

Probability Theory Refresher

Christian Oendorfer

Technische Universität München

Probability Theory

Probability Theory is the study of uncertainty. Uncertainty is all around us.

Mathematical theory of probability theory is based on *measure theory* – we do not work on this level.

Slides are mostly based on Review of Probability Theory by Arian Meleki and Tom Do.

Probability Theory

The basic problem that we study in probability theory:

Given a data generating process,
what are the properties of the outcomes?

The basic problem of statistics (or better statistical *inference*) is the *inverse* of probability theory:

Given the outcomes,
what can we say about the process that generated the data?

Statistics uses the formal language of probability theory.

Basic Elements of Probability

Sample Space Ω : The set of all outcomes of a random experiment.

Set of events (event space) \mathcal{F} : A set whose elements $A \in \mathcal{F}$ (*events*) are subsets of Ω . \mathcal{F} (σ -field) must satisfy

- ▶ $\emptyset \in \mathcal{F}$
- ▶ $A \in \mathcal{F} \rightarrow \Omega \setminus A \in \mathcal{F}$
- ▶ $A_1, A_2, \dots \in \mathcal{F} \rightarrow \cup_i A_i \in \mathcal{F}$

Probability measure $P : \mathcal{F} \rightarrow [0, 1]$, with *Axioms of Probability*:

- ▶ $P(A) \geq 0$ for all $A \in \mathcal{F}$
- ▶ $P(\Omega) = 1$
- ▶ If A_1, A_2, \dots are disjoint events ($A_i \cap A_j = \emptyset, i \neq j$) then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

The triple (Ω, \mathcal{F}, P) is called a *probability space*.

Important Properties

- ▶ If $A \subseteq B \rightarrow P(A) \leq P(B)$
- ▶ $P(A \cap B) (\equiv P(A, B)) \leq \min(P(A), P(B))$
- ▶ $P(A \cup B) \leq P(A) + P(B)$
- ▶ $P(\Omega \setminus A) = 1 - P(A)$
- ▶ If A_1, A_2, \dots, A_k are set of *disjoint* events such that $\cup_i A_i = \Omega$ then $\sum_k P(A_k) = 1$ (*Law of total probability*).

Let B be an event with non-zero probability. The probability of any event A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Two events are called *independent* if and only if $P(A, B) = P(A)P(B)$ (or $P(A|B) = P(A)$). What does that mean in words?

Random variables

We are usually only interested in some aspects of a random experiment.

Random variable $X : \Omega \rightarrow \mathbb{R}$ (actually not every function is allowed ...).

A random variable is usually just denoted by an upper case letter X (instead of $X(\omega)$). The value a random variable may take is denoted by the appropriate lower case letter (x in our case).

For a discrete random variable

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

For a continuous random variable

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$$

Note the usage of P here.

Cumulative distribution function – CDF

A probability measure P is specified by a *cumulative distribution function* (CDF), a function $F_X : \mathbb{R} \rightarrow [0, 1]$:

$$F_X(x) \equiv P(X \leq x)$$

Properties:

- ▶ $0 \leq F_X(x) \leq 1$
- ▶ $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- ▶ $\lim_{x \rightarrow \infty} F_X(x) = 1$
- ▶ $x \leq y \rightarrow F_X(x) \leq F_X(y)$

Let X have CDF F_X and Y have CDF F_Y . If $F_X(x) = F_Y(x)$ for all x , then $P(X \in A) = P(Y \in A)$ for all (measurable) A .

Probability mass function – PMF

X takes on only a *countable* set of possible values (*discrete* random variable).

A probability mass function $p_X : \Omega \rightarrow [0, 1]$ is a simple way to *represent* the probability measure associated with X :

$$p_X(x) = P(X = x)$$

(Note: We use the probability measure P on the random variable X)

Properties:

- ▶ $0 \leq p_X(x) \leq 1$
- ▶ $\sum_x p_X(x) = 1$
- ▶ $\sum_{x \in A} p_X(x) = P(X \in A)$

Probability density function – PDF

For some continuous random variables, the CDF $F_X(x)$ is differentiable everywhere. The *probability density function* is then defined as

$$f_X(x) = \frac{dF_X(x)}{dx}$$

$$P(x \leq X \leq x + \delta x) \approx f_X(x)\delta x$$

Properties:

- ▶ $f_X(x) \geq 0$
- ▶ $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- ▶ $\int_{x \in A} f_X(x) dx = P(X \in A)$

Transformation of Random Variables

Given a (continuous) random variable X and a strictly monotone (increasing *or* decreasing) function s , what can we say about $Y = s(X)$?

$$f_Y(y) = f_X(t(y))|t'(y)|$$

where t is the inverse of s .

Expectation

$$E[g(X)] = \sum_x g(x)p_X(x)$$

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

Properties

- ▶ $E[a] = a$ for any constant $a \in \mathbb{R}$
- ▶ $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$
- ▶ $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$

For a discrete random variable X : $E[\mathbb{I}[X = k]] = P(X = k)$

Variance

Variance measures the concentration of a random variable's distribution around its mean.

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Properties

- ▶ $\text{Var}[a] = 0$ for any constant $a \in \mathbb{R}$.
- ▶ $\text{Var}[af(X)] = a^2 \text{Var}[f(X)]$ for any constant $a \in \mathbb{R}$

Entropy

The *Shannon entropy* or just *entropy* of a discrete random variable X is

$$H[X] \equiv - \sum_x P(X = x) \log P(X = x) = -E[\log P(X)]$$

Given two probability mass functions p_1 and p_2 , the *Kullback-Liebler divergence* (or *relative entropy*) between p_1 and p_2 is

$$D(p_1 || p_2) \equiv - \sum_x p_1(x) \log \frac{p_2(x)}{p_1(x)}$$

Discrete random variables

$$X \sim \text{Bernoulli}(p)$$

$$X \sim \text{Binomial}(n, p)$$

$$X \sim \text{Poisson}(\lambda)$$

Continuous random variables

$$X \sim \text{Exponential}(\lambda)$$

$$X \sim \text{Normal}(\mu, \sigma^2)$$

$$X \sim \text{Beta}(a, b)$$

Two random variables – *Bivariate* random variables

If we have two random variables X and Y and we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, the *joint cumulative distribution function* of X and Y is required:

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

$F_X(x)$ and $F_Y(y)$ are the *marginal cumulative distribution function* of $F_{XY}(x, y)$.

Properties

- ▶ $0 \leq F_{XY}(x, y) \leq 1$
- ▶ $\lim_{x, y \rightarrow -\infty} F_{XY}(x, y) = 0$
- ▶ $\lim_{x, y \rightarrow \infty} F_{XY}(x, y) = 1$
- ▶ $F_X(x) = \lim_{y \rightarrow \infty} F_{XY}(x, y)$

Two *discrete* random variables

Joint probability mass function

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Properties:

- ▶ $0 \leq p_{XY}(x, y) \leq 1$
- ▶ $\sum_x \sum_y p_{XY}(x, y) = 1$

In order to get the *marginal probability mass function* $p_X(x)$, we need to *sum out* all possible y (*marginalization*).

$$p_X(x) = \sum_y p_{XY}(x, y)$$

Two *continuous* random variables

Joint probability density function

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

$$\int \int_A f_{XY}(x, y) dx dy = P((X, Y) \in A)$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$$

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

is the *marginal probability density function* or *marginal density* for short.

Conditional distributions/Bayes rule

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)} = \frac{p_{X|Y}(x|y)p_{Y}}{p_X(x)}$$

$$p_{Y|X}(y|x) = P(Y = y|X = x)$$

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

$$P(Y \in A|X = x) = \int_A f_{Y|X}(y|x)dy$$

Independence

Two random variables X , Y are *independent* if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values x and y .

Equivalently:

- ▶ $p_{XY}(x, y) = p_X(x)p_Y(y)$
- ▶ $p_{Y|X}(y|x) = p_Y(y)$
- ▶ $f_{XY}(x, y) = f_X(x)f_Y(y)$
- ▶ $f_{Y|X}(y|x) = f_Y(y)$

Expectation and covariance

Given two random variables X, Y and $g : \mathbb{R}^2 \rightarrow \mathbb{R}$.

- ▶ $E[g(X, Y)] := \sum_x \sum_y g(x, y) p_{XY}(x, y)$
- ▶ $E[g(X, Y)] := \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$

Covariance

- ▶ $Cov[X, Y] := E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- ▶ When $Cov[X, Y] = 0$, X and Y are *uncorrelated*.
- ▶ *Pearson* correlation coefficient $\rho(X, Y)$:

$$\rho(X, Y) := \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}$$

- ▶ $E[f(X, Y) + g(X, Y)] = E[f(X, Y)] + E[g(X, Y)]$
- ▶ $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$
- ▶ If X and Y are independent, then $Cov[X, Y] = 0$
- ▶ If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$

Multiple random variables – Random vectors

Generalize previous ideas to more than two random variables. Putting all these random variables together in one vector \mathbf{X} , a *random vector* ($\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$). The notions of joint CDF and PDF apply equivalently, e.g.

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

Expectation of a continuous random vector for $g : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$E[g(\mathbf{X})] = \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

If $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ then the expected value of g is the *element-wise* values of the output vector:

$$E[g(\mathbf{X})] = \begin{bmatrix} E[g_1(\mathbf{X})] \\ E[g_2(\mathbf{X})] \\ \vdots \\ E[g_m(\mathbf{X})] \end{bmatrix}$$

Covariance matrix

For a random vector $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$, the *covariance matrix* $\mathbf{\Sigma}$ is the $n \times n$ square *symmetric, positive definite* matrix whose entries are

$$\Sigma_{ij} = \text{Cov}[\mathbf{X}_i, \mathbf{X}_j]$$

$$\mathbf{\Sigma} = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$$

Multinomial distribution

The multivariate version of the Binomial is called a *Multinomial* (an urn with k different balls, drawn n times with replacement).

$$\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p}) \rightarrow p_{\mathbf{X}}(x_1, x_2, \dots, x_k) = \binom{n}{x_1 \ x_2 \ \dots \ x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

where

$$\binom{n}{x_1 \ x_2 \ \dots \ x_k} = \frac{n!}{x_1! \ x_2! \ \dots \ x_k!}$$

$$E[\mathbf{X}] = (np_1, np_2, \dots, np_k)$$

$$\text{Var}[\mathbf{X}_i] = np_i(1 - p_i)$$

$$\text{Cov}[\mathbf{X}_i, \mathbf{X}_j] = -np_i p_j$$

The marginal distribution of \mathbf{X}_i is *Binomial*(n, p_i).

Notation in the lecture

Consider

$$p_X(x), \quad x \in \mathbb{R} \quad \text{vs} \quad p_X(y), \quad y \in \mathbb{R}$$

$p_X(x)$, $f_X(x)$, $p_{XY}(x, y)$, $f_{XY}(x, y)$ are written as $p(x)$ or $p(x, y)$. Likewise $p_Y(y)$ is written as $p(y)$.