DATA8015: Mathematical Foundation of Data Science **Fall 2025**

Part 2.1 Probability and Statistics Basics

Scriber: Yingyu Liang

1 Definition of Probability

Probability theory is one of the pillars of modern data science. There are different philosophical views on the essences of probability (e.g., frequentists' and Bayesian). Fortunately, the rigorous foundation of probability theory has been well-established, which takes the modern mathematical axiom approach. Informally, this approach views probability as an abstract object in a Platonic/idealized world, specifies a few basic and intuitively-should-be-true properties (axioms), and then, on top of those axioms, derives all other properties and builds the whole probability theory. It is amazing that just a few axioms lead to the whole magnificent field of probability theory. It is even more amazing that this theory, built purely in the idealized world, aligns well with the physical world and thus finds numerous applications in practice.

Probability is meant to measure the uncertainty of observations. Such an uncertain observation is called a *random experiment*, and the set of all possible outcomes from the random experiment is called the *sample space*, usually denoted by Ω. An *event* then corresponds to a subset of the sample space. For example, for the random experiment of rolling a dice, the sample space is $\{1, 2, 3, 4, 5, 6\}$, and the event "getting an odd number" corresponds to the subset $\{1, 3, 5\}$.

Classical probability theory is intuitively developed for a finite sample space, while probability theory on possibly infinite sample spaces needs the rigorous treatment of the modern axiom approach.

1.1 Warmup: Probability on A Finite Sample Space

For a finite sample space Ω, we can associate with each outcome $\omega \in \Omega$ with a corresponding probability number $p(\omega) \geq 0$; and $\sum_{\omega\in\Omega} p(\omega) = 1$. Then we can talk about the probability of any event $E \subseteq \Omega$, by letting the probability of E to be $P(E) = \sum_{\omega\in E} p(\omega)$.

Intuitively, the probability should satisfy the following three basic properties:

- $P(E) \geq 0$ for all $E \subseteq \Omega$.
- $P(\emptyset) = 0, P(\Omega) = 1$.
- For disjoint E_1 and E_2, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

It is easy to verify that the probability defined in the previous paragraph satisfies these properties (and also some other desired properties).

Conversely, if the probability numbers for events $P(E)$ satisfy the above three properties, it turns out that they lead to all the other desired properties. Furthermore, note that an

outcome ω can be viewed as a special case of events (the singleton event $\{\omega\}$), so we can get back $p(w)$ from the probability $P(\{\omega\})$.

Therefore, we can define probability as a function $P(E)$ that is defined on subsets E of the sample space Ω and satisfies the above three properties (axioms).

1.2 Modern Definition of Probability

The sample space is often infinite, e.g., the temperature tomorrow. For such cases, we cannot follow the intuitive approach of associating with each outcome $\omega \in \Omega$ with a corresponding probability number $p(\omega)$ and letting $P(E) = \sum_{\omega \in E} p(\omega)$. After all, it could be that $p(\omega) = 0$ for all ω, and it is also unclear how to define the infinite sum properly.

Then we can follow the approach of considering a function over subsets $P(E)$ that satisfies some axioms. Unfortunately, the three simple properties above are not sufficient to derive other desired properties this time. Informally, the challenges come from 1) there could be some pathological subsets that cannot be handled properly; 2) infinite sums are needed to derive some properties.

Therefore, modern probability theory first considers a selected class of subsets (instead of the class of all subsets) of the sample space. This class should have a nice structure while supporting all those needed operations. In particular, we consider a σ-field (also called σ-algebra) on Ω.

Definition 1 (σ-field/σ-algebra)**.** A class $\mathcal{F}$ of subsets of Ω is called a σ-field if it contains Ω itself and is closed under the formation of complements and countable unions:

(i) $\Omega \in \mathcal{F}$;

(ii) $A \in \mathcal{F}$ imples $A^c \in \mathcal{F}$;

(iii) $A_1, A_2, \cdots \in \mathcal{F}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

By Demorgan's law, σ-field is also closed under countable intersection.

Based on the σ-field, we define probability by updating the three axioms in the previous subsection to handle countable sums.

Definition 2 (Probability)**.** A set function is a real-valued function defined on some class of subsets of Ω. A set function P on a σ-field $\mathcal{F}$ is a *probability measure* (abbreviated as probability), if it satisfies the following:

(i) (Nonnegativity) $P(E) \geq 0, \forall E \in \mathcal{F}$;

(ii) (Normalization) $P(\Omega) = 1$;

(iii) (Countable Additivity) if $E_i \in \mathcal{F}, i = 1, 2, \cdots$, and $E_i \cap E_j = \emptyset, \forall i \neq j$, then $P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

The triple $(\Omega, \mathcal{F}, P)$ is called a *probability measure space*, or simply a *probability space*.

This then allows us to derive many other properties of probability. Some basic ones:

1. $P(E^c) = 1 - P(E), P(\emptyset) = 0, P(E) \leq 1, \forall E \in \mathcal{F}$.
2. If $E_1, E_2, \cdots, E_n$ are disjoint, then $P\left(\cup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$.
3. For any two $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B), P(A - B) = P(A) - P(A \cap B)$.
4. If $A \subseteq B$, then $P(A) \leq P(B)$.
5. (Jordan's Formula) For any $E_1, E_2, \cdots, E_n$,

$$P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) - \sum_{1 \leq i < j \leq n} P(E_i \cap E_j) + \sum_{1 \leq i < j < k \leq n} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n+1} P(E1 \cap E_2 \cap \cdots \cap E_n).$$

$$P\left(\cup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i).$$

1.3 Independence

The fact that two events are "irrelevant" is captured formally by the notion of independence.

Definition 3 (Independence of Two Events)**.** Two events $A, B \in \mathcal{F}$ are mutually independent (or simply called independent), if $P(AB) = P(A)P(B)$.

Here AB is a shorthand for $A \cap B$.
This can be extended to more than two events.

Definition 4 (Mutual Independence of Events)**.** n events $A, B \in \mathcal{F}$ are mutually independent, if for any $k (2 \leq k \leq n)$ events $A_{i_1}, A_{i_2}, \cdots, A_{i_k} (1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n)$,

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_k}).$$

Here $A_{i_1} A_{i_2} \cdots A_{i_k}$ is a shorthand for $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$.

Note that mutual independence of n events is different from the pairwise independence. The pairwise independence only requires that any two events in the collection are independent of each other. The mutual independence informally means any combination of some events in the collection is independent of any combination of the other events. Formally, let I be any subset of $\{1, 2, \cdots, n\}$. Let $\mathcal{F}_I$ be the smallest σ-field containing $A_i (i \in I)$ (i.e., any combination of these events), and $\mathcal{F}_{\neg I}$ be the smallest σ-field containing $A_i (i \notin I)$. Then any $B_1 \in \mathcal{F}_I, B_2 \in_{\neg I}$ are independent.

1.4 Conditional Probability

Often we would like to reason about after some event C happens, what are the probabilities of other events. This is formalized as conditional probability.

Definition 5 (Conditional Probability)**.** Given an event C with $P(C) > 0$, $P(E|C) := P(EC)/P(C)$ is called the conditional probability of event E conditioned on C.

This notion is particularly important for data science and AI, since most applications essentially are formalized as a conditional probability estimation problem. For example, the problem of image classification is formalized as estimating the conditional probabilities of the class labels conditioned on the given image.

Note that $P(E|C)$ defines a new probability function $Q(E)$, so it enjoys all the properties of a probability.

2 Random Variables

The definition of probability focuses on sets, while in practice, most applications need to consider numbers. For example, when rolling a dice, the outcome will correspond to a number. We thus introduce the notion of a random variable, which associates the outcome ω with a number $X(\omega)$.

Definition 6 (Random Variable)**.** Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $X : \Omega \to \mathbb{R}$ is a real-valued function defined on Ω. If for any $a \in \mathbb{R}$, $\{\omega : X(\omega) \leq a\} \in \mathcal{F}$, then X is called a random variable.

In other words, we require that the preimage of any set $(-\infty, a]$ is in the σ-field $\mathcal{F}$. Note that a statement about X like $X \leq a$ corresponds to a subset $(-\infty, a]$ of $\mathbb{R}$, which in turn corresponds to a subset $\{\omega : X(\omega) \leq a\}$ of the sample space. The definition is to make sure we can talk about the probability of such events on X. By definition of the σ-field, $\mathcal{F}$ also contains the events $X > a$, or the event $b \leq X \leq a$. In fact, $\{\omega : X(\omega) \leq a\} \in \mathcal{F}$ for any $a \in \mathbb{R}$ is equivalent to $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ for any Borel set $B \subseteq \mathbb{R}$. The Borel set is rich enough to quantify practical statements about X.

Definition 7 (Distribution Function)**.** Suppose X is a random variable on $(\Omega, \mathcal{F}, P)$. The distribution function of X is:

$$F(x) = P(X \leq x) = P(X \in (-\infty, x]), \quad \forall x \in \mathbb{R}.$$

The distribution function essentially captures all information needed to reason about statements on X. But sometimes, it is convenient to talk about the probability density functions.

Definition 8 (Probability Density Function)**.** For the distribution function $F(x)$ of a random variable X, if there exists a nonnegative function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(u)du, \quad \forall x \in \mathbb{R},$$

then $f(x)$ is called the probability density function of X.

Clearly, if $f(x)$ is continuous, then we have $f(x) = \frac{dF(x)}{dx}$.

For more than one random variables, we can define their joint distribution.

Definition 9 (Joint Distribution)**.** The joint distribution function of n random variables $X_1, X_2, \cdots, X_n$ is:

$$F(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n).$$

Definition 10 (Marginal Distribution)**.** Given the joint distribution function F of n random variables $X_1, X_2, \cdots, X_n$, the marginal distribution of X_i is

$$F_{X_i}(x_i) = P(X_i \leq x_i) = \lim_{x_j \to +\infty, \forall j \neq i} F(x_1, x_2, \cdots, x_n).$$

We can now define mutual independence and conditioning for random variables.

Definition 11 (Independent Random Variables)**.** n random variables $X_1, \cdots, X_n$ are mutually independent, if $\forall (x_1, \cdots, x_n) \in \mathbb{R}^n$, the joint distribution function satisfies

$$F(x_1, \cdots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$$

where $F_{X_i}(x_i) = P(X_i \leq x_i)$.

It can be shown that if X, Y, Z are mutually independent, then $g_1(X, Y)$ and $g_2(Z)$ are independent, where g_1, g_2 can be piecewise monotonic or piecewise continuous functions.

Consider two random variables (X, Y), with joint distribution function $P(X \leq x, Y \leq y)$.

Definition 12 (Conditional Distribution Function)**.** Suppose D is a Borel set, and $P(Y \in D) > 0$. The conditional distribution function of X conditioned on the event $Y \in D$ is:

$$P(X \leq x | Y \in D) = \frac{P(X \leq x, Y \in D)}{P(Y \in D)}, \quad \forall x \in \mathbb{R}.$$

It is common to consider the special case when D is a singleton $\{y\}$. The above definition applies when $P(Y = y) > 0$, but we may need to consider how to define conditioning when $P(Y = y) = 0$.

Definition 13 (Conditional Distribution Function on Singleton)**.** Suppose $P(y < Y \leq y + h) > 0$ for sufficiently small $h > 0$. If

$$P(X \leq x | Y = y) := \lim_{h \to 0} P(X \leq x | y < Y \leq y + h)$$

exists, then $P(X \leq x | Y = y)$ is called the conditional distribution function of X conditioned on $Y = y$.

2.1 Numeric Characterizations of Random Variables

Riemann-Stieltjes Integral. For convenience, we introduce the Riemann-Stieltjes integral (or simply RS integral). It is a useful tool in unifying equivalent forms of statistical notions and theorems that apply to discrete and continuous probability.

Definition 14 (Riemann-Stieltjes Integral)**.** Suppose $F, g : (-\infty, +\infty) \to \mathbb{R}$, F is monotone nondecreasing and right-semicontinuous, and $a < b$.
Pick any partition $a = x_0 < x_1 < \cdots < x_{i-1} < x_i < \cdots < x_n = b, \forall u_i \in [x_{i-1}, x_i]$, define

$$\sum_{i=1}^{n} g(u_i)\Delta F(x_i) = \sum_{i=1}^{n} g(u_i)[F(x_i) - F(x_{i-1})].$$

Let $\lambda = \max_{1 \le i \le n}(x_i - x_{i-1})$. If

$$\lim_{\lambda \to 0} \sum_{i=1}^{n} g(u_i)\Delta F(x_i)$$

exists, then denote it as $\int_a^b g(x)dF(x)$ and call it the Riemann-Stieltjes integral of $g(x)$ on $[a, b]$ with respect to $F(x)$. If the limit exists when $a \to -\infty, b \to +\infty$, then denote it as $\int_{-\infty}^{+\infty} g(x)dF(x)$.

If $F(x) = x$, then the RS integral reduces to the typical Riemann integral. In probability, $F(x)$ is usually taken to be the distribution function of some random variable X. Some examples:

- If $g(x) = 1$, then $\int_a^b dF(x) = F(b) - F(a) = P(a < X \le b)$.
- If X is a discrete random variable, i.e., $P(X = c_i) = p_i, i \in \{1, 2, \cdots\}$, then $F(x) = \sum_{c_i \le x} p_i$ is a staircase function.

$$\int_{-\infty}^{+\infty} g(x)dF(x) = \sum_{n=1}^{\infty} g(c_n)[F(c_n + 0) - F(c_n - 0)] = \sum_{n=1}^{\infty} g(c_n)p_n$$

  is the average of g weighted by the probabilities on the discrete values.

Numeric Characterizations. Now we are ready to use RS integral to define some numeric characterizations, which apply to general random variables, including discrete and continuous ones.

Suppose $F(x)$ is the distribution function of a random variable X.

Definition 15 (Expectation)**.** If $\int_{-\infty}^{+\infty} |x|dF(x)$ exists, then define

$$\mathbb{E}X := \int_{-\infty}^{+\infty} xdF(x)$$

as the expectation of X.

When X is discrete (i.e., $P(X = x_n) = p_n (n \in \mathbb{N}$, then $\mathbb{E}X = \sum_{n=1}^{\infty} x_n p_n$ is the weighted average of x_i's. When X is continuous and has a probability density function $f(x)$, $\mathbb{E}X = \int_{-\infty}^{+\infty} xf(x)dx$.

Definition 16 (Variance)**.** Define

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

as the variance of X.

Definition 17 (Covariance)**.** For two random variables X, Y, define

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$$

as the covariance of (X, Y).

When X, Y are indepedent, then $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$, and $\mathrm{Cov}(X, Y) = 0$. So $\mathrm{Cov}(X, Y)$ characterizes some statistical correlation between X and Y.

Definition 18 (Moments)**.** For $k \geq 1$, define

$$\mathbb{E}(X^k) := \int_{-\infty}^{+\infty} x^k dF(x)$$

as the k-th moment of X.

3 Bayes' Rule

Bayes' rule is an important tool for statistical decision making. Consider the probability of an event H (i.e., hypothesis) after observing an event E (i.e., evidence), which is formalized as $P(H|E)$. Bayes' rule is:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where $P(H)$ is called the prior, $P(E|H)$ the likelihood, and $P(H|E)$ the posterior.

The rule comes from applying the definition of conditional probability twice:

$$P(H|E)P(E) = P(H, E) = P(E|H)P(H).$$

One can also use the definition of the conditional density function to derive the corresponding Bayes' rule.

The rule switches the conditioning from conditioned on E to conditioned on H. This is especially useful when conditioning on E is hard while conditioning on H is easy. For example, inferring the disease (H) from the symptoms of the patient (E) is usually harder than describing the symptoms of the disease. Many applications have such a structure, and thus Bayes' rule is widely useful.

Usually, we first formulate the problem as a conditional probability $P(H|E)$, apply Bayes' rule, and then estimate the three terms on the right-hand side (to be discussed in the lecture on statistical estimation).

3.1 Naïve Bayes

When estimating the terms on the right-hand side of Bayes' rule, $P(H)$ is typically easy to estimate, $P(E)$ need not be estimated or can be computed after estimating the numerator $P(H)P(E|H)$, while $P(E|H)$ is usually the focus.

In practice, E can be high-dimensional $E = (E_1, E_2, \cdots, E_k)$. For example, in the application of image classification, E is an image with millions of pixels, and H is the class label. Estimating the likelihood is then nontrivial. A simple method, Naïve Bayes, makes the following conditional independence assumption to simplify the estimation:

$$P(E_1, E_2, \cdots, E_k|H) = P(E_1|H)P(E_2|H)\cdots P(E_k|H).$$

Note that this is an *assumption*, which may not be true in practice. However, it allows simplifying the estimation drastically and works well for many applications.