Advanced Econometrics

Lecture 3: Review of Probability

Instructor: Ma, Jun

Renmin University of China

September 26, 2021

Randomness, sample space and probability

- Probability is concerned with *random experiments*.
- The outcome cannot be predicted with certainty, even if the experiment is repeated under the same conditions.
- The set of all possible outcomes is called a *sample space*, denoted by Ω. A simple example is tossing a coin. There are two outcomes, heads and tails, so we can write $\Omega = \{H, T\}$. Another simple example is rolling a dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$. A sample space may contain finite or infinite number of outcomes.
- The random experiment under the ground of statistics and econometrics should be viewed as an abstract one: nature draws a state of the world.
- A collection of elements of Ω is called an *event.* In the rolling a dice example, the event $A = \{2, 4, 6\}$ occurs if the outcome of the experiment is an even number.

- Probability function assigns probabilities (numbers between 0 and 1) to the events.
- A probability function has to satisfy the following *axioms of probability*:
    1. $\Pr(\Omega) = 1$.
    2. For any event A, $\Pr(A) \geq 0$.
    3. If $A_1, A_2, \ldots$ is a *countable* sequence of *mutually exclusive* events, then $\Pr(A_1 \cup A_2 \cup \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$.

  Some important properties of probability include:
    - If $A \subset B$ then $\Pr(A) \leq \Pr(B)$.
    - $\Pr(A) \leq 1$.
    - $\Pr(A) = 1 - \Pr(A^c)$.
    - $\Pr(\varnothing) = 0$.
    - $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

  A sample space, its collection of events and a probability function together define a probability space.

Conditional probability and independence

- If $\Pr(B) > 0$, the *conditional probability* of an event A, conditional on B is defined as follows:
$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$
- Conditional probability gives the probability of A knowing that B has occurred. For a given B, the conditional probability function $\Pr(\cdot \mid B)$ is a proper probability function.
- Conditioning can be seen as updating of the sample space based on new information.
- Probability of events A and B occurring jointly is given by the probability of their intersection $\Pr(A \cap B)$. The events A and B are called *independent* if the probability of their occurring together is equal to the product of their individual probabilities: $\Pr(A \cap B) = \Pr(A)\Pr(B)$.
- If A and B are independent, then the fact that B has occurred provides us with no information regarding occurrence of A : $\Pr(A \mid B) = \Pr(A)$.
- If A and B are independent, then so are A^c and B, A and B^c, A^c and B^c: if B cannot provide information about occurrence of A, then it also cannot tell us whether A did not occur (A^c).

Random variables

- A *random variable* is a *function* from a sample space to the real line. For every $\omega \in \Omega$, a random variable $X(\omega)$ assigns a number $x \in \mathbb{R}$.
- For example, in the tossing a coin experiment, we can define a random variable that takes on the value 0 if the outcome of the experiment is heads, and 1 if the outcome is tails: $X(H) = 0$, $X(T) = 1$.
- One can define many different random variables on the same sample space.
- A common convention is to denote random variables by capital letters, and to denote realized values by small letters.
- One can speak about the probability of a random variable taking on a particular value $\Pr(X = x)$, where $x \in \mathbb{R}$, or more generally, a probability of X taking a value in some subset of the real line $\Pr(X \in S)$, where $S \subset \mathbb{R}$, for example $S = (-\infty, 2)$. The probability of such an event is defined by the probability of the corresponding subset of the original sample space $\Omega : \Pr(X \in S) = \Pr\{\omega \in \Omega : X(\omega) \in S\}$.
- For example, suppose that in the flipping a coin example X is defined as above. Then $\Pr(X < 2)$ is given by the probability of the event $\{H, T\}$, $\Pr(X \in (0.3, 5)) = \Pr(\{T\})$, and $\Pr(X > 1.2) = \Pr(\emptyset) = 0$.

Cumulative distribution function

- For a random variable X, its *cumulative distribution function* (CDF) is defined as
$$F_X(x) = \Pr(X \le x).$$
- A CDF must be defined for all $u \in \mathbb{R}$, and satisfy the following conditions:
  1. $\lim_{u \to -\infty} F(u) = 0$, $\lim_{u \to \infty} F(u) = 1$.
  2. $F(x) \le F(y)$ if $x \le y$ (nondecreasing).
  3. $\lim_{u \downarrow x} F(u) = F(x)$ (right-continuous).

Discrete and continuous random variables

- A random variable is called *discrete* if its CDF is a step function. In this case, there exists a *countable* set of real number $X \in \{x_1, x_2, \ldots\}$ such that $\Pr(X = x_i) = p_X(x_i) > 0$ and $\sum_i p_X(x_i) = 1$. This set is called the support of a distribution, it contains all the values that X can take with probability different from zero.
- The values $p_X(x_i)$ give a *probability mass function* (PMF).
- A random variable is continuous if its CDF is a continuous function. In this case, $\Pr(X = x) = 0$ for all $x \in \mathbb{R}$, so it is impossible to describe the distribution of X by specifying probabilities at various points on the real line.
- Instead, the distribution of a continuous random variable can be described by a *probability density function* (PDF), which is defined as

$$f_X(x) = \left.\frac{dF_X(u)}{du}\right|_{u=x}.$$

Thus, $F_X(x) = \int_{-\infty}^{x} f_X(u)du$, and $\Pr(X \in (a,b)) = \int_a^b f_X(u)du$. Since the CDF is nondecreasing, $f(x) \geq 0$ for all $x \in \mathbb{R}$. Further, $\int_{-\infty}^{\infty} f_X(u)du = 1$.

Random vectors, multivariate and conditional distributions

- In economics we are usually concerned with relationships between a number of variables. Thus, we need to consider *joint* behavior of several random variables defined on the *same* probability space.
- A *random vector* is a function from the sample space Ω to $\mathbb{R}^n$.
- The random vector X is given by

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

By convention, a random vector is usually a column vector.

- Let $\boldsymbol{x} \in \mathbb{R}^n$, i.e. $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$. The CDF of a vector or a *joint* CDF of its elements is defined as follows:

$$F(x_1, x_2, \ldots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n) \text{ for all } \boldsymbol{x} \in \mathbb{R}^n.$$

  If the joint CDF is a continuous function, then the corresponding joint PDF is given by

$$f(x_1, x_2, \ldots, x_n) = \left. \frac{\partial^n F(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n} \right|_{u_1 = x_1, u_2 = x_2, \ldots, u_n = x_n},$$

  and thus,

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n) \, du_n \ldots du_2 du_1.$$

- It is possible from the joint distribution to obtain the individual distribution of a single element of the random vector (*marginal* distribution), or the joint distribution of a number of its elements.

- Consider, a bivariate case. Let X and Y be two random variables with the CDF and PDF given by $F_{X,Y}$ and $f_{X,Y}$ respectively. The marginal CDF of X is

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(X \leq x, -\infty < Y < \infty) \;\; (Y \text{ can take any value}) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u,v)\, dv du. \end{aligned}$$

- Now, the marginal PDF of X is

$$\begin{aligned} \frac{dF_X(x)}{dx} &= \frac{d}{dx} \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u,v)\, dv du \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x,v)\, dv. \end{aligned}$$

- In a discrete case, one can obtain a marginal PMF from the joint in a similar way, by using sums instead of integrals:

$$p_Y(y_j) = \sum_{i=1}^{n} p_{X,Y}\left(x_i, y_j\right).$$

- In general, it is impossible to obtain a joint distribution from the marginal distributions.

- *Conditional distribution* describes the distribution of one random variable (vector) conditional on another random variable (vector). In the continuous case, conditional PDF and CDF of X given Y is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$
$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(u \mid y)\, du,$$

  respectively, for $f_Y(y) > 0$.
- In the discrete case, suppose that with a probability greater than zero X takes values in $\{x_1, x_2, \ldots, x_n\}$, and Y takes values in $\{y_1, y_2, \ldots, y_k\}$. Let $p_{X,Y}(x_i, y_j)$ be the joint PMF. Then the conditional PMF of X conditional on Y is given by

$$p_{X|Y}\left(x \mid y_j\right) = \frac{p_{X,Y}\left(x, y_j\right)}{p_Y\left(y_j\right)} \text{ for } j = 1, 2, \ldots, k.$$

- It is important to distinguish between $f_{X|Y}(x \mid y)$ and $f_{X|Y}(x \mid Y)$. The first means that Y is fixed at some realized value y, and $f_{X|Y}(x \mid y)$ is not a random function. On the other hand, notation $f_{X|Y}(x \mid Y)$ means that uncertainty about Y remains, and, consequently, $f_{X|Y}(x \mid Y)$ is a random function.
- The concept of *independent random variables* is related to that of the events. Suppose that for *all* pairs of subsets of the real line, S_1 and S_2, we have that the events $X \in S_1$ and $Y \in S_2$ are independent, i.e.

$$\Pr(X \in S_1, Y \in S_2) = \Pr(X \in S_1)\Pr(Y \in S_2). \tag{0.1}$$

- In the continuous case, random variables are independent if and only if there joint PDF can be expressed as a product of their marginal PDFs:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \text{ for all } x \in \mathbb{R},\ y \in \mathbb{R}.$$

- Consequently, independence implies that for all $x \in \mathbb{R}$, $y \in \mathbb{R}$ such that $f_Y(y) > 0$, we have that

$$f_{X|Y}(x \mid y) = f_X(x).$$

- For any functions g and h, if X and Y are independent, then so are $g(X)$ and $h(Y)$.

Expectation and moments

- Given a random variable X its *mean*, or *expectation*, or *expected value* defined as

$$\mathrm{E}\left(X\right) = \sum_i x_i p_X\left(x_i\right) \text{ in the discrete case,}$$

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} x f_X(x) dx \text{ in the continuous case.}$$

- Note that $\int_{-\infty}^{0} x f_X(x) dx$ or $\int_{0}^{\infty} x f_X(x) dx$ can be infinite. In such cases, we say that expectation does not exist, and assign $\mathrm{E}\left(X\right) = -\infty$ if $\int_{-\infty}^{0} x f_X(x) dx = -\infty$ and $\int_{0}^{\infty} x f_X(x) dx < \infty$, and $\mathrm{E}\left(X\right) = \infty$ if $\int_{-\infty}^{0} x f_X(x) dx > -\infty$ and $\int_{0}^{\infty} x f_X(x) dx = \infty$. When $\int_{-\infty}^{0} x f_X(x) dx = -\infty$ and $\int_{0}^{\infty} x f_X(x) dx = \infty$, the expectation is not defined.
- The necessary and sufficient condition for $\mathrm{E}\left(X\right)$ to be defined and finite is that $\mathrm{E}\left|X\right| < \infty$.

- Let g be a function. The expected value of $g(X)$ is defined as

$$\mathrm{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- The k-th *moment* of a random variable X is defined as $\mathrm{E}\left(X^k\right)$. The first moment if simply the mean. The k-th *central moment* X is $\mathrm{E}(X - \mathrm{E}X)^k$. The second central moment is called the *variance*:

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}(X - \mathrm{E}X)^2 \\ &= \int_{-\infty}^{\infty} (x - \mathrm{E}X)^2 f_X(x) dx. \end{aligned}$$

  While the mean measures the center of the distribution, the variance is a measure of the spread of the distribution.

Existence of moments

- If $\mathrm{E}\,|X|^n = \infty$, we say that the n-th moment does not exist.
- Let X be a random variable, and let $n > 0$ be an integer. If $\mathrm{E}\,|X|^n < \infty$ and m is an integer such that $m \leq n$, then $\mathrm{E}\,|X|^m < \infty$.

Covariance

- For a function of two random variables, $h(X,Y)$, its expectation is defined as

$$\mathrm{E}\left[h(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y)dxdy.$$

- *Covariance* of two random variable X and Y is defined as

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\left(X-\mathrm{E}X\right)\left(Y-\mathrm{E}Y\right)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-\mathrm{E}X\right)\left(y-\mathrm{E}Y\right) f_{X,Y}(x,y)dxdy.\end{aligned}$$

- The correlation coefficient of X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{E}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- The correlation coefficient is bounded between -1 and 1. It is equal to -1 or 1 if and only if, one random variable is a *linear* function of another: $Y = a + bX$.

Let a, b and c be some constants. Some useful properties include:

- Linearity of expectation: $\mathrm{E}\left(aX + bY + c\right) = a\mathrm{E}X + b\mathrm{E}Y + c$.
- $\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(aX + bY, cZ) = ac\mathrm{Cov}(X, Z) + bc\mathrm{Cov}(Y, Z)$.
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- $\mathrm{Cov}(X, a) = 0$.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{E}(X - \mathrm{E}X) = 0$.
- $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.
- $\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - (\mathrm{E}X)^2$.
- If X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{Cov}(X, Y) = 0$. However, zero correlation (*uncorrelatedness*) does not imply independence.

Moments of random vectors (matrices)

For a random vector (matrix), the expectation is defined as a vector (matrix) composed of expected values of its corresponding elements:

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{X}\right] &= \mathrm{E}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{E}X_1 \\ \mathrm{E}X_2 \\ \vdots \\ \mathrm{E}X_n \end{pmatrix}.
\end{aligned}
$$

- The *variance-covariance matrix* of a random n-vector is a $n \times n$ matrix defined as

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{X}) &= \mathrm{E}\,(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})\,(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})' \\
&= \mathrm{E}\begin{pmatrix} X_1 - \mathrm{E}X_1 \\ X_2 - \mathrm{E}X_2 \\ \vdots \\ X_n - \mathrm{E}X_n \end{pmatrix}\begin{pmatrix} X_1 - \mathrm{E}X_1 & X_2 - \mathrm{E}X_2 & \dots & X_n - \mathrm{E}X_n \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{Var}((X_1) & \mathrm{Cov}\,(X_1, X_2) & \dots & \mathrm{Cov}\,(X_1, X_n) \\ \mathrm{Cov}\,(X_2, X_1) & \mathrm{Var}((X_2) & \dots & \mathrm{Cov}\,(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}\,(X_n, X_1) & \mathrm{Cov}\,(X_n, X_2) & \dots & \mathrm{Var}\,(X_n) \end{pmatrix}.
\end{aligned}
$$

- It is a symmetric, positive semi-definite matrix, with variances on the main diagonal and covariances off the main diagonal.
- The variance-covariance matrix is positive semi-definite (denoted by $\mathrm{Var}(\boldsymbol{X}) \geq 0$), since for any n-vector of constants $\boldsymbol{a}$, we have that $\boldsymbol{a}'\mathrm{Var}(\boldsymbol{X})\boldsymbol{a} \geq 0$.

Let $\boldsymbol{X} \in \mathbb{R}^{n}$ and $\boldsymbol{Y} \in \mathbb{R}^{k}$ be two random vectors. Their covariance of $\boldsymbol{X}$ with $\boldsymbol{Y}$ is a $n \times k$ matrix defined as

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}) &= \mathrm{E}\left(\boldsymbol{X} - \mathrm{E}\boldsymbol{X}\right)\left(\boldsymbol{Y} - \mathrm{E}\boldsymbol{Y}\right)' \\
&= \begin{pmatrix}
\operatorname{Cov}(X_1, Y_1) & \operatorname{Cov}(X_1, Y_2) & \ldots & \operatorname{Cov}(X_1, Y_k) \\
\operatorname{Cov}(X_2, Y_1) & \operatorname{Cov}(X_2, Y_2) & \ldots & \operatorname{Cov}(X_2, Y_k) \\
\ldots & \ldots & \ldots & \ldots \\
\operatorname{Cov}(X_n, Y_1) & \operatorname{Cov}(X_n, Y_2) & \ldots & \operatorname{Cov}(X_n, Y_k)
\end{pmatrix}.
\end{aligned}$$

Some useful properties:

- $\text{Var}(\boldsymbol{X}) = \text{E}(\boldsymbol{X}\boldsymbol{X}') - \text{E}(\boldsymbol{X})\,\text{E}(\boldsymbol{X})'$.
- $\text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = (\text{Cov}(\boldsymbol{Y}, \boldsymbol{X}))'$.
- $\text{Var}(\boldsymbol{X} + \boldsymbol{Y}) = \text{Var}(\boldsymbol{X}) + \text{Var}(\boldsymbol{Y}) + \text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \text{Cov}(\boldsymbol{Y}, \boldsymbol{X})$.
- If $\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{X}$, where $\boldsymbol{\alpha} \in \mathbb{R}^k$ is a fixed (non-random) vector and $\boldsymbol{\Gamma}$ is a $k \times n$ fixed matrix, then $\text{Var}(\boldsymbol{Y}) = \boldsymbol{\Gamma}(\text{Var}(\boldsymbol{X}))\boldsymbol{\Gamma}'$.
- For random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and non-random matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$: $\text{Cov}(\mathbf{A}\boldsymbol{X} + \mathbf{B}\boldsymbol{Y}, \mathbf{C}\boldsymbol{Z}) = \mathbf{A}(\text{Cov}(\boldsymbol{X}, \boldsymbol{Z}))\mathbf{C}' + \mathbf{B}(\text{Cov}(\boldsymbol{Y}, \boldsymbol{Z}))\mathbf{C}'$.

Normal distribution

- For $x \in \mathbb{R}$, the density function (PDF) of a normal distribution is given by
$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$
where μ and σ^2 are the two *parameters* determining the distribution. The common notation for a normally distributed random variable is $X \sim N(\mu,\sigma^2)$. The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard normal* distribution.
- The joint PDF of $\boldsymbol{X} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is given by
$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-n/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp\left(-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right), \boldsymbol{x} \in \mathbb{R}^n,$$
where $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.
- Let $\boldsymbol{X} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and define $\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{X}$. Then $\boldsymbol{Y} \sim N(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}')$.

Other useful statistical distributions

The following distributions are related to normal and used extensively in statistical inference:

- Suppose that $\mathbf{Z} \sim N(0, \mathbf{I}_n)$, so the elements of $\mathbf{Z}$, $Z_1, Z_2, \ldots, Z_n$ are independent identically distributed standard normal random variables. Then $X = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} Z_i^2$ has a *chi-square distribution* with n degrees of freedom. It is conventional to write $X \sim \chi_n^2$. The mean of the χ_n^2 distribution is n and the variance is $2n$. If $X_1 \sim \chi_{n_1}^2$, $X_2 \sim \chi_{n_2}^2$ and independent, then $X_1 + X_2 \sim \chi_{n_1+n_2}^2$.
- Let $Z \sim N(0,1)$ and $X \sim \chi_n^2$ be independent, then $Y = Z/\sqrt{X/n}$ has a *t distribution* with n degrees of freedom ($Y \sim t_n$). For large n, the density of t_n approaches that of $N(0,1)$. The mean of t_n does not exists for $n = 1$, and zero for $n > 1$. The variance of the t_n distribution is $n/(n-2)$ for $n > 2$.
- Let $X_1 \sim \chi_{n_1}^2$ and $X_2 \sim \chi_{n_2}^2$ be independent, then $Y = \frac{X_1/n_1}{X_2/n_2}$ has an *F distribution* with n_1, n_2 degrees of freedom ($Y \sim F_{n_1,n_2}$). $F_{1,n} = (t_n)^2$.