Lecture 19

§1 Properties of conditional distribution

Conditional pmf/pdf 满足 pmf/pdf 的所有性质，描述了给定一个随机变量的值，另一个随机变量的 probabilistic behavior.

1. Definition: Conditional distribution function (条件分布函数)

Conditional distribution function of Y given $X=x$ 被定义为

$$F_{Y|X}(y|x) = \Pr(Y \le y \mid X=x)$$

$$= \begin{cases} \sum_{i \le y} P_{Y|X}(i|x), & \text{for discrete case} \\ \int_{-\infty}^{y} f_{Y|X}(t|x)\,dt, & \text{for continuous case} \end{cases}$$

2. Definition: Conditional expectation (条件期望)

Conditional expectation of $g(Y)$ given $X=x$ 被定义为

$$E[g(Y)|X=x] = \begin{cases} \sum_i g(i) P_{Y|X}(i|x), & \text{for discrete case} \\ \int_{-\infty}^{\infty} g(t) f_{Y|X}(t|x)\,dt, & \text{for continuous case} \end{cases}$$

Conditional expectation of Y given $X=x$ 被定义为

$$E(Y|X=x)$$

3. Definition: Conditional variance (条件方差)

Conditional variance of Y given $X=x$ 被定义为

$$Var(Y|X=x) = E\{[Y-E(Y|X=x)]^2\}$$
$$= E(Y^2|X=x) - [E(Y|X=x)]^2$$

e.g. **Example 8.3.**

Suppose that the joint density of X and Y is given by

$$f(x,y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y}, & x>0, y>0; \\ 0, & \text{otherwise.} \end{cases}$$

The marginal pdf of Y is

$$f_Y(y) = \int_0^{\infty} \frac{e^{-x/y}e^{-y}}{y} \mathrm{d}x = e^{-y}\left[e^{-x/y}\right]_{\infty}^{0} = e^{-y}, \qquad y>0.$$

Hence, $Y \sim \mathrm{Exp}(1)$.

The conditional pdf of $X|(Y=y)$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{y}e^{-x/y}, \qquad x>0.$$

Hence, the conditional distribution of X given $Y=y$ is exponential with parameter $\lambda = 1/y$, or we may write

$$X|Y \sim \mathrm{Exp}(Y^{-1}), \qquad \text{or} \qquad X|(Y=y) \sim \mathrm{Exp}(y^{-1}).$$

The conditional distribution function of $X|(Y=y)$ is

$$F_{X|Y}(x|y) = \begin{cases} 0, & x \le 0; \\ 1-e^{-x/y}, & x>0. \end{cases}$$

Also, the conditional mean and variance can be determined easily as

$$\mathrm{E}(X|Y) = Y, \qquad \mathrm{Var}(X|Y) = Y^2.$$

Therefore $\mathrm{E}(X|Y)$ and $\mathrm{Var}(X|Y)$ are random variables.

Thought Question:

What is $\mathrm{E}[\mathrm{E}(X|Y)]$?

§2 Computing expectations by conditioning

1. **Theorem:** Law of total expection / Adam's law / Double expectation formula（全期望公式）

若 X 与 Y 为两个随机变量，则对任意函数 u，有

$$E[u(X)] = E\{E[u(X)|Y]\}$$

特别的，若 u 为 identity function，则

$$E(X) = E[E(X|Y)]$$

证明：

$$
\begin{aligned}
E\{E[u(X)|Y]\} &= \int_{-\infty}^{\infty} E[u(X)|Y=y]\cdot f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} u(x)\cdot f_{X|Y}(x|y)\,dx\right\} f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x) f(x,y)\,dx\,dy \\
&= E[u(X)]
\end{aligned}
$$

注：① $E(X|Y=y)$ 的另一种求解方法：

将 $X|Y=y$ 视作全期望公式中的 X，则有

$$E(X|Y=y) = E(E(X|Y=y,Z)) = \begin{cases} \sum_z E(X|Y=y,Z=z)\cdot P(Z=z) \\ \int_{-\infty}^{\infty} E(X|Y=y,Z=z)\,dF_Z(z) \end{cases}$$

② 对全期望公式背后的思想：先局部平均，再整体平均

2. **Theorem:** Law of total variance / Eve's law（全方差公式）

若 X 与 Y 为两个随机变量，则有

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

证明：

$$
\begin{aligned}
E[Var(X|Y)] &= E\{E(X^2|Y) - [E(X|Y)]^2\} \\
&= E(X^2) - E\{[E(X|Y)]^2\} \\
Var[E(X|Y)] &= E\{[E(X|Y)]^2\} - \{E[E(X|Y)]\}^2 \\
&= E(X^2) - E[Var(X|Y)] - [E(X)]^2 \\
&= Var(X) - E[Var(X|Y)]
\end{aligned}
$$

$$\therefore Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

注：$E[Var(X|Y)]$ 是每个划分下方差的均值，刻画了样本内的差异程度

$Var[E(X|Y)]$ 是不同分组下均值的方差，刻画了样本间的差异程度

因此方差刻画了样本内和样本间差异的叠加。

3. Bayesian inference on distribution

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\cdot f_X(x)}{f_Y(y)}$$

其中，$f_X(x)$ 为 prior distribution

$f_{X|Y}(x|y)$ 为 posterior distribution
$f_Y(y)$ 为 unconditional distribution
$f_{Y|X}(y|x)$ 为 conditional distribution

e.g. **Example 8.4.**

In **Example 8.3.**, the marginal pdf of Y is

$$f_Y(y) = e^{-y}, \qquad y > 0.$$

It can be easily verified that $\mathrm{E}(Y) = 1$ and $\mathrm{Var}(Y) = 1$.
Also recall that

$$X|Y \sim \mathrm{Exp}(Y^{-1}),$$

so the conditional mean and variance are respectively,

$$\mathrm{E}(X|Y) = Y, \qquad \mathrm{Var}(X|Y) = Y^2.$$

Therefore,

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}\left[\mathrm{E}(X|Y)\right] = \mathrm{E}(Y) = 1, \\
\mathrm{E}\left[\mathrm{Var}(X|Y)\right] &= \mathrm{E}(Y^2) = \mathrm{Var}(Y) + \left[\mathrm{E}(Y)\right]^2 = 1 + 1^2 = 2, \\
\mathrm{Var}\left[\mathrm{E}(X|Y)\right] &= \mathrm{Var}(Y) = 1, \\
\mathrm{Var}(X) &= \mathrm{E}\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left[\mathrm{E}(X|Y)\right] = 2 + 1 = 3.
\end{aligned}$$

Note that directly calculation of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ from $f(x, y)$ may be difficult as there is no closed form expression for the marginal pdf of X:

$$f_X(x) = \int_0^\infty \frac{e^{-x/y} e^{-y}}{y} \mathrm{d}y.$$

★

e.g. **Example 8.5.**

Suppose we have a binomial random variable X which represents the number of success in n independent Bernoulli experiments. Sometimes the success probability p is unknown. However, we usually have some understanding on the value of p, e.g., we may believe that p is a realization of another random variable P picked uniformly from $(0, 1)$, i.e., $P \sim \mathrm{U}(0, 1)$. Then we have the following *hierarchical model*:

$$P \sim \mathrm{U}(0, 1), \qquad X|P \sim \mathrm{B}(n, P) \text{ or } X|(P = p) \sim \mathrm{B}(n, p).$$

Using the formulae of expectation by conditioning, we have

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}\left[\mathrm{E}(X|P)\right] = \mathrm{E}(nP) = n\mathrm{E}(P) = \frac{n}{2}, \\
\mathrm{Var}(X) &= \mathrm{E}\left[\mathrm{Var}(X|P)\right] + \mathrm{Var}\left[\mathrm{E}(X|P)\right] \\
&= \mathrm{E}\left[nP(1-P)\right] + \mathrm{Var}\left[nP\right] \\
&= n\mathrm{E}(P) - n\mathrm{E}(P^2) + n^2\mathrm{Var}(P) \\
&= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12} \\
&= \frac{n(n+2)}{12}.
\end{aligned}$$

To find the marginal pmf of X, $p_X(x) = \Pr(X = x)$, we can let

$$\mathbf{1}_{\{X=x\}} = \begin{cases} 1, & \text{if } X = x; \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned}
\Pr(X = x) &= \mathrm{E}(\mathbf{1}_{\{X=x\}}) \\
&= \mathrm{E}\left[\mathrm{E}(\mathbf{1}_{\{X=x\}}|P)\right] \\
&= \mathrm{E}\left[\Pr(X = x|P)\right] \\
&= \mathrm{E}\left[\binom{n}{x} P^x (1-P)^{n-x}\right] \\
&= \binom{n}{x} \int_0^1 p^x (1-p)^{n-x} (1) \mathrm{d}p \qquad \because f_P(p) = 1 \text{ for } 0 < p < 1 \\
&= \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \\
&= \binom{n}{x} \frac{x!(n-x)!}{(n+1)!} \\
&= \frac{1}{n+1}, \qquad x = 0, 1, 2, \ldots, n.
\end{aligned}$$

Hence, X is distributed as discrete uniform distribution with support $\{0, 1, 2, ..., n\}$.

Using the Bayes' theorem, the conditional pdf of P given $X = x$ is given by

$$\begin{aligned} f_{P|X}(p|x) &= \frac{p_{X|P}(x|p) f_P(p)}{p_X(x)} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \times 1 \Big/ \left(\frac{1}{n+1}\right) \\ &= \frac{(n+1)!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^{(x+1)-1} (1-p)^{(n-x+1)-1}, \qquad 0 < p < 1. \end{aligned}$$

Therefore, $P|(X = x) \sim \text{Beta}(x + 1, n - x + 1)$ and

$$\mathrm{E}(P|X = x) = \frac{x+1}{(x+1) + (n-x+1)} = \frac{x+1}{n+2}.$$

In layman's terms, suppose an event happens with an unknown probability P where the values of P are equally likely between 0 and 1, then when x out of n cases of the event were observed, an appropriate estimate of P is

$$\hat{P} = \frac{x+1}{n+2}.$$

This formula is known as the *Laplace's law of succession* in the 18th century by Pierre-Simon Laplace in the course of treating the *sunrise problem* which tried to answer the question "What is the probability that the sun will rise tomorrow?"

★

e.g. **Example 8.6.**

Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ be two independent geometric random variables. Find the expected value of the proportion $\dfrac{X}{X+Y}$.

Solution:

Note that $X, Y \in \{1, 2, \ldots\}$. Let $N = X + Y \in \{2, 3, \ldots\}$. The possible values of X given $N = n$ should be from 1 to $n - 1$. Similar to **Example 8.2.**, we first note that $N \sim \text{NB}(2, p)$ as its mgf is

$$M_N(t) = M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{pe^t}{1-(1-p)e^t} \cdot \frac{pe^t}{1-(1-p)e^t} = \left[\frac{pe^t}{1-(1-p)e^t}\right]^2 \quad \text{for } t < -\ln(1-p).$$

Then, for $k = 1, 2, \ldots, n - 1$,

$$\begin{aligned} \Pr(X = k|N = n) &= \Pr(X = k|X + Y = n) = \frac{\Pr(X = k, X + Y = n)}{\Pr(X + Y = n)} = \frac{\Pr(X = k, Y = n - k)}{\Pr(X + Y = n)} \\ &= \frac{\Pr(X = k)\Pr(Y = n - k)}{\Pr(X + Y = n)} = \frac{(1-p)^{k-1} p \cdot (1-p)^{n-k-1} p}{\binom{n-1}{2-1} p^2 (1-p)^{n-2}} \qquad \because X + Y \sim \text{NB}(2, p) \\ &= \frac{1}{n-1}. \end{aligned}$$

That is, $X|(N = n) \sim \text{DU}\{1, 2, \ldots, n - 1\}$ as a discrete uniform distribution.

The conditional mean of X given $N = n$ is

$$\mathrm{E}(X|N = n) = \frac{1 + 2 + \cdots + (n-1)}{n-1} = \frac{\frac{1}{2}(n-1)(1+n-1)}{n-1} = \frac{n}{2}.$$

Thus, $\mathrm{E}\left(\dfrac{X}{X+Y}\right) = \mathrm{E}\left(\dfrac{X}{N}\right) = \mathrm{E}\left[\mathrm{E}\left(\dfrac{X}{N}\middle| N\right)\right] = \mathrm{E}\left[\dfrac{1}{N}\mathrm{E}(X|N)\right] = \mathrm{E}\left(\dfrac{1}{N} \cdot \dfrac{N}{2}\right) = \dfrac{1}{2}.$

★

e.g. **Example 8.7.** (Prediction of Y from X)

When X and Y are not independent, we can base on the observed value of X to predict the value of the unobserved random variable Y. That is, we may predict the value of Y by $g(X)$ where g is a function chosen in such a way that the mean squared error (MSE) of the prediction, $Q = \mathrm{E}\left\{[Y - g(X)]^2\right\}$, is minimized.

First we conditional on X, consider

$$\begin{aligned} \mathrm{E}\left\{[Y - g(X)]^2 |X\right\} &= \mathrm{E}(Y^2|X) - 2g(X)\mathrm{E}(Y|X) + [g(X)]^2 \\ &= \mathrm{Var}(Y|X) + [\mathrm{E}(Y|X)]^2 - 2g(X)\mathrm{E}(Y|X) + [g(X)]^2 \\ &= \mathrm{Var}(Y|X) + [g(X) - \mathrm{E}(Y|X)]^2. \end{aligned}$$

Hence, $Q = \mathrm{E}\left\{\mathrm{E}\left\{[Y - g(X)]^2 |X\right\}\right\} = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{E}\left\{[g(X) - \mathrm{E}(Y|X)]^2\right\}.$

Therefore Q is minimized if we choose $g(x) = \mathrm{E}(Y|X = x)$, i.e., the best predictor of Y given the value of X is $g(X) = \mathrm{E}(Y|X)$. The mean squared error of this predictor is

$$\mathrm{E}\left\{[Y - \mathrm{E}(Y|X)]^2\right\} = \mathrm{E}[\mathrm{Var}(Y|X)] = \mathrm{Var}(Y) - \mathrm{Var}\left[\mathrm{E}(Y|X)\right] \le \mathrm{Var}(Y).$$

★

e.g. **Example 8.8.**

Two random variables X and Y are said to have a *bivariate normal distribution* if their joint pdf is

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$, where μ_X and σ_X^2 are the mean and variance of X; μ_Y and σ_Y^2 are the mean and variance of Y; ρ is the correlation coefficient between X and Y. The distribution is denoted as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}_2\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right].$$

Consider the marginal pdf of X.

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y \\
&= C(x)\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_Y^2}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right] \exp\left[\frac{\rho(x-\mu_X)(y-\mu_Y)}{(1-\rho^2)\sigma_X\sigma_Y}\right]\mathrm{d}y \\
&\qquad\qquad \text{where } C(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right], \\
&= C(x)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)\exp\left[\frac{\rho(x-\mu_X)}{\sigma_X\sqrt{1-\rho^2}}z\right]\mathrm{d}z \qquad \text{by letting } z = \frac{y-\mu_Y}{\sigma_Y\sqrt{1-\rho^2}}, \\
&= C(x)M_Z\left(\frac{\rho(x-\mu_X)}{\sigma_X\sqrt{1-\rho^2}}\right) \qquad \text{where } M_Z(t) \text{ is the mgf of } Z \sim \mathrm{N}(0,1), \\
&= \frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]\exp\left[\frac{1}{2}\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2(1-\rho^2)}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right], \qquad -\infty < x < \infty.
\end{aligned}$$

Thus, the marginal distribution of X is $\mathrm{N}(\mu_X, \sigma_X^2)$. The conditional pdf of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2\right\}, \ -\infty < y < \infty.$$

Hence, $Y|X \sim \mathrm{N}\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X-\mu_X),\ (1-\rho^2)\sigma_Y^2\right)$.

The best predictor of Y given the value of X is

$$\mathrm{E}(Y|X) = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X-\mu_X) = \left(\mu_Y - \frac{\rho\sigma_Y}{\sigma_X}\mu_X\right) + \frac{\rho\sigma_Y}{\sigma_X}X = \alpha + \beta X.$$

This is called the *linear regression* of Y on X. (Note that $\beta = \frac{\rho\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2}$ and $\alpha = \mu_Y - \beta\mu_X$.)

★