MATH 310: Mathematical Statistics (brief notes)

Truong-Son Van

Contents

Disclaimer

This is class notes for Mathematical Statistics at Fublbright University Vietnam. I claim no mathematiacal originality in this work as it is mostly taken from the reference books. The only original contribution of mine are typos and errors.

PART 1: Background

Readings:

- Chapters 1-5 of Wasserman
- Chapters 1-5 of Rice

Chapter 1

Probability

> "If we have an atom that is in an excited state and so is going to emit a photon, we cannot say when it will emit the photon. It has a certain amplitude to emit the photon at any time, and we can predict only a probability for emission; we cannot predict the future exactly."
>
> — Richard Feynman

1.1 Review

1.1.1 Probability Space

Definition 1.1 (Sigma-algebra)**.** Let Ω be a set. A set $\Sigma \subseteq \mathcal{P}(\Omega)$ of subsets of Ω is called a σ-algebra of Ω if

1. $\Omega \in \Sigma$
2. $F \in \Sigma \implies F^C \in \Sigma$
3. If $F_n \in \Sigma$ for all $n \in \mathbb{N}$, then
$$\bigcup_n F_n \in \Sigma.$$

It is extremely convenient to deal with things called open sets. The definition of those are a bit out of the scope of this class. However, in the case of the real line $\mathbb{R}$, open sets are defined to be made of by finite intersections and arbitrary unions of open intervals (a, b). For example, $(0, 1) \cup (2, 3)$ is an open set.

Interestingly, $\mathbb{R}$ and $\emptyset$ are called clopen sets (here's a funny YouTube video about clopen sets: https://www.youtube.com/watch?v=SyD4p8_y8Kw)

A Borel σ-algebra is the smallest σ-algebra that contains all the open sets. We denote the Borel σ-algebra of a set Ω to be $\mathcal{B}(\Omega)$. This is a rather abstract definition. There is no clear way to construct a sigma algebra from a collection of sets. However, the construction is not important as the reassurance that this object does exist to give us nice domains to work with when we define a probability measure (see below definition).

Exercise 1.1. (*Challenging– not required but good for the brain*) It turns out that if Ω is a discrete set, it is typical to have the set of open sets contain every set of singletons, i.e., the set $\{a\}$ is open for every $a \in \Omega$. Take this as an assumption, show that for any discrete set Ω, $\mathcal{B}(\Omega) = \mathcal{P}(\Omega)$.

What open sets really are is not important for now. The important thing is that for $\mathbb{R}^n$ open sets are made of open intervals/ open boxes. Your typical intuitions still work.

Philosophically, the σ-algebra represents the details of information we could have access to. There are certain events that are building blocks of knowledge and that we don't have access to finer details.

Think about the σ-algebra as a consistent model of what can be known (observed). For example, you can never know what's going on in the houses on the street unless you have been to them. But somehow, together, you are still able to piece all the information you have about the houses to make sense of the world. This is related to the problem of information. How much information is enough to be useful in certain situation?!

To have a consistent system is not the same as to know everything. The system you see/invent can never be exhaustively true, but you can still say something about the reality if you can have a system that is consistent with what you observe. This is why we do sampling!!

When you have a consistent model, you now want to encode the model in such a way that it helps you with describing/predict the reality you see. A way to do that with no full knowledge of anything is to assign the certain number to measure the chance for something to happen at a given time. This encoding needs to happen on the model you constructed. This leads to the following definition of probability space.

Definition 1.2 (Probability Space)**.** A *Probability Space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set called *sample space*, $\mathcal{F}$ is a σ-algebra on Ω, $\mathbb{P} : \mathcal{F} \to [0, 1]$, called a *Probability Measure*, is a function that satisfies the following:

1. $\mathbb{P}(\Omega) = 1$,
2. If F is a disjoint union of $\{F_n\}_{n=1}^{\infty}$, then

$$\mathbb{P}(F) = \sum_{n=1}^{\infty} \mathbb{P}(F_n).$$

Each element $\omega \in \Omega$ is called an *outcome* and each subset $A \in \mathcal{F}$ is called an *event*.

Definition 1.3 (Independent Events)**.** Let $A, B \in \mathcal{F}$ be events. We say that A and B are independent if

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

Definition 1.4 (Conditional Probability)**.** Let $A, B \in \mathcal{F}$ be events such that $\mathbb{P}(B) > 0$. Then, the conditional probability of A given B is

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

Theorem 1.1 (Bayes's Theorem)**.** *Let $A, B \in \mathcal{F}$ be events such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then,*

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

In modern statistics, there are names for the above terms:

1. $\mathbb{P}(A|B)$ is called *Posterior Probability*,
2. $\mathbb{P}(B|A)$ is called *Likelihood*,
3. $\mathbb{P}(A)$ is called *Prior Probability*,
4. $\mathbb{P}(B)$ is called *Evidence*.

The theorem is often expressed in words as:

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Prior Probability}}{\text{Evidence}}$$

It is a good idea to ponder why those mathematical terms have those names.

1.1.2 Random Variables

The notion of probability alone isn't sufficient for us to describe ideas about the world. We need to have a notion of objects that associated with probabilities. This brings about the idea of *random variable*.

Definition 1.5 (Random Variable)**.** Let $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ be a probability space and $(S, \mathcal{B}(S))$ a σ-algebra. A random variable is a (Borel measurable) function from $\Omega \to S$.

- S is called the *state space* of X.

In this course, we will restrict our attentions to two types of random variables: discrete and continuous.

Definition 1.6 (Discrete RV)**.** $X : \Omega \to S$ is called a discrete RV if S is a countable set.

- A *probability function* or *probability mass function* for X is a function $f_X : S \to [0, 1]$ defined by

$$f_X(x) = \mathbb{P}(X = x)\,.$$

In contrast to the simplicity of discrete RV. Continuous RVs are a little bit messier to describe. This is because of the lack of background in measure theory so we can talk about this concept in a more precise way.

Definition 1.7 (Continuous RV)**.** A *continuous random variable* is a measurable function $X : \Omega \to S$ is continuous if it satisfies the following conditions:

1. $S = \mathbb{R}^n$ for some $n \in \mathbb{N}$.
2. There exists an (integrable) function f_X such that $f_X(x) \geq 0$ for all x, $\int_{\mathbb{R}^n} f_X(x)dx = 1$ and for every open cube $C \subseteq \mathbb{R}^n$,
$$\mathbb{P}(X \in C) = \int_C f_X(x)dV.$$

The function f_X is called the *probability density function (PDF)*.

If two RVs X and Y share the same probability function, we say that they have the same distribution and denote them by

$$X \overset{d}{=} Y.$$

In this case we also say that X and Y are **equal in distribution**.

Remark. To make the presentation more compact and clean, notationally, we will write

$$\int f(x)dx$$

to mean both integral (for continuous RV) and summation (for discrete RV).

There are more general concepts of continuous RV where we don't need to require S to be a Euclidean space as in the above definition. However, such concepts require the readers to be familiar with advanced subjects like Topology and Measure Theory. It is particularly important to know these two subjects in order to thoroughly understand Stochastic Processes.

Exercise 1.2. Create a random variable that represents the results of n coin flips.

For real-valued RV $X : \Omega \to \mathbb{R}$ we have the concept of cumulative distribution function.

Definition 1.8 (Cumulative Distribution Function)**.** Given a RV $X : \Omega \to \mathbb{R}$. The *cumulative distribution function of* X or CDF, is a function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x)\,.$$

Notationally, we use the notation $X \sim F$ to mean RV X with distribution F.

Exercise 1.3. Given a real-valued continuous RV $X : \Omega \to \mathbb{R}$, prove that if f_X is continuous then

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

is differentiable for every x and $f_X(x) = F_X'(x)$.

Exercise 1.4. Let X be an RV with CDF F and Y with CDF G. Suppose $F(x) = G(x)$ for all x. Show that for every set A that is a countable union of open intervals,

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)\,.$$

Exercise 1.5. Let $X : \Omega \to \mathbb{R}$ be an RV and F_X be its CDF. Prove the following:

1. F is non-decreasing: if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
2. F is normalized:
$$\lim_{x \to -\infty} F(x) = 0\,,$$
and
$$\lim_{x \to \infty} F(x) = 1\,.$$
3. F is right-continuous:
$$F(x) = F(x+) = \lim_{y \searrow x} F(y)\,.$$

1.1.3 Joint distribution of RVs

Let $X : \Omega \to S$ and $Y : \Omega \to S$ be RVs. We denote

$$\mathbb{P}(X \in A; Y \in B) = \mathbb{P}(\{X \in A\} \cap \{Y \in B\})\,.$$

For discrete RVs, the joint probability function of X and Y has the following meaning

$$f_{XY}(x, y) = \mathbb{P}(X = x; Y = y)$$

For continuous RVs, the situations are more complicated as we can't make sense of $\mathbb{P}(X = x; Y = y)$ (this is always 0 in most situation and in some other situation, one can't even talk about it– this is a topic of more advanced course in measure theory). However, we can have

$$\mathbb{P}(X \in A; Y \in B) = \int_{X \in A} \int_{Y \in B} f_{XY}(x, y)\, dxdy\,.$$

Another way to look at the above is the following. We can even consider $X : \Omega \to S^n$, where $n \geq 2$. Instead of thinking about this as one RV, we can think about this as a vector of RVs:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

where $X_1, \ldots X_n : \Omega \to S$ are RVs. Because of this, we have can write the density function as

$$f_X(x) = f_{X_1 X_2 \ldots X_n}(x)$$

Some people call X a random vector.

$f_{X_1 \ldots X_n}$ is called the joint probability distribution.

Exercise 1.6. True or false:

1. $f_{XY}(x,y) = f_X(x) + f_Y(y)$
2. $f_{XY}(x,y) = f_X(x) f_Y(y)$

Definition 1.9 (Marginal density)**.** The marginal density of X_i is

$$f_{X_i}(x_i) = \int f_{X_1 \dots X_n}(x_1, \dots, x_n)\, dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n$$

(integrate coordinate except the i-th coordinate.

Exercise 1.7. Can you construct f_{XY} if you know f_X and f_Y?

1.1.4 Some important random variables

1. Point mass distribution (Dirac delta): Given a discrete probability $X : \Omega \to S$. X has a point mass distribution at $a \in S$ if
$$\mathbb{P}(X = a) = 1.$$
We call X a point mass RV and write $X \sim \delta_a$.

   Question. Suppose $S = \mathbb{N}$. Write down F_X for the point mass RV X.

2. Discrete uniform distribution: $f_X(k) = \frac{1}{n}$, $k \in \{1, \dots, n\}$.

3. Bernoulli distribution: let $X : \Omega \to \{0,1\}$ be RV that represents a binary coin flip. Suppose $\mathbb{P}(X = 1) = p$ for some $p \in [0,1]$. Then X has a Bernoulli distribution, written as $X \sim \text{Bernoulli}(p)$. The probability function is
$$f_X(x) = p^x (1-p)^{1-x}.$$
We write $X \sim \text{Bernoulli}(p)$.

4. Binomial distribution: let $X : \Omega \to \mathbb{N}$ be the RV that represents the number of heads out of n independent coin flips. Then
$$f_X = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}\,, x \in \{0, 1, \dots, n\} \\ 0\,, \text{otherwise} \end{cases}$$
We write $X \sim \text{Binomial}(n,p)$.

5. Poisson distribution: $X \sim \text{Poisson}(\lambda)$.
$$f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}\,, k = 0, 1, 2 \dots$$
X is a RV that describe a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event

6. Gaussian: $X \sim N(\sigma, \mu)$.
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Exercise 1.8. Let $X_{n,p} \sim \text{Binomial}(n,p)$. Suppose that as $n \to \infty$, $p \to 0$ in such a way that $np = \lambda$ always. Let $x \in \mathbb{N}$.

1. For n very very large, what is the behaviour of
$$\frac{n!}{(n-x)!}\,.$$
(You should just get some power of n)

2. Show that
$$\lim_{n\to\infty} f_{X_{n,p}}(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

3. Interpret this result.

1.1.5 Independent random variables

Definition 1.10. Let $X : \Omega \to S$ and $Y : \Omega \to S$ be RVs. We say that X and Y are independent if, for every $A, B \in \mathcal{B}(S)$, we have
$$\mathbb{P}(X \in A; Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,,$$
and write $X \perp Y$.

So, if X and Y are independent,
$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

Definition 1.11. Let $X : \Omega \to S$ and $Y : \Omega \to S$ be RVs. Suppose that $f_Y(y) > 0$. The conditional probability mass function of X given Y is
$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}\,.$$

Exercise 1.9. Let
$$f(x, y) = \begin{cases} x + y\,, 0 \le x \le 1, 0 \le y \le 1 \\ 0\,, \text{otherwise} \end{cases}\,.$$

What is $\mathbb{P}(X < 1/4|Y = 1/3)$.

Note that the above exercise is a little bit weird and counter-intuitive. While $\mathbb{P}(X < 1/4, Y = 1/3) = 0$ (why?), $\mathbb{P}(X < 1/4|Y = 1/3) \neq 0$

A very important RV is the *multivariate Normal RV*, which obeys the following density function

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\big(-\frac{1}{2}(x - mu)^T \Sigma^{-1}(x - \mu)\big).$$

1.1.6 Transformations of RVs

Sometimes, we don't work with RV directly but certain characteristics of RVs. Those characteristics are represented by certain transformation. If the functions are nice enough, we can actually have a recipe to generate the probability density function.

Suppose $g : S^n \to \mathbb{R}$ and $Z = g(X_1, \dots, X_n)$. Let $A_z = \{(x_1, \dots, x_n) : g(x_1, \dots, x_n) \le z\}$. Then
$$F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}(g(X_1, \dots, X_n) \le z) = \int_{A_z} f_{X_1 \dots X_n}(x_1, \dots, x_n)\, dx_1 \dots dx_n,$$
and
$$f_Z(z) = F_Z'(z).$$

Exercise 1.10.

1. Let $X, Y \sim \text{Uniform}(0, 1)$ be independent RVs, i.e.,
$$f_X(x) = f_Y(y) = 1\,.$$
What is the density function for the RV $Z = X + Y$?

2. Same question but $X, Y \sim N(0,1)$.

Definition 1.12. RVs that are independent and share the same distribution are called *independent and identically distributed* RVs.

We often shorthand this by IID RVs.

1.1.7 Expectation

Definition 1.13. Let X be a RV.

1. The *expected value*, or *expectation*, or *mean*, or *first moment* of X is defined to be
$$\mathbb{E}X = \int xf(x)dx.$$
2. The *variance* of X is defined to be
$$\mathbb{E}\left(X - \mathbb{E}X\right)^2$$

We often denote μ_X to be the expectation of X, σ_X^2 ($\mathrm{Var}(X), \mathbb{V}(X)$) to be the variance of X.

The square root of the variance, σ, is called the *standard deviation*.

Theorem 1.2. *Let $X : \Omega \to S$ be a RV, $r : S \to S$ be a function and $Y = r(X)$. Then*
$$\mathbb{E}Y = \mathbb{E}(r(X)) = \int r(x)f(x)dx$$

Exercise 1.11.

1. Let $X \sim \mathrm{Uniform}(0,1)$. Compute $\mathbb{E}Y$, where
   a. $Y = e^X$.
   b. $Y = \max(X, 1-X)$
2. Let X, Y be RVs that have jointly uniform distribution on the unit square. Compute $\mathbb{E}(X^2 + Y^2)$.

Definition 1.14. Let X and Y be RVs. The *covariance* between X and Y is defined by
$$\mathrm{Cov}(X,Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right).$$

The *correlation* of X and Y is defined to be
$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Theorem 1.3.

1. Let X_i, $i = 1, \ldots, n$ be RVs and a_i's be constants. Then
$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}X_i.$$

2.
$$\mathbb{V}X_i = \mathbb{E}(X_i^2) - \mu_{X_i}^2$$

3.
$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i)$$

4.

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)\,.$$

5. Suppose further that X_i's are independent, then

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}X_i$$

and

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i)\,.$$

Definition 1.15 (Conditional Expectation)**.** Let $X, Y : \Omega \to S$, where S is either $\mathbb{N}$ or $\mathbb{R}$. The conditional expectation of X given Y is a RV $\mathbb{E}[X|Y] : \Omega \to \mathbb{R}$ that satisfies the following

$$\mathbb{E}[X|Y](y) := \mathbb{E}[X|Y = y] = \int x f_{X|Y}(x|y)\, dx.$$

If $r : S^2 \to S$ is a function, then

$$\mathbb{E}[r(X,Y)|Y = y] = \int r(x,y) f_{X|Y}(x|y)\, dx.$$

One can generalize this definition to higher dimension via the coordinate-wise conditional expectation. We will omit this definition in order to keep the presentation simple.

Theorem 1.4. *Let X and Y be Rvs. We have that*

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X.$$

1.2 Moment Generating and Characteristic Functions

Definition 1.16. Let X be a RV. 1. The *moment generating function* MGF, or *Laplace transform*, of X is $\varphi : \mathbb{R} \to \mathbb{R}$ defined by

$$\varphi_X(t) = \mathbb{E}\left(e^{tX}\right),$$

where t varies over the real numbers.

2. The *characteristic function*, or *Fourier transform* of X is $\varphi : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_X(\theta) = \mathbb{E}e^{i\theta X}.$$

Lemma 1.1.

1. Let X be a RV and $Y = aX + b$, then

$$\varphi_Y(t) = e^{bt}\varphi_X(at)$$

2.

$$\varphi_X^{(k)}(0) = \mathbb{E}(X^k)$$

3. Let X_i, $i = 1, \ldots, n$ be independent RVs and $Y = \sum_i X_i$. Then

$$\varphi_Y(t) = \prod \varphi_{X_i}(t).$$

4.

$$|\phi(\theta)| \leq 1$$

5. Denote $\overline{z}$ to be the complex conjugate of z in the complex plane.

$$\phi_{-X}(\theta) = \overline{\phi_X(\theta)}$$

6.

$$\phi_Y(\theta) = e^{ib\theta}\phi(a\theta)$$

Exercise 1.12. Prove the above lemma.

Exercise 1.13. Let $X \sim \exp(1)$, i.e,

$$f_X(x) = \begin{cases} e^{-x}\,, x \geq 0 \\ 0\,, x < 0\,. \end{cases}$$

Compute φ_X.

Recall that two RVs $X \overset{d}{=} Y$ means that $F_X(x) = F_Y(x)$. Two common ways to characterize the equality in distribution are to use the generating functions and the characteristic functions.

These ideas are not orginally from probability but from engineering/mechanics, where Laplace and Fourier transforms are understood very well since the 18th century.

Exercise 1.14. In general, differentiation is not commutative with integration, that is

$$\frac{d}{dt}\int \neq \int \frac{d}{dt}.$$

However, assuming that this is true for certain moment generating functions φ_X. Show that

$$\varphi_X^{(n)}(0) = \mathbb{E}(X^n),$$

where $f^{(n)}$ denotes the n-th derivative of f. $\mathbb{E}(X^n)$ is called the n-th moment of X and it tells you the tail behavior of f_X.

1.2.1 Moment Generating Functions

Theorem 1.5. *Let X and Y be RVs. If $\varphi_X(t) = \varphi_Y(t)$ for all t in an interval around 0, then*

$$X \overset{d}{=} Y\,.$$

The full proof of this is beyond this class (and could be a great topic for a project). However, we will prove this for finte RVs.

Proposition 1.1 (Finite RV case)**.** *Let $X, Y : \Omega \to \{1, 2, \ldots, N\}$ be RVs. If $\varphi_X(t) = \varphi_Y(t)$ in an interval around in an interval $(-\epsilon, \epsilon)$, then*

$$X \overset{d}{=} Y\,.$$

Proof. We have that

$$\varphi_X(t) = \mathbb{E}(e^{tX}) = \sum_{i=1}^{N} e^{it}\mathbb{P}(X = i)$$

and

$$\varphi_Y(t) = \mathbb{E}(e^{tX}) = \sum_{i=1}^{N} e^{it}\mathbb{P}(Y = i).$$

Therefore,

$$0 = \varphi_X(t) - \varphi_Y(t) = \sum_{i=1}^{N} (e^t)^i \left(\mathbb{P}(X = i) - \mathbb{P}(Y = i)\right)$$

for every $t \in (-\epsilon, \epsilon)$. Therefore, as the above is a polynomial,

$$\mathbb{P}(X = i) = \mathbb{P}(Y = i)$$

where $i = 1, \dots, N$. □

Note that if the above summation is infinite, then we cannot conclude that X and Y has the same distribution as easily as we just did. More work has to be done to show this.

A note of caution: the assumption that $\varphi_X = \varphi_Y$ in an interval around 0 is crucial in general.

An interesting observation arises: for analytic functions we have the Taylor series

$$f(x) = \sum_{i=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n.$$

Exercise 1.14 tells you that the n-th derivative at 0 of a moment generating function would be the n-th moment of the RV.

Question: Is knowing the moments of X enough to determine its probability distribution?

The answer is NO. One can take a look at the discussion about this problem here: https://mathoverflow.net/questions/3525/when-are-probability-distributions-completely-determined-by-their-moments.

However, things are nice for finite RVs.

Proposition 1.2. *Let $X, Y : \Omega \to \{1, 2, \dots, N\}$ be RVs. Suppose that*

$$\mathbb{E}(X^n) = \mathbb{E}(Y^n) < \infty$$

for every $n \in \mathbb{N}$. Then

$$X \overset{d}{=} Y.$$

Proof. Consider

$$\varphi_X(t) = \mathbb{E}(e^{Xt}) = \sum_{i=1}^{N} e^{it} \mathbb{P}(X = i).$$

This is a finite sums of analytic functions and is, therefore, analytic. Thus, φ_X can be expanded into Taylor series, i.e.,

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{\varphi_X^{(n)}(0)}{n!} t^n = \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)}{n!} t^n.$$

This means that the moments of X determines its moment generating function (which may not be true in general).

A similar argument can be made for φ_Y and as the coefficents of the Taylor series are the same (being the moments of X and Y), we conclude that

$$\varphi_X = \varphi_Y.$$

Therefore, by Theorem 1.5,

$$X \overset{d}{=} Y,$$

as desired. □

1.2.2 Characteristic Functions

Similar idea with the moment generating functions, but characteristic functions are easier to work with and we don't have to work with special case of finite RVs.

Theorem 1.6. *Let X and Y be RVs. If $\phi_X(t) = \phi_Y(t)$ for all t in an interval around 0, then*

$$X \stackrel{d}{=} Y\,.$$

In order to prove this theorem, we need the following important result, called inversion formula of the characteristic functions.

Theorem 1.7 (Inversion Formula)**.** *Let $X : \Omega \to S$ be a RV (either continuous or discrete) and ϕ_X be its characteristic function. Then*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \phi_X(\theta)\, d\theta = \mathbb{P}(a < X < b) + \frac{1}{2}\left(\mathbb{P}(X = a) + \mathbb{P}(X = b)\right).$$

Proof. We have

$$\begin{aligned}\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \phi_X(\theta)\, d\theta &= \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \int_{\mathbb{R}} e^{i\theta x} f_X(x)\, d\theta dx \\ &= \int_{\mathbb{R}} \frac{1}{\pi} \int_{-T}^{T} \frac{e^{i\theta(x-a)} - e^{i\theta(x-b)}}{2i\theta} f_X(x)\, d\theta dx\end{aligned}$$

Note that since $\cos(t)/t$ is odd and $\sin(t)/t$ is even, and that $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, we have

$$\frac{1}{2} \int_{-T}^{T} \frac{e^{i\theta c}}{i\theta} = \int_{0}^{T} \frac{\sin(\theta c)}{\theta}\, d\theta.$$

Therefore,

$$\frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \phi_X(\theta)\, d\theta = \frac{1}{\pi} \int_{\mathbb{R}} \int_{0}^{T} \left(\frac{\sin((x-a)\theta)}{\theta} - \frac{\sin((x-b)\theta)}{\theta} \right) f_X(x)\, d\theta dx$$

Taking the limit $T \to \infty$ and using the fact that

$$\lim_{T\to\infty} \int_{0}^{T} \frac{\sin((x-a)\theta)}{\theta} d\theta = \begin{cases} \frac{-\pi}{2}\,, & x < a\,, \\ \frac{\pi}{2}\,, & x > a\,, \\ 0\,, & x = a\,. \end{cases}$$

Therefore,

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{2\pi} \int_{\mathbb{R}} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \phi_X(\theta)\, d\theta dx &= \left(\int_{(a,\infty)} f_X(x)\, dx - \int_{(-\infty,a]} f_X(x)\, dx \right) \\ &\quad - \left(\int_{(b,\infty)} f_X(x)\, dx - \int_{(-\infty,b]} f_X(x)\, dx \right) \\ &= (\mathbb{P}(X > a) - \mathbb{P}(X \le a)) - (\mathbb{P}(X > b) - \mathbb{P}(X \le b)) \\ &= \mathbb{P}(a < X < b) + \frac{1}{2}\left(\mathbb{P}(X = a) + \mathbb{P}(X = b)\right),\end{aligned}$$

as desired. □

Exercise 1.15. Verify that

$$\lim_{T\to\infty}\int_0^\infty \frac{\sin(x)}{x}\,dx = \frac{\pi}{2}.$$

If you can't, watch this: https://www.youtube.com/watch?v=Bq5TB6cZNng.

Another way is to use contour integral from complex analysis.

Exercise 1.16. Let $X_1, \dots, X_n \sim \text{Uniform}(0,1)$ be independent and $Y_n = \max\{X_1, \dots, X_n\}$. Find $\mathbb{E}(Y_n)$.

Exercise 1.17. Let $X : \Omega \to (0,\infty)$ be continuous positive RV. Suppose $\mathbb{E}(X)$ exist. Show $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)\,dx$. (Hint: Fubini. This is called the layer cake theorem).

Exercise 1.18. The exponential distribution with parameter λ (denoted by $\exp(\lambda)$) is used to model waiting time (see https://en.wikipedia.org/wiki/Exponential_distribution). The probability density function of the exponential distribution is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

1. Find the moment-generating function of $X \sim \exp(\lambda)$.
2. Use moment-generating function to show that if X is exponential distributed, then so is cX.

Exercise 1.19. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim (\mu_2, \sigma_2^2)$ be independent. Use the moment generating function to show that $Z = c_1X + c_2Y$ is again a normal distribution. What are $\mathbb{E}(Z)$ and $\mathbb{V}(Z)$?

Exercise 1.20. Find the moment-generating function of a Bernoulli RV, and use it to find the mean, variance, and third moment.

Exercise 1.21. Let $X : \Omega \to S$ be a RV and $S = \mathbb{N}$. The *probability generating function* of X is defined to be

$$G(s) = \sum_{k=1}^{\infty} s^k \mathbb{P}(X = k).$$

a. Show that

$$\mathbb{P}(X = k) = \frac{1}{k!}\frac{d^k}{ds^k}G(s)|_{s=0}$$

b. Show that

$$\frac{dG}{ds}|_{s=1} = \mathbb{E}(X)$$

and

$$\frac{d^2G}{ds^2}|_{s=1} = \mathbb{E}[X(X-1)].$$

c. Express the probability-generating function in terms of moment-generating function.

d. Find the probability-generating function of the Poisson distribution.

1.3 Inequalities

1.3.1 Typical tail bound inequalities

Theorem 1.8 (Markov's Inequality)**.** *Let X be a non-negative RV and $\mathbb{E}(X)$ exists. Then, for each $k > 0$,*

$$\mathbb{P}(X > k) \leq \frac{\mathbb{E}(X)}{k}.$$

Theorem 1.9 (Chebyshev's Inequality)**.** *Let X be a RV such with expected value μ and standard variation σ. Then for each $k > 0$,*

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

1.3.2 Exponential concentration inequalities

Theorem 1.10 (Mill's inequality)**.** *Let $Z \sim N(0,1)$. Then, for each $t > 0$,*

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

Theorem 1.11 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent RVs such that $\mathbb{E}(X_i) = 0$, $a_i \leq Y_i \leq b_i$. For each $\epsilon > 0$ and $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^n e^{t^2(b_i-a_i)^2/8}.$$

Exercise 1.22. Let $X_1, \dots, X_n$ be independent RVs such that $\mathbb{E}(X_i) = 0$, $a_i \leq Y_i \leq b_i$. Show that for each $\epsilon > 0$ and $t > 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

1.3.3 Inequalities for expectations

Theorem 1.12 (Cauchy-Schwartz inequality)**.** *Let X, Y be RVs with finite variances. Then,*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Theorem 1.13 (Jensen's inequality)**.** *Suppose $g : \mathbb{R} \to \mathbb{R}$ is a convex function. Then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X).$$

1.4 Law of Large Numbers

Theorem 1.14. *Let X_i, $i \in \mathbb{N}$ be independent RVs such that $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$. Then, for each $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \epsilon\right) = 0\,.$$

The above kind of convergence is sometimes called *convergence in probability*. There are other modes of convergence such as convergence almost surely and uniform convergence.

1.5 Central Limit Theorem

Definition 1.17 (Convergence in distribution)**.** Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of RVs with CDF F_i. Let X be a RV with CDF F. We say that X_n convergence to X in distribution if

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

at every point at which F is continuous.

Theorem 1.15 (Continuity theorems)**.** *Let X_i, $i \in \mathbb{N}$ RVs with CDF F_i and X a RV with CDF $\bar{F}$. Suppose that either*

1. *$\varphi_{X_n}(s)$ converges to $\varphi_X(s)$ for all s in some open interval around 0.*

2. $\lim_{n\to\infty} \phi_{X_n}(s) = \phi_X(s)$ *for every* $s \in \mathbb{R}$.

Then $X_n \to X$ *in distribution.*

Theorem 1.16 (Central limit theorem)**.** *Let* $\{X_i\}_{i\in\mathbb{N}}$ *be a sequence of IID RVs with mean* 0 *and variance* $\sigma^2 < \infty$*. Define*

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sigma\sqrt{n}}.$$

Then Z_n *converges to* $Z \sim N(0,1)$ *in distribution.*

There are a few ways to go about proving this theorem. Two most common ways employ the MGF and the characteristic function. Both methods rely on the one crucial idea of using the Taylor expansion, which we will see shortly. We present the proof using MGF (adopted from Rice's book) and leave it to the reader the proof using characteristic function.

Proof. For each $n \in \mathbb{N}$, we have that

$$\varphi_{Z_n}(t) = \left(\varphi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

Observe first that $\varphi'_{X_1}(0) = \mathbb{E}X_1 = 0$ and $\varphi''_{X_1}(0) = \mathbb{E}X_1^2 = \sigma^2$. So, performing Taylor expansion for φ_{X_1}, we get

$$\begin{aligned}
\varphi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) &= 1 + \varphi'_{X_1}(0)\left(\frac{t}{\sigma\sqrt{n}}\right) + \frac{1}{2}\varphi''_{X_1}(0)\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \epsilon_n \\
&= 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \epsilon_n \\
&= 1 + \frac{1}{2}\left(\frac{t^2}{n}\right) + \epsilon_n\,.
\end{aligned}$$

where $\epsilon_n/(t^2/(n\sigma^2)) \to 0$ as $n \to \infty$. It can then be shown that

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = \lim_{n\to\infty}\left(1 + \frac{1}{2}\left(\frac{t^2}{n}\right) + \epsilon_n\right) = e^{t^2/2} = \varphi_Z.$$

Combine this with Theorem 1.15, we arrive at our result. □

Exercise 1.23. Prove the central limit theorem using the characteristic function.

PART 2: Inference

Readings:

- Rice Chapters 7, 8
- Wasserman Chapters 6, 7
- Casella-Berger Chapter 7

Statistical inference, often rebranded as learning in computer science, is the process of figuring out certain information of a distribution function F given sample $X_1, \dots, X_n \sim F$.

Typically, we don't know which distribution function our sample comes from. However, sometimes, with some background theory (or simply just to make life easier), we may assume that the data come from certain family of distributions so that we can narrow our search. This gives rise to the following definitions.

Definition 1.18. A *statistical model* $\mathcal{F}$ is a set of distributions (or densities).

A *parametric model* is a set set $\mathcal{F}$ that can be parametrized by a finite number of parameters.

A *non-parametric model* is a statistical model that is not parametric.

Example 1.1.

1. The set of Gaussians is a two parameter models:
$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}.$$
2. The set of Bernoulli distributions is a set of one parameter model:
$$\mathcal{F} = \{ \mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p, 0 \leq p \leq 1 \}.$$
3. Generally, a parametric model has the following form
$$\mathcal{F} = \{ f(x; \theta) : \theta \in \Theta \},$$
where Θ is some parameter space.

Chapter 2

Sampling

Sampling is the act of gathering data from a certain population, in order to make prediction about some property of the population of interest. Each time one goes out to the field to sample, one gets different answers for the same set of quantities of interest, making those answers random variables.

- For finite population, there are two techniques called *sampling with replacement* and *sampling without replacement*.
- Sampling without replacement is sometimes called *simple random sampling* and one needs to be careful with it. However, if the population size is very very large compared to the sample size (a very subjective judgement), it is common in practice to treat the sampling data as I.I.D. RVs.

2.1 Simple Random Sample

We will not discuss Simple Random Sampling in this class. Interested readers can consult Rice, Chapter 7.3.

2.2 Standard Random Sample

Definition 2.1 (Standard Random Sample)**.** The random variables X_i, $i = 1, \dots, n$ are called *standard random sample of size n from population $f(x)$* if X_i's are I.I.D. RVs from the same probability density function f.

There is a few nuisances regarding general practice in statistics and this definition.

1. Definition 2.1 is either for *infinite* population or finite population with *sampling with replacement*.
2. For finite population of size n, sample data X_i from *sampling without replacement* can never be independent as $\mathbb{P}(X_2 = y|X_1 = y) = 0$ and $\mathbb{P}(X_2 = y|X_1 = x) = 1/(n-1)$.

In this course, when we talk about sampling, we will understand it as in Definition 2.1.

Chapter 3

Parametric Inference (Parameter Estimation)

Notation. Given a parametric model $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$, we denote

$$\mathbb{P}(X \in A) = \int_A f(x;\theta)\, dx$$

and

$$\mathbb{E}_\theta(r(X)) = \int r(x) f(x;\theta)\, dx\,.$$

3.1 Point Estimation

(Casella - Berger Chapter 7, Wasserman Chapter 6.1)

Definition 3.1. Let $\{X_i\}$, $i = 1, \dots, n$ be a sample. A *point estimator* of $\{X_i\}$ is a function $g(X_1, \dots, X_n)$.

The purpose of the *point estimator* is to provide the "best guess" of certain quantity of interest. Those quantities could be a parameter in a parametric model, a CDF, PDF,...

Typically, the quantity of interest is denoted by θ, the point estimator is denoted by $\hat{\theta}$ or $\hat{\theta}_n$. So, combined with the above definition,

$$\hat{\theta}_n = g(X_1, \dots, X_n).$$

Note that, $\hat{\theta}_n$ is still a random variable as this is a function of your sample data, which are RVs themselves.

Of course, we know that there are cases when samples suffer from biases. A way to measure biases is to compare the expected value of $\hat{\theta}_n$ and the true value of the quantity of interest θ.

Definition 3.2. The bias of an estimator is defined by

$$b(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta.$$

We say that $\hat{\theta}_n$ is *unbiased* if $b(\hat{\theta}_n) = 0$.

We also define the variance of an estimator by

$$\mathbb{V}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2.$$

The standard error (se for short sometimes) is then

$$\mathrm{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)}.$$

Classically, unbiased estimators received a lot of attention since people wanted to have unbiased samples. However, modern statistics has a different point of view: because data is large, it doesn't matter if the samples are biased as long as the estimators converge to the true quantity of interest. This gives rise to the following definition

Definition 3.3. A point estimator $\hat{\theta}_n$ of a parameter θ is *consistent* if $\hat{\theta}_n$ converges to θ in probability.

Here comes the million-dollar question:

> How do we measure bias in the samples?

One possible approach is to use the so-called mean squared error.

Definition 3.4. The mean squared error of an estimator is defined by

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

Theorem 3.1 (Bias-Variance decomposition)**.**

$$MSE = b_\theta^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n)$$

Exercise 3.1. Prove the Bias-Variance decomposition.

Theorem 3.2. *If, as $n \to \infty$, $b_\theta^2(\hat{\theta}_n) \to 0$ and $\mathbb{V}_\theta(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n$ is consistent.*

Exercise 3.2. Prove the above theorem.

A big part of elementary statistics dealt with estimators being approximately related to the Normal distribution.

Definition 3.5. An estimator is said to be *asymptotically Normal* if

$$\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \to N(0,1)$$

in distribution.

3.2 Confidence set

In elementary statistics, given sample $X_1, \dots, X_n$, we define confidence interval with significance level α to be the interval (a, b) such that $\mathbb{P}_\theta(\theta \in (a, b)) \geq 1 - \alpha$.

Note that (a, b) depends on your sample, i.e., $a = a(X_1, \dots, X_n), b = b(X_1, \dots, X_n)$.

It must be stressed that θ is fixed and (a, b) is random.

For higher dimension / different kinds of data, the notion of confidence interval is replaced by the notion of confidence set.

Definition 3.6. Given sample $X_1, \dots, X_n$. A *confidence set* associated with significance level α is the set (random) C_n (depending on the sample) such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha.$$

Confidence set is not a probability statement about the parameter θ. It is rather a statement about the uncertainty of your data.

Example 3.1 (Example 6.14 in Wasserman)**.** Let $\theta \in \mathbb{R}$. Let X_1, X_2 RVs coming from the distribution $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Suppose $Y_i = \theta + X_i$ are your observed data. Define

$$C = \begin{cases} \{Y_1 - 1\} & Y_1 = Y_2\,, \\ \{(Y_1 + Y_2)/2\} & Y_1 \neq Y_2\,. \end{cases}$$

1. For all $\theta \in \mathbb{R}$, $\mathbb{P}_\theta(\theta \in C) = 3/4$.
2. Suppose we get $Y_1 = 9$, $Y_2 = 11$, $C = \{10\}$. Then, for sure, $\theta = 10$. Therefore, $\mathbb{P}(\theta \in C|Y_1, Y_2) = 1$.

Exercise 3.3. Recall Hoeffding's inequality

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

for $X_i \in [a_i, b_i]$ and $\mathbb{E}X_i = 0$.

Apply this to the Bernoulli parametric model

$$\mathcal{F} = \{\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p; p \in [0, 1]\}\,.$$

Question: Suppose our sample comes from a Bernoulli distribution. What is a confidence interval that gives significance level α?

Try with two approaches: Hoeffding and Chebyshev.

Exercise 3.4. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\hat{p}_n = n^{-1}\sum_{i=1}^{n} X_i$.

1. Compute $\mathbb{V}(X_i)$ and $\mathbb{V}(\hat{p}_n)$
2. Suppose we don't know $\mathbb{V}(\hat{p}_n)$, so we use an estimator of this quantity

$$\widehat{\text{se}}^2 = \hat{p}_n\left(1 - \hat{p}_n\right)/n\,.$$

Convince yourself that, by the Central Limit Theorem, $\hat{p}_n \approx N\left(p, \widehat{\text{se}}^2\right)$.

1. Find the confidence interval for the significance level α.
2. Compare this with the confidence interval in the previous exercise. You should see that the Normal-based interval is shorter but it only has approximately (when sample size is large) correct coverage.

3.3 Method of Moments

Let $l \in \mathbb{N}$, the l sample moment is

$$\hat{m}_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l\,.$$

Suppose that we want to determine k different parameters in a parametric model. The population moments are functions of those parameters:

$$\mu_l = \mu_l(\theta_1, \ldots, \theta_k)\,.$$

The method of moments says that one can construct parameters $\hat{\theta}_1, \ldots, \hat{\theta}_k$ by solving

$$\begin{aligned} \hat{m}_1 &= \mu_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ &\vdots \\ \hat{m}_k &= \mu_k(\hat{\theta}_1, \ldots, \hat{\theta}_k) \end{aligned} \tag{3.1}$$

Example 3.2. Let $X_1, \dots, X_n \sim N(\theta, \sigma^2)$. Construct estimators for the two parameters θ and σ^2.

Example 3.3. Let $X_1, \dots, X_n \sim \text{Binomial}(k, p)$, i.e.,

$$\mathbb{P}(X_i = x) = \binom{k}{x} p^x (1-p)^{k-x}.$$

Construct estimators for k and p.

Suppose the model we are considering has k parameters $\theta_j \in \mathbb{R}$, where $j = 1, \dots, k$.

Define a function $g : \mathbb{R}^k \to \mathbb{R}^k$ by

$$g(\theta) = \mu,$$

where

$$\theta = (\theta_1, \dots, \theta_k)$$

and

$$\mu = (\mu_1, \dots, \mu_k).$$

We can rephrase the above construction of the estimators as solving for $\hat{\theta}$, given $\hat{\mu}$ in the equation

$$\hat{\mu} = g(\hat{\theta}). \tag{3.2}$$

(Note that $\hat{\mu}$ and $\hat{\theta}$ depends on the sample (size))

Two natural questions arise:

1. Can we solve this equation?
2. Are the estimators consistent?

The answer to the first question is not so obvious. However, if we can solve the first problem, then the second problem is somewhat more manageable, given some reasonable assumptions.

Exercise 3.5. Suppose that $g : \mathbb{R}^k \to \mathbb{R}^k$ defined above is a bijection with continuous inverse. Then, for each $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\theta} - \bar{\theta}| > \epsilon) = 0.$$

That is, $\hat{\theta}$ is consistent.

Of course, g is nonlinear generally. So, the assumption that g is a bijection may seem to be too strong and not too satisfying. To fix this issue, let's consider a more general version of the above construction of estimators.

Define a modified version of the estimator $\hat{\theta}$ as follows

$$\tilde{\theta} = \begin{cases} \hat{\theta} & \text{if it is solvable} \\ 0 & \text{otherwise} \end{cases}$$

Theorem 3.3. *Suppose that all the moments of the underlying population are finte, g is of class C^1 and that* $\det[Dg] \neq 0$. *For each $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}(|\tilde{\theta} - \bar{\theta}| > \epsilon) = 0.$$

Proof. Let $\epsilon, \alpha > 0$.

From the weak law of large number, there exists $\bar{m}$ so that $\hat{m} \to \bar{m}$ as $n \to \infty$ in probability. Note, that $\bar{m}$ is the list of moments that are generated from the underlying population, therefore,

$$\bar{m} = g(\bar{\theta}),$$

where $\bar{\theta}$ is the list of underlying parameters.

By the inverse function theorem, there exists a $\delta > 0$ such that:

1. g is invertible in the ball $B(\bar{m}, \delta)$,
2. g^{-1} is of class C^1,
3. $g^{-1}(B(\bar{m}, \delta)) \subseteq B(\bar{\theta}, \epsilon)$.

Let N be such that for every $n > N$,

$$\mathbb{P}(|\hat{m} - \bar{m}| < \delta) \geq 1 - \alpha.$$

Since $|\hat{m} - \bar{m}| < \delta$ implies that $\hat{\theta}$ is uniquely solvable, i.e. $\hat{\theta} = g^{-1}(\hat{m})$, we have

$$\mathbb{P}(|\tilde{\theta} - \bar{\theta}| > \epsilon) \leq \mathbb{P}(|\hat{m} - \bar{m}| \geq \delta) \leq \alpha.$$

Therefore, since α is arbitrary,

$$\lim_{n \to \infty} \mathbb{P}(|\tilde{\theta} - \bar{\theta}| > \epsilon) = 0,$$

as desired. □

3.4 Maximum Likelihood Estimation

Definition 3.7. Suppose $X_1, \ldots, X_n \sim f_\theta$. The *likelihood* function is defined by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)\,.$$

The *log-likelihood function* is defined by

$$\ell_n(\theta)h = \ln \mathcal{L}_n(\theta)\,.$$

The *maximum likelihood estimator* MLE, denoted by $\hat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$.

Notation. Another common notation for the likelihood function is

$$L(X|\theta) = \mathcal{L}_n(\theta).$$

Example 3.4. Let $X_1, \ldots, X_n$ be sample from Bernoulli(p). Use MLE to find an estimator for p.

Example 3.5. Let $X_1, \ldots, X_n$ be sample from $N(\theta, 1)$. Use MLE to find an estimator for θ.

Exercise 3.6. Let $X_1, \ldots, X_n$ be sample from $Uniform([0, \theta])$, where $\theta > 0$.

1. Find the MLE for θ.
2. Find an estimator by the method of moments.
3. Compute the mean and the variance of the two estimators above.
4. Can you find the MLE if we consider $Uniform((0, \theta))$?

Theorem 3.4. *Let $\tau = g(\theta)$ be a bijective function of θ. Suppose that $\hat{\theta}_n$ is the MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.*

3.4.1 Consistency

Example 3.6 (Inconsistency of MLE)**.** Let $Y_{i,1}, Y_{i,2} \sim N(\mu_1, \sigma^2)$. Our goal is to find MLE for σ^2, which turns out to be

$$\hat{\sigma}^2 = \frac{1}{4n}\sum_{i=1}^{n}(Y_{i,1} - Y_{i,2})^2.$$

By law of large number, this will converge to

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2/2,$$

which means that the MLE is not consistent.

To discuss about the consistency of the MLE, we define the Kullback-Leibler distance between two pdf f and g.

$$D(f, g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx.$$

Abusing notation, we will write $D(\theta, \varphi)$ to mean $D(f(x;\theta), f(x;\varphi))$.

We say that a model $\mathcal{F}$ is *identifiable* if $\theta \neq \varphi$ implies $D(\theta, \varphi) > 0$.

Theorem 3.5. *Let $\theta_\star$ denote the true value of θ. Define*

$$M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\theta)}{f(X_i;\theta_\star)}$$

and $M(\theta) = -D(\theta_\star, \theta)$. Suppose that

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \to 0$$

in probability and that, for every $\epsilon > 0$,

$$\sup_{\theta:|\theta-\theta_,|\geq\epsilon} M(\theta) < M(\theta_\star).$$

Let $\hat{\theta}_n$ denote the MLE. Then $\hat{\theta}_n \to \theta_\star$ in probability.

Exercise 3.7. Let $X_1, \ldots, X_n$ be a random sample from a distribution with density:

$$p(x;\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.$$

1. Find the MLE for θ.
2. Find the Method of Moments estimator for θ.

Exercise 3.8. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

1. Find the method of moments estimator, the maximum likelihood estimator, and the Fisher information $I(\lambda)$.
2. Use the fact that the mean and variance of the Poisson distribution are both λ to propose two unbiased estimators of λ.
3. Show that one of these estimators has a larger variance than the other.

The conditions listed in the above theorem is not very easy to check. Hogg-McKean-Craig has a better theorem (this is a good theorem to read).

Theorem 3.6. *Assume that*

1. $\theta \neq \theta' \implies f_\theta \neq f_{\theta'}$

2. *f_θ has common support for all θ*
3. *θ^* is an interior point in Ω*

If $f_\theta(x)$ is differentiable with respect to θ. Then the likelihood equation

$$\frac{\partial}{\partial\theta} l_n(\theta) = 0$$

has a solution $\hat{\theta}_n$ such that

$$\lim_{n\to\infty} \hat{\theta}_n \to \theta^*$$

in distribution.

3.4.2 Asymptotic normality

Definition 3.8. Given a RV X. The score function is defined to be

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial\theta}.$$

The *Fisher information* is defined to be

$$I_n = \mathbb{V}_\theta\left(\sum_{i=1}^n s(X_i;\theta)\right) = \sum_{i=1}^n \mathbb{V}_\theta\left(s(X_i;\theta)\right).$$

Theorem 3.7. *$I_n(\theta) = nI(\theta)$. Furthermore,*

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\right).$$

The significance of this is that you can think of the Fisher information as the curvature (second derivative) on the "manifold" of parameters. So, error of the score function has certain geometric interpretation.

Theorem 3.8. *Let* $\text{se} = \sqrt{\mathbb{V}(\hat{\theta}_n)}$. *Given some regularity conditions, there exists a random variable $Z \sim N(0,1)$ such that*

$$\frac{\hat{\theta}_n - \theta}{\text{se}} \to Z.$$

3.4.3 Efficiency

As n gets large, the MLE is the most efficient estimator.

Theorem 3.9 (Cramer-Rao Inequality)**.** *Let $X_1, \ldots, X_n$ be sample with density $f(x;\theta)$. Suppose θ' is an unbiased estimator of θ, then under similar regularity conditions as in asymptotic normality,*

$$\mathbb{V}(\theta'_n) \geq \frac{1}{nI(\theta)}.$$

Note that, in the proof of asymptotic normality, we have that as n gets large, the MLE $\hat{\theta}$, if obeys the required regularity conditions, satisfies $\mathbb{V}(\hat{\theta}_n) \sim \frac{1}{nI(\theta)}$.

By consistency, $\hat{\theta} \sim \theta$ when n is very large. This means that

$$\mathbb{V}(\hat{\theta}_n - \theta) \sim \mathbb{V}(\hat{\theta}_n) \sim \frac{1}{nI(\theta)}.$$

Corollary 3.1. *Let $X_1, \dots, X_n$ be sample with density $f(x;\theta)$. Suppose θ'_n is an unbiased estimator of θ and $\hat{\theta}_n$ the MLE of θ, then, under regularity condition as in asymptotic normality, we have*

$$\lim_{n\to\infty} n\mathbb{V}(\theta'_n) \geq \lim_{n\to\infty} n\mathbb{V}(\hat{\theta}_n - \theta).$$

Note that this doesn't say that MLE (if consistent) is the most efficient for any finite n. In fact, this is a difficult question and one can only verify it for some specific estimator.

Exercise 3.9. Show that for Poisson processes, the MLE $\hat{\theta}_n$ is the most efficient for every n, compared to any other unbiased estimator θ'_n, i.e,

$$\mathbb{V}(\theta'_n) \geq \mathbb{V}(\hat{\theta} - \theta)$$

for every $n \in \mathbb{N}$.

Exercise 3.10 (Rice, 8.10.6)**.** Let $X \sim \text{Binomial}(n, p)$.

a. Find the MLE of p.

b. Show that the MLE from part (a) attains the Cramer-Rao lower bound.

3.5 (Optional) Expectation-Maximization Algorithm

Read Wasserman Section 9.13.4

3.6 Bayesian Approach

(Wasserman Chapter 11)

Please watch this great video by Philippe Rigollet (MIT) for the Bayesian approach: https://youtu.be/bFZ-0FH5hfs?si=IItsPqGD9g9kCC76

In short, we have that

$$f(\theta|X_1, \dots, X_n) \propto \mathcal{L}_n(\theta) f(\theta),$$

where $f(\theta|X_1, \dots, X_n)$ is a (believed) density distribution of the parameter θ called the posterior, $f(\theta)$ is a (believed) density distribution called the prior, and $\mathcal{L}_n(\theta)$ is the likelihood function.

One can think about this as how we update belief about the certain truth from a prior belief after seeing the evidence. This point of view is crucial in science; it is the scientific method written in mathematical form.

We can then construct a Bayesian estimator by simply taking the expectation of the posterior:

Definition 3.9 (Bayes estimator: Posterior mean)**.** Let Θ be a RV from the posterior $f(\theta|X_1, \dots, X_n)$.

$$\hat{\theta}_n = \mathbb{E}_{\theta|X}(\Theta) = \int \theta f(\theta|X_1, \dots, X_n) d\theta.$$

Note that this is one of the many candidates from Bayesian approach to talk about an estimator.

Exercise 3.11. Let $X_1, \dots, X_n$ be sample from Bernoulli distribution Bernoulli(p).

1. Suppose that at the beginning we believe that p obeys Beta(α, β) (see Definition). What is the posterior distribution of p after knowing the above sample?
2. Compute the Bayes estimator with the above posterior estimator. Compare this with MLE of p.
3. Suppose that at the beginning we believe that p obeys Uniform($[0, 1]$). What is the posterior distribution of p after knowing the above sample? Compare this with part (1). Explain what you see.

Exercise 3.12 (Rice, 8.10.4)**.** Suppose X is RV with the following distribution

$$\mathbb{P}(X=0) = \frac{2}{3}\theta$$
$$\mathbb{P}(X=1) = \frac{1}{3}\theta$$
$$\mathbb{P}(X=2) = \frac{2}{3}(1-\theta)$$
$$\mathbb{P}(X=3) = \frac{1}{3}(1-\theta)$$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$.

a. Find the method of moments estimate of θ.
b. Find an approximate standard error for your estimate.
c. What is the maximum likelihood estimate of θ ?
d. What is an approximate standard error of the maximum likelihood estimate?
e. If the prior distribution of Θ is uniform on $[0, 1]$, what is the posterior density? Plot it. What is the mode of the posterior?

3.7 Comparing Estimator / Decision Theory

Recall that we always denote: - θ: true parameter - $\hat{\theta}$: estimator of the true parameter (a function of the data)

So far, we learned a variety of ways to construct estimators: estimator by moment method, MLE, Bayes estimator. Which one would work the best? This is what we call *decision theory*. Often, within this theory, an estimator is called a *decision rule* and the possible values of the decision rule are called *actions*.

The language here is also the language used in machine learning. You will find a few repeated ideas from previous sections. However, the ideas are natural.

Before, one way to measure the discrepancy between the estimator $\hat{\theta}$ and θ is the mean square error that we learned previously. However, that is not the only way. To generalize this idea, we define the *loss function* $L(\theta, \hat{\theta})$ mapping from $\Theta \times \Theta \to \mathbb{R}$, where Θ is the set of parameters.

Example 3.7. Here is a few examples of loss functions:

$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$	Squared error loss
$L(\theta, \hat{\theta}) = \lvert\theta - \hat{\theta}\rvert^p$	L^p loss
$L(\theta, \hat{\theta}) = \begin{cases} 0, & \theta = \hat{\theta} \\ 1, & \theta \neq \hat{\theta} \end{cases}$	zero-one loss
$L(\theta, \hat{\theta}) = I(\lvert\theta - \hat{\theta}\rvert > c)$ where $I(x > c) = \begin{cases} 0, & \lvert\theta - \hat{\theta}\rvert \leq c \\ 1, & \lvert\theta - \hat{\theta}\rvert > c \end{cases}$	large deviation loss
$L(\theta, \hat{\theta}) = \int \log\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)$	Kullback-Leiber loss

Definition 3.10. The *risk* of an estimator is

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta(L(\theta, \hat{\theta})).$$

The risk from the squared error loss is the mean squared error.

Example 3.8. Compute the risks from the squared error for the MLE and Bayes estimator (with prior Beta(α, β)) for the family Bernoulli(p).

Definition 3.11. The *maximum risk* is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta} R(\theta, \widehat{\theta}).$$

The minimizer of the maximum risk is called the *minimax estimator*.

The *Bayes risk* is

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta)\, d\theta$$

where $f(\theta)$ is the prior for θ. The minimizer of the Bayes risk is called the *Bayes estimator*.

Let's do a heuristic calculation.

$$\begin{aligned} r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}) f(\theta)\, d\theta = \int \left(\int L(\theta, \widehat{\theta}) f(x^n|\theta) dx^n \right) f(\theta) d\theta \\ &= \int \int L(\theta, \widehat{\theta}) f(\theta|x^n) f(x^n) dx^n d\theta \\ &= \int \left(\int L(\theta, \widehat{\theta}) f(\theta|x^n) d\theta \right) f(x^n) dx^n. \end{aligned}$$

Denote the posterior risk as

$$r(\widehat{\theta}; \theta|x^n) = \int L(\theta, \widehat{\theta}) f(\theta|x^n) d\theta.$$

Because $r(\widehat{\theta}; \theta|x^n) \geq 0$, a minimizer for $r(\widehat{\theta}; \theta|x^n)$ for all x^n would minimize $r(f, \widehat{\theta})$.

Note that $\widehat{\theta}$ is a function that we are trying to find so this is a calculus of variations problem. Basic calculus isn't sufficient to solve this problem. However, we may proceed heuristically.

Example 3.9. Suppose that $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|^2$. Then,

$$r(\widehat{\theta}; \theta|x^n) = \int |\theta - \widehat{\theta}|^2 f(\theta|x^n) d\theta$$

Heuristically, the minimizer of the above is found when

$$0 = \frac{\partial}{\partial \widehat{\theta}} r(\widehat{\theta}; \theta|x^n) = \int (\widehat{\theta} - \theta) f(\theta|x^n) d\theta\,.$$

Therefore,

$$\widehat{\theta} = \int \theta f(\theta|x^n) d\theta.$$

So, the Bayes estimator for square loss would be the posterior mean.

The above calculation can be made rigorous via the study calculus of variations.

Exercise 3.13.

1. Let X be a continuous random variable. Show that

   $$\min_a \mathbb{E}|X - a|$$

   is achieved when a is the median of X, i.e,

   $$\mathbb{P}(X \geq a) = \mathbb{P}(X \leq a) = 1/2.$$

2. For absolute error loss, $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$, show that the Bayes estimator is the median of of the posterior distribution.

Chapter 4

Hypothesis Testing

(I wasn't entirely happy with the treatment of Wasserman and Casella & Berger. So I'm following the treatment of Hogg, McKean & Craig.)

A hypothesis test is a process to reject or not reject a well-defined statements. Intuitively, there are three components to a hypothesis test:

1. Null hypothesis H_0versus Alternative hypothesis H_1
2. Data
3. Decision rule to reject H_0 and accept H_1 or to not reject H_0 and reject H_1.

The mathematical formulation of this is a bit more restrictive because of the need for well-defined and verifiable statements. We will restrict our attention to hypotheses about either:

1. a parameter of a model, or
2. a functional of the underlying density distribution f, i.e., a mapping $T(f) \in \mathbb{R}$. For example, $\mu(f) = \int xf(x)dx$.

4.1 Procedure

Denote the quantity of interest (parameter or functional) to be θ. This parameter will be assumed to take place in a space D.

Suppose that, because of previous experience, we think that θ could only be in D_1 or D_2 (but not both), where $D_0 \cap D_1 = \emptyset$ and $D_0 \cup D_1 = D$.

The components of our test would be:

1. $H_0 : \theta \in D_0$ versus $H_1 : \theta \in D_1$
2. Sample: $X_1, \ldots, X_n$
3. Rejection (Critical) region C: determines the decision rule as follows:
   - Reject H_0 if $(X_1, \ldots, X_n) \in C$
   - Reject H_1 if $(X_1, \ldots, X_n) \notin C$

For $i = 0, 1$, if $D_i = \{\theta_i\}$ (that is, D_i is a singleton), then H_i is called a *simple* hypothesis. A hypothesis is *composite* if it is not simple. So, we could have cases where one of the hypotheses is simple and the other is composite, both are simple, or both are composite.

Due to probabilistic nature of the procedure, there are a few scenarios

	H_0 is true	H_1 is true
Reject H_0	Type I error (false positive)	Correct decision

	H_0 is true	H_1 is true
Reject H_1	Correct decision	Type II error (false negative)

One of the working assumption of science is that all models are wrong and only approximations of reality. So, the goal of science is to try to reject the hypothesis H_0. The harder it is to reject H_0, the better of a theory as an approximation of reality.

However, there is a conundrum the two types of errors will happen and it is not possible to minimize both at the same time.

Example 4.1. Let $C = \emptyset$. Then the probability for Type I error is 0 as one will never reject H_0. However, if it turns out that H_1 is true, then the probability for Type II error would be 1.

Often, we consider a false positive to be worse than a false negative (imagine a medical test says that you don't have a sickness while you do). So, we want to do the following:

1. Choose small probability α and find reasonable critical regions that make the probability for Type I error be bounded by α.
2. Among these critical regions, minimize the probability for Type II error.

Note that we have

$$1 - \mathbb{P}_\theta(\text{Type II error}) = \mathbb{P}_\theta\left[(X_1, \dots, X_n) \in C\right].$$

This inspires the following definitions

Definition 4.1 (size of critical region)**.** We say that a critical region C is of *size* α if

$$\alpha = \max_{\theta \in D_0} \mathbb{P}_\theta\left[(X_1, \dots, X_n) \in C\right].$$

Definition 4.2 (Power function of critical region)**.** A *power function* of a critical region C is a function $\gamma_C : D_1 \to [0, 1]$ so that

$$\gamma_C(\theta) = \mathbb{P}_\theta\left[(X_1, \dots, X_n) \in C\right].$$

Remark. From the above discussion, the quality of a hypothesis test is really determined by choosing the right critical region C. So a test is better than another test when the critical region of it is better than the critical region of the other one. We, thus, need to be able to compare critical regions.

Definition 4.3. Given two critical regions C_1 and C_2 of size α, C_1 is better than C_2 (denoted by $C_1 \succeq C_2$) if

$$\gamma_{C_1}(\theta) \geq \gamma_{C_2}(\theta), \forall \theta \in D_1.$$

Note that, not any pair of critical regions are comparable.

4.2 Neyman-Pearson Lemma

Assumption throughout this section: both hypotheses are simple.

When both the null and alternative hypotheses are simple, we can talk about the most powerful tests (or the *best critical region*).

Denote $X = (X_1, \dots, X_n)$ and recall that the space that this lives in is a state space S.

Definition 4.4. Let C be a subset of the state space. Then we say that C is the *best critical region* of size α for testing the simple hypothesis $H_0 : \theta = \theta_0$ against the alternative simple hypothesis $H_1 : \theta = theta_1$ if

a. $P_{\theta_0}((X_1, \dots, X_n) \in C) = \alpha$

b. And for every subset A of the state space

$$\mathbb{P}_{\theta_0}(X \in A) = \alpha \implies \mathbb{P}_{\theta_1}(X \in C) \geq \mathbb{P}_{\theta_1}(X \in A)$$

Recall the likelihood function

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta)$$

where $x = (x_1, \dots, x_n)$.

Theorem 4.1 (Neyman-Pearson Theorem)**.** *Let $X_1, \dots, X_n$ be a sample from a family of distributions $f(x; \theta)$, where $\theta \in \{\theta_0, \theta_1\}$. Let k be a positive number and C be a subset of the state space such that*

a. $\dfrac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\theta_1; x)} \leq k$ *for each point* $x \in C$.
b. $\dfrac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\theta_1; x)} \geq k$ *for each point* $x \in C^c$.
c. $\alpha = P_{\theta_0}(X \in C)$.

Then C is a bests critical region of size α for testing the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$.

4.3 Wald Test

4.4 Likelihood Ratio Test

4.5 Comparing samples

PART 3: Models

Chapter 5

Linear Least Squares

5.1 Simple Linear Regression

5.2 Matrix Approach

5.3 Statistical Properties