STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 1: Review on Probability and Statistics

Instructor: Yen-Chi Chen

1.1 Sample Space and Random Variables

1.1.1 Sample Space and Probability Measure

The *sample space* Ω is the collection of all possible outcomes of a random experiment, e.g. toss of a coin, $\Omega = \{H, T\}$. Elements $\omega \in \Omega$ are called *outcomes*, *realizations* or *elements*. Subsets $A \subseteq \Omega$ are called *events*. You should able to express events of interest using the standard set operations. For instance:

- "Not A" corresponds to the *complement* $A^c = \Omega \setminus A$;
- "A or B" corresponds to the *union* $A \cup B$;
- "A and B" corresponds to the *intersection* $A \cap B$.

We said that $A_1, A_2, ...$ are *pairwise disjoint/mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i \neq j$. A *partition* of Ω is a sequence of pairwise disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = \Omega$. We use $|A|$ to denote the number of elements in A.

The sample space defines basic elements and operations of events. But it is still too simple to be useful in describing our senses of 'probability'. Now we introduce the concept of σ-algebra.

A *σ-algebra* $\mathcal{F}$ is a collection of subsets of Ω satisfying:

(A1) (full and null set) $\Omega \in \mathcal{F}$, $\emptyset \in \mathcal{F}$ ($\emptyset$ = empty set).

(A2) (complement)$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(A3) (countably union) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The sets in $\mathcal{F}$ are said to be *measurable* and $(\Omega, \mathcal{F})$ is a *measurable space.* The intuition of a set being measurable is that we can find a function that takes the elements of $\mathcal{F}$ and output a real number; this number represents the 'size' of the input element.

Now we introduce the concept of probability. Intuitively, probability should be associated with an event – when we say a probability of something, this 'something' is an event. Using the fact that the σ-algebra $\mathcal{F}$ is a collection of events and the property that $\mathcal{F}$ is measurable, we then introduce a measure called *probability measure* $\mathbb{P}(\cdot)$ that assigns a number between 0 and 1 to every element of $\mathcal{F}$. Namely, this function $\mathbb{P}$ maps an event to a number, describing the likelihood of the event.

Formally, a probability measure is a mapping $\mathbb{P} : \mathcal{F} \mapsto \mathbb{R}$ satisfying the following three axioms

(P1) $\mathbb{P}(\Omega) = 1$.

(P2) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.

(P3) (countably additivity) $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for *mutually exclusive* events $A_1, A_2, ... \in \mathcal{F}$.

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space.*

The three axioms imply:

$$\begin{aligned}
&\mathbb{P}(\emptyset) = 0\\
&0 \leq \mathbb{P}(A) \leq 1\\
&A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B),\\
&\mathbb{P}(A^c) = 1 - \mathbb{P}(A)\\
&\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)
\end{aligned}$$

The countable additivity (P3) also implies that if a sequence of sets $A_1, A_2, \dots$ in $\mathcal{F}$ satisfying $A_n \subseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

If $A_n \supseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

How do we interpret the probability? There are two major views in statistics. The first view is called the frequentist view – the probability is interpreted as the limiting frequencies observed over repetitions in identical situations. The other view is called the Bayesian/subjective view where the probability quantifies personal belief. One way of assigning probabilities is the following. The probability of an event E is the price one is *just* willing to pay to enter a game in which one can win a unit amount of money if E is true. Example: If I believe a coin is fair and am to win 1 unit if a head arises, then I would pay $\frac{1}{2}$ unit of money to enter the bet.

Now we have a basic mathematical model for probability. This model also defines an interesting quantity called conditional probability. For two events $A, B \in \mathcal{F}$, the conditional probability of A given B is

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

Note that when B is fixed, the function $\mathbb{P}(\cdot|B) : \mathcal{F} \mapsto \mathbb{R}$ is another probability measure.

In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. This is sometimes called as the prosecutor's fallacy:

$$\mathbb{P}(\text{evidence}|\text{guilty}) \neq \mathbb{P}(\text{guilty}|\text{evidence}).$$

The probability has a power feature called *independence.* This property is probably the key property that makes the 'probability theory' distinct from measure theory. Intuitively, when we say that two events are independent, we refers to the case that the two event will not interfere each other. Two events A and B are independent if

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \text{(or equivalently, } \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)).$$

For three events A, B, C, we say events A and B are *conditional independent* given C if

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

Probability measure also has a useful property called *law of total probability.* If $B_1, B_2, ..., B_k$ forms a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

In particular, $\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$. And this further implies the famous *Bayes rule*: Let $A_1, ..., A_k$ be a partition of Ω. If $\mathbb{P}(B) > 0$ then, for $i = 1, ..., k$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

1.1.2 Random Variable

So far, we have built a mathematical model describing the probability and events. However, in reality, we are dealing with numbers, which may not be directly link to events. We need another mathematical notion that bridges the events and numbers and this is why we need to introduce random variables.

Informally, a *random variable* is a mapping $X : \Omega \mapsto \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. Fo example, we toss a coin 2 times and let X represents the number of heads. The sample space is $\Omega = \{HH, HT, TH, TT\}$. Then for each $\omega \in \Omega$, $X(\omega)$ outputs a real number: $X(\{HH\}) = 2$, $X(\{HT\}) = X(\{TH\}) = 1$, and $X(\{TT\}) = 0$.

Rigorously, a function $X(\omega) : \Omega \to \mathbb{R}$ is called a *random variable* (R.V.) if $X(\omega)$ is measurable with respect to $\mathcal{F}$, i.e.

$$X^{-1}((-\infty, c]) := \{\omega \in \Omega : X(\omega) \le c\} \in \mathcal{F}, \quad \text{for all } c \in \mathbb{R}.$$

Note that the condition is also equivalent to saying that $X^{-1}(B) \in \mathcal{F}$ for every Borel set B[1]. This means that the set $X^{-1}(B)$ is indeed an event so that it makes sense to talk about $\mathbb{P}(X \in B)$, the probability that X lies in B, for any Borel set B. The function $B \mapsto \mathbb{P}(X \in B)$ is a probability measure and is called the *(probability) distribution* of X.

A very important characteristic of a random variable is its *cumulative distribution function (CDF)*, which is defined as

$$F(x) = P(X \le x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x\}).$$

Actually, the distribution of X is completely determined by the CDF $F(x)$, regardless of X being a discrete random variable or a continuous random variable (or a mix of them).

When X takes discrete values, we may characterize its distribution using the probability mass function (PMF):

$$p(x) = P(X = x) = F(x) - F(x^-),$$

where $F(x^-) = \lim_{\epsilon \to 0} F(x - \epsilon)$. In this case, one can recover the CDF from PMF using $F(x) = \sum_{x' \le x} p(x')$.

If X is an absolutely continuous random variable, we may describe its distribution using the probability density function (PDF):

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

In this case, the CDF can be written as

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(x')dx'.$$

[1] A Borel set is a set that can be formed by countable union/intersection and complement of open sets.

However, the PMF and PDF are not always well-defined. There are situations where X does not have a PMF or a PDF. The formal definition of PMF and PDF requires the notion of the Radon-Nikodym derivative, which is beyond the scope of this course.

1.2 Common Distributions

1.2.1 Discrete Random Variables

Bernoulli. If X is a Bernoulli random variable with parameter p, then $X = 0$ or, 1 such that

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

In this case, we write $X \sim \mathsf{Ber}(p)$.

Binomial. If X is a binomial random variable with parameter (n, p), then $X = 0, 1, \cdots, n$ such that

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

In this case, we write $X \sim \mathsf{Bin}(n, p)$. Note that if $X_1, \cdots, X_n \sim \mathsf{Ber}(p)$, then the sum $S_n = X_1 + X_2 + \cdots + X_n$ is a binomial random variable with parameter (n, p).

Geometric. If X is a geometric random variable with parameter p, then

$$P(X = n) = (1 - p)^{n-1} p$$

for $n = 1, 2, \cdots$. Geometric random variable can be constructed using 'the number of trials of the first success occurs'. Consider the case we are flipping coin with a probability p that we gets a head (this is a Bernoulli (p) random variable). Then the number of trials we made to see the first head is a geometric random variable with parameter p.

Poisson. If X is a Poisson random variable with parameter λ, then $X = 0, 1, 2, 3, \cdots$ and

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

In this case, we write $X \sim \mathsf{Poi}(\lambda)$. Poisson is often used to model a counting process. For instance, the intensity of an image is commonly modeled as a Poisson random variable.

Example: Wright-Fisher Model. Recall the Wright-Fisher model: X_n is the number of A alleles in the population at generation n, with $2m$ alleles in all. We have $2m$ Bernoulli trials with $P(A) = j/2m$ where j is the number of A alleles in the previous generation (recall, assumed sampling with replacement). The probability of $X_{n+1} = k$ given $X_n = j$ is Binomial$(2m, j/2m)$:

$$P(X_{n+1} = k | X_n = j) = \binom{2m}{k} \left(\frac{j}{2m}\right)^k \left(1 - \frac{j}{2m}\right)^{2m-k},$$

for $j, k = 0, 1, ..., 2m$.

1.2.2 Continuous Random Variables

Uniform. If X is a uniform random variable over the interval $[a, b]$, then

$$p(x) = \frac{1}{b - a} I(a \leq x \leq b),$$

where $I(\mathsf{statement})$ is the indicator function such that if the $\mathsf{statement}$ is true, then it outputs 1 otherwise 0. Namely, $p(x)$ takes value $\frac{1}{b-a}$ when $x \in [a,b]$ and $p(x) = 0$ in other regions. In this case, we write $X \sim \mathsf{Uni}[a,b]$.

Normal. If X is a normal random variable with parameter (μ, σ^2), then

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In this case, we write $X \sim N(\mu, \sigma^2)$.

Exponential. If X is an exponential random variable with parameter λ, then X takes values in $[0,\infty)$ and

$$p(x) = \lambda e^{-\lambda x}.$$

In this case, we write $X \sim \mathsf{Exp}(\lambda)$. Note that we can also write

$$p(x) = \lambda e^{-\lambda x} I(x \geq 0).$$

1.3 Properties of Random Variables

1.3.1 Conditional Probability and Independence

For two random variables X, Y, the joint CDF is

$$P_{XY}(x,y) = F(x,y) = P(X \leq x, Y \leq y).$$

When both variables are absolute continuous, the corresponding joint PDF is

$$p_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}.$$

The *conditional PDF* of Y given $X = x$ is

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

where $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)dy$ is sometimes called the marginal density function.

When both X and Y are discrete, the joint PMF is

$$p_{XY}(x,y) = P(X = x, Y = y)$$

and the conditional PMF of Y given $X = x$ is

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

where $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

Random variables X and Y are *independent* if the joint CDF can be factorized as

$$F(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

For random variables, we also have the Bayes theorem:

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{XY}(x,y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \\ &= \begin{cases} \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x')p_X(x')dx'}, & \text{if } X, Y \text{ are absolutely continuous.} \\ \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}, & \text{if } X, Y \text{ are discrete.} \end{cases} \end{aligned}$$

1.3.2 Expectation

For a function $g(x)$, the expectation of $g(X)$ is

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x g(x)p(x), & \text{if } X \text{ is discrete} \end{cases}.$$

Here are some useful properties and quantities related to the expected value:

- $\mathbb{E}(\sum_{j=1}^k c_j g_j(X)) = \sum_{j=1}^k c_j \cdot \mathbb{E}(g_j(X_i))$.
- We often write $\mu = \mathbb{E}(X)$ as the mean (expectation) of X.
- $\mathsf{Var}(X) = \mathbb{E}((X-\mu)^2)$ is the variance of X.
- If $X_1, \cdots, X_n$ are independent, then

  $$\mathbb{E}\left(X_1 \cdot X_2 \cdots X_n\right) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_n).$$

- If $X_1, \cdots, X_n$ are independent, then

  $$\mathsf{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \mathsf{Var}(X_i).$$

- For two random variables X and Y with their mean being μ_X and μ_Y and variance being σ_X^2 and σ_Y^2. The covariance

  $$\mathsf{Cov}(X,Y) = \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(XY) - \mu_x\mu_y$$

  and the (Pearson's) correlation

  $$\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sigma_x\sigma_y}.$$

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X) = g(X)$ such that when $X = x$, its value is

$$\mathbb{E}(Y|X=x) = \int yp(y|x)dy,$$

where $p(y|x) = p(x,y)/p(x)$. Note that when X and Y are independent,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y), \quad \mathbb{E}(X|Y=y) = \mathbb{E}(X).$$

Law of total expectation:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X]] &= \int \mathbb{E}[Y|X=x]p_X(x)dx = \int\int y p_{Y|X}(y|x)p_X(x)dxdy \\ &= \int\int y p_{XY}(x,y)dxdy = \mathbb{E}[Y].\end{aligned}$$

Law of total variance:

$$\begin{aligned}\mathsf{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \mathbb{E}[\mathbb{E}(Y^2|X)] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(law of total expectation)} \\ &= \mathbb{E}[\mathsf{Var}(Y|X) + \mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(definition of variance)} \\ &= \mathbb{E}[\mathsf{Var}(Y|X)] + \left\{\mathbb{E}[\mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2\right\} \\ &= \mathbb{E}\left[\mathsf{Var}(Y|X)\right] + \mathsf{Var}\left(\mathbb{E}[Y \mid X]\right) \quad \text{(definition of variance).}\end{aligned}$$

1.3.3 Moment Generating Function and Characteristic Function

Moment generating function (MGF) and *characteristic function* are powerful functions that describe the underlying features of a random variable. The MGF of a RV X is

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Note that M_X may not exist. When M_X exists in a neighborhood of 0, using the fact that

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots,$$

we have

$$M_X(t) = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \cdots,$$

where $\mu_j = \mathbb{E}(X^j)$ is the j-th moment of X. Therefore,

$$\mathbb{E}(X^j) = M^{(j)}(0) = \left.\frac{d^j M_X(t)}{dt^j}\right|_{t=0}$$

Here you see how the moments of X is generated by the function M_X.

For two random variables X, Y, if their MGFs are the same, then the two random variables have the same CDF. Thus, MGFs can be used as a tool to determine if two random variables have the identical CDF. Note that the MGF is related to the Laplace transform (actually, they are the same) and this may give you more intuition why it is so powerful.

A more general function than MGF is the characteristic function. Let i be the imagination number. The characteristic function of a RV X is

$$\phi_X(t) = \mathbb{E}(e^{itX}).$$

When X is absolutely continuous, the characteristic function is the Fourier transform of the PDF. The characteristic function always exists and when two RVs have the same characteristic function, the two RVs have identical distribution.

1.4 Convergence

Let $F_1, \cdots, F_n, \cdots$ be the corresponding CDFs of $Z_1, \cdots, Z_n, \cdots$. For a random variable Z with CDF F, we say that Z_n **converges in distribution** (a.k.a. converge weakly or converge in law) to Z if for every x,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

In this case, we write

$$Z_n \overset{D}{\to} Z, \quad \text{or } Z_n \overset{d}{\to} Z.$$

Namely, the CDF's of the sequence of random variables converge to a the CDF of a fixed random variable.

For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges in probability** to another random variable Z if for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$$

and we will write

$$Z_n \overset{P}{\to} Z$$

For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges almost surely** to a random variable Z if

$$P(\lim_{n\to\infty} Z_n = Z) = 1$$

or equivalently,

$$P(\{\omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1.$$

We use the notation

$$Z_n \overset{a.s.}{\to} Z$$

to denote convergence almost surely.

Note that almost surely convergence implies convergence in probability. Convergence in probability implies convergence in distribution.

In many cases, convergence in probability or almost surely converge occurs when a sequence of RVs converging toward a fixed number. In this case, we will write (assuming that μ is the target of convergence)

$$Z_n \overset{P}{\to} \mu, \quad Z_n \overset{a.s.}{\to} \mu.$$

Later we will see that the famous Law of Large Number is describing the convergence toward a fixed number.

Examples.

- Let $\{X_1, X_2, \cdots,\}$ be a sequence of random variables such that $X_n \sim N\left(0, 1 + \frac{1}{n}\right)$. Then X_n converges in distribution to $N(0,1)$.

Continuous mapping theorem: Let g be a continuous function.

- If a sequence of random variables $X_n \overset{D}{\to} X$, then $g(X_n) \overset{D}{\to} g(X)$.
- If a sequence of random variables $X_n \overset{p}{\to} X$, then $g(X_n) \overset{p}{\to} g(X)$.

Slutsky's theorem: Let $\{X_n : n = 1, 2, \cdots\}$ and $\{Y_n : n = 1, 2, \cdots\}$ be two sequences of RVs such that $X_n \overset{D}{\to} X$ and $Y_n \overset{p}{\to} c$, where X is a RV c is a constant. Then

$$X_n + Y_n \overset{D}{\to} X + c$$

$$X_n Y_n \overset{D}{\to} cX$$

$$X_n / Y_n \overset{D}{\to} X/c \quad (\text{if } c \neq 0).$$

We will these two theorems very frequently when we are talking about the maximum likelihood estimator.

Why do we need these notions of convergences? The convergence in probability is related to the concept of statistical consistency. An estimator is statistically consistent if it converges in probability toward its target population quantity. The convergence in distribution is often used to construct a confidence interval or perform a hypothesis test.

1.4.1 Convergence theory

We write $X_1, \cdots, X_n \sim F$ when $X_1, \cdots, X_n$ are IID (independently, identically distributed) from a CDF F. In this case, $X_1, \cdots, X_n$ is called a *random sample.*

Theorem 1.1 (Law of Large Number) *Let $X_1, \cdots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$. If $\mathbb{E}|X_1| < \infty$, the sample average*

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

converges in probability to μ. i.e.,

$$\bar{X}_n \overset{a.s}{\to} \mu.$$

The above theorem is also known as Kolmogorov's Strong Law of Large Numbers.

Theorem 1.2 (Central Limit Theorem) *Let $X_1, \cdots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathsf{Var}(X_1) < \infty$. Let $\bar{X}_n$ be the sample average. Then*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \overset{D}{\to} N(0, 1).$$

Note that $N(0, 1)$ is also called standard normal random variable.

Note that there are other versions of central limit theorem that allows dependent RVs or infinite variance using the idea of 'triangular array' (also known as the Lindeberg-Feller Theorem). However, the details are beyond the scope of this course so we will not pursue it here.

In addition to the above two theorems, we often use the concentration inequality to obtain convergence in probability. Let $\{X_n : n = 1, 2, \cdots\}$ be a sequence of RVs. For a given $\epsilon > 0$, the concentration inequality aims at finding the function $\phi_n(\epsilon)$ such that

$$P(|X_n - \mathbb{E}(X_n)| > \epsilon) \leq \phi_n(\epsilon)$$

and $\phi_n(\epsilon) \to 0$. This automatically gives us convergence in probability. Moreover, the *convergence rate* of $\phi_n(\epsilon)$ with respect to n is a central quantity that describes how fast X_n converges toward its mean.

Theorem 1.3 (Markov's inequality) *Let X be a non-negative RV. Then for any $\epsilon > 0$,*

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}.$$

Example: concentration of a Gaussian mean. The Markov's inequality implies a useful bound on describing how fast the sample mean of a Gaussian converges to the population mean. For simplicity, we consider a sequence of mean 0 Gaussians: $X_1, \cdots, X_n \sim N(0, \sigma^2)$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. It is known that $\bar{X}_n \sim N(0, \sigma^2/n)$. Then

$$\begin{aligned} P(\bar{X}_n > \epsilon) &= P(e^{\bar{X}_n} > e^{\epsilon}) \\ &= P(e^{s\bar{X}_n} > e^{s\epsilon}) \\ &\leq \frac{\mathbb{E}(e^{s\bar{X}_n})}{e^{s\epsilon}} \qquad \text{by Markov's inequality} \\ &\leq e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon} \qquad \text{by the MGF of Gaussian} \end{aligned}$$

for any positive number s. In the exponent, it is a quadratic function of s and the maximal occurs at $s = \frac{n\epsilon}{\sigma^2}$, leading to

$$P(\bar{X}_n > \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

The same bound holds for the other direction $P(\bar{X}_n < -\epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}$. So we conclude

$$P(|\bar{X}_n| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

or more generally,

$$P(|\bar{X}_n - \mathbb{E}(X_1)| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

A bound like the above is often referred to as a *concentration inequality*.

Theorem 1.4 (Chebyshev's inequality) *Let X be a RV with finite variance. Then for any $\epsilon > 0$,*

$$P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathsf{Var}(X)}{\epsilon^2}.$$

Let $X_1, \cdots, X_n \sim F$ be a random sample such that $\sigma^2 = \mathsf{Var}(X_1)$. Using the Chebyshev's inequality, we know that the sample average $\bar{X}_n$ has a concentration inequality:

$$P(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

However, when the RVs are bounded, there is a stronger notion of convergence, as described in the following theorem.

Theorem 1.5 (Hoeffding's inequality) *Let $X_1, \cdots, X_n$ be IID RVs such that $0 \leq X_1 \leq 1$ and let $\bar{X}_n$ be the sample average. Then for any $\epsilon > 0$,*

$$P(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Hoeffding's inequality gives a concentration of the order of exponential (actually it is often called a Gaussian rate) so the convergence rate is much faster than the one given by the Chebyshev's inequality. Obtaining such an exponential rate is useful for analyzing the property of an estimator. Many modern statistical topics, such as high-dimensional problem, nonparametric inference, semi-parametric inference, and empirical risk minimization all rely on a convergence rate of this form.

Note that the exponential rate may also be used to obtain an almost sure convergence via the Borel-Cantelli Lemma.

Example: consistency of estimating a high-dimensional proportion. To see how the Hoeffding's inequality is useful, we consider the problem of estimating the proportion of several binary variables. Suppose that we observe IID observations

$$X_1, \cdots, X_n \in \{0,1\}^d.$$

$X_{ij} = 1$ can be interpreted as the i-th individual response 'Yes' in j-th question. We are interested in estimating the proportion vector $\pi \in [0,1]^d$ such that $\pi_j = P(X_{ij} = 1)$ is the proportion of 'Yes' response in j-th question in the population. A simple estimator is the sample proportion $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_d)^T$ such that

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}.$$

When d is much smaller than n, it is easy to see that this is a good estimator. However, if $d = d_n \to \infty$ with $n \to \infty$, will $\hat{\pi}$ still be a good estimator of π? To define a good estimator, we mean that *every proportion* can be estimated accurately. A simple way to quantify this is the vector max norm:

$$\|\hat{\pi} - \pi\|_{\max} = \max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j|.$$

We consider the problem of estimating π_j first. It is easy to see that by the Hoeffding's inequality,

$$P(|\hat{\pi}_j - \pi_j| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Thus,

$$\begin{aligned} P(\|\hat{\pi} - \pi\|_{\max} > \epsilon) &= P\left(\max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j| > \epsilon\right) \\ &\leq \sum_{j=1}^d P(|\hat{\pi}_j - \pi_j| > \epsilon) \\ &\leq 2de^{-2n\epsilon^2}. \end{aligned} \tag{1.1}$$

Thus, as long as $2de^{-2n\epsilon^2} \to 0$ for any fixed ϵ, we have the statistical consistency. This implies that we need

$$\frac{\log d}{n} \to 0,$$

which allows the number of questions/variables to increase a lot faster than the sample size n!

1.5 Estimators and Estimation Theory

Let $X_1, \cdots, X_n \sim F$ be a random sample. Here we can interpret F as the population distribution we are sampling from (that's why we are generating data from this distribution). Any numerical quantity (or even non-numerical quantity) of F that we are interested in is called the **parameter of interest**. For instance, the parameter of interest can be the mean of F, the median of F, standard deviation of F, first quartile of

F, ... etc. The parameter of interest can even be $P(X \geq t) = 1 - F(t) = S(t)$. The function $S(t)$ is called the *survival function*, which is a central topic in biostatistics and medical research.

When we know (or assume) that F is a certain distribution with some parameters, then the parameter of interest can be the parameter describing that distribution. For instance, if we assume F is an exponential distribution with an unknown parameter λ. Then this unknown parameter λ might be the parameter of interest.

Most of the statistical analysis is concerned with the following question:

> "*given the parameter of interest, how can I use the random sample to infer it?*"

Let $\theta = \theta(F)$ be the parameter of interest and let $\hat{\theta}_n$ be a statistic (a function of the random sample $X_1, \cdots, X_n$) that we use to estimate θ. In this case, $\hat{\theta}_n$ is called an *estimator*. For an estimator, there are two important quantities measuring its quality. The first quantity is the **bias**:

$$\mathsf{Bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta,$$

which captures the systematic deviation of the estimator from its target. The other quantity is the **variance** $\mathsf{Var}(\hat{\theta}_n)$, which measures the size of stochastic fluctuation.

Example. Let $X_1, \cdots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathsf{Var}(X)$. Assume the parameter of interest is the population mean μ. Then a natural estimator is the sample average $\hat{\mu}_n = \bar{X}_n$. Using this estimator, then

$$\mathbf{bias}(\hat{\mu}_n) = \mu - \mu = 0, \quad \mathsf{Var}(\hat{\mu}_n) = \frac{\sigma^2}{n}.$$

Therefore, when $n \to \infty$, both bias and variance converge to 0. Thus, we say $\hat{\mu}_n$ is a **consistent** estimator of μ. Formally, an estimator $\hat{\theta}_n$ is called a consistent estimator of θ if $\hat{\theta}_n \overset{P}{\to} \theta$.

The following lemma is a common approach to prove consistency:

Lemma 1.6 *Let $\hat{\theta}_n$ be an estimator of θ. If $\mathbf{bias}(\hat{\theta}_n) \to 0$ and $\mathsf{Var}(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \overset{P}{\to} \theta$. i.e., $\hat{\theta}_n$ is a consistent estimator of θ.*

In many statistical analysis, a common measure of the quality of the estimator is the *mean square error (MSE)*, which is defined as

$$\mathsf{MSE}(\hat{\theta}_n) = \mathsf{MSE}(\hat{\theta}_n, \theta) = \mathbb{E}\left((\hat{\theta}_n - \theta)^2\right).$$

By simple algebra, the MSE of $\hat{\theta}_n$ equals

$$\begin{aligned}
\mathsf{MSE}(\hat{\theta}_n, \theta) &= \mathbb{E}\left((\hat{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}\left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + \mathbb{E}(\hat{\theta}_n) - \theta)^2\right) \\
&= \underbrace{\mathbb{E}\left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2\right)}_{=\mathsf{Var}(\hat{\theta}_n)} + 2\underbrace{\mathbb{E}\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)}_{=0} \cdot (\mathbb{E}(\hat{\theta}_n) - \theta) + \left(\underbrace{\mathbb{E}(\hat{\theta}_n) - \theta}_{=\mathbf{bias}(\hat{\theta}_n)}\right)^2 \\
&= \mathsf{Var}(\hat{\theta}_n) + \mathbf{bias}^2(\hat{\theta}_n).
\end{aligned}$$

Namely, the MSE of an estimator is the variance plus the square of bias. This decomposition is also known as the *bias-variance tradeoff* (or bias-variance decomposition). By the Markov inequality,

$$\mathsf{MSE}(\hat{\theta}_n, \theta) \to 0 \Longrightarrow \hat{\theta}_n \overset{P}{\to} \theta.$$

i.e., if an estimator has MSE converging to 0, then it is a consistent estimator. The convergence of MSE is related to the L_2 convergence in probability theory.

Note that we write $\theta = \theta(F)$ for the parameter of interest because θ is a quantity derived from the population distribution F. Thus, we may say that the parameter of interest θ is a 'functional' (function of function; the input is a function, and the output is a real number).

1.6 O_P and o_P Notations

For a sequence of numbers a_n (indexed by n), we write $a_n = o(1)$ if $a_n \to 0$ when $n \to \infty$. For another sequence b_n indexed by n, we write $a_n = o(b_n)$ if $a_n/b_n = o(1)$.

For a sequence of numbers a_n, we write $a_n = O(1)$ if for all large n, there exists a constant C such that $|a_n| \leq C$. For another sequence b_n, we write $a_n = O(b_n)$ if $a_n/b_n = O(1)$.

Examples.

- Let $a_n = \frac{2}{n}$. Then $a_n = o(1)$ and $a_n = O\left(\frac{1}{n}\right)$.
- Let $b_n = n + 5 + \log n$. Then $b_n = O(n)$ and $b_n = o(n^2)$ and $b_n = o(n^3)$.
- Let $c_n = 1000n + 10^{-10}n^2$. Then $c_n = O(n^2)$ and $c_n = o(n^2 \cdot \log n)$.

Essentially, the big O and small o notation give us a way to compare the leading convergence/divergence rate of a sequence of (non-random) numbers.

The O_P and o_P are similar notations to O and o but are designed for random numbers. For a sequence of random variables X_n, we write $X_n = o_P(1)$ if for any $\epsilon > 0$,

$$P(|X_n| > \epsilon) \to 0$$

when $n \to \infty$. Namely, $P(|X_n| > \epsilon) = o(1)$ for any $\epsilon > 0$. Let a_n be a nonrandom sequence, we write $X_n = o_P(a_n)$ if $X_n/a_n = o_P(1)$.

In the case of O_P, we write $X_n = O_P(1)$ if for every $\epsilon > 0$, there exists a constant C such that

$$P(|X_n| > C) \leq \epsilon.$$

We write $X_n = O_P(a_n)$ if $X_n/a_n = O_P(1)$.

Examples.

- Let X_n be an R.V. (random variable) from a Exponential distribution with $\lambda = n$. Then $X_n = O_P(\frac{1}{n})$
- Let Y_n be an R.V from a normal distribution with mean 0 and variance n^2. Then $Y_n = O_P(n)$ and $Y_n = o_P(n^2)$.
- Let A_n be an R.V. from a normal distribution with mean 0 and variance $10^{100} \cdot n^2$ and B_n be an R.V. from a normal distribution with mean 0 and variance $0.1 \cdot n^4$. Then $A_n + B_n = O_P(n^2)$.

If we have a sequence of random variables $X_n = Y_n + a_n$, where Y_n is random and a_n is non-random such that $Y_n = O_P(b_n)$ and $a_n = O(c_n)$. Then we write

$$X_n = O_P(b_n) + O(c_n).$$

Examples.

- Let A_n be an R.V. from a uniform distribution over the interval $[n^2-2n, n^2+2n]$. Then $A_n = O(n^2)+O_P(n)$.
- Let X_n be an R.V from a normal distribution with mean $\log n$ and variance 10^{100}, then $X_n = O(\log n) + O_P(1)$.

The following lemma is an important property for a sequence of random variables X_n.

Lemma 1.7 *Let X_n be a sequence of random variables. If there exists a sequence of numbers a_n, b_n such that*

$$|\mathbb{E}(X_n)| \leq a_n, \quad \mathsf{Var}(X_n) \leq b_n^2.$$

Then

$$X_n = O(a_n) + O_P(b_n).$$

Examples.

- Let $X_1, \cdots, X_n$ be IID from $\mathsf{Exp}(5)$. Then the sample average
$$\bar{X}_n = O(1) + O_P(1/\sqrt{n}).$$
- Let $Y_1, \cdots, Y_n$ be IID from $N(5\log n, 1)$. Then the sample average
$$\bar{Y}_n = O(\log n) + O_P(1/\sqrt{n}).$$

Application.

- Let X_n be a sequence of random variables that are uniformly distributed over $[-n^2, n^2]$. It is easy to see that $|X_n| \leq n^2$ so $\mathbb{E}(|X_n|) \leq n^2$. Then by Markov's inequality,
$$P(|X_n| \geq t) \leq \frac{\mathbb{E}(|X_n|)}{t} \leq \frac{n^2}{t}.$$
Let $Y_n = \frac{1}{n^2}X_n$. Then
$$P(|Y_n| \geq t) = P\left(\frac{1}{n^2}|X_n| \geq t\right) = P(|X_n| \geq n^2 \cdot t) \leq \frac{n^2}{n^2 \cdot t} = t$$
for any positive t. This implies $Y_n = O_P(1)$ so $X_n = O_P(n^2)$.
- The Markov inequality and Chebeshev's inequality are good tools for deriving the O_P bound. For a sequence of random variables $\{X_n : n = 1, \cdots\}$, the Markov inequality implies
$$X_n = O_P(\mathbb{E}(|X_n|)).$$
The Chebeshev's inequality implies
$$X_n = O_P(\sqrt{\mathsf{Var}(X_n)})$$
if $\mathbb{E}(X_n) = 0$.
- If we obtain a bound like equation (1.1), we can use O_P notation to elegantly denote it as
$$\|\hat{\pi} - \pi\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right).$$