SDS7102: Linear Models and Extensions

Random Variables

Qiang Sun, Ph.D. <qiang.sun@mbzuai.ac.ae>

These slides are prepared by Eric Moulines.

August 22, 2025

MBZUAI

Joint Distribution of a random vector

Definition

The joint distribution function of a random vector $(X_1, \cdots, X_k)$ is

$$F(x_1, \cdots, x_k) = P(X_1 \le x_1, \cdots, X_k \le x_k)$$

where the event $[X_1 \le x_1, \cdots, X_k \le x_k]$ is the intersection of the events $[X_1 \le x_1], \cdots, [X_k \le x_k]$.
Given the joint distribution function of random vector $\boldsymbol{X}$, we can determine $P(\boldsymbol{X} \in A)$ for any (Borel) set $A \subset \mathbb{R}^k$.

Joint Frequency Function of a discrete random vector

Definition

Suppose that $X_1, \cdots, X_k$ are discrete random variables defined on the same sample space. Then the joint frequency function of $X = (X_1, \cdots, X_k)$ is defined to be

$$f(x_1, \cdots, x_k) = P(X_1 = x_1, \cdots, X_k = x_k)$$

Joint Density function of a random vector

Definition

Suppose that $X_1, \cdots, X_n$ are continuous random variables defined on the same sample space and that

$$P\left[X_1 \leq x_1, \cdots, X_k \leq x_k\right] = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f\left(t_1, \cdots, t_k\right) dt_1 \cdots dt_k$$

for all $x_1, \cdots, x_k$. Then $f\left(x_1, \cdots, x_k\right)$ is the joint density function of $\left(X_1, \cdots, X_k\right)$ (provided that $f\left(x_1, \cdots, x_k\right) \geq 0$).

Marginal distributions

Theorem

(a) *Suppose that* $\boldsymbol{X} = (X_1, \cdots, X_k)$ *has joint frequency function* $f(\boldsymbol{x})$*. For* $\ell < k$*, the joint frequency function of* $(X_1, \cdots, X_\ell)$ *is*

$$g(x_1, \cdots, x_\ell) = \sum_{x_{\ell+1}, \cdots, x_k} f(x_1, \cdots, x_k)$$

(b) *Suppose that* $\boldsymbol{X} = (X_1, \cdots, X_k)$ *has joint density function* $f(\boldsymbol{x})$*. For* $\ell < k$*, the joint density function of* $(X_1, \cdots, X_\ell)$ *is*

$$g(x_1, \cdots, x_\ell) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_k)\, dx_{\ell+1} \cdots dx_k$$

Uniform distribution on a disk

Suppose that X and Y are continuous random variables with joint density function

$$f(x, y) = \frac{1}{\pi} \quad \text{for } x^2 + y^2 \leq 1$$

X and Y thus have a Uniform distribution on a disk of radius 1 centered at the origin.

- Determine $P(X \leq u)$ for $-1 \leq u \leq 1$.
- Determine the probability density function (pdf) of X

Independent random variables

Definition

Let $X_1, \cdots, X_k$ be random variables defined on the same sample space. $X_1, \cdots, X_k$ are said to be independent if the events $[a_1 < X_1 \leq b_1], [a_2 < X_2 \leq b_2], \cdots, [a_k < X_k \leq b_k]$ are independent for all $a_i < b_i, i = 1, \cdots, k$.

An infinite collection $X_1, X_2, \cdots$ of random variables are independent if every finite collection of random variables is independent.

Joint density of independent random variables

Theorem

If $X_1, \cdots, X_k$ are independent and have joint density (or frequency) function $f(x_1, \cdots, x_k)$ then

$$f(x_1, \cdots, x_k) = \prod_{i=1}^{k} f_i(x_i)$$

where $f_i(x_i)$ is the marginal density (frequency) function of X_i.

Conversely, if the joint density (frequency) function is the product of marginal density (frequency) functions then $X_1, \cdots, X_k$ are independent.

Minimum and Maximum of Uniform random variables

Suppose that $X_1, \cdots, X_n$ are i.i.d. continuous random variables with common (marginal) density $f(x)$ and distribution function $F(x)$. Given $X_1, \cdots, X_n$, we can define two new random variables

$$U = \min\left(X_1, \cdots, X_n\right) \quad \text{and} \quad V = \max\left(X_1, \cdots, X_n\right)$$

(a) Determine the marginal densities of U and V.

(b) Determine the joint density of (U, V)

Transformation

Suppose that $\boldsymbol{X} = (X_1, \cdots, X_k)$ is a random vector with some joint distribution. Define new random variables $Y_i = h_i(\boldsymbol{X})(i = 1, \cdots, k)$ where $h_1, \cdots, h_k$ are real-valued functions. We would like to determine

- the (marginal) distribution of Y_i, and
- the joint distribution of $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$.

Change of Variables formulae

Objective: find the joint density of $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$ where $Y_i = h_i(X_1, \cdots, X_k)\,(i = 1, \cdots, k)$ and $\boldsymbol{X} = (X_1, \cdots, X_k)$ has a joint density f_X.

We start by defining a vector-valued function $\boldsymbol{h}$ whose elements are the functions $h_1, \cdots, h_k$:

$$\boldsymbol{h}(\boldsymbol{x}) = \begin{pmatrix} h_1(x_1, \cdots, x_k) \\ h_2(x_1, \cdots, x_k) \\ \vdots \\ h_k(x_1, \cdots, x_k) \end{pmatrix}$$

Jacobian

- Assume (that $\boldsymbol{h}$ is a one-to-one function with inverse $\boldsymbol{h}^{-1}$ that is, $\left(\boldsymbol{h}^{-1}(\boldsymbol{h}(\boldsymbol{x}))=\boldsymbol{x}\right)$.
- Define the Jacobian matrix of $\boldsymbol{h}$ to be a $k \times k$ whose i-th row and j-th column element is

$$\frac{\partial}{\partial x_j} h_i\left(x_1, \cdots, x_k\right)$$

with the Jacobian of $\boldsymbol{h}$, $J_h\left(x_1, \cdots, x_k\right)$, defined to be the determinant of this matrix.

Change-of-Variable

Theorem

Suppose that $P(\boldsymbol{X} \in S) = 1$ *for some open set* $S \subset R^k$. *If*

(a) $\boldsymbol{h}$ *has continuous partial derivatives on* S,

(b) $\boldsymbol{h}$ *is one-to-one on* S,

(c) $J_{\boldsymbol{h}}(\boldsymbol{x}) \neq 0$ *for* $\boldsymbol{x} \in S$

then $(Y_1, \cdots, Y_k)$ *has joint density function*

$$
\begin{aligned}
f_Y(\boldsymbol{y}) &= \frac{f_X\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)}{\left|J_h\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)\right|} \\
&= f_X\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)\left|J_{h^{-1}}(\boldsymbol{y})\right|
\end{aligned}
$$

for $\boldsymbol{y} \in \boldsymbol{h}(S)$. $(J_{h^{-1}}$ *is the Jacobian of* $\boldsymbol{h}^{-1}$.$)$

Sum of independent random variables

Suppose that X_1, X_2 are random variables with joint frequency function $f_X(x_1, x_2)$ and let $Y = X_1 + X_2$.

(a) Suppose that X_1, X_2 are discrete; Determine the joint frequency function of Y.

(b) Suppose that X_1, X_2 are continuous with joint density $f_X(x_1, x_2)$. Determine the density function of Y.

Gamma distribution

Suppose that X_1, X_2 are independent Gamma random variables with common scale parameters:

$$X_1 \sim \mathrm{Gamma}(\alpha, \lambda) \quad \text{and} \quad X_2 \sim \mathrm{Gamma}(\beta, \lambda)$$

Define

$$Y_1 = X_1 + X_2$$
$$Y_2 = \frac{X_1}{X_1 + X_2}$$

Show that

(a) Y_1 is independent of Y_2;
(b) Y_1 has a Gamma distribution with shape parameter $\alpha + \beta$ and scale parameter λ;
(c) Y_2 has a Beta distribution with parameters α and β ($Y_2 \sim \mathrm{Beta}(\alpha, \beta)$).

Extensions

The change-of-variable formula can be extended to the case where the transformation $\boldsymbol{h}$ is not one-to-one. Suppose that $P[\boldsymbol{X} \in S] = 1$ for some open set and that S is a disjoint union of open sets $S_1, \cdots, S_m$ where $\boldsymbol{h}$ is one-to-one on each of the S_j 's (with inverse h_j^{-1} on S_j).

The joint density of $(Y_1, \cdots, Y_k)$ is

$$f_Y(\boldsymbol{y}) = \sum_{j=1}^{m} f_X\left(h_j^{-1}(\boldsymbol{y})\right) \left| J_{h_j^{-1}}(\boldsymbol{y}) \right| \mathbb{1}_{S_j}\left(h_j^{-1}(\boldsymbol{y})\right).$$

where $J_{h_j^{-1}}$ is the Jacobian of h_j^{-1}.

Order statistics

Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with density function $f(x)$. Reorder the X_i 's so that $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$; these latter random variables are called the order statistics of $X_1, \cdots, X_n$.

- Determine the distribution of order statistics.

Expectation

Expectation

If $\boldsymbol{X} = (X_1, \cdots, X_k)$ has a joint density or frequency function; more precisely, we can define

$$E[h(\boldsymbol{X})] = \sum_{x} h(\boldsymbol{x}) f(\boldsymbol{x})$$

if $\boldsymbol{X}$ has joint frequency function $f(\boldsymbol{x})$ and

$$E[h(\boldsymbol{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{x}) f(\boldsymbol{x}) dx_1 \cdots dx_k$$

if $\boldsymbol{X}$ has joint density function $f(\boldsymbol{x})$.

Expectation

Suppose that $X_1, \cdots, X_n$ are random variables defined on some sample space and let $Y = h(X_1, \cdots, X_k)$ for some real-valued function h. The expected value of Y to be

$$E(Y) = \int_0^\infty P(Y > y)dy - \int_{-\infty}^0 P(Y \le y)dy$$

This formula implies that we need to first determine the distribution function of Y in order to evaluate $E(Y)$.

Elementary properties of the expectation

Proposition

Suppose that $X_1, \cdots, X_k$ are random variables with finite expected values.

(a) *If $X_1, \cdots, X_k$ are defined on the same sample space then*

$$E\left(X_1 + \cdots + X_k\right) = \sum_{i=1}^{k} E\left(X_i\right)$$

(b) *If $X_1, \cdots, X_k$ are independent random variables then*

$$E\left(\prod_{i=1}^{k} X_i\right) = \prod_{i=1}^{k} E\left(X_i\right)$$

Moment generating function of a sum of independent random variables

Suppose that $X_1, \cdots, X_n$ are independent random variables with moment generating functions $m_1(t), \cdots, m_n(t)$, respectively. Define $S = X_1 + \cdots + X_n$.

- Compute the MGF of S.
- Assume that $X_1, \ldots, X_n$ are Gaussian, $E(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. What is the distribution of S

Covariance

Definition

Suppose X and Y are random variables with $E\left(X^2\right)$ and $E\left(Y^2\right)$ both finite and let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. The covariance between X and Y is

$$\operatorname{Cov}(X, Y) = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] = E(XY) - \mu_X \mu_Y$$

1. For any constants a, b, c, and d,

$$\operatorname{Cov}(aX + b, cY + d) = ac \operatorname{Cov}(X, Y)$$

2. If X and Y are independent random variables (with $E(X)$ and $E(Y)$ finite) then $\operatorname{Cov}(X, Y) = 0$

Independence and correlation

The converse to 2 is not true. In fact, it is simple to find an example where $Y = g(X)$ but $\mathrm{Cov}(X, Y) = 0$.

Suppose that X has a Uniform distribution on the interval $[-1, 1]$ and let $Y = -1$ if $|X| < 1/2$ and $Y = 1$ if $|X| \geq 1/2$.

- Show that $\mathrm{Cov}(X, Y) = 0$.

Elementary property

Proposition

Suppose that $X_1, \cdots, X_n$ *are random variables with* $E\left(X_i^2\right) < \infty$ *for all* i*. Then*

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}\left(X_i\right) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} a_i a_j \operatorname{Cov}\left(X_i, X_j\right)$$

Sampling with replacement

Suppose we are sampling without replacement from a finite population consisting of N items $a_1, \cdots, a_N$. Let X_i denote the result of the i-th draw; we then have

$$P\left(X_i = a_k\right) = \frac{1}{N} \quad \text{and} \quad P\left(X_i = a_k, X_j = a_\ell\right) = \frac{1}{N(N-1)}$$

where $1 \leq i, j, k, \ell \leq N, i \neq j$ and $k \neq \ell$. Suppose we define

$$S_n = \sum_{i=1}^{n} X_i$$

where $n \leq N$.

- Determine the mean and variance of S_n (Hint: you may use the $\mathrm{Var}(S_N)$) ?
- What happens if we sample with replacement ?

Covariance matrix

Given random variables $X_1, \cdots, X_n$, it is often convenient to represent the variances and covariances of the X_i 's via a $n \times n$ matrix.

Set $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ (a column vector); then we define the variance-covariance matrix (or covariance matrix) of $\boldsymbol{X}$ to be an $n \times n$ matrix $C = \mathrm{Cov}(\boldsymbol{X})$ whose diagonal elements are $C_{ii} = \mathrm{Var}\,(X_i)\,(i = 1, \cdots, n)$ and whose off-diagonal elements are $C_{ij} = \mathrm{Cov}\,(X_i, X_j)\,(i \neq j)$.

Covariance matrix

Variance-covariance matrices can be manipulated for linear transformations of $\boldsymbol{X}$: If $\boldsymbol{Y} = B\boldsymbol{X} + \boldsymbol{a}$ for some $m \times n$ matrix B and vector $\boldsymbol{a}$ of length m then

$$\mathrm{Cov}(\boldsymbol{Y}) = B\,\mathrm{Cov}(\boldsymbol{X})B^T$$

Likewise, if we define the mean vector of $\boldsymbol{X}$ to be

$$E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

then $E(\boldsymbol{Y}) = BE(\boldsymbol{X}) + \boldsymbol{a}$.

Correlation

Definition

Suppose that X and Y are random variables where both $E\left(X^2\right)$ and $E\left(Y^2\right)$ are finite. Then the correlation between X and Y is

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{[\operatorname{Var}(X)\operatorname{Var}(Y)]^{1/2}}$$

The advantage of the correlation is the fact that it is essentially invariant to linear transformations (unlike covariance). That is, if $U = aX + b$ and $V = cY + d$ then

$$\operatorname{Corr}(U,V) = \operatorname{Corr}(X,Y)$$

if a and c have the same sign; if a and c have different signs then $\operatorname{Corr}(U,V) = -\operatorname{Corr}(X,Y)$.

Property of the correlation

Proposition

Suppose that X and Y are random variables where both $E\left(X^2\right)$ and $E\left(Y^2\right)$ are finite. Then

(a) $-1 \leq \mathrm{Corr}(X, Y) \leq 1$;

(b) $\mathrm{Corr}(X, Y) = 1$ *if, and only if,* $Y = aX + b$ *for some* $a > 0$;
$\mathrm{Corr}(X, Y) = -1$ *if, and only if,* $Y = aX + b$ *for some* $a < 0$.

Optimal linear predictor

Proposition

Suppose that X *and* Y *are random variables where both* $E\left(X^2\right)$ *and* $E\left(Y^2\right)$ *are finite and define*

$$g(a, b) = E\left[(Y - a - bX)^2\right]$$

Then $g(a, b)$ *is minimized at*

$$b_0 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \mathrm{Corr}(X, Y)\left(\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}\right)^{1/2}$$

$$\text{and} \quad a_0 = E(Y) - b_0 E(X)$$

with $g\left(a_0, b_0\right) = \mathrm{Var}(Y)\left(1 - \mathrm{Corr}^2(X, Y)\right)$.

Conditional Distribution

Conditional distribution

We are often interested in the probability distribution of a random variable (or random variables) given knowledge of some event A.

If the conditioning event A has positive probability then we can define conditional distributions, conditional density functions (marginal and joint) and conditional frequency functions using the definition of conditional probability, for example,

$$P\left(X_1 \leq x_1, \cdots, X_k \leq x_k \mid A\right) = \frac{P\left(X_1 \leq x_1, \cdots, X_k \leq x_k, A\right)}{P(A)}$$

Conditional distribution

In the case of discrete random variables, it is straightforward to define the conditional frequency function of (say) $X_1, \cdots, X_j$ given the event $X_{j+1} = x_{j+1}, \cdots, X_k = x_k$ as

$$
\begin{aligned}
& f\left(x_1, \cdots, x_j \mid x_{j+1}, \cdots, x_k\right) \\
& \quad = P\left(X_1 = x_1, \cdots, X_j = x_j \mid X_{j+1} = x_{j+1}, \cdots, X_k = x_k\right) \\
& \quad = \frac{P\left(X_1 = x_1, \cdots, X_j = x_j, X_{j+1} = x_{j+1}, \cdots, X_k = x_k\right)}{P\left(X_{j+1} = x_{j+1}, \cdots, X_k = x_k\right)}
\end{aligned}
$$

which is simply the joint frequency function of $X_1, \cdots, X_k$ divided by the joint frequency function of $X_{j+1}, \cdots, X_k$.

Capture-recapture models

Mark/recapture experiments are used to estimate the size of animal populations. Suppose that the size of the population is N (unknown).

- Initially, m_0 members of the populations are captured and tagged for future identification before being returned to the population.
- Subsequently, a similar process is repeated k times: m_i members are captured at stage i and we define a random variable X_i to be the number of captured members who were tagged previously; the $m_i - X_i$ non-tagged members are tagged and all m_i members are returned to the population.

- Derive the joint distribution of $(X_1, \cdots, X_k)$.

Conditional distribution

Definition

Suppose that $(X_1, \cdots, X_k)$ has the joint density function $g(x_1, \cdots, x_k)$. Then the conditional density function of $X_1, \cdots, X_j$ given $X_{j+1} = x_{j+1}, \cdots, X_k = x_k$ is defined to be

$$f(x_1, \cdots, x_j \mid x_{j+1}, \cdots, x_k) = \frac{g(x_1, \cdots, x_j, x_{j+1}, \cdots, x_k)}{h(x_{j+1}, \cdots, x_k)}$$

provided that $h(x_{j+1}, \cdots, x_k)$, the joint density of $X_{j+1}, \cdots, X_k$, is strictly positive.

Conditional expected value

Definition

Given an event A with $P(A) > 0$ and a random variable X with $E[|X|] < \infty$, we define

$$E(X \mid A) = \int_0^\infty P(X > x \mid A)dx - \int_{-\infty}^0 P(X < x \mid A)dx$$

to be the conditional expected value of X given A.

Law of total probability

Theorem

Suppose that $A_1, A_2, \cdots$ *are disjoint events with* $P(A_k) > 0$ *for all* k *and* $\bigcup_{k=1}^{\infty} A_k = \Omega$. *Then if* $E[|X|] < \infty$,

$$E(X) = \sum_{k=1}^{\infty} E(X \mid A_k) P(A_k)$$

Conditional expectation

- Given a continuous random vector $\boldsymbol{X}$, we would like to define $E(Y \mid \boldsymbol{X}=\boldsymbol{x})$ for a random variable Y with $E[|Y|]<\infty$.
- Since the event $[\boldsymbol{X}=\boldsymbol{x}]$ has probability 0 , this is somewhat delicate from a technical point of view, although if Y has a conditional density function given $\boldsymbol{X}=\boldsymbol{x}, f(y \mid \boldsymbol{x})$ then we can define

$$E(Y \mid \boldsymbol{X}=\boldsymbol{x})=\int_{-\infty}^{\infty} y f(y \mid \boldsymbol{x}) d y$$

- We can obtain similar expressions for $E[g(\boldsymbol{Y}) \mid \boldsymbol{X}=\boldsymbol{x}]$ provided that we can define the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}=\boldsymbol{x}$ in a satisfactory way.

Elementary property of conditional expectation

Proposition

Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors. Then

(a) *if $E\left[|g_1(\boldsymbol{Y})|\right]$ and $E\left[|g_2(\boldsymbol{Y})|\right]$ are finite,*

$$\begin{aligned} &E\left[ag_1(\boldsymbol{Y}) + bg_2(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\ &\quad = aE\left[g_1(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\right] + bE\left[g_2(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\right] \end{aligned}$$

(b) $E\left[g_1(\boldsymbol{X})g_2(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\right] = g_1(\boldsymbol{x})E\left[g_2(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\right]$ *if* $E\left[|g_2(\boldsymbol{Y})|\right]$ *is finite;*

(c) *If* $h(x) = E[g(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}]$ *then* $E[h(\boldsymbol{X})] = E[g(\boldsymbol{Y})]$ *if* $E[|g(\boldsymbol{Y})|]$ *is finite.*

Variance decomposition

Theorem

Suppose that Y is a random variable with finite variance. Then

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y \mid \boldsymbol{X})] + \mathrm{Var}[E(Y \mid \boldsymbol{X})]$$

where $\mathrm{Var}(Y \mid \boldsymbol{X}) = E\left[(Y - E(Y \mid \boldsymbol{X}))^2 \mid \boldsymbol{X}\right]$*.*