STAT 512: Statistical Inference **Autumn 2020**

Lecture 4: Conditional expectation and conditional distribution

Instructor: Yen-Chi Chen

4.1 Review: conditional distribution

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

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

In the first lecture, we have introduced the conditional distribution when both variables are continuous or discrete.

X, Y **continuous:** 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.

X, Y **discrete:** 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 $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

However, in reality, it could happen that one of them is continuous and one of them is discrete. So we have to be careful in defining conditional distribution in this case. When the variable being conditioned is a discrete, the problem is rather simple.

X **discrete,** Y **continuous:** Suppose we are interested in $Y|X = x$ where Y is continuous and X is discrete. In this case the RV $Y|X = x$ is still a continuous random variable and its PDF will be

$$p_{Y|X}(y|x) = \frac{d}{dy}P(Y \leq y|X = x) = \frac{\frac{d}{dy}P(Y \leq y, X = x)}{P(X = x)}.$$

Since x is discrete, $P(X = x) = p_X(x)$ is its PMF. Thus, we obtain

$$p_{Y|X}(y|x)p_X(x) = \frac{d}{dy}P(Y \leq y, X = x)$$

and for simplicity, we will just write it as $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$. So the joint PDF/PMF, also admits the same form when the two variables are of a mixed type.

X **continuous,** Y **discrete:** When X is continuous and Y is discrete, the random variable $Y|X = x$ is still discrete so it admits a conditional PMF $p_{Y|X}(y|x)$. However, the conditioning operation on a continuous X has to be more carefully designed. Formally, for any measurable set C (a collection of possible values of X, Y), the conditional probability $P((X,Y) \in C|X = x)$ is defined as

$$P((X,Y) \in C|X = x) \equiv \lim_{\delta\to 0} P((X,Y) \in C|x \leq X \leq x + \delta).$$

Simple algebra shows that

$$\begin{aligned} P((X,Y) \in C|X = x) &= \lim_{\delta\to 0} P((X,Y) \in C|x \leq X \leq x + \delta) \\ &= \lim_{\delta\to 0} \frac{P((X,Y) \in C, x \leq X \leq x + \delta)}{P(x \leq X \leq x + \delta)} \\ &= \lim_{\delta\to 0} \frac{P((X,Y) \in C, x \leq X \leq x + \delta)}{p_X(x)\delta} \\ &= \frac{1}{p_X(x)} \lim_{\delta\to 0} \frac{P((X,Y) \in C, x \leq X \leq x + \delta)}{\delta} \\ &= \frac{1}{p_X(x)} \frac{d}{dx} P((X,Y) \in C, X \leq x). \end{aligned}$$

With the above result, we choose $C = \{(y,x) : x \in \mathbb{R}\}$ so $\{(X,Y) \in C\} = \{Y = y\}$, which leads to

$$\begin{aligned} P(Y = y|X = x) = P((X,Y) \in C|X = x) &= \frac{1}{p_X(x)} \frac{d}{dx} P((X,Y) \in C, X \leq x) \\ &= \frac{1}{p_X(x)} \frac{d}{dx} P(Y = y, X \leq x) \\ &= \frac{p_{X,Y}(x,y)}{p_X(x)}, \end{aligned}$$

where $p_{X,Y}(x,y)$ is the same mixed PDF as defined in the previous case (X discrete and Y continuous). You can verify that using the above approach, we will obtain the same formula when both variables are continuous.

To sum up, we can extend the PDF/PMF p_{XY} from both continuous or both discrete to a mixed case where one of them is continuous and on of them is discrete. Then the conditional PDF/PMF formula will hold still so we can just use the operation that we are familiar with in this case. For simplicity, we will call p_{XY} the joint PDF even if X or Y may be discrete (or both are discrete).

Remark. Formally, PMF and PDF can both be called *density function* in a generalized way called Radon-Nikodym derivative. Roughly speaking from the measure theory, the density $p(x)$ is define as the ratio $p(x) = \frac{dP(x)}{d\mu(x)}$, where $P(x)$ is the *probability measure* and $\mu(x)$ is another *measure*. When $\mu(x)$ is the Lebesgue measure, $p(x) = \frac{dP(x)}{d\mu(x)}$ is the usual PDF and when $\mu(x)$ is the counting measure, $p(x) = \frac{dP(x)}{d\mu(x)}$ reduces to the PMF. So both PDF and PMF can be called a *density*.

Example (Poisson-Exponential-Gamma). Suppose that we have two R.V.s X, Y such that $X \in \{0, 1, 2, 3, \cdots\}$ is discrete and $Y \geq 0$ is continuous. The joint PDF is

$$p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!}.$$

What is the conditional PDF $p_{X|Y}$ and $p_{Y|X}$?

We first compute $p_{X|Y}$. Using the fact that

$$p_Y(y) = \sum_x \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} = \lambda e^{-(\lambda+1)y} \underbrace{\sum_x \frac{y^x}{x!}}_{e^y} = \lambda e^{-\lambda y}.$$

So $Y \sim \mathsf{Exp}(\lambda)$. And thus

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{\frac{\lambda y^x e^{-(\lambda+1)y}}{x!}}{\lambda e^{-\lambda y}} = \frac{y^x e^{-y}}{x!},$$

which is the PDF (PMF) of $\mathsf{Poisson}(Y)$.

Here is a trick that we can quickly compute it. We know that $p_{X|Y}(x|y)$ will be a density function of x. So we only need to keep track of how the function changes w.r.t x and treat y as a fixed constant, which leads to

$$p_{X|Y}(x|y) \propto p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \propto \frac{y^x}{x!}.$$

From this, we immediately see that it is a Poisson distribution with rate parameter y.

For the conditional distribution $p_{Y|X}$, using the same trick as the above, we just keep track of y and treat x as a fixed constant, which leads to

$$p_{Y|X}(y|x) \propto p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \propto y^x e^{-(\lambda+1)y},$$

which is the Gamma distribution with parameter $\alpha = x+1, \beta = \lambda + 1$.

4.2 Conditional expectation

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X) = g(X)$ such that when $X = x$, its value is

$$\mathbb{E}(Y|X=x) = \begin{cases} \int yp(y|x)dy, & \text{if } Y \text{ is continuous,} \\ \sum_y yp(y|x), & \text{if } Y \text{ is discrete,} \end{cases}$$

where $p(y|x) = p(x,y)/p(x)$ is the conditional PDF/PMF. Essentially, the conditional expectation is the same the regular expectation but we place the PDF/PMF $p(y)$ by the conditional PDF/PMF $p(y|x)$.

Note that when X and Y are independent,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y), \quad \mathbb{E}(X|Y=y) = \mathbb{E}(X).$$

Law of total expectation:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Y|X]] &= \int \mathbb{E}[Y|X=x]p_X(x)dx = \int\int yp_{Y|X}(y|x)p_X(x)dxdy \\ &= \int\int yp_{XY}(x,y)dxdy = \mathbb{E}[Y]. \end{aligned}$$

A more general form of this is that for any measurable function $g(x,y)$, we have

$$\mathbb{E}[g(X,Y)] = \mathbb{E}[\mathbb{E}[g(X,Y)|X]]. \tag{4.1}$$

There are many cool applications of equation (4.1).

- Suppose $g(x,y) = q(x)h(y)$. Then equation (4.1) implies

$$\mathbb{E}[q(X)h(Y)] = \mathbb{E}[\mathbb{E}[q(X)h(Y)|X]] = \mathbb{E}[q(X)\mathbb{E}[h(Y)|X]].$$

- Let $w(X) = \mathbb{E}[q(Y)|X]$. The covariance

$$\begin{aligned}\mathsf{Cov}(g(X), q(Y)) &= \mathbb{E}[g(X)q(Y)] - \mathbb{E}[g(X)]\mathbb{E}[q(Y)]\\ &= \mathbb{E}[\mathbb{E}[g(X)q(Y)|X]] - \mathbb{E}[g(X)]\mathbb{E}[\mathbb{E}[q(Y)|X]]\\ &= \mathbb{E}[g(X)w(X)] - \mathbb{E}[g(X)]\mathbb{E}[w(X)]\\ &= \mathsf{Cov}(g(X), w(X))\\ &= \mathsf{Cov}(g(X), \mathbb{E}(q(Y)|X)).\end{aligned}$$

  Namely, the covariance between $g(X)$ and $q(Y)$ is the same as the covariance between $g(X)$ and $w(X) = \mathbb{E}[q(Y)|X]$. Thus, $w(X) = \mathbb{E}[q(Y)|X]$ is sometimes viewed as the projection from $q(Y)$ onto the space of X.

Law of total variance:

$$\begin{aligned}\mathsf{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2\\ &= \mathbb{E}[\mathbb{E}(Y^2|X)] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(law of total expectation)}\\ &= \mathbb{E}[\mathsf{Var}(Y|X) + \mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(definition of variance)}\\ &= \mathbb{E}[\mathsf{Var}(Y|X)] + \left\{\mathbb{E}[\mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2\right\}\\ &= \mathbb{E}\left[\mathsf{Var}(Y|X)\right] + \mathsf{Var}\left(\mathbb{E}[Y \mid X]\right) \quad \text{(definition of variance).}\end{aligned}$$

Example (Binomial-uniform). Consider two R.V.s X, Y such that

$$X|Y \sim \mathsf{Bin}(n, Y), \qquad Y \sim \mathsf{Unif}[0,1].$$

We are interested in $E[X]$, $\mathsf{Var}(X)$. For the marginal expectation $\mathbb{E}[X]$, using the law of total expectation,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[nY] = \frac{n}{2}.$$

The variance is

$$\begin{aligned}\mathsf{Var}(X) &= \mathbb{E}\left[\mathsf{Var}(X|Y)\right] + \mathsf{Var}\left(\mathbb{E}[X \mid Y]\right)\\ &= \mathbb{E}(nY(1-Y)) + \mathsf{Var}(nY)\\ &= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{2}.\end{aligned}$$

Now we examine the distribution of $Y|X$. Using the fact that

$$p_{Y|X}(y|x) \propto p_{XY}(x,y) = p_{X|Y}(x|y)p_Y(y) = \binom{n}{x} y^x(1-y)^{n-x} \propto y^x(1-y)^{n-x},$$

we can easily see that this is the PDF of a Beta distribution with parameter $\alpha = x+1$ and $\beta = n-x+1$. This is an interesting case because the uniform distribution over $[0,1]$ is equivalent to $\mathsf{Beta}(1,1)$. And $Y|X \sim \mathsf{Beta}(X+1, n-X+1)$. Thus, it behaves like initially, Y is from $\mathsf{Beta}(1,1)$. Then after observing the data X, we update the distribution of Y to $Y|X \sim \mathsf{Beta}(X+1, n-X+1)$. This is a way of modeling how the data informs our decision and is used in the Bayesian inference.

Example (missing data). Consider a social survey where we have two variables X, Y that X is the age of a participant and Y is the income. We are interested in the average income $\mu = \mathbb{E}[Y]$. However, we may not always observe Y since people may refuse to provide their income information. We use a binary variable R to denote the response pattern of Y. When $R = 1$, we observe both X and Y. When $R = 0$, we only observe X. We assume that $R \perp Y|X$ (this is a special case of *missing at random* assumption) so the response probability $P(R = 1|X, Y) = \pi(X)$ only depends on X. We further assume that $\pi(X)$ is a known function. Consider the *inverse probability weighting* quantity:

$$W = \frac{RY}{\pi(X)}.$$

Interestingly, W can always be computed–when $R = 1$, it is $\frac{Y}{\pi(X)}$ and when $R = 0$, it is 0. A more interesting fact is that W has the same mean as Y:

$$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}\left[\frac{RY}{\pi(X)}\right]\\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[RY|X]\right]\\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[R|X]\mathbb{E}[Y|X]\right]\\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\pi(X)\mathbb{E}[Y|X]\right]\\ &= \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].\end{aligned}$$

In reality, when we observe many IID random copies of $(X, R = 1, Y)$ or $(X, R = 0)$, we estimate $\mu = \mathbb{E}[Y]$ using

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(X_i)}$$

and this is called the IPW (inverse probability weighting) estimator.

Example (survey sampling). Suppose a city government is planning to estimate the average income of the city. The city has three districts: A and B and C. 60% of population lives in district A and 30% of population lives in district B and the remaining 10% in C. Thus, the data behaves like a pair of random variables X, Y, where $X \in \{A, B, C\}$ is the indicator of the district that this individual lives and Y is the income. The average income is then

$$\mu = 0.6\mathbb{E}[Y|X = A] + 0.3\mathbb{E}[Y|X = B] + 0.1\mathbb{E}[Y|X = C].$$

However, when the government conducted the survey, they surveyed the same amount of individuals in each district. So we have $P(X = A) = P(X = B) = P(X = C) = \frac{1}{3}$. In this case, suppose we have a single observe (X, Y), how should we construct a quantity $Z = g(X, Y)$ such that $\mathbb{E}[Z] = \mu$?

It turns out that we can use a similar idea to the inverse probability weighting called the *importance weighting* to construct such $Z = g(X, Y)$. Consider

$$\begin{aligned}Z &= \frac{0.6}{1/3}I(X = A)Y + \frac{0.3}{1/3}I(X = B)Y + \frac{0.1}{1/3}I(X = C)Y\\ &= 1.8I(X = A)Y + 0.9I(X = B)Y + 0.3I(X = C)Y.\end{aligned}$$

Namely, when the observation in the data is in district A, we count it as 1.8 individuals while when the observation in the data is in district C, we only count it as 0.3 individuals. Then you can easily verify that

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[\mathbb{E}[Z|X]]\\ &= 1.8\mathbb{E}[I(X=A)]\mathbb{E}[Y|X=A] + 0.9\mathbb{E}[I(X=B)]\mathbb{E}[Y|X=B] + 0.3\mathbb{E}[I(X=C)]\mathbb{E}[Y|X=C]\\ &= 0.6\mathbb{E}[Y|X=A] + 0.3\mathbb{E}[Y|X=B] + 0.1\mathbb{E}[Y|X=C]\\ &= \mu.\end{aligned}$$

Note that we use the fact that $\mathbb{E}[I(X=A)] = P(X=A)$ (think about why; hint: Bernoulli random variable).