Review of probability theory

Based on Probability and Random Processes
Grimmet & Stirzaker
and other sources

Definitions

Sample space: Set of all possible outcomes of an experiment

σ-field: A collection $\mathcal{F}$ of subsets of a set Ω is a σ-field if

(i) $\emptyset \in \mathcal{F}$

(ii) if $A_1, A_2 \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

(iii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

Probability (Kolmogorov)

A probabililty measure is a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ satisfying

(i) $\mathbb{P}(A) \geq 0 \; \forall \; A \in \Omega$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) if $A_1, A_2 \ldots \in \mathcal{F}$ and $A_i \cap A_j = \emptyset \; \forall \; i \neq j,$ then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is referred to as a probability space

Choosing an appropriate probability function (relating theory to experiment)

- The probability of an event is the frequency with which the event occurs over many repetitions of the experiment
  - In many cases not possible to repeat experiments
- Probability function can be decided on the basis of symmetry (or fairness)
  - Let $\{H, T\}$ be the sample space associated with tossing a 'fair' coin
  - by definition of fairness $\mathbb{P}(H) = \mathbb{P}(T)$
  - $H \cup T = \Omega, \mathbb{P} = 0.5$

The ‘logical’ interpretation of probability (e.g. see Jaynes, 2003)

Considers probability as an extension of the laws of Aristotelian logic to incorporate uncertainty

Cox axioms

1. Degrees of plausibility, given the available evidence, represented by real numbers $(A|B)$
2. Plausibility should correspond qualitatively with common sense
3. Consistency (arriving at the same conclusion via different routes should yield identical results)

Rules derivable from probability axioms

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Rules derivable from probability axioms

2. if $A \subseteq B, \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A)$

Rules derivable from probability axioms

3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \ldots$ (Sum Rule)

Rules derivable from probability axioms

4. Inclusion-exclusion formula

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \ldots + (-1)^{n+1}\mathbb{P}(A_1 \cap \ldots \cap A_n)$$

Null, impossible and almost sure events

- The impossible event is represented by $\emptyset$
- A s.t. $\mathbb{P}(A) = 0$ is referred to as *null*
- A s.t. $\mathbb{P}(A) = 1$ is said to occur *almost surely*

Conditional Probability

We are frequently interested in the probability of an event A occurring, given that an event B has occurred, denoted $\mathbb{P}(A|B)$

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

$\mathbb{P}(A|B)$ is obtained by considering the probability space $(B, \mathcal{F}', \mathbb{P}')$ with $\mathcal{F}' = \{A \cap B : A \in \mathcal{F}\}$

for $A \in \mathcal{F}'$ $\mathbb{P}'(A) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$, so that $\mathbb{P}'(B) = 1$

For any $A \subset \mathcal{F}$

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

Conditional Probability (II)

Alternatively, define $\mathbb{P}(A|B)$, such that

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

Generalizes to

$$\mathbb{P}(A_1, \ldots A_n) = \mathbb{P}(A_n|A_1 \ldots A_{n-1})\mathbb{P}(A_{n-1}|A_1 \ldots A_{n-2}) \ldots \mathbb{P}(A_1)$$

Also known as *chain rule*

Bayes' Rule

Since $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$ we have

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

In application settings will will often use notation

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\mathbb{P}(\theta)}{\mathbb{P}(D)}$$

Bayes' Rule - Example

A patient goes to see a doctor. The doctor performs a test with 99 percent reliability–that is, 99 percent of people who are sick test positive and 99 percent of the healthy people test negative. The doctor knows that only 1 percent of the people in the country are sick. Now the question is: if the patient tests positive, what are the chances the patient is sick?

Scientific American, Dec 2006

Independence

A and B independent when $\mathbb{P}(A|B) = P(A)$

Alternatively, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Generalizeable to $\{A_i : i \in I\}$ an independent family if

$$\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i)$$

for finite J

Random Variable

A function $X : \Omega \to \mathbb{R}$

Distribution function: F, the probability that $X \leq x$

Necessary to restrict the definition of a random variable to functions for which $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \; \forall \, x \in \mathbb{R}$

Discrete and continuous random variables

X is discrete if it takes values in a countable subset of $\mathbb{R}$

X is continuous if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)du \qquad x \in \mathbb{R}$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$

Discrete and continuous random variables

X is discrete if it takes values in a countable subset of $\mathbb{R}$

X is continuous if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u) du \qquad x \in \mathbb{R}$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$

Probability mass function and probability denisity function

The probability mass function of a discrete random variable is the function $f : \mathbb{R} \rightarrow [0, 1]$, with $f(x) = \mathbb{P}(X(\omega) = x)$

The probability density function of a continuous random variable is the function $f : \mathbb{R} \rightarrow [0, \infty)$, with $\mathbb{P}(X < x) = \int_{-\infty}^{x} f(u)du$

Conditional Distributions

Consider two discrete random variables X and Y. If X takes on a particular value, x then the value Y takes on is a new random variable with distribution function

$$F_{Y|X}(y|x) = \mathbb{P}(Y \leq y|X = x)$$

The probability mass function is

$$f_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x)$$

Marginalization

	No cancer	Cancer
Genotype A	256	34
Genotype B	59	26

Marginalization

	No cancer	Cancer	
Genotype A	256	34	290
Genotype B	59	26	85
	315	60	375

Marginalization

B a binary variable:

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A, B) + \mathbb{P}(A, B^c) \\ &= \mathbb{P}(B)\mathbb{P}(A|B) + \mathbb{P}(B^c)\mathbb{P}(A|B^c)\end{aligned}$$

A *partition* of Ω is a set of disjoint events $\{B_i : i \in I\}$ such that $\bigcup_i B_i = \Omega$

If $\{B_i : i \in I\}$ is a partition of Ω then

$$P(A) = \sum_i \mathbb{P}(B_i) P(A \cap B_i)$$

Marginalizing a random variable

If X, Y are discrete random variables:

$$\begin{aligned}\mathbb{P}(X=x) &= \sum_y \mathbb{P}(X=x, Y=y) \\ &= \sum_y \mathbb{P}(Y=y)\mathbb{P}(X=x|Y=y) \\ &= \sum_y f(y) f_{X|Y}(x|y)\end{aligned}$$

or X, Y continuous random variables:

$$f(x) = \int_y f(y) f_{X|Y}(x|y)$$

Examples of discrete random variables

Bernoulli trial

- An experiment with a binary outcome (e.g. coin tossing)
- IID (independent and identically distributed) assumption over successive experiments

$$f(x|\mu) = \begin{cases} \mu & \text{if x = heads} \\ 1 - \mu & \text{if x = tails} \end{cases}$$

Binomial distribution

E.g. probability of observing heads exactly k times out of N Bernoulli trials (coin tosses), given that the probability (IID) in each trial is p

$$f(k|p) = \binom{N}{k} p^k (1-p)^{N-k}$$

Multinomial distribution

Equivalent for experiments with more than two possible outcomes
(E.g. dice instead of coins)

For M possible outcomes:

$$P(\mathbf{k}|\mathbf{p}) = \frac{N!}{k_1!k_2!\cdots k_M!} p_1^{k_1} p_2^{k_2} \cdots p_M^{k_M}$$

Continuous distributions

Univariate gaussian density function:

$$\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_k^2}\right)$$

Expected Value (or Mean) of a random variable

- Expected value (discrete random variable):

$$\overline{X} = E[X] = \sum_i x_i f(x_i)$$

Expected Value (or Mean) of a random variable

- Expected value (continuous random variable):

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

Expected value of a function of a random variable

If $f(x)$ is the distribution of a random variable X,
the expected value of a function $g(x)$ is (for X discrete):

$$E[g(x)] = \sum_i g(x_i) f(x_i)$$

or if X is continuous:

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

Variance

- Discrete:

$$\sigma_X^2 = E[(X - \overline{X})^2] = \sum_i (x_i - \overline{X})^2 f(x_i)$$

- Continuous:

$$\sigma_X^2 = E[(X - \overline{X})^2] = \int_{-\infty}^{\infty} (x - \overline{X})^2 f(x) dx$$

also:

$$\sigma_X^2 = E[X^2] - \overline{X}^2$$

Correlation and covariance

- Covariance:

$$\mathrm{cov}(X, Y) = E[(X - \overline{X})(Y - \overline{Y})] = \sum_i \sum_j (x - \overline{X})(y - \overline{Y}) f(x, y)$$

- Correlation (coefficient):

$$R_{XY} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Example

In a series of Bernouilli trials with unknown probability of success/failure you observe *a* successes and *b* failures. What is the posterior probability distribution of the probability of a success?

Beta distribution

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$$

(example of the kind of parameterized families of distributions that will be encountered in this course)