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 originality in this work as it is mostly taken from the reference books. However, all errors and typos are solely mine.

PART 1: Background

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: $X \sim \text{Poisson}(\lambda)$.
$$f_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}\,, x \geq 0.$$
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*

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

1.1.8 Moment Generating and Characteristics 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 *characteristics 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.

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 discrete RVs.

Proposition 1.1 (Discrete RV case)**.** *Let $X, Y : \Omega \to \mathbb{N}$ be discrete RVs. If $\varphi_X(t) = \varphi_Y(t)$ for all t in an interval around 0, then*

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

Proof. (to be written) □

We have a similar statement for characteristics function. However, the proof of this is a little more manageable.

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*

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

Proof. to be written □

1.2 Inequalities

1.3 Law of Large Numbers

1.4 Central Limit Theorem

PART 2: Inference

Chapter 2

Sampling, Estimating CDF and Statistical Functionals

Chapter 3

Parametric Inference (Parameter Estimation)

Chapter 4

Hypothesis Testing

4.1 Neyman-Pearson Lemma

4.2 Wald Test

4.3 Likelihood Ratio Test

4.4 Comparing samples

PART 3: Models

Chapter 5

Linear Least Squares

5.1 Simple Linear Regression

5.2 Matrix Approach

5.3 Statistical Properties