Part IA — Probability

Definitions

Based on lectures by R. Weber

Notes taken by Dexter Chua

Lent 2015

Basic concepts
Classical probability, equally likely outcomes. Combinatorial analysis, permutations and combinations. Stirling's formula (asymptotics for $\log n!$ proved). [3]

Axiomatic approach
Axioms (countable case). Probability spaces. Inclusion-exclusion formula. Continuity and subadditivity of probability measures. Independence. Binomial, Poisson and geometric distributions. Relation between Poisson and binomial distributions. Conditional probability, Bayes's formula. Examples, including Simpson's paradox. [5]

Discrete random variables
Expectation. Functions of a random variable, indicator function, variance, standard deviation. Covariance, independence of random variables. Generating functions: sums of independent random variables, random sum formula, moments.

Conditional expectation. Random walks: gambler's ruin, recurrence relations. Difference equations and their solution. Mean time to absorption. Branching processes: generating functions and extinction probability. Combinatorial applications of generating functions. [7]

Continuous random variables
Distributions and density functions. Expectations; expectation of a function of a random variable. Uniform, normal and exponential random variables. Memoryless property of exponential distribution. Joint distributions: transformation of random variables (including Jacobians), examples. Simulation: generating continuous random variables, independent normal random variables. Geometrical probability: Bertrand's paradox, Buffon's needle. Correlation coefficient, bivariate normal random variables. [6]

Inequalities and limits
Markov's inequality, Chebyshev's inequality. Weak law of large numbers. Convexity: Jensen's inequality for general random variables, AM/GM inequality.

Moment generating functions and statement (no proof) of continuity theorem. Statement of central limit theorem and sketch of proof. Examples, including sampling. [3]

Contents

0 Introduction

1 Classical probability

1.1 Classical probability

Definition (Classical probability)**.** *Classical probability* applies in a situation when there are a finite number of equally likely outcome.

Definition (Sample space)**.** The set of all possible outcomes is the *sample space*, Ω. We can lists the outcomes as $\omega_1, \omega_2, \cdots \in \Omega$. Each $\omega \in \Omega$ is an *outcome*.

Definition (Event)**.** A subset of Ω is called an *event*.

Definition (Set notations)**.** Given any two events $A, B \subseteq \Omega$,

- The *complement* of A is $A^C = A' = \bar{A} = \Omega \setminus A$.
- "A or B" is the set $A \cup B$.
- "A and B" is the set $A \cap B$.
- A and B are *mutually exclusive* or *disjoint* if $A \cap B = \emptyset$.
- If $A \subseteq B$, then A occurring implies B occurring.

Definition (Probability)**.** Suppose $\Omega = \{\omega_1, \omega_2, \cdots, \omega_N\}$. Let $A \subseteq \Omega$ be an event. Then the *probability* of A is

$$\mathbb{P}(A) = \frac{\text{Number of outcomes in } A}{\text{Number of outcomes in } \Omega} = \frac{|A|}{N}.$$

1.2 Counting

Definition (Sampling with replacement)**.** When we *sample with replacement*, after choosing at item, it is put back and can be chosen again. Then *any* sampling function f satisfies sampling with replacement.

Definition (Sampling without replacement)**.** When we *sample without replacement*, after choosing an item, we kill it with fire and cannot choose it again. Then f must be an injective function, and clearly we must have $x \geq n$.

Definition (Multinomial coefficient)**.** A *multinomial coefficient* is

$$\binom{n}{n_1, n_2, \cdots, n_k} = \binom{n}{n_1}\binom{n - n_1}{n_2} \cdots \binom{n - n_1 \cdots - n_{k-1}}{n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

It is the number of ways to distribute n items into k positions, in which the ith position has n_i items.

1.3 Stirling's formula

2 Axioms of probability

2.1 Axioms and definitions

Definition (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$. Ω is a set called the *sample space*, $\mathcal{F}$ is a collection of subsets of Ω, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the *probability measure*.

$\mathcal{F}$ has to satisfy the following axioms:

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$.

(iii) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

And $\mathbb{P}$ has to satisfy the following *Kolmogorov axioms*:

(i) $0 \le \mathbb{P}(A) \le 1$ for all $A \in \mathcal{F}$

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

(iii) For any countable collection of events $A_1, A_2, \cdots$ which are disjoint, i.e. $A_i \cap A_j = \emptyset$ for all i, j, we have

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i).$$

Items in Ω are known as the *outcomes*, items in $\mathcal{F}$ are known as the *events*, and $\mathbb{P}(A)$ is the *probability* of the event A.

Definition (Probability distribution)**.** Let $\Omega = \{\omega_1, \omega_2, \cdots\}$. Choose numbers $p_1, p_2, \cdots$ such that $\sum_{i=1}^\infty p_i = 1$. Let $p(\omega_i) = p_i$. Then define

$$\mathbb{P}(A) = \sum_{\omega_i \in A} p(\omega_i).$$

This $\mathbb{P}(A)$ satisfies the above axioms, and $p_1, p_2, \cdots$ is the *probability distribution*

Definition (Limit of events)**.** A sequence of events $A_1, A_2, \cdots$ is *increasing* if $A_1 \subseteq A_2 \cdots$. Then we define the *limit* as

$$\lim_{n \to \infty} A_n = \bigcup_1^\infty A_n.$$

Similarly, if they are *decreasing*, i.e. $A_1 \supseteq A_2 \cdots$, then

$$\lim_{n \to \infty} A_n = \bigcap_1^\infty A_n.$$

2.2 Inequalities and formulae

2.3 Independence

Definition (Independent events)**.** Two events A and B are *independent* if

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

Otherwise, they are said to be *dependent*.

Definition (Independence of multiple events)**.** Events $A_1, A_2, \cdots$ are said to be *mutually independent* if

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_r})$$

for any $i_1, i_2, \cdots i_r$ and $r \geq 2$.

2.4 Important discrete distributions

Definition (Bernoulli distribution)**.** Suppose we toss a coin. $\Omega = \{H, T\}$ and $p \in [0, 1]$. The *Bernoulli distribution*, denoted $B(1, p)$ has

$$\mathbb{P}(H) = p; \quad \mathbb{P}(T) = 1 - p.$$

Definition (Binomial distribution)**.** Suppose we toss a coin n times, each with probability p of getting heads. Then

$$\mathbb{P}(HHTT \cdots T) = pp(1 - p) \cdots (1 - p).$$

So

$$\mathbb{P}(\text{two heads}) = \binom{n}{2} p^2 (1 - p)^{n-2}.$$

In general,

$$\mathbb{P}(k \text{ heads}) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

We call this the *binomial distribution* and write it as $B(n, p)$.

Definition (Geometric distribution)**.** Suppose we toss a coin with probability p of getting heads. The probability of having a head after k consecutive tails is

$$p_k = (1 - p)^k p$$

This is *geometric distribution*. We say it is *memoryless* because how many tails we've got in the past does not give us any information to how long I'll have to wait until I get a head.

Definition (Hypergeometric distribution)**.** Suppose we have an urn with n_1 red balls and n_2 black balls. We choose n balls. The probability that there are k red balls is

$$\mathbb{P}(k \text{ red}) = \frac{\binom{n_1}{k}\binom{n_2}{n-k}}{\binom{n_1+n_2}{n}}.$$

Definition (Poisson distribution)**.** The *Poisson distribution* denoted $P(\lambda)$ is

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}$$

for $k \in \mathbb{N}$.

2.5 Conditional probability

Definition (Conditional probability)**.** Suppose B is an event with $\mathbb{P}(B) > 0$. For any event $A \subseteq \Omega$, the *conditional probability of A given B* is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We interpret as the probability of A happening given that B has happened.

Definition (Partition)**.** A *partition of the sample space* is a collection of disjoint events $\{B_i\}_{i=0}^{\infty}$ such that $\bigcup_i B_i = \Omega$.

3 Discrete random variables

3.1 Discrete random variables

Definition (Random variable)**.** A *random variable* X taking values in a set Ω_X is a function $X : \Omega \to \Omega_X$. Ω_X is usually a set of numbers, e.g. $\mathbb{R}$ or $\mathbb{N}$.

Definition (Discrete random variables)**.** A random variable is *discrete* if Ω_X is finite or countably infinite.

Notation. Let $T \subseteq \Omega_X$, define

$$\mathbb{P}(X \in T) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in T\}).$$

i.e. the probability that the outcome is in T.

Definition (Discrete uniform distribution)**.** A *discrete uniform distribution* is a discrete distribution with finitely many possible outcomes, in which each outcome is equally likely.

Notation. We write

$$\mathbb{P}_X(x) = \mathbb{P}(X = x).$$

We can also write $X \sim B(n, p)$ to mean

$$\mathbb{P}(X = r) = \binom{n}{r} p^r (1-p)^{n-r},$$

and similarly for the other distributions we have come up with before.

Definition (Expectation)**.** The *expectation* (or *mean*) of a real-valued X is equal to

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} p_\omega X(\omega).$$

provided this is *absolutely convergent*. Otherwise, we say the expectation doesn't exist. Alternatively,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \Omega_X} \sum_{\omega : X(\omega) = x} p_\omega X(\omega) \\
&= \sum_{x \in \Omega_X} x \sum_{\omega : X(\omega) = x} p_\omega \\
&= \sum_{x \in \Omega_X} x P(X = x).
\end{aligned}$$

We are sometimes lazy and just write $\mathbb{E}X$.

Definition (Variance and standard deviation)**.** The *variance* of a random variable X is defined as

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

The *standard deviation* is the square root of the variance, $\sqrt{\operatorname{var}(X)}$.

Definition (Indicator function)**.** The *indicator function* or *indicator variable* $I[A]$ (or I_A) of an event $A \subseteq \Omega$ is

$$I[A](\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

Definition (Independent random variables)**.** Let $X_1, X_2, \cdots, X_n$ be discrete random variables. They are *independent* iff for any $x_1, x_2, \cdots, x_n$,

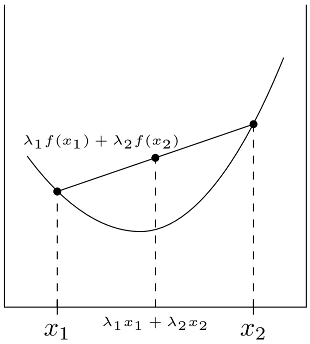
$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n).$$

3.2 Inequalities

Definition (Convex function)**.** A function $f : (a, b) \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in (a, b)$ and $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$,

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geq f(\lambda_1 x_1 + \lambda_2 x_2).$$

It is *strictly convex* if the inequality above is strict (except when $x_1 = x_2$ or λ_1 or $\lambda_2 = 0$).



A function is *concave* if $-f$ is convex.

3.3 Weak law of large numbers

3.4 Multiple random variables

Definition (Covariance)**.** Given two random variables X, Y, the *covariance* is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Definition (Correlation coefficient)**.** The *correlation coefficient* of X and Y is

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Definition (Conditional distribution)**.** Let X and Y be random variables (in general not independent) with joint distribution $\mathbb{P}(X = x, Y = y)$. Then the *marginal distribution* (or simply *distribution*) of X is

$$\mathbb{P}(X = x) = \sum_{y \in \Omega_y} \mathbb{P}(X = x, Y = y).$$

The *conditional distribution* of X given Y is

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

The *conditional expectation* of X given Y is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \Omega_X} x\mathbb{P}(X = x \mid Y = y).$$

We can view $\mathbb{E}[X \mid Y]$ as a random variable in Y: given a value of Y, we return the expectation of X.

3.5 Probability generating functions

Definition (Probability generating function (pgf))**.** The *probability generating function (pgf)* of X is

$$p(z) = \mathbb{E}[z^X] = \sum_{r=0}^{\infty} \mathbb{P}(X = r)z^r = p_0 + p_1 z + p_2 z^2 \cdots = \sum_0^\infty p_r z^r.$$

This is a power series (or polynomial), and converges if $|z| \leq 1$, since

$$|p(z)| \leq \sum_r p_r |z^r| \leq \sum_r p_r = 1.$$

We sometimes write as $p_X(z)$ to indicate what the random variable.

4 Interesting problems

4.1 Branching processes

4.2 Random walk and gambler's ruin

Definition (Random walk)**.** Let $X_1, \cdots, X_n$ be iid random variables such that $X_n = +1$ with probability p, and -1 with probability $1 - p$. Let $S_n = S_0 + X_1 + \cdots + X_n$. Then $(S_0, S_1, \cdots, S_n)$ is a *1-dimensional random walk.*

If $p = q = \frac{1}{2}$, we say it is a *symmetric random walk.*

5 Continuous random variables

5.1 Continuous random variables

Definition (Continuous random variable)**.** A random variable $X : \Omega \to \mathbb{R}$ is *continuous* if there is a function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\;\mathrm{d}x.$$

We call f the *probability density function*, which satisfies

- $f \geq 0$
- $\int_{-\infty}^{\infty} f(x) = 1$.

Definition (Cumulative distribution function)**.** The *cumulative distribution function* (or simply *distribution function*) of a random variable X (discrete, continuous, or neither) is

$$F(x) = \mathbb{P}(X \leq x).$$

Definition (Uniform distribution)**.** The *uniform distribution* on $[a, b]$ has pdf

$$f(x) = \frac{1}{b-a}.$$

Then

$$F(x) = \int_a^x f(z)\;\mathrm{d}z = \frac{x-a}{b-a}$$

for $a \leq x \leq b$.

If X follows a uniform distribution on $[a, b]$, we write $X \sim U[a, b]$.

Definition (Exponential random variable)**.** The *exponential random variable with parameter* λ has pdf

$$f(x) = \lambda e^{-\lambda x}$$

and

$$F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$.

We write $X \sim \mathcal{E}(\lambda)$.

Definition (Expectation)**.** The *expectation* (or *mean*) of a continuous random variable is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\;\mathrm{d}x,$$

provided not both $\int_0^\infty xf(x)\;\mathrm{d}x$ and $\int_{-\infty}^0 xf(x)\;\mathrm{d}x$ are infinite.

Definition (Variance)**.** The *variance* of a continuous random variable is

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

Definition (Mode and median)**.** Given a pdf $f(x)$, we call $\hat{x}$ a *mode* if

$$f(\hat{x}) \geq f(x)$$

for all x. Note that a distribution can have many modes. For example, in the uniform distribution, all x are modes.

We say it is a median if

$$\int_{-\infty}^{\hat{x}} f(x)\,\mathrm{d}x = \frac{1}{2} = \int_{\hat{x}}^{\infty} f(x)\,\mathrm{d}x.$$

For a discrete random variable, the median is $\hat{x}$ such that

$$\mathbb{P}(X \leq \hat{x}) \geq \frac{1}{2},\quad \mathbb{P}(X \geq \hat{x}) \geq \frac{1}{2}.$$

Here we have a non-strict inequality since if the random variable, say, always takes value 0, then both probabilities will be 1.

Definition (Sample mean)**.** If $X_1, \cdots, X_n$ is a random sample from some distribution, then the *sample mean* is

$$\bar{X} = \frac{1}{n}\sum_1^n X_i.$$

5.2 Stochastic ordering and inspection paradox

Definition (Stochastic order)**.** The *stochastic order* is defined as: $X \geq_{\mathrm{st}} Y$ if $\mathbb{P}(X > t) \geq \mathbb{P}(Y > t)$ for all t.

5.3 Jointly distributed random variables

Definition (Joint distribution)**.** Two random variables X, Y have *joint distribution* $F : \mathbb{R}^2 \mapsto [0, 1]$ defined by

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

The *marginal distribution* of X is

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = F(x, \infty) = \lim_{y\to\infty} F(x, y)$$

Definition (Jointly distributed random variables)**.** We say $X_1, \cdots, X_n$ are *jointly distributed continuous random variables* and have *joint pdf* f if for any set $A \subseteq \mathbb{R}^n$

$$\mathbb{P}((X_1, \cdots, X_n) \in A) = \int_{(x_1,\cdots x_n)\in A} f(x_1, \cdots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

where

$$f(x_1, \cdots, x_n) \geq 0$$

and

$$\int_{\mathbb{R}^n} f(x_1, \cdots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

Definition (Independent continuous random variables)**.** Continuous random variables $X_1, \cdots, X_n$ are independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n)$$

for all $A_i \subseteq \Omega_{X_i}$.

If we let F_{X_i} and f_{X_i} be the cdf, pdf of X, then

$$F(x_1, \cdots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

and

$$f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

are each individually equivalent to the definition above.

5.4 Geometric probability

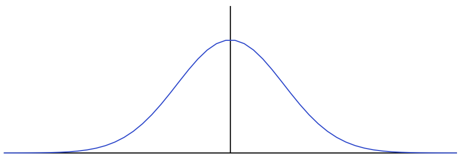
5.5 The normal distribution

Definition (Normal distribution)**.** The *normal distribution* with parameters μ, σ^2, written $N(\mu, \sigma^2)$ has pdf

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

for $-\infty < x < \infty$.

It looks approximately like this:



The *standard normal* is when $\mu = 0, \sigma^2 = 1$, i.e. $X \sim N(0, 1)$.

We usually write $\phi(x)$ for the pdf and $\Phi(x)$ for the cdf of the standard normal.

5.6 Transformation of random variables

Definition (Jacobian determinant)**.** Suppose $\frac{\partial s_i}{\partial y_j}$ exists and is continuous at every point $(y_1, \cdots, y_n) \in S$. Then the *Jacobian determinant* is

$$J = \frac{\partial(s_1, \cdots, s_n)}{\partial(y_1, \cdots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

Definition (Order statistics)**.** Suppose that $X_1, \cdots, X_n$ are some random variables, and $Y_1, \cdots, Y_n$ is $X_1, \cdots, X_n$ arranged in increasing order, i.e. $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. This is the *order statistics*.

We sometimes write $Y_i = X_{(i)}$.

5.7 Moment generating functions

Definition (Moment generating function). The *moment generating function* of a random variable X is

$$m(\theta) = \mathbb{E}[e^{\theta X}].$$

For those θ in which $m(\theta)$ is finite, we have

$$m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\,\mathrm{d}x.$$

Definition (Moment). The rth *moment* of X is $\mathbb{E}[X^r]$.

6 More distributions

6.1 Cauchy distribution

Definition (Cauchy distribution)**.** The *Cauchy distribution* has pdf

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

for $-\infty < x < \infty$.

6.2 Gamma distribution

Definition (Gamma distribution)**.** The *gamma distribution* $\Gamma(n, \lambda)$ has pdf

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}.$$

We can show that this is a distribution by showing that it integrates to 1.

6.3 Beta distribution*

Definition (Beta distribution)**.** The *beta distribution* $\beta(a, b)$ has pdf

$$f(x; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1}$$

for $0 \leq x \leq 1$.

This has mean $a/(a + b)$.

6.4 More on the normal distribution

6.5 Multivariate normal

7 Central limit theorem

8 Summary of distributions

8.1 Discrete distributions

8.2 Continuous distributions