LECTURE 4: MULTIVARIATE RANDOM VARIABLES II

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
SEPTEMBER 24, 2025

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 $f_X(x)$.

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

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

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,

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

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

Comparing $\mathbb{E}[Y|X]$ and $\mathbb{E}[Y|Z]$, we can think of $\mathbb{E}[Y|Z]$ as conditioning on a smaller information set. Because if we know the outcome of X, then we would know Z, but the converse is not true.

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

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

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:

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

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

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 $\text{Var}(Y|X)$ 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 $\{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

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

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

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

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

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

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:

$$\int_0^1 [2x]_0^{1-y}\,dx = \int_0^1 2(1-y)\,dy$$

$$= 2\left[y - \frac{y^2}{2}\right]_0^1 = 2(1-\frac{1}{2}) = 1$$

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 $\text{Cov}(X,Y)$

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

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

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

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

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 $Y|X=x$:

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

Conditional expectation:

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

Conditional variance:

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

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

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

Alternatively,

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

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

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

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:

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

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

The joint pdf of (U,V) is:

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

$|\det(\mathbf{J})|$ 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:

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

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

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

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 Chebyshev

Concentration inequalities give worst–case upper bounds on the probability that a random variable deviates from a typical value (often its mean). They are distribution-free: the bounds hold without assuming normality or smooth densities, and they work for discrete, continuous, or mixed random variables as long as the relevant moments exist.

We start with Markov's inequality, which applies to any nonnegative random variable. Chebyshev's inequality then follows by applying Markov to a squared deviation from the mean.

Theorem 1 (Markov's inequality). *Let $Y \geq 0$ be a random variable with $\mathbb{E}[Y] < \infty$. For any $\varepsilon > 0$,*

$$\Pr(Y \geq \varepsilon) \;\leq\; \frac{\mathbb{E}[Y]}{\varepsilon}.$$

Proof. Since $Y \geq \varepsilon\,\mathbf{1}_{\{Y\geq\varepsilon\}}$, taking expectations gives

$$\mathbb{E}[Y] \;\geq\; \varepsilon\,\mathbb{E}[\mathbf{1}_{\{Y\geq\varepsilon\}}] \;=\; \varepsilon\,\Pr(Y \geq \varepsilon).$$

$\square$

Note that the proof does not assume a density and holds for discrete, continuous, or mixed Y.

A common corollary (take $Y = |X|$) is

$$\Pr(|X| \geq \varepsilon) \;\leq\; \frac{\mathbb{E}[|X|]}{\varepsilon}, \qquad \text{provided } \mathbb{E}[|X|] < \infty.$$

More generally, for any $r > 0$ and $t > 0$,

$$\Pr\big(|X| \geq t\big) \;\leq\; \frac{\mathbb{E}\big[|X|^r\big]}{t^r},$$

by applying Markov to $|X|^r$.

Corollary 1 (Chebyshev's inequality)**:** Let X have mean μ and variance $\sigma^2 < \infty$. For any $t > 0$,

$$\Pr\left(|X-\mu| \geq t\right) \;\leq\; \frac{\sigma^2}{t^2}.$$

Equivalently, for $k > 0$,

$$\Pr\left(|X-\mu| \geq k\sigma\right) \;\leq\; \frac{1}{k^2}.$$

Proof. Apply Markov to the nonnegative random variable $(X-\mu)^2$ with threshold t^2:

$$\Pr\left((X-\mu)^2 \geq t^2\right) \;\leq\; \frac{\mathbb{E}[(X-\mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

$\square$

Setting $k = 2$ gives

$$\Pr\left(|X-\mu| \geq 2\sigma\right) \;\leq\; \frac{1}{4} \quad\Rightarrow\quad \Pr\left(|X-\mu| < 2\sigma\right) \;\geq\; \frac{3}{4}.$$

So *at least* 75% of the mass lies within two standard deviations of the mean—no distributional shape assumed. (For bell-shaped distributions like the normal, the true coverage is much higher—about 95%—illustrating that Chebyshev is conservative but universal.)

Tightness (sharpness). These inequalities are *best possible* in general: one cannot replace the right-hand side by a uniformly smaller function of the moments alone.

- **Markov:** Fix $\varepsilon > 0$ and $m = \mathbb{E}[Y]$. The two-point distribution with $\Pr(Y = \varepsilon) = m/\varepsilon$ and $\Pr(Y = 0) = 1 - m/\varepsilon$ satisfies $\Pr(Y \geq \varepsilon) = m/\varepsilon$, i.e., equality.
- **Chebyshev:** For any $k > 0$, the three-point distribution
$$\Pr(X = \mu) = 1 - \tfrac{1}{k^2}, \quad \Pr\big(X = \mu \pm k\sigma\big) = \tfrac{1}{2k^2} \text{ each},$$
has mean μ, variance σ^2, and achieves $\Pr(|X - \mu| \geq k\sigma) = 1/k^2$.

Examples.

(i) **Discrete (Poisson).** If $X \sim \text{Poisson}(\lambda)$ then $\mu = \sigma^2 = \lambda$. For $k = 2$,
$$\Pr\Big(|X - \lambda| \geq 2\sqrt{\lambda}\Big) \;\leq\; \frac{1}{4}.$$
For instance, with $\lambda = 1$, the actual tail is $\Pr(|X - 1| \geq 2) = \Pr(X \geq 3) \approx 0.0803 < 0.25$, which is rather conservative.

(ii) **Normal.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\Pr(|X - \mu| < 2\sigma) \approx 0.954$, far tighter than the Chebyshev guarantee ≥ 0.75.

If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then for any $\varepsilon > 0$,
$$\Pr\Big(\big|\bar{X}_n - \mu\big| \geq \varepsilon\Big) \;\leq\; \frac{\operatorname{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2} \;\xrightarrow[n\to\infty]{}\; 0,$$
a one-line proof of the WLLN (Weak Law of Large Numbers) via Chebyshev.

Theorem 2 (Cantelli's one-sided Chebyshev inequality)**.** *Let X have mean μ and variance $\sigma^2 < \infty$. For any $t > 0$,*
$$\Pr\big(X - \mu \geq t\big) \;\leq\; \frac{\sigma^2}{\sigma^2 + t^2} \quad \textit{and} \quad \Pr\big(\mu - X \geq t\big) \;\leq\; \frac{\sigma^2}{\sigma^2 + t^2}.$$
Equivalently, with $k = t/\sigma$, $\Pr(X - \mu \geq k\sigma) \leq \frac{1}{1+k^2}$.

Proof. Write $Y = X - \mu$ so that $\mathbb{E}[Y] = 0$ and $\operatorname{Var}(Y) = \sigma^2$. Fix any $a \geq 0$. Then $\{Y \geq t\} \subseteq \{Y + a \geq t + a\}$ and hence
$$\Pr(Y \geq t) \;=\; \Pr\big((Y + a)^2 \geq (t + a)^2\big) \;\leq\; \frac{\mathbb{E}[(Y + a)^2]}{(t + a)^2} \;=\; \frac{\sigma^2 + a^2}{(t + a)^2},$$

by Markov's inequality applied to the nonnegative $(Y+a)^2$. We now minimize the right-hand side over $a \geq 0$:

$$\frac{\sigma^2+a^2}{(t+a)^2} - \frac{\sigma^2}{\sigma^2+t^2} = \frac{(\sigma^2-ta)^2}{(t+a)^2(\sigma^2+t^2)} \geq 0,$$

with equality at $a^* = \sigma^2/t$. Therefore $\Pr(Y \geq t) \leq \frac{\sigma^2}{\sigma^2+t^2}$. Applying the same bound to $-Y$ gives the left-tail version. One can also show that this bound is sharp in that there exists a random variable for which this bound is achieved with equality. □

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. The k diagonal terms of Σ describe the variances of each individual random variable, while the $(k^2-k)/2$ 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:

(5)
$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

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

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

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

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

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

In addition, we can compute $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ 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:

$$(Y|X = x) \sim \mathcal{N}\Big(\mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(x - \mathbb{E}[X]), (1 - \rho^2)\sigma_Y^2\Big) \tag{6}$$

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 7). The location parameters μ_X and μ_Y merely shift the center of the distribution around. Then we have:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) \tag{7}$$

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

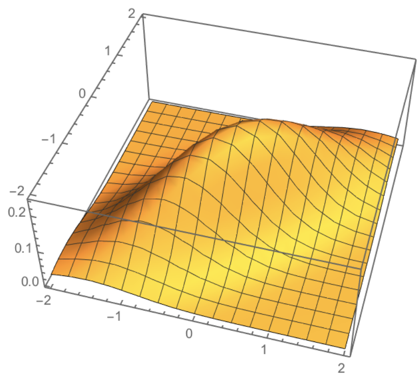
Now we derive the conditional distribution of Y given X.

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

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

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

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



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

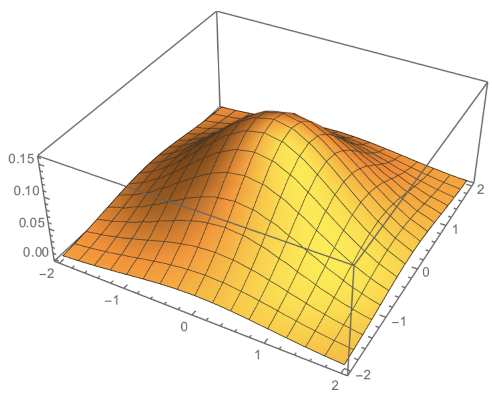


FIGURE 1

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:

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

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]

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:

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

[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 $0 < \delta < 1$.

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, and 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 $x \in (a, b)$.

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

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

For instance, if $X \sim \mathcal{N}(0, \sigma^2)$, then we can use the above formula to show that $\mathbb{E}[X|X > 0] = 2\int_0^\infty x\phi(x)\, dx = 2\frac{\sigma}{\sqrt{2\pi}} \approx 0.7978\sigma$.

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. We can show that the joint density of $(X, Y)|(X, Y) \in A$ is $\frac{1}{\Pr((X,Y)\in A)} f_{X,Y}(x, y)\mathbb{1}((x, y) \in A)$, hence, the density of $X|(X, Y) \in A$ is $\int_{-\infty}^\infty \frac{1}{\Pr((X,Y)\in A)} f_{X,Y}(x, y)\mathbb{1}((x, y) \in A)\, dy$

$$\begin{aligned}
f_{X|Y>X}(x) &= \int_0^1 \frac{1}{\Pr(Y > X)} f_{X,Y}(x, y)\mathbb{1}(y > x)\, dy \\
&= \int_x^1 2\, dy = 2(1 - x), \quad \text{for } x \in [0, 1]
\end{aligned}$$

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

*Let $f_{X,Y}(x, y)$ be the joint density of (X, Y). We want to find the conditional density $f_{(X,Y)\,|\,(X,Y)\in A}(x, y)$, which must satisfy the following for all measurable set $B \subseteq \mathbb{R}^2$.

$$P\left((X,Y) \in B \mid (X,Y) \in A\right) = \int\int_B f_{(X,Y)\,|\,(X,Y)\in A}(x,y)\,dx\,dy \tag{8}$$

By the definition of conditional probability:

$$P\left((X,Y) \in B \mid (X,Y) \in A\right) = \frac{P\left((X,Y) \in B \cap A\right)}{P\left((X,Y) \in A\right)} \tag{9}$$

Now,

$$\begin{aligned} P\left((X,Y) \in B \cap A\right) &= \int\int_{A\cap B} f_{X,Y}(x,y)\,dx\,dy \\ &= \int\int_B f_{X,Y}(x,y)\mathbb{1}((x,y) \in A)\,dx\,dy \end{aligned}$$

Equating 8 and 9, which holds for all B, the integrands must be equal almost everywhere, we then have $f_{(X,Y)\,|\,(X,Y)\in A}(x,y) = \frac{1}{\Pr((X,Y)\in A)} f_{X,Y}(x,y)\mathbb{1}((x,y) \in A)$, for $(x,y) \in \mathbb{R}^2$.