MATH 5010 –Foundations of Statistical Theory and Probability

❖ Conditional Expectations

Instructor: He Wang
Department of Mathematics
Northeastern University

❖ Outline:

1. **Conditional Distributions**
2. **Conditional Expectations**

❖ Conditional Distribution

If X and Y are *discrete* r.v.'s then we can compute conditional probabilities:

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

The total probability formula

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$$

where the sum runs over all possible values of Y.

Remark: Conditioning is a very useful method for solving problems in probability, because it is often much easier to compute conditional probabilities and then sum over the result to find the 'unconditioned' probability.

Example: Best prize.

n distinct prizes arrive in sequence, all have different values, and one is the best. You must pick a prize or else move on to the next one (no going back to earlier ones). Your knowledge consists of the values of the previous prizes. You want to use a strategy that will maximize the probability of selecting the best prize. The prizes are randomly arranged in sequence.

***Strategy**: reject the first k prizes, then select the first one which is better than all of these previous ones.*

Let X be the position of the best prize. Use

$$P_k(Best) = \sum_{i=1}^{n} P_k(Best|X = i)P(X = i)$$

So

$$P_k(Best) \approx \frac{k}{n}\log\frac{n}{k}$$

Find the value of k to maximize this

Conditioning with respect to a continuous random variable

Let X be a continuous random variable, then for any event A we have

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x) f_X(x) dx$$

It is often convenient to use a shorthand and write this as

$$P(A) = E[P(A|X)]$$

where it is understood that the quantity $P(A|X)$ is a random variable which is a function of X.

Many interesting examples arise when the event A involves another random variable.

The Gambler's Ruin Problem

Setup:

- A gambler starts with k dollars
- Each round: win 1 with probability p , lose 1 with probability $q = 1 - p$.
- Game ends when gambler reaches N dollars (wins) or 0 dollars (ruins)
- $X =$ The result of the first bet $(+1\ or\ -1)$
- $A =$ Event "gambler eventually reaches N before going broke"

The Calculation Using Conditioning

Let P_k denote the probability of reaching N starting from k dollars.

Using the law of total probability:

$$P_k = P(A) = P(A \mid X = +1) \cdot P(X = +1) + P(A \mid X = -1) \cdot P(X = -1)$$

$$P_k = P_{k+1} \cdot p + P_{k-1} \cdot q$$

Mixed type of conditional distributions

Let X and Y be (either discrete or continuous) random variables.

Then we can compute conditional pdf/pmf:

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

The total probability formula

$$p_X(x) = \int_y p(x|y)p_Y(y)dy$$

where the integral/sum runs over all possible values of Y.

Example. Suppose that X, Y are independent exponentials with mean 1 and we want $P(X+Y \geq \mathrm{z})$, where $z \geq 0$. Now

$$P(X + Y \geq z \,|X = x) = P(Y \geq z - x \,|X = x) = P(Y \geq z - x)$$

because they are independent. Thus

$$P(Y \geq z - x) = \begin{cases} e^{-(z-x)} & \text{If } z - x \geq 0 \\ 1 & \text{If } z - x < 0 \end{cases}$$

And
$$P(X + Y \geq z) = \int_0^\infty P(X + Y \geq z \,|X = x)e^{-x}dx$$

$$= \int_0^\infty P(Y \geq z - x)e^{-x}dx$$

$$= \int_0^z e^{-z}dx + \int_z^\infty e^{-x}dx = ze^{-z} + e^{-z}$$

The same technique can be applied even when the random variables are dependent.

Example. Suppose X is uniform on [0, 1] and Y is uniform on $[0, X]$.

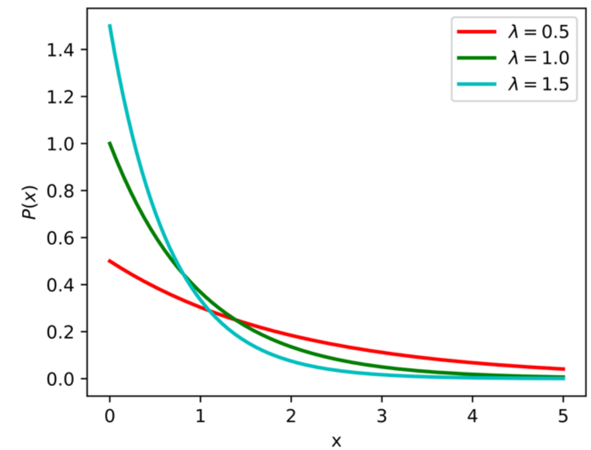
Calculate $E[Y]$.

Memoryless property of exponential rv

Conditioning can have quite unexpected effects on the distributions of random variables. One well-known example is the memoryless property of the exponential random variable.

Suppose that X is exponential with rate λ, so that its pdf is

$$f_X(t) = \lambda e^{-\lambda t} \qquad \text{for } t \geq 0$$



Then a calculation shows that

$$P(X \geq t) = e^{-\lambda t}$$

If we condition on this event we find that

$$P(X \geq t + s \mid X > s) = e^{-\lambda t}$$

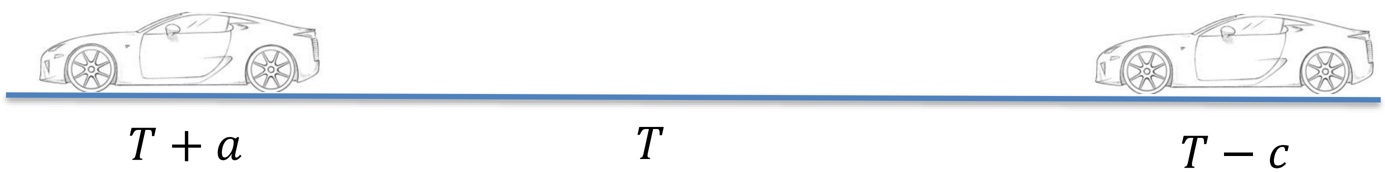
This can be interpreted as a memoryless property by viewing X as the time to failure of a device.

Conditioning on the event $\{X > s\}$ means that we condition on the device not having failed up to time s.

The result above says that given this event, the subsequent lifetime of the device has the same distribution as a fresh lifetime.

Example Cars pass a point on a highway. The times between successive cars are independent exponential random variables with the same mean m.

Suppose at a random time you stand at the point on the highway.
What is the mean time until the next car passes?



Example (Poisson-Exponential-Gamma).

Suppose that we have two random variables X, Y such that $X \in \{0, 1, 2, 3, \cdots\}$ is discrete and $Y \geq 0$ is continuous. The joint PDF is

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

The marginal distribution:

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

So, $Y \sim Exponential(\lambda)$

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

$$X|Y = y \sim Poisson(y)$$

Another method: We know that $p_{X|\mathrm{Y}}(x|y)$ will be a density function of x. So we consider y as a constant.

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

So, $X|Y = y$ is a Poisson distribution with rate parameter y

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

So, $Y|X = x$ is the Gamma distribution with parameter $\alpha = x + 1, \beta = \lambda + 1$

❖ Conditional Expectations

Define the **conditional expectation of X conditioned on the value $Y = y$** as

$$E[X|Y = y] := \sum_x x p_{X|Y}(x|y) = \sum_x x\, P(X = x|Y = y) \qquad X \text{ discrete}$$

$$E[X|Y = y] := \int_x x p_{X|Y}(x|y)\, dx \qquad X \text{ continuous}$$

Here, $p_{X|Y}(x|y) = \frac{p(x,y)}{p(x)}$ is the conditional pdf/pmf.

When X and Y are independent,

$$E(XY) = E(X)E(Y)$$

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

Conditional expectation $E[X|Y = y]$ is defined for each possible value of Y.

We get the random variable $E[X|Y]$ as a function of Y .

Think of $E[X|Y]$ as a random variable which is determined by the random variable Y , like Y^2 or e^{tY} : if you know the value of Y, then you know the value of $E[X|Y]$.

There is a very useful relation between the conditional expectation $E[X|Y]$ and the ‘unconditioned’ expectation $E[X]$.

Law of Total Expectation:

Theorem (Law of total expectation). $\boldsymbol{E}\big[\boldsymbol{E}[\boldsymbol{X}|\boldsymbol{Y}]\big] = \boldsymbol{E}[\boldsymbol{X}]$

Note that on the left side $E_Y\big[E_X[X|Y]\big]$ we are first averaging over X, with Y fixed, and then we average over Y.

$$E_Y\big[E_X[X|Y]\big] = \sum_y E(X|Y=y)\,P(Y=y) = \sum_y \left\{\sum_x x\,P(X=x|Y=y)\right\} P(Y=y)$$

$$= \sum_x \left\{\sum_y x\,P(X=x|Y=y)\right\} P(Y=y)$$

$$= \sum_x x \left\{\sum_y P(X=x|Y=y)\right\} P(Y=y)$$

$$= \sum_x x \left\{\sum_y P(X=x, Y=y)\right\} = \sum_x x\,P(X=x) = E[X]$$

Similarly for continuous case:

$$E_Y\big[E_X[X|Y]\big] = \int_y E[X|Y = y]p_Y(y)dy = \int_y \int_x xp_{X|Y}(x|y)\, p_Y(y)dxdy$$

$$= \int_y \int_x xp_{X,Y}(x,y)\, dxdy = \int_x \int_y xp_{X,Y}(x,y)\, dydx$$

$$= \int_x xP_X(x)dx = E[X]$$

More generally, given a measurable function $g(x,y)$

$$E[g(X,Y)] = E_X\big[E[g(X,Y)|X]\big]$$

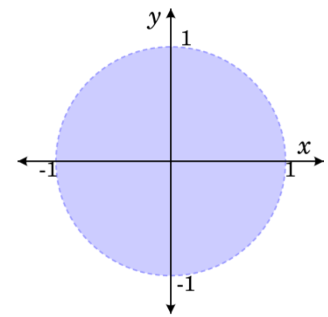
If $g(x,y) = f(x)h(y), then$

$$E[f(X)h(Y)] = E_X\big[E_Y[f(X)h(Y)|X]\big] = E_X\big[f(X)E_Y[h(Y)|X]\big]$$

Example:

Let (X, Y) be uniformly distributed over the **unit disk** $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$

The joint PDF is $f_{XY}(x,y) = \begin{cases} c & \text{if } (x,y) \in D \\ 0 & \text{otherwise} \end{cases}$



Are X and Y uncorrelated?

We need to check $Cov(X, Y)$

We already know that X and Y are not independent.

$$X|Y \sim \text{Uniform}\left(-\sqrt{1-Y^2}, \sqrt{1-Y^2}\right)$$

Example

Suppose $X \sim \text{Uniform}(1,2)$, and , $Y|X = x$ is exponential with parameter $\lambda = x$. Find $Cov(X, Y)$.

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Example (Random Sum of Random Variables):

Let $N, X_1, X_2, \dots$ be independent, where X_i are IID with $E[X_i] = \mu$. Define

$$Y = \sum_{i=1}^{N} X_i$$

Then

$$E[Y] = E[X]E[N]$$

For example, N is the number of insurance claims in a month, and X_i is the size of the i-th claim.

$$E[Y] = E\left[\sum_{i=1}^{N} X_i\right] = E[X_1 + X_2 + \cdots + X_N]$$

$$? = E[X_1] + E[X_2] + \cdots + E[X_N]$$

Conditional on N, this fixes the number of terms:

$$E[Y|N=n] = E\left[\sum_{i=1}^{N} X_i \mid N=n\right] = E\left[\sum_{i=1}^{n} X_i \mid N=n\right]$$

$$= E\left[\sum_{i=1}^{n} X_i\right]$$ Reason?

$$= E[X_1] + E[X_2] + \cdots + E[X_n]$$

$$= n\mu$$

$$E[Y] = E_N\big[E[Y|N]\big] = \sum_n E(Y|N=n)\,P(N=n) = \sum_n n\mu\,P(N=n)$$

$$= \mu \sum_n n\,P(N=n) = \mu E[N]$$

Law of total variance

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$$

$$\mathrm{Var}(Y) = E[Y^2] - (E[Y]\,)^2$$

$$= E\big[E[Y^2|X]\big] \; - (E[E[Y]\,|X]\,)^2$$

$$= E[Var(Y|X) + E[Y|X]^2] \; - E[E[Y]^2|X]$$

$$= E[Var(Y|X)] + E[E[Y|X]^2] \; - E[E[Y]^2|X]$$

$$= E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$$

Covariance

$$Cov\big(g(X),h(Y)\big) = Cov(g(X),E[h(Y)|X])$$

The random variable $E[h(Y)|X]$ is viewed as the projection of $h(Y)$ onto space of X

$$Cov\big(g(X),h(Y)\big) = E[g(X)h(Y)] - E[g(X)]E[h(Y)]$$

$$= E\big[E[g(X)h(Y)|X]\big] - E[g(X)]E\big[E[h(Y)|X]\big]$$

$$= E\big[g(X)E[h(Y)|X]\big] - E[g(X)]E\big[E[h(Y)|X]\big]$$

$$= Cov(g(X),E[h(Y)|X])$$

Example (Binomial-uniform).

Suppose $X|Y \sim Binomial(n, Y)$ and $Y \sim Uniform[0,1]$

Find $E[X]$ and $Var(X)$

Solution:

Using the law of total expectation,

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

Using the law of total variance,

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

$$= E[nY(1-Y)] + Var(nY)$$

$$= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{2}$$

Determine the distribution of $Y\ |X$

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

$$\propto y^x(1-y)^{n-x}$$

$Y\ |X$ is a Beta distribution with parameters $\alpha = x + 1\ and\ \beta = n - x + 1$.

Remark: Y~Uniform[0, 1] is equivalent to Beta(1, 1).

After observing the data X, we update the distribution of $Y\ to$

$Y|X \sim$ Beta($X + 1, n - X + 1$)

Bayesian inference: modeling how the data informs our decision.

Example (Missing data)

Consider a survey with two variables:

$X =$ the age of a participant
$Y =$ the income of a participant

We are interested in the average income $\mu = E[Y]$.

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 R and Y are conditionally independent given 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:

$$W = \frac{RY}{\pi(X)} = \begin{cases} \dfrac{Y}{\pi(X)} & \text{when } R = 1 \\ 0 & \text{when } R = 0 \end{cases}$$

So, W is computable and further more $E[W] = E[Y]$

$$E[W] = E\left[\frac{RY}{\pi(X)}\right] = E\left[\frac{1}{\pi(X)} E[RY|X]\right]$$

$$= E\left[\frac{1}{\pi(X)} E[R|X] E[Y|X]\right]$$

$$= E\left[\frac{1}{\pi(X)} \pi(X) E[Y|X]\right]$$

$$= E\big[E[Y|X]\big] = E[Y]$$

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

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

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

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

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 $E[Z] = \mu$?

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

$$Z = \frac{0.6}{1/3} I(X = A)Y + \frac{0.3}{1/3} I(X = B)Y + \frac{0.1}{1/3} I(X = C)Y$$

$$= 1.8I(X = A)Y + 0.9I(X = B)Y + 0.3I(X = C)Y$$

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.

$$E[Z] = E[E[Z|X]]$$

$$= 1.8E[I(X = A)]E[Y|X = A] + 0.9E[I(X = B)]E[Y|X = B] + 0.3E[I(X = C)]E[Y|X = C]$$

$$= 0.6E[Y|X = A] + 0.3E[Y|X = B] + 0.1E[Y|X = C]$$

Here we use $E[I(X = A)] = P(X = A)$ for Bernoulli random variable.

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).