

## Background

Lecturer: Xiaojin Zhu

jerryzhu@cs.wisc.edu

## 1 Probability Measure

A *sample space*  $\Omega$  is the set of all possible outcomes. Elements  $\omega \in \Omega$  are called *sample outcomes*, while subsets  $A \subseteq \Omega$  are called *events*. For example, for a die roll,  $\Omega = \{1, 2, 3, 4, 5, 6\}$ ,  $\omega = 5$  is an outcome,  $A_1 = \{5\}$  is the event that the outcome is 5, and  $A_2 = \{1, 3, 5\}$  is the event that the outcome is odd. Ideally, one would like to be able to assign probabilities to all  $A$ s. This is trivial for finite  $\Omega$ . However, when  $\Omega = \mathbb{R}$  strange things happen: it is no longer possible to consistently assign probabilities to all subsets of  $\mathbb{R}$ .

Instead, we will restrict ourselves to only some of the events. A  $\sigma$ -algebra  $\mathcal{B}$  is a set of  $\Omega$  subsets satisfying:

1.  $\emptyset \in \mathcal{B}$
2. if  $A \in \mathcal{B}$  then  $A^c \in \mathcal{B}$  (complementarity)
3. if  $A_1, A_2, \dots \in \mathcal{B}$  then  $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$  (countable unions)

The sets in  $\mathcal{B}$  are called *measurable*. The pair  $(\Omega, \mathcal{B})$  is called a *measurable space*. For  $\Omega = \mathbb{R}$ , we take the smallest  $\sigma$ -algebra that contains all the open subsets and call it the *Borel  $\sigma$ -algebra*.

A *measure* is a function  $P : \mathcal{B} \mapsto \mathbb{R}$  satisfying:

1.  $P(A) \geq 0$  for all  $A \in \mathcal{B}$
2. If  $A_1, A_2, \dots$  are disjoint then  $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Note these imply  $P(\emptyset) = 0$ . The triple  $(\Omega, \mathcal{B}, P)$  is called a *measure space*. The *Lebesgue measure* is a uniform measure over the Borel  $\sigma$ -algebra with the usual meaning of length, area or volume, depending on dimensionality. For example, for  $\mathbb{R}$  the Lebesgue measure of the interval  $(a, b)$  is  $b - a$ .

A *probability measure* is a measure satisfying additionally the normalization constraint:

- 3  $P(\Omega) = 1$ .

Such a triple  $(\Omega, \mathcal{B}, P)$  is called a *probability space*.

## 2 Random Variables

Let  $(\Omega, \mathcal{B}, P)$  be a probability space. Let  $(\mathbb{R}, \mathcal{R})$  be the usual measurable space of reals and its Borel  $\sigma$ -algebra. A *random variable* is a function  $X : \Omega \mapsto \mathbb{R}$  such that the preimage of any set  $A \in \mathcal{R}$  is measurable in  $\mathcal{B}$ :  $X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{B}$ . This allows us to define the following (the first  $P$  is the new definition, while the 2nd and 3rd  $P$ s are the already-defined probability measure on  $\mathcal{B}$ ):

$$P(X \in A) = P(X^{-1}(A)) = P(\{\omega : X(\omega) \in A\}) \quad (1)$$

$$P(X = x) = P(X^{-1}(x)) = P(\{\omega : X(\omega) = x\}) \quad (2)$$

**Example 1** Let  $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$  be the unit disk. Consider drawing a point at random from  $\Omega$ . The outcome is of the form  $\omega = (x, y)$ . Some example random variables are  $X(\omega) = x, Y(\omega) = y, Z(\omega) = xy$ , and  $W(\omega) = \sqrt{x^2 + y^2}$ .

**Example 2** Let  $X = \omega$  be a uniform random variable (to be defined later) with the sample space  $\Omega = [0, 1]$ . A sequence of different random variables  $\{X_n\}_{n=1}^{\infty}$  can be defined as follows, where  $1\{z\} = 1$  if  $z$  is true, and 0 otherwise:

$$X_1(\omega) = \omega + 1\{\omega \in [0, 1]\} \quad (3)$$

$$X_2(\omega) = \omega + 1\{\omega \in [0, \frac{1}{2}]\} \quad (4)$$

$$X_3(\omega) = \omega + 1\{\omega \in [\frac{1}{2}, 1]\} \quad (5)$$

$$X_4(\omega) = \omega + 1\{\omega \in [0, \frac{1}{3}]\} \quad (6)$$

$$X_5(\omega) = \omega + 1\{\omega \in [\frac{1}{3}, \frac{2}{3}]\} \quad (7)$$

$$X_6(\omega) = \omega + 1\{\omega \in [\frac{2}{3}, 1]\} \quad (8)$$

...

Given a random variable  $X$ , the *cumulative distribution function* (CDF) is the function  $F_X : \mathbb{R} \mapsto [0, 1]$

$$F_X(x) = P(X \leq x) = P(X \in (-\infty, x]) = P(\{\omega : X(\omega) \in (-\infty, x]\}). \quad (9)$$

A random variable  $X$  is *discrete* if it takes countably many values. We define the probability mass function  $f_X(x) = P(X = x)$ .

A random variable  $X$  is *continuous* if there exists a function  $f_X$  such that

1.  $f_X(x) \geq 0$  for all  $x \in \mathbb{R}$
2.  $\int_{-\infty}^{\infty} f_X(x) dx = 1$
3. for every  $a \leq b$ ,  $P(X \in [a, b]) = \int_a^b f_X(x) dx$ .

The function  $f_X$  is called the *probability density function* (PDF) of  $X$ . CDF and PDF are related by

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (10)$$

$$f_X(x) = F'_X(x) \text{ at all points } x \text{ at which } F_X \text{ is differentiable.} \quad (11)$$

### 3 Some Random Variables

#### 3.1 Discrete

**Dirac** or point mass distribution  $X \sim \delta_a$  if  $P(X = a) = 1$  with CDF  $F(x) = 0$  if  $x < a$  and 1 if  $x \geq a$ .

**Binomial.** A random variable  $X$  has a binomial distribution with parameters  $n$  (number of trials) and  $p$  (head probability)  $X \sim \text{Binomial}(n, p)$  with probability mass function

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

If  $X_1 \sim \text{Binomial}(n_1, p)$  and  $X_2 \sim \text{Binomial}(n_2, p)$  then  $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$ . Think this as merging two coin flip experiments on the same coin.

**Bernoulli.** Binomial with  $n = 1$ .

**Multinomial.** The  $d$ -dimensional version of binomial. The parameter  $p = (p_1, \dots, p_d)^\top$  is now the probabilities of a  $d$ -sided die, and  $x = (x_1, \dots, x_d)^\top$  is the counts of each face.

$$f(x) = \begin{cases} \binom{n}{x_1, \dots, x_d} \prod_{k=1}^d p_k^{x_k} & \text{if } \sum_{k=1}^d x_k = n \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

**Poisson.**  $X \sim \text{Poisson}(\lambda)$  if  $f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$  for  $x = 0, 1, 2, \dots$ . If  $X_1 \sim \text{Poisson}(\lambda_1)$  and  $X_2 \sim \text{Poisson}(\lambda_2)$  then  $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$ .

**Geometric.**  $X \sim \text{Geom}(p)$  if  $f(x) = p(1-p)^{x-1}$  for  $x = 1, 2, \dots$ .

### 3.2 Continuous

**Gaussian (also called Normal distribution).**  $X \sim N(\mu, \sigma^2)$  with parameters  $\mu \in \mathbb{R}$  (the mean) and  $\sigma^2$  (the variance) if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (14)$$

The square root of variance  $\sigma > 0$  is called the standard deviation. If  $\mu = 0, \sigma = 1$ ,  $X$  has a *standard normal distribution*. In this case,  $X$  is usually written as  $Z$ . Some useful properties:

- (Scaling) If  $X \sim N(\mu, \sigma^2)$ , then  $Z = (X - \mu)/\sigma \sim N(0, 1)$
- (Independent sum) If  $X_i \sim N(\mu_i, \sigma_i^2)$  are independent, then  $\sum_i X_i \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$

**$\chi^2$  distribution.** If  $Z_1, \dots, Z_k$  are independent standard normal random variables, then  $Y = \sum_i^k Z_i^2$  has a  $\chi^2$  distribution with  $k$  degrees of freedom. If  $X_i \sim N(\mu_i, \sigma_i^2)$  are independent, then  $\sum_i \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$  has a  $\chi^2$  distribution with  $k$  degrees of freedom. The PDF for the  $\chi^2$  distribution with  $k$  degrees of freedom is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0. \quad (15)$$

**Multivariate Gaussian.** Let  $x, \mu \in \mathbb{R}^d$ ,  $\Sigma \in S_+^d$  a symmetric, positive definite matrix of size  $d \times d$ . Then  $X \sim N(\mu, \Sigma)$  with PDF

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \quad (16)$$

Here,  $\mu$  is the mean vector,  $\Sigma$  is the covariance matrix,  $|\Sigma|$  its determinant, and  $\Sigma^{-1}$  its inverse (exists because  $\Sigma$  is positive definite).

If we have two (groups of) variables that are jointly Gaussian:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right) \quad (17)$$

then we have:

- (Marginal)  $x \sim N(\mu_x, A)$
- (Conditional)  $y|x \sim N(\mu_y + C^\top A^{-1}(x - \mu_x), B - C^\top A^{-1}C)$

**Exponential.**  $X \sim \text{Exp}(\beta)$  with parameter  $\beta > 0$ , if  $f(x) = \frac{1}{\beta} e^{-x/\beta}$ .

**Gamma.** The *Gamma function* (not distribution) is defined as  $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$  with  $\alpha > 0$ . The Gamma function is an extension to the factorial function:  $\Gamma(n) = (n-1)!$  when  $n$  is a positive integer. It

also satisfies  $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$  for  $\alpha > 0$ .  $X$  has a *Gamma distribution* with shape parameter  $\alpha > 0$  and scale parameter  $\beta > 0$ , denoted by  $X \sim \text{Gamma}(\alpha, \beta)$ , if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0. \quad (18)$$

$\text{Gamma}(1, \beta)$  is the same as  $\text{Exp}(\beta)$ .

**Beta.**  $X \sim \text{Beta}(\alpha, \beta)$  with parameters  $\alpha, \beta > 0$ , if

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in (0, 1). \quad (19)$$

$\text{Beta}(1, 1)$  is uniform in  $[0, 1]$ .  $\text{Beta}(\alpha < 1, \beta < 1)$  has a U-shape.  $\text{Beta}(\alpha > 1, \beta > 1)$  is unimodal with mean  $\alpha/(\alpha + \beta)$  and mode  $(\alpha - 1)/(\alpha + \beta - 2)$ .

**Dirichlet.** The Dirichlet distribution is the multivariate version of the beta distribution.  $X \sim \text{Dir}(\alpha_1, \dots, \alpha_d)$  with parameters  $\alpha_i > 0$ , if

$$f(x) = \frac{\Gamma(\sum_i^d \alpha_i)}{\prod_i^d \Gamma(\alpha_i)} \prod_i^d x_i^{\alpha_i-1}, \quad (20)$$

where  $x = (x_1, \dots, x_d)$  with  $x_i > 0$ ,  $\sum_i^d x_i = 1$ . This support is called the open  $(d - 1)$  dimensional simplex.

**$t$  (or Student's  $t$ ) Distribution**  $X \sim t_\nu$  has a  $t$  distribution with  $\nu$  degrees of freedom, if

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \quad (21)$$

It is similar to a Gaussian distribution, but with heavier tails (more likely to produce extreme values). When  $\nu = \infty$ , it becomes a standard Gaussian distribution.

**Cauchy.** The Cauchy distribution is a special case of  $t$ -distribution with  $\nu = 1$ . The PDF is

$$f(x) = \frac{1}{\pi\gamma \left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)} \quad (22)$$

where  $x_0$  is the location parameter for the mode, and  $\gamma$  is the scale parameter. It is notable for the lack of mean:  $\mathbb{E}(X)$  does not exist because  $\int_x |x| dF_X(x) = \infty$ . Similarly, it has no variance or higher moments. However, the mode and median are both  $x_0$ .

## 4 Convergence of Random Variables

Let  $X_1, X_2, \dots, X_n, \dots$  be a sequence of random variables, and  $X$  be another random variable.

$X_n$  **converges to  $X$  in distribution**, written  $X_n \rightsquigarrow X$ , if

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F(t) \quad (23)$$

at all  $t$  where  $F$  is continuous. Here,  $F_{X_n}$  is the CDF of  $X_n$ , and  $F$  is the CDF of  $X$ . We expect to see the next outcome in a sequence of random experiments becoming better and better modeled by the probability distribution of  $X$ . In other words, the probability for  $X_n$  to be in a given interval is approximately equal to the probability for  $X$  to be in the same interval, as  $n$  grows.

**Example 3** Let  $X_1, \dots, X_n$  be iid continuous random variables. Then trivially  $X_n \rightsquigarrow X_1$ . But note  $P(X_1 = X_n) = 0$ .

**Example 4** Let  $X_2, \dots, X_n$  be identical (but not independent) copies of  $X_1 \sim N(0, 1)$ . Then  $X_n \rightsquigarrow Y = -X_1$ .

**Example 5** Let  $X_n \sim \text{uniform}[0, 1/n]$ . Then  $X_n \rightsquigarrow \delta_0$ . This is often written as  $X_n \rightsquigarrow 0$ .

**Example 6** Let  $X_n$  has the PDF  $f_n(x) = (1 - \cos(2\pi nx))$ ,  $x \in (0, 1)$ . Then  $X_n \rightsquigarrow \text{uniform}(0, 1)$ .

**Theorem 1** (The Central Limit Theorem). Let  $X_1, \dots, X_n$  be iid with finite mean  $\mu$  and finite variance  $\sigma^2 > 0$ . Let  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1). \quad (24)$$

**Example 7** Let  $X_i \sim \text{uniform}(-1, 1)$  for  $i = 1, \dots$ . Note  $\mu = 0, \sigma^2 = 1/3$ . Then  $Z_n = \sqrt{\frac{3}{n}} \sum_{i=1}^n X_i \rightsquigarrow N(0, 1)$ .

**Example 8** Let  $X_i \sim \text{beta}(\frac{1}{2}, \frac{1}{2})$ . Note  $\mu = \frac{1}{2}, \sigma^2 = \frac{1}{8}$ . Then  $Z_n = \sqrt{8n} (\frac{1}{n} \sum_{i=1}^n X_i - 1/2) \rightsquigarrow N(0, 1)$ .

$X_n$  converges to  $X$  in probability, written  $X_n \xrightarrow{P} X$ , if for any  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0. \quad (25)$$

Convergence in probability implies convergence in distribution. The reverse is not true in general. However, convergence in distribution to a point mass distribution implies convergence in probability.

**Example 9** Example 2 converges in probability to  $X$  there.

$X_n$  converges almost surely to  $X$ , written  $X_n \xrightarrow{as} X$ , if

$$P\left(\left\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1 \quad (26)$$

**Example 10** Example 2 does not converge almost surely to  $X$  there. This is because  $\lim_{n \rightarrow \infty} X_n(\omega)$  does not exist for any  $\omega$ . Pointwise convergence is central to  $\xrightarrow{as}$ .

$\xrightarrow{as}$  implies  $\xrightarrow{P}$ , which in turn implies  $\rightsquigarrow$ .

**Example 11**  $X_n \xrightarrow{as} 0$  (and hence  $X_n \xrightarrow{P} 0$  and  $X_n \rightsquigarrow 0$ ) does not imply convergence in expectation  $\mathbb{E}(X_n) \rightarrow 0$ . To see this, let

$$P(X_n) = \begin{cases} 1/n, & \text{if } X_n = n^2 \\ 1 - 1/n, & \text{if } X_n = 0 \end{cases} \quad (27)$$

Then  $X_n \xrightarrow{as} 0$ . However,  $\mathbb{E}(X_n) = \frac{1}{n}n^2 = n$  does not converge.

$X_n$  converges in  $r$ th mean where  $r \geq 1$ , written as  $X_n \xrightarrow{L^r} X$ , if

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^r) = 0. \quad (28)$$

$\xrightarrow{L^r}$  implies  $\xrightarrow{P}$ .

**Theorem 2** (The Weak Law of Large Numbers). If  $X_1, \dots, X_n$  are iid, then  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu(X_1)$ .

**Theorem 3** (The Strong Law of Large Numbers). If  $X_1, \dots, X_n$  are iid, and  $\mathbb{E}(|X_1|) < \infty$ , then  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{as} \mu(X_1)$ .