**STAT 535: Statistical Machine Learning** **Autumn 2019**

# 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* $\Omega$ 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 $\Omega$ 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 $\sigma$-algebra.

A *$\sigma$-algebra* $\mathcal{F}$ is a collection of subsets of $\Omega$ 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 $\sigma$-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 $A_1, A_2, \ldots$ 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 $\Omega$, 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 $\Omega$. 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.

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

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

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

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 $F(x) = \sum_{x' \leq x} p(x')$.

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

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

[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 $\lambda$, 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

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

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 $(\mu, \sigma^2)$, 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 $\lambda$, 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 $\mu_X$ and $\mu_Y$ and variance being $\sigma_X^2$ and $\sigma_Y^2$. 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*:

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

*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

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

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

$$Z_n \xrightarrow{P} Z$$

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

$$Z_n \xrightarrow{a.s.} Z$$

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 $\mu$ is the target of convergence)

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

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

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

$$\begin{aligned} X_n + Y_n &\xrightarrow{D} X + c \\ X_n Y_n &\xrightarrow{D} cX \\ X_n / Y_n &\xrightarrow{D} X/c \quad (\text{if } c \neq 0). \end{aligned}$$

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 $\mu$. i.e.,*

$$\bar{X}_n \xrightarrow{a.s} \mu.$$

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*

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

*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 $\pi$? 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 $\pi_j$ 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 $\epsilon$, 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 $\lambda$. Then this unknown parameter $\lambda$ 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 $\theta$. 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 $\mu$. 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 $\mu$. Formally, an estimator $\hat{\theta}_n$ is called a consistent estimator of $\theta$ if $\hat{\theta}_n \xrightarrow{P} \theta$.

The following lemma is a common approach to prove consistency:

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

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,

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

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 $\theta$ is a quantity derived from the population distribution $F$. Thus, we may say that the parameter of interest $\theta$ 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*

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

*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 $|X_n| \le n^2$ so $\mathbb{E}(|X_n|) \le n^2$. Then by Markov's inequality,

$$P(|X_n| \ge t) \le \frac{\mathbb{E}(|X_n|)}{t} \le \frac{n^2}{t}.$$

  Let $Y_n = \frac{1}{n^2} X_n$. Then

$$P(|Y_n| \ge t) = P\left(\frac{1}{n^2}|X_n| \ge t\right) = P(|X_n| \ge n^2 \cdot t) \le \frac{n^2}{n^2 \cdot t} = t$$

  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

$$\|\hat{\pi} - \pi\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right).$$

## 1.7 Introduction: Statistical Inference

Statistical inference is about drawing conclusions from the data. This process often involves with estimating some parameters of interest. In a parametric model, the parameter of interest are often the parameters of the corresponding parametric family. For a parametric model, there are three common tasks in statistical inference – estimating the underlying parameter, providing an interval inference about the underlying parameter, and testing if the underlying parameter satisfy certain conditions.

In statistics, there are two major paradigm for making inference – the frequentist paradigm and the Bayesian paradigm[2]. We will talk about their principles of estimation, interval inference, and testing.

Note that there is no right or wrong about each paradigm – they are just different ways of making arguments. Each paradigm is a self-consistent way to making logical arguments and has its own advantages and limitations.

## 1.8 Frequentist Approach

The Frequentist approach is the paradigm we learn from Statistics 101. It interprets probability as the long term frequency. In Frequentist approach, the parameter of interest is a fixed and unknown number.

### 1.8.1 Estimation

In a parametric model, we often estimate the parameter of interest using the so-called **maximum likelihood estimator (MLE)**. The idea is very simple. Suppose we observe only one observation $X$ from a PDF/PMF $p(x)$. The parametric model assumes that such a PDF/PMF can be written as $p(x) = p(x;\theta)$, where $\theta$ is the parameter of the model ($\theta$ is often the parameter of interest) inside a parameter space $\Theta$ ($\theta \in \Theta$). The idea of MLE is to ask the following question: given the observation $X$, which $\theta$ is the *most likely* parameter that generates $X$? To answer this question, we can vary $\theta$ and examine the value of $p(X;\theta)$.

Because we are treating $X$ as fixed and $\theta$ being something that we want to optimize, we can view the problem as finding the best $\theta$ such that the **likelihood function** $L(\theta|X) = p(X;\theta)$ is maximized. The MLE uses the $\theta$ that maximizes the likelihood value. Namely,

$$\hat{\theta}_{MLE} = \mathsf{argmax}_\theta L(\theta|X).$$

When we have multiple observations $X_1, \cdots, X_n$, the likelihood function can be defined in a similar way – we use the joint PDF/PMF to define the likelihood function. Let $p(x_1, \cdots, x_n;\theta)$ be the joinr PDF/PMF. Then the likelihood function is

$$L_n(\theta) = L(\theta|X_1, \cdots, X_n) = p(X_1, \cdots, X_n;\theta).$$

Note that when we assume IID observations,

$$L_n(\theta) = \prod_{i=1}^n L(\theta|X_i) = \prod_{i=1}^n p(X_i;\theta).$$

In many cases, instead of using the likelihood function, we often work with the **log-likelihood function**

$$\ell_n(\theta) = \log L_n(\theta).$$

[2]There are also other paradigms such as the fiducial inference (https://en.wikipedia.org/wiki/Fiducial_inference) but the Frequentists and Bayesian are the two major paradigm.

Because taking the logarithmic does not change the maximizer of a function, the maximizer of the log-likelihood function is the same as the maximizer of the likelihood function. There are both computational and mathematical advantages of using a log-likelihood function over likelihood function. To see this, we consider the case of IID sample. Computationally, the likelihood function often has a very small value due to the product form of PDF/PMFs. So it is very likely that the number if too small, making the computation very challenging. Mathematically, when we take log of the likelihood function, the product of PDF/PMFs becomes an additive form

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log p(X_i; \theta).$$

Under IID assumption, each $\log p(X_i; \theta)$ is an IID random variable so the central limit theorem and the law of large number can be applied to the average, making it possible to analyze it asymptotic behavior.

Since under the IID assumptions, we have many advantages, we will assume IID from now on. Because MLE finds the maximum of $\ell_n(\theta)$, a common trick to find MLE is to study the gradient of the log-likelihood function, which is also known as the **score function**:

$$S_n(\theta) = \frac{\partial}{\partial \theta} \ell_n(\theta) = \sum_{i=1}^{n} s(\theta | X_i),$$

where $s(\theta|X_i) = \frac{\partial}{\partial \theta} \ell(\theta|X_i) = \frac{\partial}{\partial \theta} \log p(X_i; \theta)$. Under suitable conditions, the MLE satisfies the *score equation*:

$$S_n(\hat{\theta}_{MLE}) = 0.$$

Note that if there are more than one parameter, say $\theta \in \mathbb{R}^p$, the score equation will be a system of $p$ equations.

Because the MLE is at the maximal point of the likelihood function, the curvature of the likelihood function around the maximal will determine its stability. To measure the curvature, we use the **Fisher's information matrix**:

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^T} \ell_n(\theta)\right] = n \cdot I_1(\theta) = n \cdot -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^T} p(X_1; \theta)\right].$$

If the data is generated from a PDF/PMF $p(x; \theta_0)$ and some regularity conditions are satisfied,

$$\mathbb{E}(S_n(\theta_0)) = 0, I_1(\theta_0) = \mathbb{E}(S_1(\theta_0) S_1^T(\theta_0)).$$

Moreover,

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) \xrightarrow{D} N(0, I_1^{-1}(\theta)).$$

Namely, the MLE is asymptotically normally distributed around the true parameter $\theta_0$ and the covariance is determined by the Fisher's information matrix. Note that the asymptotic normality also implies that $\hat{\theta}_{MLE} - \theta_0 \xrightarrow{P} 0$.

**Example 1: Binomial Distribution.** Assume that we obtain a single observation $Y \sim \mathsf{Bin}(n, p)$, and we assume that $n$ is known. The goal is to estimate $p$. The log-likelihood function is

$$\ell(p) = Y \log p + (n - Y) \log(1 - p) + C_n(Y),$$

where $C_n(Y) = \log \binom{n}{Y}$ is independent of $p$. The score function is

$$S(p) = \frac{Y}{p} - \frac{n - Y}{1 - p}$$

so solving the score equation gives us $\hat{p}_{MLE} = \frac{Y}{n}$. Moreover, the Fisher's information is

$$I(p) = \mathbb{E}\left\{\frac{\partial}{\partial p} S(p)\right\} = -\frac{\mathbb{E}(Y)}{p^2} - \frac{n - \mathbb{E}(Y)}{(1-p)^2} = \frac{n}{p(1-p)}.$$

**Example 2: Multinomial Distribution.** Let $X_1, \cdots, X_n$ be IID from a multinomial distribution such that $P(X_1 = j) = p_j$ for $j = 1, \cdots, s$ and $\sum_{j=1}^{s} p_j = 1$. Note that the parameter space is $\Theta = \{(p_1, \cdots, p_s) : 0 \le p_j, \sum_{j=1}^{s} p_j = 1\}$. By setting $N_j = \sum_{i=1}^{n} I(X_i = j)$ for each $j = 1, \cdots, s$, we obtain the random vector $(N_1, \cdots, N_s) \sim \mathsf{Multinomial}(n, p)$, where $p = (p_1, \cdots, p_s)$. The parameters of interest are $p_1, \cdots, p_s$. In this case, the likelihood function is

$$L_n(p_1, \cdots, p_s) = \frac{n!}{N_1! \cdots N_s!} p_1^{N_1} \cdots p_s^{N_s}$$

and the log-likelihood function is

$$\ell_n(p_1, \cdots, p_s) = \sum_{j=1}^{s} N_j \log p_j + C_n,$$

where $C_n$ is independent of $p$. Note that naively computing the score function and set it to be 0 will not grant us a solution (think about why) because we do not use the constraint of the parameter space – the parameters are summed to 1. To use this constraint in our analysis, we consider adding the Lagrange multipliers and optimize it:

$$F(p, \lambda) = \sum_{j=1}^{s} N_j \log p_j + \lambda \left(1 - \sum_{j=1}^{s} p_j\right).$$

Differentiating this function with respect to $p_1, \cdots, p_s$, and $\lambda$ and set it to be 0 gives

$$\frac{\partial F}{\partial p_j} = \frac{N_j}{p_j} - \lambda 0 \Rightarrow N_j = \lambda \hat{p}_{MLE,j}$$

and $1 - \sum_{j=1}^{s} p_j = 0$. Thus, $n = \sum_{j=1}^{s} N_j = \lambda \sum_{j=1}^{p} = \lambda$ so $\hat{p}_{MLE,j} = \frac{N_j}{n}$.

### 1.8.2 Confidence Intervals

In some analysis, we not only want to have just a point estimate of the parameter of interest, but also want to use an interval to infer the parameter of interest. And we also want to assign a level to this interval to describe how 'accurate' this interval is. Note that here the concept of accuracy is not well-defined – we will talk about it later. Ideally, given an accuracy level, we want an interval as small as possible.

The Frequentist and the Bayesian defines the accuracy differently so their construction of intervals are also different. In short, the Frequentists defines the accuracy as the *long term frequency coverage* of the underlying true parameter of interest whereas the Bayesian defines the accuracy in terms of covering the posterior probability. In this section, we will talk about the Frequentist approach and the interval is known as the **confidence interval**. The accuracy that Frequentists are using is called the *confidence level*.

Formally, given a confidence level $1 - \alpha$, a confidence interval of $\theta_0$ is a random interval $C_{n,\alpha}$ that can be constructed solely from the data (i.e., can be constructed using $X_1, \cdots, X_n$) such that

$$P(\theta \in C_{n,\alpha}) \ge 1 - \alpha + o(1).$$

Beware, what is random is not $\theta$ but the interval $C_{n,\alpha}$. The quantity $P(\theta \in C_{n,\alpha})$ is also called the (Frequentist) coverage. Note that we allow the coverage to be *asymptotically* $1 - \alpha$; when there is no $o(1)$

term, we will say that the confidence interval has a finite sample coverage. A confidence interval with the above property is also called a (asymptotically) valid confidence interval.

**Normal confidence interval.** A traditional approach to constructing a confidence interval of $\theta_0$ is based on the asymptotic normality of the MLE:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{D} N(0, I_1^{-1}(\theta_0)).$$

When the dimension of the parameter is 1, a simple confidence interval is

$$\hat{\theta}_{MLE} \pm z_{1-\alpha/2} \cdot \sigma_{\theta_0},$$

where $\sigma^2_{\theta_0} = I_1^{-1}(\theta_0)$ Such interval is not a confidence interval because $\theta_0$ is unknown. We can modify it using a plug-in estimate of the Fisher's information:

$$C_{n,\alpha} = [\hat{\theta}_{MLE} - z_{1-\alpha/2} \cdot \sigma_{\hat{\theta}_{MLE}}, \hat{\theta}_{MLE} + z_{1-\alpha/2} \cdot \sigma_{\hat{\theta}_{MLE}}],$$

where $z_\beta$ the $\beta$-percentile of the standard normal distribution. Using the slutsky's theorem, you can easily show that this confidence interval has the asymptotic coverage.

When the dimension of the parameter is greater than 1, there are multiple ways we can construct a confidence interval. Note that in this case, the set $C_{n,\alpha}$ is no longer an interval but a region/set so it is often called a confidence region/set. A simple approach of constructing a confidence set is via an ellipse. Note that the asymptotic normality also implies that (using continuous mapping theorem)

$$n(\hat{\theta}_{MLE} - \theta_0)^T I_1(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{D} \chi^2_p,$$

where $\chi^2_p$ denotes the $\chi^2$ distribution with a degree of freedom $p$. So we construct the confidence set using

$$C_{n,\alpha} = \left\{\theta : n(\hat{\theta}_{MLE} - \theta)^T I_1(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta) \le \chi^2_{p,1-\alpha}\right\},$$

where $\chi^2_{p,\beta}$ is the $\beta$-percentile of the $\chi^2$ distribution with a degree of freedom $p$.

**Bootstrap confidence interval.** Bootstrap approach is an Monte Carlo method for assessing the uncertainty of an estimator. It can be used to compute the variance of an estimator (not necessarily the MLE) and construct a confidence interval. In the case of likelihood inference, the bootstrap approach has an advantage that we do not need to know the closed-form of $I_1(\theta)$ to construct the confidence interval or to approximate the variance of the MLE.

While there are many variant of bootstrap methods, we introduce the simplest one – the empirical bootstrap. For simplicity, we assume that the dimension of $\theta$ is 1 (the bootstrap works for higher dimensions as well). Let $X_1, \cdots, X_n$ be the original sample. We then *sample with replacement* from the original sample to obtain a new sample of each size $X_1^*, \cdots, X_n^*$. This new sample is called a bootstrap sample. We find the MLE using the bootstrap sample and let $\hat{\theta}^*_{MLE}$ denote the bootstrap MLE. Now we repeat the bootstrap process $B$ times, leading to $B$ bootstrap MLEs

$$\hat{\theta}^{*(1)}_{MLE}, \cdots, \hat{\theta}^{*(B)}_{MLE}.$$

Let $t_\beta$ denotes the $\beta$-percentile of these $B$ values, i.e.,

$$\hat{t}_\beta = \hat{G}^{-1}(\beta), \quad \hat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}^{*(b)}_{MLE} \le t).$$

Then the bootstrap confidence interval of $\theta_0$ is

$$C_{n,\alpha} = [\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2}].$$

One can prove that under very mild conditions, the bootstrap confidence interval has asymptotic coverage.

The power of the bootstrap method is that *we do not use anything about the Fisher's information*! As long as we can compute the estimator, we can construct an asymptotically valid confidence interval. Note that if we do know the Fisher's information, the bootstrap method can be modified using the bootstrap $t$-distribution method, which provides a better asymptotic coverage (namely, the $o(1)$ decays faster to 0 than the above method and the normal confidence interval)[3].

### 1.8.3 Test of Significance

Statistical test is about how to design a procedure that allows us to make scientific discovery. Such a procedure has to be able to handle the uncertain nature of our data. In statistics, we model the data as random variables so the testing procedure needs to account for the randomness.

Let $\mathcal{D}_n = \{X_1, \cdots, X_n\}$ denotes our data. The testing procedure involves two competing hypotheses:

- *Null hypothesis* $H_0$: the hypothesis that we want to challenge. It is often related to the current scientific knowledge.
- *Alternative hypothesis* $H_a$: the hypothesis that complements to the null hypothesis. It is the hypothesis we would like to prove to be plausible using our data.

The goal is to see if we have strong enough evidence (from $\mathcal{D}_n$) that we can argue the alternative hypothesis is more reasonable than the null hypothesis. If we do have enough evidence, then we will reject the null hypothesis. When the null hypothesis reflects the scenarios that can be explained by the current scientific knowledge, rejecting the null hypothesis means that we have discovered something new.

To design a testing procedure, we need to quantify the notion of evidence. The Frequentists and the Bayesian use different ways to measure the evidence. The Frequentist approach is the *p-value* whereas the Bayesian approach is the *Bayes factor*. We will talk about Bayes factor later so here we focus on the p-value.

Here is a summary on Frequentist approach of hypothesis test.

1. Based on the model and null hypothesis, design a test statistic.
2. Compute the distribution of the test statistics under the null hypothesis.
3. Plug-in the data into the test statistic, compute the probability of observing a more extreme data against the null hypothesis. Such a probability is the p-value.
4. Compare p-value to the significance level. If p-value is less than the significance level, we reject the null hypothesis.

The central idea of hypothesis test is to control the *type-1 error*, the probability of falsely rejecting $H_0$ when $H_0$ is correct. Essentially, the p-value can be interpreted as *if we reject the null hypothesis (under this p-value), then our type-1 error is the same as the p-value*. The significance level reflects the amount of type-1 error we can tolerate so when p-value is less than the significance level, we can reject $H_0$. Due to the construction of p-value, a small p-value means that the null hypothesis does not fit to the data very well (so we are seeing an extreme event if $H_0$ is true). Thus, small p-value or rejecting $H_0$ under a small significance level means that we have more evidence against $H_0$.

[3] see, e.g., Chapter 2 of *All of nonparametric statistics* by Larry Wasserman and Chapter 3.5 of *The bootstrap and Edgeworth expansion* by Peter Hall.

Note that there is another quantity called *type-2 error*, the probability of not rejecting $H_0$ when $H_0$ is false. Namely, type-2 error is concerned with the case that we fail to reject $H_0$ when we should.

In statistics, we often control type-1 error first and the hope that the type-2 error is also small. When do we put more emphasis on type-1 error? This has something to do with the philosophy of scientific research. The scientific approach is a systematic way to acquire reliable knowledge. Thus, every discovery we made should be accompanied with sufficient evidences. In Frequentist approach, the measure of evidence against $H_0$ is the p-value – the smaller p-value, the more evidence. Thus, controlling type-1 error means that we put requirements on the amount of evidence we need to claim a scientific discovery.

While there are many possible ways to construct a test statistic, here we consider two common approaches: Wald test and the likelihood ratio test.

**Wald test.** Assume in a simple case where we model the data as IID from a parametric model $p(x;\theta)$ with a $1D$ parameter. Suppose that we are comparing two hypotheses

$$H_0 : \theta = \theta_1, \qquad H_a : \theta \neq \theta_1.$$

Since the MLE is a good estimate of $\theta$, we can design our test statistic using the MLE. Because we know that the MLE has asymptotic normality,

$$\sqrt{n} I_1(\theta)^{-1/2}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0,1),$$

we can then use

$$T_n(\theta) = \sqrt{n} I_1(\theta)^{-1/2}(\hat{\theta}_n - \theta)$$

as our test statistic. If $H_0$ is true, $T_n(\theta_1)$ should behaves like a standard normal distribution. When we observe the actual data, we can then compare the observed value of $T_n(\theta_1)$ against 0 (since $T_n(\theta_1) = 0$ means a perfect match with $H_0$). We reject $H_0$ if $T_n(\theta_1)$ is either too large or too small. The p-value is then

$$p(\mathcal{D}_n) = 1 - 2 \cdot \Phi^{-1}(|T_n(\theta_1)|),$$

where $\Phi(t) = P(Z \leq t)$ is the CDF of a standard normal distribution. Note that the above method is also known as the **Wald test**.

**Likelihood ratio test (LRT).** LRT is another popular way to conduct hypothesis test under a parametric model. It can be easily applied to multivariate parameters so now we assume that there are $p$ parameters, i.e., $\theta \in \Theta \subset \mathbb{R}^p$. The null hypothesis and the alternative hypothesis are

$$H_0 : \theta \in \Theta_0, \qquad H_a : \theta \in \Theta \backslash \Theta_0.$$

Let $\ell_n(\theta)$ be the log-likelihood function and let

$$\hat{\theta}_{MLE} = \mathsf{argmax}_{\theta \in \Theta} \ell_n(\theta), \quad \hat{\theta}_{MLE,0} = \mathsf{argmax}_{\theta \in \Theta_0} \ell_n(\theta)$$

be the global MLE and the constrained MLE (in the null hypothesis space). The LRT starts with a test statistic

$$\mathsf{LRT} = \frac{\sup_{\theta \in \Theta_0} L_n(\theta)}{\sup_{\theta \in \Theta} L_n(\theta)} = \frac{L_n(\hat{\theta}_{MLE,0})}{L_n(\hat{\theta}_{MLE})}$$

and reject $H_0$ if $\mathsf{LRT}$ is too small. To get the p-value of the LRT, we often use the following fact:

$$-2 \log \mathsf{LRT} = 2\left(\ell_n(\hat{\theta}_{MLE}) - \ell_n(\hat{\theta}_{MLE,0})\right) \xrightarrow{D} \chi^2_r,$$

where $r = \mathsf{dim}(\Theta) - \mathsf{dim}(\Theta_0)$ is the difference of the dimensions/degree-of-freedoms between the full parameter space and constrained parameter space. The above result is also known as the likelihood ratio approximation

or Wilk's theorem. The LRT proposes to use the $T_n = 2\left(\ell_n(\hat{\theta}_{MLE}) - \ell_n(\hat{\theta}_{MLE,0})\right)$ as our test statistic and compared it with the CDF of $\chi^2_r$ to obtain the p-value.

**Remarks.**

- *Why $r$ degrees of freedom?* Some of you may be wondering why we are getting a $\chi^2$ distribution with a degree of freedom $r$. Here is a simple explanation using geometry. Recall that the MLE behaves like a normal distribution around the true parameter $\theta_0$. If $H_0$ is true, $\theta_0 \in \Theta_0$ so $\hat{\theta}_{MLE}$ will be asymptotically normally distributed around $\theta_0$ . Since $\hat{\theta}_{MLE}$ is the maximizer over $\Theta$ so it has $p$ degrees of freedom (it can moves in each of the $p$ dimensions). The constrained MLE $\hat{\theta}_{MLE,0}$ is the maximizer under $\Theta_0$, which has $r$ constraints. Thus, $\Theta_0$ has $p - r$ degrees of freedom, which implies that its maximizer $\hat{\theta}_{MLE,0}$ uses $p - r$ degree of freedom. The remaining degrees of freedom is $p - (p - r) = r$ so this is why we are obtaining a $\chi^2$ distribution with $r$ degrees of freedom. Note that one can replace the $r$ constraints to be saying that $\Theta_0$ is a $p - r$ dimensional manifold.
- *Equivalence between likelihood ratio test and Wald test.* Asymptotically, one can show that the likelihood ratio test and the Wald test are the same under classical assumptions on the MLE. Also, there is another test that is closely related to them called the *score test*, which is based on the value of score function as a test statistic. Again, asymptotically the score test and other two tests are equivalent. However, when the likelihood function is more complex, such as having multiple local maxima. These three tests may not be the same (they can be quiet different)[4].
- *Relation to confidence interval.* Hypothesis test can be used to construct a confidence interval. Consider testing the null hypothesis: $H_0 : \theta = \theta_1$ for a specific value $\theta_1$ under a significance level $\alpha$. For each $\theta_1 \in \Theta$, we can do a hypothesis test. Some parameters will be rejected whereas the others will not. Let $\hat{A}_{n,\alpha}$ be the collection of parameters that the null hypothesis will not be rejected. Then you can show that $\hat{A}_{n,\alpha}$ is a confidence interval of $\theta$. This approach is also known as the confidence interval from inverting a hypothesis test. Note that if we are given a $1 - \alpha$ confidence interval of the parameter of interest $\theta$, we can use it to test the null hypothesis $H_0 : \theta \in \Theta_0$. If the confidence interval intersects with $\Theta_0$, we cannot reject $H_0$. If they are disjoint, then we can reject $H_0$ under a significance level $\alpha$. Thus, hypothesis test problem and confidence intervals are highly related to each other. However, testing *a given hypothesis* is often an easier problem than constructing a confidence interval because confidence interval requires a procedure that is valid regardless of the null hypothesis being correct or not. On the other hand, hypothesis test problem only requires a procedure that works when $H_0$ is true.

There are many misconceptions about p-values. To clarify what p-value stands for and what it does NOT stand for, I obtain the following 6 principles from American Statistical Association's website[5]:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

[4] see https://arxiv.org/abs/1807.04431

[5] https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

### 1.8.4 Model Mis-specification

Many theory about the MLE assumes that the population distribution function belongs to our parametric family. However, this is a very strong assumption in reality. It is very likely that the population distribution function does not belong to our parametric family (e.g., the population PDF is not Gaussian but we fit a Gaussian to it). What will happen in this case for our MLE? Will it still converge to something? If so, what will be the quantity that it is converging?

Model mis-specification studies the situation like this – we assume a wrong model for the population distribution function. Let $p_0(x)$ be the population PDF and we assume that the population PDF can be written as $p(x;\theta)$. However, $p_0 \neq p(x;\theta)$ for every $\theta \in \Theta$. It turns out that the MLE $\hat{\theta}_{MLE}$ still converges under mild assumptions to a quantity $\theta^*$ in probability. Moreover, the corresponding PDF/PMF $p(x;\theta^*)$ has an interesting relation with $p_0(x)$. Assume that the RV $X$ has a PDF/PMF $p_0$. Then

$$\mathbb{E}\left\{\log\left(\frac{p_0(X)}{p(X;\theta^*)}\right)\right\} = \inf_{\theta\in\Theta} \mathbb{E}\left\{\log\left(\frac{p_0(X)}{p(X;\theta)}\right)\right\} = \inf_{\theta\in\Theta} \mathsf{KL}(p_0, p_\theta),$$

where $\mathsf{KL}$ is also known as the *Kullback-Liebler (KL)* divergence and $p_\theta(x) = p(x;\theta)$. Namely, the MLE corresponds to the parametric distribution in the specified family that minimizes the KL divergence to the population distribution.

In model mis-specification case, the MLE still satisfies the score equation (under appropriate assumption) but the Fisher's information may not reflect the actual curvature of the likelihood function around $\theta^*$. The asymptotic covariance (related to the curvature) of $\hat{\theta}_{MLE}$ will be $\Sigma = I_1^{-1}(\theta^*)\mathbb{E}(S_1(\theta^*)S_1^T(\theta^*))I_1^{-1}(\theta^*)$ and we still have

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta^*) \xrightarrow{D} N(0, \Sigma).$$

## 1.9 Bayesian Approach

The Bayesian inference is an alternative statistical paradigm to the Frequentist approach. The Bayesian approach interprets the probability in a broader sense that include subjective probability, which allows us to assign probability to almost every quantity in our model (including the parameter of interest and even a statistical model). The Bayesian inference relies on a simple decision theoretic rule – if we are competing two or more choices, we always choose the one with higher probability. This simple rule allows us to design an estimator, construct an interval, and perform hypothesis test.

In the Bayesian analysis, we assign a probability to every parameter in our model. For a parametric model $p(x;\theta)$, the parameter of interest $\theta$ is given a **prior distribution** $\pi(\theta)$ that reflects our belief about the value of $\theta$. In a sense, the prior distribution quantifies our subjective belief about the parameter $\theta$. The higher value of $\pi(\theta)$ indicates that we believe that $\theta$ is a more likely value of it.

How do we interpret this prior distribution? Here is a decision theoretic way of viewing it. To simplify the problem we assume that $\Theta = \{0, 1, 2\}$. Even without any data at hand, we can ask ourselves about our belief about each parameter value. Some people may think that 1 is the most likely one; some may think that 2 is the most likely one. To make our belief more precise, we use probability to work on it. Let $\pi(j)$ be the number that reflects our belief about $\theta = j$. We interpret the numerical value of $\pi(j)$ as follows. We are forced to guess the answer of $\theta = j$ versus $\theta \neq j$. If the answer is $\theta = j$ and we indeed guess it correctly, we will be rewarded $\delta$ dollar. If the answer is $\theta \neq j$ and we get it correct, we will be rewarded 1 dollar. If

we get it wrong, we do not lose anything. Our principle is to maximize our expected reward. Now assume that the true value of $\theta$ has equal probability of being $j$ or not $j$. Then what should we choose? $\theta = j$ or $\theta \neq j$? Now we think about this problem by varying $\delta$ from 0 to infinity. When $\delta$ is small, unless *we have very strong belief on* $\theta = j$, we will not bid on it. When increasing $\delta$, at certain threshold we will switch our decision from bidding on $\theta \neq j$ to $\theta = j$. Let this threshold be $\eta_j$. $\eta_j$ is a number that reflects our belief about $\theta = j$ and we associate it with our prior

$$\pi(j) = \frac{1}{1+\eta_j} \Leftrightarrow \eta_j = \frac{1-\pi(j)}{\pi(j)} \quad \text{(odds of } \theta = j\text{)}.$$

Here, you see that we only use one simple decision rule – bidding on the one with a higher expected outcome. This allows us to quantify our belief.

Using the prior distribution, the Bayesian probability model can be written as follows:

$$\begin{aligned} X_1, \cdots, X_n | \theta &\overset{IID}{\sim} p(x|\theta) \\ \theta &\sim \pi. \end{aligned}$$

The Bayesian inference focuses on the distribution of $\theta$ after observing $X_1, \cdots, X_n$:

$$\pi(\theta|X_1, \cdots, X_n) = \frac{p(X_1, \cdots, X_n, \theta)}{p(X_1, \cdots, X_n)} \propto \underbrace{p(X_1, \cdots, X_n|\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}.$$

This distribution is also known as the **posterior distribution.**

The posterior distribution informs us about how our prior belief is updated after seeing the data. It is the central quantity in Bayesian inference – all our decisions will be related to it. In Bayesian's point of view, probability models are just mathematical tools for analyzing data. We do not assume that the data is generated from a probability distribution. We just *view* the data as generated from $p(x;\theta)$. Given that we do not assume the probability model to be the *true* model, there is NO true parameter so we cannot talk about conventional statistical errors. However, Bayesian does have another way to expressing the error in our inference – the posterior distribution. The posterior distribution reflects our belief about the parameter after seeing the data, we can use it as a measure of *uncertainty* about $\theta$. If the posterior distribution is more spread out, then the uncertainty in our inference is larger. On the other hand, if the posterior distribution is very concentrated, then there is very little (Bayesian) uncertainty.

### 1.9.1 Bayesian Estimation

There are two common estimator in Bayesian inference: the posterior mean and the maximum a posteriori estimation (MAP).

**Posterior mean.** Just like we often use the sample mean as an estimator of the population mean, the mean of the posterior distribution is a common quantity that was used as an estimator of $\theta$:

$$\hat{\theta}_\pi = \mathbb{E}(\theta|X_1, \cdots, X_n) = \int \theta \cdot \pi(\theta|X_1, \cdots, X_n) d\theta.$$

It represents the average location of our belief about the parameter after seeing the data.

**Maximum a posteriori estimation (MAP).** Another common estimator of $\theta$ is the MAP; it relies on the similar principle as the MLE – we choose the one that is the most likely. Here the 'likely' is interpreted as our posterior belief about the parameter of interest $\theta$. Formally, MAP is defined as

$$\hat{\theta}_{MAP} = \mathsf{argmax}_\theta \pi(\theta|X_1, \cdots, X_n).$$

**Example: Binomial Sampling.** Assume that we have an observation $Y \sim \mathsf{Bin}(N, \theta)$ where $N$ is known and the parameter of interest is $\theta$:

$$P(Y = y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}.$$

We use a Beta distribution with parameters $(\alpha, \beta)$ as our prior distribution for $\theta$. Namely,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function and $\alpha, \beta > 0$. Note that $(\alpha, \beta)$ are called the hypter-parameters and are known quantities (because we know our belief about the data). For a Beta distribution with parameter $\alpha, \beta$, the mean is $\frac{\alpha}{\alpha+\beta}$.

The posterior distribution is

$$\begin{aligned}\pi(\theta|Y) &= \frac{\binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &\propto \theta^{Y+\alpha-1}(1-\theta)^{N-Y+\beta-1}\end{aligned}$$

so it is a Beta distribution with parameters $(Y+\alpha, N-Y+\beta)$. Then the posterior mean and MAP are

$$\hat{\theta}_\pi = \frac{Y+\alpha}{N+\alpha+\beta}, \qquad \hat{\theta}_{MAP} = \frac{Y+\alpha-1}{N+\alpha+\beta-2}$$

(these are the mean and the mode of a Beta distribution).

Note that in this problem, the MLE is $\hat{\theta}_{MLE} = \frac{Y}{N}$. Thus, the posterior mean has an interesting decomposition:

$$\begin{aligned}\hat{\theta}_\pi &= \frac{Y+\alpha}{N+\alpha+\beta} \\ &= \hat{\theta}_\pi = \frac{Y}{N+\alpha+\beta} + \frac{\alpha}{N+\alpha+\beta} \\ &= \frac{Y}{N}\times\frac{N}{N+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\times\frac{\alpha+\beta}{N+\alpha+\beta} \\ &= \hat{\theta}_{MLE}\times W + \text{Prior mean}\times(1-W),\end{aligned}$$

where $W = \frac{N}{N+\alpha+\beta}$ is a weight that is tending to 1 when $N \to \infty$. This phenomenon – the posterior mean can be written as the weighted average of the MLE and the prior mean – occurs in several scenarios. Moreover, the fact that the weights $W \to 1$ as the sample size $N \to \infty$ means that when we have more and more data, the prior distribution seems to be irrelevant. Thus, the posterior mean would have a similar asymptotic property as the sample mean. However, this is not a general phenomenon; often only certain combination of prior and likelihood models will have this feature.

**Remark**

- *Choice of prior and conjugate prior.* The choice of prior reflects our belief about the parameter before seeing any data. Sometimes people want to choose a prior distribution such that the posterior distribution is in the same family as the prior distribution, just like what we have observed in the above example. If a prior distribution and a likelihood function leads to a posterior that belongs to the same family as the prior, we call this prior **conjugate prior**. There are several conjugate priors know to date, see https://en.wikipedia.org/wiki/Conjugate_prior for an incomplete list of cases.

  Another common choice of prior is called the **Jeffreys prior**[6], which chooses a prior $\pi(\theta) \propto \sqrt{\det(I_1(\theta))}$,

[6] see https://en.wikipedia.org/wiki/Jeffreys_prior for more details.

where $I_1(\theta)$ is the Fisher's information matrix. One can view the Jeffreys prior as the prior that *we do not have any prior belief about* $\theta$; or more formally, an uninformative prior.

- *Challenge of computing the posterior.* In general, if we do not choose a conjugate prior, the posterior distribution could be difficult to compute. The challenge often comes from the normalization quantity $p(X_1, \cdots, X_n)$ in the denominator of the posterior $\pi(\theta|X_1, \cdots, X_n)$ (the numerator is just the prior times the likelihood). In practice we will use Monte Carlo method to compute the posterior – we generate points from $\pi(\theta|X_1, \cdots, X_n)$ and as we generate enough points, these points should approximate the true posterior distribution well. We will talk more about this later in the lecture of MCMC (Monte Carlo Markov Chain).

- *Consistency.* In pure Bayesian's point of view, statistical consistency is not an important property because probability model is a working model to describe the data and we do not need to assume that there exists an actual parameter that generates the data. Thus, the posterior distribution is the quantity that we really need to make our inference. However, sometimes Bayesian estimators, such as the posterior mean or MAP, does have statistical consistency. Namely, $\hat{\theta}_\pi \xrightarrow{P} \theta_0$ and $\hat{\theta}_{MAP} \xrightarrow{P} \theta_0$, where the data $X_1, \cdots, X_n \overset{IID}{\sim} p(x; \theta_0)$. This is often related to the Bernstein-von Mises theorem[7]. Although statistical consistency was not an important property in Bayesian paradigm (because Bayesian does not assume the data is indeed from a probability model; probability models are just a mathematical model to help us analyze the data), still many researchers would prove consistency when proposing a Bayesian approach.

### 1.9.2 Bayesian Interval: Credible Interval

The Bayesian's interval is very straight forward – since the posterior contains our belief about the parameter after seeing the data, we just construct the interval using the posterior distribution. Given a credible level $1 - \alpha$, a credible interval $D_{n,\alpha}$ satisfies

$$1 - \alpha = \int_{D_{n,\alpha}} \pi(\theta|X_1, \cdots, X_n) d\theta.$$

Often we will choose the credible interval such that it is shortest. It turns out that (actually it is not hard to prove) such a credible interval is related to the upper level set of the posterior. Given a level $\lambda$, we define

$$L_\lambda = \{\theta \in \Theta : \pi(\theta|X_1, \cdots, X_n) \geq \lambda\}.$$

Define $V(\lambda) = \int_{L_\lambda} \pi(\theta|X_1, \cdots, X_n) d\theta$. When $\lambda$ is very large, $V(\lambda)$ is very small. When we decrease $\lambda$, $V(\lambda)$ will increase since we are including more regions. At a critical level $\lambda_\alpha$, we will have exactly

$$V(\lambda_\alpha) = 1 - \alpha.$$

We will then use the level set $L_{\lambda_\alpha}$ as the credible interval. You can show that under appropriate conditions ($\pi(\theta|X_1, \cdots, X_n)$ has not flat region and is Lipschitz), $L_{\lambda_\alpha}$ is a shortest credible interval with a credible level $1 - \alpha$.

Because the above (shortest) credible interval is very straight forward, when people are talking about credible intervals, they are often referring to this interval.

[7] https://en.wikipedia.org/wiki/Bernstein%E2%80%93von_Mises_theorem

### 1.9.3 Bayesian Testing: Bayes Factor

Bayesian hypothesis testing is also very straight forward – as you can guess, it is based on the posterior distribution. Instead of just putting priors on parameter, using our decision theoretic way, we can also *put priors on the hypotheses*.

Recall that $H_0$ and $H_a$ are the null and alternative hypotheses. Let $\pi(H_0)$ and $\pi(H_a) = 1-\pi(H_0)$ denote the prior distribution on the two hypotheses and let $\pi(H_0|X_1, \cdots, X_n)$ and $\pi(H_a|X_1, \cdots, X_n)$ be the posterior distribution given the data.

The testing procedure is very simple – we reject $H_0$ if

$$\frac{\pi(H_0|X_1, \cdots, X_n)}{\pi(H_1|X_1, \cdots, X_n)} < 1 \Leftrightarrow \frac{p(X_1, \cdots, X_n|H_0)}{p(X_1, \cdots, X_n|H_a)} \cdot \frac{\pi(H_0)}{\pi(H_a)} < 1. \tag{1.2}$$

Namely, we reject $H_0$ if our posterior belief about $H_0$ is less than that of $H_a$.

In many scenarios, the hypotheses will not directly give us a probability model related to the data. In a parametric model, they often put some constraints on the parameter. Thus, we can then rewrite the posterior $p(X_1, \cdots, X_n|H_0)$ as

$$p(X_1, \cdots, X_n|H_0) = \int p(X_1, \cdots, X_n, \theta|H_0)d\theta = \int p(X_1, \cdots, X_n|\theta, H_0)\pi(\theta|H_0)d\theta.$$

Using the above equality, we define the **Bayes factor** as

$$\mathsf{BF}(X_1, \cdots, X_n) = \frac{p(X_1, \cdots, X_n|H_0)}{p(X_1, \cdots, X_n|H_a)} = \frac{\int p(X_1, \cdots, X_n|\theta, H_0)\pi(\theta|H_0)d\theta}{\int p(X_1, \cdots, X_n|\theta, H_a)\pi(\theta|H_a)d\theta}. \tag{1.3}$$

Using the Bayes factor, we can rewrite the decision rule as: we reject $H_0$ if

$$\mathsf{BF}(X_1, \cdots, X_n) \cdot \underbrace{\frac{\pi(H_0)}{\pi(H_a)}}_{\text{odds}} < 1.$$

Namely, if the Bayes factor is less than the inverse of the odds of $H_0$, we reject $H_0$.

The Bayes factor can be viewed as a *Bayesian version of p-value* – the smaller, the less favor for $H_0$. To see how the Bayes factor is like the p-value, note that the Frequentist way of rejecting $H_0$ is if the p-value is less than a pre-specified significance level $\alpha$. The Bayesian's threshold is given by the odds of $H_0$ from the prior distribution. If the inverse of the Bayes factor is greater than the odds of $H_0$, we reject $H_0$. In Bayesian, the threshold in testing (significance level) has a simple interpretation – the odds of our prior belief about the null hypothesis.

**Example: Binomial Sampling.** Now we come back to the Binomial sampling example where we have an observation $Y \sim \mathsf{Bin}(N, \theta)$ where $N$ is known and the parameter of interest is $\theta$:

$$P(Y = y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}.$$

We use a Beta distribution with parameters $(\alpha, \beta)$ as our prior distribution for $\theta$. Now we consider testing

$$H_0 : \theta = 0.5, \qquad H_a : \theta \neq 0.5.$$

Our goal is to compute the Bayes factor. First, we compute the numerator:

$$p(Y|H_0) = \int_0^1 p(Y|\theta, H_0)\pi(\theta|H_0)d\theta = \int_0^1 p(Y|\theta)\delta(\theta = 0.5)d\theta = p(Y|\theta = 0.5) = \binom{N}{Y}0.5^Y 0.5^{N-Y} = \binom{N}{Y}0.5^N.$$

The denominator of the Bayes factor will be

$$\begin{aligned} p(Y|H_a) &= \int_0^1 p(Y|\theta, H_a)\pi(\theta|H_a)d\theta \\ &= \int_0^1 p(Y|\theta, H_a)\pi(\theta)d\theta \qquad \text{(Note that } \theta = 0.5 \text{ is just a single point so it does not affect the integral)} \\ &= \int_0^1 \binom{N}{Y}\theta^Y\theta^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &= \binom{N}{Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(Y+a)\Gamma(N-Y+b)}{\Gamma(N+a+b)}. \end{aligned}$$

Therefore, the Bayes factor is

$$\mathsf{BF}(Y) = \frac{0.5^N}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(Y+a)\Gamma(N-Y+b)}{\Gamma(N+a+b)}}.$$

As you can see, the computation of Bayes factor does not involve the prior on $H_0$. However, it does involve the prior of the parameter $\theta$ (and this is why the value of it depends on the hyperparameters $\alpha, \beta$).

There is an interesting relation between the Bayes factor and the likelihood ratio test. In the above Binomial example, if we use the likelihood ratio test, we obtain a test statistic:

$$\mathsf{LRT} = \frac{p(Y|\theta = 0.5)}{p(Y|\theta = \hat{\theta}_{MLE})} = \frac{\binom{N}{Y}0.5^N}{\sup_{\theta\in[0,1]}\binom{N}{Y}\theta^Y(1-\theta)^{N-Y}} = \frac{0.5^N}{(Y/N)^Y(1-Y/N)^{N-Y}}.$$

The numerator is the same while the denominator is slightly different – the LRT only uses the maximum likelihood value whereas the Bayes factor uses the average value within $H_a$! An interesting fact: if the alternative hypothesis is a simple hypothesis $H_a : \theta = \theta_a$ for a fixed quantity $\theta_a$, then the Bayes factor and the LRT statistic coincides.

### 1.9.4 Bayesian nonparametric

Bayesian approach can also be applied to nonparametric estimation. Nonparametric estimation refers to the case where we do not assume the parameter of interest is a vector or number. One common problem is density estimation. If our goal is to estimate the underlying PDF or CDF, we can do it without assuming the data being from a parametric family such as a Gaussian. For instance, histogram can be used as a density estimator without assuming the data is from a Gaussian or any other parametric family (there are many more advanced techniques such as the kernel density estimator, orthonormal basis approach, wavelet approach, ...etc).

The challenge of Bayesian inference in nonparametric problem is that we need to put a prior on 'function space' (a collection of functions). Often a function space does not admit a density function. However, there are some tricks that we can put a prior on function space.

- *Dirichlet process.* The Dirichlet process[8] is a stochastic process that generates random probability distributions. It also has several other interesting names such as the Chinese buffet process and stick-breaking process. The appealing feature of Dirichlet process is that it is constructed from an algorithmic scheme so there is a simple way to sampling points from a random distribution generated from this process.

[8] https://en.wikipedia.org/wiki/Dirichlet_process

- *Sequential mixture model.* Another way to assign priors on distributions is to use a sequential mixture model. For instance, a Gaussian mixture model with parameters drawn from some distribution. When we allow the number of mixture to increase with sample size, we can approximate many distributions in the function space. Note that when we allow the number of mixture to increase, this implies that our prior distribution is changing with respect to the sample size.

- *Basis approach.* A smooth density function with a compact support may be written as $p_0(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x)$, where $\{\phi_k(x) : k = 1, 2, \cdots\}$ is a basis and $\{\theta_k : k = 1, 2, \cdots\}$ are the coefficients. For instance, one may choose the cosine basis or spline basis. Putting priors on $\theta_k$ leads to a prior distribution over functions. Note that in practice we often had to truncate the basis at certain level , i.e., we use only $\{\phi_k(x) : k = 1, 2, \cdots, N\}$ for some $N = N$. Often we allow $N = N_n$ to increase with respect to the sample size so that our prior covers most part of the function space.

- *$\epsilon$-cover/bracketing.* There is another approach to assign priors over function space using the covering/bracketing of function space. An $\epsilon$-cover is a collection of functions in a function space such that every function inside the space has a distance at most $\epsilon$ to the nearest element inside the cover (bracketing is a generalization of this concept). There are many $\epsilon$-covers but we often use those with the minimal number of elements. Now consider a sequence of $\epsilon$-cover with $\epsilon = \epsilon_n \to 0$ as $n \to \infty$. Let $\Pi_n$ be a uniform distribution over each element of $\epsilon_n$-cover. Then the mixture distribution $\Pi = \sum_{n=1}^{\infty} \beta_n \Pi_n$, with an appropriate choice of $\{\beta_n\}$ forms a prior distribution that can well-approximate almost every function in the function space (when the function space is sufficiently smooth). This idea is a powerful tool in constructing a prior with amazing theoretical properties[9] although it is often hard to numerically compute this prior.

## 1.10 Prediction

Prediction is about the question that given observations $X_1, \cdots, X_n$, what will be the possible range of the next observation $X_{n+1}$? Using probability model, we know that the next observation $X_{n+1}$ will have a distribution. So the prediction problem is about finding the distribution $p(x_{n+1}|X_1, \cdots, X_n)$. Here we will show how the Frequentist prediction method and the Bayesian prediction method are different in a parametric model.

The Frequentist approach is very simple – given that we have already assigned a parametric model $p(x; \theta)$ for the observations, we should just use it as our predictive distribution. To use this model, we need to choose the parameter $\theta$. A simple choice is the MLE $\hat{\theta}_{MLE}$. Thus, we will predict the distribution of $X_{n+1}$ as

$$p(x_{n+1}|X_1, \cdots, X_n) = p(x_{n+1}; \theta = \hat{\theta}_{MLE}).$$

The Bayesian approach again relies on the posterior. The predictive distribution $p(x_{n+1}|X_1, \cdots, X_n)$ can be written as

$$\begin{aligned} p(x_{n+1}|X_1, \cdots, X_n) &= \int p(x_{n+1}, \theta|X_1, \cdots, X_n) d\theta \\ &= \int p(x_{n+1}|\theta, X_1, \cdots, X_n) \pi(\theta|X_1, \cdots, X_n) d\theta \\ &= \int p(x_{n+1}|\theta) \underbrace{\pi(\theta|X_1, \cdots, X_n)}_{\text{posterior distribution}} d\theta. \end{aligned}$$

[9] see, e.g., the famous paper Ghosal et al. (2000) "Convergence rates of posterior distributions": https://projecteuclid.org/euclid.aos/1016218228

Thus, the predictive distribution is the *averaged* distribution of $p(x|\theta)$ where we average $\theta$ over the posterior distribution.

Here, as you can see, the two paradigms make prediction using different principles – the Frequetists use only the *most likely* model to make predictions whereas the Bayesians use the *averaged* model over the posterior distribution to make predictions.

## 1.11 Comments: Frequentist versus Bayesian

Both Frequentist and Bayesian approaches are self-consistent. They start with probability models and design their own procedure for estimation, interval inference, and hypothesis test. In practice, it is hard to really say if any method truly describes the reality because we do not even know if our data are indeed generated from a random process. Some people believe that our data should be viewed as realizations from a complex dynamic system and they can still construct estimators and derive consistency without introducing any probability (often they would use the Ergodic theory[10]). Thus, we cannot say if Frequentist or Bayesian approach is the right approach to analyze data – they are just principles that allows us to analyze data in a well-established way. In what follows, I briefly comments on the criticisms and defends of the two paradigms.

Many people support Frequentist paradigm because it is more *objective* – we do not introduce any subjective belief (prior) on the parameter of interest. Moreover, the way Frequentist views the parameter of interest, an unknown but fixed quantity, fits into the what most scientists think about parameters in model – these parameters are some fixed numbers but we just do not know it. Moreover, the Frequentist's view of the probability – long term frequency – is very intuitive for most people.

The Bayesian paradigm is very clean. From estimation, interval inference, to hypothesis test, every inference just depends on one single quantity – the posterior distribution. And we only have one single guiding principle – decision theoretic rule – we choose the one that has a higher posterior distribution. On the other hand, in the case of Frequentist approach, estimation requires a principle (such as the MLE principle), confidence interval relies on another principle (coverage), and the hypothesis test introduces another principle (p-value and significance level). So Bayesian is a very elegant way to analyzing the data.

A major criticism of Bayesian is the concept of *subjectivity* – the prior distribution. Many people prefer Frequentist approaches over Bayesian approaches because they think that scientific studies should be objective. However, this argument is actually not valid if we really think deep about Frequentist approach – when we use a probability model to describe the data, we are already making a subjective choice of how we model the data! Why not use a dynamic system approach? why not use a Bayesian? The choice of Frequentist approach itself is a subjective decision made by scientists. Moreover, the choice of estimator, the choice of confidence level, and the choice of significance level along with the testing procedures, are all subjective decisions. None of them are truly objective. One attractive feature of Bayesian paradigm is that Bayesians not only accept the fact that we are making many subjective choices in analyzing data but also they have a well-defined mathematical framework – the probability model – to describe how these subjectivity are propagating throughout the analysis.

[10] https://en.wikipedia.org/wiki/Ergodic_theory