MATH/STAT395: Probability II Spring 2020

Review of MATH/STAT394

Chapters 1, 2, 3, Sections 4.4, 4.5, 4.6 of ASV

Instructor: Vincent Roulet *Teaching Assistant: Zhenman Yuen*

Review of probability distributions

This lecture note serves as reference about the material you should know from MATH/STAT394. Starred items are advanced topics, you don't need to know but it is preferable.

1 Probability space, conditional probability, independence

1.1 Probability space

Definition 1 (Probability space)**.** *A* ***probability space*** $(\Omega, \mathcal{F}, \mathbb{P})$ *consists of*

- *A sample space* Ω*, the set of all possible outcomes of a random action,*
- *A set of events* $\mathcal{F}$*, where each event* $E \in \mathcal{F}$ *is a subset of* Ω*, (*$\mathcal{F} \subset 2^\Omega$ *must be a* σ*-algebra)*
- *A probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that assigns probabilities to events.*

Axioms of probability

1. *For all* $A \in \mathcal{F}$*,* $0 \leq \mathbb{P}(A) \leq 1$*,*
2. $\mathbb{P}(\emptyset) = 0$*,* $\mathbb{P}(\Omega) = 1$
3. *For any sequence* $A_1, A_2, \ldots \in \mathcal{F}$ *of disjoint sets,*

$$\mathbb{P}\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i)$$

Definition 2 (σ-algebra*)**.** *Let* Ω *be a set. A* σ*-algebra* $\mathcal{F}$ *on* Ω *is a subset of* $2^\Omega = \{B \subset \Omega\}$ *such that*

1. $\Omega \in \mathcal{F}$
2. *For any* $A \in \mathcal{F}$*,* $A^c \triangleq \Omega \setminus A \in \mathcal{F}$
3. *For any* $A_1, A_2, \ldots \in \mathcal{F}$*,* $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$

The smallest σ*-algebra that contains all intervals of* $\mathbb{R}^n$ *is called the Borel algebra of* $\mathbb{R}^n$*.*

1.2 Conditional probability

Definition 3 (Conditional probability)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $B \in \mathcal{F}$ *s.t.* $\mathbb{P}(B) \neq 0$*, the* ***conditional probability of*** A ***given*** B *is defined as*

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

Definition 4. $B_1, \ldots, B_n \subset \Omega$ *is a partition of* Ω *if* $\bigcup_{i=1}^n B_i = \Omega$ *and* $B_i \cap B_j = \emptyset$ *for any* $i \neq j$*.*

Property 5. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space*

1. *For* $B \in \mathcal{F}$ *s.t.* $\mathbb{P}(B) \neq 0$*,* $\mathbb{P}(\cdot|B)$ *satisfies the axioms of probability*

2. *For* $A_1 \dots A_n \in \mathcal{F}$,

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_n | A_{n-1} \cap \dots \cap A_1)\, \mathbb{P}(A_{n-1} | A_{n-2} \cap \dots \cap A_1) \dots \mathbb{P}(A_1)$$

3. *Let* $B_1, \dots, B_n \in \mathcal{F}$ *a partition of* Ω *such that* $\mathbb{P}(B_i) > 0$ *for all* i*, then we have*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)$$

Theorem 6 (Bayes Formula)**.** *Let* $B_1, \dots, B_n \in \mathcal{F}$ *a partition of* Ω *such that* $\mathbb{P}(B_i) > 0$ *for all* i*, then we have for any* $k \in \{1, \dots, n\}$*,*

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_k)\, \mathbb{P}(B_k)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)}$$

1.3 Independence

Definition 7 (Independence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. Two events* $A, B \in \mathcal{F}$ *are said to be* ***independent*** *if*

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

n *events* $A_1, \dots A_n \in \mathcal{F}$ *are* ***independent*** *or* ***mutually independent*** *if for any* $2 \leq k \leq n$ *and* $1 \leq i_1 \leq \dots \leq i_k \leq n$*,*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k})$$

Property 8. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. If* A, B *are independent, then any pair of events* $(A^*, B^*) \in \{(A, B), (A^c, B), (A, B^c), (A^c, B^c)\}$ *is a pair of independent events.*

Definition 9 (Conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $B \in \mathcal{F}$ *s.t.* $\mathbb{P}(B) \neq 0$*, events* $A_1, \dots A_n$ *are* ***conditionally independent*** *if they are independent with respect to the probability* $\mathbb{P}(\cdot|B)$*.*

Definition 10. *Let* $X_1, \dots X_n$ *be r.v. (see definition below) on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*.* $X_1, \dots X_n$ *are independent if for any*[1] $B_1, \dots, B_n \subset 2^{\Omega}$*,*

$$\mathbb{P}(X_1^{-1}(B_1) \cap \dots \cap X_n^{-1}(B_n)) = \prod_{i=1}^{n} \mathbb{P}(X_i^{-1}(B_i))$$

2 Random variables

2.1 Probability distribution

Definition 11 (Probability distribution of a random variable)**.** *Given a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, a (real-valued)* ***random variable*** *(r.v.)* X *is defined as a mapping* $X : \Omega \to \mathbb{R}$ *such that for any*[1] *subset* $B \subset \mathbb{R}$*,*

$$\{X \in B\} \triangleq X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F}.$$

Denoting $2^{\mathbb{R}} = \{B \subset \mathbb{R}\}$*, the* ***probability distribution*** *of* X *is the mapping*

$$\mathbb{P}_X : \begin{cases} 2^{\mathbb{R}} & \to [0, 1] \\ B & \mapsto \mathbb{P}_X(B) \triangleq \mathbb{P}(\{X \in B\}) \end{cases}$$

We that "X *follows a distribution* $\mathbb{P}_X$*" and denote it by* $X \sim \mathbb{P}_X$*.*

Definition 12 (Discrete Random variable)**.** *A r.v.* X *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is said to be* ***discrete*** *if it takes values in a finite or countably infinite set* $\mathcal{X} = X(\Omega)$ *s.t.* $\sum_{k \in \mathcal{X}} \mathbb{P}(X = k) = 1$

[1] A formal definition requires to restrict the subsets considered in the definition to belong to the Borel algebra of $\mathbb{R}$ defined above.

2.2 Probability functions

Definition 13 (Probability mass function)**.** *Let X be a **discrete** r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **probability mass function** (p.m.f.) p of X is defined by :*

$$p : \begin{cases} X(\Omega) & \to [0,1] \\ k & \to p(k) \triangleq \mathbb{P}(X = k) \end{cases}$$

Definition 14 (Probability density function)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If a function f satisfies*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \qquad \text{for any } a, b \in \mathbb{R} \cup \{-\infty, +\infty\},$$

*then f is called the **probability density function** (p.d.f.) of X. X is then called a **continuous** r.v.*

Note: From now on, in the definitions, we consider without loss of generality, that if X is a discrete random variable, then $X(\Omega) = \mathbb{Z}$, that is, we identify any countable set to the set of integers. For random variables taking values in a finite set $\mathcal{X}$, it means that we assume this set to be a set of integers and that we consider $\mathbb{P}(X = k) = 0$ for any $k \in \mathbb{Z} \setminus \mathcal{X}$. Similarly, for continuous random variables, we consider $\mathcal{X}(\Omega) = \mathbb{R}$, such that if the random variable takes values in a bounded set $\mathcal{X}$, then $f(x) = 0$ for any $x \in \mathbb{R} \setminus \mathcal{X}$.

Property 15. *Let f be a p.d.f. of a r.v. X then*

1. $\int_{-\infty}^{+\infty} f(x)dx = 1$, $f(x) \geq 0$ *for all* $x \in \mathbb{R}$
2. $\mathbb{P}(X = k) = \int_k^k f(x)dx = 0$

Definition 16 (Cumulative distribution function)**.** *The **cumulative distribution function** (c.d.f.) of a r.v. X on $(\Omega, \mathcal{F}, \mathbb{P})$ is*

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}_X([-\infty, t])$$

Property 17. *Let F be the c.d.f. of a r.v. then*

1. $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a)$
2. $\lim_{t \to -\infty} F(t) = 0$, $\lim_{t \to +\infty} F(t) = 1$
3. *If* $s \leq t$, $F(s) \leq F(t)$
4. $F(t) = \lim_{s \to t^+} F(s)$

2.3 Expectation

Definition 18 (Expectation)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

1. *(Discrete case) If X has a p.m.f p s.t. $\sum_{k \in \mathbb{Z}} |k|p(k) < \infty$, the **expectation** (or **expected value**) of X exists and reads*
$$\mathbb{E}[X] = \sum_{k \in \mathbb{Z}} kp(k)$$
2. *(Continuous case) If X has a p.d.f. f s.t. $\int_{-\infty}^{+\infty} |x|f(x)dx < +\infty$ the **expectation** (or **expected value**) of X exists and reads*
$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

Property 19 (Linearity of Expectation)**.** *Let X, Y be two (discrete/continuous) r.v. and $a \in \mathbb{R}$,*

$$\mathbb{E}[aX + Y] = a\,\mathbb{E}[X] + \mathbb{E}[Y]$$

Proof. If X, Y are two discrete continuous random variables the result comes from the linearity of the sum. If X, Y are two continuous random variables, the result comes from the linearity of the integral. □

Property 20 (Expectation of a function of a random variable)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : X(\Omega) \to \mathbb{R}$. Then $g(X)$ is a r.v. and*

1. *(Discrete case) if X has a p.m.f. p, and $\sum_{k\in\mathbb{Z}} |g(k)|p(k) < +\infty$, then*

$$\mathbb{E}[g(X)] \text{ exists} \quad \text{and} \quad \mathbb{E}[g(X)] = \sum_{k\in\mathbb{Z}} g(k)p(k)$$

2. *(Continuous case) if X has a p.d.f. f, and $\int_{-\infty}^{+\infty} |g(x)|f(x)dx < +\infty$, then*

$$\mathbb{E}[g(X)] \text{ exists} \quad \text{and } \mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

Property 21. *Let X be a r.v. with probability distribution $\mathbb{P}_X$ and c.d.f. F_X, then*

$$\mathbb{E}[\mathbf{1}_B(X)] = \mathbb{P}[X \in B] = \mathbb{P}_X(B), \qquad \mathbb{E}[\mathbf{1}_{[-\infty,t]}(X)] = \mathbb{P}(X \le t) = F_X(t)$$

where $\mathbf{1}_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$

2.4 Moments, Variance

Definition 22 (Moment)**.** *For a r.v. X and $m \in \mathbb{N}$, if $\mathbb{E}[|X|^m] < +\infty$, then*

1. *the m^{th} moment of X exists and is defined as $\mathbb{E}(X^m)$*
2. *the m^{th} centered moment is defined as $\mathbb{E}((X - \mathbb{E}(X))^m)$*

Definition 23 (Variance–Standard Deviation)**.** *Let X be a discrete r.v. on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathbb{E}[|X|^2] < +\infty$, the **variance** of X is defined by*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

*The **standard deviation** of X is defined by $\sigma_X = \sqrt{\mathrm{Var}(X)}$*

Definition 24 (Degenerate random variable)**.** *A r.v. X is said to be degenerate if $\exists a \in \mathbb{R}$ s.t. $\mathbb{P}(X = a) = 1$.*

Property 25. *If X is a degenerate r.v. as defined in Def. 24, then $\mathbb{E}[X] = a$. Moreover, for any r.v. X we have $\mathrm{Var}(X) = 0 \Leftrightarrow X$ is degenerate.*

Remark 26. *In the course, for any $b \in \mathbb{R}$, we define e.g. $\mathbb{E}[b]$ by identifying b to the associated degenerate r.v.* $X : \begin{cases} \Omega & \to \mathbb{R} \\ \omega & \mapsto b \end{cases}$

Property 27. *For any r.v. X and $a, b \in \mathbb{R}$,*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

2.5 Common discrete random variables

In the following we emphasize the set of values that can take the random variable as $X(\Omega) = \{k \in \mathbb{Z} : \mathbb{P}(X = k) \neq 0\}$.

2.5.1 Bernoulli

Model Models the success of a trial (1 for success, 0 for fail)
Example Can model that the flip of a coin will be tail.
Range $X(\Omega) = \{0, 1\}$
Parameters $p \in [0, 1]$
Notation $X \sim \text{Ber}(p)$
Probability mass function $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$
Expectation, Variance $\mathbb{E}[X] = p$, $\text{Var}(X) = p(1 - p)$

2.5.2 Binomial

Model Model the number of success among n trials, each trial being independent and identically distributed as a Bernoulli r.v. with parameter p
Example Models the number of tails among n flips of a coin
Range $X(\Omega) = \{0, \dots, n\}$
Parameters $n \in \mathbb{N}, p \in [0, 1]$
Notation $X \sim \text{Bin}(n, p)$
Probability mass function $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}$ for $k \in \{0, \dots n\}$
Expectation, Variance $\mathbb{E}[X] = np, \text{Var}[X] = np(1 - p)$

Proof. Proof done for expectation. For the variance the proof can be found in the book page 115. We will provide a much simpler proof later. □

Remark Can be written as $X = \sum_{i=1}^{n} B_i$, where $B_i \sim \text{Ber}(p)$ are n independent Bernoulli r.v.

2.5.3 Poisson

Model Models the number of success among an infinite number of trials, with an average number of success λ
Example Models the number of typos in an infinite document
Range $X(\Omega) = \mathbb{N}$
Parameters $\lambda > 0$
Notation $X \sim \text{Poisson}(\lambda)$
Probability mass function $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ for $k \in \mathbb{N}$.
Expectation, Variance $\mathbb{E}[X] = \lambda$, $\text{Var}[X] = \lambda$.
Remark Consider a sequence of binomial random variables $B_n \sim \text{Bin}(n, \lambda/n)$ defined for $n > \lambda$, such that the average number of success of all those random variables is independent of n, then this sequence of random variables converge in distribution to a Poisson distribution as n goes to infinity. That is we retrieve the model of a Poisson distribution as the number of successes among an infinite number of trials.

2.5.4 Geometric

Model Models the number of trials of Bernoulli random variable with proba of success p before getting one success
Example Number of times you play an armed bandit before getting some money
Range $X(\Omega) = \mathbb{N}$
Parameters $p \in [0, 1]$
Notation $X \sim \text{Geom}(p)$
Probability mass function $\mathbb{P}(X = k) = (1 - p)^{k-1} p$
Expectation, Variance $\mathbb{E}(X) = \frac{1}{p}, \text{Var}(X) = \frac{1-p}{p^2}$

2.5.5 Hypergeometric*

Model Models sampling without replacement with order not mattering. Specifically denote K the number of items A in a total number of items N and assume we draw n items from this set. The random variable $X =$" number of items A in the n items that we sampled from the set" is distributed as a hypergeometric random variable

Range $X(\Omega) = \{0, \dots K\}$

Parameters $K, N, n \in \mathbb{N}$ with $1 \leq n \leq N$ and $1 \leq K \leq N$

Notation $X \sim \text{Hypergeom}(N, Kn)$

Probability mass function $\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Expectation $\mathbb{E}[X] = n\frac{K}{N}$

2.6 Common continuous random variables

In the following we emphasize the set of values that can take the random variable as $X(\Omega) = \{x \in \mathbb{R} : f(x) \neq 0\}$.

2.6.1 Uniform

Model Uniform probability on an interval $[a, b]$, with $a < b$

Example Models the reaching point of a bowling ball

Range $X(\Omega) = [a, b]$

Parameters $a, b \in \mathbb{R}$, $a < b$

Notation $X \sim \text{Unif}([a, b])$

Probability density function $f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

Expectation, Variance $\mathbb{E}(X) = \frac{a+b}{2}, \text{Var}(X) = \frac{(b-a)^2}{12}$

2.6.2 Gaussian

Model Standard continuous distribution to model a continuous random variable centered around a point μ with variance σ^2

Range $X(\Omega) = \mathbb{R}$

Parameters μ, σ^2

Notation $X \sim \mathcal{N}(\mu, \sigma^2)$

Probability density function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Expectation, Variance $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \sigma^2$

Remark Appears as the asymptotic behavior of the empirical mean of independent and identically distributed random variables, see central limit theorem studied later in the course.

2.6.3 Exponential

Model Models the waiting time before an event occurs, with an average of waiting time λ

Example Waiting time for a phone call

Range $X(\Omega) = [0, +\infty)$

Parameters $\lambda > 0$

Notation $X \sim \text{Exp}(\lambda)$

Probability density function $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Expectation, Variance $\mathbb{E}(X) = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}$

Remark Can be seen as the continuous time counterpart of the geometric distribution see lecture 4

2.6.4 Gamma distribution*

Model Versatile family of distribution that can model for example the time needed for a n^{th} phone call

Range $X(\Omega) = [0, +\infty)$

Parameters $\lambda > 0, r > 0$

Notation $X \sim \text{Gamma}(r, \lambda)$

Probability density function $f(x) = \begin{cases} \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$ where $\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx$

Expectation, Variance $\mathbb{E}(X) = \frac{r}{\lambda}, \text{Var}(X) = \frac{r}{\lambda^2}$