LECTURE 4: MULTIVARIATE RANDOM VARIABLES II

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
SEPTEMBER 20, 2023

1. Important identities

1.1. Law of iterated expectations

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

Now $\mathbb{E}[Y|X]$ is a scalar random variable, and inhabits the same probability space as X. Therefore, the outer expectation on the right-hand side is taken with respect to $f_X(x)$.

$$\begin{aligned}\mathbb{E}[Y|X=x] &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y|x)\, dy \\ &= g(x)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X]] &= \mathbb{E}[g(X)] \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy \right) f(x)\, dx\end{aligned}$$

Intuitively, suppose we use realizations of the variable X to predict Y. Then the average of the predicted values over X equals to the average of Y.

Example:

Recall the pdf $f(x,y) = x + y$ with the support on $\{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Previously, we found that:

$$\mathbb{E}[Y|X] = \frac{2+3X}{3+6X}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X]] &= \int \frac{2+3x}{3+6x} f_X(x)\,dx \\ &= \int_0^1 \frac{2+3x}{3+6x}\left(\frac{1}{2}+x\right)\,dx \\ &= \frac{1}{6}\left(\frac{3x^2}{2}+2x\right)\Big|_0^1 \\ &= \frac{7}{12} \\ &= \mathbb{E}[Y]\end{aligned}$$

1.2. Important properties of conditional expectations

This section is adapted from Chapter 2 of "Econometric Analysis of Cross Section and Panel Data" by Jeffrey M. Wooldridge.

Let Y, W be random variables. Let X be the random variable such that $X = g(W)$, for some function g.

Comparing $\mathbb{E}[Y|W]$ and $\mathbb{E}[Y|X]$, we can think of $\mathbb{E}[Y|X]$ as conditioning on a set of events that is a subset of the set of events being conditioned on in $\mathbb{E}[Y|W]$. Because if we know the outcome of W, then we would know X, but the converse is not true.

$$\mathbb{E}[\mathbb{E}[Y|X]|W] = \mathbb{E}[Y|X] \tag{1}$$

$$\mathbb{E}[\mathbb{E}[Y|W]|X] = \mathbb{E}[Y|X] \tag{2}$$

A phrase useful for remembering both equations above: "The smaller information set always dominates". This is also known as the Tower Property of conditional expectations, which can be demonstrated more formally with measure-theoretic notations.

Some consequences of this useful property:

$$\mathbb{E}[\mathbb{E}[Y|X]|X^2] = \mathbb{E}[\mathbb{E}[Y|X^2]|X] = \mathbb{E}[Y|X^2] \tag{3}$$

$$\mathbb{E}[\mathbb{E}[Y|X,Z]|X] = \mathbb{E}[\mathbb{E}[Y|X]|X,Z] = \mathbb{E}[Y|X] \tag{4}$$

1.3. Conditional variance identity

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

$\mathbb{E}[Y|X]$ and $\mathrm{Var}(Y|X)$ are each scalar random variable that is a transformation of X and has the same probability space as X. Therefore, the expectation and variance on the right-hand side is taken with respect to the pdf $f_X(x)$.

Example:

Using the same example as before, we have the pdf $f(x,y) = x+y$ with the support on $\{(x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 1\}$.

$$\begin{aligned}
\mathbb{E}[Y|X] &= \frac{2+3X}{3+6X} \\
\mathrm{Var}(\mathbb{E}[Y|X]) &= \mathbb{E}[(\mathbb{E}[Y|X])^2] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 \\
&= \int_0^1 \left(\frac{2+3x}{3+6x}\right)^2 f_X(x)\, dx - \mathbb{E}[Y]^2 \\
&= \int_0^1 \left(\frac{2+3x}{3+6x}\right)^2 \left(\frac{1}{2}+x\right) dx - \left(\frac{7}{12}\right)^2 \\
&= \frac{1}{288}(96+\log(9)) - \frac{49}{144}
\end{aligned}$$

We can derive $\mathrm{Var}(Y|X)$ by:

$$\begin{aligned}
\mathrm{Var}[Y|X=x] &= \mathbb{E}[Y^2|X=x] - (\mathbb{E}[Y|X=x])^2 \\
&= \int_0^1 y^2 f_{Y|X=x}(y|x)\, dy - (\mathbb{E}[Y|X=x])^2 \\
&= \int_0^1 y^2 \frac{2(x+y)}{1+2x}\, dy - \left(\frac{2+3x}{3+6x}\right)^2 \\
&= \frac{4x+3}{12x+6} - \left(\frac{2+3x}{3+6x}\right)^2 \\
&= \frac{1}{36}\left(3 - \frac{1}{(2x+1)^2}\right)
\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[Y|X]] &= \int_0^1 \frac{1}{36}\left(3 - \frac{1}{(2x+1)^2}\right) f_X(x)\,dx \\ &= \frac{1}{144}(12 - \log(3))\end{aligned}$$

Therefore, $\mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}(\mathbb{E}[Y|X]) = \frac{11}{144} = \mathrm{Var}(Y)$.

2. Example: putting everything together

Suppose X and Y are distributed uniformly on the triangle $(0,0), (0,1), (1,0)$. That is:

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

1.) Is this a valid pdf?

$$\int_0^1 \int_0^{1-y} 2\,dx\,dy$$

Performing the inner integral first with respect to x:

$$\begin{aligned}\int_0^1 [2x]_0^{1-y}\,dx &= \int_0^1 2(1-y)\,dy \\ &= 2\left[y - \frac{y^2}{2}\right]_0^1 = 2(1 - \frac{1}{2}) = 1\end{aligned}$$

2.) Derive the marginal pdfs.

$$f_X(x) = \int_0^{1-x} 2\,dy = 2(1-x) \ \text{ for } x \in [0,1]$$

$$f_Y(y) = \int_0^{1-y} 2\,dy = 2(1-y) \ \text{ for } y \in [0,1]$$

3.) Calculate $\mathrm{Cov}(X,Y)$

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

$$\mathbb{E}[X] = \int_0^1 2(1-x)x\,dx = \frac{1}{3}$$

$$\mathbb{E}[Y] = \int_0^1 2(1-y)y\,dy = \frac{1}{3}$$

$$\begin{aligned}
\mathbb{E}[XY] &= \int\int xyf(x,y)\,dx\,dy \\
&= \int_0^1 \int_0^{1-y} 2xy\,dx\,dy \\
&= \int_0^1 [x^2y]_0^{1-y}\,dy \\
&= \int_0^1 (1-y)^2 y\,dy \\
&= \left[\frac{y^2}{2} - \frac{2y^3}{3} + \frac{y^4}{4}\right]_0^1 \\
&= \frac{1}{12}
\end{aligned}$$

Hence $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{12} - (\frac{1}{3})(\frac{1}{3}) = -\frac{1}{36}$

4.) Calculate $P(Y \leq 1-2X)$:

$$\begin{aligned}
P(Y \leq 1-2X) &= \int_0^{1/2} \int_0^{1-2x} f(x,y)\,dy\,dx \\
&= \int_0^{1/2} 2 - 4x\,dx \\
&= \left[2x - 2x^2\right]_0^{1/2} \\
&= \frac{1}{2}
\end{aligned}$$

5.) Derive $\mathbb{E}[Y|X=x]$ and $\mathrm{Var}(Y|X=x)$:

First, the density of $Y|X=x$:

$$f_{Y|X=x}(y|x) = \frac{f(x,y)}{f(x)} = \frac{2}{2(1-x)}, \text{ for } 0 \leq x \leq 1,\ 0 \leq y \leq 1,\ x+y \leq 1$$

Conditional expectation:

$$\mathbb{E}(Y|X=x) = \int_0^{1-x} y f_{Y|X=x}(y|x)\, dy = \int_0^{1-x} \frac{y}{(1-x)}\, dy = \frac{1-x}{2}$$

Conditional variance:

$$\begin{aligned}
\text{Var}(Y|X=x) &= \mathbb{E}[Y^2|X=x] - \mathbb{E}[Y|X=x]^2 \\
&= \int_0^{1-x} y^2 f_{Y|X=x}(y|x)\, dy - \left(\frac{1-x}{2}\right)^2 \\
&= \frac{1}{3}(1-x)^2 - \left(\frac{1-x}{2}\right)^2 \\
&= \frac{1}{12}(1-x)^2
\end{aligned}$$

6.) Derive $\text{Var}(\mathbb{E}[Y|X])$ and $\mathbb{E}[\text{Var}(Y|X)]$:

$$\begin{aligned}
\text{Var}(\mathbb{E}[Y|X]) &= \mathbb{E}[(\mathbb{E}[Y|X])^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\
&= \int_0^1 \left(\frac{1-x}{2}\right)^2 2(1-x)\, dx - \mathbb{E}[Y]^2 \\
&= \frac{1}{8} - \frac{1}{9} \\
&= \frac{1}{72}
\end{aligned}$$

Alternatively,

$$\begin{aligned}
\text{Var}(\mathbb{E}[Y|X]) &= \text{Var}\left(\frac{1-X}{2}\right) \\
&= \frac{1}{4}\text{Var}(X) \\
&= \frac{1}{4}\left(\int x^2 2(1-x)dx - \mathbb{E}[X]^2\right) \\
&= \frac{1}{4}\left(\frac{1}{6} - \frac{1}{9}\right) \\
&= \frac{1}{4} \times \frac{1}{18} = \frac{1}{72}
\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\text{Var}[Y|X]) &= \int_0^1 \frac{1}{12}(1-x)^2 \cdot 2(1-x)\,dx \\ &= \frac{1}{24}\end{aligned}$$

Indeed, we see that the Conditional Variance Identity holds true here. $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$, where $\text{Var}(Y) = \int_0^1 y^2 2(1-y)\,dy - \frac{1}{9} = \frac{1}{18}$.

3. Transformation of bivariate random variables

Let (X,Y) be a bivariate random vector. Consider a new bivariate random vector (U,V) defined by $U = g_1(X,Y)$, $V = g_2(X,Y)$. What is the probability distribution of (U,V)?

Let $\mathcal{A}$ denote the support of the (X,Y), i.e. $\mathcal{A} = \{(x,y) \in \mathbb{R}^2 : f_{X,Y}(x,y) > 0\}$.

The transformation is $U = g_1(X,Y)$ and $V = g_2(X,Y)$. The support of (U,V) is then $\mathcal{B} = \{(u,v) \in \mathbb{R}^2 : u = g_1(x,y), v = g_2(x,y) \text{ for some } (x,y) \in \mathcal{A}\}$.

Assume that g_1 and g_2 are functions such that the relationship between $\mathcal{A}$ and $\mathcal{B}$ is one-to-one and onto (a bijection). For each $(u,v) \in \mathcal{B}$, there is only one $(x,y) \in \mathcal{A}$ such that $u = g_1(x,y)$ and $v = g_2(x,y)$.

As such, we can solve the equations $u = g_1(x,y)$ and $v = g_2(x,y)$ in terms of x and y. That is, there is an inverse transformation such that $x = h_1(u,v)$ and $y = h_2(u,v)$, where h_1 and h_2 are differentiable functions.

Define the Jacobian matrix:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{bmatrix}$$

The determinant of the Jacobian matrix is:

$$J = \det(\mathbf{J}) = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}$$

That is, $J = \frac{\partial h_1}{\partial u}\frac{\partial h_2}{\partial v} - \frac{\partial h_1}{\partial v}\frac{\partial h_2}{\partial u}$.

The joint pdf of (U,V) is:

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}(h_1(u,v), h_2(u,v))\,|J| & \text{for } (u,v) \in \mathcal{B} \\ 0 \quad \text{otherwise} \end{cases}$$

$|\det(\mathbf{J})|$ is often called the Jacobian, or the Jacobian of the transformation, or the Jacobian determinant. Note that $\det(\mathbf{J})$ is a function of u, v. Moreover, $\det(\boldsymbol{J}) \neq 0$ since there is an inverse transformation such that $x = h_1(u,v)$ and $y = h_2(u,v)$, where h_1 and h_2 are differentiable functions. The Jacobian is also used during change-of-variables in multiple integrals.

3.1. Example

Let X and Y be independent, standard Normal random variables.

Consider the transformation $U = X + Y$ and $V = X - Y$. What is the joint pdf of (U, V)?

The joint pdf of (X, Y) is just $f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi}e^{-\frac{x^2}{2}}e^{-\frac{y^2}{2}}$ since X and Y are independent.

The support of (X, Y) is $\mathbb{R}^2$. It follows that U and V can also take any value from $-\infty$ to ∞.

The inverse transformation is $x = h_1(u,v) = \frac{u+v}{2}$ and $y = h_2(u,v) = \frac{u-v}{2}$.

The Jacobian of the transformation is:

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Hence the joint pdf of (U, V) is:

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(h_1(u,v), h_2(u,v))\,|J| \\ &= \frac{1}{2\pi}e^{-\frac{(\frac{u+v}{2})^2}{2}}e^{-\frac{(\frac{u-v}{2})^2}{2}}\frac{1}{2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{u^2}{4}}\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{v^2}{4}}\right) \end{aligned}$$

Note that the pdf of $N(\mu, \sigma^2)$ is $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Hence the joint pdf of (U, V) can be factored into two functions $f_U(u)$ and $f_V(v)$. Moreover, $f_U(u)$ is the pdf of $N(0, 2)$. That is, $U \sim N(0, 2)$ and $V \sim N(0, 2)$. The sum (U) and difference (V) of independent normal random variables are independent normal random variables, as long as $\text{Var}(X) = \text{Var}(Y)$.

We can also consider the ratio and the product of Normal variables. Consider the transformation $U = X/Y$ and $V = X$. What is the joint pdf of (U, V)? What about the product $V = XY$?

3.2. Discrete bivariate random vectors

Let (X, Y) be a a discrete bivariate random vector. Let $\mathcal{A}$ be the support of (X, Y), i.e. the set of points where the joint pmf of (X, Y) takes strictly positive values. Note that $\mathcal{A}$ must be a countable set (either finite or countably infinite).

The joint pmf of (U, V) is:

$$f_{U,V}(u, v) = P(U = v, V = v) = \sum_{(x,y)\in\mathcal{A}:g_1(x,y)=u,g_2(x,y)=v} f_{X,Y}(x, y)$$

4. Some important inequalities

4.1. Jensen's Inequality

A function $g(x)$ is convex if and only if $\lambda g(x) + (1 - \lambda)g(y) \geq g(\lambda x + (1 - \lambda)y)$ for $0 < \lambda < 1$. Graphically, a straight line connecting any two points of the convex function lies above the function.

Jensen's Inequality: For any random variable X, if $g(X)$ is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.

For example: take $g(X) = X^2$, then $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, which implies that $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$.

4.2. Concentration inequalities (Markov and Chebychev's Inequalities)

Concentration inequalities provide bounds on the probabilities of a random variable deviating from a certain value. Markov's inequality and Chebyshev's inequality are examples of concentration inequalities. Let X be a random variable and $g(X)$ be a non-negative function. Chebyshev's inequality: for any $\epsilon > 0$,

$$P(g(X) \geq \epsilon) \leq \frac{\mathbb{E}[g(X)]}{\epsilon}$$

Proof:

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\ &\geq \int_{x:g(x)\geq\epsilon}^{\infty} g(x)f(x)\,dx \\ &\geq \int_{x:g(x)\geq\epsilon}^{\infty} \epsilon f(x)\,dx \\ &= \epsilon P(g(X)\geq\epsilon)\end{aligned}$$

Markov's inequality is just $P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$.

Now let $g(x) = \frac{(x-\mu)^2}{\sigma^2} \geq 0$, where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}(X)$. Note that g is always positive. By the Chebyshev's inequality,

$$\begin{aligned}P(g(X) \geq \epsilon^2) &\leq \frac{\mathbb{E}[g(X)]}{\epsilon^2} \\ P(\frac{(X-\mu)^2}{\sigma^2} \geq \epsilon^2) &\leq \frac{\mathbb{E}[\frac{(X-\mu)^2}{\sigma^2}]}{\epsilon^2} \\ P(\frac{(X-\mu)^2}{\sigma^2} \geq \epsilon^2) &\leq \frac{1}{\epsilon^2} \\ P(|X-\mu| \geq \epsilon\sigma) &\leq \frac{1}{\epsilon^2} \qquad (5)\end{aligned}$$

If we take $\epsilon = 2$, then $P(|x-\mu| \geq 2\sigma) \leq 0.25$ or $P(|x-\mu| < 2\sigma) > 0.75$. That is, there is at least 75% chance that a random variable (any random variable!) will be within 2 standard deviation of its mean.

In general, the Chebyshev's inequality can be used to show that as $\text{Var}(X_n) \to 0$, $P(|X_n - \mu| \geq \epsilon) \to 0$, by taking $g(X) = (X-\mu)^2$.

As such, Chebyshev's inequality can be used to prove the Weak Law of Large Numbers. Let $X_1, \ldots, X_n$ be n independent random variables, each with the same density f. Define the sample mean as the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that $\bar{X}$ has expectation $\mathbb{E}[X] \equiv \mu$, and variance $\frac{\text{Var}(X)}{n} \equiv \frac{\sigma^2}{n}$.

By the inequality in (5), we have:

$$P(|\bar{X}-\mu| \geq \epsilon\frac{\sigma}{\sqrt{n}}) \leq \frac{1}{\epsilon^2}$$

Now if we let $\epsilon = v\frac{\sqrt{n}}{\sigma}$,

$$P(|\bar{X}-\mu| \geq v) \leq \frac{\sigma^2}{nv^2}$$

Therefore, as $n \to \infty$, $P(|\bar{X}-\mu| \geq v) = 0$ for any $v > 0$, which is the Weak Law of Large Numbers.

5. Common families of statistical distributions

5.1. Multivariate Normal

We are already familiar with the one-dimensional Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, which has the pdf $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ with the support over the entire real line.

The k-dimensional Gaussian random variable is described as:

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

$\boldsymbol{X}$ is a k-dimensional random vector. $\boldsymbol{\mu}$ is a k-dimensional vector, Σ is a k-by-k symmetric matrix called the variance-covariance matrix. A matrix Σ is symmetric if $\Sigma^T = \Sigma$, as such Σ has $k+(k^2-k)/2 = (k^2+k)/2$ number of parameters. Intuitively, k diagonal terms of Σ describe the variances of each individual random variable, and $(k^2-k)/2$ off-diagonal terms of Σ describe the pairwise correlations between each of the variable.[1]

Therefore a k-dimensional Gaussian variable has $\frac{3k+k^2}{2}$ number of parameters. For example, a 2-dimensional multivariate Gaussian has 5 parameters.

For the bivariate Normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right]$$

The pdf of (X, Y) is:

(6)

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

for $x, y \in \mathbb{R}^2$. Check that the marginal pdf of X is just the univariate Normal pdf:

---

[1] In additional Σ also has to be positive semi-definite, that is, $\boldsymbol{x}^T\Sigma\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^k$.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}, \quad x \in \mathbb{R}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}, \quad y \in \mathbb{R}$$

Hence, the moments of (X, Y) are described by the parameters of the pdf, i.e. $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\text{Var}(X) = \sigma_X^2$, $\text{Var}(Y) = \sigma_Y^2$.

In addition, we can compute $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ from the joint pdf, which turns out to be $\rho\sigma_X\sigma_Y$. As such the correlation of X and Y is just ρ.

If we set $\rho = 0$, i.e. zero correlation between X and Y, then:

$$f(x, y) = f_X(x) f_Y(y)$$

Hence, for Multivariate Normals, zero correlation implies independence. Also, if X and Y are independent with univariate Normal distributions, then (X, Y) trivially has a bivariate Normal distribution.

However in general, if two random variables X and Y are univariate Normals, it is not true that (X, Y) has a bivariate Normal distribution. Can you work out an example?

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

$$(Y|X = x) \sim \mathcal{N}\Big(\mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(x - \mathbb{E}[X]), (1 - \rho^2)\sigma_Y^2\Big) \tag{7}$$

This implies that the conditional expectation of Y given X is:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}[X])$$

It is a **linear** function of X and has a normal pdf. The fact that $\mathbb{E}[Y|X]$ is linear in X means that the best prediction of Y using X is some linear function of X. That is, we can't do better than a linear regression of Y on X if (Y, X) is a bivariate Normal.

The conditional variance of Y given X is $\text{Var}[Y|X] = (1 - \rho^2)\sigma_Y^2$, which does not depend on X.

In general, the joint density of a k-th dimensional multivariate Normal distribution is:

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

Where $\boldsymbol{\Sigma}$ is a k-by-k variance-covariance matrix of $\boldsymbol{X}$, and $\boldsymbol{\mu}$ is a k-dimensional vector. We say that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

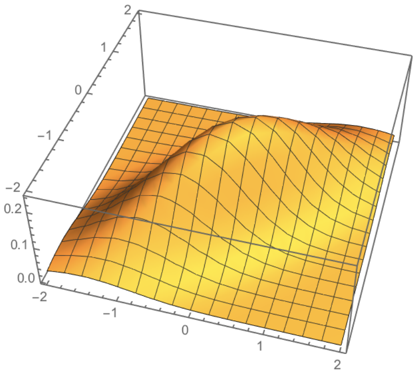
5.2. Example

For example, let $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$ in the joint pdf of Bivariate Normal (Equation 8). The location parameters μ_X and μ_Y merely shift the center of the distribution around. Then we have:

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)}\right) \tag{8}$$

Visualize this joint pdf at various values of ρ as in Figure 1.

$$\texttt{Plot3D}\Big[\texttt{ReplaceAll}\Big[\frac{1}{2\,\pi\,\sqrt{(1-\rho^2)}}\ \texttt{Exp}\Big[-\frac{x^2+y^2-2\,\rho\,x\,y}{2\,(1-\rho^2)}\Big],\ \rho\to 0.75\Big],\ \{x,\ -2,\ 2\},\ \{y,\ -2,\ 2\}\Big]$$



$$\texttt{Plot3D}\Big[\texttt{ReplaceAll}\Big[\frac{1}{2\,\pi\,\sqrt{(1-\rho^2)}}\ \texttt{Exp}\Big[-\frac{x^2+y^2-2\,\rho\,x\,y}{2\,(1-\rho^2)}\Big],\ \rho\to 0\Big],\ \{x,\ -2,\ 2\},\ \{y,\ -2,\ 2\}\Big]$$

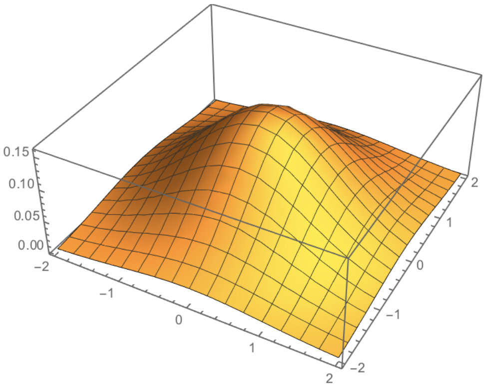


Figure 1

Now we derive the conditional distribution of X given Y.

$$
\begin{aligned}
f_{X|Y=y}(x|y) &= \frac{f(x,y)}{f(y)} \\
&= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right)}{\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)} + \frac{y^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+\rho^2 y^2-2\rho xy}{2(1-\rho^2)}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right)
\end{aligned}
$$

The last line is the pdf of a univariate Normal distribution with mean ρy and variance $1-\rho^2$. Therefore,

$$(X|Y=y) \sim N(\rho y, 1-\rho^2)$$

Let's also check whether the joint pdf integrates to the marginal pdfs (which can be evaluated analytically by completing the squares):

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{-y^2}{2(1-\rho^2}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2-2\rho xy}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{\rho^2 y^2-y^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{-y^2}{2}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}
\end{aligned}
$$

5.3. Sampling from a multivariate Normal

To sample from a scalar random variable, we learned how to use the probability integral transform. We can use the conditional distribution to sample from a multivariate distribution. For instance, to sample from a bivariate Normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \ \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right]$$

First, we sample from the marginal of X, which is just $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$.

Recall the conditional distribution of Y given $X = x$ is:

$$(Y|X = x) \sim \mathcal{N}\Big(\mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(x - \mathbb{E}[X]), (1 - \rho^2)\sigma_Y^2\Big)$$

For every draw of x_i from the marginal distribution $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, we then sample y_i from $Y|X = x_i$. The sample $(x_i, y_i)_{i=1}^n$ will be a valid sample from the the bivariate Normal distribution.

This approach is called *Gibbs Sampling.*[2] More generally, to sample from a trivariate distribution $f(x, y, z)$, we first draw x_i from the marginal of X, then draw y_i from $Y|X = x_i$, then finally, draw z_i from $Z|Y = y_i, X = x_i$. Now, the density of $Z|Y, X$ can be derived as $f(x, y, z)/f(x, y)$.

Let's try to implement Gibbs sampling using R or Python.

5.4. Beta distribution

Beta distribution is used to model random variables that lie within the unit interval $[0, 1]$. For example, if we want to model fractions or probabilities, then we use the Beta distribution.

The Beta distribution is controlled by two parameters $\alpha > 0$ and $\beta > 0$, that is, $X \sim Beta(\alpha, \beta)$.

The pdf is $f_X(x) \propto x^{\alpha-1}(1 - x)^{\beta-1}$ for $x \in [0, 1]$. The constant of proportionality is $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where Γ is the Gamma function.[3]

---

[2] More specifically, this is the Collapsed Gibbs Sampling

[3] The Gamma function is an interesting function. It is defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. The Gamma function satisfies the following recurrence relation: $\Gamma(z) = (z - 1)\Gamma(z - 1)$. As such, when z is an integer, $\Gamma(z) = (z - 1)!$. We can think of the Gamma function as an extension of the factorial function to non-negative real numbers. For non-integers $z > 1$, it must be that $\Gamma(z) = (z - 1)(z - 2)\ldots\delta\Gamma(\delta)$ where $0 < \delta < 1$.

The Beta distribution is a very flexible class of distributions that can generate distributions that are positively or negatively skewed, varying modes and medians. The mean is given by $\frac{\alpha}{\alpha+\beta}$.

The Dirichlet distribution generalizes the Beta distribution to multiple dimensions:

$$f\left(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K\right) \propto \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

Where $\{x_k\}_{k=1}^{k=K}$ belong to the standard $K-1$ simplex, or in other words: $\sum_{i=1}^{K} x_i = 1$ and $x_i \geq 0$ for all $i \in \{1, \ldots, K\}$. The normalizing constant is the multivariate beta function, which can be expressed in terms of the gamma function

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

5.5. Gamma distribution

The Gamma distribution is used to model random variables that takes positive values. It is a general form of the Exponential distribution. It is also used in Bayesian statistics as conjugate priors. Moreover, it is used in the frequentist setting for hypothesis testing.

Let $X_1, X_2, \ldots, X_n$ be n independent Exponential distribution with parameter λ. Then, $\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda)$. Therefore the Gamma distribution gives the duration it takes until n number of event occurrences, where the rate of an event occurrence is λ.

More generally, the Gamma distribution is a two-parameter distribution. $X \sim \text{Gamma}(\alpha, \beta)$ where X takes only positive real values and $\alpha, \beta > 0$. The pdf is given by $f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)}$ for $x \geq 0$.

If $X \sim \text{Gamma}(1, \lambda)$, then X has an exponential distribution with mean $\frac{1}{\lambda}$. If $X \sim \text{Gamma}(v/2, 1/2)$, then X is identical to $\chi(v)$, the chi-squared distribution with v degrees of freedom.

5.6. Bernoulli and Binomial Distribution

X is a Bernoulli distribution with parameter p if $X = 1$ with probability p, and $X = 0$ with probability $1 - p$.

Let $X_1, X_2, \ldots, X_n$ be n independent Bernoulli random variables with parameter p. $Y = \sum_{i=1}^{n} X_i$ is a Binomial distribution with parameters (n, p).

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Y is the number of successes in n independent trials, where p is the probability of a success in a trial. The mean of Y is np, and the variance of Y is $np(1-p)$, can you prove this?

6. A note on truncated random variables

Consider a random variable X with density $f_X(x)$. What is $\mathbb{E}[X|X > a]$? $X > a$ is an event, not a random variable, so do not confuse with the formula for deriving conditional density. The density of $X|X > a$ is $\frac{1}{1-F_X(a)} f_X(x)\mathbb{1}(x > a)$ with the support truncated to $x > a$. Note this density integrates to one.

In general, the density of $X|X \in (a, b)$ is $\frac{1}{F_X(b)-F_X(a)} f_X(x)\mathbb{1}(x \in (a, b))$ with the support truncated to $x \in (a, b)$. Note this density integrates to one as well.

Now let $X \sim U[0, 1]$, and $a \in (0, 1)$, what is $\mathbb{E}[X|X > a]$?

$$\begin{aligned}
\mathbb{E}[X|X > a] &= \frac{\mathbb{E}[X\mathbb{1}_{\{X>a\}}]}{1 - F_X(a)} \\
&= \frac{\int_a^\infty x f_X(x)\, dx}{1 - F_X(a)} \\
&= \frac{\int_a^1 x\, dx}{1 - a} \\
&= \frac{\int_a^1 x\, dx}{1 - a} \\
&= \frac{a+1}{2} \quad \text{for } a \in (0, 1)
\end{aligned}$$

For instance, if $X \sim \mathcal{N}(0, \sigma^2)$, then we can use the above formula to show that $\mathbb{E}[X|X > 0] \approx 0.7978\sigma$.

Now consider the random variables (X, Y) which are joint uniformly distributed on the unit square. That is, $f(x, y) = 1$ for $0 < x < 1$ and $0 < y < 1$. Show that $\mathbb{E}[X|Y > X] = \frac{1}{3}$. Note that $Y > X$ is an event, not a random variable. As such, the formula to compute this conditional expectation is $\mathbb{E}[X|Y >$

$X] = \frac{\mathbb{E}[X\mathbb{1}_{\{Y>X\}}]}{P(Y>X)}$, and *not* $\mathbb{E}[Y|X = x] = \int y\frac{f(x,y)}{f(x)}\,dy$, which is the formula when conditioning on a random variable. In general, the density of $X|(X,Y) \in A$ is $\int_{-\infty}^{\infty} \frac{1}{\Pr((X,Y)\in A)} f_{X,Y}(x,y)\mathbb{1}((x,y) \in A)$.

$$\begin{aligned}
\mathbb{E}[X|Y > X] &= \frac{\mathbb{E}[X\mathbb{1}_{\{Y>X\}}]}{\Pr(Y > X)} \\
&= \frac{\int_0^1 x \int_x^1 f_{X,Y}(x,y)\,dy\,dx}{1/2} \\
&= \frac{\int_0^1 x(1-x)\,dx}{1/2} \\
&= \frac{1}{3}
\end{aligned}$$