1. Fundamentals of Probability Theory and Mathematical Statistics

Statistics are exploring methodology to derive conclusions from observed data of experiments while any uncertainty exists. Statistical decision and learning based on mathematical statistics creates foundation of state-of-the-art information communication technology (ICT), and of course statistical communication theory. We shall minimize mathematical theory (such as measure theory) to introduce probability theory and mathematical statistics.

1.1 Probability and Radom Variables

Probability has been used to model uncertainty, including one major application of gambling. Classical probability proceeds on intuition. Following progress in modern mathematical analysis, axiomatic probability theory has been developed and applied in many aspects of engineering and computer science. More applications can be used in nature science, biology and ecology, and social science. We are summarizing some fundamental probability theory that is useful in statistical decision, inference, and learning, in this section.

To construct probability, our first encounter would be the outcomes from an experiment. The space of elementary outcomes, usually a non-empty set, is denoted as Ω, whose elements $\omega \in \Omega$ are called *elementary outcomes*. Ω is called *sample space*.

Example: We roll a dice and there are 6 outcomes. $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definition: A collection of $\mathcal{G}$ of subsets of Ω is called *algebra* with the following properties:

(i) $\Omega \in \mathcal{G}$

(ii) $C \in \mathcal{G}$ implies $C^C \in \mathcal{G}$

(iii) $C_1, \ldots, C_n \in \mathcal{G}$ implies $\sum_{i=1}^{n} C_i \in \mathcal{G}$

Definition: A σ-algebra or σ-field $\mathfrak{B}$ is a non-empty collection of subsets of Ω with the following properties:

(i) $\Omega \in \mathfrak{B}$

(ii) If $F \in \mathfrak{B}, F^c = \{\omega: \omega \notin F\} \in \mathfrak{B}$

(iii) If $F_i \in \mathfrak{B}\ i = 1, 2, \ldots, \bigcup_i F_i \in \mathfrak{B}$

Definition: A measureable space (Ω, $\mathfrak{B}$) is a pair consisting of a sample space Ω and a σ-algebra $\mathfrak{B}$ of subsets of Ω (that is also known as the event space).

Definition: A probability space $(\Omega, \mathfrak{B}, P)$ is a triple consisting of sample space Ω, a σ-field $\mathfrak{B}$ of subsets of Ω, and a probability measure P defined on the σ-field. That is, $P(A)$ assigns a real number to every member A of $\mathfrak{B}$ such that the following conditions are satisfied:

(i) $P(A) \geq 0, \forall\ A \in \mathfrak{B}$

(ii) $P(\Omega) = 1$

(iii) $\forall$ disjoint $A_i \in \mathfrak{B}, i = 1, 2, \ldots, P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Lemma: If $A_n \downarrow \emptyset$ (empty or null set), that is, $A_{n+1} \subset A_n\ \forall n$, and $\bigcap_{n=1}^{\infty} A_n = \emptyset$.

Remark: This Lemma suggests *continuity* at $\emptyset$. Therefore, it is straightforward to derive the following two lemmas: *continuity from below* and *continuity from above*, by setting $F_n = A_n - A_{n-1}$.

Lemma: If $F_n \uparrow F$, then $\lim_{n\to\infty} P(F_n) = P(F)$.
Lemma: If $F_n \downarrow F$, then $\lim_{n\to\infty} P(F_n) = P(F)$.

Remark: To prove that P is a probability measure, the following inequality must be hold.

$$P\left(\bigcup_{i=0}^{\infty} F_i\right) \leq \sum_{i=0}^{\infty} P(F_i)$$

Definition: A random variable (i.e. a measureable function) defined on (Ω, $\mathfrak{B}$) and taking values on $(\mathcal{R}, \mathfrak{B}_{\mathcal{R}})$, is a mapping (or function) f: $\Omega \to \mathcal{R}$ with the property that

$$\text{if } F \in \mathfrak{B}_{\mathcal{R}}, \text{ then } f^{-1}(F) = \{\omega: f(\omega) \in F\} \in \mathfrak{B}$$

Remark: There involve two measureable spaces here: (Ω, $\mathfrak{B}$) and $(\mathcal{R}, \mathfrak{B}_{\mathcal{R}})$. We may treat the first measureable space as an input space and the second one as output space. A random variable is just a mapping whose inverse images of the input events

are events in the original measureable space. If we consider in $\Re^n$ (Euclidean space), we are dealing random vectors.

Definition: A collection of random variables $\{X_t\}_{t\in\mathcal{I}}$, where $\mathcal{I}$ is an index set, defined on the common probability space is called a *random process*.

Remark: If $\mathcal{I}$ is countable, it is a discrete-time random process. Otherwise, it is a continuous-time random process.

1.2 Convergence of Random Variables and Limit Theorems

Before going to details of mathematical statistics, we have to look into convergence behaviors of random variables.

Definition: A sequence of random variables $\{X_n\}$ converges *in distribution* to X, $X_n \xrightarrow{D} X$ if $F_{X_n}(t) \to F_X(t)\ \forall t$ such that F_X is continuous at t.

Definition: A sequence of random variables $\{X_n\}$ converges *in probability* to X, $X_n \xrightarrow{P} X$ if $P[|X_n - X| \geq \epsilon] \to 0$ as $n \to \infty\ \forall \epsilon > 0$.

Remark: Above definition suggests that $X_n \xrightarrow{P} X$ if the chances that X_n and X differ by any given amount is negligible when n is large.

Theorem: If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

Corollary: If $X = c$ (a constant), $X_n \xrightarrow{D} X$ can imply $X_n \xrightarrow{P} X$.

Corollary: If $X_n \xrightarrow{P} c$ (a constant), and g is continuous at c, then $g(X_n) \xrightarrow{P} g(c)$.

Corollary: If $X_n \xrightarrow{D} X$ and g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Theorem: If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$ (a constant), then

(a) $X_n + Y_n \xrightarrow{D} X + c$

(b) $X_n Y_n \xrightarrow{D} cX$

Proof: We prove (a) and the proof of (b) is similar.

$$F_{X_n+Y_n}(t) = P[X_n + Y_n \le t, Y_n \ge c - \epsilon] + P[X_n + Y_n \le t, Y_n < c - \epsilon]$$

Let t be a point of continuity of F_{X+c}. Because a distribution function has at most countably many points of continuity, for any t, we select $\epsilon > 0$ such that $t \pm \epsilon$ are both points of continuity of F_{X+c}.

$$\begin{aligned} F_{X_n+Y_n}(t) &\le P[X_n \le t - c + \epsilon] + P[|Y_n - c| > \epsilon] \\ &\le F_{X_n+c}(t + \epsilon) + P[|Y_n - c| > \epsilon] \end{aligned}$$

Because $F_{X_n+c}(t) = P[X_n \le t - c] = F_{X_n}(t - c)$, $X_n + c \xrightarrow{D} X + c$.

$$\limsup_n F_{X_n+Y_n}(t) \le \lim_n F_{X_n+c}(t + \epsilon) + \lim_n P[|Y_n - c| > \epsilon] = F_{X+c}(t + \epsilon)$$

Similarly, $1 - F_{X_n+Y_n}(t) = P[X_n + Y_n > t] \le P[X_n > t - c - \epsilon] + P[|Y_n - c| > \epsilon]$

$$\liminf_n F_{X_n+Y_n}(t) \ge \lim_n F_{X_n+c}(t - \epsilon) = F_{X+c}(t - \epsilon)$$

Consequently,

$$F_{X+c}(t - \epsilon) \le \liminf_n F_{X_n+Y_n}(t) \le F_{X+c}(t + \epsilon) \le F_{X+c}(t + \epsilon)$$

$\epsilon \to 0$ and F_{X+c} is continuous at t. Then, (a) follows. ¶

Definition: A sequence of random variables $\{X_n\}$ converges *almost everywhere* (or, *almost surely, or with probability 1*) to X, $X_n \xrightarrow{a.e.} X$ if $P(\lim_{n\to\infty} X_n = X) = 1$

Definition: A sequence of random variables $\{X_n\}$ converges *in the* pth *mean* (or, in L^p) to X, $X_n \xrightarrow{L^p} X$ if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$$

In particular, $\{X_n\}$ converges *in mean square* to X, $X_n \xrightarrow{m.s.} X$ if $p = 2$, which is of interests in many engineering problems as L^2 stands for the square of Euclidean distance.

Theorem: If $X_n \xrightarrow{a.e.} X$, then $X_n \xrightarrow{P} X$.

Remark: The converse of this theorem is not true. A good counter example shall consider sample space. Let $\Omega = [0,1]$ and the event space $\mathfrak{B}[0,1]$. We consider $A_m^i = \left[\frac{i-1}{m}, \frac{i}{m}\right], i = 1,2,\dots,m, m \geq 1,$ and define random variables $\varphi_m^i = 1_{A_m^i}(\omega)$ via the indicator function. We re-arrange the random variables $\varphi_1^1, \varphi_2^1, \varphi_2^2, \dots$ as $X_1, X_2, X_3, \dots$. It is clearly $X_n \xrightarrow{P} X$, and also in mean square, but does not converge almost everywhere.

Theorem: If $X_n \xrightarrow{m.s.} X$, then $X_n \xrightarrow{P} X$.

Remark: The following figure similar to the well known Venn's diagram may represent the relationship of these 4 kinds of convergence.

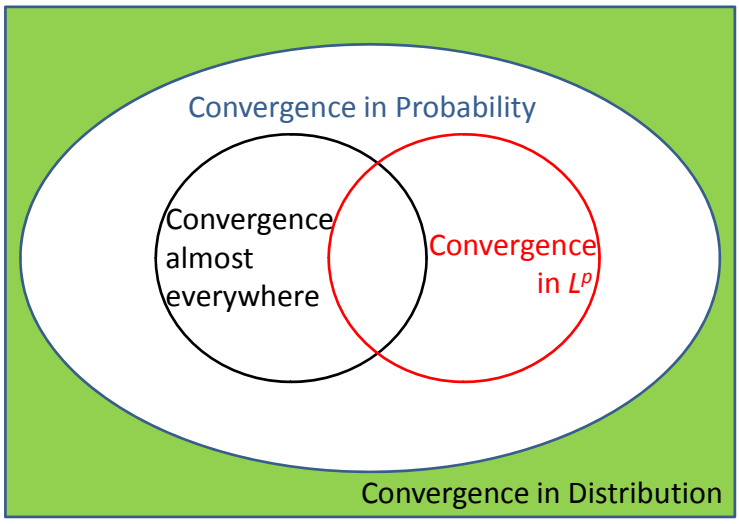

Figure 1: Relationship of Convergence

In the following, we are going to introduce different versions of *Law of Large Numbers* to study the convergence behaviors.

Theorem (Bernoulli's Weak Law of Large Numbers): If $\{X_n\}$ is a sequence of random variables such that $X_n \sim \mathcal{B}(n,p)$ $n \geq 1$, then

$$\frac{X_n}{n} \xrightarrow{P} p$$

Theorem (Chebychev's Inequality): If X is any random variable, then

$$P[|X| \geq a] \leq \frac{E(X^2)}{a^2}$$

Corollary: Let g be a nonnegative function on $\Re$ such that g is non-decreasing on the range of a random variable Z. Then,

$$P[Z \geq a] \leq \frac{E[g(Z)]}{g(a)}$$

Proof:

$$g(a)1_{Z \geq a} \leq g(Z)1_{Z \geq a} \leq g(Z)$$
$$g(a)P[Z \geq a] = E[g(a)1_{Z \geq a}] \leq E(g[Z]) \quad \P$$

Corollary (Markov Inequality): For a random variable $X, P[|X| \geq t] \leq \frac{E[X]}{t}$.

Corollary (Chernoff Bound): For a random variable X and a constant c,

$$P[X \geq c] \leq \min_{s \geq 0} e^{-sc}\varphi_X(s)$$

Remark: These inequalities are useful in information theory. The *Chernoff bound* usually offers a tighter bound than Chebychev inequality.

Theorem (Weak Law of Large Numbers): Let $\{X_n\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean.

$$P[|\overline{X_n} - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}$$

and $P[|\overline{X_n} - \mu| > \epsilon] \to 0$ as $n \to \infty$.

Theorem (Khintchin's Weak Law of Large Numbers): Let $\{X_i\}\, i \geq 1$ be a sequence of i.i.d. random variables with mean $\mu < \infty$ and $S_n = \sum_{i=1}^{n} X_i$. Then,

$$\frac{S_n}{n} \xrightarrow{P} \mu$$

Theorem (De Moivre-Laplace): $\{X_n\}$ is a sequence of random variables such that $\forall n$, X_n has a Bernoulli distribution $\mathcal{B}(n,p)\ 0 < p < 1$. Then,

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} \mathcal{N}(0,1)$$

Central Limit Theorem: Let $\{X_i\}$ be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. If $S_n = \sum_{i=1}^{n} X_i$,

$$\frac{S_n - n\mu}{\sqrt{n\sigma}} \xrightarrow{D} \mathcal{N}(0,1)$$

Theorem (Kolmoorov's Strong Law of Large Numbers): Let $\{X_i\}\, i \geq 1$ be a sequence of i.i.d. random variables with mean $\mu < \infty$. Then,

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{a.e} \mu$$

1.3 Mathematical Statistics

To understand unknown behaviors, we must conduct some experiments to acquire useful information from observations (or data). A typical experiment is to sample. Given a random experiment with sample space Ω, we define a random vector $\mathbb{X} = (X_1, X_2, \dots, X_n)$. When ω denotes the outcome of the experiment, $\mathbb{X}(\omega)$ is referred as the *observations* or *data*. Since we only observe $\mathbb{X}$, we only care its probability distribution, which is assumed to be a member of a family $\mathcal{P}$ of probability distributions on $\Re^n$. Therefore, $\mathcal{P}$ is known as the *model*.

We usually describe $\mathcal{P}$ by *parameterization* (but not necessary), that is, by a mapping $\theta \to P_\theta$ from the parameter space Θ (a space of label) to $\mathcal{P}$. In other words, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and models $\mathcal{P}$ are called *parametric* if Θ is a "nice" subject of $\Re^n$ such that $\theta \to P_\theta$ is a smooth mapping.

Definition: A function of one or more random variables that does not depend on any unknown parameter is called a *statistic*.

Definition: Let $X_1, X_2, \dots, X_n$ denote a random sample of size n from a given (known or unknown) distribution. The statistic

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

is called the *sample mean* and the statistic

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

is called the *sample variance*.

Remark: If we inappropriately design an experiment, we may induce errors in sampling, in addition to observation errors.

However, please keep in mind that the value of statistics lays in deriving useful information or conclusion from data in presence of uncertainty. If this reasoning is based on random variables, it is called *statistical inference*. The following example gives the most straightforward construction of statistical inference.

Example (Regression): We observe pairs of random variables $(\mathbb{z}_1, Y_1), \dots, (\mathbb{z}_n, Y_n)$. $\mathbb{z}_i$ is a d-dimensional vector representing various characteristics of the ith subject in the experiment or study, such as height, weight, age, etc. The distribution of response Y_i for the ith subject is postulated to depend on characteristics $\mathbb{z}_i$. In general, $\mathbb{z}_i$ is a nonrandom vector called *covariate vector* or a vector of *explanatory variables*. Y_i is random and known as *response variable* or *dependent* variable as its distribution depending on $\mathbb{z}_i$. If $f(y_i|\mathbb{z}_i)$ denotes the density of Y_i for a subject with covariate vector $\mathbb{z}_i$, then the model is

$$p(y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i|\mathbb{z}_i)$$

Let $\mu(\mathbb{z})$ be an unknown function from $\Re^d$ to $\Re$ of interests, to denote the expectation of a response with a given covariate vector $\mathbb{z}$. We can then write

$$Y_i = \mu(\mathbb{z}_i) + \epsilon_i, i = 1, \dots, n$$

where $\epsilon_i = Y_i - E(Y_i), i = 1, \dots, n$. As a special case, by defining $\mathbb{b} = (b_1, \dots, b_d)^T$ and identically distributed ϵ_i, a linear regression model thus becomes

$$Y_i = \mathbb{z}_i^T \mathbb{b} + \epsilon_i, 1 \le i \le n$$

Please note that we usually conduct mean-squared error criterion in EE&CS, and same to derive linear regression using ϵ_i.

Lemma: There are some useful properties of conditional probability for statistical inference, where $\mathcal{H}$ stands for a hypothesis.

(Product or Chain Rule) $P(x, y|\mathcal{H}) = P(x|y, \mathcal{H})P(y|\mathcal{H}) = P(y|x, \mathcal{H})P(x|\mathcal{H})$

(Sum Rule) $P(x|\mathcal{H}) = \sum_y P(x, y|\mathcal{H}) = \sum_y P(x|y, \mathcal{H})P(y|\mathcal{H})$

(Bayes) $P(y|x,\mathcal{H}) = \frac{P(x|y,\mathcal{H})P(y|\mathcal{H})}{P(x|\mathcal{H})} = \frac{p(x|y,H)p(y|H)}{\sum_{\varphi} p(x|\varphi,H)p(\varphi|H)}$

(Independence) $p(x,y) = p(x)p(y)$

Actually, probability can be used to model more general problems including human knowledge systems (for belief, cognition, trust, etc.), which involves reasoning in the presence of uncertainty. It can be linked by the famous *Cox Axioms*.

Cox Axioms: Let "the degree of belief in proposition x" be denoted by $B(x)$. The negation of x is $\bar{x}$. $B(x|y)$ means that the degree of belief in a conditional proposition x assuming proposition y to be true.

(a) Degree of belief can be ordered. That is, if $B(x) > B(y)$ and $B(y) > B(z)$, then $B(x) > B(z)$.

(b) The degree of belief in a proposition x and its negation $\bar{x}$ are related. There is a function f such that $B(x) = f[B(\bar{x})]$.

(c) The degree of belief in a conjunction of propositions x, y $(x \wedge y)$ is related to the degree of belief in the conditional proposition $x|y$ and the degree of belief in proposition y. There exists a function g such that $B(x,y) = g[B(x|y), B(y)]$.

If above 3 conditions are satisfied, we can model the belief as a probability.

Remark: Belief can be mapped into real numbers.

Derivations of probability or statistics often fall into two scenarios: forward and inverse. Forward problems usually involve a generative model describing a process to deliver data.

Example: An urn contains K balls, among which B are blue and $W = K - B$ are white. Danny randomly draws a ball from the urn and replaces it for N times. (a) What is the probability distribution of the number of times a blue ball is drawn, n_B? (b) What is the expectation of n_B? What is the variance of n_B?

Answer:

Define $\rho = B/K$. $P(n_B|\rho,N) = \binom{N}{n_B}\rho^{n_B}(1-\rho)^{N-n_B}$ and thus $\mathbb{E}[n_B] = \rho N$ $\mathrm{var}[n_B] = \rho(1-\rho)N$. ¶

Inverse probability problems again involve a generative model. However, instead of computing probability distribution of certain quantity induced by the model, we

derive the conditional probability of *unobserved* variable(s) in the model, given observed variable(s). *Bayes theorem* this plays a dominating role.

Lemma (Bayes' Theorem): Suppose $(\mathbb{Y}, \mathbb{Z})$ is a random vector with density function $p(\mathbb{y}, \mathbb{z})$. The conditional density function of $\mathbb{Y}$ given $\mathbb{Z} = \mathbb{z}$ is $p(\mathbb{y}|\mathbb{z}) = \frac{p(\mathbb{y},\mathbb{z}))}{p_{\mathbb{Z}}(\mathbb{z})}$ if $p_{\mathbb{Z}}(\mathbb{z}) > 0$. Similarly, q is the conditional density of $\mathbb{Z}$ given $\mathbb{Y} = \mathbb{y}$.

$$p(\mathbb{y}|\mathbb{z}) = \frac{q(\mathbb{z}|\mathbb{y})p_{\mathbb{Y}}(\mathbb{y})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(\mathbb{z}|\mathbb{t})p_{\mathbb{Y}}(\mathbb{t})dt_1 \cdots dt_n}$$

Example: There are 11 urns labeled by 0, 1, 2, …, 10, and each urn contains 10 balls. Urn u contains u blue balls and $10 - u$ white balls. Danny randomly selects an urn and draws N times with replacement, to obtain n_B blue balls and $N - n_B$ white balls. Chloé examines the results without knowing the selected urn. In $N = 10$ draws, $n_B = 3$ blue balls have been drawn. For Chloé, what is the probability that Danny is using urn u?

Answer:

The conditional probability of u given n_B is

$$P(u|n_B, N) = \frac{P(u)P(n_B|u, N)}{P(n_B|N)} = \frac{1}{P(n_B|N)} \frac{1}{11} \binom{N}{n_B} \left(\frac{u}{10}\right)^{n_B} \left(1 - \frac{u}{10}\right)^{N-n_B}$$

We may denote $\frac{u}{10} = f_u$. ¶

Remark: Above inverse probability problem is useful to illustrate some terminologies in statistical inference or statistics. The marginal probability $P(u)$ is called the *prior* probability of u. $P(n_B|u, N)$ is called the *likelihood*. It is vital to note the difference between probability and likelihood. $P(n_B|u, N)$ is a function of both n_B and u. For fixed u, $P(n_B|u, N)$ defines a *probability* over n_B. For fixed n_B, $P(n_B|u, N)$ defines the *likelihood* of u. The conditional probability $P(u|n_B, N)$ is called the *posterior probability* of u given n_B. $P(n_B|N)$ is a constant independent of u and is known as the *evidence* or *marginal likelihood*.

Proposition: If θ denotes the unknown parameter, $\mathcal{D}$ denotes the data, and $\mathcal{H}$ denotes the overall hypothesis space, then we have

$$P(\theta|\mathcal{D}, \mathcal{H}) = \frac{P(\mathcal{D}|\theta, \mathcal{H})P(\theta|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})}$$

Remark: In high-level language, above proposition means

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Example (continue): Chloé observes Danny's drawing $n_B = 3$ blue balls in $N = 10$ draws. Supposing that Danny draws another ball from the same urn, what is the probability that next drawn ball is blue? Please note that this is so-called *prediction*.

Answer:

Let β represents the event of ball $N+1$ being blue.

$$P(\beta|n_B, N) = \sum_u P(\beta|u, n_B, N)\, P(u|n_B, N) = \sum_u f_u\, P(u|n_B, N)$$

u	$P(u\|n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.099
8	0.00086
9	0.0000096
10	0

Table: Conditional probability of u given $n_B = 3$ and $N = 10$

$P(\text{ball } N+1 \text{ being blue}|n_B = 3 \text{ and } N = 10)$=0.333 ¶

Example (Laplace Formula): Suppose we have a bent coin with uneven probabilities for head and tail in a series of independent tossing as a series of Bernoulli trials. For the first $L_h + L_t = L$ trials, there are L_h times to have head and L_t times for tail. The famous *Laplace Formula* tells that the probability of next independent tossing to be head is

$$\frac{L_h + 1}{L_h + L_t + 2} = \frac{L_h + 1}{L + 2}$$

Proposition (Likelihood Principle): Given a generative model for data D given parameter θ, $P(D|\theta)$, and having observed a particular outcome d, all inference and predictions should depend only on the function $P(d|\theta)$.

Remark: Although likelihood principle is simple, many classical statistical methods violate it.

1.4 Information and Entropy

As a matter of fact, source coding (e.g. data compression) has close relationship with data modeling and inverse probability problem. To deal with information transmission, we may consider an ensemble X as a triple (x, A_X, P_X), where x denotes outcomes from a random variable; A_X denotes possible values (that is, alphabets in information transmission or information theory); P_X denotes probabilities (or possibilities).

Definition: Shannon's information content of an outcome x

$$h(x) = \log_2 \frac{1}{p(x)}$$

Definition: Entropy of an ensemble X is

$$H(X) = -\sum_{x \in \mathcal{A}_X} p(x) \log p(x)$$

Lemma: $H(X) = 0$ with equality if an only if $p(x_0) = 1$ for one x_0.

Lemma: Entropy is maximized if $p(x)$ is uniformly distributed. Then, $H(X) \leq \log |\mathcal{A}_X|$ with equality if and only if $p(x) = \frac{1}{|\mathcal{A}_X|}$ $\forall x$.

Corollary: The joint entropy of X and Y is

$$H(X,Y) = -\sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_y} p(x,y) \log p(x,y)$$

Definition: The *Relative Entropy* or *Kullback-Leibler Divergence* between two probability distributions $p(x), q(x)$ over the same $\mathcal{A}_X$ is

$$D_{KL}(p||q) = \sum_{\mathrm{x}} p(x) \log \frac{p(x)}{q(x)}$$

Corollary (Gibb's Inequality): $D_{KL}(p||q) \geq 0$

Remark: $D_{KL}(p||q) \neq D_{KL}(q||p)$ in general, and thus D_{KL} is *not* a distance in

general.

In the following, we are going to introduce the well known and useful inequality for convex functions.

Definition: A function $f(x)$ is *convex up* over (a,b) if every chord of the function lies above the function. In other words, $\forall x_1, x_2 \in (a,b)$ and $0 \le \gamma \le 1$,

$$f(\gamma x_1 + (1-\gamma)x_2) \le \gamma f(x_1) + (1-\gamma)x_2 f(x_2)$$

A function $f(x)$ is strictly convex up if $\forall x_1, x_2 \in (a,b)$. The equality holds only for $\gamma = 0$ and $\gamma = 1$.

Lemma (Jensen's Inequality): If f is a convex up function, and X is a random variable, then

$$\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$$

Remark: *Jensen's inequality* can be rewritten for a convex (or concave) down function by reversing the inequality. A physical meaning of Jensen's inequality: In case masses p_i are placed on a convex up curve $f(x)$, i.e. at locations $(x_i, f(x_i))$, the center of gravity of these masses lies above the curve.

Exercises:

1. The outer-measure of an interval $[a,b]$ is defined to be $b-a$. Let $\mathfrak{Q}$ denote the set of all rational numbers within $[0,1]$ and $\mathfrak{Q}^c$ denote the set of all irrational numbers. Please find the outer-measure for $\mathfrak{Q}$ and $\mathfrak{Q}^c$.
2. The condition (iii) in the definition of σ-algebra, it is equivalent to that *If* $F_i \in \mathfrak{B}\ i = 1, 2, \ldots, \bigcap_i F_i \in \mathfrak{B}$.
3. Suppose X_n takes values only on $\mathfrak{N}$ and $p_{X_n}(x) \to p_X(x)\ \forall x$. Then, $X_n \xrightarrow{D} X$.
4. Suppose X_n, X are continuous and $p_{X_n}(x) \to p_X(x)\ \forall x$. Then, $X_n \xrightarrow{D} X$.
5. Please find another example that $X_n \xrightarrow{P} X$ does not imply $X_n \xrightarrow{a.e.} X$.
6. (Berry-Esséen Theorem) Let $\{X_n\}$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then,

$$\sup_t \left| P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le t\right) - \Phi(t) \right| \le \frac{33}{4}\frac{E|X_1-\mu|^3}{\sqrt{n}\sigma^3}$$

7. Suppose $X_1, X_2, \ldots, X_n, \ldots$ are independently sampled from $G(\mu, \sigma^2)$ distribution. Let $M_n = \max(X_1, X_2, \ldots, X_n)$. (a) Is M_n a monotonically

increasing function of n? (b) For $G(0,1)$, 3 of $X_{101},\dots,X_{110}$ are larger than M_{100}. Is it likely?

8. Let X be a continuous random variable such that $F_X(0)=0$ and $F_X(\epsilon)>0\ \forall\epsilon>0$. Let $X_1,X_2,\dots$ denote independent and identically distributed (i.i.d.) random variables of probability density function (PDF) $f_X(x)$. Suppose $Y_n=\min(X_1,\dots,X_n)$. Please show that $Y_n \xrightarrow{P} 0$.
9. Please derive the Laplace formula in Section 1.3.
10. ABC company manufactures N computers for XYZ brand and D of them are defective. The quality control team in XYZ samples n computers to check and finds k defectives. For $\max(0,n-(N-D))\le k\le\min(n,D)$, find the distribution for each $k\in\Re$, which is known as hyper-geometric $\mathcal{H}(D,N,n)$.
11. The entropy $H(X)$ of a random variable X is defined as

$$H(X)=-\sum_x p(x)\log p(x)$$

(a) Please find an example of X that $H(X)=0$.

(b) The information divergence between two probability distributions p and q on a common alphabet $\mathcal{X}$ is defined as

$$D(p||q)=\sum_{\mathrm{x}} p(x)\log\frac{p(x)}{q(x)}$$

Please prove that $D(p||q)\ge 0$, which is know as *Gibbs' inequality*. Hint: $\forall a>0, ln\, a\le a-1$, with equality if and only if $a=1$.

(c) Please prove that for any random variable X,

$$H(X)\le \log\ |\mathcal{X}|$$

Where $|\mathcal{X}|$ denotes the size of alphabet $\mathcal{X}$. Hint: (b) is useful. This inequality leads to the famous *Fano inequality*.

(d) Please find the condition and thus an example to tight the upper bound in (c).

12. A source randomly produces a character x from the alphabet $\mathcal{A}=\{0,1,..,9,\mathrm{a},\mathrm{b},\dots,\mathrm{z}\}$. x is a numeral (i.e. from $\{0,1,..,9\}$) equally probable with total probability 1/3; x is a vowel (i.e. from $\{\mathrm{a},\mathrm{e},\mathrm{i},\mathrm{o},\mathrm{u}\}$) equally probable with total probability 1/3; x is one of the 21 consonants equally probable with total probability 1/3. Please find the "quickest" estimate of the number of bits to represent $\mathcal{A}$ in binary bits. Hint: Please consider the *decomposability of the entropy*.