Chapter 6

Joint Distributions

6.1 Joint, marginal and conditional distributions

A joint distribution is a statistical concept used to describe the likelihood of two or more random variables occurring together. When we talk about joint distribution, we are considering the probability of different values of these variables happening simultaneously, rather than in isolation. Suppose we toss a coin and roll a die. Joint probability represents the probability that the two events happening simultaneously, e.g. $P(\text{Coin} = H, \text{Die} = 6)$.

Given the joint distribution, we are interested in: (i) the distribution of multi-variables simultaneously (joint probability); (ii) the distribution of one variable ignoring other variables (marginal probability); (iii) the distribution of one variable given the value of other variables (conditional probability).

	Discrete	Continuous
Joint CDF	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$	$F_{XY}(x,y) = P(X \leq x, Y \leq y)$
Joint PMF/PDF	$p_{XY}(x,y) = P(X = x, Y = y)$	$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$
	$\sum_x \sum_y P(X = x, Y = y) = 1$	$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y) dx dy = 1$
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y) dy$
Conditional PMF/PDF	$P(X = x \mid Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$	$f_{X\mid Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}$
Independence	$P(X = x, Y = y) = P(X = x) P(Y = y)$	$f_{XY}(x,y) = f_X(x) f_Y(y)$
	$P(X = x \mid Y = y) = P(X = x)$	$f_{X\mid Y}(x \mid y) = f_X(x)$
	$F_{XY}(x,y) = F_X(x) F_Y(y)$	$F_{XY}(x,y) = F_X(x) F_Y(y)$
Bayes' rule	$P(Y = y \mid X = x) = \frac{P(X=x \mid Y=y) P(Y=y)}{P(X=x)}$	$f_{Y\mid X}(y \mid x) = \frac{f_{X\mid Y}(x \mid y) f_Y(y)}{f_X(x)}$
LOTP	$P(X = x) = \sum_y P(X = x \mid Y = y) P(Y = y)$	$f_X(x) = \int_{-\infty}^{+\infty} f_{X\mid Y}(x \mid y) f_Y(y) dy$
LOTUS	$E(g(X,Y)) = \sum_x \sum_y g(x,y) P(X = x) P(y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy$

Table 6.1: Joint, marginal and conditional distributions

Example 6.1. Let X be an indicator of an individual being a current smoker. Let Y be the indicator of his developing lung cancer at some point in his life. The joint PMF of X and Y is as specified in the table below.

	$Y = 1$	$Y = 0$	**Total**
$X = 1$	0.05	0.20	**0.25**
$X = 0$	0.03	0.72	**0.75**
Total	**0.08**	**0.92**	**1**

The marginal PMF for having lung cancer is

$$P(Y = 1) = P(Y = 1, X = 0) + P(Y = 1, X = 1) = 0.08,$$
$$P(Y = 0) = P(Y = 0, X = 0) + P(Y = 0, X = 1) = 0.92.$$

The conditional PMF of having lung cancer conditioned on being a smoker is

$$P(Y = 1|X = 1) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{0.05}{0.25} = \frac{1}{5}.$$

In this example, X, Y are not independent, because

$$P(X = 1, Y = 1) \neq P(X = 1)P(Y = 1).$$

Example 6.2. Suppose X and Y are uniformly distributed on a disk $\{(x, y) : x^2 + y^2 \leq 1\}$. Find the joint PDF, marginal distributions and conditional distributions. Are X and Y independent?

Solution: The area of the disk is π, therefore

$$f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal distributions are

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1+x^2}} \frac{1}{\pi} dy = \frac{2}{\pi}\sqrt{1-x^2}, \qquad -1 \leq x \leq 1$$
$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1+y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}, \qquad -1 \leq y \leq 1$$

The conditional distributions are

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-x^2}} = \frac{1}{2\sqrt{1-x^2}}$$

Therefore, $Y|X \sim \text{Unif}(-\sqrt{1-x^2}, \sqrt{1-x^2})$.

Since $f(x,y) \neq f_X(x)f_Y(y)$, X and Y are not independent. This is because knowing the value of X constrains the value of Y.

Example 6.3. Suppose $X, Y \overset{iid}{\sim} Unif(0,1)$. Find the probability $P\left(Y \leq \frac{1}{2X}\right)$.

Solution: The joint distribution is

$$f(x,y) = \begin{cases} 1 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P\left(Y \leq \frac{1}{2X}\right) = \int_0^{1/2}\int_0^1 1dydx + \int_{1/2}^1\int_0^{1/2x} 1dydx = \frac{1}{2} + \int_{1/2}^1 \frac{1}{2x}dx = \frac{1}{2} + \ln\sqrt{2}.$$

Example 6.4. For $X, Y \overset{iid}{\sim} \text{Unif}(0,1)$, find $E(|X-Y|)$.

Solution: Apply 2D LOTUS:

$$\begin{aligned} E(|X-Y|) &= \int_0^1\int_0^1 |x-y|dxdy \\ &= \int_0^1\int_y^1 (x-y)dxdy + \int_0^1\int_0^y (y-x)dxdy \\ &= 2\int_0^1\int_y^1 (x-y)dxdy \\ &= \frac{1}{3}. \end{aligned}$$

Example 6.5. $X, Y \overset{iid}{\sim} N(0,1)$, find $E(|X-Y|)$.

Solution: Since the sum or difference of independent Normals is Normal, $X - Y \sim N(0,2)$. Let $Z = X - Y$. Then $Z \sim N(0,1)$, and $E(|X-Y|) = \sqrt{2}E(|Z|)$. Apply LOTUS,

$$E(|Z|) = \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 2\int_{0}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \sqrt{\frac{2}{\pi}},$$

Therefore, $\mathbb{E}(|X-Y|) = \frac{2}{\sqrt{\pi}}$.

6.2 Joint normal distribution

Definition 6.1. (X, Y) is said to have a **Bivariate Normal** distribution if the joint PDF satisfies

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right)$$

where $\rho \in (-1,1)$ is the correlation between X and Y.

A **Multivariate Normal (MVN)** is fully specified by knowing the mean of each component, the variance of each component, and the covariance or correlation between any two components. In other words, the parameters of an MVN random vector $(X_1, ..., X_k)$ are as follows:

- the mean vector $(\mu_1, ..., \mu_k)$, where $E(X_j) = \mu_j$;
- the covariance matrix $Cov(X_i, X_j)$ for $1 \le i, j \le k$.

If $(X_1, ..., X_k)$ is MVN, then the marginal distribution of every X_j is Normal. However, the converse is false: it is possible to have Normally distributed $X_1, ..., X_k$ such that $(X_1, ..., X_k)$ is not Multivariate Normal.

Theorem 6.1. *A random vector $(X_1, ..., X_k)$ is Multivariate Normal if every linear combination of the X_j has a Normal distribution. That is, we require $t_1X_1 + \cdots + t_kX_k$ to have a Normal distribution for any choice of constants $t_1, ..., t_k$.*

Theorem 6.2. *Within an MVN random vector, uncorrelated implies independent. In particular, if (X, Y) is Bivariate Normal and $Corr(X, Y) = 0$, then X and Y are independent.*

This is a special property of MVN random variables. In general, uncorrelated does not imply independent.

Theorem 6.3. *If (X, Y) is Bivariate Normal, then the conditional expectation satisfies*

$$E(Y|X) = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E(X)).$$

This is also a special property of MVN —$E(Y|X)$ is a linear function of X. This is not the case in general.

6.3 Conditional expectation

Theorem 6.4. *For any random variable X and Y,*

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

This is known as the ***law of iterated expectation****.*

Proof. Note that $E(Y|X) = g(X)$ is a function of X. Apply LOTUS:

$$\begin{aligned} E(E(Y|X)) &= \int g(x)f(x)dx \\ &= \int \left(\int yf(y|x)dy \right) f(x)dx \\ &= \int \int yf(y|x)f(x)dydx \\ &= \int y \int f(y, x)dx\, dy \\ &= \int_{-\infty}^{\infty} yf(y)dy \\ &= E(Y). \end{aligned}$$

□

Theorem 6.5. *For any random variable X and Y, and any function g,*

$$E(g(X)Y|X) = g(X)E(Y|X).$$

Proof. For any specific value of $X = x$, $g(x)$ is a constant. Thus, $E(g(x)Y|X = x) = g(x)E(Y|X = x)$. This is true for all values of x. □

Example 6.6 (PG exam)**.** Suppose $X \sim \text{Unif}(0, 1)$, and $Y|X \sim N(X, X^2)$, meaning that for a given $X = x$, Y is normally distributed with mean x and variance x^2. Find $E(Y)$, $Var(Y)$ and $Cov(X, Y)$.

Solution:

Since $Y|X \sim N(X, X^2)$, we know $E(Y|X) = X$. By the law of iterated expectation,

$$E(Y) = E(E(Y|X)) = E(X) = \frac{1}{2}.$$

For the variance,

$$Var(Y) = E(Y^2) - (E(Y))^2 = E(E(Y^2|X)) - \frac{1}{4}.$$

Since

$$Var(Y|X) = E(Y^2|X) - E^2(Y|X) = E(Y^2|X) - X^2 = X^2,$$

we have $E(Y^2|X) = 2X^2$. Meanwhile, $E(X^2) = \int_0^1 x^2 \cdot 1dx = \frac{1}{3}$. Therefore,

$$Var(Y) = \frac{2}{3} - \frac{1}{4} = \frac{5}{12}.$$

For the covariance,

$$E(XY) = E(E(XY|X)) = E(XE(Y|X)) = E(X^2) = \frac{1}{3},$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Theorem 6.6. *Conditional expectation $E(Y|X)$ is the best predictor for Y using X (minimized the square loss function).*

Proof. Let $g(X)$ be a predictor for Y using X. We want to find the g such that minimizes $E(Y - g(X))^2$.

$$\begin{aligned}E(Y - g(X))^2 &= E(Y - E(Y|X) + E(Y|X) - g(X))^2 \\ &= E(Y - E(Y|X))^2 + 2\underbrace{E(Y - E(Y|X)}_{E(Y)=E(E(Y|X))}((E(Y|X) - g(X)) + E(E(Y|X) - g(X))^2 \\ &= E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2 \\ &\geq E(Y - E(Y|X))^2.\end{aligned}$$

Therefore, $E(Y - g(X))^2$ is minimized when $g(X) = E(Y|X)$. □

6.4 Linear conditional expectation model

Definition 6.2. An extremely widely used method for data analysis in statistics is linear regression. In its most basic form, we want to predict the mean of Y using a single explanatory variable X. **A linear conditional expectation model** assumes that $E(Y|X)$ is linear in X:

$$E(Y|X) = a + bX,$$

or equivalently,

$$Y = a + bX + \epsilon,$$

with $E(\epsilon|X) = 0$. The intercept and the slope is given by

$$b = \frac{Cov(X, Y)}{Var(X)}, a = E(Y) - bE(X).$$

We first show the equivalence of the two expressions of the model. Let $Y = a + bX + \epsilon$, with $E(\epsilon|X) = 0$. Then by linearity,

$$E(Y|X) = E(a|X) + E(bX|X) + E(\epsilon|X) = a + bX.$$

Conversely, suppose that $E(Y|X) = a + bX$, and define

$$\epsilon = Y - (a + bX).$$

Then $Y = a + bX + \epsilon$, with

$$E(\epsilon|X) = E(Y|X) - E(a + bX|X) = E(Y|X) - (a + bX) = 0.$$

To derive the expression for a and b, take covariance between X and Y,

$$\begin{aligned} Cov(X,Y) &= Cov(X, a + bX + \epsilon) \\ &= Cov(X,a) + bCov(X,X) + Cov(X,\epsilon) \\ &= bVar(X) + Cov(X,\epsilon) \end{aligned}$$

Note that $Cov(X,\epsilon) = 0$ because

$$\begin{aligned} Cov(X,\epsilon) &= E(X\epsilon) - E(X)E(\epsilon) \\ &= E(E(X\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= E(XE(\epsilon|X)) - E(X)E(E(\epsilon|X)) \\ &= 0 \end{aligned}$$

Therefore,

$$Cov(X,Y) = bVar(X)$$

Thus,

$$\begin{aligned} b &= \frac{Cov(X,Y)}{Var(X)}, \\ a &= E(Y) - bE(X) = E(Y) - \frac{Cov(X,Y)}{Var(X)}E(X). \end{aligned}$$

6.5 Change of variables

Theorem 6.7. *Let X be a continuous r.v. with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly increasing (or strictly decreasing). Then the PDF of Y is given by*

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|,$$

where $x = g^{-1}(y)$.

Proof. Let g be strictly increasing. The CDF of Y is

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)) = F_X(x)$$

By the chain rule, the PDF of Y is

$$f_Y(y) = f_X(x)\frac{dx}{dy}.$$

If g is strictly decreasing,

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) = 1 - F_X(x)$$

Then the PDF of Y is

$$f_Y(y) = -f_X(x)\frac{dx}{dy}.$$

But in this case, $dx/dy < 0$. So taking absolute value covers both cases. □

Example 6.7 (Log-Normal PDF). Let $X \sim N(0,1)$, $Y = e^X$. Then the distribution of Y is called the **Log-Normal distribution**. Find the PDF of Y.

Since $g(x) = e^x$ is strictly increasing. Let $y = e^x$, so $x = \log y$ and $dy/dx = e^x$. Then

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \varphi(x)\frac{1}{e^x} = \varphi(\log y)\frac{1}{y}, \quad y > 0.$$

Note that after applying the change of variables formula, we write everything on the right-hand side in terms of y, and we specify the support of the distribution. To determine the support, we just observe that as x ranges from $-\infty$ to∞, e^x ranges from 0 to ∞.

Example 6.8 (Chi-Square PDF). Let $X \sim N(0,1)$, $Y = X^2$. The distribution of Y is an example of a **Chi-Square distribution**. Find the PDF of Y.

In this case, we can no longer apply the change of variables formula because $g(x) = x^2$ is not one-to-one. Instead, we use the CDF:

$$F_Y(y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

Therefore,

$$f_Y(y) = 2\varphi(\sqrt{y}) \cdot \frac{1}{2}y^{-1/2} = \varphi(\sqrt{y})y^{-1/2}, \quad y > 0.$$

Theorem 6.8. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a continuous random vector with joint PDF* $f_{\mathbf{X}}$*, and let* $\mathbf{Y} = g(\mathbf{X})$ *where* g *is an invertible function from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$*. Let* $\mathbf{y} = g(\mathbf{x})$*. Define the Jacobian matrix:*

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}.$$

Also assume that the determinant of the Jacobian matrix is never 0. Then the joint PDF of $\mathbf{Y}$ *is*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|,$$

where $\left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|$ *is the absolute value of the determinant of the Jacobian matrix.*

Example 6.9. Suppose $X, Y \stackrel{iid}{\sim} Expo(1)$. Find the distribution of $X/(X+Y)$.

Solution: Let $U = \frac{X}{X+Y}$, $V = X + Y$. Then $X = UV$, $Y = V - UV$. The determinant of the Jacobian matrix is

$$\left| \frac{\partial(x, y)}{\partial(u, v)} \right| = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Thus, the joint distribution of (U, V) is

$$f_{UV}(u, v) = f_{XY}(x, y)|v| = f_X(x) f_Y(y) v = e^{-(x+y)} v = e^{-v} v.$$

The distribution of $X/(X+Y)$ is equivalent to the marginal distribution of U:

$$f_U(u) = \int_0^\infty e^{-v} v dv = 1$$

for $0 \leq u \leq 1$. Hence U is a Uniform distribution over [0,1].