1 Joint, marginal and conditional distributions for discrete random variables

The individual distributions of random variables provide no information about whether the random variables are independent or dependent. We refer to the individual distribution of a random variable X as the *marginal* distribution of X. The *joint* and *marginal* distributions express the dependence or independence of X with respect to another random variable Y. The joint distribution of X and Y provides complete information about the probability of the vector (X, Y) falling into any subset of the plane. The conditional distribution of X given $Y = y$ is the updated distribution of X after observing $Y = y$.

Definition (Joint CDF)**.** The *joint* CDF of random variables X and Y is the function $F_{X,Y}$ given by

$$F_{X,Y}(x, y) = \mathsf{P}(X \leq x, Y \leq y).$$

The joint CDF of n random variables $X_1, \ldots, X_n$ is the function $F_{X_1,\ldots,X_n}$ given by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathsf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Definition (Joint PMF)**.** The *joint* PMF of discrete random variables X and Y is the function $p_{X,Y}$ given by

$$p_{X,Y}(x, y) = \mathsf{P}(X = x, Y = y).$$

The joint PMF of n discrete random variables $X_1, \ldots, X_n$ is the function $p_{X_1,\ldots,X_n}$ given by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Valid joint PMFs must be nonnegative and sum to 1, where the sum is taken over all possible values of X and Y:

$$\sum_x \sum_y \mathsf{P}(X = x, Y = y) = 1.$$

See Figure 1. The joint PMF determines the distribution because it can be used to find the probability of the event $(X, Y) \in A$ for any set A in the plane $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ by summing the joint PMF over A:

$$\mathsf{P}((X, Y) \in A) = \sum_{(x,y)\in A} \mathsf{P}(X = x, Y = y).$$

From the joint distribution of X and Y, we can get the distribution of X alone by summing over the possible values of Y. In this context, we will call it the marginal or unconditional distribution of X – see Figure 2.

Definition (Marginal PMF)**.** For discrete random variables X and Y, the marginal PMF of X is

$$\mathsf{P}(X = x) = \sum_y \mathsf{P}(X = x, Y = y).$$

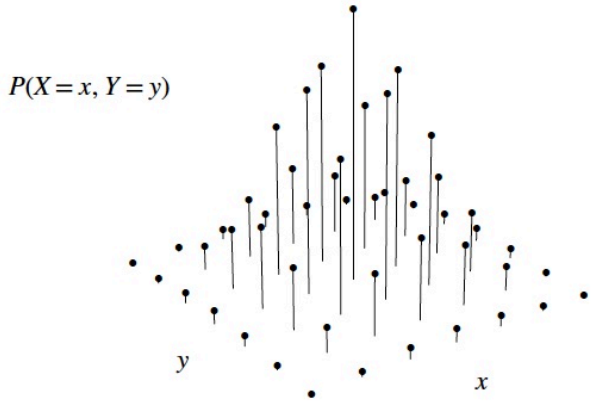

Figure 1: Joint PMF of discrete random variables X and Y. The height of a vertical bar at (x, y) represents the joint probability $\mathsf{P}(X = x, Y = y)$. For the joint PMF to be valid, the total height of the vertical bars must be 1.

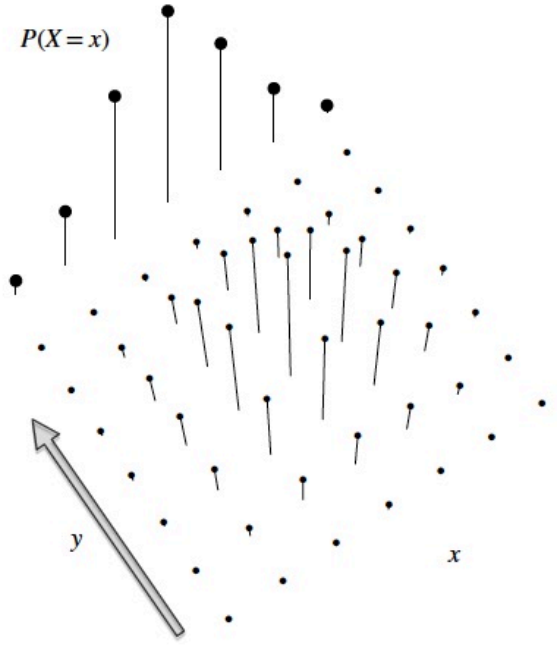

Figure 2: This is the same joint PMF as in Figure 2. The marginal PMF $\mathsf{P}(X = x)$ is obtained by summing over the joint PMF in the y-direction, as indicated by the arrow.

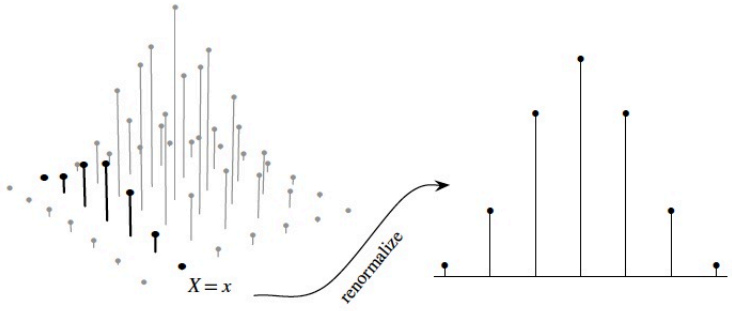

Figure 3: Conditional PMF of Y given $X = x$. To condition on the event $X = x$, we first take the joint PMF and focus in on the vertical bars where $X = x$ (shown in bold). All the other vertical bars are inconsistent with the knowledge that $X = x$ has occurred. Since the total height of the bold bars is the marginal probability $\mathsf{P}(X = x)$, we then renormalized the conditional PMF by dividing by $\mathsf{P}(X = x)$.

Another way to obtain marginal distributions from joint distributions is via the joint CDF:

$$F_X(x) = \mathsf{P}(X \leq x) = \lim_{y\to\infty} \mathsf{P}(X \leq x, Y \leq y) = \lim_{y\to\infty} F_{X,Y}(x, y).$$

This holds for discrete as well as for continuous random variables.

Now suppose that we observe the value of X, and want to update the distribution of Y to reflect this information, i.e. to reflect that the event $X = x$ has occurred. To this end, we should use a PMF that conditions on the event $X = x$.

Definition (Conditional PMF)**.** For discrete random variables X and Y, the *conditional PMF* of Y given $X = x$ is

$$\mathsf{P}(Y = y \mid X = x) = \frac{\mathsf{P}(X = x, Y = y)}{\mathsf{P}(X = x)}.$$

This is viewed as a function of y for fixed x. Note that the conditional PMF (for fixed x) is a valid PMF. Conditional PMFs are PMFs. See Figure 3.

We can relate the conditional distribution of Y given $X = x$ to that of X given $Y = y$ using Bayes' rule:

$$\mathsf{P}(Y = y \mid X = x) = \frac{\mathsf{P}(X = x \mid Y = y)\mathsf{P}(Y = y)}{\mathsf{P}(X = x)}.$$

The law of total probability tells that that the marginal of X is a weighted average of conditional PMFs $\mathsf{P}(X = x \mid Y = y)$, where the weights are the probabilities $\mathsf{P}(Y = y)$:

$$\mathsf{P}(X = x) = \sum_y \mathsf{P}(X = x \mid Y = y)\mathsf{P}(Y = y).$$

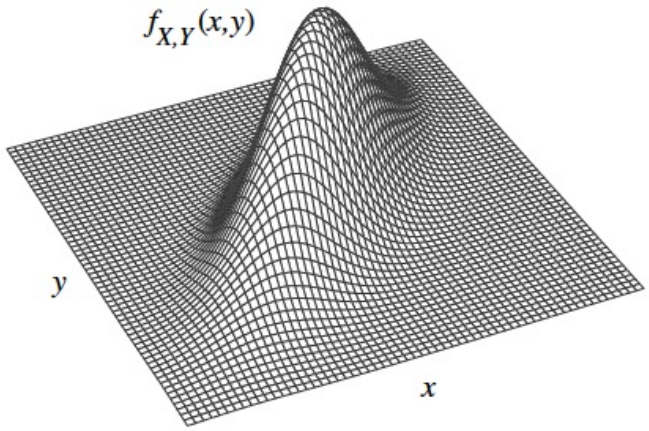

Figure 4: Joint PDF of continuous random variables X and Y. The height of the surface $f_{X,Y}(x,y)$ at a single point does not represent a probability. The probability of any specific point in the plane is 0. The probability of any line or curve in the plane is also 0. When we integrate over a region of positive area in the plane we get a non-zero probability.

2 Joint, marginal and conditional distributions for continuous random variables

Definition (Joint PDF)**.** If X and Y are continuous with joint CDF

$$F_{X,Y}(x,y) = \mathsf{P}(X \le x, Y \le y),$$

their *joint* PDF is the derivative of the joint CDF with respect to x and y:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

We require valid joint PDFs to be nonnegative and integrate to 1:

$$f_{X,Y}(x,y) \ge 0, \text{ and } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$

The joint PDF of two random variable is the function we integrate to get the probability of a two-dimensional region $A \subseteq \mathbb{R}^2$:

$$\mathsf{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dx\,dy.$$

When we integrate the joint PDF over an area A, we are calculating the volume under the surface of the joint PDF and above A. See Figure 4.

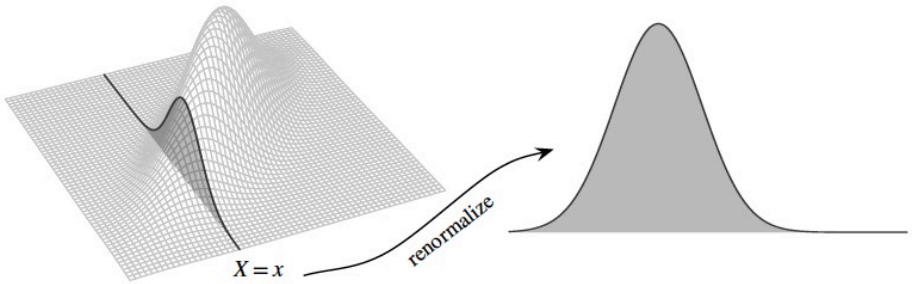

Figure 5: Conditional PDF of Y given $X = x$. We take the vertical slice of the joint PDF corresponding to $X = x$. Since the total area under this slice is $f_X(x)$, we then renormalize (divide by) $f_X(x)$ to ensue that the conditional PDF $f_{Y|X}(y \mid x)$ has an area of 1.

Definition (Marginal PDF)**.** For continuous random variables X and Y with joint PDF $f_{X,Y}$, the *marginal* PDF of X is

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

This is the PDF of X, viewing X individually rather than jointly with Y.

Definition (Conditional PDF)**.** For continuous random variables X and Y with joint PDF $f_{X,Y}$, the *conditional* PDF of Y given $X = x$ is

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

This is considered as a function of y for fixed x. Conditional PDFs satisfy the properties of valid PDFs – see Figure 5:

$$f_{Y|X}(y \mid x) \geq 0, \quad \int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, dy = 1.$$

Note that the following relations among the joint, marginal and conditional PDFs hold:

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

Theorem (Continuous version of Bayes' rule and LOTP)**.** *For continuous random variables X and Y, we have*

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)}, \quad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\, dy.$$

Definition (Independence of random variables)**.** Random variables X and Y are *independent* if for all x and y, we have

$$F_{X,Y}(x, y) \quad = \quad F_X(x) F_Y(y). \tag{1}$$

If X and Y are discrete, condition (1) is equivalent with

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x)\mathsf{P}(Y = y),$$

for all x and y, and it is also equivalent with

$$\mathsf{P}(Y = y \mid X = x) = \mathsf{P}(Y = y).$$

for all y and all x such that $\mathsf{P}(X = x) > 0$.

If X and Y are continuous with joint PDF $f_{X,Y}$, condition (1) is equivalent with

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

for all x and y, and it is also equivalent with

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

for all y and all x such that $f_X(x) > 0$.

Theorem. *Suppose that the joint PDF $f_{X,Y}$ of X and Y factors as*

$$f_{X,Y}(x, y) = g(x)h(y),$$

for all x and y in the xy plane $\mathbb{R}^2$, where g and h are nonnegative functions. Then X and Y are independent. Moreover, if either g or h is a valid PDF, then the other one is a valid PDF too, and g and h are the marginal PDFs of X and Y, respectively. Remember that the analogous result for X and Y discrete random variables also holds.

Example: Uniform on a region in the plane

Let (X, Y) be a completely random point in the square $\{(x, y) : x, y \in [0, 1]\}$, in the sense that the joint PDF of X and Y is constant over the square and 0 outside it:

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } x, y \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

This is called the Uniform distribution on the square. The marginal distribution of X is $\mathsf{Unif}(0, 1)$:

$$f_X(x) = \int_0^1 f_{X,Y}(x, y)\, dy = \int_0^1 1\, dy = 1.$$

The marginal distribution of Y is also $\mathsf{Unif}(0, 1)$. Is X independent of Y? We have

$$f_{X,Y}(x, y) = 1 = 1 \cdot 1 = f_X(x) f_Y(y),$$

for any $x, y \in [0, 1]$. Moreover, we also have

$$f_{X,Y}(x, y) = 0 = 0 \cdot 0 = f_X(x) f_Y(y),$$

for any $x, y \in (-\infty, 0) \cup (1, \infty)$. From these two relations we can conclude that X and Y are indeed independent. However, the key thing to note is that these two relations hold because the value of X does not constrain the possible values of Y, and vice versa.

Next we let (X, Y) be a completely random point in the unit disk $\{(x, y) : x^2 + y^2 = 1\}$, with joint PDF:

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

In this case X and Y are not independent because knowing the values of X constrains the possible values of Y: larger values of $|X|$ restrict Y to be in a smaller range – see Figure 6. It would be incorrect to argue that X and Y are independent because

$$f_{X,Y}(x, y) = g(x)h(y),$$

for all (x, y) in the unit disk, where $g(x) = 1/\pi$ and $h(y) = 1$ are constant functions.

The marginal distribution of X is not $\mathsf{Unif}(0, 1)$:

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy = \frac{2}{\pi}\sqrt{1 - x^2}, \text{ for } -1 \leq x \leq 1.$$

Similarly, the PDF of Y is

$$f_Y(y) = \frac{2}{\pi}\sqrt{1 - y^2}, \quad \text{for } -1 \leq y \leq 1.$$

Note that X and Y are more likely to fall near 0 than near -1 or 1.

The conditional distribution of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{1}{2\sqrt{1 - x^2}}, \text{ for } -\sqrt{1 - x^2} \leq y \leq \sqrt{1 - x^2}.$$

Since the conditional PDF is constant as a function of y, the conditional distribution of Y given $X = x$ is $\mathsf{Unif}[-\sqrt{1 - x^2}, \sqrt{1 - x^2}]$.

Theorem (2D LOTUS)**.** *Let g be a function from $\mathbb{R}^2$ to $\mathbb{R}$. If X and Y are discrete, then*

$$\mathsf{E}(g(X, Y)) = \sum_x \sum_y g(x, y)\mathsf{P}(X = x, Y = y).$$

If X and Y are continuous with joint PDF $f_{X,Y}$, then

$$\mathsf{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

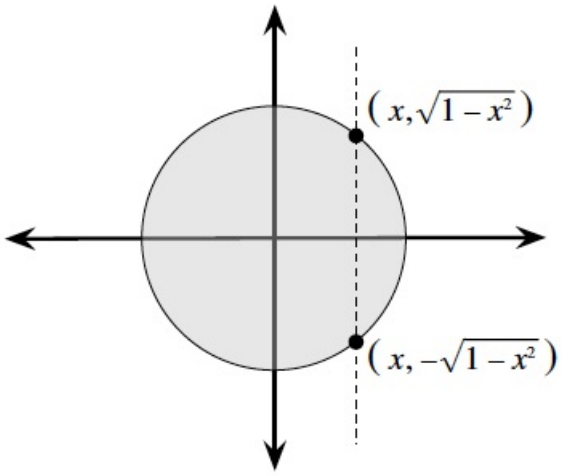

Figure 6: Uniform joint PDF on the unit disk. Conditional on $X = x$, Y is restricted to the interval $[-\sqrt{1-x^2}, \sqrt{1-x^2}]$.

Example: Expected distance between two Normals
For X and Y i.i.d. $\mathsf{N}(0,1)$, find $\mathsf{E}(|X-Y|)$.

Solution: 2D LOTUS gives:

$$\mathsf{E}(|X-Y|) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x-y| \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dx\, dy.$$

While it is possible to calculate this double integral, we can take an easier path by exploiting the properties of Normal distributions. Since $X, Y \sim \mathsf{N}(0,1)$, and X and Y are independent, we have $X - Y \sim \mathsf{N}(0,2)$. Thus $X - Y = \sqrt{2}Z$ with $Z \sim \mathsf{N}(0,1)$. Thus

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

3 Covariance and correlation

Covariance measures a tendency of two random variables to go up and down together, relative to their expected values: positive covariance between X and Y indicates that when X goes up, Y also tends to go up, and negative covariance indicates that when X goes up, Y tends to go down.

Definition (Covariance)**.** The *covariance* between the random variables X and Y is

$$\mathsf{Cov}(X,Y) = \mathsf{E}((X - \mathsf{E}X)(Y - \mathsf{E}Y)) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y).$$

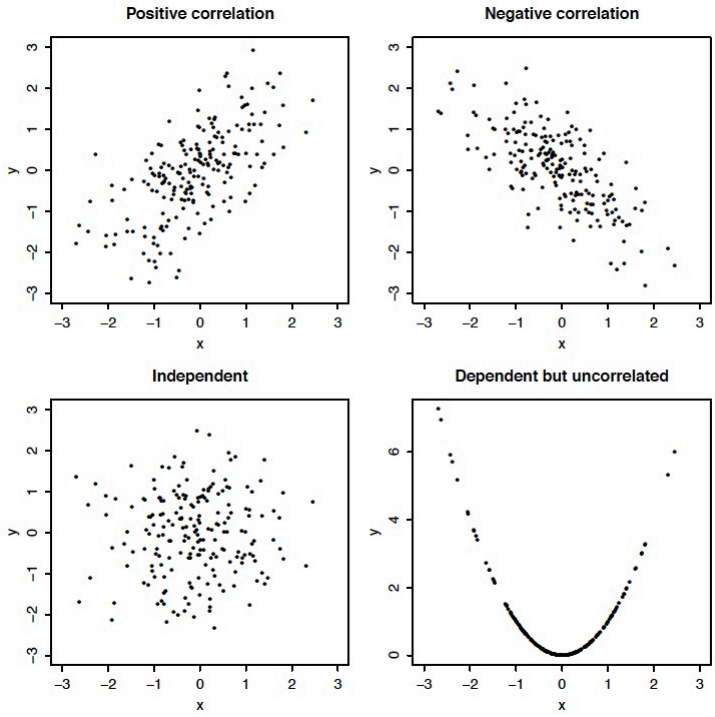

Figure 7: Samples from the joint distribution of (X, Y) under four dependence structures. Top left: X and Y are positively correlated. Top right: X and Y are negatively correlated. Bottom left: X and Y are independent, hence uncorrelated. Bottom right: Y is a deterministic function of X: $Y = X^2$, and $X \sim \mathsf{N}(0, 1)$. Since $\mathsf{E}(XY) = \mathsf{E}(X^3) = 0$, $\mathsf{Cov}(X, Y) = \mathsf{E}(X^3) - \mathsf{E}(X)\mathsf{E}(X^2) = 0$, hence X and Y are uncorrelated. However, X and Y are certainly not independent: knowing $X = x$ gives us perfect information about Y, namely $Y = x^2$.

Theorem. *If X and Y are independent, then their covariance is zero. We say that X and Y are uncorrelated.*

The converse of this result is false: there are variables that are uncorrelated and dependent – see Figure 7.

Covariance has the following key properties:

1. $\mathsf{Cov}(X, X) = \mathsf{Var}(X)$.
2. Symmetry: $\mathsf{Cov}(X, Y) = \mathsf{Cov}(Y, X)$.
3. $\mathsf{Cov}(X, c) = 0$ for any constant $c \in \mathbb{R}$.
4. $\mathsf{Cov}(aX, Y) = a\mathsf{Cov}(X, Y)$ for any constant $a \in \mathbb{R}$.
5. $\mathsf{Cov}(X + Y, Z) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, Z)$.
6. $\mathsf{Cov}(X + Y, Z + W) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(X, W) + \mathsf{Cov}(Y, Z) + \mathsf{Cov}(Y, W)$.
7. $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X, Y)$.

8. For any random variables $X_1, \ldots, X_n$, we have

$$\mathsf{Var}(X_1 + \ldots + X_n) = \mathsf{Var}(X_1) + \ldots + \mathsf{Var}(X_n) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j).$$

Example: Hypergeometric variance

Let $X \sim \mathsf{HGeom}(w, b, n)$. Find $\mathsf{Var}(X)$.

Solution: X can be considered to be the number of white balls in a sample of size n drawn without replacement from an urn that contains w white balls and b black balls. We represent X as a sum of indicator random variables:

$$X = \mathsf{I}_1 + I_2 + \ldots + I_n,$$

where I_j is the indicator of the jth ball in the sample being white. Then $\mathsf{I}_j \sim \mathsf{Bern}(p)$ with $p = \frac{w}{w+b}$, i.e., $\mathsf{I}_1, \mathsf{I}_2, \ldots, \mathsf{I}_n$ are identically distributed. But $\mathsf{I}_1, \mathsf{I}_2, \ldots, \mathsf{I}_n$ are certainly not independent. The properties of covariance help:

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{Var}(\mathsf{I}_1 + I_2 + \ldots + I_n), \\
&= \mathsf{Var}(\mathsf{I}_1) + \mathsf{Var}(\mathsf{I}_2) + \ldots + \mathsf{Var}(\mathsf{I}_n) + 2\sum_{i<j} \mathsf{Cov}(\mathsf{I}_i, \mathsf{I}_j), \\
&= np(1-p) + 2\binom{n}{2}\mathsf{Cov}(\mathsf{I}_1, \mathsf{I}_2).
\end{aligned}$$

By the fundamental bridge:

$$\begin{aligned}
\mathsf{Cov}(\mathsf{I}_1, \mathsf{I}_2) &= \mathsf{E}(\mathsf{I}_1\mathsf{I}_2) - \mathsf{E}(\mathsf{I}_1)\mathsf{E}(\mathsf{I}_2), \\
&= \mathsf{P}(\text{first and second balls are both white}) - \mathsf{P}(\text{first ball is white})\mathsf{P}(\text{second ball is white}), \\
&= \frac{w}{w+b} \cdot \frac{w-1}{w+b-1} - p^2, \\
&= p\left(\frac{Np-1}{N-1} - p\right),
\end{aligned}$$

where $N = w + b$. It follows that

$$\mathsf{Var}(X) = \frac{N-n}{N-1} np(1-p).$$

Definition (Correlation)**.** The *correlation* between random variables X and Y is

$$\mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}$$

Since shifting (adding a constant to a random variable) does not affect $\mathsf{Cov}(X, Y)$, $\mathsf{Var}(X)$ and $\mathsf{Var}(Y)$, it will leave $\mathsf{Corr}(X, Y)$ unchanged. Scaling (multiplying a random variable by a constant) also leaves the correlation unchanged. For $c \in \mathbb{R}$, we have:

$$\mathsf{Corr}(cX, Y) = \frac{c\mathsf{Cov}(X, Y)}{\sqrt{c^2\mathsf{Var}(X)\mathsf{Var}(Y)}} = \mathsf{Corr}(X, Y).$$

Moreover, correlation is easy to interpret because it does not depend on the units of measurement of the two random variables.

Theorem (Correlation bounds)**.** *For any random variables X and Y, we have*

$$-1 \leq \mathsf{Corr}(X, Y) \leq 1.$$

Proof. Since scaling does not change correlation, we can assume $\mathsf{Var}(X) = \mathsf{Var}(Y) = 1$. Denote $\rho = \mathsf{Corr}(X, Y) = \mathsf{Cov}(X, Y)$. We write

$$\begin{aligned} 0 \leq \mathsf{Var}(X + Y) &= \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X, Y) = 2 + 2\rho, \\ 0 \leq \mathsf{Var}(X - Y) &= \mathsf{Var}(X) + \mathsf{Var}(Y) - 2\mathsf{Cov}(X, Y) = 2 - 2\rho. \end{aligned}$$

Thus $-1 \leq \rho \leq 1$. □

4 The Multinomial distribution

The Multinomial distribution is a generalization of the Binomial. Whereas the Binomial distribution counts the successes in a fixed number of trials that can only be categorized as success or failure, the Multinomial distribution keeps track of trials whose outcomes can fall into multiple categories, such as excellent, adequate, poor; or red, yellow, green, blue.

More concretely, we assume that n objects are independently places into one of k categories. An object is placed in category j with probability $p_j \geq 0$. We must have $\sum_{j=1}^{k} p_j = 1$. Let X_j the number of objects in category j, $j = 1, 2, \ldots, k$ such that $X_1 + X_2 + \ldots + X_k = n$. The random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is said to have a *Multinomial distribution* with parameters n and $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$. We write $\mathbf{X} \sim \mathsf{Mult}_k(n, \boldsymbol{p})$.

Theorem (Multinomial joint PMF)**.** *If* $\mathbf{X} \sim \mathsf{Mult}_k(n, \boldsymbol{p})$*, then the joint PMF of the random vector* $\mathbf{X}$ *is*

$$\mathsf{P}(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

where $n_1 + n_2 + \ldots + n_k = n$*.*

Theorem (Multinomial margins)**.** *If* $\mathbf{X} \sim \mathsf{Mult}_k(n, \boldsymbol{p})$*, then* $X_j \sim \mathsf{Bin}(n, p_j)$*.*

Theorem (Multinomial lumping)**.** *If* $\mathbf{X} \sim \mathsf{Mult}_k(n, \boldsymbol{p})$*, then for any distinct i and j,* $X_i + X_j \sim \mathsf{Bin}(n, p_i + p_j)$*. The random vector of counts obtained from merging categories i and j is still Multinomial. For example, merging categories* 1 *and* 2 *gives*

$$(X_1 + X_2, X_3, \ldots, X_k) \sim \mathsf{Mult}_{k-1}(n, (p_1 + p_2, p_3, \ldots, p_k)).$$

Theorem (Multinomial conditioning)**.** *If* $\mathbf{X} \sim \mathsf{Mult}_k(n, \boldsymbol{p})$*, then*

$$(X_2, \ldots, X_k) \mid X_1 = n_1 \sim \mathsf{Mult}_{k-1}(n - n_1, (p_2', \ldots, p_k')),$$

where $p_j' = p_j/(p_2 + \ldots + p_k)$*.*

Proof. Given that there are n_1 objects in category 1, the remaining $n - n_1$ objects fall into categories 2 to k, independently of one another. By Bayes' theorem, the conditional probability of falling into category j is

$$\mathsf{P}(\text{in category } j \mid \text{not in category 1}) = \frac{\mathsf{P}(\text{in category } j)}{\mathsf{P}(\text{not in category 1})} = \frac{p_j}{p_2 + \ldots + p_k}, \quad j = 2, \ldots, k.$$

□

Theorem (Covariance in a Multinomial)**.** *Let* $(X_1, \ldots, X_k) \sim \mathsf{Mult}_k(n, \boldsymbol{p})$*, where* $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$*. For* $i \neq j$*,* $\mathsf{Cov}(X_i, X_j) = -np_ip_j$*.*

Proof. We know that $X_i + X_j \sim \mathsf{Bin}(n, p_i + p_j)$, $X_i \sim \mathsf{Bin}(n, p_i)$, $X_j \sim \mathsf{Bin}(n, p_j)$. Thus

$$\begin{aligned} \mathsf{Var}(X_i + X_j) &= \mathsf{Var}(X_i) + \mathsf{Var}(X_j) + 2\mathsf{Cov}(X_i, X_j), \\ n(p_i + p_j)(1 - (p_i + p_j)) &= np_i(1 - p_i) + np_j(1 - p_j) + 2\mathsf{Cov}(X_i, X_j). \end{aligned}$$

Then $\mathsf{Cov}(X_i, X_j) = -np_ip_j$. □

5 The Multivariate Normal distribution

The Multivariate Normal distribution is a continuous multivariate distribution that generalizes the Normal into higher dimensions.

Definition (Multivariate Normal distribution)**.** A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is said to have a *Multivariate Normal* (MVN) distribution if every linear combination of the X_j has a Normal distribution. That is, we require

$$t_1X_1 + t_2X_2 + \ldots + t_kX_k$$

to have a Normal distribution for any $t_1, t_2, \ldots, t_k \in \mathbb{R}$. For $k = 2$ this distribution is called the *Bivariate Normal* (BVN). The marginal distribution of any component X_j of the random vector $\mathbf{X}$ is Normal which can be seen by taking $t_j = 1$ and $t_{j'} = 0$ for $j' \neq j$.

The next example shows that it is possible to have Normally distributed random variables $X_1, \ldots, X_k$ such that $\mathbf{X} = (X_1, \ldots, X_k)$ is not Multivariate Normal.

Example:
Let $X \sim \mathsf{N}(0, 1)$ and let

$$S = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

be a *random sign* independent of X. Then, from the symmetry of the standard Normal distribution, it follows that $Y = SX \sim \mathsf{N}(0, 1)$:

$$\mathsf{P}(Y \leq x) = \mathsf{P}(SX \leq x) = \mathsf{P}(X \leq x)\mathsf{P}(S = 1) + \mathsf{P}(-X \leq x)\mathsf{P}(S = -1) = \mathsf{P}(X \leq x),$$

since $\mathsf{P}(-X \leq x) = \mathsf{P}(X \geq -x) = \mathsf{P}(X \leq x)$. However, we have:

$$\mathsf{P}(X + Y = 0) = \mathsf{P}(S = -1) = \frac{1}{2},$$

which implies that $X + Y$ does not have a continuous distribution, and hence $X + Y$ cannot have a Normal distribution. This implies that the random vector (X, Y) cannot have a Bivariate Normal distribution although both X and Y have a standard Normal distribution.

Example:
Consider Z and W to be i.i.d. $\mathsf{N}(0, 1)$. Then (Z, W) is Bivariate Normal since the sum of independent Normals is also Normal. Furthermore, $(Z + 2W, 3Z + 5Z)$ is also Bivariate Normal since any linear combination of its components can be expressed as a linear combination of Z and W:

$$t_1(Z + 2W) + t_2(3Z + 5W) = (t_1 + 3t_2)Z + (2t_1 + 5t_2)W.$$

Theorem. *If (X_1, X_2, X_3) is Multivariate Normal, then so is the subvector (X_1, X_2). In fact, any subvector of a Multivariate Normal random vector $(X_1, X_2, \ldots, X_k)$ is also Multivariate Normal.*

Theorem. *If $\mathbf{X} = (X_1, \ldots, X_k)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ are Multivariate Normal random vectors with $\mathbf{X}$ independent of $\mathbf{Y}$, then the concatenated random vector*

$$\mathbf{W} = (\mathbf{X}, \mathbf{Y}) = (X_1, \ldots, X_k, Y_1, \ldots, Y_m)$$

also has a Multivariate Normal distribution.

A Multivariate Normal distribution is fully specified by knowing the mean of each component, the variance of each component, and the covariance or correlation between any two components. The parameters of an MVN random vector $\mathbf{X} = (X_1, \ldots, X_k)$ are as follows:

- the *mean vector* $(\mu_1, \ldots, \mu_k)$, where $\mathsf{E}(X_j) = \mu_j$.
- the *covariance matrix*, which is a $k \times k$ matrix of covariances between components, arranged so that the row i, column j entry is $\mathsf{Cov}(X_i, X_j)$.

In order to specify a Bivariate Normal distribution for (X, Y), we need to know five parameters – see Figure 8:

- the means $\mathsf{E}(X)$, $\mathsf{E}(Y)$.
- the variances $\mathsf{Var}(X)$, $\mathsf{Var}(Y)$.
- the correlation $\mathsf{Corr}(X, Y)$.

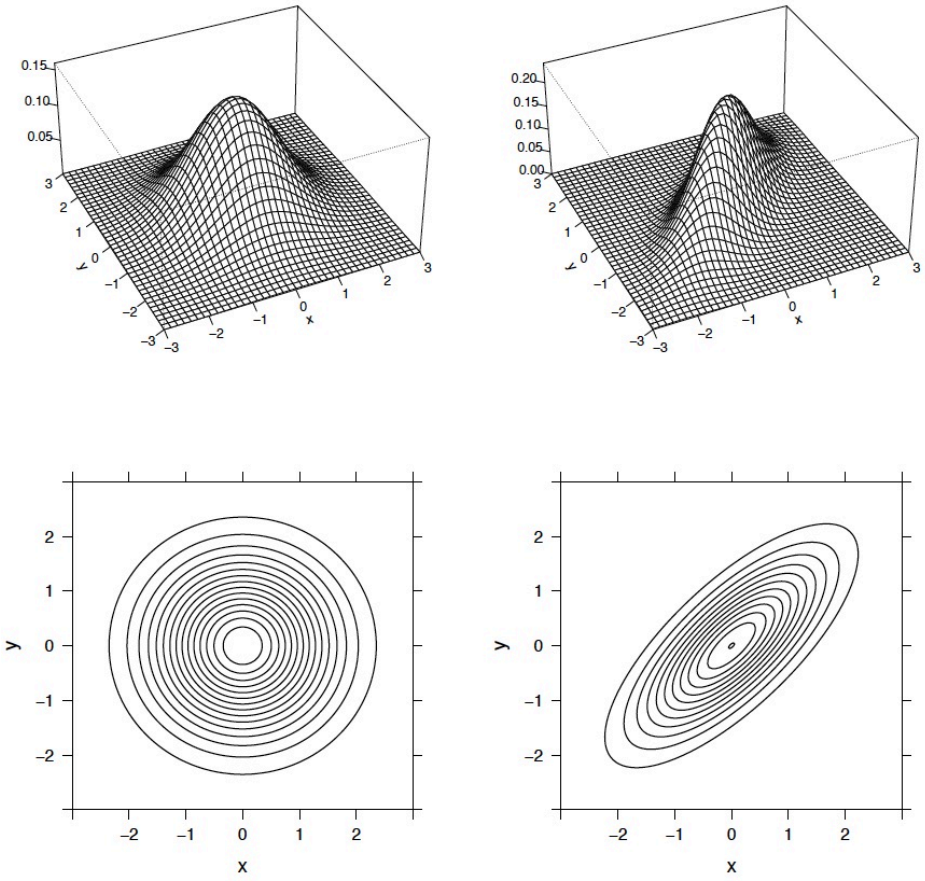

Figure 8: Joint PDFs of two Bivariate Normal distributions. In the two plots on the left, X and Y are marginally $\mathsf{N}(0, 1)$ and have zero correlation, hence the contours of the joint PDF are shaped like circles. In the two plots on the right, X and Y are marginally $\mathsf{N}(0, 1)$ and have correlation 0.75, hence the contour plots are ellipsoidal, reflecting the fact that X tends to be large when Y is large and vice versa.

Definition (Joint Moment Generating Function (MGF))**.** The *joint* MGF of a random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is the function which takes a vector of constants $\mathbf{t} = (t_1, \ldots, t_k)$ and returns

$$M(\mathbf{t}) = \mathsf{E}\left(e^{\mathbf{t'X}}\right) = \mathsf{E}\left(e^{t_1X_1+\ldots+t_kX_k}\right).$$

Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a Multivariate Normal random vector. Marginally, each component X_j of $\mathbf{X}$ follows a Normal distribution, hence the MGF of X_j is

$$\mathsf{E}\left(e^{t_jX_j}\right) = e^{t_j\mathsf{E}(X_j)+\frac{1}{2}t_j^2\mathsf{Var}(X_j)}.$$

But, by definition, $t_1X_1 + \ldots + t_kX_k$ is Normal, hence

$$\mathsf{E}\left(e^{t_1X_1+\ldots+t_kX_k}\right) = e^{t_1\mathsf{E}(X_1)+\ldots+t_k\mathsf{E}(X_k)+\frac{1}{2}\mathsf{Var}(t_1X_1+\ldots+t_kX_k)}.$$

which represents the joint MGF of the random vector $\mathbf{X} = (X_1, \ldots, X_k)$.

Theorem. *Within a Multivariate Normal random vector, uncorrelated implies independence. That is, if* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *is Multivariate Normal, where* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are subvectors, and every component of* $\mathbf{X}_1$ *is uncorrelated with every component of* $\mathbf{X}_2$*, then* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent. In particular, if* (X, Y) *is Bivariate Normal with* $\mathsf{Corr}(X, Y) = 0$*, then X and Y are independent.*

Proof. Let (X, Y) be Bivariate Normal with $\mathsf{E}(X) = \mu_1$, $\mathsf{E}(Y) = \mu_2$, $\mathsf{Var}(X) = \sigma_1^2$, $\mathsf{Var}(Y) = \sigma_2^2$, and $\mathsf{Corr}(X, Y) = \rho$. The joint MGF of (X, Y) is

$$\begin{aligned} M_{X,Y}(t_1, t_2) = \mathsf{E}\left(e^{t_1X+t_2Y}\right) &= \exp\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\mathsf{Var}(t_1X + t_2Y)\right), \\ &= exp\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2t_1t_2\sigma_1\sigma_2\rho)\right) \end{aligned}$$

When $\rho = 0$, we have

$$M_{X,Y}(t_1, t_2) = exp\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2)\right).$$

But this is also the joint MGF of the vector (Z, W) where $Z \sim \mathsf{N}(\mu_1, \sigma_1^2)$, $W \sim \mathsf{N}(\mu_2, \sigma_2^2)$, and Z is independent of W. Since the joint MGF determines the joint distribution, it must be that (X, Y) has the same distribution as (Z, W). Consequently, X and Y must be independent. □

Example: Independence of sum and difference
Let X, Y be i.i.d. $\mathsf{N}(0, 1)$. Find the joint distribution of $(X + Y, X - Y)$.

Solution: We have $X + Y \sim \mathsf{N}(0, 2)$ and $X - Y \sim \mathsf{N}(0, 2)$. Also,

$$\mathsf{Cov}(X + Y, X - Y) = \mathsf{Var}(X) - \mathsf{Cov}(X, Y) + \mathsf{Cov}(Y, X) - \mathsf{Var}(Y) = 0,$$

Since $(X + Y, X - Y)$ is Bivariate Normal, it follows that $X + Y$ and $X - Y$ are independent.