Chapter 1

Probability

1.1 Basic concepts and results

A **random experiment** is when a set of all possible outcomes is known, but it is impossible to predict the actual outcome of the experiment. A **sample space**, denoted as Ω, contains all possible outcomes of the experiment. An **event** is a subset of Ω. We say that $A \subset \Omega$ has occrred if and only if the outcome of the experiment is an element of A. Formally, the family of events forms a σ-algebra of subsets of Ω that we denote by $\mathcal{A}$.

Note:

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A} \Rightarrow \bar{A} \in \bar{\mathcal{A}}$,where $\bar{A}$ indicates the compliment of A
- $A_1, A_2, \cdots \in \mathcal{A}$
- $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

1.1.1 Probability measures

Definition 1.1.1: Kolmogorov's axioms

- $P(A) \geq 0$
- $P(\Omega) = 1$
- If $A_i \bigcap A_j = \emptyset, i \neq j$, then $P(\cup_i A_i) = \sum_i P(A_i)$

Probability measure $P : \mathcal{A} \rightarrow \mathbb{R}$ satisfying Kolmogorov's axioms has the following properties:

- $P(\varnothing) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A - B) = P(A \cap \bar{B}) = P(A) - P(A \cap B)$

Definition 1.1.2: Conditional probability

If $P(B) > 0$,

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

We are re-evaluating the probability of A given the B space.

Let $\{A_1, A_2, \cdots\}$ denote a partition of $\Omega : \cup_i A_i = \Omega$; $A_i \cap A_j = \emptyset, i \neq j$. Meaning union makes up Ω and are mutually exclusive. Then if $P(A_i) > 0$ for all i

Theorem 1.1.1 Total probability theorem

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

$B = B \cap \Omega = B \cap [\cup_i A_i] = \cup_i (B \cap A_i)$ and $P(\cup_i B \cap A_i) = \sum_i P(B \cap A_i)$

Theorem 1.1.2 Bayes' theorem

If $P(B) > 0$

$$P(A_i|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}$$

$$P(\underbrace{A_j}_{\text{explanation}} | \underbrace{B}_{\text{evidence}}) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\underbrace{P(B)}_{\text{substitute with total probability theorem}}}$$

1.1.2 Random variables

Definition 1.1.3: Random variable

Function defined in Ω and taking values in $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

$$\omega \mapsto X(\omega) = x$$

A random variable induces a probability measure in $\mathbb{R}$ that we denote by P_X: if $B \subset \mathbb{R}$, $P_X(B) = P(A)$, where $A = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. Formally, there must be a σ-algebra of subsets of $\mathbb{R}$,$\mathcal{B}$, and we have to verify that for every set $B \in \mathcal{B}$ we have $X^{-1}(B) \in \mathcal{A}$. Typically, $\mathcal{B}$ is the so called Borel σ-algebra and it suffices to make sure that X satisfies $X^{-1}((-\infty, x]) \in \mathcal{A}$,$\forall x \in \mathbb{R}$.

Basically what it means is that we don't know if $X^{-1}(B) \in \mathcal{A}$ and for which B can I compute $P_X(B)$. If $X^{-1}(B) \in \mathcal{A}$ for B is in the Borel σ-algebra, then X is measurable.

Definition 1.1.4: Distribution function of a random variable

X: for all $x \in \mathbb{R}$

$$F_X(x) = P_X((-\infty, x]) = P(X \leq x)$$

It is suffice to know $F_X(\cdot)$ to be able to compute $P_X(B)$ for all $B \in \mathcal{B}$.

- For all $a > b$, $P(a < X \leq b) = F_X(b) - F_X(a)$
- $F_X(-\infty) = 0$; $F_X(\infty) = 1$

- F_X is right-continuous and non-decreasing
- The set of points at which F_X is discontinuous is either finite or countable (at most countable)

Definition 1.1.5: Discrete random variable

X is a discrete random variable if D_X is such that $P_X(D_X) = 1$

The probability mass function of X is defined as $f_X(x) = F_X(x) - \lim_{y \to x} F_X(y) = \begin{cases} P(X = x) & \text{if } x \in D_X \\ 0 & \text{otherwise} \end{cases}$

Any f satisfying the following is a probability mass function

- $f(x) \geq 0$ for all x
- $f(x) > 0$ iff $x \in D$, where $D \subset \mathbb{R}$ is finite or countable
- $\sum_{x \in D} f(x) = 1$

For any event $B \subset \mathbb{R}$, $P(X \in B) = \sum\limits_{x \in B \cap D_X} f_X(x)$.

Note:

$F_X(x) = \sum_{y \leq x} f_X(y)$

$F_X(x) = P(X \leq x)$ cumulative distribution function
$\downarrow$
$f_X(x) = P(X = x)$ probability mass function
where $0 \leq f_X(x) \leq 1$

Discrete distribution include Bernoulli, binomial, Poisson, geometric, negative binomial, multinomial, hypergeometric, etc.

Definition 1.1.6: Continuous random variable

X is continuous if $P_X(D_X) = 0, D_X = \varnothing$ and if additionally there is f_X such that for all $x \in \mathbb{R}$

- $f_X(x) \geq 0 \longrightarrow$ probability density function
- $F_X(x) = \int_{-\infty}^{+\infty} f(x)\, dx = 1$

At the points where F_X is differentiable, we have $F'_X(x) = f_X(x)$.

Any f satisfying the following conditions is a probability density function

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$

Continuous distributions include uniform, exponential, gamma, chi-squared, normal. t-student, F-Snedcor, beta, Pareto, Weibull, log-normal, etc.

1.1.3 Functions of a random variable

Let X be a r.v. and $Y = h(X)$ where $h : \mathbb{R} \longrightarrow \mathbb{R}$
In general, if $X = g(Y)$ with g invertible and differentiable, and X continuous, we have

$$f_Y(y) = |g'(y)|\, f_x(g(y))$$

Proof. $\dfrac{\partial F_X(x)}{\partial x} = f_X(x)$
Using chain rule: $(f \circ g)'(x) = [f(g(x))]' = f'(g(x))g'(x)$ ■

Definition 1.1.7: Expected value

Let $Y = h(X)$, a linear function.

The expected value of Y is defined by $E[Y] = \begin{cases} \sum_x h(x)\, f_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x)\, f_X(x)\, dx & \text{if } X \text{ continuous} \end{cases}$

Formally, we must addtionally verify that the integral or series are absolutely convergent. $E[Y]$ may not exist.

There are two ways to compute $E[Y]$ with $Y = h(X)$, either use the definition above, or first obtain the distribution of Y and compute $E[Y] = \begin{cases} \sum_y y\, f_Y(y) & \text{if } Y \text{ discrete} \\ \int_{-\infty}^{+\infty} y\, f_Y(y)\, dy & \text{if } Y \text{ continuous} \end{cases}$. The two methods are equivalent.

Definition 1.1.8: Raw moment of oder k

$\mu'_k = E[X^k]$

Definition 1.1.9: Central moment of order k

$\mu_k = E[(X-\mu)^k]$, $\mu = E[X]$

Definition 1.1.10: Moment generating function

$M_X(s) = E[e^{sX}]$ whenever the expectation exists for s in a neighborhood of the origin.

- If $M_X(s)$ exists, then X has moments of all orders and $M^{(k)}(0) = E[X^k]$
- The moment generating function, when it exists, identifies the probability distribution

Some useful **properties**:

- $E[h_1(X) + h_2(X)] = E[h_1(X)] + E[h_2(X)]$
- If $c \in \mathbb{R}$, then $E[cX] = cE[X]$; $E[c] = c$
- If $c \in \mathbb{R}$, then $Var(cX + b) = c^2 Var(X)$
- $Var(X) = E[X^2] - (E[X])^2$
- $Var(X) \geq 0$; $Var(X) = 0 \Leftrightarrow P(X = c) = 1$ for some $c \in \mathbb{R}$

1.1.4 Bivariate random variables

$$\begin{aligned}(X, Y) : \Omega &\to \mathbb{R}^2 \\ \omega &\mapsto (X(\omega), Y(\omega)) = (x, y)\end{aligned}$$

If (X, Y) discrete, we define the joint probability mass function as $f(x, y) = P(X = x, Y = y)$. If $(X; Y)$ continuous, then there exists the joint probability density function, $f(x, y)$ such that for all $(x, y) \in \mathbb{R}^2$,

- $f(x, y) \geq 0$
- $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du$

Example 1.1.1

X = weight, Y = height $\Rightarrow$ Z = BMI

Definition 1.1.11: Marginal distributions

$$f_X(x) = \begin{cases} \sum_y f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} f(x,y)\,dy & \text{if } (X,Y) \text{ continuous} \end{cases}$$

Definition 1.1.12: Expectation of $Z = h(X,Y)$

$$E[Z] = \begin{cases} \sum_x \sum_y h(x,y)\,f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x,y)\,f(x,y)\,dy\,dx & \text{if } (X,Y) \text{ continuous} \end{cases}$$

Definition 1.1.13: Conditional didstributions

$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$, y fixed: $f_Y(y) > 0$

function of x for every y where $f_Y(y) > 0$

Definition 1.1.14: Raw moment of order (r,s)

$\mu'_{(r,s)} = E[X^r\,Y^s]$

Definition 1.1.15: Central moment of order (r,s)

$\mu_{(r,s)} = E[(X-\mu_X)^r\,(Y-\mu_Y)^s]$

Definition 1.1.16: Covariance

$\operatorname{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \mu_{(1,1)}$

If x and y are positively associated $\longrightarrow \operatorname{Cov}(x,y) > 0 \longrightarrow$ If x is larger than its mean, then typically y is larger than its mean.

Some useful **properties**:

- $\operatorname{Cov}(X,Y) = E[X,Y] - E[X]E[Y]$
- $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$
- $\operatorname{Cov}(cX,Y) = c\operatorname{Cov}(X,Y), c \in \mathbb{R}$
- $\operatorname{Cov}(X+Y,Z) = \operatorname{Cov}(X,Z) + \operatorname{Cov}(Y,Z)$
- $\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2\operatorname{Cov}(X,Y)$

Example 1.1.2 (Portfolio management)

$\operatorname{Cov}(x,y) < 0$
$\operatorname{Var}(x,y) < \operatorname{Var}(x) + \operatorname{Var}(y)$

Theorem 1.1.3 Law of iterated expectation

If $Z = h(X,Y)$ then $E[Z] = E_X[E[Z|X]]$

Theorem 1.1.4 Law of total variance

$\operatorname{Var}(Y) = \operatorname{Var}_X(E[Y|X]) + E_X[\operatorname{Var}(Y|X)]$

Other useful tricks:

- $E[h(X)\,Y\,|\,X = x] = h(x)\,E[Y\,|\,X = x]$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, E[Y|X])$

  Proof.

$$\begin{aligned}\mathrm{Cov}(X, E[Y|X]) &= E[X\,E[Y|X]] - E[X]\,E[E[Y|X]] \\ &= E[E[XY|X]] - E[X]\,E[Y] \\ &= E[XY] - E[X]\,E[Y] \\ &= \mathrm{Cov}(X, Y)\end{aligned}$$

  ■

1.1.5 Independence

Definition 1.1.17: Stochastic independence

X and Y are stochastically independent if and only if $\forall(x, y) \in \mathbb{R}^2$, $f(x, y) = f_X(x)\,f_Y(y)$

If X and Y are independent, then

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

  Proof. $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2 \times \underbrace{\mathrm{Cov}(X, Y)}_{\to 0}$ ■

- $M_{X+Y}(s) = M_X(s)\,M_Y(s)$

  Proof. $M_{X+Y}(s) = E[e^{s(X+Y)}] = E[\underbrace{e^{sx}}_{u}\ \underbrace{e^{sy}}_{v}]$

  x and y independent stochastically $\Rightarrow$ u and v independent

  $M_{X+Y}(s) = E[e^{sx}]\,E[e^{sy}] = M_X(s)\,M_Y(s)$ ■

- $\mathrm{Cov}(X, Y) = 0$

  Proof. $\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \underbrace{E[XY]}_{X,Y \text{uncorrelated}} - E[X]E[Y] = E[X]E[Y] - E[x]E[Y] = 0$ ■

- $E[X^r Y^s] = E[X^r]\,E[Y^s]$
- $E[Y\,|\,X = x] = E[Y]$; $E[X\,|\,Y = y] = E[X]$
- $f_{X|Y=y}(x) = f_X(x)$; $f_{Y|X=x}(y) = f_Y(y)$

  Proof. $f_{X|Y=y}(x) = \dfrac{f(x, y)}{f_Y(y)} = \dfrac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x)$ ■

Definition 1.1.18: Mean independence

Y is mean independent of X iff $E[Y\,|\,X = x]$ does not depend on x for all x.

Proof. $E[Y|X = x] = c$

$E[Y|X] = c \Rightarrow E[E[Y|X]] = c \Rightarrow E[Y] = c \longrightarrow$ conditional is equal to marginal ■

Definition 1.1.19: Uncorrelatedness

X and Y are uncorrelated iff $\text{Cov}(X, Y) = 0$

Useful **results**:

- If X and Y are stochastically independent, then Y is mean-independent of X, and X is mean independent of Y.
- If Y is mean-independet of X, then X and Y are uncorrelated. The converse is not true.

  Proof. Y mean independence of $X \Rightarrow \text{Cov}(X, Y) = \text{Cov}(X, E[Y|X]) = \text{Cov}(X, c) = 0 \Rightarrow$ uncorrelated ■

- If Y is uncorrelated with X, then $E[XY] = E[X]E[Y]$
- If Y is mean-independent of X, then $E[X^k Y] = E[X^k]E[Y]$ for all k
- If Y and X are stochastically independent, then $E[X^k Y^r] = E[X^k]E[Y^r]$ for all k, r

Note:

stochastic independence $\Rightarrow$ mean independence $\Rightarrow$ uncorrelatedness

1.2 Convergence of sequences of random variables

- Notions of Convergence
  1. Pointwise convergence
  2. Uniform convergence
  3. Convergence in L^P
  4. Convergence in measure
- Convergence for random variables
  1. Almost surely
  2. In the rth mean
  3. In probability
  4. In distribution
- Skorokhod representation theorem
- Continuous mapping theorem
- Slutsky theorem

1.2.1 Notions of convergence

If $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables and X is a random variable,

$$X_n : \underbrace{\Omega}_{\text{exists probability, } \sigma\text{-algebra}} \rightarrow \mathbb{R}$$

$$X_n \longrightarrow X \qquad \text{as } n \longrightarrow +\infty$$

n can be population size, or can be the number of iterations for Monte Carlo simulation.

Notions of **convergence**: let $f_n, f : [0, 1] \longrightarrow \mathbb{R}$

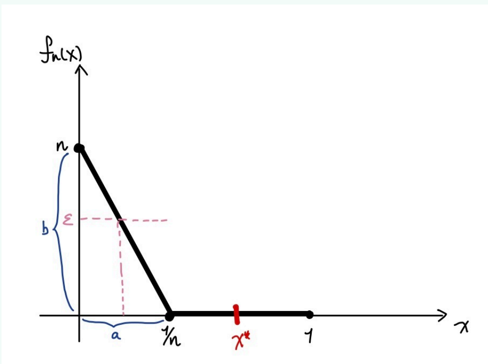
- Point wise convergence: $f_n(x) \longrightarrow f(x)$ for all $x \in [0, 1]$

- Uniform convergence: $\sup_{x\in[0,1]} |f_n(x) - f(x)| \longrightarrow 0$
- Convergence in L^P: $\int_0^1 |f_n(x) - f(x)|^P \, dx \longrightarrow 0$
- Convergence in measure: $\mu(A_{n,\epsilon}) \longrightarrow 0$ for all $\epsilon > 0$ where $A_{n,\epsilon} = \{x \in [0,1] : |f_n(x) - f(x)| > \epsilon\}$

Example 1.2.1

$f_n : [0,1] \longrightarrow \mathbb{R}$

$$f_n(x) = \begin{cases} 0 & 1/n \leq x \leq 1 \\ n - n^2 x & 0 \leq x < 1/n \end{cases}$$



As $n \longrightarrow \infty$, a becomes smaller, b becomes bigger.

- Point wise convergence

$$\forall x \in [0,1]$$

$\forall x^* > 0, f_n(x^*) = 0 \qquad \text{for } n > N \qquad \text{except } f_n(0) = 0 \longrightarrow \infty$

- Uniform convergence
- Convergence in L^1
- Convergence in measure

1.3 Important asymptotic results

- Weak law of large numbers
- Strong law of large numbers
- Central limit theorem
- Lévy's continuity theorem
- Applications
  1. Bernoulli
  2. Simple Monte Carlo
- Delta method and its applications

1. Log odds
2. Variance stabalizing

Chapter 2

Classical Statistical Model

2.1 Probability versus statistical inference

- Probability theory
- Statistical inference

2.2 Model specification

- Random sample
- Sampling
- IID random sampling

2.3 Statistics

- Statistic definition

2.4 Sampling distribution

- Definition
- Methods to obtain the sampling distribution of a statistic
  - Monte Carlo simulation
- Sample distribution of the sample moments
  - Sample moments
  - Properties of the sample mean
  - Properties of the sample variance
  - Properties of the bias-corrected sample variance
  - Properties of central sample moments
  - Asymptotic distribution of $\bar{X}$
- Order statistics