# Lecture 1: Preliminaries I. Probability

School of Mathematical Sciences, Xiamen University

## 1. Discrete probability

- $\Omega$: the sample space
- The *probability measure* or *probability function* $\mathbb{P}[\omega]$ maps the sample space $\Omega$ to the interval $[0, 1]$. This function has the so-called normalization property,

$$\sum_{\omega \in \Omega} \mathbb{P}[\omega] = 1.$$

- If $\mathcal{E} \subset \Omega$ is an event, then

$$\mathbb{P}[\mathcal{E}] = \sum_{\omega \in \mathcal{E}} \mathbb{P}[\omega],$$

namely the probability of an event is the sum of the probabilities of its elements.

- The union of two events:

$$\mathbb{P}[\mathcal{E}_1 \cup \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2] - \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2].$$

- The union bound

$$\mathbb{P}\left[\bigcup_{i=1}^{n} \mathcal{E}_i\right] \leq \sum_{i=1}^{n} \mathbb{P}[\mathcal{E}_i].$$

- Disjoint or mutually exclusive events:

$$\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset.$$

  This can be generalized to any number of events by necessitating that the events are all pairwise disjoint.

- Let $\overline{\mathcal{E}}$ denote the complement of the event $\mathcal{E}$

$$\mathbb{P}[\overline{\mathcal{E}}] = 1 - \mathbb{P}[\mathcal{E}].$$

## 1.1 Conditional probability

- For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$, the conditional probability $\mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_2]$ is the probability that $\mathcal{E}_1$ occurs given that $\mathcal{E}_2$ occurs. Formally,

$$\mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_2] = \frac{\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2]}{\mathbb{P}[\mathcal{E}_2]}.$$

- Bayes rules: for any two events $\mathcal{E}_1$ and $\mathcal{E}_2$ such that $\mathbb{P}[\mathcal{E}_1] > 0$ and $\mathbb{P}[\mathcal{E}_2] > 0$,

$$\mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1] = \frac{\mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_2] \cdot \mathbb{P}[\mathcal{E}_2]}{\mathbb{P}[\mathcal{E}_1]}.$$

- By $\Omega = \mathcal{E}_2 \cup \overline{\mathcal{E}_2}$, we have

$$\begin{aligned} \mathbb{P}[\mathcal{E}_1] &= \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] + \mathbb{P}[\mathcal{E}_1 \cap \overline{\mathcal{E}_2}] \\ &= \mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_2] \cdot \mathbb{P}[\mathcal{E}_2] + \mathbb{P}[\mathcal{E}_1 \mid \overline{\mathcal{E}_2}] \cdot \mathbb{P}[\overline{\mathcal{E}_2}]. \end{aligned}$$

## 1.2 Independent events

- Two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are called independent if

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] \cdot \mathbb{P}[\mathcal{E}_2].$$

  This can be generalized to more than two events by necessitating that the events are all pairwise independent.

- For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$ such that $\mathbb{P}[\mathcal{E}_1] > 0$ and $\mathbb{P}[\mathcal{E}_2] > 0$ the following three statements are equivalent:

$$\mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1],$$

$$\mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1] = \mathbb{P}[\mathcal{E}_2],$$

  and

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] = \mathbb{P}[\mathcal{E}_1] \cdot \mathbb{P}[\mathcal{E}_2].$$

## 2. Random variables

- Random variables are functions mapping the sample space $\Omega$ to the real numbers $\mathbb{R}$.

  Note that even though they are called variables, in reality they are functions.

- Let $\Omega$ be the sample space of a random experiment. A formal definition for the random variable $X$ would be as follows: let $\alpha \in \mathbb{R}$ be a real number (not necessarily positive) and note that the function

$$X^{-1}(\alpha) = \{\omega \in \Omega : X(\omega) = \alpha\}$$

  returns a subset of $\Omega$ and thus is an event. Therefore, the function $X^{-1}(\alpha)$ has a probability.

- We will abuse notation and write $\mathbb{P}[X = \alpha]$ instead of the more proper notation $\mathbb{P}[X^{-1}(\alpha)]$, i.e.,

$$\begin{aligned} \mathbb{P}[X = \alpha] &= \mathbb{P}[X^{-1}(\alpha)] \\ &= \mathbb{P}[\omega \in \Omega : X(\omega) = \alpha]. \end{aligned}$$

  This function of $\alpha$ is of great interest and it is easy to generalize as follows:

$$\begin{aligned} \mathbb{P}[X \leq \alpha] &= \mathbb{P}[X^{-1}(\beta) : \beta \in (-\infty, \alpha]] \\ &= \mathbb{P}[\omega \in \Omega : X(\omega) \leq \alpha]. \end{aligned}$$

- **Independent random variables**: Two random variables $X$ and $Y$ are independent if for all $a, b \in \mathbb{R}$,

$$\mathbb{P}[X = a \text{ and } Y = b] = \mathbb{P}[X = a] \cdot \mathbb{P}[Y = b].$$

## 2.1 PMF and CDF

- Probability mass function (PMF) measures the probability that a random variable $X$ takes a particular value $\alpha \in \mathbb{R}$:

$$f(\alpha) = \mathbb{P}[X = \alpha].$$

- Cumulative distribution function (CDF) measures the probability that a random variable $X$ takes any value below $\alpha \in \mathbb{R}$:

$$F(\alpha) = \mathbb{P}[X \leq \alpha].$$

It is obvious from the above definitions that

$$F(\alpha) = \sum_{x \leq \alpha} f(x).$$

## 2.2 Expectation (mean)

- Given a random variable $X$, its expectation $\mathbb{E}[X]$ is defined as

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}[X = x] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}[\omega],$$

  where $X(\Omega)$ is the image of the random variable $X$ over the sample space $\Omega$. Note that $\mathbb{E}[f(X)] = \sum_{x \in X(\Omega)} f(x) \cdot \mathbb{P}[X = x]$.

- The most important property is linearity of expectation: for any random variables $X$ and $Y$ and real number $\lambda$,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \mathbb{E}[\lambda X] = \lambda \mathbb{E}[X].$$

- If two random variables $X$ and $Y$ are independent then we can manipulate the expectation of their product as follows:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

## 2.3 Variance and covariance

- Given a random variable $X$, its variance $\mathbb{V}[X]$ is defined as

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

  Obviously, it holds $\mathbb{V}[X] \leq \mathbb{E}[X^2]$.

- The covariance of two random variables $X$ and $Y$ is defined as

$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

  $X$ and $Y$ are said to be uncorrelated if $\mathbb{C}\mathrm{ov}(X, Y) = 0$.

- If the random variables $X$ and $Y$ are independent, then

$$\mathbb{C}\mathrm{ov}(X, Y) = 0 \quad \text{and} \quad \mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y].$$

  Also, for any real $\lambda$, it holds $\mathbb{V}[\lambda X] = \lambda^2 \mathbb{V}[X]$.

- The standard deviation is the square root of the variance and is often denoted by $\mathbb{S}\mathrm{td}(X) = \sqrt{\mathbb{V}(X)}$.

## 2.4 Markov's inequality

- Let $X$ be a non-negative random variable. For any $\alpha > 0$,

$$\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}.$$

*Proof.* For any $\alpha > 0$, define the following function

$$f(X) = \begin{cases} 1, & \text{if } X \geq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(X) \leq X/\alpha$, which yields $\mathbb{E}[f(X)] \leq \mathbb{E}[X]/\alpha$. It follows from

$$\mathbb{E}[f(X)] = 1 \cdot \mathbb{P}[X \geq \alpha] + 0 \cdot \mathbb{P}[X < \alpha] = \mathbb{P}[X \geq \alpha]$$

that

$$\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}. \quad \square$$

## 3. Random vectors

- The expectation of an $n \times 1$ random vector, $\boldsymbol{X}$, is the vector of expectations of each entry (provided they exist):

$$\mathbb{E}\boldsymbol{X} = (\mathbb{E}X_1, \cdots, \mathbb{E}X_n)^\top.$$

  The expectation of a random matrix is also defined as the matrix consisting of the expectations of each entry.

- The variance of $\boldsymbol{X}$ is defined as the $n \times n$ symmetric matrix:

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top\right] = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}_{\boldsymbol{X}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top =: \boldsymbol{\Sigma},$$

  with $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}\boldsymbol{X}$. The covariance of two random variables $X_i$ and $X_j$ is the $(i, j)$ entry of $\boldsymbol{\Sigma}$, i.e.,

$$\mathbb{C}\mathrm{ov}(X_i, X_j) = \Sigma_{ij}.$$

  We also call $\boldsymbol{\Sigma}$ the covariance matrix of $\boldsymbol{X}$.

- The covariance (or cross-covariance) of $\boldsymbol{X}$ with a second $m \times 1$ random vector, $\boldsymbol{Y}$, of mean $\boldsymbol{\mu_Y}$ is the $n \times m$ matrix,

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{Y}) &= \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^\top\right] \\ &= \mathbb{E}(\boldsymbol{X}\boldsymbol{Y}^\top) - \boldsymbol{\mu_X}\boldsymbol{\mu_Y}^\top,\end{aligned}$$

and, as in the scalar case,

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X}) = \mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{X}).$$

Note that

$$\mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{Y}) = (\mathbb{C}\mathrm{ov}(\boldsymbol{Y}, \boldsymbol{X}))^\top.$$

## 4. Properties of expectation, variance, and covariance

- $\mathbb{E}$ is order preserving:

$$\mathbb{E}X \leq \mathbb{E}Y, \quad \text{if} \quad X \leq Y.$$

- Cauchy-Schwarz inequality:

  If $X$ and $Y$ have finite variances, then $|\mathbb{E}(XY)| < \infty$ and

$$|\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

  In particular,

$$|\mathbb{C}\mathrm{ov}(X, Y)| \leq \mathbb{S}\mathrm{td}(X)\mathbb{S}\mathrm{td}(Y).$$

  More generally,

$$|\mathbb{E}(\boldsymbol{X}^\top \boldsymbol{Y})| \leq \mathbb{E}|\boldsymbol{X}^\top \boldsymbol{Y}| \leq \sqrt{\mathbb{E}(\|\boldsymbol{X}\|^2)\mathbb{E}(\|\boldsymbol{Y}\|^2)}.$$

- $\mathbb{E}$ is linear: For any constants $a$ and $b$,

$$\mathbb{E}(a\boldsymbol{X} + b\boldsymbol{Y}) = a\mathbb{E}\boldsymbol{X} + b\mathbb{E}\boldsymbol{Y}.$$

- $\mathbb{C}\mathrm{ov}$ is bilinear and shift invariant:

  For any constants $a$ and $b$ and fixed vector $\mathbf{c}$,

$$\mathbb{C}\mathrm{ov}(a\boldsymbol{X} + b\boldsymbol{Y} + \mathbf{c}, \boldsymbol{Z}) = a\mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{Z}) + b\mathbb{C}\mathrm{ov}(\boldsymbol{Y}, \boldsymbol{Z}),$$

$$\mathbb{C}\mathrm{ov}(\boldsymbol{Z}, a\boldsymbol{X} + b\boldsymbol{Y} + \mathbf{c}) = a\mathbb{C}\mathrm{ov}(\boldsymbol{Z}, \boldsymbol{X}) + b\mathbb{C}\mathrm{ov}(\boldsymbol{Z}, \boldsymbol{Y}).$$

  In particular,

$$\mathbb{V}\mathrm{ar}(X \pm Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) \pm 2\mathbb{C}\mathrm{ov}(X, Y),$$

  and

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X} \pm \boldsymbol{Y}) = \mathbb{V}\mathrm{ar}(\boldsymbol{X}) + \mathbb{V}\mathrm{ar}(\boldsymbol{Y}) \pm (\mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathbb{C}\mathrm{ov}(\boldsymbol{Y}, \boldsymbol{X})).$$

- Covariance transformation:

  For any matrices $\mathbf{A}$ and $\mathbf{B}$ (of appropriate sizes),

$$\mathbb{C}\mathrm{ov}(\mathbf{A}\boldsymbol{X}, \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\mathbb{C}\mathrm{ov}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}^\top.$$

  In particular,

$$\mathbb{V}\mathrm{ar}(aX) = a^2\mathbb{V}\mathrm{ar}(X), \quad \mathbb{V}\mathrm{ar}(\mathbf{A}\boldsymbol{X}) = \mathbf{A}\mathbb{V}\mathrm{ar}(\boldsymbol{X})\mathbf{A}^\top.$$

- Expectation of a quadratic form:

  If $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, then

$$\mathbb{E}(\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}) = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}[\mathbf{A}\mathbb{V}\mathrm{ar}(\boldsymbol{X})],$$

  where tr denotes the trace of the matrix.

- The law of total expectation

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]\right]$$

- Jensen's inequality:

  If $\psi$ is a convex function, then

$$\psi(\mathbb{E}X) \leq \mathbb{E}\psi(X).$$

  In particular, $|\mathbb{E}X| \leq \mathbb{E}|X|$ and $\|\mathbb{E}\boldsymbol{X}\| \leq \mathbb{E}\|\boldsymbol{X}\|$.

- Markov's inequality:

  If $X$ is a random variable with $\mathbb{E}|X| < \infty$, then for any $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq \mathbb{E}|X|/t.$$

- Association inequality:

  If $X$ is a random variable and $f$ and $g$ are nondecreasing functions, then

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)].$$