7. Multivariate Probability

Chris Piech and Mehran Sahami

May 2017

Often you will work on problems where there are several random variables (often interacting with one another). We are going to start to formally look at how those interactions play out.

For now we will think of joint probabilities with two random variables X and Y.

1 Discrete Joint Distributions

In the discrete case a joint probability mass function tells you the probability of any combination of events $X = a$ and $Y = b$:

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

This function tells you the probability of all combinations of events (the "," means "and"). If you want to back calculate the probability of an event only for one variable you can calculate a "marginal" from the joint probability mass function:

$$p_X(a) = P(X = a) = \sum_y P_{X,Y}(a,y)$$
$$p_Y(b) = P(Y = b) = \sum_x P_{X,Y}(x,b)$$

In the continuous case a joint probability density function tells you the relative probability of any combination of events $X = a$ and $Y = y$.

In the discrete case, we can define the function $p_{X,Y}$ non-parametrically. Instead of using a formula for p we simply state the probability of each possible outcome.

2 Continuous Joint Distributions

Random variables X and Y are Jointly Continuous if there exists a Probability Density Function (PDF) $f_{X,Y}$ such that:

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)dy\,dx$$

Using the PDF we can compute marginal probability densities:

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy$$
$$f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)dx$$

Let $F(a,b)$ be the Cumulative Density Function (CDF):

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F(a_2,b_2) - F(a_1,b_2) + F(a_1,b_1) - F(a_2,b_1)$$

3 Multinomial Distribution

Say you perform n independent trials of an experiment where each trial results in one of m outcomes, with respective probabilities: $p_1, p_2, \ldots, p_m$ (constrained so that $\sum_i p_i = 1$). Define X_i to be the number of trials with outcome i. A multinomial distribution is a closed form function that answers the question: What is the probability that there are c_i trials with outcome i. Mathematically:

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \ldots p_m^{c_m}$$

Example 1

A 6-sided die is rolled 7 times. What is the probability that you roll: 1 one, 1 two, 0 threes, 2 fours, 0 fives, 3 sixes (disregarding order).

$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3) &= \frac{7!}{2!3!} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 \\ &= 420 \left(\frac{1}{6}\right)^7 \end{aligned}$$

Fedaralist Papers

In class we wrote a program to decide whether or not James Madison or Alexander Hamilton wrote Fedaralist Paper 49. Both men have claimed to be have written it, and hence the authorship is in dispute. First we used historical essays to estimate p_i, the probability that Hamilton generates the word i (independent of all previous and future choices or words). Similarly we estimated q_i, the probability that Madison generates the word i. For each word i we observe the number of times that word occurs in Fedaralist Paper 49 (we call that count c_i). We assume that, given no evidence, the paper is equally likely to be written by Madison or Hamilton.

Define three events: H is the event that Hamilton wrote the paper, M is the event that Madison wrote the paper, and D is the event that a paper has the collection of words observed in Fedaralist Paper 49. We would like to know whether $P(H|D)$ is larger than $P(M|D)$. This is equivalent to trying to decide if $P(H|D)/P(M|D)$ is larger than 1.

The event $D|H$ is a multinomial parameterized by the values p. The event $D|M$ is also a multinomial, this time parameterized by the values q.

Using Bayes Rule we can simplify the desired probability.

$$\begin{aligned} \frac{P(H|D)}{P(M|D)} &= \frac{\frac{P(D|H)P(H)}{P(D)}}{\frac{P(D|M)P(M)}{P(D)}} = \frac{P(D|H)P(H)}{P(D|M)P(M)} = \frac{P(D|H)}{P(D|M)} \\ &= \frac{\binom{n}{c_1, c_2, \ldots, c_m} \prod_i p_i^{c_i}}{\binom{n}{c_1, c_2, \ldots, c_m} \prod_i q_i^{c_i}} = \frac{\prod_i p_i^{c_i}}{\prod_i q_i^{c_i}} \end{aligned}$$

This seems great! We have our desired probability statement expressed in terms of a product of values we have already estimated. However, when we plug this into a computer, both the numerator and denominator come out to be zero. The product of many numbers close to zero is too hard for a computer to represent. To fix this problem, we use a standard trick in computational probability: we apply a log to both sides and apply

some basic rules of logs.

$$\begin{aligned}\log\Big(\frac{P(H|D)}{P(M|D)}\Big) &= \log\Big(\frac{\prod_i p_i^{c_i}}{\prod_i q_i^{c_i}}\Big)\\ &= \log(\prod_i p_i^{c_i}) - \log(\prod_i q_i^{c_i})\\ &= \sum_i \log(p_i^{c_i}) - \sum_i \log(q_i^{c_i})\\ &= \sum_i c_i \log(p_i) - \sum_i c_i \log(q_i)\end{aligned}$$

This expression is "numerically stable" and my computer returned that the answer was a negative number. We can use exponentiation to solve for $P(H|D)/P(M|D)$. Since the exponent of a negative number is a number smaller than 1, this implies that $P(H|D)/P(M|D)$ is smaller than 1. As a result, we conclude that Madison was more likely to have written Federalist Paper 49.

4 Expectation with Multiple RVs

Expectation over a joint isn't nicely defined because it is not clear how to compose the multiple variables. However, expectations over functions of random variables (for example sums or multiplications) are nicely defined: $E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$ for any function $g(X,Y)$. When you expand that result for the function $g(X,Y) = X+Y$ you get a beautiful result:

$$\begin{aligned}E[X+Y] = E[g(X,Y)] &= \sum_{x,y} g(x,y)p(x,y) = \sum_{x,y}[x+y]p(x,y)\\ &= \sum_{x,y} xp(x,y) + \sum_{x,y} yp(x,y)\\ &= \sum_x x\sum_y p(x,y) + \sum_y y\sum_x p(x,y)\\ &= \sum_x xp(x) + \sum_y yp(y)\\ &= E[X] + E[Y]\end{aligned}$$

This can be generalized to multiple variables:

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

Expectations of Products Lemma

Unfortunately the expectation of the product of two random variables only has a nice decomposition in the case where the random variables are independent of one another.

$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ if and only if X and Y are independent

Example 3

A disk surface is a circle of radius R. A single point imperfection is uniformly distributed on the disk with joint PDF:

$$f_{X,Y}(x,y) = \begin{cases}\frac{1}{\pi R^2} & \text{if } x^2+y^2 \le R^2\\ 0 & \text{else}\end{cases}$$

Let D be the distance from the origin: $D = \sqrt{X^2+Y^2}$. What is $E[D]$? Hint: use the lemmas

5 Independence with Multiple RVs

Discrete

Two discrete random variables X and Y are called independent if:

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x, y$$

Intuitively: knowing the value of X tells us nothing about the distribution of Y. If two variables are not independent, they are called dependent. This is a similar conceptually to independent events, but we are dealing with multiple *variables*. Make sure to keep your events and variables distinct.

Continuous

Two continuous random variables X and Y are called independent if:

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b) \text{ for all } a, b$$

This can be stated equivalently as:

$$F_{X,Y}(a,b) = F_X(a)F_Y(b) \text{ for all } a, b$$
$$f_{X,Y}(a,b) = f_X(a)f_Y(b) \text{ for all } a, b$$

More generally, if you can factor the joint density function then your continuous random variable are independent:

$$f_{X,Y}(x,y) = h(x)g(y) \text{ where } -\infty < x, y < \infty$$

Example 2

Let N be the # of requests to a web server/day and that $N \sim Poi(\lambda)$. Each request comes from a human (probability = p) or from a "bot" (probability = $(1–p)$), independently. Define X to be the # of requests from humans/day and Y to be the # of requests from bots/day.

Since requests come in independently, the probability of X conditioned on knowing the number of requests is a Binomial. Specifically:

$$(X|N) \sim Bin(N, p)$$
$$(Y|N) \sim Bin(N, 1-p)$$

Calculate the probability of getting exactly i human requests and j bot requests. Start by expanding using the chain rule:

$$P(X = i, Y = j) = P(X = i, Y = j | X + Y = i + j)P(X + Y = i + j)$$

We can calculate each term in this expression:

$$P(X = i, Y = j | X + Y = i + j) = \binom{i+j}{i} p^i (1-p)^j$$

$$P(X + Y = i + j) = e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

Now we can put those together and simplify:

$$P(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

As an exercise you can simplify this expression into two independent Poisson distributions.

Symmetry of Independence

Independence is symmetric. That means that if random variables X and Y are independent, X is independent of Y and Y is independent of X. This claim may seem meaningless but it can be very useful. Imagine a sequence of events $X_1, X_2, \ldots$. Let A_i be the event that X_i is a "record value" (eg it is larger than all previous values). Is A_{n+1} independent of A_n? It is easier to answer that A_n is independent of A_{n+1}. By symmetry of independence both claims must be true.

6 Convolution of Distributions

Convolution is the result of adding two different random variables together. For some particular random variables computing convolution has intuitive closed form equations. Importantly convolution is the sum of the random variables themselves, not the addition of the probability density functions (PDF)s that correspond to the random variables.

Independent Binomials with equal p

For any two Binomial random variables with the same "success" probability: $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$ the sum of those two random variables is another binomial: $X + Y \sim Bin(n_1 + n_2, p)$. This does not hold when the two distribution have different parameters p.

Independent Poissons

For any two Poisson random variables: $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$ the sum of those two random variables is another Poisson: $X + Y \sim Poi(\lambda_1 + \lambda_2)$. This holds when λ_1 is not the same as λ_2.

Independent Normals

For any two normal random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ the sum of those two random variables is another normal: $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

General Independent Case

For two general independent random variables (aka cases of independent random variables that don't fit the above special situations) you can calculate the CDF or the PDF of the sum of two random variables using the following formulas:

$$F_{X+Y}(a) = P(X + Y \leq a) = \int_{y=-\infty}^{\infty} F_X(a - y) f_Y(y) dy$$

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a - y) f_Y(y) dy$$

There are direct analogies in the discrete case where you replace the integrals with sums and change notation for CDF and PDF.

Example 1

Calculate the PDF of $X+Y$ for independent uniform random variables $X \sim Uni(0,1)$ and $Y \sim Uni(0,1)$? First plug in the equation for general convolution of independent random variables:

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a-y) f_Y(y) dy$$

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a-y) dy \qquad \text{Because } f_Y(y) = 1$$

It turns out that is not the easiest thing to integrate. By trying a few different values of a in the range $[0,2]$ we can observe that the PDF we are trying to calculate is discontinuous at the point $a = 1$ and thus will be easier to think about as two cases: $a < 1$ and $a > 1$. If we calculate f_{X+Y} for both cases and correctly constrain the bounds of the integral we get simple closed forms for each case:

$$f_{X+Y}(a) = \begin{cases} a & \text{if } 0 < a \leq 1 \\ 2-a & \text{if } 1 < a \leq 2 \\ 0 & \text{else} \end{cases}$$

7 Conditional Distributions

Before we looked at conditional probabilities for events. Here we formally go over conditional probabilities for random variables. The equations for both the discrete and continuous case are intuitive extensions of our understanding of conditional probability:

Discrete

The conditional probability mass function (PMF) for the discrete case:

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

The conditional cumulative density function (CDF) for the discrete case:

$$F_{X|Y}(a|y) = P(X \leq a|Y = y) = \frac{\sum_{x \leq a} p_{X,Y}(x,y)}{p_Y(y)} = \sum_{x \leq a} p_{X|Y}(x|y)$$

Continuous

The conditional probability density function (PDF) for the continuous case:

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

The conditional cumulative density function (CDF) for the continuous case:

$$F_{X|Y}(a|y) = P(X \leq a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y) dx$$

Mixing Discrete and Continuous

These equations are straightforward once you have your head around the notation for probability density functions ($f_X(x)$) and probability mass functions ($p_X(x)$).

Let X be continuous random variable and let N be a discrete random variable. The conditional probabilities of X given N and N given X respectively are:

$$f_{X|N}(x|n) = \frac{p_{N|X}(n|x)f_X(x)}{p_N(n)} \qquad p_{N|X}(n|x) = \frac{f_{X|N}(x|n)p_N(n)}{f_X(x)}$$

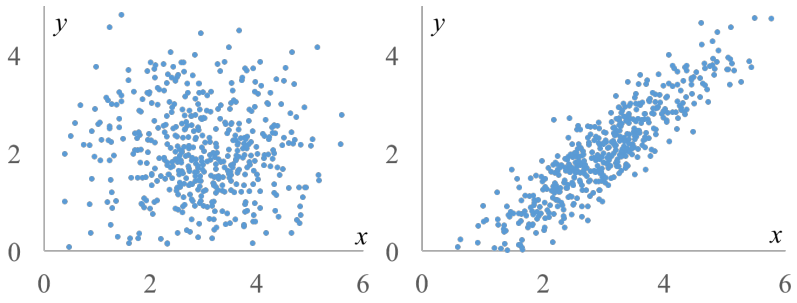
Example 2

Let's say we have two independent random Poisson variables for requests received at a web server in a day: X = # requests from humans/day, $X \sim Poi(\lambda_1)$ and Y = # requests from bots/day, $Y \sim Poi(\lambda_2)$. Since the convolution of Poisson random variables is also a Poisson we know that the total number of requests $(X+Y)$ is also a Poisson $(X+Y) \sim Poi(\lambda_1+\lambda_2)$. What is the probability of having k human requests on a particular day given that there were n total requests?

$$\begin{aligned}
P(X=k|X+Y=n) &= \frac{P(X=k,Y=n-k)}{P(X+Y=n)} = \frac{P(X=k)P(Y=n-k)}{P(X+Y=n)} \\
&= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{1(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n} \\
&= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \\
&\sim Bin\left(n, \frac{\lambda_2}{\lambda_1+\lambda_2}\right)
\end{aligned}$$

8 Covariance and Correlation

Consider the two multivariate distributions shown bellow. In both images I have plotted one thousand samples drawn from the underlying joint distribution. Clearly the two distributions are different. However, the mean and variance are the same in both the x and the y dimension. What is different?



Covariance is a quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean. It is a mathematical relationship that is defined as:

$$\text{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$$

That is a little hard to wrap your mind around (but worth pushing on a bit). The outer expectation will be a weighted sum of the inner function evaluated at a particular (x,y) weighted by the probability of (x,y). If x and y are both above their respective means, or if x and y are both below their respective means, that term will be positive. If one is above its mean and the other is below, the term is negative. If the weighted sum of terms is positive, the two random variables will have a positive correlation. We can rewrite the above equation to

get an equivalent equation:

$$\text{Cov}(X,Y) = E[XY] - E[Y]E[X]$$

Using this equation (and the product lemma) is it easy to see that if two random variables are independent their covariance is 0. The reverse is *not* true in general.

Properties of Covariance

Say that X and Y are arbitrary random variables:

$$\text{Cov}(X,Y) = \text{Cov}(Y,X)$$
$$\text{Cov}(X,X) = E[X^2] - E[X]E[X] = \text{Var}(X)$$
$$\text{Cov}(aX+b,Y) = a\text{Cov}(X,Y)$$

Let $X = X_1 + X_2 + \cdots + X_n$ and let $Y = Y_1 + Y_2 + \cdots + Y_m$. The covariance of X and Y is:

$$\text{Cov}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i,Y_j)$$
$$\text{Cov}(X,X) = \text{Var}(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i,X_j)$$

That last property gives us a third way to calculate variance. You could use this definition to calculate the variance of the binomial.

Correlation

Covariance is interesting because it is a quantitative measurement of the relationship between two variables. Correlation between two random variables, $\rho(X,Y)$ is the covariance of the two variables normalized by the variance of each variable. This normalization cancels the units out and normalizes the measure so that it is always in the range [0, 1]:

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)Var(Y)}}$$

Correlation measure linearity between X and Y.

$\rho(X,Y) = 1$	$Y = aX + b$ where $a = \sigma_y/\sigma_x$
$\rho(X,Y) = -1$	$Y = aX + b$ where $a = -\sigma_y/\sigma_x$
$\rho(X,Y) = 0$	absence of linear relationship

If $\rho(X,Y) = 0$ we say that X and Y are "uncorrelated." If two varaibles are independent, then their correlation will be 0. However, it doesn't go the other way. A correlation of 0 does not imply independence.

When people use the term correlation, they are actually referring to a specific type of correlation called "Pearson" correlation. It measures the degree to which there is a linear relationship between the two variables. An alternative measure is "Spearman" correlation which has a formula almost identical to your regular correlation score, with the exception that the underlying random variables are first transformed into their rank. "Spearman" correlation is outside the scope of CS109.