Jointly continuous distributions and the multivariate Normal

Márton Balázs[*] and Bálint Tóth[*]

October 13, 2014

This little write-up is part of important foundations of probability that were left out of the unit Probability 1 due to lack of time and prerequisites. Elementary linear algebra and multivariate calculus is required.

1 Basics of joint distributions

We give the fundamental tools to investigate the joint behaviour of several random variables. For most cases *several* will mean two, but everything can easily be generalised for any (possibly countably infinite) number of variables.

Let X and Y be random variables. The most important object to describe their joint behaviour is the following.

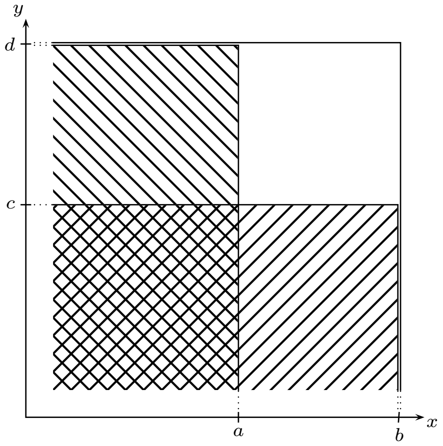
Definition 1 *The* joint distribution function *of the random variables X and Y is*

$$F(x,\,y):=\mathbf{P}\{X\le x,\ Y\le y\}.$$

This function has an answer to every meaningful question one wants to know about the distribution of X and Y. For example, let $a<b$ and $c<d$ be fixed real numbers. Then

$$\mathbf{P}\{a<X\le b,\ c<Y\le d\}=F(b,\,d)-F(a,\,d)-F(b,\,c)+F(a,\,c), \tag{1}$$

which can be seen in the below picture after realising that $F(x,\,y)$ is the probability of our random point $(X,\,Y)$ falling in the lower-left quarter with corner $(x,\,y)$. Looking for the probability of the upper-right (non shaded) rectangle, we can exactly express that using the above combination.



Here is a formal derivation of the same thing:

$$\begin{aligned}
&\{a<X\le b,\ c<Y\le d\}\\
&\qquad=\{X>a\}\cap\{X\le b\}\cap\{Y>c\}\cap\{Y\le d\}\\
&\qquad=\big(\{X>a\}\cap\{Y\le d\}\big)\cap\big(\{Y>c\}\cap\{X\le b\}\big)\\
&\qquad=\big((\{X>a\}\cap\{Y\le d\})\cup(\{Y>d\}\cap\{Y\le d\})\big)\cap\big((\{X>b\}\cap\{X\le b\})\cup(\{Y>c\}\cap\{X\le b\})\big)\\
&\qquad=\{X\le b\}\cap\{Y\le d\}\cap\big(\{X>a\}\cup\{Y>d\}\big)\cap\big(\{X>b\}\cup\{Y>c\}\big)\\
&\qquad=\{X\le b\}\cap\{Y\le d\}\cap\big(\{X\le a\}\cap\{Y\le d\}\big)^c\cap\big(\{X\le b\}\cap\{Y\le c\}\big)^c\\
&\qquad=\big(\{X\le b\}\cap\{Y\le d\}\big)\cap\big[\big(\{X\le a\}\cap\{Y\le d\}\big)\cup\big(\{X\le b\}\cap\{Y\le c\}\big)\big]^c\\
&\qquad=\big(\{X\le b\}\cap\{Y\le d\}\big)-\big[\big(\{X\le a\}\cap\{Y\le d\}\big)\cup\big(\{X\le b\}\cap\{Y\le c\}\big)\big].
\end{aligned}$$

[*] University of Bristol / Budapest University of Technology and Economics

As the subtracted set is subset of the previous one, inclusion exclusion gives

$$\begin{aligned}
&\mathbf{P}\{a < X \le b,\ c < Y \le d\} \\
&\qquad\qquad = \mathbf{P}\bigl\{\{X \le b\} \cap \{Y \le d\}\bigr\} - \mathbf{P}\bigl\{\bigl(\{X \le a\} \cap \{Y \le d\}\bigr) \cup \bigl(\{X \le b\} \cap \{Y \le c\}\bigr)\bigr\} \\
&\qquad\qquad = \mathbf{P}\bigl\{\{X \le b\} \cap \{Y \le d\}\bigr\} \\
&\qquad\qquad\quad - \bigl[\mathbf{P}\bigl\{\{X \le a\} \cap \{Y \le d\}\bigr\} + \mathbf{P}\bigl\{\{X \le b\} \cap \{Y \le c\}\bigr\} - \mathbf{P}\bigl\{\{X \le a\} \cap \{Y \le c\}\bigr\}\bigr] \\
&\qquad\qquad = F(b,\, d) - F(a,\, d) - F(b,\, c) + F(a,\, c).
\end{aligned}$$

Now, every element of the σ-algebra generated by rectangles, that is, every Borel measurable set has a probability that can be expressed in terms of F. Non-negativity of such probabilities give nontrivial conditions for a function F of two variables to make sure that this function is a distribution function. For more variables the situation is even more complicated.

One of the fundamental questions regarding joint distributions is the behaviour of one single variable:

Definition 2 *The* marginal distribution *of the random variable X is*

$$F_X(x) := \mathbf{P}\{X \le x\}.$$

Here we do not care at all about Y. A similar definition applies to the marginal of Y, when X is completely disregarded.

Proposition 3 *The marginal distributions can be expressed in terms of the joint distribution as*

$$F_X(x) = \lim_{y\to\infty} F(x,\, y), \qquad F_Y(y) = \lim_{x\to\infty} F(x,\, y).$$

It is very important to notice that marginal distributions do not contain enough information to restore the whole joint distribution.

Proof. The events $\{Y \le n\}$ are non-decreasing in n, and $\lim\limits_{n\to\infty} \{Y \le n\} = \bigcup\limits_{n>0} \{Y \le n\} = \Omega$. Therefore

$$\begin{aligned}
F_X(x) = \mathbf{P}\{X \le x\} = \mathbf{P}\bigl\{\{X \le x\} \cap \lim_{n\to\infty} \{Y \le n\}\bigr\} &= \mathbf{P}\bigl\{\lim_{n\to\infty} (\{X \le x\} \cap \{Y \le n\})\bigr\} \\
= \lim_{n\to\infty} \mathbf{P}\bigl\{\{X \le x\} \cap \{Y \le n\}\bigr\} &= \lim_{y\to\infty} F(x,\, y),
\end{aligned}$$

as the intersection of the two events is also non-decreasing in n, and finally we can make use of the monotonicity of F to change to the limit in the real values y from that of n's. □

The case of joint discrete random variables is covered in the Probability 1 slides. Here we concentrate on the *jointly continuous* case.

Definition 4 *The pair $(X,\, Y)$ has* jointly continuous distribution*, if there exists a* probability density function *f of two variables such that for all $C \subseteq \mathbb{R}^2$ measurable sets*

$$\mathbf{P}\{(X,\, Y) \in C\} = \iint\limits_C f(x,\, y)\, \mathrm{d}x\, \mathrm{d}y.$$

It is important to notice that marginally continuous random variables are *not* necessarily jointly continuous. The pair $X = Y \sim \mathrm{E}(0,\, 1)$ has a distribution that is concentrated on a line, and therefore cannot have a joint density function.

A few simple propositions follow easily from the definition:

Proposition 5 *For all $A,\, B \subseteq \mathbb{R}$ measurable sets,*

$$\mathbf{P}\{X \in A,\ Y \in B\} = \int\limits_A \int\limits_B f(x,\, y)\, \mathrm{d}y\, \mathrm{d}x.$$

In particular, the intuitive meaning of the joint density is demonstrated by

$$\mathbf{P}\{X \in (a,\, a+\varepsilon),\ Y \in (b,\, b+\delta)\} = \int\limits_a^{a+\varepsilon} \int\limits_b^{b+\delta} f(x,\, y)\, \mathrm{d}y\, \mathrm{d}x \simeq f(a,\, b) \cdot \varepsilon\delta.$$

Also,

Proposition 6

$$F(a,\,b)=\mathbf{P}\{X\le a,\ Y\le b\}=\int\limits_{-\infty}^{a}\int\limits_{-\infty}^{b}f(x,\,y)\,\mathrm{d}y\,\mathrm{d}x,\qquad \textit{from which}$$

$$f(a,\,b)=\frac{\partial^2F(a,\,b)}{\partial a\,\partial b}=\frac{\partial^2F(a,\,b)}{\partial b\,\partial a}.$$

Proposition 7 *The* marginal density *of the random variables X and Y, respectively, is*

$$f_X(x)=\int\limits_{-\infty}^{\infty}f(x,\,y)\,\mathrm{d}y\qquad \textit{and}\qquad f_Y(y)=\int\limits_{-\infty}^{\infty}f(x,\,y)\,\mathrm{d}x.$$

Proof. For all measurable $A\subseteq\mathbb{R}$,

$$\mathbf{P}\{X\in A\}=\mathbf{P}\{X\in A,\ Y\in\mathbb{R}\}=\int\limits_{A}\int\limits_{-\infty}^{\infty}f(x,\,y)\,\mathrm{d}y\,\mathrm{d}x,$$

which can be compared to the definition

$$\mathbf{P}\{X\in A\}=\int\limits_{A}f_X(x)\,\mathrm{d}x$$

of the marginal density to conclude the statement. □

Example 8 *Let the joint density of X and Y be given by*

$$f(x,\,y)=\begin{cases}\mathrm{e}^{-x/y}\cdot\mathrm{e}^{-y}/y, & \textit{if } x,\,y>0,\\ 0, & \textit{otherwise.}\end{cases}$$

Determine the marginal distribution of Y.

Let $y>0$, as otherwise the marginal density is surely zero. According to the above,

$$f_Y(y)=\int\limits_{-\infty}^{\infty}f(x,\,y)\,\mathrm{d}x=\int\limits_{0}^{\infty}\mathrm{e}^{-x/y}\cdot\mathrm{e}^{-y}/y\,\mathrm{d}x=\mathrm{e}^{-y}.$$

Therefore $Y\sim\mathrm{Exp}(1)$, its marginal distribution function is $F_Y(y)=1-\mathrm{e}^{-y}$ for positive y's, and zero otherwise. Observe that finding the marginal of X would be much more troublesome.

2 Transforming joint distributions

In several dimensions there are various cases of transforming variables. If we are after the distribution of a real-valued function $Z=g(X_1,\,X_2)$ or a pair $(X_1,\,X_2)$, then we can write

$$F_Z(z)=\mathbf{P}\{Z\le z\}=\mathbf{P}\{g(X_1,\,X_2)\le z\}=\iint\limits_{(x_1,\,x_2)\,:\,g(x_1,\,x_2)\le z}f(x_1,\,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2. \tag{2}$$

If, after integration, F_Z is differentiable then the density f_Z of Z can be obtained this way. An example follows in Section 6.

If we deal with an $\mathbb{R}^2$-valued function $(Y_1,\,Y_2)=\big(g_1(X_1,\,X_2),\,g_2(X_1,\,X_2)\big)$ of the pair $(X_1,\,X_2)$, we can still proceed along the definitions. For simplicity, we use a vector notation:

$$\begin{aligned}F_{\underline{Y}}(\underline{y})&=\mathbf{P}\{Y_1\le y_1,\ Y_2\le y_2\}=\mathbf{P}\{g_1(X_1,\,X_2)\le y_1,\ g_2(X_1,\,X_2)\le y_2\}\\&=\iint\limits_{(x_1,\,x_2)\,:\,g_1(x_1,\,x_2)\le y_1,\ g_2(x_1,\,x_2)\le y_2}f_{\underline{X}}(x_1,\,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2.\end{aligned}\tag{3}$$

If the mixed partial derivative exists, it gives the joint density of $(Y_1,\,Y_2)$.

Assuming $\underline{g}$ is nice enough, one can summarise the above as:

Proposition 9 (two dimensional transformations) *Let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be such that*

- *it is one-to-one,*
- *it has continuous partial derivatives,*
- *its Jacobean* $|\underline{\underline{J}}| = \begin{vmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 \end{vmatrix} \neq 0$ *on* $\mathbb{R}^2$.

Then $\underline{g}$ is invertible, and the random variable $\underline{Y} := \underline{g}(\underline{X})$ is jointly continuous with density

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}\big(\underline{g}^{-1}(\underline{y})\big) \cdot |\underline{\underline{J}}|^{-1}.$$

The last term is the reciprocal of the Jacobi determinant (taken at $\underline{x} = \underline{g}^{-1}(\underline{y})$).

Compare this to the one dimensional statement (from Probability 1).

Proof. Continue from (3) and perform the two dimensional change $\underline{a} := \underline{g}(\underline{x})$ of variables:

$$F_{\underline{Y}}(\underline{y}) = \iint\limits_{a_1<y_1,\ a_2<y_2} f_{\underline{X}}\big(g^{-1}(\underline{a})\big) \cdot |\underline{\underline{J}}|^{-1}\,\mathrm{d}a_1\,\mathrm{d}a_2 = \int\limits_{-\infty}^{y_1}\int\limits_{-\infty}^{y_2} f_{\underline{X}}\big(g^{-1}(\underline{a})\big) \cdot |\underline{\underline{J}}|^{-1}\,\mathrm{d}a_1\,\mathrm{d}a_2,$$

from which

$$f_{\underline{Y}}(\underline{y}) = \frac{\partial^2 F_{\underline{Y}}(y_1,\, y_2)}{\partial y_1\,\partial y_2} = f_{\underline{X}}\big(g^{-1}(\underline{y})\big) \cdot |\underline{\underline{J}}|^{-1}.$$

□

A nice and important example will be given in Proposition 16 below.

3 Independent random variables

The discrete case of independent random variables has been covered in Probability 1. Here we concentrate on continuous random variables.

Definition 10 *The random variables $X_1,\, X_2,\, \dots,\, X_n$ are* independent*, if marginal events formulated with them are so, that is, for any measurable sets $A_1,\, A_2,\, \dots,\, A_n$ in $\mathbb{R}$,*

$$\mathbf{P}\{X_1 \in A_1,\ X_2 \in A_2,\ \dots\ X_n \in A_n\} = \mathbf{P}\{X_1 \in A_1\} \cdot \mathbf{P}\{X_2 \in A_2\} \cdots \mathbf{P}\{X_n \in A_n\}.$$

Countably infinitely many variables are independent*, if all finite families of them are so in the above sense. The abbreviation* i.i.d. *stands for* ***i****ndependent* ***i****dentically* ***d****istributed.*

We shall mostly consider two variables, but everything generalises to any countable number of variables.

Proposition 11 *Independence of X and Y*

1. *is equivalent to the factorisation of the distribution functions:*

$$F(x,\, y) = F_X(x) \cdot F_Y(y) \qquad (x,\, y \in \mathbb{R});$$

2. *in the jointly continuous case is equivalent to the density function factorising:*

$$f(x,\, y) = f_X(x) \cdot f_Y(y) \qquad (x,\, y \in \mathbb{R});$$

3. *in the jointly continuous case is equivalent to the product form of the joint density:*

$$f(x,\, y) = h(x) \cdot g(y) \qquad (x,\, y \in \mathbb{R}).$$

Similar statements are valid in the discrete case with the probability mass function, those were covered in Probability 1.

Proof.

1. That independence implies 1 is trivial. To prove the reverse, let $a < b$ and $c < d$. Then, see (1),

$$\begin{aligned}\mathbf{P}\{a < X \le b,\ c < Y \le d\} &= F(b,\,d) - F(a,\,d) - F(b,\,c) + F(a,\,c)\\ &= F_X(b)\cdot F_Y(d) - F_X(a)\cdot F_Y(d) - F_X(b)\cdot F_Y(c) + F_X(a)\cdot F_Y(c)\\ &= \big(F_X(b) - F_X(a)\big)\cdot\big(F_Y(d) - F_Y(c)\big)\\ &= \mathbf{P}\{a < X \le b\}\cdot\mathbf{P}\{c < Y \le d\}.\end{aligned}$$

   This is the defining property of independence for intervals, the extension to all Borel-measurable sets is done in the usual measure-theoretic way.

2. If X and Y are jointly continuous and independent, then

$$f(x,\,y) = \frac{\partial^2}{\partial x\,\partial y}F(x,\,y) = \frac{\partial^2}{\partial x\,\partial y}F_X(x)\cdot F_Y(y) = \frac{\partial}{\partial x}F_X(x)\cdot\frac{\partial}{\partial y}F_Y(y) = f_X(x)\cdot f_Y(y).$$

   To see the reverse, if $f(x,\,y) = f_X(x)\cdot f_Y(y)$, then for any A and B measurable sets,

$$\begin{aligned}\mathbf{P}\{X\in A,\ Y\in B\} &= \int\limits_A\int\limits_B f(x,\,y)\,\mathrm{d}y\,\mathrm{d}x = \int\limits_A\int\limits_B f_X(x)\cdot f_Y(y)\,\mathrm{d}y\,\mathrm{d}x\\ &= \Big(\int\limits_A f_X(x)\,\mathrm{d}x\Big)\cdot\Big(\int\limits_B f_Y(y)\,\mathrm{d}y\Big) = \mathbf{P}\{X\in A\}\cdot\mathbf{P}\{Y\in B\}.\end{aligned}$$

3. This depends on the right grouping of multiplicative constants, exactly the same way as done in Probability 1 for the discrete case.

□

It now follows that knowing the marginal distributions and the fact that the random variables are independent uniquely determines the distribution.

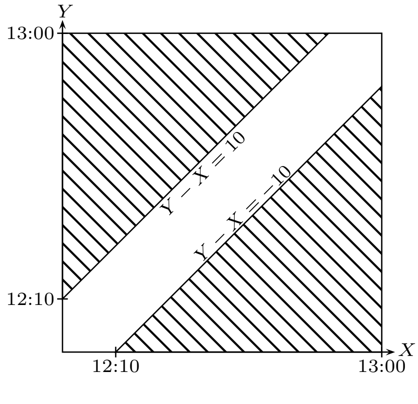
Let $X \sim \mathrm{U}(a,\,b)$ and $Y \sim \mathrm{U}(c,\,d)$ be independent and uniformly distributed. Then their marginal densities are constant ($= 1/(b-a)$ and $1/(d-c)$) on the respective intervals, therefore the joint density is also constant on the Cartesian product of the intervals. In other words, the density is constant on the rectangle $(a,\,b)\times(c,\,d)$, and the value of the constant is the reciprocal of the area of the rectangle. This motivates the following

Definition 12 *Fix a measurable subset C of the plane with positive and finite Lebesgue measure $|C|$. The distribution with constant density on C and zero outside C is called* uniform on C. *Then the value of the constant density is necessarily* $1/|C|$. *The probability of falling in a measurable $A \subseteq C$ is $|A|/|C|$, the proportion of areas.*

The first thing to do in a problem involving two dimensional uniforms (or independent one dimensional uniforms) is drawing a picture. The problem is then often solved by comparing areas.

Example 13 *Juliet and John arrive randomly and independently to a rendez-vous between 12:00 and 13:00. What is the probability that the first to arrive waits more than 10 minutes for the second one?*

We can assume that the respective arrival times X and Y are i.i.d. U(12:00, 13:00) random variables. Therefore the joint distribution is uniform on the square (12:00, 13:00) × (12:00, 13:00). Plotting the event $\{|X-Y| > 10\}$ in question:



The probability of the event is the fraction of areas of the event and of the full square: $2 \cdot \frac{50^2}{2}/60^2 = 25/36$. Of course the problem can also be solved by two dimensional integration (this is basically a calculation of the areas of the two triangles).

Example 14 *Let the joint distribution function of X and Y be given by*

$$f(x,\, y) = \begin{cases} 24xy, & \textit{if } \; 0 < x < 1, \quad 0 < y < 1, \quad 0 < x + y < 1, \\ 0, & \textit{otherwise.} \end{cases}$$

Are X and Y independent?

At first sight one might think that this function is of product form. However, including the conditions given:

$$f(x,\, y) = 24xy \cdot \mathbf{1}\{0 < x < 1\} \cdot \mathbf{1}\{0 < y < 1\} \cdot \mathbf{1}\{0 < x + y < 1\}$$

clearly shows that this is not the case due to the last indicator. Any pair of variables the distribution of which is not concentrated on a product set cannot be independent, and this is the case here.

4 Multivariate Normal

The following definition is given in general, d dimensions.

Definition 15 *Let $\underline{\underline{A}}$ be a $d \times d$ positive definite symmetric real matrix (that is, for any $\underline{x} \in \mathbb{R}^d$, $\underline{x}^T \underline{\underline{A}}\, \underline{x} > 0$), and $\underline{m} \in \mathbb{R}$ a fixed vector. The vector $\underline{X} \in \mathbb{R}^d$ of random variables is said to have the* Multivariate Normal (or Multivariate Gauss) distribution*, if its density is given by*

$$f(\underline{x}) = \frac{\sqrt{\det \underline{\underline{A}}}}{\sqrt{2\pi}^{\,d}} \cdot \mathrm{e}^{-\frac{1}{2}(\underline{x}-\underline{m})^T \underline{\underline{A}}(\underline{x}-\underline{m})} \qquad (\underline{x} \in \mathbb{R}^d). \tag{4}$$

The matrix $\underline{\underline{A}}$ will be given a very natural meaning below in Proposition 20.

Proposition 16 *The d dimensional multivariate normal is obtained as the affine transform of d i.i.d. standard normal variables.*

Proof. With the above notations the matrix $\underline{\underline{A}}$ is symmetric, therefore it has real eigenvalues, and we can pick d orthonormal eigenvectors $\underline{v}^{(1)}$, $\underline{v}^{(2)}$, $\dots$, $\underline{v}^{(d)}$. The matrix $\underline{\underline{P}}$ formed with these vectors as columns diagonalises $\underline{\underline{A}}$-t:

$$\underline{\underline{P}}^{-1}\, \underline{\underline{A}}\, \underline{\underline{P}} = \underline{\underline{D}} = \begin{pmatrix} \lambda^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & \lambda^{(2)} & 0 & \cdots & 0 \\ 0 & 0 & \lambda^{(3)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda^{(d)} \end{pmatrix}.$$

As $\underline{\underline{A}}$ is positive definite, the eigenvalues are positive, and we can consider the diagonal matrix $\sqrt{\underline{\underline{D}}}$ formed by the square-roots of the eigenvalues. Define the affine transformation

$$\underline{g} \,:\, \mathbb{R}^d \to \mathbb{R}^d, \qquad \underline{x} \mapsto \underline{g}(\underline{x}) = \sqrt{\underline{\underline{D}}}\, \underline{\underline{P}}^{-1}(\underline{x} - \underline{m})$$

(shift by $\underline{m}$, rotation according to the basis $\{\underline{v}^{(i)}\}_i$, then stretching by various factors in various directions). Due to

$$\underline{\underline{A}} = \underline{\underline{P}}\, \underline{\underline{D}}\, \underline{\underline{P}}^{-1} = \big(\underline{\underline{P}}^{-1}\big)^T \sqrt{\underline{\underline{D}}} \sqrt{\underline{\underline{D}}}\, \underline{\underline{P}}^{-1} = \big(\sqrt{\underline{\underline{D}}}\, \underline{\underline{P}}^{-1}\big)^T \sqrt{\underline{\underline{D}}}\, \underline{\underline{P}}^{-1}$$

we can write (4) as

$$f_{\underline{X}}(\underline{x}) = \frac{\sqrt{\det \underline{\underline{D}}}}{\sqrt{2\pi}^{\,d}} \cdot \mathrm{e}^{-\frac{1}{2}(\underline{g}(\underline{x}))^T \cdot \underline{g}(\underline{x})}.$$

The function $\underline{g}$ satisfies the conditions of Proposition 9, its Jacobi determinant is $\det \sqrt{\underline{\underline{D}}} \cdot \det \underline{\underline{P}}^{(-1)} = \pm \det \sqrt{\underline{\underline{D}}} = \pm\sqrt{\det \underline{\underline{D}}}$, hence the density of the random variable $\underline{Y} := \underline{g}(\underline{X})$ is

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}\big(\underline{g}^{-1}(\underline{y})\big) \cdot |\underline{\underline{J}}|^{-1} &= \frac{\sqrt{\det \underline{\underline{D}}}}{\sqrt{2\pi}^{\,d}} \cdot \mathrm{e}^{-\frac{1}{2}(\underline{g}(\underline{g}^{-1}(\underline{y})))^T \cdot \underline{g}(\underline{g}^{-1}(\underline{y}))} \cdot \Big(\sqrt{\det \underline{\underline{D}}}\Big)^{-1} \\ &= \frac{1}{\sqrt{2\pi}^{\,d}} \cdot \mathrm{e}^{-\frac{1}{2}\underline{y}^T \cdot \underline{y}} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-y_i^2/2}, \end{aligned}$$

which exactly means that the variables $\underline{Y} = (Y_1, Y_2, \dots, Y_d)$ are i.i.d. $\mathcal{N}(0, 1)$ distributed. As

$$\underline{X} = \underline{g}^{-1}(\underline{Y}) = \underline{\underline{P}}\,\underline{\underline{D}}^{-1/2}\underline{Y} + \underline{m},$$

the proof is complete. □

The reversed statement is also true:

Proposition 17 *Any affine transform of d i.i.d. Standard Normal random variables with nonzero Jacobi determinant results in a d dimensional multivariate normal distribution.*

Proof. Substituting the i.i.d. Normals into the transformation proposition 9 we easily recognise the joint density with its $\underline{\underline{A}}$ positive definite symmetric matrix and $\underline{m}$ vector. □

Proposition 18 *The joint distribution of i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables is rotationally invariant.*

Proof. We have $\underline{m} = \underline{0}$ and $\underline{\underline{A}}$ is the $1/\sigma^2$ multiple of the identity matrix, therefore with any orthogonal transformation

$$\underline{g} : \mathbb{R}^d \to \mathbb{R}^d, \qquad \underline{x} \mapsto \underline{g}(\underline{x}) = \underline{\underline{P}}^{-1}\underline{x}$$

it follows that

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}\big(\underline{g}^{-1}(\underline{y})\big) \cdot |\underline{\underline{J}}|^{-1} = \frac{1}{\sqrt{2\pi}^d\,\sigma^d} \cdot \mathrm{e}^{-\frac{1}{2\sigma^2}(\underline{\underline{P}}\,\underline{y})^T \cdot (\underline{\underline{P}}\,\underline{y})} \cdot 1^{-1} = \frac{1}{\sqrt{2\pi}^d\,\sigma^d} \cdot \mathrm{e}^{-\frac{1}{2\sigma^2}\underline{y}^T \cdot \underline{y}}$$

which is still the product of $\mathcal{N}(0, \sigma^2)$ density functions. □

5 The covariance matrix

Definition 19 *Let $X_1, X_2, \dots, X_n$ be random variables. Their* covariance matrix *$\underline{\underline{C}}$ is the one formed of*

$$C_{ij} := \mathbf{Cov}(X_i, X_j)$$

as entries ($1 \le i, j \le n$).

Due to the properties of the covariance, this matrix is symmetric, its diagonal consists of the variances. Let $\underline{a}$ be a fixed vector, then

$$\underline{a}^T\underline{\underline{C}}\,\underline{a} = \sum_{i,j} a_i \mathbf{Cov}(X_i, X_j) a_j = \mathbf{Cov}\Big(\sum_i a_i X_i, \sum_j a_j X_j\Big) = \mathbf{Var}\Big(\sum_i a_i X_i\Big) \ge 0,$$

which shows that *the covariance matrix is positive semidefinite.*

Proposition 20 *The covariance matrix of the Multivariate Normal distribution in Definition 15 is $\underline{\underline{A}}^{-1}$.*

Proof. We have seen that $\underline{X} = \underline{\underline{P}}\,\underline{\underline{D}}^{-1/2}\underline{Y} + \underline{m}$, where $\underline{\underline{P}}$ is an orthogonal matrix, $\underline{\underline{D}} = \underline{\underline{P}}^{-1}\underline{\underline{A}}\,\underline{\underline{P}}$, and $\underline{Y}$ are i.i.d. standard normal variables. As the covariance matrix is translation-invariant, we can immediately forget about $\underline{m}$. The i, j entry of the matrix is

$$\begin{aligned}
C_{ij} = \mathbf{Cov}(X_i, X_j) &= \mathbf{Cov}\Big(\sum_{k,\ell} P_{ik}(\underline{\underline{D}}^{-1/2})_{k\ell} Y_\ell, \sum_{m,h} P_{jm}(\underline{\underline{D}}^{-1/2})_{mh} Y_h\Big) \\
&= \sum_{k,\ell,m,h} P_{ik}(\underline{\underline{D}}^{-1/2})_{k\ell} P_{jm}(\underline{\underline{D}}^{-1/2})_{mh} \mathbf{Cov}(Y_\ell, Y_h) \\
&= \sum_{k,\ell,m,h} P_{ik}(\underline{\underline{D}}^{-1/2})_{k\ell} P_{jm}(\underline{\underline{D}}^{-1/2})_{mh} \delta_{\ell h} \\
&= \sum_{k,\ell,m} P_{ik}(\underline{\underline{D}}^{-1/2})_{k\ell} P_{jm}(\underline{\underline{D}}^{-1/2})_{m\ell} \\
&= \sum_{k,\ell,m} P_{ik}(\underline{\underline{D}}^{-1/2})_{k\ell}(\underline{\underline{D}}^{-1/2})_{\ell m}(\underline{\underline{P}}^{-1})_{mj} \\
&= (\underline{\underline{P}}\,\underline{\underline{D}}^{-1}\underline{\underline{P}}^{-1})_{ij} = \big((\underline{\underline{P}}\,\underline{\underline{D}}\,\underline{\underline{P}}^{-1})^{-1}\big)_{ij} = \big(\underline{\underline{A}}^{-1}\big)_{ij},
\end{aligned}$$

□

where $\delta_{h\ell} = \mathbf{1}\{h = \ell\}$ is Kronecker's delta.

6 Continuous convolutions

Discrete integer-valued convolutions were covered in Probability 1. Here we derive the continuous convolution formula, and show a few applications. The convolution of Exponential and Gamma distributions were also covered in Probability 1.

Let X and Y be independent continuous random variables. We start with the distribution function of the sum with the help of (2):

$$F_{X+Y}(a)=\mathbf{P}\{X+Y\le a\}=\iint\limits_{x+y\le a} f(x,\,y)\,\mathrm{d}x\,\mathrm{d}y=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{a-y} f_X(x)\cdot f_Y(y)\,\mathrm{d}x\,\mathrm{d}y=\int\limits_{-\infty}^{\infty}F_X(a-y)\cdot f_Y(y)\,\mathrm{d}y.$$

This formula is called the *convolution of distribution functions.* Differentiating it gives

Proposition 21 *If X and Y are continuous independent random variables, then*

$$f_{X+Y}(a)=\int\limits_{-\infty}^{\infty}f_X(a-y)\cdot f_Y(y)\,\mathrm{d}y,$$

which formula is known as the convolution of density functions*.*

Example 22 *Determine the density of the sum of two i.i.d. $U(0,\,1)$ random variables.*

With independent uniform variables it is always advisable to work with pictures and distribution functions. Instead, here we apply the convolution formula. The marginal densities are $f_X(z)=f_Y(z)=\mathbf{1}\{0<z<1\}$. Therefore with $a\in(0,\,2)$ (the density is zero elsewhere)

$$f_{X+Y}(a)=\int\limits_{-\infty}^{\infty}\mathbf{1}\{0<a-y<1\}\cdot\mathbf{1}\{0<y<1\}\,\mathrm{d}y=\int\limits_{(a-1)\vee 0}^{a\wedge 1}1\,\mathrm{d}y=\begin{cases}a, & \text{if } 0<a<1,\\ 2-a, & \text{if } 1<a<2.\end{cases}$$

The sum of independent uniforms is *not* uniform. However, the sum of independent Normals is Normal:

Proposition 23 *Let $X\sim\mathcal{N}(\mu_x,\,\sigma_x^2)$ and $Y\sim\mathcal{N}(\mu_y,\,\sigma_y^2)$ be independent. Then $X+Y\sim\mathcal{N}(\mu_x+\mu_y,\,\sigma_x^2+\sigma_y^2)$.*

Proof. Consider the case $\mu_x=\mu_y=0$ first. The density of the sum is

$$f_{X+Y}(a)=\int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma_x}\mathrm{e}^{-\frac{(a-y)^2}{2\sigma_x^2}}\cdot\frac{1}{\sqrt{2\pi}\,\sigma_y}\mathrm{e}^{-\frac{y^2}{2\sigma_y^2}}\,\mathrm{d}y=\frac{1}{2\pi\sigma_x\sigma_y}\int\limits_{-\infty}^{\infty}\mathrm{e}^{-\frac12\left[\frac{(a-y)^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right]}\,\mathrm{d}y.$$

The strategy is to complete the square in the bracket so that the integral can be computed.

$$\begin{aligned}\frac{(a-y)^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}&=\frac{y^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2\,\frac{ay}{\sigma_x^2}+\frac{a^2}{\sigma_x^2}=\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot y^2-2\,\frac{a}{\sigma_x^2}\cdot y+\frac{a^2}{\sigma_x^2}\\&=\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot\Big(y^2-2\,a\,\frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\cdot y\Big)+\frac{a^2}{\sigma_x^2}\\&=\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot\Big(y-a\,\frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\Big)^2-\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot a^2\,\frac{\sigma_y^4}{(\sigma_x^2+\sigma_y^2)^2}+\frac{a^2}{\sigma_x^2}\\&=\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot\Big(y-a\,\frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\Big)^2+a^2\cdot\frac{-\sigma_y^2+\sigma_x^2+\sigma_y^2}{(\sigma_x^2+\sigma_y^2)\sigma_x^2}\\&=\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\cdot\Big(y-a\,\frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\Big)^2+\frac{a^2}{\sigma_x^2+\sigma_y^2}.\end{aligned}$$

Introducing the integration variable $z=\frac{\sqrt{\sigma_x^2+\sigma_y^2}}{\sigma_x\sigma_y}\cdot\Big(y-a\,\frac{\sigma_y^2}{\sigma_x^2+\sigma_y^2}\Big)$,

$$f_{X+Y}(a)=\frac{1}{2\pi\sqrt{\sigma_x^2+\sigma_y^2}}\int\limits_{-\infty}^{\infty}\mathrm{e}^{-z^2/2}\,\mathrm{d}z\cdot\mathrm{e}^{-\frac12\frac{a^2}{\sigma_x^2+\sigma_y^2}}=\frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}}\cdot\mathrm{e}^{-\frac12\frac{a^2}{\sigma_x^2+\sigma_y^2}}.$$

This shows that $X+Y\sim\mathcal{N}(0,\,\sigma_x^2+\sigma_y^2)$. For the general case, we write

$$X+Y=\underbrace{\underbrace{X-\mu_x}_{\sim\mathcal{N}(0,\,\sigma_x^2)}+\underbrace{Y-\mu_y}_{\sim\mathcal{N}(0,\,\sigma_y^2)}}_{\sim\mathcal{N}(0,\,\sigma_x^2+\sigma_y^2)}+\mu_x+\mu_y\sim\mathcal{N}(\mu_x+\mu_y,\,\sigma_x^2+\sigma_y^2).$$

□

Of course the statement implies $X-Y\sim\mathcal{N}(\mu_x-\mu_y,\,\sigma_x^2+\sigma_y^2)$ as well (notice that the variances still add up!).

Example 24 *A basketball team plays $n_A=26$ games against class A teams, and $n_B=18$ games against class B teams. The probability of winning against class A teams is, independently, $p_A=0.4$, while it is $p_B=0.7$ against class B teams. What is the probability that our teams wins 25 or more games? What is the probability that it wins more against class A teams than class B teams?*

The number of games won against class A teams is $X\sim\text{Binom}(n_A,\,p_A)$, against class B teams is $Y\sim\text{Binom}(n_B,\,p_B)$, and these variables are independent. The problem is that due to $p_A\neq p_B$, the sum of these variables is not Binomial. However, the number of games is enough for the DeMoivre-Laplace Theorem to give a usable approximation. That is,

$$\frac{X-n_Ap_A}{\sqrt{n_Ap_A(1-p_A)}}\sim\mathcal{N}(0,\,1)\qquad\text{and}\qquad\frac{Y-n_Bp_B}{\sqrt{n_Bp_B(1-p_B)}}\sim\mathcal{N}(0,\,1).$$

We therefore norm our centered variable with a yet unknown constant C:

$$\begin{aligned}&\mathbf{P}\{X+Y\geq 25\}\\&\quad=\mathbf{P}\{X+Y\geq 24.5\}\\&\quad=\mathbf{P}\Big\{\frac{X-n_Ap_A}{C}+\frac{Y-n_Bp_B}{C}\geq\frac{24.5-n_Ap_A-n_Bp_B}{C}\Big\}\\&\quad=\mathbf{P}\Big\{\underbrace{\underbrace{\underbrace{\frac{X-n_Ap_A}{\sqrt{n_Ap_A(1-p_A)}}}_{\sim\mathcal{N}(0,\,1)}\cdot\frac{\sqrt{n_Ap_A(1-p_A)}}{C}}_{\sim\mathcal{N}(0,\,n_Ap_A(1-p_A)/C^2)}+\underbrace{\underbrace{\frac{Y-n_Bp_B}{\sqrt{n_Bp_B(1-p_B)}}}_{\sim\mathcal{N}(0,\,1)}\cdot\frac{\sqrt{n_Bp_B(1-p_B)}}{C}}_{\sim\mathcal{N}(0,\,n_Bp_B(1-p_B)/C^2)}}_{\sim\mathcal{N}(0,\,[n_Ap_A(1-p_A)+n_Bp_B(1-p_B)]/C^2)}\geq\frac{24.5-n_Ap_A-n_Bp_B}{C}\Big\}.\end{aligned}$$

From here we see that

$$C=\sqrt{n_Ap_A(1-p_A)+n_Bp_B(1-p_B)}$$

is a reasonable choice as then the left hand-side will be close to a standard normal. The DeMoivre-Laplace approximation remains valid if the factors $\sqrt{n_Ap_A(1-p_A)}/C$ and $\sqrt{n_Bp_B(1-p_B)}/C$ are not too large compared to $\sqrt{n_Ap_A(1-p_A)}$ and $\sqrt{n_Bp_B(1-p_B)}$, that is, $\sqrt{n_Ap_A(1-p_A)}$ and $\sqrt{n_Bp_B(1-p_B)}$ do not have different orders of magnitude. In our case their values are approximately 2.50 and 1.94. (Of course these are intuitive arguments, precise statements would require estimates of error terms.) With this choice of C,

$$\begin{aligned}\mathbf{P}\{X+Y\geq 25\}&\simeq 1-\Phi\Big(\frac{24.5-n_Ap_A-n_Bp_B}{C}\Big)\\&=1-\Phi\Big(\frac{24.5-n_Ap_A-n_Bp_B}{\sqrt{n_Ap_A(1-p_A)+n_Bp_B(1-p_B)}}\Big)\simeq 1-\Phi(0.474)\simeq 0.32.\end{aligned}$$

Similarly, the probability of winning more games against class A teams than against class B teams is

$$\begin{aligned}&\mathbf{P}\{X-Y>0\}\\&\quad=\mathbf{P}\{X-Y\geq 0.5\}\\&\quad=\mathbf{P}\Big\{\frac{X-n_Ap_A}{C}-\frac{Y-n_Bp_B}{C}\geq\frac{0.5-n_Ap_A+n_Bp_B}{C}\Big\}\\&\quad=\mathbf{P}\Big\{\underbrace{\underbrace{\underbrace{\frac{X-n_Ap_A}{\sqrt{n_Ap_A(1-p_A)}}}_{\sim\mathcal{N}(0,\,1)}\cdot\frac{\sqrt{n_Ap_A(1-p_A)}}{C}}_{\sim\mathcal{N}(0,\,n_Ap_A(1-p_A)/C^2)}-\underbrace{\underbrace{\frac{Y-n_Bp_B}{\sqrt{n_Bp_B(1-p_B)}}}_{\sim\mathcal{N}(0,\,1)}\cdot\frac{\sqrt{n_Bp_B(1-p_B)}}{C}}_{\sim\mathcal{N}(0,\,n_Bp_B(1-p_B)/C^2)}}_{\sim\mathcal{N}(0,\,[n_Ap_A(1-p_A)+n_Bp_B(1-p_B)]/C^2)}\geq\frac{0.5-n_Ap_A+n_Bp_B}{C}\Big\},\end{aligned}$$

hence we can use the same norming constant C as before, and

$$\mathbf{P}\{X - Y > 0\} \simeq 1 - \Phi\Big(\frac{0.5 - n_A p_A + n_B p_B}{\sqrt{n_A p_A(1-p_A) + n_B p_B(1-p_B)}}\Big) \simeq 1 - \Phi(0.853) \simeq 0.20.$$