Probabilistic Modeling and Statistical Computing Fall 2015

October 6, 2015

Conditional Distribution

... for the case of two discrete random variables X, Y with joint pmf $f_{XY}(x, y)$. Suppose the marginal distribution of X_1 has pmf $f_X(x)$.

Given x such that $\mathcal{P}(X = x) = f_X(x) > 0$, define the conditional pmf of $Y|X = x$ as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

What does this mean in terms of actual probabilities?

Independence

The random variables X and Y are called independent if

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

for all possible choices of x, y.
Since $f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x)$, this means

$$f_{Y|X}(y|x) = f_Y(y).$$

X and Y are independent if the joint pmf can be factored algebraically into the marginal pmfs.

Burglar Example

Are X and N independent? Intuition?

$$f_{NX}(n,k) = \binom{n}{k} q^k (1-q)^{n-k} \cdot p \cdot (1-p)^n$$

Is $f_{NX}(n,k) = f_X(k) f_N(n)$? Does it look like this can be factored?

Independence of >2 random variables

The random variables $X_1, \ldots, X_m$ are independent if their joint pmf can be factored into a product of all the marginal pmfs.

$$\boxed{f(x_1, x_2, \ldots, x_m) = f_{X_1}(x_1) f_{X_2}(x_2) \cdot \cdots \cdot f_{X_m}(x_m)}$$

Strong requirement - more than just pairwise independence!

Multinomial Distribution

Given an integer $n > 0, n \in \mathbb{Z}$ and probabilities $p_1, p_2, \ldots, p_k$ with $\sum_i p_i = 1$, the pmf of the **multinomial distribution** is

$$\mathcal{P}(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}$$

for $(X_1, \ldots X_k)$. The range is

$$\mathcal{R} = \{(n_1, \ldots, n_k) | n_i \geq 0,\ n_i \in \mathbb{Z}, \sum_i n_i = n\}$$

That is, the n_i must be nonnegative integers whose sum equals n.

Multinomial Distribution

Interpretation: carry out n independent trials. Each trial can have one of k outcomes. Outcome i occurs with probability p_i. Then $X_i =$ number of outcomes i.

The probabilities occur as terms when

$$(p_1 + p_2 + \cdots + p_k)^n$$

is expanded. The values of the X_i are the powers of the p_i.

Special case: $k = 2$. Then $X_1 \sim B(n, p_1)$.

Multinomial Distribution

Are the X_i independent? What is your intuition? What if $k = 2$? What if k is very large? If $k = 2$, then X_1 is completely determined by X_2. But if k is very large (think $k \approx 10^6$ for an online retailer), then any two X_i might be approximately independent.

Multinomial Distribution - Marginal and Conditional

Given the multinomial distribution, the **marginal distribution** of each X_i is $B(n, p_i)$.

We can also describe conditional distributions. Let's say $k = 4$ and we want the joint distribution of

$$(X_1, X_2, X_3) | X_4 = n_4$$

for some fixed n_4.
This is simply the multinomial distribution with $k = 3$, $n - n_4$ trials, and probabilities $\frac{p_i}{1-p_4}$.
Where does all this come from?

Conditional Expectations

... for the case of two random variables X, Y with joint pmf $f_{XY}(x, y)$. Suppose the marginal distribution of X has pmf $f_X(x)$.
We can define $\mathcal{E}(Y|X = x)$ for all choices of x, using the conditional distribution:

$$\mathcal{E}(Y|x) = \sum_y y \cdot f_{Y|X}(y|x)$$

This doesn't depend on y. It's just a function of the variable x. **Therefore, we can think of $\mathcal{E}(Y|X)$ as a random variable.**

Conditional Expectations

The conditional expectation $\mathcal{E}(Y|X)$ tells us the mean of Y once X is observed. **It depends therefore on X but not on Y.**

In general, $\mathcal{E}(\mathcal{E}(Y|X)) = \mathcal{E}(Y)$.

On the right: Take the expectation over this marginal distribution.

On the left: $\mathcal{E}(Y|X)$ is taken over the conditional distribution of Y, for each value of X. The "outer" expectation then is taken over the marginal distribution of X.

Burglar Example

Since $X|N$ has a binomial distribution, $B(N, q)$, we can say that

$$\mathcal{E}(X|N) = Nq\,.$$

We think of N as a random variable (it hasn't been observed yet). $\mathcal{E}(X|N)$ *a function of N.* Take the "outer" expected value:

$$\mathcal{E}(X) = \mathcal{E}(\mathcal{E}(X|N)) = q \cdot \mathcal{E}(N) = q \cdot \frac{1-p}{p}$$

Questions

- What is $\mathcal{E}(Y|X)$ if X and Y are independent?
- What is $\mathcal{E}(X|X)$?
- What is $\mathcal{E}\left(\mathcal{E}(Y|X)|X\right)$?
- What is $\mathcal{E}(X|X+1)$?
- Suppose $\mathcal{P}(Y = 1000) = 1$. What is $\mathcal{P}(X|Y)$? What is $\mathcal{P}(Y|X)$?
- Suppose $h(X)$ is some function of X alone. Why is $\mathcal{E}(h(X)Y|X) = h(X)\mathcal{E}(Y|X)$?

Variance and Covariance

.. for the case of two random variables X, Y with joint pmf $f_{XY}(x, y)$.
Define $\mathcal{E}(X)$, $\mathcal{E}(Y)$ simply in terms of marginal distributions.
Define $var(X) = \mathcal{E}((X - \mathcal{E}(X))^2)$ etc. in terms of marginal distributions only.
However, for $\mathcal{E}(XY)$ we need the joint pmf:

$$\mathcal{E}(XY) = \sum_{x,y} xyf_{XY}(x, y)$$

Variance and Covariance

If X and Y are independent, the joint pmf factors, $f_{XY}(x, y) = f_X(x) f_Y(y)$ and therefore

$$
\begin{aligned}
\boxed{\mathcal{E}(XY)} &= \sum_{x,y} xy f_{XY}(x, y) \\
&= \sum_{x,y} xy f_X(x) f_Y(y) \\
&= \sum_{x} x f_X(x) \cdot \sum_{y} y f_Y(y) \\
&\boxed{= \mathcal{E}(X)\mathcal{E}(Y)}
\end{aligned}
$$

Covariance and Correlation Coefficient

The covariance of X and Y is defined as

$$cov(X, Y) = \mathcal{E}((X - \mathcal{E}(X)) \cdot (Y - \mathcal{E}(Y)))$$

Can be simplified to $\mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y)$. Therefore, if X, Y are independent, then

$$cov(X, Y) = 0\,.$$

The correlation coefficient is defined as

$$\rho_{XY} = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

Covariance and Correlation

- $cov(X, Y) = cov(Y, X)$
- $cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$
- If $\alpha \in \mathbb{R}$, then $cov(\alpha X, Y) = \alpha cov(X, Y) = cov(X, \alpha Y)$
- $cov(X, X) = var(X)$
- If $y = const.$, then $cov(X, Y) = 0$
- $-1 \leq \rho_{XY} = \rho_{YX} \leq 1$
- If X is replaced by αX and Y is replaced by βY with $\alpha, \beta > 0$ (change of units) then ρ_{XY} does not change.

Addition Formula for Variances

Given two random variables X, Y, how can we compute $var(X+Y)$?
We need to know $var(X)$, $var(Y)$, $cov(X,Y)$.
Don't need the entire joint pdf!

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$$

If X and T are independent, then $cov(X,Y) = 0$ and therefore $var(X+Y) = var(X) + var(Y)$.
What if $Y = X$? What if $Y = -X$?

Addition Formula for Variances

Given two random variables X, Y, how can we compute $var(X+Y)$?
We need to know $var(X)$, $var(Y)$, $cov(X, Y)$.
Don't need the entire joint pdf!

$$var(X+Y) = var(X) + var(Y) + 2cov(X, Y)$$

If X and T are independent, then $cov(X, Y) = 0$ and therefore $var(X+Y) = var(X) + var(Y)$.
What if $Y = X$? What if $Y = -X$?

Burglar Example

We know

$$\mathcal{E}(X) = q \cdot \frac{1-p}{p}, \quad \mathcal{E}(N) = \frac{1-p}{p}.$$

We can compute

$$\mathcal{E}(XN) = \mathcal{E}(\mathcal{E}(XN|N)) = \mathcal{E}(N\mathcal{E}(X|N))$$
$$= \mathcal{E}(N \cdot Nq) = q\mathcal{E}(N^2)$$

which can be worked out to become $q\frac{(2-p)(1-p)}{p^2}$

and therefore $\boxed{cov(X,N) = \frac{q(1-p)}{p^2}}$.

Check this!

Urn Example

A box contains n_1 red balls, n_2 black balls and n_3 green balls. Let $n = n_1 + n_2 + n_3$. We draw $k \leq n$ balls **with replacement**. Let $X_{1,2,3} =$ number of red/black/green balls. Then (X_1, X_2, X_3) have a multinomial distribution. What are the parameters?

- Write an **R** function of n_1, n_2, n_3, k that simulates a single sample.
- It is known that $var(X_1) = \frac{kn_1(n-n_1)}{n^2}$ and $cov(X_1, X_2) = -\frac{kn_1n_2}{n^2}$. Check this with a simulation for some k, n_i of your choice.

Continuous Multivariate Distribution

Assume that (X, Y) is a 2-dim random vector of continuous random variables. The joint probability density function (pdf) $f_{XY}(x, y)$ is defined by the property

$$\mathcal{P}(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) dy\, dx\,.$$

Can also be defined for $n > 2$ random variables.
The integration becomes difficult if $n \gtrapprox 5$.
But: *Cases like $n \approx 10^6$ occur e.g. in imaging.*

Example: Arrival Times

You and your friend Zoe want to get together. You will arrive at some random time between 1PM and 1:30PM. Zoe is running late and will arrive at some random time between 1:10PM and 1:30PM. Let $X, \ Y$ be the two arrival times, in minutes after 1PM.

Joint pdf:

$$f(x, y) = \begin{cases} \frac{1}{30 \cdot 20} & (0 < x < 30, 10 < y < 30) \\ 0 & otherwise \end{cases}$$

Marginal Distributions

Assume that (X, X) is a 2-dim random vector with joint pdf $f_{XY}(x, y)$). The marginal pdf of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy\,.$$

*Integrate over all possible values of the **other** component of the random vector.*

Example: Arrival Times

Marginal pdf's for your and your friend Zoe's arrival times:

$$f_X(x) = \begin{cases} \frac{1}{30} & (0 < x < 30) \\ 0 & \textit{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{20} & (10 < y < 30) \\ 0 & \textit{otherwise} \end{cases}$$

Conditional Distribution

Suppose the joint pdf of X, Y is $f_{XY}(x, y)$ and the marginal pdf of X_1 has pmf $f_X(x)$.

For x such that $f_X(x) > 0$, define the conditional pdf of $Y|X = x$ as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

Example: Arrival Times

Conditional pdf for Zoe's arrival times: Given a possible arrival time $x \in [0, 30]$, Zoe's conditional pdf is Defined for all $0 < x < 30$ and is given by

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{20} & (10 < y < 30) \\ 0 & otherwise \end{cases}$$

This does not depend on y. What does that mean?
This does not depend on x. What does that mean?

Independence

The random variables X and Y are called independent if

$$\boxed{f_{XY}(x,y) = f_X(x)f_Y(y)}$$

for all possible choices of x, y.
Since $f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x)$, this means

$$\boxed{f_{Y|X}(y|x) = f_Y(y)}\ .$$

Zoe's and your arrival times are independent.

Another worked out example

Consider two continuous random variables X, Y with joint pdf

$$f_{XY}(x,y) = \begin{cases} \frac{1}{2}(x+3y) & (0 \leq x, y \leq 1) \\ 0 & \text{otherwise} \end{cases}$$

This is a legitimate joint pdf. What must we check to verify this?

Example - Continued

Find the marginal distributions:

$$f_X(x) = \int_0^1 f_{XY}(x,y)dy = \int_0^1 \frac{x+3y}{2}dy = \frac{x}{2} + \frac{3}{4}$$
$$f_Y(y) = \int_0^1 f_{XY}(x,y)dx = \int_0^1 \frac{x+3y}{2}dx = \frac{1}{4} + \frac{3y}{2}$$

Check that each of these is a legitimate pdf.

Example - Continued

Find the conditional distribution of Y given $X = x \in [0, 1]$:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{(x + 3y)/2}{x/2 + 3/4}$$

Check that this is a legitimate pdf wrt. y.
Is it a legitimate pdf wrt. x? Should it be one?

Example - Continued

Find the conditional expectation $\mathcal{E}(Y|X=x)$ for each $x \in [0,1]$: This is given by the integral

$$\int_0^1 y \cdot f_{Y|X}(y|x)dy = \int_0^1 y \cdot \frac{(x+3y)/2}{x/2+3/4}dy = \frac{x+2}{2x+3}$$

Therefore $\mathcal{E}(Y|X=x) = \frac{x+2}{2x+3}$. Then

$$\mathcal{E}(\mathcal{E}(Y|X)) = \int_0^1 f_X(x)\mathcal{E}(Y|X=x)dx = \frac{5}{8} = \mathcal{E}(Y).$$

Example - Continued

Since for $0 \leq x, y \leq 1$

$$f_{XY}(x, y) = (x + 3y)/2 \neq f_X(x) f_Y(y)$$

X and Y are **not** independent. This can also be seen from the fact that $\mathcal{E}(Y|X)$ depends on X. Find the correlation!

- Compute $\mathcal{E}(X), \mathcal{E}(Y)$ from marginal pdf's.
- Compute $var(X) = \mathcal{E}(X^2) - \mathcal{E}(X)^2$ and $var(Y)$ from marginal pdf's.
- Compute $cov(X, Y) = \mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y)$ from joint pdf. Then compute ρ_{XY}.

Example - Continued

Results:

- $\mathcal{E}(X) = \int_0^1 x f_X(x) dx = \frac{13}{24}$, $\mathcal{E}(Y) = \frac{5}{8}$
- $\mathcal{E}(X^2) = \int_0^1 x^2 f_X(x) dx = \frac{3}{8}$
  $var(X) = \mathcal{E}(X^2) - \mathcal{E}(X)^2 = \frac{47}{576}$
- $var(Y) = \frac{13}{192}$
- $\mathcal{E}(XY) = \int\int_0^1 xy \cdot f_{XY}(x, y) dy dx = \frac{1}{3}$
  $cov(X, Y) = \mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y) = -\frac{1}{192}$
- $\rho_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = -.07007\ldots.$

Bivariate Normal Distribution

Want to make continues random variables X, Y such that

- The marginal dist. of X is given, $N(0, \sigma_x^2)$.
- The marginal dist. of Y is given, $N(0, \sigma_y^2)$.
- The correlation coefficient $\rho \in (-1, 1)$ is given.

Easy to simulate. Joint pdf is messy to write down. Joint pdf is easy to write down if you know Linear Algebra.

Bivariate Normal - Construction

Start with two independent $Z_1, Z_2 \sim N(0, 1)$.

- Set $X = \sigma_x Z_1$.
- Set $Y = \sigma_y \left(\rho Z_1 + \sqrt{1 - \rho^2} Z_2\right)$.
- **Why does this work?**
- Theoretical result implies: X and Y have normal distributions.

Bivariate Normal - Construction

Start with two independent $Z_1, Z_2 \sim N(0,1)$.

- Set $X = \sigma_x Z_1$.
- Set $Y = \sigma_y \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2 \right)$.
- **Why does this work?**
- Theoretical result implies: X and Y have normal distributions.

Bivariate Normal Distribution

Compute means, variances, and covariances of X and Y:

- $\mathcal{E}(X) = \sigma_x \mathcal{E}(Z_1) = 0$, $\mathcal{E}(Y) = 0$
- $var(X) = \sigma_x^2 \cdot var(Z_1) = \sigma_x^2$.
- We know that $cov(Z_1, Z_2) = 0$. Therefore,

$$\begin{aligned} var(Y) &= \sigma_y^2 \left(var(\rho Z_1) + var(\sqrt{1-\rho^2} Z_2 \right) \\ &= \sigma_y^2 \left(\rho^2 + (1-\rho^2) \right) = \sigma_y^2 \end{aligned}$$

- Therefore $X \sim N(0, \sigma_x^2)$ and $Y \sim N(0, \sigma_y^2)$.

Bivariate Normal Distribution

Compute the covariance of X and Y: We know that $cov(Z_1, Z_1) = 1,\ cov(Z_1, Z_2) = 0$.

$$\begin{aligned} cov(X, Y) &= cov(\sigma_x Z_1, \sigma_y(\rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= \sigma_x \sigma_y cov(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= \sigma_x \sigma_y (\rho cov(Z_1, Z_1) + \sqrt{1-\rho^2} cov(Z_1, Z_2) \\ &= \sigma_x \sigma_y \rho = \rho \sqrt{var(X) var(Y)} \end{aligned}$$

and therefore

$$\rho_{XY} = \frac{cov(X, Y)}{\sqrt{var(X) var(y)}} = \rho\,.$$

The case of general means

To make $X \sim N(\mu_x, \sigma_x^2),\ Y \sim N(\mu_y, \sigma_y^2)$ such that $\rho_{XY} = \rho$, just make $X,\ Y$ as above and then add μ_x to X and μ_y to Y.

Simulations

The case of general means

To make $X \sim N(\mu_x, \sigma_x^2),\ Y \sim N(\mu_y, \sigma_y^2)$ such that $\rho_{XY} = \rho$, just make $X,\ Y$ as above and then add μ_x to X and μ_y to Y.
Simulations

The joint pdf

Use Linear Algebra to write down the joint pdf.
Let

$$\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Here $\vec{\mu}$ is the vector of means, and Σ is called "covariance matrix" for obvious reasons. One can compute its determinant $\det\Sigma = \sigma_x^2\sigma_y^2(1-\rho^2)$. The inverse is

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_x^{-2} & \rho\sigma_x^{-1}\sigma_y^{-1} \\ \rho\sigma_x^{-1}\sigma_y^{-1} & \sigma_y^{-2} \end{pmatrix}$$

Joint pdf using matrix algebra

Write $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$. Then $(X, Y) \sim N(\vec{\mu}, \Sigma)$ and the joint pdf of X and Y is given by

$$f_{XY}(\vec{x}) = \frac{1}{2\pi\sqrt{\det\Sigma}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

Compare this to the 1-d case: If $X \sim N(\mu, \sigma^2)$, then μ is like $\vec{\mu}$, σ^2 is like Σ, and the pdf of X is

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

The following material was briefly done near the end of the class on the whiteboard, but we did not look at the slides.

Best Predictor I

Suppose X and Y are distributed jointly.
That means that they have something to do with each other: $X \sim$ age and $Y \sim$ preference for pizza, $X \sim$ income and $Y \sim$ credit worthiness, etc.
We observe X and want to predict Y with some function $g(X)$.
What is the best such predictor?
Which function g minimizes the "quadratic loss"

$$\boxed{L_g = \mathcal{E}((Y - g(X))^2)}$$

Best Predictor II

Find g that minimizes

$$\begin{aligned} L_g &= \mathcal{E}((Y - g(X))^2) \\ &= \mathcal{E}(([Y - \mathcal{E}(Y|X)] + [\mathcal{E}(Y|X) - g(X)])^2) \\ &= \mathcal{E}((A+B)^2) = \mathcal{E}(A^2) + 2\mathcal{E}(AB) + \mathcal{E}(B^2) \end{aligned}$$

with $A = [Y - \mathcal{E}(Y|X)]$, $B = [\mathcal{E}(Y|X) - g(X)]$.
Now $\mathcal{E}(A^2)$ does not depend on g (nothing to minimize).
We can make $\mathcal{E}(B^2) = 0$ with the choice

$$\boxed{g(X) = \mathcal{E}(Y|X)}\,.$$

Best Predictor III

For our choice $g(X) = \mathcal{E}(Y|X)$, what is

$$\mathcal{E}(AB) = \mathcal{E}([Y - \mathcal{E}(Y|X)] \cdot [\mathcal{E}(Y|X) - g(X)])?$$

With our choice of g, this is zero. But could it be negative?
Note that A depends on X and Y, but $B = B(X)$ depends only on X. Then:

$$\mathcal{E}(AB) = \mathcal{E}(\mathcal{E}(AB|X)) = \mathcal{E}(B(X) \cdot \mathcal{E}(A|X))$$

and $\mathcal{E}(A|X) = \mathcal{E}([Y - \mathcal{E}(Y|X)]|X) = 0$.
Therefore $\mathcal{E}(AB) = 0$ no matter what g is.

Best Predictor IV

With the choice $g(X) = \mathcal{E}(Y|X)$ the quadratic loss becomes
$L_g = \mathcal{E}((Y - g(X))^2) = \mathcal{E}((Y - \mathcal{E}(Y|X)^2)$ and it cannot become smaller.
That is the best predictor for this quadratic loss function.

Now if we can somehow apply the plug-in principle, all our prediction problems are in principle solved.

Where are the practical difficulties? Is this at all useful?

Best Predictor IV

With the choice $g(X) = \mathcal{E}(Y|X)$ the quadratic loss becomes
$L_g = \mathcal{E}((Y - g(X))^2) = \mathcal{E}((Y - \mathcal{E}(Y|X)^2)$ and it cannot become smaller.
That is the best predictor for this quadratic loss function.

Now if we can somehow apply the plug-in principle, all our prediction problems are in principle solved.

Where are the practical difficulties? Is this at all useful?