STAT 512: Statistical Inference **Autumn 2020**

Lecture 1: Introduction to Probability and Statistics

Instructor: Yen-Chi Chen

These notes are partially based on those of Michael Perlman.

Reference: Casella and Berger Chapter 1.

1.1 Sample Space and Probability Measure

The *sample space* Ω is the collection of all possible outcomes of a random experiment, e.g. toss of a coin, $\Omega = \{H, T\}$. Elements $\omega \in \Omega$ are called *outcomes*, *realizations* or *elements*. Subsets $A \subseteq \Omega$ are called *events*. You should able to express events of interest using the standard set operations. For instance:

- "Not A" corresponds to the *complement* $A^c = \Omega \setminus A$;
- "A or B" corresponds to the *union* $A \cup B$;
- "A and B" corresponds to the *intersection* $A \cap B$.

We said that $A_1, A_2, ...$ are *pairwise disjoint/mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i \neq j$. A *partition* of Ω is a sequence of pairwise disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = \Omega$. We use $|A|$ to denote the number of elements in A.

The sample space defines basic elements and operations of events. But it is still too simple to be useful in describing our senses of 'probability'. Now we introduce the concept of σ-algebra.

A *σ-algebra* $\mathcal{F}$ is a collection of subsets of Ω satisfying:

(A1) (full and null set) $\Omega \in \mathcal{F}$, $\emptyset \in \mathcal{F}$ ($\emptyset$ = empty set).

(A2) (complement)$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(A3) (countably union) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The sets in $\mathcal{F}$ are said to be *measurable* and $(\Omega, \mathcal{F})$ is a *measurable space*. The intuition of a set being measurable is that we can find a function that takes the elements of $\mathcal{F}$ and output a real number; this number represents the 'size' of the input element.

Now we introduce the concept of probability. Intuitively, probability should be associated with an event – when we say a probability of something, this 'something' is an event. Using the fact that the σ-algebra $\mathcal{F}$ is a collection of events and the property that $\mathcal{F}$ is measurable, we then introduce a measure called *probability measure* $\mathbb{P}(\cdot)$ that assigns a number between 0 and 1 to every element of $\mathcal{F}$. Namely, this function $\mathbb{P}$ maps an event to a number, describing the likelihood of the event.

Formally, a probability measure is a mapping $\mathbb{P} : \mathcal{F} \mapsto \mathbb{R}$ satisfying the following three axioms

(P1) $\mathbb{P}(\Omega) = 1$.

(P2) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.

(P3) (countably additivity) $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for *mutually exclusive* events $A_1, A_2, ... \in \mathcal{F}$.

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

The three axioms imply:

$$\begin{aligned}
&\mathbb{P}(\emptyset) = 0\\
&0 \leq \mathbb{P}(A) \leq 1\\
&A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B),\\
&\mathbb{P}(A^c) = 1 - \mathbb{P}(A)\\
&\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)
\end{aligned}$$

The countable additivity (P3) also implies that if a sequence of sets $A_1, A_2, \dots$ in $\mathcal{F}$ satisfying $A_n \subseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

If $A_n \supseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

How do we interpret the probability? There are two major views in statistics. The first view is called the frequentist view – the probability is interpreted as the limiting frequencies observed over repetitions in identical situations. The other view is called the Bayesian/subjective view where the probability quantifies personal belief. One way of assigning probabilities is the following. The probability of an event E is the price one is *just* willing to pay to enter a game in which one can win a unit amount of money if E is true. Example: If I believe a coin is fair and am to win 1 unit if a head arises, then I would pay $\frac{1}{2}$ unit of money to enter the bet.

1.2 Random Variable

So far, we have built a mathematical model describing the probability and events. However, in reality, we are dealing with numbers, which may not be directly link to events. We need another mathematical notion that bridges the events and numbers and this is why we need to introduce random variables.

Informally, a *random variable* is a mapping $X : \Omega \mapsto \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. Fo example, we toss a coin 2 times and let X represents the number of heads. The sample space is $\Omega = \{HH, HT, TH, TT\}$. Then for each $\omega \in \Omega$, $X(\omega)$ outputs a real number: $X(\{HH\}) = 2$, $X(\{HT\}) = X(\{TH\}) = 1$, and $X(\{TT\}) = 0$.

Rigorously, a function $X(\omega) : \Omega \to \mathbb{R}$ is called a *random variable* (R.V.) if $X(\omega)$ is measurable with respect to $\mathcal{F}$, i.e.

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

Note that the condition is also equivalent to saying that $X^{-1}(B) \in \mathcal{F}$ for every Borel set B[1]. This means that the set $X^{-1}(B)$ is indeed an event so that it makes sense to talk about $\mathbb{P}(X \in B)$, the probability that

[1] A Borel set is a set that can be formed by countable union/intersection and complement of open sets.

X lies in B, for any Borel set B. The function $B \mapsto \mathbb{P}(X \in B)$ is a probability measure and is called the *(probability) distribution* of X.

A very important characteristic of a random variable is its *cumulative distribution function (CDF)*, which is defined as

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

Actually, the distribution of X is completely determined by the CDF $F(x)$, regardless of X being a discrete random variable or a continuous random variable (or a mix of them).

When X takes discrete values, we may characterize its distribution using the probability mass function (PMF):

$$p(x) = P(X = x) = F(x) - F(x^-),$$

where $F(x^-) = \lim_{\epsilon \to 0} F(x - \epsilon)$. In this case, one can recover the CDF from PMF using $F(x) = \sum_{x' \leq x} p(x')$.

If X is an absolutely continuous random variable, we may describe its distribution using the probability density function (PDF):

$$p(x) = F'(x) = \frac{d}{dx} F(x).$$

In this case, the CDF can be written as

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

However, the PMF and PDF are not always well-defined. There are situations where X does not have a PMF or a PDF. The formal definition of PMF and PDF requires the notion of the Radon-Nikodym derivative, which is beyond the scope of this course.

Example 1 (discrete). Suppose X takes only three possible values: $1, 2, 3$, with equal probabilities. Then the PMF $p(x) = \frac{1}{3}$ for $x = 1, 2, 3$ and $p(x) = 0$ otherwise. The CDF will be

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{3}, & 1 \leq x < 2 \\ \frac{2}{3}, & 2 \leq x < 3 \\ 1, & x \geq 3 \end{cases} \tag{1.1}$$

Example 2 (continuous). Consider a random variable X that has a uniform PDF over the interval $[1, 10]$. Namely, there is a constant c such that $p(x) = c$ for $x \in [1, 10]$ and $p(x) = 0$ otherwise. What will c be? Using the fact that $1 = \int p(x)dx$, you can easily see that $c = \frac{1}{9}$. What will the CDF be in this case? By definition,

$$F(x) = \int_{-\infty}^{x} p(u) du = \begin{cases} 0, & x \leq 1 \\ \frac{x-1}{9}, & 1 < x \leq 10 \\ 1, & x > 10. \end{cases}$$

Example 3 (no PDF or PMF). Consider a random variable X such that with a probability of 0.5, it always takes a fixed value 2 and with a probability of 0.5, it is from a uniform PDF over $[0, 1]$. In this case, can we define PDF or PMF? It turns out that this random variable X does not have a PDF (since it has a point mass at $x = 2$) but it has a strange PMF (that takes a value of 0 except at $x = 2$). So we cannot characterize its distribution well using the PDF or PMF. However, using the definition of CDF $F(x) = P(X \leq x)$, you can easily see that it has a well-defined CDF:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x \leq 0 \\ \frac{x}{2}, & 0 < x \leq 1 \\ \frac{1}{2}, & 1 < x < 2 \\ 1, & x \geq 2. \end{cases}$$

If you take a more advanced probability theory course, you will find that the CDF is the formal definition of a distribution function–it is always well-defined unlike the PMF or PDF.

1.3 Common Distributions

1.3.1 Discrete Random Variables

Bernoulli. If X is a Bernoulli random variable with parameter p, then $X = 0$ or, 1 such that

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

In this case, we write $X \sim \mathsf{Ber}(p)$.

Binomial. If X is a binomial random variable with parameter (n, p), then $X = 0, 1, \cdots, n$ such that

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

In this case, we write $X \sim \mathsf{Bin}(n, p)$. Note that if $X_1, \cdots, X_n \sim \mathsf{Ber}(p)$, then the sum $S_n = X_1 + X_2 + \cdots + X_n$ is a binomial random variable with parameter (n, p).

Geometric. If X is a geometric random variable with parameter p, then

$$P(X = n) = (1 - p)^{n-1} p$$

for $n = 1, 2, \cdots$. Geometric random variable can be constructed using 'the number of trials of the first success occurs'. Consider the case we are flipping coin with a probability p that we gets a head (this is a Bernoulli (p) random variable). Then the number of trials we made to see the first head is a geometric random variable with parameter p.

Poisson. If X is a Poisson random variable with parameter λ, then $X = 0, 1, 2, 3, \cdots$ and

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

In this case, we write $X \sim \mathsf{Poi}(\lambda)$. Poisson is often used to model a counting process. For instance, the intensity of an image is commonly modeled as a Poisson random variable.

1.3.2 Continuous Random Variables

Uniform. If X is a uniform random variable over the interval $[a, b]$, then

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

where $I(\mathsf{statement})$ is the indicator function such that if the $\mathsf{statement}$ is true, then it outputs 1 otherwise 0. Namely, $p(x)$ takes value $\frac{1}{b-a}$ when $x \in [a,b]$ and $p(x) = 0$ in other regions. In this case, we write $X \sim \mathsf{Uni}[a,b]$.

Normal. If X is a normal random variable with parameter (μ, σ^2), then

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In this case, we write $X \sim N(\mu, \sigma^2)$.

Exponential. If X is an exponential random variable with parameter λ, then X takes values in $[0, \infty)$ and

$$p(x) = \lambda e^{-\lambda x}.$$

In this case, we write $X \sim \mathsf{Exp}(\lambda)$. Note that we can also write

$$p(x) = \lambda e^{-\lambda x} I(x \geq 0).$$

A *double exponential* random variable X is that $|X|$ is an exponential random variable. So its PDF will be

$$p(x) = \frac{\lambda}{2} e^{-\lambda |x|}.$$

Cauchy. If $X \in \mathbb{R}$ is a Cauchy random variable with parameter μ, σ^2, then it has a PDF

$$p(x) = \frac{1}{\pi\sigma} \frac{1}{1 + (x-\mu)^2/\sigma^2}.$$

Interesting fact: the Cauchy distribution has *no* mean (average); the parameter μ is the median.

Gamma. A Gamma random variable $X \geq 0$ has two parameters $\alpha, \lambda > 0$ and has a PDF

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I(x \geq 0).$$

The function $\Gamma(\alpha) = \int x^{\alpha-1} e^{-x} dx$ is known as the Gamma function.

Beta. The Beta distribution is a continuous distribution on $[0, 1]$. So it is often used to model a ratio or a probability. If X is a Beta random variable with parameter α, β, then

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} I(0 \leq x \leq 1),$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Logistic. The logistic distribution is a random variable whose CDF follows from the logit function. It has two parameter $\alpha \in \mathbb{R}, \beta > 0$ and has a CDF

$$F(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} = \frac{1}{1 + e^{-\alpha-\beta x}}.$$

The PDF is

$$p(x) = \frac{\beta e^{-\alpha-\beta x}}{(1 + e^{-\alpha-\beta x})^2} = \frac{\beta e^{\alpha+\beta x}}{(1 + e^{\alpha+\beta x})^2}$$

1.4 Conditional Probability

Now we have a basic mathematical model for probability. This model also defines an interesting quantity called conditional probability. For two events $A, B \in \mathcal{F}$, the conditional probability of A given B is

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

Note that when B is fixed, the function $\mathbb{P}(\cdot|B) : \mathcal{F} \mapsto \mathbb{R}$ is another probability measure.

In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. This is sometimes called as the prosecutor's fallacy:

$$\mathbb{P}(\text{evidence}|\text{guilty}) \neq \mathbb{P}(\text{guilty}|\text{evidence}).$$

Example (Exponential). Let X be an exponential random variable with parameter $\lambda > 0$ and consider two positive numbers $x, y > 0$. What is the probability $P(X > x + y|X > y)$? In this case the two events $A = \{X > x + y\}$ (formally, $A = \{\omega : X(\omega) > x + y\}$) and $B = \{X > y\}$. It is easy to see that $A \subset B$ so $A \cap B = A$. Thus,

$$P(X > x + y|X > y) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{P(X > x + y)}{P(X > y)}.$$

It is easy to see that for an exponential RV X, $P(X > y) = e^{-\lambda y}$, which implies

$$P(X > x + y|X > y) = \frac{P(X > x + y)}{P(X > y)} = e^{-\lambda x} = P(X > x).$$

Thus, the probability only depends on the *increment* x, not y. This is known as the *memoryless* property.

1.5 Conditional Distribution

For two random variables X, Y, the joint CDF is

$$P_{XY}(x, y) = F(x, y) = P(X \leq x, Y \leq y).$$

When both variables are absolute continuous, the corresponding joint PDF is

$$p_{XY}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

The *conditional PDF* of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy$ is sometimes called the marginal density function.

When both X and Y are discrete, the joint PMF is

$$p_{XY}(x, y) = P(X = x, Y = y)$$

and the conditional PMF of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

Example (triangle uniform). Consider two random variables (X, Y) that has a uniform PDF over the region $D = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$. It is easy to see that $p(x, y) = 2$ when $(x, y) \in D$ and 0 otherwise. What is the conditional PDF of $p_{Y|X}(y|x)$? Because the joint PDF is a constant, one can easily see that $p_{Y|X}(y|x)$ will also be a constant. The key is to identify what is the feasible range of y when $X = x$. We have two constraints $y \geq 0$ and $y \leq 1 - x$ so the feasible range of y is $[0, 1 - x]$. Thus,

$$p_{Y|X}(y|x) = \frac{1}{1-x} I(0 \leq y \leq 1 - x).$$

Example (Beta-Bernoulli). Consider two random variables $X \in \{0, 1\}$ and $Y \in [0, 1]$ such that given Y, the random variable X is a Bernoulli random variable with parameter $p = Y$. Namely,

$$P(X = 1|Y) = Y, \quad P(X = 0|Y) = 1 - Y.$$

Also, assume that Y follows from a Beta distribution with parameter α, β. We are interested in the conditional distribution of Y given $X = x$. The conditional PDF

$$p(x|y) = y^x(1 - y)^{1-x}.$$

Thus, the joint PDF/PMF

$$p(x, y) = p(x|y)p(y) = y^x(1 - y)^{1-x} \cdot \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1 - y)^{\beta-1}.$$

Here is a trick: because $p(y|x) = \frac{p(x,y)}{p(x)} \propto p(x, y)$, we only need to focus on the part of $p(x, y)$ that involves y. The above product shows that

$$p(y|x) \propto p(x, y) \propto y^{\alpha+x-1}(1 - y)^{\beta-x}$$

which is proportional to the PDF of a Beta distribution with parameter $(\alpha' = \alpha + x, \beta' = \beta + (1 - x))$. Namely, when $X = 1$, we will increase the α parameter by 1 while keeping β parameter the same. When $X = 0$, we will increase β by 1 and keep the same α. You can think about what will be the conditional distribution of $X|Y = y$.

The Beta-Bernoulli example illustrates the fact that the conditioning operation can be viewed as an information flow. Suppose that Y is a variable of interest that is unobserved but it implicitly determines the distribution of X. And X is something that we can measure/observe (think of it as the data). Before seeing X, we place a model that Y is from a Beta distribution with parameter α, β. After observing X, this information should improve our knowledge about Y. A simple mathematical model to describe the improvement from the information is the conditional distribution. As is shown in the above example, the conditional distribution of Y given X is a Beta distribution with parameter $\alpha + X, \beta + (1 - X)$. The change of parameter is an example of how the observed data X improves our understanding of an unobserved quantity Y.

1.6 Independence

The probability has a power feature called *independence*. This property is probably the key property that makes the 'probability theory' distinct from measure theory. Intuitively, when we say that two events are independent, we refers to the case that the two event will not interfere each other. Two events A and B are independent if

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \text{(or equivalently, } \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)).$$

For three events A, B, C, we say events A and B are *conditional independent* given C if

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

Random variables X and Y are *independent* if the joint CDF can be factorized as

$$F(x, y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

When X and Y are independent, we often write $X \perp Y$. A direct result from the factorization of CDF is that the PDF/PMF will also be factorized under independence:

$$p(x, y) = p(x)p(y).$$

This implies that the conditional PDF

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p(x)} = p(y).$$

Now we use the information interpretation (in the Beta-Bernoulli example) to think of the independence. The independence implies $p_{Y|X}(y|x) = p(y)$, which can be interpreted that *knowing X does NOT change the distribution of Y*. This is essentially what an intuitive meaning of independence should be–knowing the outcome of one variable does not provide any information about another variable.

When we have many random variables $X_1, \cdots, X_n$, they are (mutually) independent if the joint CDF

$$F(x_1, x_2, \cdots, x_n) = F(x_1)F(x_2)\cdots F(x_2),$$

which implies

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n).$$

Example (Uniform on a disk). Consider two random variables X, Y such that they jointly follow from a distribution that is uniform over the unit disk $S_0 = \{(x, y) : x^2 + y^2 \leq 1\}$. Clearly, X and Y are not independent because when $X = 0$, the feasible range of Y is $[-1, 1]$ while when $X = 1$, the only possible value of Y is 0. Now suppose we reparametrize the two random variable using polar coordinate $(R, \Theta) \in [0, 1] \times [0, 2\pi]$. Then

$$\begin{aligned} F(r, \theta) &= P(R \leq r, \Theta \leq \theta) \\ &= \frac{1}{\pi} \cdot \pi r^2 \cdot \frac{\theta}{2\pi} \\ &= r^2 \cdot \frac{\theta}{2\pi} \\ &= F_R(r)F_\Theta(\theta), \\ F_R(r) &= r^2, \quad 0 \leq r \leq 1 \\ F_\Theta(\theta) &= \frac{\theta}{2\pi}, \quad 0 \leq \theta \leq 2\pi. \end{aligned}$$

So $R \perp \Theta$, i.e., they are independent.

Example (Independence and information). In the Beta-Bernoulli example, we have seen a probabilistic approach to infer an unobserved variable Y using the information from another random variable X. That idea is something related to the so-called *Bayesian inference*. Here we will introduce another approach to infer an unobserved quantity θ without assuming that θ is random. Suppose we observe $X_1, \cdots, X_n \in \{0, 1\}$ that are independent. We assume that they are all from the same Bernoulli distribution with an unknown parameter

$\theta_0 = P(X_i = 1)$. In this case, we say $X_1, \cdots, X_n$ are IID (independently and identically distributed). Given $X_1, \cdots, X_n$, how do we infer θ_0? Under a probabilistic model, any parameter θ would implies a joint PMF

$$p(x_1, \cdots, x_n; \theta) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta)$$

due to the independence. Since it is a product term, we take a logarithm, which leads to

$$\log p(x_1, \cdots, x_n; \theta) = \log p(x_1; \theta) + \log p(x_2; \theta) + \cdots + \log p(x_n; \theta).$$

Since $X_1, \cdots, X_n$ are observed, we can viewed the above function as a function of θ, and this function is known as the log-likelihood function

$$\underbrace{\ell(\theta|X_1, \cdots, X_n)}_{\text{Total information}} = \log p(X_1, \cdots, X_n; \theta) = \sum_{i=1}^n \log p(X_i; \theta) = \sum_{i=1}^n \underbrace{\ell(\theta|X_i)}_{\text{Information of the } i\text{-th obs.}}.$$

Informally, we can call $\ell(\theta|X_1, \cdots, X_n)$ as the total information from $X_1, \cdots, X_n$ on θ. The independence assumption implies the above equality, which means that *under independence, the total information is the addition of all individual information.* In the *likelihood framework*, information about θ is determined by the log-likelihood function (Total information term). Note that unlike the Beta-Bernoulli example, here we did not specify any distribution of θ–it is just an unknown quantity and we use the likelihood function to infer plausible value of it. The famous *maximal likelihood estimator* (MLE) finds an estimated value of θ by maximizing the log-likelihood value.

1.7 Total probability and Bayes theorem

Probability measure also has a useful property called *law of total probability.* If $B_1, B_2, ..., B_k$ forms a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^k \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

In particular, $\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$. And this further implies the famous *Bayes rule*: Let $A_1, ..., A_k$ be a partition of Ω. If $\mathbb{P}(B) > 0$ then, for $i = 1, ..., k$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^k \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

For random variables, we also have the Bayes theorem:

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{XY}(x,y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \\ &= \begin{cases} \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x')p_X(x')dx'}, & \text{if } X, Y \text{ are absolutely continuous.} \\ \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}, & \text{if } X, Y \text{ are discrete.} \end{cases} \end{aligned}$$

Example (Poisson-Binomial). Consider two random variables X and Y such that $X \sim \mathsf{Poisson}(\lambda)$ and $Y|X = x$ is from a Binomial distribution with parameters (X, p). What will the marginal distribution of Y

be? To study this, we attempt to compute the probability $P(Y=y)$.

$$\begin{aligned} P(Y=y) &= \sum_x P(Y=y, X=x) \\ &= \sum_{x\geq y} P(Y=y|X=x)P(X=x) \quad \text{(Think about why } x\geq y) \\ &= \sum_{x\geq y} \binom{x}{y} p^y(1-p)^{x-y}\frac{\lambda^x e^{-\lambda}}{x!}. \end{aligned}$$

Using the fact that $\binom{x}{y} = \frac{x!}{(x-y)!y!}$ and set $k = x-y$, we can rewrite the above as

$$\begin{aligned} P(Y=y) &= \sum_{x\geq y} \binom{x}{y} p^y(1-p)^{x-y}\frac{\lambda^x e^{-\lambda}}{x!} \\ &= p^y e^{-\lambda}\sum_{x\geq y}\frac{x!}{(x-y)!y!}(1-p)^{x-y}\lambda^x\frac{1}{x!} \\ &= \frac{p^y e^{-\lambda}}{y!}\sum_{k=0}^{\infty}\frac{1}{k!}(1-p)^k\lambda^{y+k} \\ &= \frac{(\lambda p)^y e^{-\lambda p}}{y!}\underbrace{\sum_{k=0}^{\infty}\frac{1}{k!}(1-p)^k\lambda^k e^{-\lambda(1-p)}}_{=1}, \end{aligned}$$

which is the PMF of Poisson(λp). Thus, Y follows from a Poisson distribution with parameter λp.

1.8 Conditional independence

For three RVs X, Y, and Z, we say X, Y are conditional independent given Z if

$$P(X\leq x, Y\leq y|Z=z) = P(X\leq x|Z=z)P(Y\leq y|Z=z)$$

for every x and y and P_Z-almost everywhere of z. P_Z-almost everywhere of z means that the above equality holds for all z except for a set of values that has 0 probability. It is a slightly weaker notion than 'for every z'. We use the notation

$$X \perp Y|Z$$

for denote the case where X, Y are conditional independent given Z.

Note that $X\perp Y|Z$ also implies

$$P(X\leq x|Y=y, Z=z) = P(X\leq x|Z=z)$$

for every x and $P_{Y,Z}$-almost everywhere of (y,z).

Beware! Independence is not the same as conditional independence, i.e., $X\perp Y \not\Leftrightarrow X\perp Y|Z$.

Example (conditional independence $\not\Rightarrow$ indepedence). Assume $X\perp Y|Z$ and $Z\in\{0,1\}$ such that when $Z=0$, X and Y are both from a uniform distribution over $[0,1]$ and when $Z=1$, X and Y are from a uniform distribution over $[2,3]$. Assume that Z has an equal probability of being 0 or 1. Marginally, both X and Y are from a uniform distribution over the set $[0,1]\cup[2,3]$. However, if we observe $X=2.5$, we know that Y has to be from a uniform distribution over $[2,3]$ so $P(Y\in[0,1]|X=2.5)=0\neq P(Y\in[0,1])=0.5$.

The following is a theorem about different ways of saying *conditional independence*.

Theorem 1.1 *Let p_{XYZ} be the joint PDF/PMF of X, Y, and Z. Then the followings are equivalent:*

(i) $X \perp Y|Z$.

(ii) $p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$ *a.e.*

(iii) $p_{X|YZ}(x|y,z) = p_{X|Z}(x|z)$ *a.e.*

(iv) $p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)}$ *a.e.*

(v) $p_{XYZ}(x,y,z) = g(x,z)h(y,z)$*, where* g *and* h *are some (measurable) functions.*

(vi) $p_{X|YZ}(x|y,z) = w(x,z)$*, where* w *is some (measurable) function.*

Proof: The equivalence between (i), (ii), (iii), and (iv) are trivial so we focus on case (v) and (vi).

(ii) ⇒ (v):
Because

$$p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z),$$

we have

$$\frac{p_{XYZ}(x,y,z)}{p_Z(z)} = \frac{p_{XZ}(x,z)}{p_Z(z)}\frac{p_{YZ}(y,z)}{p_Z(z)}$$

so

$$p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)} = h(x,z)g(y,z),$$

which proves (v).

(v) ⇒ (vi):
Based on (v), we have

$$p_{YZ}(y,z) = \int p_{XYZ}(x,y,z)dx = h(y,z)\int g(x,z)dx = h(y,z)q(z).$$

Thus,

$$p_{X|YZ}(x|y,z) = \frac{p_{XYZ}(x,y,z)}{p_{YZ}(y,z)} = \frac{g(x,z)h(y,z)}{h(y,z)q(z)} = \frac{g(x,z)}{q(z)} = w(x,z).$$

Finally, we show that (vi) ⇒ (iii):

$$\begin{aligned} p_{X|Z}(x|z) = \int p_{XY|Z}(x,y|z)dy &= \int p_{X|YZ}(x|y,z)p_{Y|Z}(y|z)dy \\ &= w(x,z)\int p_{Y|Z}(y|z)dy = w(x,z) = p_{X|YZ}(x|y,z). \end{aligned}$$

■

Here are five important properties of conditional independence. Let X, Y, Z, W be RVs.

(C1) (symmetry) $X \perp Y|Z \Longleftrightarrow Y \perp X|Z$.

(C2) (decomposition) $X \perp Y|Z \Longrightarrow h(X) \perp Y|Z$ for any (measurable) function h.
A special case is: $(X, W) \perp Y|Z \Longrightarrow X \perp Y|Z$.

(C3) (weak union) $X \perp Y|Z \Longrightarrow X \perp Y|Z, h(X)$ for any (measurable) function h.
A special case is: $(X, W) \perp Y|Z \Longrightarrow X \perp Y|(Z, W)$

(C4) (contraction)

$$X \perp Y|Z \text{ and } X \perp W|(Y, Z) \Longleftrightarrow X \perp (W, Y)|Z.$$

(C5) If the joint PDF $p_{XYZW}(x, y, z, w)$ satisfies $f_{YZW}(y, z, w) > 0$ almost everywhere. Then

$$X \perp Y|(W, Z) \text{ and } X \perp W|(Y, Z) \Longleftrightarrow X \perp (W, Y)|Z.$$