

Random Vectors

Paolo Zacchia

Probability and Statistics

Lecture 3

Random vectors

It is important to analyze how different random variables *relate to one another*. The starting point is the following concept.

Definition 1

Random Vector. A random vector \mathbf{x} of length K is a collection of K random variables X_1, \dots, X_K :

$$\mathbf{x} = \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}$$

each with support $\mathbb{X}_k \subseteq \mathbb{R}$ for $k = 1, \dots, K$.

The *realizations* of random vectors are denoted here with roman, bold lower-case letters, e.g. \mathbf{x} .

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}$$

Joint distributions

Certain concepts about random variables extend quite naturally to the multivariate case.

Definition 2

Support of a Random Vector. The support $\mathbb{X} \subseteq \mathbb{R}^K$ of a random vector \mathbf{x} is the Cartesian product of all the supports of the random variables featured in \mathbf{x} .

$$\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_K$$

Definition 3

Joint Probability Cumulative Distribution. Given some random vector \mathbf{x} , its joint probability *cumulative* distribution is defined as the following function.

$$F_{\mathbf{x}}(\mathbf{x}) = \mathbb{P}(\mathbf{x} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1 \cap \cdots \cap X_K \leq x_K)$$

Joint discrete distributions

Definition 4

Joint Probability Mass Function. Given some random vector \mathbf{x} composed by *discrete* random variables *only*, its joint probability *mass* function $f_{\mathbf{x}}(\mathbf{x})$ is defined as follows, for all $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{R}^K$.

$$f_{\mathbf{x}}(x_1, \dots, x_K) = \mathbb{P}(X_1 = x_1 \cap \dots \cap X_K = x_K)$$

- A joint p.m.f. is related to the joint c.d.f. via the following relationship:

$$\mathbb{P}(\mathbf{x} \leq \mathbf{x}) = F_{\mathbf{x}}(\mathbf{x}) = \sum_{\mathbf{t} \in \mathbb{X}: \mathbf{t} \leq \mathbf{x}} f_{\mathbf{x}}(\mathbf{t})$$

- ... and the total probability mass is obviously 1.

$$\mathbb{P}(\mathbf{x} \in \mathbb{X}) = \sum_{\mathbf{t} \in \mathbb{X}} f_{\mathbf{x}}(\mathbf{t}) = 1$$

Joint continuous distributions (1/2)

Definition 5

Joint Probability Density Function. Given some random vector \mathbf{x} composed by *continuous* random variables *only*, its joint probability density function $f_{\mathbf{x}}(\mathbf{x})$ is defined as the function that satisfies the following relationship, for all $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{R}^K$.

$$F_{\mathbf{x}}(x_1, \dots, x_K) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_K} f_{\mathbf{x}}(t_1, \dots, t_K) dt_1 \dots dt_K$$

- Given two vectors $\mathbf{a} = (a_1, \dots, a_K)$ and $\mathbf{b} = (b_1, \dots, b_K)$ with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ and $b_k \geq a_k$ for $k = 1, \dots, K$, it is:

$$\begin{aligned} \mathbb{P}(a_1 \leq x_1 \leq b_1 \cap \cdots \cap a_K \leq x_K \leq b_K) &= \\ &= F_{\mathbf{x}}(b_1, \dots, b_K) - F_{\mathbf{x}}(a_1, \dots, a_K) = \\ &= \int_{a_1}^{b_1} \cdots \int_{a_K}^{b_K} f_{\mathbf{x}}(t_1, \dots, t_K) dt_1 \dots dt_K \end{aligned}$$

(Continues...)

Joint continuous distributions (2/2)

Definition 6

Joint Probability Density Function. Given some random vector \mathbf{x} composed by *continuous* random variables *only*, its joint probability density function $f_{\mathbf{x}}(\mathbf{x})$ is defined as the function that satisfies the following relationship, for all $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{R}^K$.

$$F_{\mathbf{x}}(x_1, \dots, x_K) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_K} f_{\mathbf{x}}(t_1, \dots, t_K) dt_1 \dots dt_K$$

- **(Continued.)** Clearly, the joint density integrates to 1 over the entire support of \mathbf{x} .

$$\mathbb{P}(\mathbf{x} \in \mathbb{X}) = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_K} f_{\mathbf{x}}(t_1, \dots, t_K) dt_1 \dots dt_K = 1$$

Marginal discrete distributions

Definition 7

Marginal Distribution (discrete). For some given random vector \mathbf{x} made of *discrete* random variables *only*, the probability mass function of X_k – the k -th random variable in \mathbf{x} – is obtained as:

$$f_{X_k}(x_k) = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_{k-1} \in \mathbb{X}_{k-1}} \sum_{x_{k+1} \in \mathbb{X}_{k+1}} \cdots \sum_{x_K \in \mathbb{X}_K} f_{\mathbf{x}}(x_1, \dots, x_K)$$

and thus $F_{X_k}(x_k) = \sum_{t=\inf \mathbb{X}_k}^{x_k} f_{X_k}(t)$.

Note: the above summation proceeds over all the values in the support of \mathbf{x} , *except* those of X_k . If $k = 1$ or $k = K$, (that is to say, X_k is either first or last in the list) then the summation is to be reformulated accordingly.

Example: marginal demographics (1/2)

Recall the example about an imperfect medical treatment with an imperfect take-up in the population.

- Let $X \in \{0, 1\}$ indicate treatment **take-up** and $Y \in \{0, 1\}$ **health status**. Clearly, here (X, Y) is a random vector.
- Let $x = 1$ for a taker, $x = 0$ for an hesitant, $y = 1$ if one stays healthy, $y = 0$ if one gets sick.
- The entire joint p.m.f. is expressed as follows:

$$\begin{aligned}f_{X,Y}(x = 1, y = 1) &= 0.40, & f_{X,Y}(x = 1, y = 0) &= 0.20, \\f_{X,Y}(x = 0, y = 1) &= 0.15, & f_{X,Y}(x = 0, y = 0) &= 0.25.\end{aligned}$$

- This is a *bivariate Bernoulli* distribution.

Example: marginal demographics (2/2)

- The **marginal** distributions are obtained as follows:

$$f_X(x) = f_{X,Y}(x, y = 1) + f_{X,Y}(x, y = 0) \quad \text{for } x = 0, 1$$

$$f_Y(y) = f_{X,Y}(x = 1, y) + f_{X,Y}(x = 0, y) \quad \text{for } y = 0, 1$$

- ...and they can easily be represented with a table.

	$Y = 0$	$Y = 1$	<i>Total</i>
$X = 0$	0.25	0.15	0.40
$X = 1$	0.20	0.40	0.60
<i>Total</i>	0.45	0.55	1

Unsurprisingly, marginal distributions lie at the margins of the table!

Marginal continuous distributions

Definition 8

Marginal Distribution (continuous). For some given random vector \mathbf{x} composed by *continuous* random variables *only*, the probability density function of X_k – the k -th random variable in \mathbf{x} – is obtained as:

$$f_{X_k}(x_k) = \int_{\times_{\ell \neq k} \mathbb{X}_\ell} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}_{-k}$$

and thus $F_{X_k}(x_k) = \int_{-\infty}^{x_k} f_{X_k}(t) dt$.

Note: here, $\times_{\ell \neq k} \mathbb{X}_\ell$ indicates the **Cartesian product** of all the supports of each random variable in \mathbf{x} excluding X_k : e.g. for $k \neq 1, K$ it is $\times_{\ell \neq k} \mathbb{X}_\ell = \mathbb{X}_1 \times \dots \times \mathbb{X}_{k-1} \times \mathbb{X}_{k+1} \times \dots \times \mathbb{X}_K$.

Similarly, the expression $d\mathbf{x}_{-k}$ for the differential of the integral is interpreted as the product of all differentials excluding that of x_k : $d\mathbf{x}_{-k} = dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_K$.

Example: the bivariate normal distribution

A two-dimensional random vector $\mathbf{x} = (X_1, X_2)$ is said to follow a **bivariate** normal distribution if, given some parameters:

$$\mu_1 \in \mathbb{R}, \quad \mu_2 \in \mathbb{R}, \quad \sigma_1 \in \mathbb{R}_{++}, \quad \sigma_2 \in \mathbb{R}_{++}, \quad \rho \in [-1, 1]$$

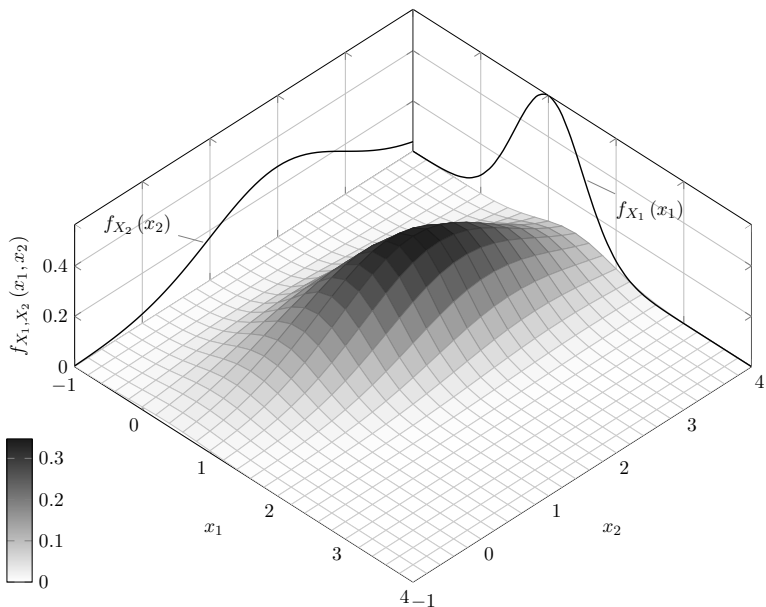
the joint density function $f_{X_1, X_2}(x_1, x_2)$ is expressed as follows.

$$f_{X_1, X_2}(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2(1-\rho^2)} - \frac{(x_2 - \mu_2)^2}{2\sigma_2^2(1-\rho^2)} + \frac{\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2(1-\rho^2)}\right)$$

See e.g. the figure in the next slide for $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 0.5$, $\sigma_2 = 1$ and $\rho = 0.4$.

- Note: the two **marginal** distributions are obtained through integration as $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Drawing the bivariate normal: $\rho = 0.4$



Joint discrete-continuous random variables (1/2)

Most ‘real’ random vectors mix discrete and continuous random variables. The definitions of joint p.m.f. and p.d.f. are not valid for such random vectors as a whole.

Here is an example.

- Let $\mathbf{x} = (H, G)$ where H indicates **height** and $G \in \{0, 1\}$ **gender** (say $G = 1$ for females and $G = 0$ for males).
- A full description of this population can be given as follows.

$$f_{H,G}(h, g = 1; \mu_F, \mu_M, \sigma_F, \sigma_M, p) = \frac{1}{\sigma_F} \phi\left(\frac{h - \mu_F}{\sigma_F}\right) \cdot p$$

$$f_{H,G}(h, g = 0; \mu_F, \mu_M, \sigma_F, \sigma_M, p) = \frac{1}{\sigma_M} \phi\left(\frac{h - \mu_M}{\sigma_M}\right) \cdot (1 - p)$$

(Continues...)

Joint discrete-continuous random variables (2/2)

- **(Continued.)** Summing the two expressions delivers the p.d.f. of H alone:

$$\begin{aligned} f_H(h; \mu_F, \mu_M, \sigma_F, \sigma_M, p) &= \\ &= \frac{1}{\sigma_F} \phi\left(\frac{h - \mu_F}{\sigma_F}\right) \cdot p + \frac{1}{\sigma_M} \phi\left(\frac{h - \mu_M}{\sigma_M}\right) \cdot (1 - p) \end{aligned}$$

- likewise, the marginal distribution of G is returned through integration of both expressions over h ; clearly, $G \sim \text{Be}(p)$.

$$f_G(g = 1; p) = p$$

$$f_G(g = 0; p) = 1 - p$$

- This example is stylized! A joint p.m.f. or p.d.f. can still be defined in a subset of a random vector's components.

Transformations of random vectors (1/3)

- The analysis of transformations extends to random vectors. Let $\mathbf{y} = \mathbf{g}(\mathbf{x})$ where $\mathbf{g}(\cdot)$ is a function taking K arguments and returning J values, with possibly $J \neq K$ (so $|\mathbf{y}| = J$).
- Interest falls on the joint distribution of \mathbf{y} . Such an analysis is tractable if the transformation is invertible, that is, there is a sequence of functions $g_1^{-1}(\cdot), \dots, g_K^{-1}(\cdot)$ such that:

$$X_k = g_k^{-1}(Y_1, \dots, Y_J)$$

for $k = 1, \dots, K$.

- If \mathbf{x} is discrete, the p.m.f. of \mathbf{y} is obtained as:

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}\left(g_1^{-1}(\mathbf{y}), \dots, g_K^{-1}(\mathbf{y})\right)$$

and the c.d.f. $F_{\mathbf{y}}(\mathbf{y})$ is derived consequently.

Transformations of random vectors (2/3)

In the continuous case, the results obtained in Lecture 1 for the univariate case can be extended if the transformation is $\mathbf{g}(\cdot)$ is **bijjective** (one-to-one and onto).

Theorem 1

Joint Density of Transformed Random Vectors. *Consider \mathbf{x} and $\mathbf{y} = \mathbf{g}(\mathbf{x})$: two random vectors of length K that are related through a bijective transformation $\mathbf{g}(\cdot)$ which preserves vector length, \mathbb{X} and \mathbb{Y} their respective supports, and $f_{\mathbf{x}}(\mathbf{x})$ the joint probability density function of \mathbf{x} , which is continuous on \mathbb{X} . If the inverse of the transformation function, $g_k^{-1}(\cdot)$, is continuously differentiable on \mathbb{Y} for $k = 1, \dots, K$, the joint probability density function of \mathbf{y} can be calculated as:*

$$f_{\mathbf{y}}(\mathbf{y}) = \begin{cases} f_{\mathbf{x}}(g_1^{-1}(\mathbf{y}), \dots, g_K^{-1}(\mathbf{y})) \cdot \left| \det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}^{-1}(\mathbf{y}) \right) \right| & \text{if } \mathbf{y} \in \mathbb{Y} \\ 0 & \text{if } \mathbf{y} \notin \mathbb{Y} \end{cases}$$

where $\mathbf{g}^{-1}(\mathbf{y}) = (g_1^{-1}(\mathbf{y}), \dots, g_K^{-1}(\mathbf{y}))^T$.

Transformations of random vectors (3/3)

- Note: in the above statement, the notation

$$\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}^{-1}(\mathbf{y}) = \begin{bmatrix} \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_K} \\ \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_K} \end{bmatrix}$$

indicates the $K \times K$ **Jacobian matrix** of $\mathbf{g}^{-1}(\mathbf{y})$, while

$$\left| \det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}^{-1}(\mathbf{y}) \right) \right|$$

is the **absolute value** of its **determinant**.

- This theorem is an application of Jacobian transformations from multivariate calculus. It can be further extended when $\mathbf{g}(\cdot)$ is bijective in a partition of the support of \mathbf{x} .

Example: bivariate lognormal distribution (1/2)

Consider the previous bivariate normal distribution, and let

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \mathbf{g} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \exp(X_1) \\ \exp(X_2) \end{pmatrix}$$

be a transformed random vector $\mathbf{y} = (Y_1, Y_2)$. Note that:

$$\mathbf{x} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{g}^{-1} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \log(Y_1) \\ \log(Y_2) \end{pmatrix}$$

hence, here the absolute value of the determinant of the inverse transformation is as follows.

$$\det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}^{-1}(y_1, y_2) \right) = \det \begin{pmatrix} 1/y_1 & 0 \\ 0 & 1/y_2 \end{pmatrix} = \frac{1}{y_1 y_2} > 0$$

Example: bivariate lognormal distribution (2/2)

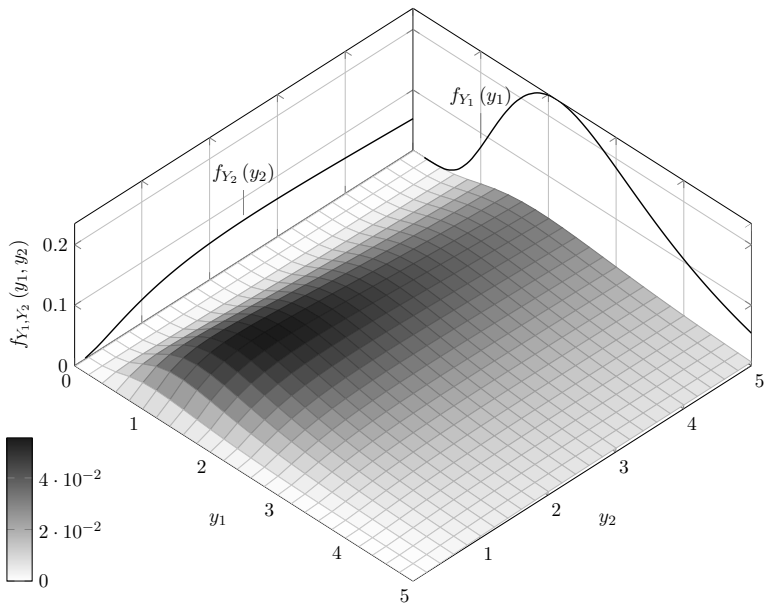
Thus, the joint p.d.f. of the **bivariate lognormal** distribution (with support in \mathbb{R}_{++}^2) can be written as follows.

$$f_{Y_1, Y_2}(y_1, y_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \frac{1}{y_1 y_2} \cdot \exp\left(-\frac{(\log y_1 - \mu_1)^2}{2\sigma_1^2(1-\rho^2)} - \frac{(\log y_2 - \mu_2)^2}{2\sigma_2^2(1-\rho^2)} + \frac{\rho(\log y_1 - \mu_1)(\log y_2 - \mu_2)}{\sigma_1\sigma_2(1-\rho^2)}\right)$$

The distribution is graphically shown in the next figure for the parameters $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 0.5$, $\sigma_2 = 1$ and $\rho = 0.4$.

(Note: this example is relatively simple because the Jacobian is diagonal.)

Drawing the bivariate lognormal



Random matrices

- Random variables can also be arrayed in **random matrices**: collections of random vectors of equal length.
- If a random matrix \mathbf{X} has dimension $K \times J$, it is:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_J \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1J} \\ X_{21} & X_{22} & \dots & X_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ X_{K1} & X_{K2} & \dots & X_{KJ} \end{bmatrix}$$

- ...and the matrix \mathbf{X} of its *realizations* looks alike.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_J \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1J} \\ x_{21} & x_{22} & \dots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K1} & x_{K2} & \dots & x_{KJ} \end{bmatrix}$$

- All previous concepts also extend to random matrices.

Notation for random objects: summary

- X, Y , upper-case italic (slanted) letter: a **random variable**.
- x, y , lower-case italic (slanted) letter: the **realization** of a **random variable**.
- \mathbf{x}, \mathbf{y} , lower-case italic (slanted) letter: a **random vector**.
- \mathbf{x}, \mathbf{y} , lower-case roman letter: the **realization** of a **random vector**.
- \mathbf{X}, \mathbf{Y} , upper-case italic (slanted) letter: a **random matrix**.
- \mathbf{X}, \mathbf{Y} , upper-case roman letter: the **realization** of a **random matrix**.

Independent random variables

Some random variables that are possibly collected in a random vector are said to be **independent**: intuitively, the realization of one provides no information about those of the others.

In other words, the events described by these random variables are **statistically independent**.

Definition 9

Independent Random Variables. Let $\mathbf{x} = (X, Y)$ be a random vector with joint probability mass or density function $f_{X,Y}(x, y)$, and marginal mass or density functions $f_X(x)$ and $f_Y(y)$. Lastly, let uppercase F denote corresponding cumulative distributions instead (joint or marginal). The two random variables X and Y are *independent* if the two equivalent conditions below hold.

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \iff F_{X,Y}(x, y) = F_X(x) F_Y(y)$$

Mutually independent random variables

The idea extends naturally to grouped random variables.

Definition 10

Mutually – or Pairwise – Independent Random Variables. Let $\mathbf{x} = (X_1, \dots, X_K)$ be a random vector with joint probability mass or density function $f_{\mathbf{x}}(\mathbf{x})$, and marginal mass or density functions $f_{X_1}(x_1), \dots, f_{X_K}(x_K)$. Instead, let $F_{\mathbf{x}}(\mathbf{x})$ denote corresponding cumulative distributions (either joint or marginal). The random variables X_1, \dots, X_K are *pairwise independent* if every pair of random variables listed in \mathbf{x} are independent, and they are *mutually independent* if the two equivalent conditions below hold.

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{k=1}^K f_{X_k}(x_k) \iff F_{\mathbf{x}}(\mathbf{x}) = \prod_{k=1}^K F_{X_k}(x_k)$$

Note that while mutual independence implies pairwise independence, the converse is not true.

Independent random vectors (1/2)

Definitions that are specific to random vectors follow.

Definition 11

Independent Random Vectors. Let $(\mathbf{x}_1, \dots, \mathbf{x}_J)$ be a sequence of J random vectors whose joint probability mass or density function is written as $f_{\mathbf{x}_1, \dots, \mathbf{x}_J}(\mathbf{x}_1, \dots, \mathbf{x}_J)$. Let the joint probability mass or density functions of an individual nested random vector be $f_{\mathbf{x}_i}(\mathbf{x}_i)$, where $i = 1, \dots, J$, and the joint probability mass or density functions of any two random vectors indexed i, j (with $i \neq j$) as $f_{\mathbf{x}_i, \mathbf{x}_j}(\mathbf{x}_i, \mathbf{x}_j)$. Lastly, let $F_{\mathbf{x}}(\mathbf{x})$ denote corresponding cumulative distributions instead (joint or marginal). Any pair of random vectors indexed i and j are *independent* if the two equivalent conditions below hold.

$$\begin{aligned}f_{\mathbf{x}_i, \mathbf{x}_j}(\mathbf{x}_i, \mathbf{x}_j) &= f_{\mathbf{x}_i}(\mathbf{x}_i) f_{\mathbf{x}_j}(\mathbf{x}_j) \\ F_{\mathbf{x}_i, \mathbf{x}_j}(\mathbf{x}_i, \mathbf{x}_j) &= F_{\mathbf{x}_i}(\mathbf{x}_i) F_{\mathbf{x}_j}(\mathbf{x}_j)\end{aligned}$$

If the above holds for any i, j distinct pair, the J random vectors are said to be *pairwise independent*. (**Continues...**)

Independent random vectors (2/2)

Definition 11

(Continued.) If the above holds for any i, j distinct pair, the J random vectors are said to be *pairwise independent*. The J random vectors are *mutually independent* if the two equivalent conditions below hold.

$$f_{\mathbf{x}_1, \dots, \mathbf{x}_J}(\mathbf{x}_1, \dots, \mathbf{x}_J) = \prod_{i=1}^J f_{\mathbf{x}_i}(\mathbf{x}_i)$$
$$F_{\mathbf{x}_1, \dots, \mathbf{x}_J}(\mathbf{x}_1, \dots, \mathbf{x}_J) = \prod_{i=1}^J F_{\mathbf{x}_i}(\mathbf{x}_i)$$

Note that within each random vector the underlying random variables are not necessarily independent. In addition, if all the random vectors in question have length one, all these definitions reduce to those given above for random variables.

Independent random variables and events (1/2)

It now gets easier to appreciate the connection with statistically independent events.

Theorem 2

Independence of Events. *Any two events that are mapped by two independent random variables X and Y are statistically independent.*

Proof.

(Outline.) This requires to show that, for any two events $\mathbb{A} \subset \mathbb{S}_X$ and $\mathbb{B} \subset \mathbb{S}_Y$ – where \mathbb{S}_X and \mathbb{S}_Y are the primitive sample spaces of X and Y respectively – it is:

$$\mathbb{P}(X \in X(\mathbb{A}) \cap Y \in Y(\mathbb{B})) = \mathbb{P}(X \in X(\mathbb{A})) \cdot \mathbb{P}(Y \in Y(\mathbb{B}))$$

which follows from the definitions of (joint) cumulative distribution, mass and density functions, and that of independent events. \square

Independent random variables and events (2/2)

The logic applies to multiple random variables at once.

Theorem 2

Generalization: Mutual Independence between Events. *Any combination of events that are mapped by a sequence of J mutually independent random vectors $(\mathbf{x}_1, \dots, \mathbf{x}_J)$ are mutually independent.*

Proof.

(Outline.) Extending the reasoning above, consider a collection of J events denoted by $\mathbb{A}_i \subset \mathbb{S}_{\mathbf{x}_i}$ for $i = 1, \dots, J$, where $\mathbb{S}_{\mathbf{x}_i}$ is the primitive sample space of \mathbf{x}_i . It must be shown that:

$$\mathbb{P} \left(\bigcap_{i=1}^J (\mathbf{x}_i \in \mathbf{x}_i(\mathbb{A}_i)) \right) = \prod_{i=1}^J \mathbb{P}(\mathbf{x}_i \in \mathbf{x}_i(\mathbb{A}_i))$$

which follows by analogous considerations. □

Independent functions of random variables (1/2)

Here is a useful result for later: independence between random variables is preserved by transformations.

Theorem 3

Independence of Functions of Random Variables. *Consider two independent random variables X and Y , and let $U = g_X(X)$ be a transformation of X and $V = g_Y(Y)$ a transformation of Y . The two transformed random variables U and V are independent.*

Proof.

(*Outline.*) This requires to show that,

$$f_{U,V}(u,v) = f_U(u) f_V(v) \iff F_{U,V}(u,v) = F_U(u) F_V(v)$$

which is achieved by manipulating the inverse mappings $g_X^{-1}([a,b])$ and $g_Y^{-1}([a,b])$ for any appropriate interval $[a,b] \subset \mathbb{R}$, with $a \leq b$. \square

Independent functions of random variables (2/2)

This also applies to a multivariate environment, of course.

Theorem 3

Generalization: Independence of Functions of Random Vectors. *Consider a sequence of mutually independent random vectors $(\mathbf{x}_1, \dots, \mathbf{x}_J)$, as well as a sequence of transformations $(\mathbf{y}_1, \dots, \mathbf{y}_J)$ such that $\mathbf{y}_i = \mathbf{g}_i(\mathbf{x}_i)$ for $i = 1, \dots, J$. The J transformed random vectors $(\mathbf{y}_1, \dots, \mathbf{y}_J)$ are also themselves mutually independent.*

Proof.

(Outline.) The proof extends the logic behind the previous result from the bivariate case to higher dimensions; it requires manipulating the J Jacobian transformations at hand. □

Random products and random ratios

- Notions of independence are central in statistics, first and foremost for specifying a framework for **estimation**.
- They have many applications, including the derivation of the distribution of **random products** or **random ratios** of two independent random variables.
- Specifically, let X_1 and X_2 be two **independent** random variables. Consider transformations like the following:

$$Y = X_1X_2 \quad \text{or} \quad Y = \frac{X_1}{X_2}.$$

- The distribution of Y can be derived via some steps based upon multivariate transformations. This is illustrated via examples that are also relevant for statistical inference.

Cauchy distribution as a random ratio (1/4)

Observation 1

If $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(0, 1)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as

$$Y = \frac{X_1}{X_2}$$

is such that $Y \sim \text{Cauchy}(0, 1)$.

Proof.

This is shown in three steps:

1. derive the joint distribution of (X_1, X_2) ;
2. derive the joint distribution of (Y, Z) , where $Z = |X_2|$;
3. derive the marginal distribution of Y accordingly.

The first step is the easiest. Here is where independence is applied.

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$$

Cauchy distribution as a random ratio (2/4)

Proof.

The second step is not immediate because this transformation is not bijective. Yet Theorem 1 can be applied to partitions of the support where the transformation is bijective. Note that if $\mathbb{X} = \mathbb{R}^2$ is split as:

$$\mathbb{X}_0 = \{(x_1, x_2) : x_2 = 0\}$$

$$\mathbb{X}_1 = \{(x_1, x_2) : x_2 < 0\}$$

$$\mathbb{X}_2 = \{(x_1, x_2) : x_2 > 0\}$$

the transformation is bijective on \mathbb{X}_1 & \mathbb{X}_2 with image $\mathbb{Y} = \mathbb{R} \times \mathbb{R}_+$. Also, $\mathbb{P}(X_2 = 0) = 0$ so \mathbb{X}_0 can be ignored. Consider \mathbb{X}_1 : there, it is $Z = -X_2$ and thus:

$$X_1 = g_{1, \mathbb{X}_1}^{-1}(Y, Z) = -YZ \quad \text{and} \quad X_2 = g_{2, \mathbb{X}_1}^{-1}(Y, Z) = -Z;$$

while in \mathbb{X}_2 it is $Z = X_2$ with the following inverse transformation:

$$X_1 = g_{1, \mathbb{X}_2}^{-1}(Y, Z) = YZ \quad \text{and} \quad X_2 = g_{2, \mathbb{X}_2}^{-1}(Y, Z) = Z.$$

Cauchy distribution as a random ratio (3/4)

Proof.

Thus, the determinant of the Jacobian in \mathbb{X}_1 is:

$$\det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}_{\mathbb{X}_1}^{-1}(y, z) \right) = \det \begin{pmatrix} -z & -y \\ 0 & -1 \end{pmatrix} = z > 0$$

which is always positive; in \mathbb{X}_2 it is identical.

$$\det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}_{\mathbb{X}_2}^{-1}(y, z) \right) = \det \begin{pmatrix} z & y \\ 0 & 1 \end{pmatrix} = z > 0$$

The second step is accomplished by deriving the joint distribution of (Y, Z) ; specifically this is obtained by separately applying Theorem 1 on both \mathbb{X}_1 and \mathbb{X}_2 and summing up the resulting density functions. Hence:

$$f_{Y,Z}(y, z) = \frac{z}{\pi} \exp \left(-\frac{(y^2 + 1) z^2}{2} \right)$$

and all is left to do is to integrate out z (step three).

Cauchy distribution as a random ratio (4/4)

Proof.

To proceed, note that the support of z is \mathbb{R}_+ .

$$\begin{aligned} f_Y(y) &= \int_0^{+\infty} f_{Y,Z}(y, z) dz \\ &= \int_0^{+\infty} \frac{z}{\pi} \exp\left(-\frac{(y^2+1)z^2}{2}\right) dz \\ &= \int_0^{+\infty} \frac{1}{2\pi} \exp\left(-\frac{(y^2+1)u}{2}\right) du \\ &= \frac{1}{\pi(y^2+1)} \int_0^{+\infty} \frac{(y^2+1)}{2} \exp\left(-\frac{(y^2+1)u}{2}\right) du \\ &= \frac{1}{\pi(y^2+1)} \end{aligned}$$

The third line applies the change of variable $u = z^2$ while the integral in the fourth line vanishes: it is the total probability of an exponential distribution. The result is the standard Cauchy density. \square

Student's t -distribution as a random ratio (1/3)

Observation 2

If $Z \sim \mathcal{N}(0, 1)$ and $X \sim \chi^2(\nu)$, and the two random variables Z and X are independent, the random variable Y obtained as

$$Y = \frac{Z}{\sqrt{\frac{X}{\nu}}}$$

is such that $Y \sim \mathcal{T}(\nu)$.

Proof.

The steps here are identical as in the 'normals-ratio-to-Cauchy' case; however in place of X_2 here is $W = \sqrt{X/\nu}$ where $X \sim \chi^2(\nu)$. Thus, one should first derive the p.d.f. of W . Note that the transformation that defines W is increasing; the inverse is $X = g^{-1}(W) = \nu W^2$ and thus $\frac{dx}{dw} = 2\nu w > 0$; the support stays \mathbb{R}_{++} and hence:

$$f_W(w; \nu) = \frac{\nu^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\frac{\nu}{2}-1}} w^{\nu-1} \exp\left(-\frac{\nu w^2}{2}\right) \quad \text{for } w > 0.$$

Student's t -distribution as a random ratio (2/3)

Proof.

Because of independence, the joint density of $\mathbf{w} = (Z, W)$ is:

$$f_{\mathbf{w}}(z, w; \nu) = \frac{1}{\sqrt{2\pi}} \frac{\nu^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\frac{\nu}{2}-1}} w^{\nu-1} \exp\left(-\frac{z^2 + \nu w^2}{2}\right)$$

and this completes the (longer) first step. Yet the second step is easier now: it is necessary to derive the joint distribution of $\mathbf{y} = (Y, W)$, but this specific transformation is already bijective and there is no need to split the support. Similarly to the Cauchy case, the determinant of the Jacobian is $w > 0$; and since $Z = YW$, the joint p.d.f. of interest is:

$$\begin{aligned} f_{\mathbf{y}}(y, w; \nu) &= w \cdot f_{\mathbf{w}}(yw, w; \nu) \\ &= \frac{\nu^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\frac{\nu-1}{2}}} \frac{1}{\sqrt{\pi\nu}} w^{\nu} \exp\left(-\frac{(y^2 + \nu) w^2}{2}\right) \end{aligned}$$

for $y \in \mathbb{R}$ and $w \in \mathbb{R}_{++}$. The last step is about integrating out w ; this requires a change of variable: $u = w^2$ so as to recognize the integral as the total probability of a Gamma-distributed random variable.

Student's t -distribution as a random ratio (3/3)

Proof.

$$\begin{aligned}f_Y(y; \nu) &= \int_0^{+\infty} f_{Y,W}(y, w) dw \\&= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \int_0^{+\infty} \frac{\nu^{\frac{\nu+1}{2}}}{2^{\frac{\nu-1}{2}}} w^\nu \exp\left(-\frac{(\nu + y^2) w^2}{2}\right) dw \\&= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \int_0^{+\infty} \left(\frac{\nu}{2}\right)^{\frac{\nu+1}{2}} u^{\frac{\nu-1}{2}} \exp\left(-\frac{\nu}{2} \left(1 + \frac{y^2}{\nu}\right) u\right) du \\&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}} \times \\&\times \int_0^{+\infty} \frac{1}{\Gamma\left(\frac{\nu+1}{2}\right)} \left(\frac{\nu + y^2}{2}\right)^{\frac{\nu+1}{2}} u^{\frac{\nu-1}{2}} \exp\left(-\left(\frac{\nu + y^2}{2}\right) u\right) du \\&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}\end{aligned}$$

The result is the p.d.f. of a t -distribution with parameter ν . □

Other important random ratios

Observation 3

If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as

$$Y = \frac{X_1/\nu_1}{X_2/\nu_2}$$

is such that $Y \sim \mathcal{F}(\nu_1, \nu_2)$.

Observation 4

If $X_1 \sim \Gamma(\alpha, \gamma)$ and $X_2 \sim \Gamma(\beta, \gamma)$, and the two random variables X_1 and X_2 are independent, the random variables Y and W obtained as

$$Y = \frac{X_1}{X_1 + X_2}$$
$$W = X_1 + X_2$$

are independent and such that $Y \sim \text{Beta}(\alpha, \beta)$ and $W \sim \Gamma(\alpha + \beta, \gamma)$.

Multivariate moments

The concepts of **moments** extend straightforwardly to all the marginal distributions of a random vector. Therefore, the r -th **uncentered** moment of the k -th random variable in a random vector \mathbf{x} also obtains as:

$$\mathbb{E} [X_k^r] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} x_k^r f_{\mathbf{x}}(\mathbf{x})$$

$$\mathbb{E} [X_k^r] = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_K} x_k^r f_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x}$$

if \mathbf{x} is all-discrete or all-continuous respectively (where $d\mathbf{x}$ is the product of all differentials); **centered** moments are analogous.

$$\mathbb{E} [(X_k - \mathbb{E} [X_k])^r] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} (x_k - \mathbb{E} [X_k])^r f_{\mathbf{x}}(\mathbf{x})$$

$$\mathbb{E} [(X_k - \mathbb{E} [X_k])^r] = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_K} (x_k - \mathbb{E} [X_k])^r f_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x}$$

Covariance

Multivariate distributions allow for important “cross-moments” that express to what extent two random variables tend to move together in a probabilistic sense.

Definition 12

Covariance. For any two random variables X_k and X_ℓ belonging to a random vector \mathbf{x} , their **covariance** is defined as the expectation of a particular function of X_k and X_ℓ , i.e. the product of both variables’ deviations from their respective means.

$$\text{Cov} [X_k, X_\ell] = \mathbb{E} [(X_k - \mathbb{E} [X_k]) (X_\ell - \mathbb{E} [X_\ell])]$$

For respectively all-discrete and all-continuous random vectors, covariances can also be expressed as follows.

$$\text{Cov} [X_k, X_\ell] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} (x_k - \mathbb{E} [X_k]) (x_\ell - \mathbb{E} [X_\ell]) f_{\mathbf{x}} (\mathbf{x})$$

$$\text{Cov} [X_k, X_\ell] = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_K} (x_k - \mathbb{E} [X_k]) (x_\ell - \mathbb{E} [X_\ell]) f_{\mathbf{x}} (\mathbf{x}) d\mathbf{x}$$

Covariance and Correlation

A covariance takes positive values if the two variables X_k and X_ℓ move together, negative values if those proceed in opposite directions. It can be expressed as more fundamental moments.

$$\begin{aligned}\mathbb{Cov} [X_k, X_\ell] &= \mathbb{E} [(X_k - \mathbb{E} [X_k]) (X_\ell - \mathbb{E} [X_\ell])] \\ &= \mathbb{E} [X_k X_\ell] - \mathbb{E} [X_k \mathbb{E} [X_\ell]] \\ &\quad - \mathbb{E} [X_\ell \mathbb{E} [X_k]] + \mathbb{E} [X_k] \mathbb{E} [X_\ell] \\ &= \mathbb{E} [X_k X_\ell] - \mathbb{E} [X_k] \mathbb{E} [X_\ell]\end{aligned}$$

Covariances can be *normalized*, thus becoming *correlations*.

Definition 13

Correlation. For any two random variables X_k and X_ℓ belonging to a random vector \mathbf{x} , their *population* correlation is defined as follows.

$$\mathbb{Corr} [X_k, X_\ell] = \frac{\mathbb{Cov} [X_k, X_\ell]}{\sqrt{\mathbb{Var} [X_k]} \sqrt{\mathbb{Var} [X_\ell]}}$$

Properties of Correlations (1/3)

Theorem 4

Properties of Correlation. *For any two random variables X and Y , it is:*

- $\text{Corr}[X, Y] \in [-1, 1]$, and
- $|\text{Corr}[X, Y]| = 1$ if and only if there are some real numbers $a \neq 0$ and b such that $\mathbb{P}(Y = aX + b) = 1$. If $\text{Corr}[X, Y] = 1$ then it is $a > 0$, while if $\text{Corr}[X, Y] = -1$ it is $a < 0$.

Proof.

Define the following function:

$$\begin{aligned} C(t) &= \mathbb{E} \left[[(X - \mathbb{E}[X]) \cdot t + (Y - \mathbb{E}[Y])]^2 \right] \\ &= \text{Var}[X] \cdot t^2 + 2 \text{Cov}[X, Y] \cdot t + \text{Var}[Y] \end{aligned}$$

which must be nonnegative, because it is defined as the expectation of the square of a random variable. **(Continues...)**

Properties of Correlations (2/3)

Theorem 4

Proof.

(Continued.) As $\mathcal{C}(t)$ is nonnegative, its discriminant must be non-positive:

$$(2 \operatorname{Cov}[X, Y])^2 - 4 \operatorname{Var}[X] \operatorname{Var}[Y] \leq 0.$$

or, equivalently:

$$-\sqrt{\operatorname{Var}[X]} \sqrt{\operatorname{Var}[Y]} \leq \operatorname{Cov}[X, Y] \leq \sqrt{\operatorname{Var}[X]} \sqrt{\operatorname{Var}[Y]}.$$

thus **a.** is proved. Next, consider that $|\operatorname{Corr}[X, Y]| = 1$ only if:

$$(2 \operatorname{Cov}[X, Y])^2 - 4 \operatorname{Var}[X] \operatorname{Var}[Y] = 0.$$

that is, the discriminant is just zero, and $\mathcal{C}(t)$ has exactly one solution for t . Proving **b.** is thus about finding this solution in terms of moments of X and Y . **(Continues...)**

Properties of Correlations (3/3)

Theorem 4

Proof.

(Continued.) As $\mathcal{C}(t)$ is nonnegative, such a solution for t must also satisfy:

$$\mathbb{P}\left(\left[(X - \mathbb{E}[X])t + (Y - \mathbb{E}[Y])\right]^2 = 0\right) = 1$$

or equivalently:

$$\mathbb{P}\left((X - \mathbb{E}[X])t + (Y - \mathbb{E}[Y]) = 0\right) = 1$$

which occurs if and only if $Y = aX + b$ for:

$$a = -t \quad \text{and} \quad b = \mathbb{E}[X] \cdot t + \mathbb{E}[Y]$$

while, as the discriminant is zero, the solution reads:

$$t = -\frac{\text{Cov}[X, Y]}{\text{Var}[X]}$$

which proves **b.** since a and $\text{Corr}[X, Y]$ must share the same sign. \square

Cross-expectation and independence (1/2)

Theorem 5

Cross-expectation of independent random variables. *Given two independent random variables X and Y , it is*

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$$

if the above moments exist.

Proof.

The left hand side of the above relationship is the expectation of a random variable which is defined as the product of X and Y :

$$\begin{aligned} \int_{\mathbf{X}} \int_{\mathbf{Y}} xy f_{X,Y}(x,y) dx dy &= \int_{\mathbf{X}} \int_{\mathbf{Y}} xy f_X(x) f_Y(y) dx dy \\ &= \int_{\mathbf{X}} x f_X(x) dx \cdot \int_{\mathbf{Y}} y f_Y(y) dy \end{aligned}$$

falling back to the product of two expressions that correspond to the definition of mean (for X and Y respectively); observe that the first equality exploits the definition of independent random variables. \square

Cross-expectation and independence (2/2)

Corollary 1

(Theorem 5.) *Both the covariance and the correlation between two independent random variables X and Y equal zero.*

Corollary 2

(Theorem 5.) *Given any transformations $U = g_X(X)$, $V = g_Y(Y)$ of two independent random variables X and Y whose moments exist, it is:*

$$\mathbb{E}[UV] = \mathbb{E}[U] \mathbb{E}[V]$$

because U and V are also independent; this also implies that U and V have zero covariance and correlation and that also all higher moments of X and Y inherit this property. For example:

$$\begin{aligned}\text{Var}[XY] &= \mathbb{E} \left[(X - \mathbb{E}[X])^2 (Y - \mathbb{E}[Y])^2 \right] \\ &= \mathbb{E} \left[(X - \mathbb{E}[X])^2 \right] \mathbb{E} \left[(Y - \mathbb{E}[Y])^2 \right] \\ &= \text{Var}[X] \text{Var}[Y]\end{aligned}$$

which is best seen by setting $U = (X - \mathbb{E}[X])^2$ and $V = (Y - \mathbb{E}[Y])^2$.

Covariance in the bivariate normal (1/4)

In a bivariate normal distribution, the following holds.

$$\mathbb{E}[X_1 X_2] = \rho \sigma_1 \sigma_2 + \mu_1 \mu_2$$

This shown in a few steps. First, define the transformation from $\mathbf{x} = (X_1, X_2)$ to $\mathbf{y} = (Y, Z)$ as follows.

$$Y = \frac{X_1 - \mu_1}{\sigma_1} \frac{X_2 - \mu_2}{\sigma_2}$$
$$Z = \frac{X_1 - \mu_1}{\sigma_1}$$

Note that the support of \mathbf{y} is $\mathbb{Y} = \mathbb{R}^2$ and the transformation is clearly bijective. The inverse transformation is as follows.

$$X_1 = g_1^{-1}(Y, Z) = \sigma_1 Z + \mu_1$$
$$X_2 = g_2^{-1}(Y, Z) = \sigma_2 \frac{Y}{Z} + \mu_2$$

Covariance in the bivariate normal (2/4)

Here the Jacobian has the following absolute determinant:

$$\left| \det \left(\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}^{-1}(y, z) \right) \right| = \left| \det \begin{pmatrix} 0 & \sigma_1 \\ \frac{\sigma_2}{z} & -\frac{\sigma_2 y}{z} \end{pmatrix} \right| = \frac{\sigma_1 \sigma_2}{|z|} = \frac{\sigma_1 \sigma_2}{\sqrt{z^2}}$$

thus, by Theorem 1 the joint p.d.f. of $\mathbf{y} = (Y, Z)$ is as follows.

$$\begin{aligned} f_{Y,Z}(y, z; \rho) &= \frac{1}{2\pi \sqrt{(1-\rho^2)} z^2} \exp \left(-\frac{z^2 - 2\rho y + y^2 z^{-2}}{2(1-\rho^2)} \right) \\ &= \phi(z) \cdot \frac{1}{\sqrt{2\pi(1-\rho^2)} z^2} \exp \left(-\frac{(y - \rho z^2)^2}{2(1-\rho^2) z^2} \right) \end{aligned}$$

Recall that $\phi(z)$ is the standard normal's p.d.f.; note that

$$z^2 - 2\rho y + y^2 z^{-2} = (1 - \rho^2) z^2 + (y - \rho z^2)^2 z^{-2}$$

hence the second line.

Covariance in the bivariate normal (3/4)

The next step is to calculate the mean of Y .

$$\begin{aligned}\mathbb{E}[Y] &= \int_{-\infty}^{+\infty} \phi(z) \\ &\quad \times \underbrace{\left[\int_{-\infty}^{+\infty} \frac{y}{\sqrt{2\pi(1-\rho^2)}z^2} \exp\left(-\frac{(y-\rho z^2)^2}{2(1-\rho^2)z^2}\right) dy \right]}_{=\rho z^2} dz \\ &= \rho \int_{-\infty}^{+\infty} z^2 \phi(z) dz \\ &= \rho\end{aligned}$$

- The simplification in the second line occurs because there, the integral is the mean of a normally distributed random variable with mean ρz^2 and variance $(1-\rho^2)z^2$.
- The integral in the second line vanishes since $Z \sim \mathcal{N}(0, 1)$ and $\mathbb{E}[Z^2] = 1$.

Covariance in the bivariate normal (4/4)

In summary, one can conclude that:

$$\begin{aligned}\mathbb{E}[X_1 X_2] &= \mathbb{E}\left[(\sigma_1 Z + \mu_1) \left(\sigma_2 \frac{Y}{Z} + \mu_2\right)\right] \\ &= \sigma_1 \sigma_2 \mathbb{E}[Y] + \sigma_1 \mu_2 \mathbb{E}[Z] + \sigma_2 \mu_1 \mathbb{E}\left[\frac{Y}{Z}\right] + \mu_1 \mu_2 \\ &= \rho \sigma_1 \sigma_2 + \mu_1 \mu_2\end{aligned}$$

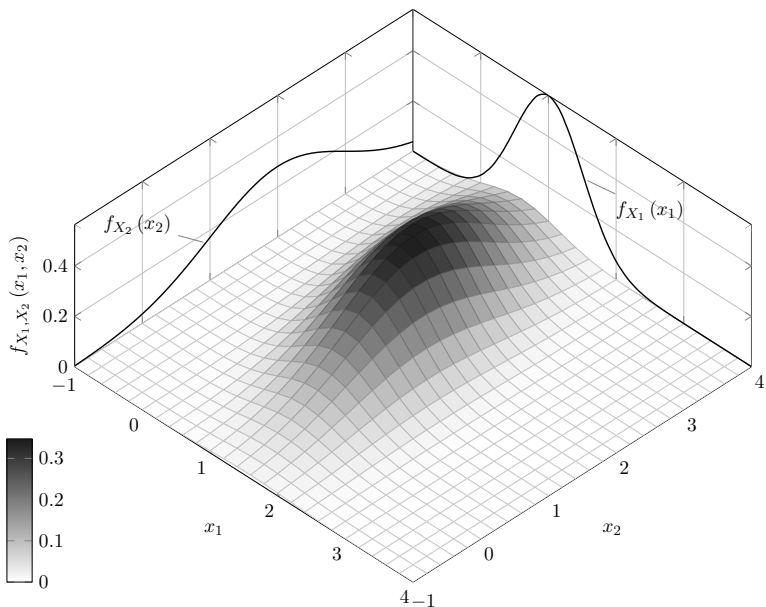
as postulated, because $\mathbb{E}[Z] = \mathbb{E}\left[\frac{Y}{Z}\right] = 0$ are both expectations of random variables following the standard normal distribution. In light of this result, it is:

$$\text{Cov}[X_1, X_2] = \rho \sigma_1 \sigma_2$$

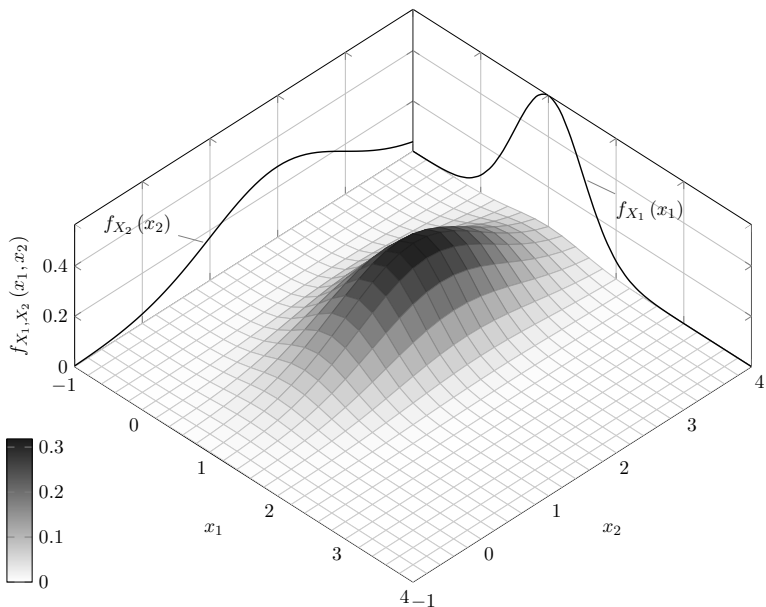
$$\text{Corr}[X_1, X_2] = \rho$$

hence, parameter ρ has an interpretation as *correlation* (and in fact its range is confined in the $[-1, 1]$ interval).

Drawing the bivariate normal: $\rho = -0.4$



Drawing the bivariate normal: $\rho = 0$



Collecting terms

It is convenient to summarize all moments of a random vector of the same order through **compact notation**. To begin with, the **mean vector** $\mathbb{E}[\mathbf{x}]$ gathers all the means.

$$\mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_K] \end{bmatrix}$$

The **variance-covariance matrix** $\text{Var}[\mathbf{x}]$ instead collects its namesakes; can be seen as the expectation of a random matrix.

$$\begin{aligned} \text{Var}[\mathbf{x}] &= \mathbb{E} \left[(\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \right] \\ &= \begin{bmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \dots & \text{Cov}[X_1, X_K] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \dots & \text{Cov}[X_2, X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_K, X_1] & \text{Cov}[X_K, X_2] & \dots & \text{Var}[X_K] \end{bmatrix} \end{aligned}$$

Properties of variance-covariance matrices

- $\text{Var}[\mathbf{x}]$ it has dimension $K \times K$ and is symmetric.
- The elements along the diagonal of $\text{Var}[\mathbf{x}]$ (the variances), are always nonnegative; the elements outside the diagonal (the covariances), can be negative.
- The random matrix expression can be simplified

$$\begin{aligned}\text{Var}[\mathbf{x}] &= \mathbb{E} \left[(\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \right] \\ &= \mathbb{E} \left[\mathbf{x}\mathbf{x}^T \right] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{x}]^T \\ &\quad - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{x}]^T + \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{x}]^T \\ &= \mathbb{E} \left[\mathbf{x}\mathbf{x}^T \right] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{x}]^T\end{aligned}$$

- $\text{Var}[\mathbf{x}]$ is positive semi-definite: for any non-zero vector \mathbf{a} of length K , the quadratic form $\mathbf{a}^T \text{Var}[\mathbf{x}] \mathbf{a} \geq 0$ cannot be negative. This property is demonstrated later.

Example: moments of the bivariate normal

- The mean vector of the bivariate normal is usually denoted as follows.

$$\boldsymbol{\mu} \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mathbb{E}[\mathbf{x}]$$

- Instead, here the notation for the variance-covariance matrix is the following.

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \text{Var}[\mathbf{x}]$$

- Note: $\boldsymbol{\Sigma}$ satisfies all the properties of a variance-covariance matrix just outlined.
- If $\mathbf{x} = (X_1, X_2)$ follows a bivariate normal distribution with parameters specified as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, this is denoted as follows.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Moments of linear transformations (1/3)

Like in the univariate case, it is useful to derive the moment of a transformed random vector in terms of the original moments.

Consider simple **linear transformations** of a random vector \mathbf{x} that return a random variable Y :

$$Y = \mathbf{a}^T \mathbf{x} = a_1 X_1 + \cdots + a_K X_K$$

where $\mathbf{a} = (a_1, \dots, a_K)^T$ has length K like $\mathbf{x} = (X_1, \dots, X_K)^T$.

The mean of Y is obtained as follows.

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[\mathbf{a}^T \mathbf{x}] \\ &= \mathbb{E}[a_1 X_1 + \cdots + a_K X_K] \\ &= a_1 \mathbb{E}[X_1] + \cdots + a_K \mathbb{E}[X_K] \\ &= \mathbf{a}^T \mathbb{E}[\mathbf{x}] \end{aligned}$$

Moments of linear transformations (2/3)

The variance of Y is instead obtained as follows.

$$\begin{aligned}\text{Var}[Y] &= \text{Var}[\mathbf{a}^T \mathbf{x}] \\ &= \mathbb{E} \left[\left(\mathbf{a}^T \mathbf{x} - \mathbb{E}[\mathbf{a}^T \mathbf{x}] \right) \left(\mathbf{a}^T \mathbf{x} - \mathbb{E}[\mathbf{a}^T \mathbf{x}] \right)^T \right] \\ &= \mathbb{E} \left[\mathbf{a}^T (\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \mathbf{a} \right] \\ &= \mathbf{a}^T \mathbb{E} \left[(\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \right] \mathbf{a} \\ &= \mathbf{a}^T \text{Var}[\mathbf{x}] \mathbf{a}\end{aligned}$$

This shows that variance-covariance matrices must be positive semi-definite: any quadratic form $\mathbf{a}^T \text{Var}[\mathbf{x}] \mathbf{a}$ corresponds with a variance $\text{Var}[Y] \geq 0$. This can be expanded as follows.

$$\text{Var}[Y] = \sum_{k=1}^K \left[a_k^2 \text{Var}[X_k] + 2 \sum_{\ell=1}^{k-1} a_k a_\ell \text{Cov}[X_k, X_\ell] \right]$$

Moments of linear transformations (3/3)

What if a linear transformation returns another random vector?

Let:

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x} = (Y_1, \dots, Y_J)^T$$

be a J -dimensional vector obtained via a **linear** transformation of \mathbf{x} where vector \mathbf{a} has length J and matrix \mathbf{B} has size $J \times K$.

The mean of \mathbf{y} is obtained as:

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{a} + \mathbf{B}\mathbf{x}] = \mathbf{a} + \mathbf{B} \mathbb{E}[\mathbf{x}]$$

while its variance-covariance is as follows.

$$\text{Var}[\mathbf{y}] = \text{Var}[\mathbf{a} + \mathbf{B}\mathbf{x}] = \mathbf{B} \text{Var}[\mathbf{x}] \mathbf{B}^T$$

To appreciate this expression, note that if \mathbf{b}_i and \mathbf{b}_j are the i -th and the j -th rows of \mathbf{B} , then the ij -th element of $\text{Var}[\mathbf{y}]$ equals $\text{Cov}[\mathbf{b}_i^T \mathbf{x}, \mathbf{b}_j^T \mathbf{x}] = \mathbf{b}_i^T \text{Var}[\mathbf{x}] \mathbf{b}_j$.

Moments of non-linear transformations

Consider instead **non-linear transformations** like:

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) = (Y_1, \dots, Y_J)^T$$

is obtained from a generic J -valued function $\mathbf{g}(\cdot)$ of \mathbf{x} . Like in the univariate case, a Taylor expansion helps the analysis:

$$\begin{aligned}\mathbf{g}(\mathbf{x}) &\approx \mathbf{g}(\mathbb{E}[\mathbf{x}]) + \frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}(\mathbb{E}[\mathbf{x}]) [\mathbf{x} - \mathbb{E}[\mathbf{x}]] \\ &\approx \left[\mathbf{g}(\mathbb{E}[\mathbf{x}]) - \frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}(\mathbb{E}[\mathbf{x}]) \mathbb{E}[\mathbf{x}] \right] + \frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}(\mathbb{E}[\mathbf{x}]) \mathbf{x}\end{aligned}$$

showing that $\mathbb{E}[\mathbf{g}(\mathbf{x})] \approx \mathbf{g}(\mathbb{E}[\mathbf{x}])$ can be a **bad** approximation. However:

$$\text{Var}[\mathbf{g}(\mathbf{x})] \approx \left[\frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}(\mathbb{E}[\mathbf{x}]) \right] \text{Var}[\mathbf{x}] \left[\frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}(\mathbb{E}[\mathbf{x}]) \right]^T$$

is a generally **good** approximation for the variance of \mathbf{y} .

Cross-covariance matrix

Sometimes it is useful to collect all the covariances between the elements of a random vector \mathbf{x} of dimension $K_{\mathbf{x}}$ and those of a random vector \mathbf{y} of dimension $K_{\mathbf{y}}$.

The resulting $K_{\mathbf{x}} \times K_{\mathbf{y}}$ matrix is named the **cross-covariance matrix** and is denoted as $\text{Cov}[\mathbf{x}, \mathbf{y}]$.

Just like the variance-covariance matrix, it is also defined as the expectation of a random matrix.

$$\begin{aligned} \text{Cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T] \\ &= \begin{bmatrix} \text{Cov}[X_1, Y_1] & \text{Cov}[X_1, Y_2] & \dots & \text{Cov}[X_1, Y_{K_{\mathbf{y}}}] \\ \text{Cov}[X_2, Y_1] & \text{Cov}[X_2, Y_2] & \dots & \text{Cov}[X_2, Y_{K_{\mathbf{y}}}] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_{K_{\mathbf{x}}}, Y_1] & \text{Cov}[X_{K_{\mathbf{x}}}, Y_2] & \dots & \text{Cov}[X_{K_{\mathbf{x}}}, Y_{K_{\mathbf{y}}}] \end{bmatrix} \end{aligned}$$

Properties of cross-covariance matrices

- Like other expressions involving second-order moments, the one defining the cross-covariance matrix can be simplified.

$$\begin{aligned}\text{Cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E} \left[(\mathbf{x} - \mathbb{E}[\mathbf{x}]) (\mathbf{y} - \mathbb{E}[\mathbf{y}])^T \right] \\ &= \mathbb{E} \left[\mathbf{x} \mathbf{y}^T \right] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}]^T \\ &\quad - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}]^T + \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}]^T \\ &= \mathbb{E} \left[\mathbf{x} \mathbf{y}^T \right] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}]^T\end{aligned}$$

- If \mathbf{x} and \mathbf{y} are independent, the cross-covariance matrix is a collection of zeroes, since $\mathbb{E} \left[\mathbf{x} \mathbf{y}^T \right] = \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}]^T$.
- The cross-covariance matrix between \mathbf{x} and \mathbf{y} relates with the respective variance-covariance matrices as follows:

$$\text{Var}[\mathbf{x}] - \text{Cov}[\mathbf{x}, \mathbf{y}] [\text{Var}[\mathbf{y}]]^{-1} \text{Cov}[\mathbf{x}, \mathbf{y}]^T \geq \mathbf{0}$$

meaning that the left-hand side is positive semi-definite.

Cross-covariance matrices of transformation

Sometimes one is interested in the cross-covariance matrices of transformed random vectors.

For **linear** transformations, if $\mathbf{u} = \mathbf{a}_x + \mathbf{B}_x \mathbf{x}$ and $\mathbf{v} = \mathbf{a}_y + \mathbf{B}_y \mathbf{y}$ are obtained from \mathbf{x} and \mathbf{y} respectively and have length J_u and J_v , their $J_u \times J_v$ cross-covariance matrix is as follows.

$$\begin{aligned}\text{Cov}[\mathbf{u}, \mathbf{v}] &= \text{Cov}[\mathbf{a}_x + \mathbf{B}_x \mathbf{x}, \mathbf{a}_y + \mathbf{B}_y \mathbf{y}] \\ &= \mathbf{B}_x \text{Cov}[\mathbf{x}, \mathbf{y}] \mathbf{B}_y^T\end{aligned}$$

At the same time, if $\mathbf{u} = \mathbf{g}_x(\mathbf{x})$ and $\mathbf{v} = \mathbf{g}_y(\mathbf{y})$ are the result of some **non-linears** transformations, the following approximation applies.

$$\text{Cov}[\mathbf{u}, \mathbf{v}] \approx \left[\frac{\partial}{\partial \mathbf{x}^T} \mathbf{g}_x(\mathbb{E}[\mathbf{x}]) \right] \text{Cov}[\mathbf{x}, \mathbf{y}] \left[\frac{\partial}{\partial \mathbf{y}^T} \mathbf{g}_y(\mathbb{E}[\mathbf{y}]) \right]^T$$

Multivariate moment generation

Definition 14

Moment generating function (multivariate). Given a random vector $\mathbf{x} = (X_1, \dots, X_K)$ with support \mathbb{X} , the *moment-generating function* $M_{\mathbf{x}}(\mathbf{t})$ is given by the expectation of the transformation $g(X) = \exp(\mathbf{x}^T \mathbf{t})$, for $\mathbf{t} = (t_1, \dots, t_K) \in \mathbb{R}^K$.

$$M_{\mathbf{x}}(\mathbf{t}) = \mathbb{E} [\exp(\mathbf{t}^T \mathbf{x})] = \mathbb{E} \left[\exp \left(\sum_{k=1}^K t_k X_k \right) \right]$$

Definition 15

Characteristic function (multivariate). Given a random vector $\mathbf{x} = (X_1, \dots, X_K)$ with support \mathbb{X} , the *characteristic function* $\varphi_{\mathbf{x}}(\mathbf{t})$ is given by the expectation of the transformation $g(X) = \exp(i\mathbf{t}^T \mathbf{x})$, for $\mathbf{t} = (t_1, \dots, t_K) \in \mathbb{R}^K$.

$$\varphi_{\mathbf{x}}(\mathbf{t}) = \mathbb{E} [\exp(i\mathbf{t}^T \mathbf{x})] = \mathbb{E} \left[\exp \left(i \sum_{k=1}^K t_k X_k \right) \right]$$

Generating multivariate moments

The r -th centered moments for each k -th element of the random vector \mathbf{x} can be calculated in analogy with the univariate case.

$$\mathbb{E} [X_k^r] = \left. \frac{\partial^r M_{\mathbf{x}}(\mathbf{t})}{\partial t_k^r} \right|_{\mathbf{t}=\mathbf{0}} = \frac{1}{i^r} \cdot \left. \frac{\partial^r \varphi_{\mathbf{x}}(\mathbf{t})}{\partial t_k^r} \right|_{\mathbf{t}=\mathbf{0}}$$

Furthermore, the *cross-moments* are obtained, for $r, s \in \mathbb{N}$:

$$\mathbb{E} [X_k^r X_\ell^s] = \left. \frac{\partial^{r+s} M_{\mathbf{x}}(\mathbf{t})}{\partial t_k^r \partial t_\ell^s} \right|_{\mathbf{t}=\mathbf{0}} = \frac{1}{i^{r+s}} \cdot \left. \frac{\partial^{r+s} \varphi_{\mathbf{x}}(\mathbf{t})}{\partial t_k^r \partial t_\ell^s} \right|_{\mathbf{t}=\mathbf{0}}$$

which can be shown as follows for the case of m.g.f.s.

$$\begin{aligned} \frac{\partial^{r+s} M_{\mathbf{x}}(\mathbf{t})}{\partial t_k^r \partial t_\ell^s} &= \mathbb{E} \left[\frac{\partial^{r+s}}{\partial t_k^r \partial t_\ell^s} \exp \left(\sum_{k=1}^K t_k X_k \right) \right] \\ &= \mathbb{E} \left[X_k^r X_\ell^s \exp \left(\sum_{k=1}^K t_k X_k \right) \right] \end{aligned}$$

Generating the bivariate normal covariance

The m.g.f. of the bivariate normal distribution is:

$$\begin{aligned}M_{X_1, X_2}(t_1, t_2) &= \mathbb{E}[\exp(t_1 X_1 + t_2 X_2)] \\ &= \exp\left(t_1 \mu_1 + t_2 \mu_2 + \frac{1}{2} \left(t_1^2 \sigma_1^2 + 2t_1 t_2 \rho \sigma_1 \sigma_2 + t_2^2 \sigma_2^2\right)\right)\end{aligned}$$

to be proven later. The first moment of $X_1 X_2$ is obtained from:

$$\begin{aligned}\frac{\partial^2}{\partial t_1 \partial t_2} M_{X_1, X_2}(t_1, t_2) &= \\ &= \left[\left(\mu_1 + t_1 \sigma_1^2 + t_2 \rho \sigma_1 \sigma_2 \right) \left(\mu_2 + t_2 \sigma_2^2 + t_1 \rho \sigma_1 \sigma_2 \right) + \rho \sigma_1 \sigma_2 \right] \times \\ &\quad \times \exp\left(t_1 \mu_1 + t_2 \mu_2 + \frac{1}{2} \left(t_1^2 \sigma_1^2 + 2t_1 t_2 \rho \sigma_1 \sigma_2 + t_2^2 \sigma_2^2\right)\right)\end{aligned}$$

and by evaluating the above expression for $t_1 = 0$ and $t_2 = 0$.

$$\mathbb{E}[X_1 X_2] = \frac{\partial^2}{\partial t_1 \partial t_2} M_{X_1, X_2}(t_1, t_2) \Big|_{t_1, t_2=0} = \rho \sigma_1 \sigma_2 + \mu_1 \mu_2$$

Moment generation and independence (1/2)

Theorem 6

Moment generating functions and characteristic functions of independent random variables. *If the random variables belonging to a random vector $\mathbf{x} = (X_1, \dots, X_K)$ are mutually independent, the moment generating function of \mathbf{x} (if it exists) and the characteristic function of \mathbf{x} equal the product of the K moment generating functions (if they exist) and the K characteristic functions of the K random variables involved, respectively.*

$$M_{\mathbf{x}}(\mathbf{t}) = \prod_{k=1}^K M_{X_k}(t_k)$$

$$\varphi_{\mathbf{x}}(\mathbf{t}) = \prod_{k=1}^K \varphi_{X_k}(t_k)$$

Moment generation and independence (2/2)

Proof.

This is an application Theorem 3 and Theorem 5 upon a sequence of K transformed random variables: $\exp(t_1 X_1), \dots, \exp(t_K X_K)$ which are themselves mutually independent. For m.g.f.s:

$$\begin{aligned} M_{\mathbf{x}}(\mathbf{t}) &= \mathbb{E}[\exp(\mathbf{t}^T \mathbf{x})] = \mathbb{E}\left[\exp\left(\sum_{k=1}^K t_k X_k\right)\right] \\ &= \mathbb{E}\left[\prod_{k=1}^K \exp(t_k X_k)\right] \\ &= \prod_{k=1}^K \mathbb{E}[\exp(t_k X_k)] \\ &= \prod_{k=1}^K M_{X_k}(t_k) \end{aligned}$$

and the case of characteristic functions is analogous. □

Moment generation, sums, independence (1/2)

Theorem 7

Moment generating and characteristic functions of linear combinations of independent random variables. *Consider a random variable Y obtained as the sum of N linearly transformed, mutually independent random variables $\mathbf{x} = (X_1, \dots, X_N)$:*

$$Y = \sum_{i=1}^N (a_i + b_i X_i)$$

where $(a_i, b_i) \in \mathbb{R}^2$ for $i = 1, \dots, N$. The moment generating and characteristic functions of Y are obtained as follows.

$$M_Y(t) = \exp\left(t \sum_{i=1}^N a_i\right) \prod_{i=1}^N M_{X_i}(b_i t)$$
$$\varphi_Y(t) = \exp\left(t \sum_{i=1}^N a_i\right) \prod_{i=1}^N \varphi_{X_i}(b_i t)$$

Moment generation, sums, independence (2/2)

Proof.

For moment generating functions this result is obtained as:

$$\begin{aligned}M_Y(t) &= \mathbb{E}[\exp(tY)] = \mathbb{E}\left[\exp\left(t \cdot \sum_{i=1}^N (a_i + b_i X_i)\right)\right] \\&= \exp\left(t \sum_{i=1}^N a_i\right) \mathbb{E}\left[\exp\left(\sum_{i=1}^N t b_i X_i\right)\right] \\&= \exp\left(t \sum_{i=1}^N a_i\right) \mathbb{E}\left[\prod_{i=1}^N \exp(t b_i X_i)\right] \\&= \exp\left(t \sum_{i=1}^N a_i\right) \prod_{i=1}^N \mathbb{E}[\exp(t b_i X_i)] \\&= \exp\left(t \sum_{i=1}^N a_i\right) \prod_{i=1}^N M_{X_i}(b_i t)\end{aligned}$$

where the second-to-last line exploits mutual independence. The case of characteristic functions is analogous. \square

Sums of independent random variables (1/6)

This result is extremely powerful: it allows to quickly derive the **distribution of a sum of independent random variables** in a number of cases. Consider the following one.

Observation 5

If all the N random variables in the vector (X_1, \dots, X_N) are pairwise independent and $X_i \sim \text{Be}(p)$ for $i = 1, \dots, N$ then:

$$\sum_{i=1}^N X_i \sim \text{BN}(p, N).$$

Proof.

If $M_{X_i}(t) = p \exp(t) + (1 - p)$, it suffices to multiply the N identical moment generating functions:

$$M_{\sum_{i=1}^N X_i}(t) = [p \exp(t) + (1 - p)]^N$$

which implies the statement by Theorem 7. □

Sums of independent random variables (2/6)

The following observations all assume that the random variables in (X_1, \dots, X_N) are pairwise independent. Proofs are omitted.

Observation 6

If $X_i \sim \text{NB}(p, 1)$, it is $\sum_{i=1}^N X_i \sim \text{NB}(p, N)$.

Observation 7

If $X_i \sim \text{Pois}(\lambda)$, it is $\sum_{i=1}^N X_i \sim \text{Pois}(N\lambda)$.

Observation 8

If $X_i \sim \text{Exp}(\lambda)$, it is $\sum_{i=1}^N X_i \sim \Gamma(N, \lambda^{-1})$.

Observation 9

If $X_i \sim \chi^2(\kappa_i)$, it is $\sum_{i=1}^N X_i \sim \chi^2\left(\sum_{i=1}^N \kappa_i\right)$.

Observation 10

If $X_i \sim \Gamma(\alpha_i, \beta)$, it is $\sum_{i=1}^N X_i \sim \Gamma\left(\sum_{i=1}^N \alpha_i, \beta\right)$.

Sums of independent random variables (3/6)

Observation 11

If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for all real a_i, b_i it is as follows.

$$Y = \sum_{i=1}^N (a_i + b_i X_i) \sim \mathcal{N}\left(\sum_{i=1}^N (a_i + b_i \mu_i), \sum_{i=1}^N b_i^2 \sigma_i^2\right)$$

Proof.

This requires few steps:

$$\begin{aligned} M_Y(t) &= \exp\left(t \sum_{i=1}^K a_i\right) \prod_{i=1}^K M_{X_i}(b_i t) \\ &= \exp\left(t \sum_{i=1}^K a_i\right) \exp\left(t \sum_{i=1}^K b_i \mu_i + t^2 \sum_{i=1}^K \frac{b_i^2 \sigma_i^2}{2}\right) \\ &= \exp\left(t \sum_{i=1}^K (a_i + b_i \mu_i) + t^2 \sum_{i=1}^K \frac{b_i^2 \sigma_i^2}{2}\right) \end{aligned}$$

the third line expresses a familiar m.g.f. for Y . □

Sums of independent random variables (4/6)

- Note: the previous observations states that a linear combination of independent normal random variables is itself normal.
- Later, this result is generalized while relaxing the independence requirement.
- This result also easily extends to lognormal distributions, but in its own way.

Observation 12

If $\log(X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for all real a_i, b_i it is as follows.

$$\log\left(\prod_{i=1}^N \exp(a_i) X_i^{b_i}\right) \sim \mathcal{N}\left(\sum_{i=1}^N (a_i + b_i \mu_i), \sum_{i=1}^N b_i^2 \sigma_i^2\right)$$

Proof.

Since $\log\left(\prod_{i=1}^N \exp(a_i) X_i^{b_i}\right) = \sum_{i=1}^N [a_i + b_i \log(X_i)]$, the previous observation extends easily. \square

Sums of independent random variables (5/6)

Observation 13

If $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ are independent, the random variable $Y = X_1/\lambda_1 - X_2/\lambda_2$ is such that $Y \sim \text{Laplace}(0, 1)$.

Proof.

Define the two random variables $W_1 = X_1/\lambda_1$ and $W_2 = -X_2/\lambda_2$, which obviously are independent. By the properties of m.g.f.s for linear transformations, the m.g.f. of the two transformed random variables is:

$$M_{W_1}(t) = (1 - t)^{-1}$$

$$M_{W_2}(t) = (1 + t)^{-1}$$

and since $Y = W_1 + W_2$, the moment generating function of Y is:

$$M_Y(t) = M_{W_1}(t) M_{W_2}(t) = (1 - t^2)^{-1}$$

that is, that of a standard Laplace distribution. □

Sums of independent random variables (6/6)

Observation 14

If $X_1 \sim \text{Gumbel}(\mu_1, \sigma)$ and $X_2 \sim \text{Gumbel}(\mu_2, \sigma)$ are independent, the random variable $Y = X_1 - X_2$ is such that $Y \sim \text{Logistic}(\mu_1 - \mu_2, \sigma)$.

Proof.

The m.g.f. of X_i (for $i = 1, 2$) is $M_{X_i}(t) = \exp(\mu_i t) \Gamma(1 - \sigma t)$. Thus, the transformed random variables $W_i = -X_i$ (for $i = 1, 2$) have m.g.f.

$$M_{W_i}(t) = \exp(-\mu_i t) \Gamma(1 + \sigma t)$$

while X_1 is independent of W_2 and vice versa. Since $Y = X_1 + W_2$:

$$\begin{aligned} M_Y(t) &= M_{X_1}(t) M_{W_2}(t) \\ &= \exp(\mu_1 t) \Gamma(1 - \sigma t) \cdot \exp(-\mu_2 t) \Gamma(1 + \sigma t) \\ &= \exp(\mu_1 t - \mu_2 t) \frac{\Gamma(1 - \sigma t) \Gamma(1 + \sigma t)}{\Gamma(2)} \\ &= \exp((\mu_1 - \mu_2) t) \cdot \text{B}(1 - \sigma t, 1 + \sigma t) \end{aligned}$$

i.e. the logistic's m.g.f. sought after (note that $\Gamma(2) = 1! = 1$). □

Fixing realizations

- It is often useful to analyze certain random variables of a random vector when the realizations of the other random variables are held constant, or “fixed.”
- This leads to the analysis of **conditional distributions** and **conditional moments**.
- One can also “fix” subsets of the support – not just single realizations; this case is left aside for the moment.
- This discussion is based on two random vectors \mathbf{x} and \mathbf{y} of dimension $K_{\mathbf{x}} \geq 1$ & $K_{\mathbf{y}} \geq 1$ (with possibly $K_{\mathbf{x}} \neq K_{\mathbf{y}}$) and supports \mathbb{X} and \mathbb{Y} respectively, expressed as follows.

$$\mathbf{x} = \left(X_1 \quad \dots \quad X_{K_{\mathbf{x}}} \right)^T$$

$$\mathbf{y} = \left(Y_1 \quad \dots \quad Y_{K_{\mathbf{y}}} \right)^T$$

Conditional mass or density

- In what follows, assume that \mathbf{x} contains either only discrete r.v.s, or only continuous ones, but not both. Same with \mathbf{y} .
- Definitions of joint mass/density would adjust accordingly.

Definition 16

Conditional mass or density function. Consider the combined random vector (\mathbf{x}, \mathbf{y}) with joint mass/density function $f_{\mathbf{x}, \mathbf{y}}(\mathbf{x}, \mathbf{y})$. Suppose that the random vectors \mathbf{x} has a probability mass or density function $f_{\mathbf{x}}(\mathbf{x})$. The *conditional* mass or density function of \mathbf{y} , *given* $\mathbf{x} = \mathbf{x}$, is defined as follows for all $\mathbf{x} \in \mathbb{X}$:

$$f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x} = \mathbf{x}) = \frac{f_{\mathbf{x}, \mathbf{y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{x}}(\mathbf{x})}$$

It is a conditional mass function if all the random variables in \mathbf{y} are discrete, and a conditional density function if they are all continuous.

Example: conditional normal distribution

Return to the bivariate normal distribution. Whenever one fixes any $X_2 = x_2 \in \mathbb{R}$, the resulting conditional p.d.f. is:

$$\begin{aligned} f_{X_1|X_2}(x_1|X_2 = x_2) &= \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{\left[x_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)\right]^2}{2\sigma_1^2(1-\rho^2)}\right) \end{aligned}$$

so that we write the resulting conditional *distribution* as:

$$X_1|X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

and this is symmetrical if one fixes any $X_1 = x_1 \in \mathbb{R}$ instead.

Conditional cumulative distribution

Definition 17

Conditional cumulative distribution. Consider the combined random vector (\mathbf{x}, \mathbf{y}) with joint mass/density function $f_{\mathbf{x}, \mathbf{y}}(\mathbf{x}, \mathbf{y})$. The *conditional* cumulative distribution of \mathbf{y} , given $\mathbf{x} = \mathbf{x}$ is defined as:

$$F_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x} = \mathbf{x}) = \sum_{\mathbf{t} \in \mathbb{Y}: \mathbf{t} \leq \mathbf{y}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{t}|\mathbf{x} = \mathbf{x})$$

if all the random variables in \mathbf{y} are discrete, and

$$F_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x} = \mathbf{x}) = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_{K_{\mathbf{y}}}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{t}|\mathbf{x} = \mathbf{x}) dt$$

if all the random variables in \mathbf{y} are continuous.

- If \mathbf{x} is indeterminate/unspecified, the short notation $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and $F_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ can also be used.
- In the previous example about the normal distribution, one may denote the conditional p.d.f. as $f_{X_1|X_2}(x_1|x_2)$.

Interpreting conditional distributions

- Adapting the general definitions of (joint) p.m.f. and c.d.f., a **conditional** p.m.f. or c.d.f. has a direct interpretation in terms of **conditional probability**.
- For conditional p.d.f.s, the interpretation holds for specified **intervals** of the non-fixed vector.

$$\begin{aligned}\mathbb{P}(a_1 \leq Y_1 \leq b_1 \cap \dots \cap a_{K_y} \leq Y_{K_y} \leq b_{K_y} | X_1 = x_1 \cap \dots \cap X_{K_x} = x_{K_x}) \\ = \int_{a_1}^{b_1} \dots \int_{a_{K_y}}^{b_{K_y}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y} | \mathbf{x} = \mathbf{x}) d\mathbf{y}\end{aligned}$$

- If \mathbf{x} is discrete whereas \mathbf{y} is continuous – or vice versa – the interpretation must (intuitively) adjust accordingly.
- Recall the example about height across genders. Describing the height of one gender only is an exercise in conditioning.

$$f_{H|G=1}(h | g = 1) = \frac{1}{\sigma_F} \phi\left(\frac{h - \mu_F}{\sigma_F}\right)$$

Conditional distributions and independence

- All these definitions are somehow “moot” for independent random variables/vectors. If \mathbf{x} and \mathbf{y} are independent:

$$f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{y}}(\mathbf{y})$$

and thus the following holds.

$$f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = f_{\mathbf{y}}(\mathbf{y})$$

$$f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = f_{\mathbf{x}}(\mathbf{x})$$

- In words, the conditional distribution of \mathbf{y} *given* \mathbf{x} is equal to the *unconditional* distribution of \mathbf{y} , and vice versa.
- This is straightforward given the interpretations in terms of conditional probability.

Conditional moments: expectation

- Moments are easily defined on conditional distributions too.
- The **conditional expectation**, also called **regression**, is defined for discrete random variables as:

$$\mathbb{E}[\mathbf{y}|\mathbf{x}] = \sum_{y_1 \in \mathbb{Y}_1} \cdots \sum_{y_{K_{\mathbf{y}}} \in \mathbb{Y}_{K_{\mathbf{y}}}} \mathbf{y} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$$

and for continuous random variables as follows.

$$\mathbb{E}[\mathbf{y}|\mathbf{x}] = \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_{K_{\mathbf{y}}}} \mathbf{y} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

- Note: it is a scalar if $K_{\mathbf{y}} = 1$, a vector if $K_{\mathbf{y}} > 1$.
- It represents the mean of \mathbf{y} when $\mathbf{x} = \mathbf{x}$ is “fixed.”

Conditional moments: variance-covariance (1/2)

- The **conditional variance** is defined in the discrete case as:

$$\text{Var}[\mathbf{y}|\mathbf{x}] = \sum_{y_1 \in \mathbb{Y}_1} \cdots \sum_{y_{K_y} \in \mathbb{Y}_{K_y}} (\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}]) (\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}])^T \times \\ \times f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$$

and in the continuous case as follows.

$$\text{Var}[\mathbf{y}|\mathbf{x}] = \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_{K_y}} (\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}]) (\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}])^T \times \\ \times f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

- Note: it is a scalar if $K_y = 1$, a square matrix if $K_y > 1$. In the latter case it is typically more appropriate to call it the **conditional variance-covariance**.

Conditional moments: variance-covariance (2/2)

- Similarly with the conditional expectation, the conditional variance-covariance denotes the dispersion or co-movement of/between random variables in \mathbf{y} when $\mathbf{x} = \mathbf{x}$ is “fixed.”
- The conditional variance-covariances inherit all the typical properties of their unconditional counterparts.
- For example, they can be recast as follows.

$$\begin{aligned}\text{Var} [\mathbf{y} | \mathbf{x}] &= \mathbb{E} \left[(\mathbf{y} - \mathbb{E} [\mathbf{y} | \mathbf{x}]) (\mathbf{y} - \mathbb{E} [\mathbf{y} | \mathbf{x}])^T \middle| \mathbf{x} \right] \\ &= \mathbb{E} \left[\mathbf{y} \mathbf{y}^T \middle| \mathbf{x} \right] - \mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T \\ &\quad - \mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T + \mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T \\ &= \mathbb{E} \left[\mathbf{y} \mathbf{y}^T \middle| \mathbf{x} \right] - \mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T\end{aligned}$$

More about conditional moments

- An example on conditional moments: the conditional normal distribution of $X_1|X_2 = x_2$ that is derived earlier from the bivariate normal has the following mean and variance.

$$\mathbb{E}[X_1|X_2 = x_2] = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2)$$

$$\text{Var}[X_1|X_2 = x_2] = \sigma_1^2 (1 - \rho^2)$$

- Conditional moments can be recast as specific **functions** of the conditioned random variable. In this case, the notation for random variables/vectors/matrices (as opposed to their realizations) may be used. Examples of this notation follow.

$$\mathbb{E}[X_1|X_2]$$

$$\mathbb{E}[\mathbf{y}|\mathbf{x}]$$

$$\text{Var}[X_1|X_2]$$

$$\text{Var}[\mathbf{y}|\mathbf{x}]$$

Law of Iterated Expectations

Theorem 8

Law of Iterated Expectations. *Given any two random vectors \mathbf{x} and \mathbf{y} , it is:*

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}_{\mathbf{x}}[\mathbb{E}[\mathbf{y}|\mathbf{x}]]$$

where $\mathbb{E}_{\mathbf{x}}[\cdot]$ denotes an expectation taken over the support of \mathbf{x} .

Proof.

In the continuous case, apply the following decomposition:

$$\begin{aligned}\mathbb{E}[\mathbf{y}] &= \int_{\mathbf{X}} \int_{\mathbf{Y}} \mathbf{y} f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} d\mathbf{x} \\ &= \int_{\mathbf{X}} \int_{\mathbf{Y}} \mathbf{y} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{y} d\mathbf{x} \\ &= \int_{\mathbf{X}} f_{\mathbf{x}}(\mathbf{x}) \left[\int_{\mathbf{Y}} \mathbf{y} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \right] d\mathbf{x} \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}[\mathbf{y}|\mathbf{x}]]\end{aligned}$$

and the discrete case is analogous (sums substitute integrals). □

Example: linear regression (1/4)

- The **linear regression** model is a cornerstone of statistics and econometrics and it is intimately linked to conditional expectations.
- The model relates an **endogenous** or **dependent** variable Y_i to some **exogenous** or **independent** variables X_i (also denoted as Z_i).
- Here i denotes the *unit of observation* (more on this later).
- The conditional expectation function of Y_i given a sequence of X_{ki} (or Z_{ki}) for $k = 1, \dots, K$ is modeled as linear.

$$\mathbb{E}[Y_i | X_{1i}, \dots, X_{Ki}] = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki}$$

Here $(\beta_0, \beta_1, \dots, \beta_K)$ are the **parameters of interest**.

Example: linear regression (2/4)

- The simplest linear regression model is the **bivariate** one.

$$\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i$$

This is equivalent to the following expression.

$$\mathbb{E}[Y_i - \beta_0 - \beta_1 X_i | X_i] = 0$$

- Parameter β_0 is given the interpretation as the conditional mean of Y_i given $X_i = 0$, or equivalently, as the “constant” coefficient that satisfies the following relationship.

$$\mathbb{E}[Y_i - \beta_0 - \beta_1 X_i] = 0$$

- Note this application of the Law of Iterated Expectations.

$$\begin{aligned}\mathbb{E}[X_i (Y_i - \beta_0 - \beta_1 X_i)] &= \mathbb{E}_X [\mathbb{E}[X_i (Y_i - \beta_0 - \beta_1 X_i) | X_i]] \\ &= \mathbb{E}_X [X_i \cdot \mathbb{E}[(Y_i - \beta_0 - \beta_1 X_i) | X_i]] \\ &= 0\end{aligned}$$

Example: linear regression (3/4)

- One can thus put together two equations for two unknowns.

$$\mathbb{E}[Y_i - \beta_0 - \beta_1 X_i] = 0$$

$$\mathbb{E}[X_i (Y_i - \beta_0 - \beta_1 X_i)] = 0$$

- After some manipulation, the solution for β_0 and β_1 is:

$$\beta_0 = \mathbb{E}[Y_i] - \frac{\text{Cov}[X_i, Y_i]}{\text{Var}[X_i]} \cdot \mathbb{E}[X_i]$$

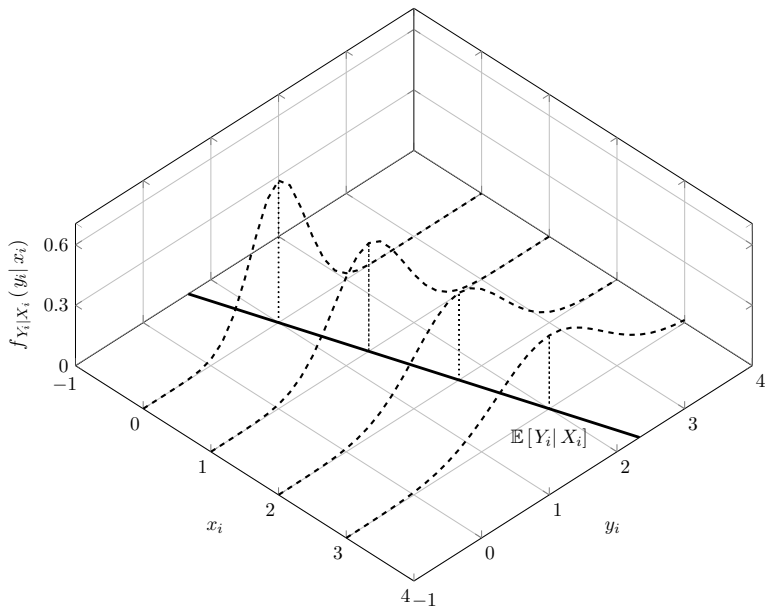
$$\beta_1 = \frac{\text{Cov}[X_i, Y_i]}{\text{Var}[X_i]}$$

although β_0 is commonly written as $\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$.

- Parameter β_1 is named the **regression slope**; it expresses the average response of Y_i to X_i , and it is closely related to a measure of *correlation*.

$$\beta_1 = \text{Corr}[X_i, Y_i] \cdot \frac{\sqrt{\text{Var}[Y_i]}}{\sqrt{\text{Var}[X_i]}}$$

Example: linear regression (4/4)



Law of Total Variance

Theorem 9

Law of Total Variance (variance decomposition). *Given any two random vectors \mathbf{x} and \mathbf{y} :*

$$\text{Var} [\mathbf{y}] = \text{Var}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]] + \mathbb{E}_{\mathbf{x}} [\text{Var} [\mathbf{y} | \mathbf{x}]]$$

where $\text{Var}_{\mathbf{x}} [\cdot]$ and $\mathbb{E}_{\mathbf{x}} [\cdot]$ denote sums/integrals taken over the support of \mathbf{x} for every element of the argument vectors/matrices.

Proof.

$$\begin{aligned} \text{Var} [\mathbf{y}] &= \mathbb{E} [\mathbf{y}\mathbf{y}^T] - \mathbb{E} [\mathbf{y}] \mathbb{E} [\mathbf{y}]^T \\ &= \mathbb{E}_{\mathbf{x}} [\mathbb{E} [\mathbf{y}\mathbf{y}^T | \mathbf{x}] - \mathbb{E}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]] [\mathbb{E}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]]]^T] \\ &= \mathbb{E}_{\mathbf{x}} \left[\mathbb{E} [\mathbf{y}\mathbf{y}^T | \mathbf{x}] - \mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T \middle| \mathbf{x} \right] \\ &\quad + \mathbb{E}_{\mathbf{x}} \left[\mathbb{E} [\mathbf{y} | \mathbf{x}] \mathbb{E} [\mathbf{y} | \mathbf{x}]^T \middle| \mathbf{x} \right] - \mathbb{E}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]] [\mathbb{E}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]]]^T \\ &= \mathbb{E}_{\mathbf{x}} [\text{Var} [\mathbf{y} | \mathbf{x}]] + \text{Var}_{\mathbf{x}} [\mathbb{E} [\mathbf{y} | \mathbf{x}]] \end{aligned}$$

Note the various applications of the Law of Iterated Expectations. \square

Example: income groups (1/4)

- Here is an application for the Law of Total Variance.
- Let $X = 1, 2, 3, 4$ be a **discrete** random variable whose role is to identify *groups* of a population (e.g. ethnicities).
- Let Y be a **continuous** random variable (e.g. log-income) that is distributed over the population in question.
- Assume the following conditional distributions.

$$Y|X = 1 \sim \mathcal{N}(3, 1.5)$$

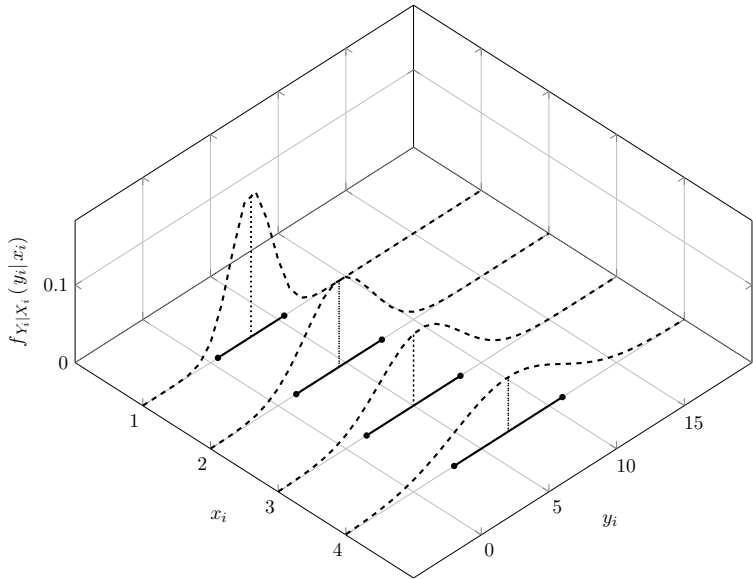
$$Y|X = 2 \sim \mathcal{N}(4.5, 2.5)$$

$$Y|X = 3 \sim \mathcal{N}(5.5, 3)$$

$$Y|X = 4 \sim \mathcal{N}(7, 4)$$

- Groups have equal size: $\mathbb{P}(X = x) = 0.25$ for $x = 1, 2, 3, 4$.

Example: income groups (2/4)



Example: income groups (3/4)

- By the Law of Iterated Expectations, $\mathbb{E}[Y]$ is as follows.

$$\mathbb{E}[Y] = \mathbb{E}_X [\mathbb{E}[Y|X]] = \frac{1}{4} \sum_{x=1}^4 \mathbb{E}[Y|X=x] = 5$$

- Let interest be on the analysis of income inequality $\text{Var}[Y]$.
- The **between group variation** $\text{Var}_X [\mathbb{E}[Y|X]]$ is:

$$\begin{aligned} \text{Var}_X [\mathbb{E}[Y|X]] &= \frac{1}{4} \sum_{x=1}^4 (\mathbb{E}[Y|X=x] - \mathbb{E}_X [\mathbb{E}[Y|X]])^2 \\ &= 2.125 \end{aligned}$$

- ...here, this is interpreted as the dispersion of the average (log-)income across the population's groups.

Example: income groups (4/4)

- The **within group variation** $\mathbb{E}_X [\text{Var} [Y | X]]$ is:

$$\mathbb{E}_X [\text{Var} [Y | X]] = \frac{1}{4} \sum_{x=1}^4 \text{Var} [Y | X = x] = 2.75$$

- ... which instead is interpreted as the average dispersion of (log-)income as calculated separately for each group.
- By the Law of Total Variance, here it is:

$$\text{Var} [Y] = \text{Var}_X [\mathbb{E} [Y | X]] + \mathbb{E}_X [\text{Var} [Y | X]] = 4.875$$

- ... hence, the two components of inequality have about the same effect on the overall income inequality.

Two multivariate distributions

- This lecture concludes with the analysis of two important multivariate distributions: one **discrete**, one **continuous**.
- The discrete distribution is the **multinomial** distribution: it generalizes the binomial distribution to allow for multiple outcomes of a trial.
- The continuous distribution is the **multivariate normal** distribution: it extends the bivariate normal to allow for a random vector of length $K \geq 2$ collecting random variables that are normally distributed and possibly correlated.
- The analysis proceeds along the lines of Lecture 2: for both analyzed distributions, their support, parameters, p.m.f. or p.d.f., m.g.f. et cetera are specified.

The multinomial distribution (1/4)

- Consider a variation of the binomial experiment: there are n trials but these are not Bernoulli.
- Instead, any trial can return one out of K alternatives (e.g. colors of the balls in an urn with replacement), with $K \geq 2$.
- Each alternative has probability $p_k \in [0, 1]$ in each trial (for $k = 1, \dots, K$), with $\sum_{k=1}^K p_k = 1$.
- The experiment delivers a list of **success counts** for every alternative, which can be written as $\mathbf{x} = (X_1, \dots, X_K)$.
- Given that $X_k \in \{0, 1, \dots, n\}$ for $k = 1, \dots, K$, the support of this random vector is the following set.

$$\mathbb{X} = \left\{ \mathbf{x} = (x_1, \dots, x_K) \in \{0, 1, \dots, n\}^n : \sum_{k=1}^K x_k = n \right\}$$

The multinomial distribution (2/4)

- The joint p.m.f. of the multinomial distribution is:

$$f_{\mathbf{x}}(x_1, \dots, x_K; n, p_1, \dots, p_K) = \frac{n!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k}$$

where the *multinomial coefficient* $n! \cdot \prod_{k=1}^K (x_k!)^{-1}$ counts the number of realizations containing exactly (x_1, \dots, x_K) successes for each alternative out of n draws.

- The joint c.d.f. sums the mass function over points in the support as follows, for $\mathbf{t} = (t_1, \dots, t_K)$.

$$F_{\mathbf{x}}(x_1, \dots, x_K; n, p_1, \dots, p_K) = \sum_{\mathbf{t} \in \mathbb{X}: \mathbf{t} \leq \mathbf{x}} \frac{n!}{\prod_{k=1}^K t_k!} \prod_{k=1}^K p_k^{t_k}$$

The multinomial distribution (3/4)

- The distribution owes its name to the *multinomial theorem*, which helps show that the total probability mass equals 1.

$$\mathbb{P}(\mathbf{x} \in \mathbb{X}) = \sum_{\mathbf{x} \in \mathbb{X}} \frac{n!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k} = \left(\sum_{k=1}^K p_k \right)^n = 1$$

- This helps calculate the m.g.f. too.

$$\begin{aligned} M_{\mathbf{x}}(t_1, \dots, t_K) &= \sum_{\mathbf{x} \in \mathbb{X}} \exp\left(\sum_{k=1}^K t_k x_k\right) \frac{n!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k} \\ &= \sum_{\mathbf{x} \in \mathbb{X}} \frac{n!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K (p_k \cdot \exp(t_k))^{x_k} \\ &= \left(\sum_{k=1}^K p_k \cdot \exp(t_k) \right)^n \end{aligned}$$

The multinomial distribution (4/4)

- Through the m.g.f. one can show that for all $k = 1, \dots, K$:

$$\mathbb{E}[X_k] = np_k$$

$$\text{Var}[X_k] = np_k(1 - p_k)$$

and for all $k, \ell = 1, \dots, K$:

$$\text{Cov}[X_k, X_\ell] = -np_k p_\ell$$

and the covariance is always negative because an increasing number of successes for one alternative implies a decreasing number for another alternative.

- These moments can be expressed in **compact notation** as:

$$\mathbb{E}[\mathbf{x}] = n\mathbf{p}$$

$$\text{Var}[\mathbf{x}] = n \left(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T \right)$$

where $\mathbf{p} \equiv (p_1 \ p_2 \ \dots \ p_K)^T$.

The multivariate normal distribution (1/8)

- A random vector $\mathbf{x} = (X_1, \dots, X_K)$ of length K follows the **multivariate** normal distribution with support $\mathbb{X} = \mathbb{R}^K$ if its joint probability density function is as follows.

$$f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- The distribution's **parameters** are collected by a vector $\boldsymbol{\mu}$ of length K and a symmetric, positive semi-definite matrix $\boldsymbol{\Sigma}$ having size $K \times K$ and full rank:

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_{KK} \end{pmatrix}$$

with $\sigma_{ij} = \sigma_{ji}$ for $i, j = 1, \dots, K$ and where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

The multivariate normal distribution (2/8)

- The notation expressing that \mathbf{x} follows a multivariate normal distribution with given parameters is as follows.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- A special case is the *standardized* multivariate normal, with $\boldsymbol{\mu} = \mathbf{0}$ & $\boldsymbol{\Sigma} = \mathbf{I}$. If a random vector \mathbf{z} follows the standard multivariate normal distribution, this is written as follows.

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Because $\boldsymbol{\Sigma}$ is a symmetric and positive semi-definite matrix, eigendecomposing it returns a matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}$ with properties:

$$\boldsymbol{\Sigma}^{\frac{1}{2}} \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \right)^{\text{T}} = \boldsymbol{\Sigma} \quad \text{and} \quad \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} \right)^{\text{T}} = \mathbf{I}.$$

- Hence, \mathbf{x} and \mathbf{z} are related via mutual transformations.

$$\mathbf{z} = \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{x} - \boldsymbol{\mu})$$

$$\mathbf{x} = \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{z} + \boldsymbol{\mu}$$

The multivariate normal distribution (3/8)

- The c.d.f. obtains from integrating the p.d.f.: similarly as in the univariate case, it has no closed form solution. Showing that the total p.d.f. integrates to 1 is tedious here too.

$$F_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_K} f_{\mathbf{x}}(t; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dt$$

- However, obtaining the m.g.f. is relatively easy if one starts from the standardized case $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

$$\begin{aligned} M_{\mathbf{z}}(\mathbf{t}) &= \int_{\mathbb{R}^K} \exp(\mathbf{t}^T \mathbf{z}) \frac{1}{\sqrt{(2\pi)^K}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z}\right) d\mathbf{z} \\ &= \exp\left(\frac{\mathbf{t}^T \mathbf{t}}{2}\right) \int_{\mathbb{R}^K} \frac{1}{\sqrt{(2\pi)^K}} \exp\left(-\frac{(\mathbf{z} - \mathbf{t})^T (\mathbf{z} - \mathbf{t})}{2}\right) d\mathbf{z} \\ &= \exp\left(\frac{\mathbf{t}^T \mathbf{t}}{2}\right) \end{aligned}$$

The multivariate normal distribution (4/8)

- The general version of the m.g.f. is then:

$$\begin{aligned}M_{\mathbf{x}}(\mathbf{t}) &= \mathbb{E} \left[\exp \left(\mathbf{t}^T \mathbf{x} \right) \right] \\&= \mathbb{E} \left[\exp \left(\mathbf{t}^T \left(\Sigma^{\frac{1}{2}} \mathbf{z} + \boldsymbol{\mu} \right) \right) \right] \\&= \exp \left(\mathbf{t}^T \boldsymbol{\mu} \right) \cdot \mathbb{E} \left[\exp \left(\mathbf{t}^T \Sigma^{\frac{1}{2}} \mathbf{z} \right) \right] \\&= \exp \left(\mathbf{t}^T \boldsymbol{\mu} + \frac{\mathbf{t}^T \Sigma \mathbf{t}}{2} \right)\end{aligned}$$

because if $\mathbf{d} = \left(\Sigma^{\frac{1}{2}} \right)^T \mathbf{t}$, it is $\mathbf{d}^T \mathbf{d} = \mathbf{t}^T \Sigma \mathbf{t}$.

- This allows to derive all key moments as:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \\ \text{Var}[\mathbf{x}] &= \Sigma\end{aligned}$$

and the covariances lie in the off-diagonal elements of Σ .

The multivariate normal distribution (5/8)

- If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, what is the distribution of $\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}$?
Note that the transformation \mathbf{y} may have length $J \neq K$.
- The m.g.f. of \mathbf{y} is:

$$\begin{aligned}M_{\mathbf{y}}(\mathbf{t}) &= \mathbb{E} \left[\exp \left(\mathbf{t}^T \mathbf{y} \right) \right] \\&= \mathbb{E} \left[\exp \left(\mathbf{t}^T (\mathbf{a} + \mathbf{B}\mathbf{x}) \right) \right] \\&= \exp \left(\mathbf{t}^T \mathbf{a} \right) \cdot \mathbb{E} \left[\exp \left(\mathbf{t}^T \mathbf{B}\mathbf{x} \right) \right] \\&= \exp \left(\mathbf{t}^T (\mathbf{B}\boldsymbol{\mu} + \mathbf{a}) + \frac{\mathbf{t}^T \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T \mathbf{t}}{2} \right)\end{aligned}$$

because if $\mathbf{b} = \mathbf{B}^T \mathbf{t}$, it is $\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b} = \mathbf{t}^T \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T \mathbf{t}$.

- Therefore:

$$\mathbf{y} \sim \mathcal{N} \left(\mathbf{B}\boldsymbol{\mu} + \mathbf{a}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T \right)$$

even if the random variables in \mathbf{x} are dependent.

The multivariate normal distribution (6/8)

- Suppose that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ can be split into two subvectors \mathbf{x}_1 and \mathbf{x}_2 of length K_1 and K_2 respectively; $K_1 + K_2 = K$. What are the marginal, conditional distributions of \mathbf{x}_1 , \mathbf{x}_2 ?
- Partition the original collection of parameters as follows:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where:

- $\boldsymbol{\mu}_1$ is a vector of length K_1 , $\boldsymbol{\mu}_2$ one of length K_2 ;
- $\boldsymbol{\Sigma}_{11}$ is a symmetric $K_1 \times K_1$ matrix, $\boldsymbol{\Sigma}_{22}$ is a symmetric $K_2 \times K_2$ matrix;
- whereas $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ are two matrices, where one is the transpose of the other, having dimension $K_1 \times K_2$ and $K_2 \times K_1$ respectively.

The multivariate normal distribution (7/8)

- (*Algebraic digression.*) By the analysis of partitioned inverse matrices:

$$\Sigma^{-1} = \begin{pmatrix} \bar{\Sigma}_1^{-1} & -\bar{\Sigma}_1^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\bar{\Sigma}_2^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \bar{\Sigma}_2^{-1} \end{pmatrix}$$

where:

$$\bar{\Sigma}_1 \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\bar{\Sigma}_2 \equiv \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

and:

$$|\Sigma| = |\bar{\Sigma}_1| \cdot |\Sigma_{22}| = |\bar{\Sigma}_2| \cdot |\Sigma_{11}|$$

relating the determinant of Σ to those of the matrices that express its partitioned inverse.

The multivariate normal distribution (8/8)

- All this lets rewrite the p.d.f. of \mathbf{x} in (very) “long” form as:

$$f_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}) = \frac{1}{\sqrt{(2\pi)^K |\bar{\boldsymbol{\Sigma}}_1| \cdot |\boldsymbol{\Sigma}_{22}|}} \times \\ \times \exp\left(\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \bar{\boldsymbol{\Sigma}}_1^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) - \frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \bar{\boldsymbol{\Sigma}}_1^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) + \right. \\ \left. + \frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \bar{\boldsymbol{\Sigma}}_2^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \bar{\boldsymbol{\Sigma}}_2^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)\right)$$

- ...so the two “marginalized” distributions for $\mathbf{x}_1, \mathbf{x}_2$ are:

$$\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

$$\mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

- ...while the conditional ones are, reciprocally, as follows.

$$\mathbf{x}_1 | \mathbf{x}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}\right)$$

$$\mathbf{x}_2 | \mathbf{x}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}\right)$$