Lecture 4
Multiple Random Variables

Summary: We first introduce the joint probability distribution of a bivariate random vector (X, Y) via the characterization of the joint cumulative distribution function, the joint probability mass function (when (X, Y) are discrete), and the joint probability density function (when (X, Y) are continuous). We then characterize the relationships between X and Y using the conditional distributions, correlation, and conditional expectations. The concept of independence and its implications on the joint distributions are also discussed. Finally, we also introduce a class of bivariate normal distribution and examine its important properties. We study more on the conditional distributions and conditional expectations. Therefore, we discuss more on the concept of independence and introduce the law of the iterated expectations.

Key words: Joint probability distribution, Marginal distribution, Conditional Distribution, Correlation, Independence, Conditional Mean, Conditional Variance, Law of Iterated Expectations, Bivariate Transformation.

1. Random Vector and Its Distribution

Random Vector

Definition: Let $\{S, \mathcal{F}, P\}$ be a probability space. A random vector $X \in \mathbb{R}^k$ defined on $\{S, \mathcal{F}, P\}$ is a $\mathbb{R}^k$-valued function $X(\cdot)$ on S which assigns one and only one $x = X(s)$ to each $s \in S$, and is such that for each $x_1, \cdots, x_k \in \mathbb{R}$,
$A \equiv \{s \in S : X_1(s) \leq x_1, \cdots, X_k(s) \leq x_k\} \in \mathcal{F}$. □

Remark: The above definition implies that the components of X are random variables, $X = (X_1, X_2, \cdots, X_k)'$. A vector of random variables is a random vector. Similarly to the univariate case, each random vector induces a probability space $\left\{\mathbb{R}^k, \mathcal{B}, \mu\right\}$, where μ is defined as $\mu(B) = \Pr(X \in B) = \Pr(A)$, where
$B = \times_{i=1}^{k}(-\infty, x] \in \mathcal{B}$. □

Distribution Function

Consider a probability measure μ on $\mathcal{B}$ which is a collection of the sets of the form

$$\times_{j=1}^{k}(-\infty, x_j].$$

Then we can define a real function $F(\cdot)$ on $\mathbb{R}^k$. We can similarly define a multivariate CDF:

$$F\left(x_1, \cdots, x_k\right) = \mu\left(\times_{j=1}^{k}(-\infty, x_j]\right) = \Pr(X_1 \leq x_1, \cdots, X_k \leq x_k),$$

which has the following properties:

1. $F(x_1, \cdots, x_k)$ is monotonically non-decreasing in each of its arguments:

$$F(x_1, \cdots, x_k) \geq F(y_1, \cdots, y_k) \qquad \text{if } x_j \geq y_j,\ j = 1, \cdots, k.$$

2. $F(x_1, \cdots, x_k)$ is right continuous in each of its arguments:

$$\lim_{h \downarrow 0} F(x_1, \cdots, x_i + h, \cdots x_k) = F(x_1, \cdots, x_k).$$

3. $\lim_{\min x_j \to -\infty} F(x_1, \cdots, x_k) = 0,$
$\lim_{\min x_j \to +\infty} F(x_1, \cdots, x_k) = 1.$

Let $X = (X_1', X_2')'$ be a random vector in $\mathbb{R}^{k_1+k_2}$ with $X_1 \in \mathbb{R}^{k_1}$ and $X_2 \in \mathbb{R}^{k_2}$. Let F be the (joint) distribution function of X and F_1, F_2 be the distribution functions of X_1, X_2, respectively. The distribution functions F_1 and F_2 are called **marginal distribution** functions of F. They can be constructed from F by setting the arguments of F corresponding to X_1 and X_2, respectively, equal to ∞. Thus, $F_1(x_1, \cdots, x_{k_1}) = F(x_1, \cdots, x_{k_1}, \infty, \cdots, \infty)$, $F_2(x_{k_1+1}, \cdots, x_{k_1+k_2}) = F(\infty, \cdots, \infty, x_{k_1+1}, \cdots, x_{k_1+k_2})$.

PDF/PMF

Definition: A distribution function F on $\mathbb{R}^k$ is *continuous* if there exists a non-negative function f on $\mathbb{R}^k$ such that for $x_1, \cdots, x_k \in \mathbb{R}$,

$$F(x_1, x_2, \cdots, x_k) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_k} f(y_1, y_2, \cdots, y_k)\, dy_1 dy_2 \cdots dy_k.$$

This function f is called the density (probability density function, PDF) of F. □

Definition: A distribution function F on $\mathbb{R}^k$ is *discrete* if there exists a countable subset X of $\mathbb{R}^k$ and a positive function f on $\mathbb{R}^k$ such that for $x_1, \cdots, x_k \in \mathbb{R}$,

$$F(x_1, x_2, \cdots, x_k) = \sum_{\substack{(y_1, y_2, \cdots, y_k)' \in X \\ y_j \leq x_j,\ j=1, \cdots, k}} f(y_1, y_2, \cdots, y_k).$$

This f is also called the density (probability density function, PDF) or probability mass function (PMF) of F. □

Independence of Random Variables

Definition: Let $X_1, X_2, X_3, \cdots$ be random variables defined on a common probability space $\{S, \mathcal{F}, P\}$. Then X_j, $j = 1, 2, \cdots, n$ (n may be ∞) are independent if

$$F(x_1, \cdots, x_n) = \prod_{j=1}^{n} F_j(x_j),$$

where $F(x_1, \cdots, x_n)$ is the joint CDF and $F_j(x_j)$, $j = 1, 2, \cdots, n$, are the marginal CDFs. □

Theorem: If X_j, $j = 1, 2, \cdots, n$ (n may be ∞) are independent, then

$$f(x_1, \cdots, x_n) = \prod_{j=1}^{n} f_j(x_j).$$

□

2. Linear Multivariate Transformations

Distributions of Functions of Random Vectors

- Let $\mathbf{X}$ be a random vector in $\mathbb{R}^k$ with continuous distribution function $F(x)$, $\mathbf{x} \in \mathbb{R}^k$, and density $f(\mathbf{x})$.
- Recall the case in LEC 2 when $k = 1$.
- Consider the following example.
  Let X be a random variable with density $f(x) = 2x$ for $0 < x < 1$; zero, otherwise.
- Consider the two types of linear transformation of X.
  $Y = g(X) = X + 1$
  $Y = g(X) = 2X$
  What is the difference in the above two transformations? Any linear transformation can be decomposed into these two types.

(1) Let $Y = g(X) = X + 1$. Calculate the density $h(y)$ of y. The transformation $Y = X + 1$ maps $A = \{x : 0 < x < 1\}$ onto $B = \{y : 1 < y < 2\}$. Note that this transformation shifts the domain but preserves the distance.

$$H(y) = \Pr(Y \leq y) = \Pr(X+1 \leq y) = \Pr(X \leq y-1) = F(y-1)$$

$$\begin{aligned} h(y) &= \frac{dH(y)}{dy} = \frac{d}{dy}F(y-1) = f(y-1) \cdot 1 = f(y-1) \\ &= \begin{cases} 2y-2 & \text{for } 1 < y < 2 \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

(2) Consider instead $Y = g(X) = 2X$. Calculate the density $h(y)$ of y. The transformation $Y = 2X$ maps $A = \{x : 0 < x < 1\}$ onto $B = \{y : 0 < y < 2\}$. Note that this transformation stretches the domain and changes the distance of the domain.

$$H(y) = \Pr(Y \leq y) = \Pr(2X \leq y) = \Pr\left(X \leq \frac{1}{2}y\right) = F\left(\frac{1}{2}y\right)$$

$$\begin{aligned} h(y) &= \frac{dH(y)}{dy} = \frac{d}{dy}F\left(\frac{1}{2}y\right) = f\left(\frac{1}{2}y\right) \cdot \frac{1}{2} \\ &= \begin{cases} \frac{1}{2}y & \text{for } 0 < y < 2 \\ 0 & \text{elsewhere} \end{cases} \end{aligned}$$

Remarks:

- What is the difference in the above two transformations? Any linear transformation can be decomposed into the two steps.
- We will consider two transformations separately for the multivariate transformations and combine the steps.

(1) Orthogonal Transformation

- First, we consider a random vector $\mathbf{Y} \in \mathbb{R}^k$ defined by $\mathbf{Y} = A\mathbf{X}$, where A is an orthogonal matrix.
- If A is orthogonal, $A'A = I$, $AA' = I$, and $A' = A^{-1}$.
- $\mathbf{Y} = A\mathbf{X}$ implies $\mathbf{X} = A'\mathbf{Y} = A^{-1}\mathbf{Y}$.
- Since A is orthogonal, the orthogonal mapping $\mathbf{x} \rightarrow \mathbf{y} = A\mathbf{x}$ maps a hypercube B to another hypercube C with the same shape and magnitude. An orthogonal mapping *rotates* the axes but leaves the angles and distances between vectors unchanged (next page).

- In $\mathbb{R}^k$, let the vectors $\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_k^{(1)})'$ and $\mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_k^{(2)})'$ represent two points. The squared distance between the two points is

$$D(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})'(\mathbf{x}^{(1)} - \mathbf{x}^{(2)}).$$

- If A is orthogonal, the transformation is distance-preserving:

$$\begin{aligned} D(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}) &= D(A\mathbf{x}^{(1)}, A\mathbf{x}^{(2)}) \\ &= (A\mathbf{x}^{(1)} - A\mathbf{x}^{(2)})'(A\mathbf{x}^{(1)} - A\mathbf{x}^{(2)}) \\ &= (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})'A'A(\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) \\ &= (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})'(\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) \\ &= D(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}). \end{aligned}$$

- Also, since the angles are determined by the lengths of its sides, the orthogonal transformation also preserves angles. An orthogonal transformation is distance-preserving as well as shape-preserving, and can be thought of as a change of coordinate axes in $\mathbb{R}^k$.

Theorem 1: Let $\mathbf{X}$ be a random vector in $\mathbb{R}^k$ with continuous distribution function $F(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$, and density $f(\mathbf{x})$. Let A be an orthogonal matrix and let $\mathbf{Y} = A\mathbf{X}$. The the distribution of $\mathbf{Y}$ has density $h(\mathbf{y})$ satisfying $h(\mathbf{y}) = f\left(A^{-1}\mathbf{y}\right) = f(A'\mathbf{y})$. □

(2) Diagonal Transformation

Next, let $\mathbf{X}$ and $f(\mathbf{x})$ be as before but let $\mathbf{Y} = \Lambda\mathbf{X}$, where Λ is a $k \times k$ diagonal matrix: $\Lambda = \text{diag}(\lambda_i)$, $\lambda_i \neq 0$, $i = 1, \ldots, k$. Then the distribution function of $Y = (Y_1, \ldots, Y_k)'$ is

$$\begin{aligned} H(\mathbf{y}) &= \Pr(Y_1 \leq y_1, \ldots, Y_k \leq y_k) \\ &= \Pr(\lambda_1 X_1 \leq y_1, \ldots, \lambda_k X_k \leq y_k) \\ &= \Pr\left(X_i \leq \frac{y_i}{\lambda_i}, \text{ if } \lambda_i > 0; X_i \geq \frac{y_i}{\lambda_i}, \text{ if } \lambda_i < 0, \ i = 1, \ldots, k\right). \end{aligned}$$

Hence, analogous to Theorem 4 of Lecture 2, the density of $\mathbf{Y}$ is

$$h(\mathbf{y}) = f\left(\frac{y_1}{\lambda_1}, \cdots, \frac{y_k}{\lambda_k}\right) |\lambda_1^{-1}| \cdots |\lambda_1^{-1}|.$$

Thus we have proved that:

Theorem 2: Let $\mathbf{X}$ and $f(\mathbf{x})$ be as before but let $\mathbf{Y} = \Lambda\mathbf{X}$, where Λ is a $k \times k$ diagonal matrix: $\Lambda = diag(\lambda_i)$, $\lambda_i \neq 0$, $i = 1, \cdots, k$. Then the distribution function of Y has density

$$h(\mathbf{y}) = f\left(\Lambda^{-1}\mathbf{y}\right)\left|\det \Lambda^{-1}\right|.$$

□

Remark on notation: $|\cdot|$ denotes the absolute value, det(·) denotes the determinant, and |det(·)| is the absolute value of the det(·).

(3) General Linear Transformation

Now, note that every nonsingular $k \times k$ matrix M can be written as (by the singular value decomposition)

$$M = P\Lambda Q,$$

where P and Q are orthonormal matrices ($PP' = QQ' = I$) and Λ is diagonal.

Theorem 3: Let $\mathbf{X}$ be a random vector in $\mathbb{R}^k$ with continuous distribution function $F(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$, and density $f(\mathbf{x})$, and let $\mathbf{Y} = M\mathbf{X}$, where M is a nonsingular $k \times k$ matrix. Then the distribution function of $\mathbf{Y}$ has density

$$h(\mathbf{y}) = f\left(M^{-1}\mathbf{y}\right)\left|\det M^{-1}\right|.$$

□

Theorem (Singular value decomposition, SVD): Given $n \times p$ matrix M, $n \geq p$, there exist an $n \times n$ orthonormal matrix P, a $p \times p$ orthonormal matrix Q, and an $n \times p$ matrix Λ consisting of a $p \times p$ diagonal positive semidefinite matrix and an $(n-p) \times p$ zero matrix such that $M = P\Lambda Q$.

Remark: Since P and Q are orthogonal, $|\det P| = |\det Q| = 1$, because $PP' = I$ and so $(\det P)(\det P') = (\det P)^2 = \det I = 1$. Thus

$$\det M = \det(P\Lambda Q) = (\det P)(\det \Lambda)(\det Q) = \det \Lambda$$

and $|\det M| = |\det \Lambda|$.

Sum and difference of two independent $unif\,orm$ random variables

Question for you:

Let $\mathbf{X} = (X_1\ X_2)'$ have joint density $f(\mathbf{x}) = 1$ for $0 < x_1 < 1$ and $0 < x_2 < 1$; zero elsewhere. Let $\mathbf{Y} = M\mathbf{X}$ where $\mathbf{Y} = (Y_1\ Y_2)'$ and $M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Find the joint density of $\mathbf{Y}$.

- $\mathbf{y} = M\mathbf{x}$:

$$\begin{aligned} y_1 &= x_1 + x_2 \\ y_2 &= x_1 - x_2 \end{aligned}$$

- $\mathbf{x} = M^{-1}\mathbf{y}$:

$$\begin{aligned} x_1 &= g_1^{-1}(y_1, y_2) = (y_1 + y_2)/2 \\ x_2 &= g_2^{-1}(y_1, y_2) = (y_1 - y_2)/2 \end{aligned}$$

- Let

$$\begin{aligned} S &= \{(x_1, x_2) : 0 < x_1 < 1 \text{ and } 0 < x_2 < 1\}, \\ T &= \{(y_1, y_2) : 0 < y_1 + y_2 < 2 \text{ and } 0 < y_1 - y_2 < 2\}. \end{aligned}$$

- $\det M = -2$, $|\det M| = 2$.
- By Theorem 3,

$$h(\mathbf{y}) = f\left(M^{-1}\mathbf{y}\right)\left|\det M^{-1}\right| = f(M^{-1}\mathbf{y}) \cdot \left|\frac{1}{2}\right| = \frac{1}{2}.$$

Hence, $h(\mathbf{y}) = \frac{1}{2}$ if $(y_1, y_2) \in T$; zero otherwise. Draw a figure for the new domain T (Exercise).

Remark: Now let B be a subset of S, and $C \subset T$ be the set onto which B is mapped by the one-to-one transformation. So the events $(X_1, X_2) \in B$ and $(Y_1, Y_2) \in C$ are equivalent. Therefore

$$\begin{aligned} \Pr\left[(Y_1, Y_2) \in C\right] &= \Pr\left[(X_1, X_2) \in B\right] \\ &= \int\int_B f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2. \end{aligned}$$

For this multivariate case, $\det M = -2$ and so

$$\begin{aligned}
\Pr\left[(Y_1, Y_2) \in C\right] &= \int\int_C h(\mathbf{y})\mathrm{d}\mathbf{y} = \int\int_C h(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2 \\
&= \int\int_C f\left(M^{-1}\mathbf{y}\right)\left|\det M^{-1}\right| \ \mathrm{d}y_1\mathrm{d}y_2 \\
&= \int\int_C 1 \cdot \frac{1}{2}\mathrm{d}y_1\mathrm{d}y_2
\end{aligned}$$

This implies that the joint density of (Y_1, Y_2) is $h(\mathbf{y}) = \frac{1}{2}$ if $(y_1, y_2) \in T$; zero otherwise.
Note that the effect of the value of the Jacobian determinant is to adjust the height of the density so that the probability over the new support T is unity. The probabilities of equivalent events in S and T are equal.

Question for you:

Consider the example in the previous few pages (sum and difference of two independent uniform random variables). Use a computer program to answer the following.

1. Find the singular value decomposition for $M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = P\Lambda Q$, i.e., find P, Λ, Q.
2. Show that $|\det M| = |\det \Lambda|$.
3. Show the densities of the three separate transformations via P, Λ, Q, respectively.
4. Draw the 3D graphs of $f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{if } \mathbf{x} \notin S \end{cases}$ and $h(\mathbf{y}) = \begin{cases} \frac{1}{2} & \text{if } \mathbf{y} \in T \\ 0 & \text{if } \mathbf{y} \notin T \end{cases}$.
5. Demonstrate, by the 3D graphs, how the three separate transformations via P, Λ, Q lead to the transformation by M.

Sum and difference of two independent $normal$ random variables

Question for you:

Let $\mathbf{X} = (X_1\ X_2)'$. Let X_1 and X_2 be independent, standard normal random variables. Let $\mathbf{Y} = M\mathbf{X}$ where $\mathbf{Y} = (Y_1\ Y_2)'$ and $M = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

1. Find the joint density of $\mathbf{Y}$.
2. Find the marginal distribution of Y_1 and Y_2. Use CB Theorem 4.2.14 and/or CB Theorem 4.2.12.
3. Show that Y_1 and Y_2 are independent. Use CB Lemma 4.2.7.
4. Use a computer program to draw the 3D graphs of $f(\mathbf{x})$ and $h(\mathbf{y})$.

Remark: (a) See CB Example 4.3.4. (b) Read CB p. 160. (c) See CB Exercise 4.27.

Example (CB Theorem 4.3.2) Distribution of the sum of Poisson variables:
If $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ and X and Y are independent, then $X + Y \sim \text{Poisson}(\theta + \lambda)$.

Proof: CB Example 4.3.1

3. Nonlinear Multivariate Transformations

Nonlinear Multivariate Transformation

So far, we have considered $Y = MX$, which is a linear transformation. More generally, we have:

Theorem: Let X as before and let $Y = g(X)$, where $g(\cdot)$ is a one-to-one mapping from $\mathbb{R}^k$ to $\mathbb{R}^k$ with differentiable inverse $x = g^{-1}(y)$. Then the distribution function of Y has density

$$h(y) = f\left(g^{-1}(y)\right)\left|\det(\partial g^{-1}(y)/\partial y)\right|.$$

□

Remark: CB equation (4.3.2) in p. 158

$$h(y) = f\left(g^{-1}(y)\right)|J|$$

where $J := \det(\partial g^{-1}(y)/\partial y)$ and $|J|$ is the absolute value of J.

- $J := \det(\partial g^{-1}(y)/\partial y)$ is called the Jacobian of the transformation. It is the determinant of a matrix of partial derivatives.
- The matrix of partial derivatives $\partial g^{-1}(y)/\partial y$ is defined as

$$\frac{\partial g^{-1}(y)}{\partial y} = \begin{bmatrix} \frac{\partial g_1^{-1}(y_1,\cdots,y_k)}{\partial y_1} & \cdots & \frac{\partial g_1^{-1}(y_1,\cdots,y_k)}{\partial y_k} \\ \vdots & & \vdots \\ \frac{\partial g_k^{-1}(y_1,\cdots,y_k)}{\partial y_1} & \cdots & \frac{\partial g_k^{-1}(y_1,\cdots,y_k)}{\partial y_k} \end{bmatrix},$$

  where $g_i^{-1}(y_1,\cdots,y_k)$ is the i-th component of $g^{-1}(y_1,\cdots,y_k)$.
- Read CB page 158 for $k = 2$.

Example (Nonlinear Bivariate Transformation):
Recall from LEC 3:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x = 1.$$

Question for you: How do we prove this? Use the nonlinear bivariate transformation... (next 3 pages)

$$
\begin{aligned}
& \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} \, \mathrm{d}x \right)^2 \\
= \quad & \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} \, \mathrm{d}x \right) \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} \, \mathrm{d}y \right) \\
= \quad & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} \, \mathrm{d}x \; \mathrm{d}y \\
= \quad & \int_{0}^{\infty} \int_{0}^{2\pi} e^{-\frac{1}{2}r^2} r \; \mathrm{d}\theta \; \mathrm{d}r \qquad \text{(Why?)} \\
= \quad & \left(\int_{0}^{2\pi} \mathrm{d}\theta \right) \int_{0}^{\infty} e^{-\frac{1}{2}r^2} r \; \mathrm{d}r \\
= \quad & 2\pi \int_{0}^{\infty} e^{-\frac{1}{2}r^2} r \; \mathrm{d}r \\
= \quad & 2\pi \left[-e^{r^2/2} \Big|_0^{\infty} \right] \\
= \quad & 2\pi.
\end{aligned}
$$

See CB p.104, line 1.
Why $dx\ dy = r\ dr\ d\theta$? In other words, why is r there?
It is the Jacobian of the polar transformation $g : (x, y) \to (r, \theta)$.

Consider the bivariate transformation $g : (x, y) \to (r, \theta)$ or $(x, y) = g^{-1}(r, \theta)$. Set $x = r\cos(\theta) = g_1^{-1}(r, \theta)$ and $y = r\sin(\theta) = g_2^{-1}(r, \theta)$. The Jacobian of the polar transformation $J = \det\left(\partial g^{-1}(r, \theta)/\partial(r, \theta)\right) = r$ is obtaineded as follows:

$$\begin{aligned}
\frac{\partial g^{-1}(r,\theta)}{\partial(r,\theta)} &= \begin{bmatrix} \frac{\partial g_1^{-1}(r,\theta)}{\partial r} & \frac{\partial g_1^{-1}(r,\theta)}{\partial \theta} \\ \frac{\partial g_2^{-1}(r,\theta)}{\partial r} & \frac{\partial g_2^{-1}(r,\theta)}{\partial \theta} \end{bmatrix}, \\
\frac{\partial(x,y)}{\partial(r,\theta)} &= \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \frac{\partial(r\cos(\theta))}{\partial r} & \frac{\partial(r\cos(\theta))}{\partial \theta} \\ \frac{\partial(r\sin(\theta))}{\partial r} & \frac{\partial(r\sin(\theta))}{\partial \theta} \end{bmatrix} \\
&= \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}.
\end{aligned}$$

Therefore the Jacobian of the transformation is

$$\det\left(\frac{\partial g^{-1}(r,\theta)}{\partial(r,\theta)}\right) = r,$$

and

$$\mathrm{d}x\ \mathrm{d}y = r\ \mathrm{d}r\ \mathrm{d}\theta.$$

Then the distribution function of (r,θ) has density

$$h(r,\theta) = f\left(g^{-1}(r,\theta)\right)\left|\det\left(\frac{\partial g^{-1}(r,\theta)}{\partial(r,\theta)}\right)\right| = f\left(g^{-1}(r,\theta)\right)\cdot r.$$

Digression. It is also good to know how to obtain $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

Note that

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx &= 1 \\ \int_0^\infty e^{-\frac{1}{2}x^2} dx &= \sqrt{\frac{\pi}{2}} \end{aligned}$$

implies that

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt = \sqrt{\pi},$$

with setting $t = \frac{1}{2}x^2$.

More examples

1. (CB Example 4.3.3) Distribution of the product of beta variables
2. (CB Example 4.3.6) Distribution of the ratio of normal variables

Discrete Case

So far, we have assumed that $F(x)$ is continuous, i.e., X is a continuous random vector. The discrete case differs from the continuous case in that the Jacobian of the transformation does not play a role. Thus, in discrete case, the previous theorem becomes:

Theorem: Let $X \in \mathbb{R}^k$ be a discrete random vector with density $f(x)$. Let $Y = g(X)$, where $g(\cdot)$ is a one-to-one mapping from $\mathbb{R}^k$ to $\mathbb{R}^k$ with differentiable inverse $x = g^{-1}(y)$. Then the distribution function of Y has density $h(y) = f\left(g^{-1}(y)\right)$. □

4. Joint, Marginal, and Conditional Distributions

Joint distribution

Any economy is a system which consists of different parts. These parts are generally related to each other. As a consequence, economic variables are interrelated. Given events A and B, their joint probability $\Pr(A \cap B)$ describe their relation. Such relation can be exploited to predict one using the other.

Definition [Random Vector]: An n-dimensional random vector, denoted as $Z = (Z_1, Z_2, \ldots, Z_n)'$, is a function from a sample space S into $\mathbb{R}^n$; the n-dimensional Euclidean space. For each outcome $s \in S$, $Z(s)$ is a n-dimensional real-valued vector and is called a realization of the random vector Z.

We often focus on bivariate probability distributions, which can illustrate most (but not all) essentials of multivariate probability distributions. We now consider two random variables (X, Y) in most of the subsequent discussion, where both X and Y are defined on the same probability space $(S, \mathcal{F}, P)$. A realization of (X, Y) will be a pair $(x, y) \in \mathbb{R}^2$.

Having defined a random vector (X, Y), we can now discuss probabilities of events that are defined in terms of (X, Y). How to characterize the joint probability distribution of X and Y?

Definition [Joint Cumulative Distribution Function]:

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y)$$

for any pair (x, y) in the xy-plane. □

Properties of $F_{XY}(x, y)$:
$F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0$
$F_{XY}(x, y)$ is non-decreasing in both x and y
$F_{XY}(x, y)$ is right continuous in both x and y

Theorem: $F_X(x) = F_{XY}(x, +\infty)$ and $F_Y(y) = F_{XY}(+\infty, y)$. □

Joint distribution

discrete case

Definition [Discrete Joint Distribution]: Let X and Y be two d.r.v., then their joint probability mass function (pmf) is defined as

$$f_{XY}(x, y) = \Pr(X = x, Y = y)$$

for any pair (x, y) in the xy-plane. □

Properties of $f_{XY}(x, y)$:
(i) $f_{XY}(x, y) \geq 0$ for all (x, y)
(ii) $\sum_x \sum_y f_{XY}(x, y) = 1$.

The pmf $f_{XY}(x, y)$ can be used to calculate the probability of any event defined in terms of (X, Y). For any subset A in the xy-plane, we have

$$\Pr\{(X, Y) \in A\} = \sum_{(x,y)\in A} f_{XY}(x, y).$$

The support of (X, Y) is defined as

$$\text{support}(X, Y) = \{(x, y) \in \mathbb{R}^2 : f_{XY}(x, y) > 0\}.$$

Joint distribution

continuous case

Definition [Continuous joint distribution]: Two random variables X and Y are said to have a continuous joint distribution if there exists a nonnegative function $f_{XY}(x, y)$ such that for any subset A on the xy-plane,

$$\Pr[(X, Y) \in A] = \int \int_{(x,y)\in A} f_{XY}(x, y)\mathrm{d}x\mathrm{d}y,$$

The function $f_{XY}(x, y)$ is called a joint probability density function. □

Properties: The function $f_{XY}(x, y)$ satisfies the two properties:
(i) $f_{XY}(x, y) \geq 0$ for all (x, y)
(ii) $\int \int_{(x,y)\in\mathbb{R}^2} f_{XY}(x, y)\mathrm{d}x\mathrm{d}y = 1$. (This follows from $F_{XY}(\infty, \infty) = 1$.)

Remarks: Selecting $A = \{(u, v) : u \leq x, v \leq y\}$ for any given pair (x, y), we have

$$\begin{aligned} \Pr[(X, Y) \in A] &= \Pr(X \leq x, Y \leq y) \\ &= F_{XY}(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v. \end{aligned}$$

Then at the points of (x, y) where $F_{XY}(x, y)$ is differentiable, we have

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \geq 0$$

as $F_{XY}(x, y)$ is nondecreasing in (x, y).

Marginal distributions

Discrete case

Definition [Discrete Marginal Distribution]: Suppose X and Y have a joint discrete distribution with joint pmf $f_{XY}(x, y)$. Then the marginal pmf's of X and Y are defined as

$$\begin{aligned} f_X(x) &= \Pr(X = x) = \sum_y f_{XY}(x, y), \\ f_Y(y) &= \Pr(Y = y) = \sum_x f_{XY}(x, y). \end{aligned}$$

Remarks: The marginal probability of X is the probability that X assumes a given value x regardless of the values taken by Y. By taking into account of all possibilities of X (or Y), we get rid of all the information of X (or Y).

Properties of $f_X(x)$ and $f_Y(y)$:

$f_X(x) \geq 0$, $\sum_x f_X(x) = 1$.

$f_Y(y) \geq 0$, $\sum_y f_Y(y) = 1$.

Question for you:

Suppose X and Y have the joint distribution

$$f_{XY}(x,y) = \frac{1}{5}|x+y|, \text{ for } x = -1, 0, 1 \text{ and } y = 0, 1.$$

Find (a) $f_X(x)$ and (b) $f_Y(y)$.

Solution: For $x = -1$, the event "$X = -1$" contains two basic outcomes: "$X = -1, Y = 0$" and "$X = -1, Y = 1$". These basic outcomes are mutually exclusive. Thus, it follows that

$$\begin{aligned} f_X(-1) &= \Pr(X = -1) = f_{XY}(-1, 0) + f_{XY}(-1, 1) = \frac{1}{5} \\ f_X(0) &= \Pr(X = 0) = f_{XY}(0, 0) + f_{XY}(0, 1) = \frac{1}{5} \\ f_X(1) &= \Pr(X = 1) = f_{XY}(1, 0) + f_{XY}(1, 1) = \frac{3}{5} \end{aligned}$$

Then

$$f_X(x) = \begin{cases} \frac{1}{5} & \text{if } x = -1 \\ \frac{1}{5} & \text{if } x = 0 \\ \frac{3}{5} & \text{if } x = 1. \end{cases}$$

Marginal distributions

Continuous case

Next, consider the case of c.r.v. Consider the cdf of X :

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(X \leq x, -\infty < Y < \infty) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(u, y) \mathrm{d}u \mathrm{d}y \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{XY}(u, y) \mathrm{d}y \right] \mathrm{d}u \\ &= \int_{-\infty}^{x} f_X(u) \mathrm{d}u, \end{aligned}$$

it follows by taking the derivatives of the above equation that

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}y.$$

Definition [Continuous Marginal Distribution]: Suppose X and Y have a joint continuous distribution with joint pdf $f_{XY}(x, y)$, then the marginal pdf's of X and Y are defined as

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}y, \text{ where } -\infty < x < \infty, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}x, \text{ where } -\infty < y < \infty. \end{aligned}$$

□

Properties of $f_X(x)$ and $f_Y(y)$:
$f_X(x) \geq 0$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
$f_Y(y) \geq 0$, $\int_{-\infty}^{\infty} f_Y(y)dy = 1$.

Example. $f_{XY}(x, y) = 4xy$ if $0 < x < 1$ and $0 < y < 1$. Then find $f_X(x)$ and $f_Y(y)$.

Solution. $f_X(x) = 2x$, $f_Y(y) = 2y$.

Question for you:

Compare CB Example 4.1.5 and CB Example 4.1.9. What do you learn?

Example (CB Example 4.1.11): Consider a joint pdf

$$f(x,y) = \begin{cases} 6xy^2 & 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Verify that $f(x,y)$ is the density.
- Find $\Pr(X + Y \geq 1)$.
- Find the marginal pdf's $f_X(x)$ of X and $f_Y(y)$ of Y.
- Find $\Pr(\frac{1}{2} < X < \frac{3}{4})$ and $\Pr(\frac{1}{2} < Y < \frac{3}{4})$.

Example (CB Example 4.1.12): Consider the joint pdf

$$f(x, y) = e^{-y}, \qquad 0 < x < y < \infty.$$

Find $\Pr(X + Y \geq 1)$.

Also, for this example:

- Next, we compute the conditional pdfs. (See also CB Example 4.2.4.)
- Are X and Y independent? Answer is No. Why?
- At the end of this lecture (LEC 4), we will also compute $\mathbb{E}(Y|X)$ and $var(Y|X)$.

Conditional distributions

Oftentimes when two random variables, (X, Y), are observed, the values of the two variables are related. Knowledge about the value of X gives us some information about the value of Y even if it does not tell us the value of Y exactly. How to characterize the relationship between X and Y? We use conditional probability distribution of Y given knowledge of the X value.
Recall $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$.

Conditional distributions

discrete case

Definition [Discrete Conditional Distributions]: Let X and Y have a joint discrete distribution with joint pmf $f_{XY}(x, y)$ and marginal pmf's $f_X(x)$ and $f_Y(y)$. Then the conditional pmf of X given $Y = y$ is defined as

$$f_{X|Y}(x|y) = \Pr(X = x|Y = y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

provided $f_Y(y) > 0$. If $f_Y(y) = 0$, we define $f_{X|Y}(x|y) = 0$ for all $y \in \mathbb{R}$.
Similarly, the conditional pmf of Y given $X = x$ is defined as

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

provided $f_X(x) > 0$. If $f_X(x) = 0$, we define $f_{Y|X}(y|x) = 0$ for all $x \in \mathbb{R}$. □

Properties of $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$:

(i) $f_{X|Y}(x|y) \geq 0$ for all x
(ii) $\sum_x f_{X|Y}(x|y) = 1$.

Similarly, for Y :
(i) $f_{Y|X}(y|x) \geq 0$ for all y
(ii) $\sum_y f_{Y|X}(y|x) = 1$.

$f_{X|Y}(x|y)$ is the probability that the random variable X will take any particular value x given that the value y of the random variable Y has been observed. If we put $A = \{X = x\}$ and $B = \{Y = y\}$, then

$$f_{X|Y}(x|y) = \Pr(X = x|Y = y) = \Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{f_{XY}(x, y)}{f_Y(y)}.$$

What happens if $f_Y(y) = 0$? In this case, the conditional pmf, $f_{X|Y}(x|y)$, is not well-defined, because it does not make any sense in practice to condition on something that is impossible to occur.

Different y's can be associated with different conditional pmf's for X. For example, Y can be a state variable taking two possible values: 1 and 2. When $Y = 1$ (a bear market), the distribution for the stock return X may have a large dispersion; when $Y = 0$ (a bull market), the distribution for the stock return X may have a small dispersion.

Conditional distributions

continuous case

Definition [Continuous Conditional Distribution]: Let X and Y have a joint continuous distribution with joint pdf $f_{XY}(x, y)$ and marginal pdf's $f_X(x)$ and $f_Y(y)$. Then the conditional pdf of X given $Y = y$ is defined as

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} \text{ if } f_Y(y) > 0.$$

If $f_Y(y) = 0$ we define $f_{X|Y}(x|y) = 0$ for all $x \in \mathbb{R}$. Similarly, we can define the conditional pdf of Y given $X = x$. □

Properties of $f_{X|Y}(x|y)$:
(i) $f_{X|Y}(x|y) \geq 0$ for all $x \in \mathbb{R}$
(ii) $\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$.
Similarly, for $f_{Y|X}(y|x)$.

Remark: Given each y, $f_{X|Y}(x|y)$ is a pdf for X. Different y's will give different distributions for X. Suppose $f_Y(y) > 0$. Then

$$\begin{aligned}\int_{-\infty}^{\infty} f_{X|Y}(x|y)\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{f_{XY}(x,y)}{f_Y(y)}\mathrm{d}x \\ &= \frac{1}{f_Y(y)}\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x \\ &= \frac{1}{f_Y(y)} f_Y(y) \\ &= 1.\end{aligned}$$

In other words, $f_{X|Y}(x|y)$ is a valid pdf given each y. Obviously, it is possible that there will be generally different pdf's for different values of y.

Granger causality: The conditional pmf/pdf $f_{X|Y}(x|y)$ is very useful because it tells how the information of y can be used to predict the probability of x.

Independence

- The marginal distributions of X and Y, described by the marginal pmfs/pdfs $f_X(x)$ and $f_Y(y)$ do not completely describe the joint distribution of X and Y. Indeed, there are many different joint distributions that have the same marginal distributions. Recall CB Example 4.1.9.
- Thus, it is hopeless to try to determine the joint pmf/pdf, $f_{XY}(x, y)$, from knowledge of only the marginal pmfs/pdfs, $f_X(x)$ and $f_Y(y)$.
- However, there is a special but important case in which we can use the marginal distributions to determine the joint distributions. This occurs when knowledge about X gives us no information about Y. This is the case of independence between X and Y.

Definition [Independence]: Two random variables X and Y are independent if and only if

$$F_{XY}(x, y) = F_X(x)F_Y(y) \text{ for all } -\infty < x, y < \infty$$

where $F_{XY}(x, y), F_X(x), F_Y(y)$ are the joint and marginal cdf's. $\square$

Remark: The above definition of independence is equivalent to the definition that two random variables X and Y defined on the same sample space are independent if

$$\Pr(X \in A, Y \in B) = \Pr(X \in A) \Pr(Y \in B)$$

for all subsets of A and B of real numbers. X and Y can be discrete or continuous.

Theorem: Two discrete random variables (X, Y) are independent if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \text{ for all } -\infty < x, y < \infty$$

where $f_{XY}(x, y), f_X(x), f_Y(y)$ are the joint and marginal pmf's. □

Theorem: Suppose X and Y are two continuous random variables. Then X and Y are independent if and only if

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v = \int_{-\infty}^{x} f_X(u) \mathrm{d}u \int_{-\infty}^{y} f_Y(y) \mathrm{d}y \text{ for all } (x, y),$$

where $f_{XY}(x, y), f_X(x), f_Y(y)$ are the joint and marginal pdf's. □

Proof: We first show $\Longrightarrow$. Suppose (X, Y) are independent, then

$$F_{XY}(x, y) = F_X(x)F_Y(y) \text{ for all } -\infty < x, y < \infty$$

Taking derivatives of both sides with respect to x and y respectively, we obtain

$$\frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y) = \frac{\partial}{\partial x} F_X(x) \frac{\partial}{\partial y} F_Y(y)$$

which implies

$$f_{XY}(x, y) = f_X(x)f_Y(y) \text{ for all } -\infty < x, y < \infty.$$

Next, we show $\Longleftarrow$. Suppose $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all $-\infty < x, y < \infty$. By integration,

$$\begin{aligned}\int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)\mathrm{d}u\mathrm{d}v &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{x} f_X(u)\mathrm{d}u \int_{-\infty}^{y} f_Y(v)\mathrm{d}v.\end{aligned}$$

Using proper integration ranges, we have

$$F_{XY}(x,y) = F_X(x)F_Y(y) \text{ for all } -\infty < x, y < \infty$$

which implies that X and Y are independent.

Theorem: The two random variables X and Y are independent if and only if the joint pmf/pdf can be written as

$$f_{XY}(x,y) = g(x)h(y) \text{ for all } -\infty < x, y < \infty.$$

□

Proof: First we show $\Longrightarrow$. If X and Y are independent, then

$$\begin{aligned} f_{XY}(x,y) &= f_X(x)f_Y(y) \text{ for all } -\infty < x, y < \infty \\ &= g(x)h(y) \end{aligned}$$

where $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.

Next, we show $\Longleftarrow$. Suppose $f_{XY}(x,y) = g(x)h(y)$ for all $-\infty < x, y < \infty$. Then

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y = \int_{-\infty}^{\infty} g(x)h(y)\mathrm{d}y = g(x)\int_{-\infty}^{\infty} h(y)\mathrm{d}y \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x = \int_{-\infty}^{\infty} g(x)h(y)\mathrm{d}x = h(y)\int_{-\infty}^{\infty} g(x)\mathrm{d}x \end{aligned}$$

and

$$\begin{aligned} f_X(x)f_Y(y) &= g(x)h(y)\int_{-\infty}^{\infty} g(u)\mathrm{d}u\int_{-\infty}^{\infty} h(v)\mathrm{d}v \\ &= g(x)h(y)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)h(v)\mathrm{d}u\mathrm{d}v \\ &= g(x)h(y)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(u,v)\mathrm{d}u\mathrm{d}v \\ &= g(x)h(y) \\ &= f_{XY}(x,y), \end{aligned}$$

which implies that X and Y are independent. See CB Lemma 4.2.7.

Example (CB Example 4.2.8): Consider the joint density $f_{XY}(x, y) = \frac{1}{384}x^2y^4e^{-y-(x/2)}$, $x > 0$, $y > 0$. Are X and Y independent?

Remark: See CB page 154, paragraph before Example 4.2.9.

Example: Consider the joint density $f_{XY}(x,y) = 8xy$, $0 \le x \le y \le 1$. Are X and Y independent?
Solution: $f_X(x) = 4x(1-x^2)$, $0 \le x \le 1$, and $f_Y(y) = 4y^3$, $0 \le y \le 1$. Thus, X and Y are not independent since $f_{XY}(x,y) \neq f_X(x) f_Y(y)$.
Example: Consider the joint pdf $f(x,y) = e^{-y}$, $0 < x < y < \infty$. Are X and Y independent?
Solution: *Question for you:*
Example: Suppose X and Y have a joint pmf or pdf $f_{XY}(x,y)$ that can be decomposed as the product of two functions $h(x)g(y)$, where $h(x)$ and $g(y)$ are not necessarily a pmf or pdf of X and Y (i.e., $f_{XY}(x,y) = h(x)g(y)$ for all x, y). Are X and Y independent? Answer: Yes.

5. Mathematical Expectations

Mathematical Expectations

Definition 1: Let $g(\cdot)$ be a measurable real function on $\mathbb{R}^k$ and let X be a random vector in $\mathbb{R}^k$ with distribution function F. Then the mathematical expectation of $g(X)$ is defined by
$\mathbb{E}g(X) = \int g(x)\, dF(x)$. □

Moments of Distribution

Let X be a random variable with distribution function $F(x)$.

Definition 2: The m-th moment of X is defined as $\mathbb{E}X^m = \int x^m dF(x)$ provided that $\mathbb{E}|X|^m < \infty$.

Definition 3: The first $(m = 1)$ moment of X, $\mathbb{E}X$, is called the *mean* of (the distribution of) X. This is usually denoted by μ_X.

Definition 4: The *variance* of X is $\mathbb{E}(X - \mu_X)^2$. It is usually denoted by σ_X^2 or by $Var(X)$. The square root of σ_X^2 (thus σ_X) is called the *standard deviation* of X.

Next, let Y be another random variable defined on the same probability space as X.

Definition 5: The *covariance* of X and Y is $Cov(X,Y) = \mathbb{E}\left(X - \mu_X\right)\left(Y - \mu_Y\right)$, which is also denoted by σ_{XY}.

Theorem 1 [CB 4.5.3]. Show that $\sigma_{XY} = \mathbb{E}(XY) - \mu_X\mu_Y$.

Theorem 2 [CB 4.5.5]. If X and Y are independent, then $Cov(X,Y) = 0$. □

The converse of Theorem 2, however, is not true. If $Cov(X, Y) = 0$ then X and Y need not be independent. Here are 3 counter-examples.

Example A: Let X be such that $\mathbb{E}(X) = 0$, $\mathbb{E}(X^3) = 0$, and let $Y = X^2$. Then

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2) = 0$$

whereas X and Y are clearly not independent.

Example B [CB p. 171]: If $X \sim f(x - \theta)$, symmetric around 0 with $\mathbb{E}(X) = \theta$, and $Y = \mathbf{1}(|X - \theta| < 2)$, then X, Y are not independent but

$$\begin{aligned} \mathbb{E}(XY) &= \int_{-\infty}^{\infty} x\mathbf{1}(|x - \theta| < 2) f(x - \theta)\mathrm{d}x \\ &= \int_{-2}^{2} (t + \theta) f(t)\mathrm{d}t = \theta \int_{-2}^{2} f(t)\mathrm{d}t = \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Example C [CB Example 4.5.9].

Definition 6: The quantity $\rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$ is called the *correlation coefficient* between X and Y.

Theorem 3 [CB 4.5.7a]: $-1 \leq \rho_{XY} \leq 1$. □

Proof: Consider a random variable $(U - kV)^2$ where U and V are random variables and $k \in \mathbb{R}$. Since

$$\mathbb{E}(U - kV)^2 = \mathbb{E}U^2 - 2k\mathbb{E}(UV) + k^2\mathbb{E}V^2 \geq 0$$

for all $k \in \mathbb{R}$, the discriminant of this second order quadratic equation is

$$D/4 = (\mathbb{E}UV)^2 - (\mathbb{E}U^2) \cdot (\mathbb{E}V^2) \leq 0,$$

which is called the Cauchy-Schwartz inequality (CB Theorem 4.7.3, CB Example 4.7.4). Let $U = X - \mu_X$ and $V = Y - \mu_Y$. Then

$$\begin{aligned} [Cov(X,Y)]^2 &\leq Vax(X) \cdot Var(Y). \\ \sigma_{XY}^2 &\leq \sigma_X^2 \cdot \sigma_Y^2 \\ \rho_{XY}^2 &\leq 1 \\ -1 &\leq \rho_{XY} \leq 1. \end{aligned}$$

□

Theorem 4 [CB 4.5.7b]: (a) If $|\rho_{XY}| = 1$, then $Y = \alpha + \beta X$ for some constants α and β. (b) If $\rho_{XY} = 1$, then $\beta > 0$. If $\rho_{XY} = -1$, then $\beta < 0$. □

Proof. (a) If $\rho_{XY}^2 = 1$, then from the above proof $\mathbb{E}(U - kV)^2 = 0$, i.e., $U = kV$, or $(X - \mu_X) = k(Y - \mu_Y)$. Thus $X = kY + (\mu_X - k\mu_Y)$. Then $\alpha = -\frac{1}{k}\mu_X + \mu_Y$ and $\beta = \frac{1}{k}$.
(b) Solving the quadratic equation $\mathbb{E}(U - kV)^2 = \mathbb{E}U^2 - 2k\mathbb{E}(UV) + k^2\mathbb{E}V^2 = 0$ for k,

$$k = \frac{\mathbb{E}(UV) \pm D/4}{\mathbb{E}V^2} = \frac{\mathbb{E}(UV)}{\mathbb{E}V^2} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\frac{\sigma_Y}{\sigma_X} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}.$$

As $k = \rho_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{1}{\beta}$, the result follows.

Next, let $X_1, X_2, X_3, \cdots$ be a sequence of random variables defined on a common probability space.

Definition 7: The sequence $\{X_j\}$ is said to be *uncorrelated* if $Cov(X_i, X_j) = 0$ for all $i \neq j$.

Theorem 5: If $X_1, \cdots, X_n$ be uncorrelated, $Var\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n Var(X_j)$. □

Proof.

$$\begin{aligned} Var\left(\sum_{j=1}^n X_j\right) &= \mathbb{E}\left(\sum_{j=1}^n X_j - \mathbb{E}\left(\sum_{j=1}^n X_j\right)\right)^2 \\ &= \sum_{j=1}^n \mathbb{E}\left(X_j - \mathbb{E}(X_j)\right)^2 + 2\sum_{i\neq j} Cov(X_i, X_j) \\ &= \sum_{j=1}^n Var(X_j). \end{aligned}$$

See CB Exercise 4.44.

Mean and Covariance of a Random Vector

Let $X = (X_1, \cdots, X_k)'$ be a random vector in $\mathbb{R}^k$.
Definition 8: The mean vector of X, denoted by $\mathbb{E}X$, is the vector of the means of $X_1, \cdots, X_k$:

$$\mathbb{E}X = (\mathbb{E}X_1, \cdots, \mathbb{E}X_k)'.$$

Definition 9: The covariance matrix of X, denoted by $Var(X)$, is the matrix of covariances of the components of X :

$$Var(X) = \begin{pmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_k) \\ \vdots & & \vdots \\ Cov(X_k, X_1) & \cdots & Cov(X_k, X_k) \end{pmatrix}.$$

Remark. $Var(X)$ is symmetric.

Theorem 6: Let X be a random vector in $\mathbb{R}^k$ with mean vector μ and covariance matrix Σ. Let

$$Y = BX + b, \qquad Y \in \mathbb{R}^m,$$

where B is an $m \times k$ matrix with nonrandom elements and b is a nonrandom vector in $\mathbb{R}^m$. Then Y has mean vector $B\mu + b$ and covariance matrix $B\Sigma B'$. □

Theorem 7: $Var\,(X)$ is always positive semidefinite. If $Var(X)$ is nonsingular, then it is positive definite. □

Theorem 8: If $Var\,(X)$ is nonsingular, then its determinant is positive. □

For proofs, see e.g., Dhrymes (1974).

Let

$$\mathbb{E}XX' = \begin{pmatrix} \mathbb{E}X_1X_1 & \cdots & \mathbb{E}X_1X_k \\ \vdots & & \vdots \\ \mathbb{E}X_kX_1 & \cdots & \mathbb{E}X_kX_k \end{pmatrix}.$$

Then we have

Theorem 9: $\mathbb{E}XX' = Var\,(X) + (\mathbb{E}X)(\mathbb{E}X)'$. □

Expectations under bivariate distributions

Definition [Expected Value]: Let $g(X, Y)$ be a real valued function. Then the expected value of $g(X, Y)$ is defined as

$$\begin{aligned} \mathbb{E}g(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)dF_{XY}(x,y) \\ &= \begin{cases} \sum_x \sum_y g(x,y) f_{XY}(x,y) & \text{d.r.v.} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)dxdy & \text{c.r.v.} \end{cases} \end{aligned}$$

provided the double sum or the double integral exists.
Remark: Like the univariate case, we say that $\mathbb{E}[g(X, Y)]$ exists if $\mathbb{E}|g(X, Y)| < \infty$.

$\mathbb{E}g(X,Y)$

- $g(X,Y) = X$: The mean of X is $\mu_X = \mathbb{E}(X)$.
- $g(X,Y) = X^k$: The k-th moment of X is $\mu_k = \mathbb{E}(X^k)$.
- $g(X,Y) = (X - \mu_X)^2$: The variance of $X = \sigma_X^2 = \mathbb{E}(X^2) - \mu_X^2$.
- $g(X,Y) = X^r Y^s$: The rth and sth product moment of (X,Y) about the origin.
- $g(X,Y) = (X - \mu_X)(Y - \mu_Y)$: Covariance

When X and Y are not independent, we say that there exists a relationship between them. However, if there is a relationship, the relationship may be weak or strong. How to measure the strength of a relationship between X and Y?

Covariance: Suppose $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. The covariance between two variables X and Y is defined as

$$cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) dF_{XY}(x, y).$$

Correlation: Suppose $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. The correlation between two variables X and Y is defined as

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation coefficient is the standardized covariance. This is analogous to the definitions of skewness and kurtosis.

Theorem: $|\rho_{XY}| \leq 1$

Theorem: When $X = aY + b, a > 0$, then $\rho_{XY} = 1$. When $X = aY + b, a < 0$, then $\rho_{XY} = -1$.

Remarks:

- When there is a perfect linear relationship between X and Y, we always have $\rho_{XY} = \pm 1$.
- Suppose we have $X = aY + b + u$, where u is a random variable with $\mathbb{E}(u) = 0$, $var(u) = \sigma_u^2 > 0$, and it is orthogonal to Y in the sense $cov(Y, u) = 0$. Then $|\rho_{XY}| < 1$. The deviation from unity depends on the magnitude of σ_u^2. Find the expression of ρ_{XY} in this case?
- Correlation does not necessarily imply causality.

Question for you:

CB Example 4.5.4: Let the joint pdf of (X, Y) be $f(x, y) = 1, 0 < x < 1, x < y < x + 1$. Find ρ_{XY}.

$f_X(x) = 1,\ 0 < x < 1$

$\mu_X = \frac{1}{2},\ \sigma_X^2 = \frac{1}{12}$

$$f_Y(y) = \begin{cases} y, & 0 < y < 1 \\ 2 - y, & 1 \leq y < 2 \end{cases}$$

$\mu_Y = 1, \sigma_Y^2 = \frac{1}{6}$

$\mathbb{E}(XY) = \int_0^1 \int_x^{x+1} xy \, \mathrm{d}y\mathrm{d}x = \frac{7}{12}$

$\sigma_{XY} = \frac{7}{12} - \left(\frac{1}{2}\right)(1) = \frac{1}{12}$

$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{1/12}{\sqrt{1/12}\sqrt{1/6}} = \frac{1}{\sqrt{2}}$

CB Example 4.5.8: Let the joint pdf of (X, Y) be $f(x, y) = 10,\ 0 < x < 1, x < y < x + \frac{1}{10}$.

1. Find the marginal distribution of X.
2. Find the marginal distribution of Y.
3. Show that $\rho_{XY} = \sqrt{\frac{100}{101}}$.

CB Example 4.5.9: Let the joint pdf of (X, Y) be $f(x, y) = 5,\ -1 < x < 1, x^2 < y < x^2 + \frac{1}{10}$. Show that $\rho_{XY} = 0$.

Question for you to think about: : Compare the above two examples (CB Example 4.5.8 and CB Example 4.5.9). They look similar in some sense, but the values of ρ_{XY} are totally different. What makes this difference? It shows the limitation of the linear concept. Nevertheless, we consider more on a linear model based on ρ_{XY}, as discussed next:

Theorem [linear regression model]: Suppose X and Y are two random variables with finite second moments. Consider a linear regression function

$$Y = g(X) + u,$$

where $g(X) = \alpha + \beta X$, and u is called the prediction error. Then the optimal coefficient $(\alpha^* \ \beta^*)$ that minimizes the mean square error

$$MSE(\alpha, \beta) = \mathbb{E}[Y - g(X)]^2 = \mathbb{E}[Y - (\alpha + \beta X)]^2$$

is given by

$$\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \begin{bmatrix} 1 & \mu_X \\ \mu_X & \mathbb{E}X^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_Y \\ \mathbb{E}XY \end{bmatrix}.$$

In particular,

$$\begin{aligned} \alpha^* &= \mu_Y - \frac{cov(X,Y)}{var(X)}\mu_X, \\ \beta^* &= \frac{cov(X,Y)}{var(X)}. \end{aligned}$$

Proof: Find the first order conditions (FOC) and the second order conditions (SOC).
FOC:

$$\begin{aligned} \frac{\partial}{\partial \alpha} MSE(\alpha, \beta) &= 0 \\ \frac{\partial}{\partial \beta} MSE(\alpha, \beta) &= 0 \end{aligned}$$

From these two equations, we can solve for α^* and β^*.

Linear regression analysis is a covariance analysis: β^* is proportional to $cov(X, Y)$. This is always so no matter whether the true relationship between X and Y is linear or nonlinear.

Many economic and financial theories are based on correlation analysis.

Example [Capital Asset Pricing Model (CAPM)]: Let R_p be the return on a portfolio during a certain horizon, r_f is the risk-free interest rate, R_m is the return on the market portfolio (i.e., the return on the S&P500 index). Then CAPM says that

$$\mathbb{E}(R_p - r_f) = \beta \mathbb{E}(R_m - r_f),$$

where $R_p - r_f$ is called the excess return on portfolio, $R_m - r_f$ is the excess return on the market portfolio, and the investment beta

$$\beta = \frac{cov(R_p - r_f, R_m - r_f)}{var(R_m - r_f)}.$$

β is called "investment beta", which is a measure of undiversifiable systematic risk.

- If $\beta = 1$, the portfolio is equally risky to the market portfolio.
- If $\beta > 1$, the portfolio is more risky than the market portfolio.
- If $\beta < 1$, the portfolio is less risky than the market portfolio.

Example [Phillip's Curve]: $cov(X, Y) < 0$ where $X =$ inflation rate, $Y =$ unemployment rate.

$\mathbb{E}g(X,Y)$ with $g(X,Y) = aX + bY + c$:

Theorem [CB 4.5.6]: Suppose $Z = aX + bY + c$. Then

$$\begin{aligned} E(Z) &= a\mu_X + b\mu_Y + c \\ Var(Z) &= a^2 Var(X) + b^2 Var(Y) + 2ab \cdot Cov(X,Y). \end{aligned}$$

□

Remarks:

(a) It implies $var(X+Y) > var(X) + var(Y)$ if $cov(X,Y) > 0$. When $cov(X,Y) > 0$, the variation of the sum is greater than the sum of variation. This is because both X and Y move in the same direction, and as a result, there are more chances that $X + Y$ tends to be extremely large or small.

(b) In contrast, when $cov(X,Y) < 0$, the variation of the sum is less than the sum of individual variations. This is because both X and Y move in opposite direction so that $X + Y$ will not vary much.

(c) When X and are uncorrelated, then $var(aX + bY) = a^2 var(X) + b^2 var(Y)$.

Conditional Distributions, Conditional Densities, and Conditional Expectations

For convenience, let $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. Let $F_{XY}(x, y)$ be the joint CDF and let $F_X(x)$, $F_Y(y)$ be the marginals. We shall derive the conditional distribution function of X given the event that $Y = y$. If $P(Y = y) > 0$, then the conditional distribution of X given $Y = y$ is

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \frac{\Pr(X \leq x \text{ and } Y = y)}{\Pr(Y = y)}.$$

Next, suppose that $F_{X|Y}(x, y)$ is continuous with density $f_{X|Y}(x, y)$. Moreover, let $f_Y(y)$ be the marginal density, i.e.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx.$$

Then $f_Y(y)$ is the density of the marginal distribution

$$F_Y(y) = F_{XY}(\infty, y).$$

Furthermore, the conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{\partial F_{X|Y}(x|y)}{\partial x} = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Proof:

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{\partial F_{X|Y}(x|y)}{\partial x} \\
&= \frac{\partial}{\partial x} \lim_{\varepsilon \to 0} \Pr(X \leq x | y \leq Y \leq y + \varepsilon) \\
&= \frac{\partial}{\partial x} \lim_{\varepsilon \to 0} \frac{\Pr\{(X \leq x) \cap (y \leq Y \leq y + \varepsilon)\}}{\Pr(y \leq Y \leq y + \varepsilon)} \\
&= \frac{\partial}{\partial x} \lim_{\varepsilon \to 0} \frac{F_{XY}(x, y+\varepsilon) - F_{XY}(x,y)}{F_Y(y+\varepsilon) - F_Y(y)} \\
&= \frac{\partial}{\partial x} \lim_{\varepsilon \to 0} \frac{[F_{XY}(x, y+\varepsilon) - F_{XY}(x,y)]/\varepsilon}{[F_Y(y+\varepsilon) - F_Y(y)]/\varepsilon}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial x} \frac{\lim_{\varepsilon \to 0} \left[F_{XY}\left(x, y+\varepsilon\right) - F_{XY}\left(x, y\right) \right] / \varepsilon}{\lim_{\varepsilon \to 0} \left[F_{Y}(y+\varepsilon) - F_{Y}\left(y\right) \right] / \varepsilon} \\
&= \frac{\partial}{\partial x} \frac{\partial F_{XY}\left(x, y\right) / \partial y}{f_{Y}(y)} \\
&= \frac{\partial^2 F_{XY}\left(x, y\right) / \partial y \partial x}{f_{Y}(y)} \\
&= \frac{f_{XY}\left(x, y\right)}{f_{Y}\left(y\right)}.
\end{aligned}
$$

Remark (notation): $F_{XY}(x,y)$, $F_X(x)$, $F_Y(y)$, $F_{X|Y}(x|y)$, $F_{Y|X}(y|x)$, $f_{XY}(x,y)$, $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$, $f_{Y|X}(y|x)$.

In discrete case, the same applies. Denoting

$$\begin{aligned} f_{XY}(x,y) &= \Pr(X = x \text{ and } Y = y) \\ f_Y(y) &= \Pr(Y = y), \end{aligned}$$

we have

$$f_{X|Y}(x|y) = \frac{\Pr(X = x \text{ and } Y = y)}{\Pr(Y = y)} = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Conditional Expectation

We have seen that for a measurable function g, the mathematical expectation $\mathbb{E}\left(g\left(X\right)\right)$ can be written as a Stieltje's integral $\int g\left(x\right) dF_X\left(x\right)$, F is the distribution function of X. The same applies to conditional expectations. Thus if X and Y are discrete or continuous, we have

$$\begin{aligned}
\mathbb{E}\left(g\left(X\right)|Y=y\right) &= \int g\left(x\right) dF_{X|Y}\left(x|y\right) \\
&= \begin{cases} \int g\left(x\right) f_{X|Y}\left(x|y\right) dx & \text{in continuous case} \\ \sum_{x\in C} g\left(x\right) f_{X|Y}\left(x|y\right) & \text{in discrete case} \end{cases}
\end{aligned}$$

provided that $f_Y\left(y\right) > 0$.

Finally, if we replace the non-random y on the right-hand side above by Y, we get

$$\begin{aligned}\mathbb{E}\left(g\left(X\right)|Y\right) &= \int g\left(x\right) dF_{X|Y}\left(x|Y\right) \\ &= \begin{cases} \int g\left(x\right) f_{X|Y}\left(x|Y\right) dx & \text{in continuous case} \\ \sum_{x\in C} g\left(x\right) f_{X|Y}\left(x|Y\right) & \text{in discrete case} \end{cases}\end{aligned}$$

which is a random variable.

Exercise. Show that if X and Y are independent then $\mathbb{E}\left(g\left(X\right)|Y\right) = \mathbb{E}\left(g\left(X\right)\right).$

Solution. $\mathbb{E}\left(g\left(X\right)|Y\right) = \int g\left(x\right) f_{X|Y}\left(x|Y\right) dx =$ $\int g\left(x\right) f_X\left(x\right) dx = \mathbb{E}\left(g\left(X\right)\right)$, because $f_{X|Y}\left(x|y\right) = \frac{f_{XY}(x,y)}{f_Y(y)} =$ $\frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X\left(x\right).$

Characteristic functions of multivariate distributions

Let $\mathbf{X}$ be a random vector on $\mathbb{R}^k$ with distribution function $F(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$.

Definition: The characteristic function of F is $\varphi(\mathbf{t}) = \mathbb{E}e^{i\mathbf{t}'\mathbf{x}}$, where $\mathbf{t} \in \mathbb{R}^k$. □

Theorem: If F_1 and F_2 are distributions on $\mathbb{R}^k$ with the same characteristic functions φ_1 and φ_2, then $F_1 \equiv F_2$ if and only iff $\varphi_1 \equiv \varphi_2$. □

Theorem: Let $\mathbf{X}$ be a random vector on $\mathbb{R}^k$ and let $\mathbf{Y}$ be a random vector on $\mathbb{R}^m$. $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if for all $\mathbf{t}_1 \in \mathbb{R}^k$, $\mathbf{t}_2 \in \mathbb{R}^m$, $\mathbb{E}e^{(i\mathbf{t}_1'\mathbf{X} + i\mathbf{t}_2'\mathbf{Y})} = \mathbb{E}e^{i\mathbf{t}_1'\mathbf{X}}\mathbb{E}e^{i\mathbf{t}_2'\mathbf{Y}}$. □

6. What is the relationship between uncorrelatedness and independence?

Theorem A: Suppose $g(X,Y) = h(X)q(Y)$ and (X,Y) are independent. Then

$$\mathbb{E}g(X,Y) = \mathbb{E}h(X) \cdot \mathbb{E}q(Y).$$

Proof:

$$\begin{aligned} \mathbb{E}g(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)dF_{XY}(x,y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(X)q(Y)dF_X(x)dF_Y(y) \\ &= \mathbb{E}h(X) \cdot \mathbb{E}q(Y). \end{aligned}$$

Corollary: If X and Y are independent, then $cov(X,Y) = 0$.
Proof: $h(X) = X - \mu_X$ and $q(Y) = Y - \mu_Y$.

Corollary: If X and Y are independent and their marginal mgfs exist for t in a neighborhood of 0, then

$$M_{X+Y}(t) = M_X(t)M_Y(t) \text{ for all } t \text{ in a neighborhood of } (0,0).$$

□

Proof: Applying the above theorem, we have

$$g(X,Y) = e^{t(X+Y)} = e^{tX}e^{tY} = h(X)q(Y).$$

It follows that if X and Y are independent, then

$$\begin{aligned} \mathbb{E}[e^{t(X+Y)}] &= \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}), \\ M_{X+Y}(t) &= M_X(t)M_Y(t). \end{aligned}$$

CB Theorem 4.2.12.
Remark: This property of mgf is rather useful in characterizing the distribution for some random variable which itself is the sum of other independent random variables.

Example: Suppose $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and X, Y are independent. Then

$$X \pm Y \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2).$$

Proof: The mgfs of X and Y are

$$\begin{aligned} M_X(t) &= \mathbb{E}e^{tX} = \exp\left(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right), \\ M_Y(t) &= \mathbb{E}e^{tY} = \exp\left(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right). \end{aligned}$$

By independence,

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) M_Y(t) \\ &= \exp\left(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right) \cdot \exp\left(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right) \\ &= \exp\left((\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2) t^2\right). \end{aligned}$$

It follows that

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

That is, *the sum of two independent normals is also normal*. This result can be extended to many independent normal random variables. See the next section on the multivariate normal distribution. See CB Example 4.2.13 and CB Theorem 4.2.14.

Now, consider the converse of Theorem A.

Theorem B: Suppose we have $g(X,Y) = h(X)q(Y)$ and

$$\mathbb{E}g(X,Y) = \mathbb{E}h(X) \cdot \mathbb{E}q(Y)$$

for any measurable $h(X)$ and $q(Y)$. Then (X,Y) are independent.

□

Consider $h(X) = 1(X \leq x)$, $q(Y) = 1(Y \leq y)$ for any given (x,y).

$$\begin{aligned} \mathbb{E}h(X) &= \mathbb{E}1(X \leq x) = F_X(x), \\ \mathbb{E}q(Y) &= \mathbb{E}1(Y \leq y) = F_Y(y), \\ \mathbb{E}h(X)q(Y) &= \mathbb{E}1(X \leq x)1(Y \leq y) = F_{XY}(x,y). \end{aligned}$$

Suppose $\mathbb{E}h(X)q(Y) = \mathbb{E}h(X) \cdot \mathbb{E}q(Y)$. Then

$$F_{XY}(x,y) = F_X(x)F_Y(y).$$

It follows that X and Y must be independent. We can use mgfs to characterize independence.

Theorem: Suppose $M_{XY}(t_1, t_2) = \mathbb{E}e^{t_1 X + t_2 Y}$ exists for t_1, t_2 in some neighborhood of the origin. Then (X, Y) are independent if and only if

$$M_{XY}(t_1, t_2) = M_X(t_1) M_Y(t_2),$$

for all (t_1, t_2) in the neighborhood of $(0, 0)$. □

Proof: Suppose (X, Y) are independent, then

$$F_{XY}(x, y) = F_X(x)F_Y(y).$$

By definition

$$\begin{aligned}
M_{XY}(t_1, t_2) &= \mathbb{E}e^{t_1X+t_2Y} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1X+t_2Y} dF_{XY}(x, y) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1X+t_2Y} dF_X(x)dF_Y(y) \\
&= \int_{-\infty}^{\infty} e^{t_1X} dF_X(x) \cdot \int_{-\infty}^{\infty} e^{t_2Y} dF_Y(y) \\
&= \mathbb{E}e^{t_1X} \cdot \mathbb{E}e^{t_2Y} \\
&= M_X(t_1)M_Y(t_2).
\end{aligned}$$

- In the previous section, we have learned limitations of using the correlation ρ_{XY} (a linear measure, of uncorrelatedness) in charaterizing independence.
- We may like a generalization, which permits the correlation to caputer nonlinear dependence.
- An alternative representation of the mgf-based characterization of correlation, $\sigma_{XY}(t_1, t_2) \equiv cov(e^{t_1 X}, e^{t_2 Y})$, does provide such a generalization, to capture not only linear correlation but also nonlinear dependence.
- It is to measure the departure from the equality of the last theorem

$$M_{XY}(t_1, t_2) = M_X(t_1) M_Y(t_2).$$

- It turns out (next page):

$$\sigma_{XY}(t_1, t_2) \equiv cov(e^{t_1 X}, e^{t_2 Y}) = M_{XY}(t_1, t_2) - M_X(t_1) M_Y(t_2)$$

Theorem: Suppose $M_{XY}(t_1, t_2)$ exists. Then (X, Y) are independent if and only if

$$\sigma_{XY}(t_1, t_2) \equiv cov(e^{t_1 X}, e^{t_2 Y}) = 0$$

for all (t_1, t_2) in the neighborhood of $(0, 0)$. □

Proof: Using the formula $cov(U, V) = \mathbb{E}(UV) - \mathbb{E}(U)\mathbb{E}(V)$,

$$\begin{aligned} \sigma_{XY}(t_1, t_2) &\equiv cov(e^{t_1 X}, e^{t_2 Y}) \\ &= \mathbb{E}(e^{t_1 X} e^{t_2 Y}) - \mathbb{E}(e^{t_1 X})\mathbb{E}(e^{t_2 Y}) \\ &= M_{XY}(t_1, t_2) - M_X(t_1) M_Y(t_2). \end{aligned}$$

Theorem: Suppose $M_{XY}(t)$ exists in a neighborhood of $(0,0)$. Then

$$\sigma_{XY}^{(1,1)}(0,0) = \frac{\partial^2}{\partial t_1 \partial t_2}\sigma_{XY}(t_1,t_2)\Big|_{t=(0,0)} = cov(X,Y).$$

Moreover,

$$\sigma_{XY}^{(k,l)}(0,0) = \frac{\partial^{k+l}}{\partial^k t_1 \partial^l t_2}\sigma_{XY}(t_1,t_2)\Big|_{t=(0,0)} = cov(X^k,Y^l).$$

□

Remarks:

- $\sigma_{XY}(t_1,t_2)$ can be viewed as a covariance generating function.
- If X and Y are independent, then $\sigma_{XY}^{(k,l)}(0,0) = cov(X^k,Y^l) = 0$ for all $k,l > 0$. In general, $cov(X,Y) = 0$ is only one of an infinite set of implications for independence.
- *Question for you:* Suppose $cov(X^k,Y^l) = 0$ for all $k,l > 0$. Are X and Y independent?

Theorem: Suppose (X, Y) are independent. Then $cov(X, Y) = 0$. However, the converse is not true. □

Remark: $\rho_{XY} = 0$ does not imply that X and Y are independent. Because ρ_{XY} is a measure of the linear association between X and Y, it is often referred to as the linear correlation coefficient. And, to avoid confusion, the word "linearly" is often used to describe the types of correlation with respect to ρ_{XY}. For instance, the term "positively linearly correlated" is often used in place of positively correlated.

Example: Recall CB Example 4.5.9 as discussed in the previous section in comparison with CB Example 4.5.8.

The question is, under what conditions, will we have that $cov(X, Y) = 0$ if and only if (X, Y) are independent?
A special case is when (X, Y) are jointly normally distributed. We now show this.

Theorem: Suppose X, Y are jointly normally distributed. Then $cov(X, Y) = 0$ if and only if X and Y are independent. □

Remark: This characterization provides a very simple way to check independence between two random variables. The proof of this theorem is provided in Section 8 on the bivariate normal distribution. See Theorem 3 in Section 7.

7. Multivariate Normal Distribution

Multivariate Normal Distribution

Let $X_1, \cdots, X_k$ be iid $N(0,1)$, and let

$$X = (X_1, \cdots, X_k)' \in \mathbb{R}^k.$$

The joint density of X is the product of the densities of $X_1, \cdots, X_k$ due to the independence assumption. Thus the density of X is

$$f(x) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_j^2}{2}} = \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{x'x}{2}},$$

where $x = (x_1, \cdots, x_k)' \in \mathbb{R}^k$. We call this distribution the k-variate standard normal distribution, denoted by $N_k(0, I_k)$, where 0 is the $k \times 1$ vector of zeros and I_k is the $k \times k$ identity matrix.

Next, let $Y = AX + \mu$, where μ is a nonrandom vector in $\mathbb{R}^k$ and A is a nonsingular $k \times k$ matrix with nonrandom elements. Then it follows that

$$\mathbb{E}Y = \mu, \quad Var(Y) = AA' \equiv \Sigma.$$

Note that $Y = g(X) = AX + \mu$ and $X = g^{-1}(Y) = A^{-1}(Y - \mu)$. The density of Y is

$$\begin{aligned} h(y) &= f(g^{-1}(y)) \left| \det \frac{dg^{-1}(y)}{dy} \right| \\ &= f(A^{-1}(y-\mu)) |\det A^{-1}| \\ &= \frac{1}{(\sqrt{2\pi})^k} \exp\left(-\frac{1}{2} (y-\mu)' (A^{-1})' A^{-1} (y-\mu) \right) \cdot \frac{1}{\sqrt{|\det \Sigma|}} \end{aligned}$$

Since

$$(A^{-1})' A^{-1} = (A')^{-1} A^{-1} = (AA')^{-1}$$

and

$$\begin{aligned} |\det A| &= \sqrt{(\det A)^2} = \sqrt{\det(A)\det(A)} = \sqrt{\det(A)\det(A')} \\ &= \sqrt{\det AA'} = \sqrt{\det \Sigma}, \end{aligned}$$

$$h(y) = \frac{1}{(\sqrt{2\pi})^k \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} (y-\mu)' \Sigma^{-1} (y-\mu) \right).$$

This distribution is called the k-variate normal, denoted by $N_k(\mu, \Sigma)$, with mean vector μ and covariance matrix Σ.

The characteristic function of the k-variate standard normal distribution $N_k(0, I_k)$ is:

$$\begin{aligned}
\varphi_X(t) &= \mathbb{E}e^{it'\mathbf{X}} = \mathbb{E}e^{i\sum_{j=1}^k t_j \mathbf{X}_j} = \mathbb{E}\prod_{j=1}^k e^{it_j \mathbf{X}_j} \\
&= \prod_{j=1}^k \mathbb{E}e^{it_j \mathbf{X}_j} = \prod_{j=1}^k e^{-\frac{1}{2}t_j^2} = \exp\left(-\frac{1}{2}\sum_{j=1}^k t_j^2\right) \\
&= \exp\left(-\frac{1}{2}t't\right)
\end{aligned}$$

where $t = (t_1, \ldots, t_k)' \in \mathbb{R}^k$.

Theorem 1: The characteristic function of the k-variate normal distribution $N_k(\mu, \Sigma)$ is

$$\varphi_Y(t) = \exp\left(it'\mu - \frac{1}{2}t'\Sigma t\right), \qquad t \in \mathbb{R}^k.$$

□

Proof. Dhrymes (1974, p. 14, Lemma 4).

Theorem 2: Let X be a random vector in $\mathbb{R}^k$ with mean vector μ and covariance matrix Σ. If every linear combination of the components of X is normally distributed, then X is distributed $N_k(\mu, \Sigma)$. □

Proof. Dhrymes (1974, p. 19, Proposition 1).

Theorem 3: Let $X = (X_1, \cdots, X_k)'$ be k-variate normally distributed. If $X_1, \cdots, X_k$ are uncorrelated, then they are independent. □

Proof. Dhrymes (1974, p. 18, Lemma 8). Show this.

Theorem 4: Let $X \sim N_k(\mu, \Sigma)$ and let $Y = BX + b$, where b is a nonrandom vector in $\mathbb{R}^m$ and B is a $m \times k$ matrix of nonrandom elements. Then

$$Y \sim N_m\left(B\mu + b, B\Sigma B'\right).$$

□

Proof. Dhrymes (1974, p. 15, Lemma 5).

Theorem 5: If $X \sim N_k(0, \Sigma)$ with Σ nonsingular, then $X'\Sigma^{-1}X \sim \chi^2_k$. □

Proof. Since Σ is symmetric, $\Sigma = Q\Lambda Q'$, where Λ is the diagonal element with the eigenvalues of Σ on the diagonal and Q is the orthogonal $k \times k$ matrix of corresponding eigenvectors. Then

$$\begin{aligned}
X'\Sigma^{-1}X &= X'(Q\Lambda Q')^{-1}X \\
&= X'(Q')^{-1}\Lambda^{-1}Q^{-1}X \\
&= X'Q\Lambda^{-1}Q'X \qquad \text{(because } QQ' = Q'Q = I_k\text{)} \\
&= X'Q\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}Q'X \\
&= Y'Y \qquad \text{(letting } Y = \Lambda^{-\frac{1}{2}}Q'X\text{)}.
\end{aligned}$$

If $Y = (Y_1, \ldots, Y_k)'$ then $Y'Y = \sum_{j=1}^{k} Y_j$. By Theorem 4,

$$Y = \Lambda^{-\frac{1}{2}} Q' X \sim N_k \left(\mathbf{0}, \Lambda^{-\frac{1}{2}} Q' \Sigma Q \Lambda^{-\frac{1}{2}} \right) \equiv N_k \left(\mathbf{0}, I_k \right)$$

where "$\equiv$" denotes for the same distribution, and

$$\Lambda^{-\frac{1}{2}} Q' \Sigma Q \Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}} Q' Q \Lambda Q' Q \Lambda^{-\frac{1}{2}} = I_k.$$

Thus $Y_j \sim N_1(0, 1)$ for all $j = 1, \ldots, k$. Since $Y_1, \ldots, Y_k$ follow normal distribution and uncorrelated (as I_k is diagonal), by Theorem 3, they are independent. And since $Y_j^2 \sim \chi_1^2$ for all $j = 1, \ldots, k$,

$$Y'Y = \sum_{j=1}^{k} Y_j^2 \sim \chi_k^2.$$

Theorem 6: Let X be distributed $N_k(0, I_k)$ and let $Y_1 = B_1X$ and $Y_2 = B_2X$, where B_1 and B_2 are nonrandom $m_1 \times k$ and $m_2 \times k$ matrices. If $B_1B_2' = 0$, then Y_1 and Y_2 are independent. □

Proof. Let $B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$ and $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = BX$. Then it follows from Theorem 4 that

$$Y \sim N_{m_1+m_2}\left(0, BB'\right).$$

The covariance matrix BB' can be written as

$$BB' = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \begin{pmatrix} B_1' & B_2' \end{pmatrix} = \begin{pmatrix} B_1B_1' & B_1B_2' \\ B_2B_1' & B_2B_2' \end{pmatrix}.$$

The characteristic function of Y is

$$\begin{aligned}
\varphi_{_Y}(t) &= \mathbb{E}\exp(it'Y) \\
&= \exp\left(-\frac{1}{2}t'BB't\right) \\
&= \exp\left(-\frac{1}{2}t_1'B_1B_1't_1 - \frac{1}{2}t_2'B_2B_2't_2 - t_1'B_1B_2't_2\right)
\end{aligned}$$

where $t = (t_1' \; t_2')' \in \mathbb{R}^{m_1+m_2}$ with $t_1 \in \mathbb{R}^{m_1}$ and $t_2 \in \mathbb{R}^{m_2}$.

Moreover, the characteristic functions of Y_1 and Y_2 are

$$\begin{aligned}\varphi_{Y_1}(t_1) &= \mathbb{E}\exp(it_1'Y_1) = \exp\left(-\frac{1}{2}t_1'B_1B_1't_1\right),\\ \varphi_{Y_2}(t_2) &= \mathbb{E}\exp(it_2'Y_2) = \exp\left(-\frac{1}{2}t_2'B_2B_2't_2\right).\end{aligned}$$

since $Y_1 \sim N_{m_1}(0, B_1B_1')$ and $Y_2 \sim N_{m_2}(0, B_2B_2')$.
Thus we have

$$\varphi_Y(t) = \varphi_{Y_1}(t_1)\varphi_{Y_2}(t_2)\exp\left(-t_1'B_1B_2't_2\right).$$

Consequently, if B_1B_2' is a zero $m_1 \times m_2$ matrix, then Y_1 and Y_2 are independent.

Marginal and Conditional Normal Distributions

Let $X \sim N_k(\mu, \Sigma)$ and consider the following partition

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $X_1, \mu_1 \in \mathbb{R}^{k_1}$, $X_2, \mu_2 \in \mathbb{R}^{k_2}$, $k_1 + k_2 = k$.

Theorem 7: (1) The marginal distribution of X_1 and X_2 are $N_k(\mu_1, \Sigma_{11})$ and $N_{k_2}(\mu_2, \Sigma_{22})$, respectively.
(2) The conditional distribution of X_2 given $X_1 = x_1$ is

$$N_{k_1}(\Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) + \mu_2, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

provided that Σ_{11} is nonsingular. □

Proof. (1) Dhrymes (1974, p. 15, Lemma 6).
(2) Dhrymes (1974, p. 16, Lemma 7).

Theorem 8: Let $X \in \mathbb{R}^k$ and $Y \in \mathbb{R}$ be jointly normally distributed. Then

$$\mathbb{E}(Y|X = x) = \alpha + \beta' x,$$

where $\beta = (Var(X))^{-1} Cov(X, Y)$ and $\alpha = \mathbb{E}Y - \beta' \mathbb{E}X$. □

Proof. Use Theorem 7.

8. Bivariate Normal Distribution

Bivariate Normal Distribution

Joint density

Two random variables (X, Y) follows a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ if their pdf $f_{XY}(x, y)$ is given by

$$f_{XY}(x,y) = \frac{1}{(\sqrt{2\pi})^2\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(y-\mu)' \Sigma^{-1} (y-\mu)\right)$$

where $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and

$$\begin{aligned} \Sigma &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}. \end{aligned}$$

Bivariate Normal Distribution

Joint density (alternative representation)

$$
\begin{aligned}
f_{XY}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\begin{array}{l} \left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 \\ -2\rho \left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) \end{array} \right] \right\} \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{x-\mu_1}{\sigma_1}\right)^2 \right\} \\
&\quad \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{y-\mu_2}{\sigma_2}\right)^2 \right\} \\
&\quad \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[-2\rho \left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) \right] \right\}
\end{aligned}
$$

Bivariate Normal Distribution

Marginal density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}, \quad -\infty < x < \infty$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right\}, \quad -\infty < y < \infty$$

Bivariate Normal Distribution

Conditional density

Question for you: Write the conditional density of Y given X, when (X, Y) follows a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

$$
\begin{aligned}
f_{Y|X}(y) &= \frac{f_{XY}(x,y)}{f_X(x)} \\
&= \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left\{ -\frac{\left[(y-\mu_2) - \frac{\rho\sigma_2}{\sigma_1}(x_1-\mu_1)\right]^2}{2\sigma_2^2(1-\rho^2)} \right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_{2|1}^2}} \exp\left\{ -\frac{\left(y-\mu_{2|1}\right)^2}{2\sigma_{2|1}^2} \right\}, \quad -\infty < y < \infty
\end{aligned}
$$

where

$$\begin{aligned}
\mu_{2|1} &= \mathbb{E}(Y|X) \\
&= \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) + \mu_2 \\
&= \frac{\rho\sigma_1\sigma_2}{\sigma_1^2}(x_1 - \mu_1) + \mu_2 \\
&= \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1) + \mu_2,
\end{aligned}$$

$$\begin{aligned}
\sigma_{2|1}^2 &= var(Y|X) \\
&= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= \sigma_2^2 - \left(\rho\sigma_1\sigma_2\right)^2/\sigma_1^2 \\
&= \sigma_2^2(1 - \rho^2).
\end{aligned}$$

Bivariate Normal Distribution

Conditional density

Question for you: Write the conditional density of X given Y, when (X, Y) follows a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Bivariate Normal Distribution

Correlation

Question for you: Find the correlation ρ_{XY}.
Answer: $\rho_{XY} = \rho$. [We will show below. Or, see CB p. 176. Do not copy. Find your own way.]

Let $u = \left(\frac{x-\mu_1}{\sigma_1}\right)$ and $v = \left(\frac{y-\mu_2}{\sigma_2}\right)$. Then

$$\begin{aligned}
\rho_{XY} &= \frac{cov(X,Y)}{\sigma_1\sigma_2} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) f_{XY}(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv\frac{1}{2\pi\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[u^2-2\rho uv+v^2\right]\right\}\mathrm{d}u\mathrm{d}v
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv \frac{1}{2\pi\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[u^2 - 2\rho uv + v^2\right]\right\} \mathsf{d}u\mathsf{d}v \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv \frac{1}{2\pi\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right\} \exp\left\{-\frac{(1-\rho^2)v^2}{2(1-\rho^2)}\right\} \mathsf{d}u\mathsf{d}v \\
&= \int_{-\infty}^{\infty} \frac{v}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} \\
&\quad \times \left[\int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right\} \mathsf{d}u\right] \mathsf{d}v
\end{aligned}
$$

The inner integral

$$\int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right\} du = E(u)$$

where $u \sim N(\rho v, 1-\rho^2)$. Thus the inner integral is ρv. Hence we have

$$\rho_{XY} = \rho \int_{-\infty}^{\infty} \frac{v^2}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} \mathrm{d}v$$

This integral is

$$\int_{-\infty}^{\infty} \frac{v^2}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} dv = E(v^2)$$

where $v \sim N(0,1)$. Thus the integral is 1. Therefore,

$$\rho_{XY} = \rho$$

Remark: ρ_{XY} is only a particular kind of linear relationship. Two random variables X and Y may have a strong relationship but whose correlation is zero, because their relationship is nonlinear. In other words, ρ_{XY} may not capture some nonlinear association.
Suppose $X \sim N(0, \sigma^2)$ and $Y = X^2$. Then
$cov(X, Y) = \mathbb{E}XY = \mathbb{E}X^3 = 0$, although Y is a deterministic function of X. They are uncorrelated.
See CB Example 4.5.9.

Theorem: Suppose X, Y are jointly normally distributed. Then $cov(X, Y) = 0$ if and only if X and Y are independent. $\square$

Proof:

$$
\begin{aligned}
f_{XY}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right]\right\}
\end{aligned}
$$

First, we show $\Longrightarrow$. If $\rho = 0$, then

$$\begin{aligned} f_{XY}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2} \\ &\quad \times \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} \\ &\quad \times 1 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} \\ &= f_X(x)f_Y(y) \text{ for all } -\infty < x, y < \infty. \end{aligned}$$

Therefore (X, Y) are independent.

Next, we show $\Longleftarrow$. When (X, Y) are independent, $\rho = 0$, by Theorem A in Section 6.

Question for you:

1. Generate 1000 obervations of (X, Y) from $N(0, 0, 1, 1, \rho)$ with $\rho = -0.7$, -0.3, 0, 0.3, and 0.7. Do not use a built-in command for the bivariate normal random number generation. Use the univariate normal random number generator and a matrix A such that $AA' = \Sigma$. Recall that $Y = g(X) = AX + \mu$ and $X = g^{-1}(Y) = A^{-1}(Y - \mu)$. The density of Y is

$$\begin{aligned} h(y) &= f(g^{-1}(y)) \left| \det \frac{dg^{-1}(y)}{dy} \right| \\ &= f(A^{-1}(y-\mu)) |\det A^{-1}| \\ &= \frac{1}{(\sqrt{2\pi})^2 \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} (y-\mu)' \Sigma^{-1} (y-\mu) \right). \end{aligned}$$

2. Draw 3D graphs of the bivariate normal pdf $f_{XY}(x, y)$ with $\rho = -0.7$, -0.3, 0, 0.3, and 0.7.
3. Draw the contours of the 3D graphs of the joint bivariate normal density with $\rho = -0.7$, -0.3, 0, 0.3, and 0.7.

9. Conditional Expectations

Conditional Expectations

What information can we extract from a conditional distribution $f_{Y|X}(y|x)$?

Definition [Conditional Expectation]: The conditional expectation of $g(X,Y)$ given $X = x$ is defined as

$$\begin{aligned}\mathbb{E}[g(X,Y)|X=x] &= \mathbb{E}[g(X,Y)|x] \\ &= \begin{cases} \sum_y g(x,y) f_{Y|X}(y|x) & \text{d.r.v.} \\ \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x) dy & \text{c.r.v.} \end{cases}\end{aligned}$$

Remarks:

- When taking a conditional expectation, one can treat x as fixed.
- Conditional expectation is simply an expectation with respect to a conditional distribution instead of an unconditional expectation.
- This conditional expectation is a function of x only, because y has been integrated out.
- The conditional expectation $\mathbb{E}[g(X,Y)|x]$ is a function of x. That is, for each value of x, $\mathbb{E}[g(X,Y)|x]$ is a number obtained by computing the appropriate integral or sum. Thus, $\mathbb{E}[g(X,Y)|X]$ is a random variable whose value depends on the value of X. If $X = x$, the value of the random variable $\mathbb{E}[g(X,Y)|X]$ is $\mathbb{E}[g(X,Y)|x]$.

Theorem [Law of Iterated Expectations]:

$$\begin{aligned}\mathbb{E}[g(X,Y)] &= \mathbb{E}_X\left(\mathbb{E}[g(X,Y)|X]\right) \\ &= \mathbb{E}_Y\left(\mathbb{E}[g(X,Y)|Y]\right).\end{aligned}$$

□

Proof: Immediately by partition of the joint probability. Suppose (X,Y) have a pdf $f_{XY}(x,y)$. Then

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

Remark: The law of iterated expectations provides a two stage procedure to compute an unconditional expectation. In practice, we often write $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(Y|X)]$, by abusing the notation "$\mathbb{E}$". The same notation $\mathbb{E}$ stands for different expectations in the same equation. The inside notation $\mathbb{E}$ is the expectation with respect to the conditional distribution of $Y|X$, and the outside notation $\mathbb{E}$ is the expectation with respect to the marginal distribution of X.

Consider a special Case: $g(X, Y) = Y$.

Definition [Conditional Mean of Y given $X = x$]:

$$\begin{aligned}\mathbb{E}[Y|X = x] &= \mathbb{E}[Y|x] \\ &= \begin{cases} \sum_y y f_{Y|X}(y|x) & \text{d.r.v.} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy & \text{c.r.v.} \end{cases}\end{aligned}$$

Remark. $\mathbb{E}[Y|x]$ is the average value of Y conditional on when X takes the value of x. Let $X =$ Gender ($X = 0$ for female and $X = 1$ for male). Then $\mathbb{E}[Y|X = 0]$ is the average wage of the female. If $\mathbb{E}[Y|X = x] = \mathbb{E}[Y|x]$ is a function of x, then it implies gender discrimination. Overall average wage

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}_X[\mathbb{E}(Y|X)] \\ &= \Pr(X = 0)\mathbb{E}(Y|X = 0) + \Pr(X = 1)\mathbb{E}(Y|X = 1).\end{aligned}$$

where $Y =$ Wage of an employee and $X =$ Gender ($X = 0$ female, $X = 1$ male).

Remark. $\mathbb{E}[Y|x]$ is a function of x only. It is also called the regression function of Y given X.

Theorem [Mean Squared Error Criterion]: Let X and Y be random variables defined on the same sample space and suppose Y has a finite variance. Then

$$\mathbb{E}(Y|X) = \arg\min_{g(\cdot)} \mathbb{E}[Y - g(X)]^2,$$

where the minimization is over all measurable and square-integrable functions. □

Proof: Put $g_0(X) = \mathbb{E}(Y|X)$. Then we have

$$\begin{aligned} MSE(g) &= \mathbb{E}[Y - g(X)]^2 \\ &= \mathbb{E}[Y - g_0(X) + g_0(X) - g(X)]^2 \\ &= \mathbb{E}[Y - g_0(X)]^2 + \mathbb{E}[g_0(X) - g(X)]^2 \\ &\quad +2\mathbb{E}[Y - g_0(X)][g_0(X) - g(X)] \\ &= \mathbb{E}[Y - g_0(X)]^2 + \mathbb{E}[g_0(X) - g(X)]^2 + 0 \\ &= \text{variance } + \text{ (bias)}^2, \end{aligned}$$

where $2\mathbb{E}[Y - g_0(X)][g_0(X) - g(X)] = 0$, because $\mathbb{E}(\cdot) = \mathbb{E}_X[\mathbb{E}(\cdot|X)]$ and

$$\begin{aligned}
& \mathbb{E}[Y - g_0(X)][g_0(X) - g(X)] \\
= \; & \mathbb{E}_X \left\{ \mathbb{E}[Y - g_0(X)][g_0(X) - g(X)]|X \right\} \\
= \; & \mathbb{E}_X \left\{ [g_0(X) - g(X)] \mathbb{E}[Y - g_0(X)]|X \right\} \\
= \; & \mathbb{E}_X \left\{ [g_0(X) - g(X)] \left(\mathbb{E}[Y|X] - g_0(X) \right) \right\} \\
= \; & \mathbb{E}_X \left\{ [g_0(X) - g(X)] \left(g_0(X) - g_0(X) \right) \right\} \\
= \; & 0,
\end{aligned}$$

by the law of iterated expectations.
Now, note that the first term (the variance term) is solely determined by the probability distribution of (X, Y). The second term (the squared bias term) can be made to zero, by selecting $g(X) = g_0(X)$. This will minimize $MSE(g)$.

Remarks:
(a) Because Y is correlated with X, it is a function of X: We use $g(X)$ to approximate Y.
(b) Criterion of approximation $MSE(g) = \mathbb{E}[Y - g(X)]^2$ is called the mean squared error of $g(X)$. It is a loss function used to measure how good the approximation of $g(X)$. The smaller $\mathbb{E}[Y - g(X)]^2$ is, the better the approximation of $g(X)$.
(c) The theorem shows that the best predictor in terms of MSE is the conditional mean $g_0(X) = \mathbb{E}(Y|X)$. This is why $\mathbb{E}(Y|X)$ is so important in econometrics.

(d) In many cases, one may try to solve the *constrained* minimization problem

$$\mathbb{E}(Y|X) = \arg \min_{g(\cdot)\in\mathbb{A}} \mathbb{E}[Y - g(X)]^2,$$

where

$$\mathbb{A} = \{g : \mathbb{R} \to \mathbb{R} \mid g(x) = \alpha + \beta x\}.$$

This is the linear least squares approximation. The optimal linear function $g(X) = \alpha^* + \beta^* X$ may not be equal to the conditional mean $\mathbb{E}(Y|X)$.
If $\Pr[\mathbb{E}(Y|X) = \alpha^* + \beta^* X] = 1$ for some α^*, β^*, then we say the linear model $\mathbb{A}$ is correctly specified fot the conditional mean. Otherwise, we say the linear model $\mathbb{A}$ suffers from the *neglected nonlinearity*. See Lee, White, and Granger (1993 JoE).

Theorem: Suppose that $\mathbb{E}(Y|X)$ exists. Then there is a random variable ε such that

$$Y = \mathbb{E}(Y|X) + \varepsilon,$$

where ε is called the regression disturbance or regression error, with

$$\mathbb{E}(\varepsilon|X) = 0.$$

□

Remarks: (a) This is called the regression equation of Y given X. The regression disturbance ε reflects the degree of uncertainty about the relationship between Y and X. When $\varepsilon = 0$, we have a perfect deterministic relationship between X and Y. An example is $Y = g(X)$.
(b) What is meant by $\mathbb{E}(\varepsilon|X) = 0$? Put it simply, ε contains no systematic information of X that can be used to predict the expected value of Y. All systematic information of X that can be used to predict the expected value of Y has been incorporated in $\mathbb{E}(Y|X)$.
(c) If $\mathbb{E}(Y|X) = a + bX$, then the regression becomes a linear regression, i.e. $Y = a + bX + \varepsilon$, where $\mathbb{E}(\varepsilon|X) = 0$.

Remark: Suppose $Y = a + bX + \varepsilon$, where $\mathbb{E}(\varepsilon|X) = 0$. Then $\mathbb{E}(Y|X) = a + bX$?

$$\begin{aligned}
\mathbb{E}(Y|X) &= \mathbb{E}[a + bX + \varepsilon|X] \\
&= \mathbb{E}(a|X) + \mathbb{E}(bX|X) + \mathbb{E}(\varepsilon|X) \\
&= a + bX + 0 \\
&= a + bX.
\end{aligned}$$

Remark: Suppose $Y = a + bX + \varepsilon$, where $\mathbb{E}(\varepsilon|X) = 0$. Then $\mathbb{E}(X\varepsilon) = 0$?

$$\begin{aligned}\mathbb{E}(X\varepsilon) &= \mathbb{E}[\mathbb{E}(X\varepsilon|X)] \\ &= \mathbb{E}[X\mathbb{E}(\varepsilon|X)] \\ &= \mathbb{E}(X \cdot 0) \\ &= 0.\end{aligned}$$

This implies $cov(X, \varepsilon) = 0$. (Why?) Since ε has no information on X that can be used to predict the mean of Y, it should be orthogonal to X. By orthogonality, we mean that ε and X are uncorrelated, i.e., $cov(X, \varepsilon) = 0$.

Remark: $\mathbb{E}(\varepsilon|X) = 0$ implied $\mathbb{E}(X\varepsilon) = 0$. But $\mathbb{E}(\varepsilon|X) \neq 0$ does not imply $\mathbb{E}(X\varepsilon) \neq 0$.

Remark: Suppose X and Y are two random variables with finite second moments. Then we always have

$$Y = \alpha^* + \beta^* X + \varepsilon,$$

where ε satisfies $\mathbb{E}(X\varepsilon) = 0$ and the optimal coefficient $(\alpha^* \ \beta^*)$ that minimizes the mean square error

$$MSE(\alpha, \beta) = \mathbb{E}[Y - (\alpha + \beta X)]^2$$

is given by

$$\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \begin{bmatrix} 1 & \mu_X \\ \mu_X & \mathbb{E}X^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_Y \\ \mathbb{E}XY \end{bmatrix},$$

i.e.,

$$\begin{aligned} \alpha^* &= \mu_Y - \frac{cov(X,Y)}{var(X)}\mu_X, \\ \beta^* &= \frac{cov(X,Y)}{var(X)}. \end{aligned}$$

FOC:

$$\begin{aligned} \frac{\partial}{\partial \alpha} MSE(\alpha, \beta) &= 0 \\ \frac{\partial}{\partial \beta} MSE(\alpha, \beta) &= 0 \end{aligned}$$

From these two equations, we can solve for α^* and β^*. These conditions are equivalent to

$$\begin{aligned} \mathbb{E}(Y - \alpha^* - \beta^* X) &= 0, \\ \mathbb{E}\left[X(Y - \alpha^* - \beta^* X)\right] &= 0. \end{aligned}$$

Put $\varepsilon = Y - \alpha^* - \beta^* X$.

Law of Iterated Expectations [CB 4.4.3]: If X and Y are random variables, then

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

provided that the expectations exist. □

Remark. $\mathbb{E}[Y|x]$ is the average value of Y conditional on when X takes the value of x. Let $X =$ Gender ($X = 0$ for female and $X = 1$ for male). Then $\mathbb{E}[Y|X = 0]$ is the average wage of the female. Overall average wage

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}_X[\mathbb{E}(Y|X)] \\ &= \Pr(X = 0)\mathbb{E}(Y|X = 0) + \Pr(X = 1)\mathbb{E}(Y|X = 1). \end{aligned}$$

where $Y =$ Wage of an employee and $X =$ Gender ($X = 0$ female, $X = 1$ male). This is a two-stage (iterated) procedure. In Stage 1, for each x, find the average of Y, that is $\mathbb{E}(Y|X = x)$. Do this for all x's. In Stage 2, sum over these averages to get the overall averages to get the overall average of Y. This two-stage calculation provides more insight into the income distribution of the economy.

Proof: We only prove for the continuous case. Noting that $\mathbb{E}(Y|X)$ is a function of X only, and so

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

Example [CB 4.4.1, CB 4.4.2] An insect lays a large number of eggs, each surviving w.p. p. On average how many eggs will survive? Let $X =$ number of survivors and $Y =$ number of eggs laid. Then

$$\begin{aligned} X|Y &\sim \text{Binomial}(Y, p) \\ Y &\sim \text{Poisson}(\lambda). \end{aligned}$$

On average how many eggs will survive? That is to find

$$E(X) = E_Y[E(X|Y)] = E_Y[Y\ p] = E_Y[Y]\,p = \lambda p.$$

Remark: The distribution of X, Poisson(λp), in this example is a *mixture* distribution, because the distribution of X depends on Y that also has a distribution. See CB 4.4.4.

Question for you: Let $f_{XY}(x,y) = e^{-y}$ for $0 < x < y < \infty$. Find $\mathbb{E}(Y|X = x)$.

First, note that

$$f_{Y|X}(y|x) = e^{-(y-x)} \quad \text{for } 0 < x < y < \infty.$$

Then

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

$$
\begin{aligned}
&= -e^x \int_x^\infty y de^{-y} \\
&= -e^x \int_x^\infty u dv \qquad (u = y, v = e^{-y}) \\
&= -e^x \left[uv|_x^\infty - \int_x^\infty v du \right] \\
&= -e^x \left[ye^{-y}|_x^\infty - \int_x^\infty e^{-y} dy \right] \\
&= -e^x \left[-xe^{-x} - e^{-x} \right] \\
&= 1 + x.
\end{aligned}
$$

This implies that the regression equation of Y given X is

$$Y = \mathbb{E}(Y|X) + \varepsilon = 1 + X + \varepsilon = \alpha_0 + \alpha_1 X + \varepsilon,$$

where $\mathbb{E}(\varepsilon|X) = 0$ and $\alpha_0 = \alpha_1 = 1$.

Remark: This is an example that the regression function $\mathbb{E}(Y|X)$ is linear in X even if $(Y\ X)$ are not jointly normal and the conditional distribution of Y given X is not normal.

Example. [Bivariate Normal Distribution] Suppose (X, Y) follows a bivariate normal distribution $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Then the conditional mean

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

is a linear function of X.

Application [Efficient Market Hypothesis (EMH)]: If $\mathbb{E}(Y|X) = \mathbb{E}(Y)$, then Y is not predictable in mean using the information of X. An important application of this statistical hypothesis is the efficient market hypothesis. Let Y_t be the asset return at time t, and I_{t-1} is an information available at time $t-1$.
(a) The weak form EMH:

$$\mathbb{E}(Y_t|I_{t-1}) = \mathbb{E}(Y_t),$$

if I_{t-1} only contains the information of past asset returns.
(b) The semi-strong form EMH:

$$\mathbb{E}(Y_t|I_{t-1}) = \mathbb{E}(Y_t),$$

if I_{t-1} contains all the publicly available information at time $t-1$.
(c) The strong form EMH:

$$\mathbb{E}(Y_t|I_{t-1}) = \mathbb{E}(Y_t),$$

if I_{t-1} contains not only publicly available information but also some insider information.

How to test $\mathbb{E}(Y|X) = \mathbb{E}(Y)$?

Theorem: Suppose $\mathbb{E}(Y|X) = \mathbb{E}(Y)$. Then $cov(X, Y) = 0$. $\square$

Proof: Put $\varepsilon = Y - \mathbb{E}(Y|X) = Y - \mathbb{E}(Y)$. Then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ implies

$$\mathbb{E}(\varepsilon|X) = \mathbb{E}(Y - \mathbb{E}(Y|X)|X) = \mathbb{E}(Y - \mathbb{E}(Y)|X) = \mathbb{E}(Y|X) - \mathbb{E}(Y|X) = 0$$

By the law of iterated expectations, we have

$$\begin{aligned} cov(X, Y) &= \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \mathbb{E}\left[(X - \mu_X)\varepsilon\right] \\ &= \mathbb{E}\left(\mathbb{E}\left[(X - \mu_X)\varepsilon|X\right]\right) \\ &= \mathbb{E}\left[(X - \mu_X)\mathbb{E}(\varepsilon|X)\right] \\ &= \mathbb{E}\left[(X - \mu_X) \cdot 0\right] \\ &= 0. \end{aligned}$$

Remark: Therefore, we can check if $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ by checking whether $cov(X, Y) = 0$. This is indeed the basic idea behind the regression

$$Y_t = \alpha + \sum_{j=1}^{p} \beta_j Y_{t-j} + \varepsilon_t,$$

and check whether all coefficients β's are jointly zero.

However, we cannot conclude $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ if $cov(X, Y) = 0$. It is possible that $cov(X, Y) = 0$ but $\mathbb{E}(Y|X) \neq \mathbb{E}(Y)$.

Example. $Y = X^2 + u$, where $X \sim N(0, 1)$, $u \sim N(0, 1)$, and X and u are independent. Then $cov(X, Y) = 0$, but $\mathbb{E}(Y|X) = X^2 \neq \mathbb{E}(Y) = 1$.

Conditional Variance:

Most dynamic economic theories (e.g., efficient market hypothesis, dynamic asset pricing, consumption and tax smoothing, rational expectations) have implications on and only on the conditional mean dynamics of the underlying economic process. However, we can consider the conditional variance of Y given X.
Now, let $g(X,Y) = [Y - \mathbb{E}(Y|X)]^2 = \varepsilon^2$.

Definition: [Conditional Variance]: The conditional variance of Y given X is defined as

$$\begin{aligned} var(Y|x) &= \mathbb{E}(\varepsilon^2|X) \\ &= \int_{-\infty}^{\infty} [Y - \mathbb{E}(Y|X)]^2 dF_{Y|X}(y|x). \end{aligned}$$

Remarks:

- When $var(\varepsilon|X) = var(Y|X) = \sigma^2$ (a constant independent of X), then ε is called a conditionally homoskedastic disturbance (homoskedasticity); otherwise, it is called a conditionally heteroskedastic disturbance (heteroskedasticity). Classical regression analysis usually assumes conditional homoskedasticity. However, this may not be a realistic assumption in many economic and financial applications.
- The uncertainty of Y depends on the level of X. When there exists heteroskedasticity, we can use X to predict the uncertainty of Y. When there is homoskedasticity, the information of X is useless in explaining uncertainty of Y.

- Conditional heteroskedasticity is very useful in economics and finance. For example,
  - Large firms have large variations in output.
  - "Level effect" in interest rate volatility: The interest rate volatility depends on the interest rate level – the higher the interest rate level r_t, the higher the interest rate volatility:

$$var(r_t|I_{t-1}) = \alpha r_{t-1}^{\rho},$$

    where $\alpha, \rho > 0$, $I_{t-1} = \{r_{t-1}, r_{t-2}, \ldots\}$.
  - GDP growth rate and its volatility: $Y_t = \alpha\sigma_t^{\rho}$, where $\sigma_t^2 = var(Y_t|I_{t-1})$, where $I_{t-1} = \{Y_{t-1}, Y_{t-2}, \ldots\}$ is oberved past information available at time $t-1$.
  - It can be used to explain and predict volatility of stock returns. This is the well-known AutoRegressive Conditional Heteroskedasticity model (ARCH) proposed by Engle (1982, *Econometrica*).

Example: What is an ARCH model? Let $Y_t = 100 \ln(P_t/P_{t-1})$ be the asset return. Then

$$var(Y_t|I_{t-1}) = \alpha + \beta Y_{t-1}^2,$$

where $\alpha, \beta > 0$. This model can be used to explain the so-called volatility clustering (i.e., a large volatility tends to be followed by another large volatility, and a small volatility tends to be followed by another small volatility).

Example (RiskMetrics):

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \varepsilon_t &\sim \text{i.i.d. N(0,1)} \\ \sigma_t^2 &= var(Y_t|I_{t-1}) \\ &= (1-\lambda)\sum_{j=1}^{\infty} \lambda^{j-1} Y_{t-1}^2 \end{aligned}$$

where $I_{t-1} = \{Y_{t-1}, Y_{t-2}, \ldots\}$.

Theorem:

$$var(Y|x) = \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2.$$

□

Example. Let $f_{XY}(x,y) = e^{-y}$ for $0 < x < y < \infty$. Find $var(Y|X = x)$. Then

$$\begin{aligned}
var(Y|x) &= \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2 \\
&= \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)dy - (1+x)^2 \\
&= \int_{x}^{\infty} y^2 e^{-(y-x)}dy - (1+x)^2 \\
&= e^x \int_{x}^{\infty} y^2 e^{-y}dy - (1+x)^2 \\
&= e^x \int_{x}^{\infty} -y^2 de^{-y} - (1+x)^2 \\
&= -e^x \int_{x}^{\infty} udv - (1+x)^2 \qquad (u = y^2, v = e^{-y})
\end{aligned}$$

$$
\begin{aligned}
&= -e^x \int_x^\infty u dv - (1+x)^2 \qquad (u = y^2, v = e^{-y}) \\
&= -e^x \left[uv|_x^\infty - \int_x^\infty v du \right] - (1+x)^2 \\
&= -e^x \left[y^2 e^{-y}\big|_x^\infty - \int_x^\infty e^{-y} dy^2 \right] - (1+x)^2 \\
&= -e^x \left[0 - x^2 e^{-x} - \int_x^\infty 2y e^{-y} dy \right] - (1+x)^2 \\
&= x^2 + 2 \int_x^\infty y e^{-(y-x)} dy - (1+x)^2 \\
&= x^2 + 2(1+x) - (1+x)^2 \\
&= 1.
\end{aligned}
$$

The variable $\varepsilon = Y - \mathbb{E}(Y|X)$ is a homoskedastic disturbance.

Theorem [Conditional Variance Identity, CB 4.4.7] For any two random variables X and Y with finite second moments,

$$var(Y) = \mathbb{E}[var(Y|X)] + var[\mathbb{E}(Y|X)].$$

□

Remark: What is the interpretation for this formula? The total variance (i.e., the unconditional variance) of Y consists of two components. One is the average of the unexpected variation ε. The other is the variation of the expected component $\mathbb{E}(Y|X)$.

Proof:

$$
\begin{aligned}
& var(Y) \\
= \; & \mathbb{E}(Y-\mu_Y)^2 \\
= \; & \mathbb{E}(Y-g_0(X)+g_0(X)-\mu_Y)^2 && (1)\\
= \; & \mathbb{E}(\varepsilon+g_0(X)-\mu_Y)^2 && (2)\\
= \; & \mathbb{E}(\varepsilon^2)+\mathbb{E}(g_0(X)-\mu_Y)^2+2\mathbb{E}\left[\varepsilon(g_0(X)-\mu_Y)\right] \\
= \; & \mathbb{E}\left[\mathbb{E}(\varepsilon^2|X)\right]+\mathbb{E}(g_0(X)-\mu_Y)^2 && (3)\\
& +2\mathbb{E}\left[\mathbb{E}\left[\varepsilon(g_0(X)-\mu_Y)\right]|X\right] \\
= \; & \mathbb{E}\left[var(\varepsilon|X)\right]+\mathbb{E}\left[g_0(X)-\mathbb{E}(g_0(X))\right]^2 \\
& +2\mathbb{E}\left[(g_0(X)-\mu_Y)\mathbb{E}\left(\varepsilon|X\right)\right] \\
= \; & \mathbb{E}\left[var(Y|X)\right]+var(g_0(X))+2\mathbb{E}\left[(g_0(X)-\mu_Y)\cdot 0\right] \\
= \; & \mathbb{E}\left[var(Y|X)\right]+var(\mathbb{E}(Y|X)).
\end{aligned}
$$

where (1) takes $g_0(X)=\mathbb{E}(Y|X)$, (2) takes $\varepsilon=Y-g_0(X)$, and (3) takes $\mu_Y=\mathbb{E}\left[\mathbb{E}(Y|X)\right]=\mathbb{E}(g_0(X))$.