$*. Then:*

(i) $\bar{X}$ *and* S^2 *are independent;*

(ii) The distribution of $\frac{(n-1)S^2}{\sigma^2}$ *is* χ_{n-1}^2*, the* χ^2 *distribution with degree of freedom* $n-1$*;*

(iii) Thus we have $T := \frac{\sqrt{n}(\bar{X}-\mu)}{S}$ *satisfies the* t *distribution with degree of freedom* $n-1$*. (proof left as homework).*

References

Agresti, Alan and Maria Kateri (2021). *Foundations of Statistics for Data Scientists: With R and Python*. Chapman and Hall/CRC.

Shao, Jun (2003). *Mathematical statistics*. Second. Springer Texts in Statistics. Springer-Verlag, New York, pp. xvi+591.