Lecture notes for Part A Probability

Notes written by James Martin, updated by Matthias Winkel

Oxford, Michaelmas Term 2018

winkel@stats.ox.ac.uk

Version of 30 September 2018
(Chapters 1-2 of 7)
Updates for remaining chapters to follow

1

Review: probability spaces, random variables, distributions, independence

1.1 Probability spaces and random variables

We start by reviewing the basic idea of a **probability space** introduced in last year's course. This framework underlies modern probability theory, even though we will seldom need to appeal to it directly in this course.

A **probability space** is a collection $(\Omega, \mathcal{F}, \mathbb{P})$ with the following structure:

(i) Ω is a set, which we call the **sample space**.

(ii) $\mathcal{F}$ is a collection of subsets of Ω. An element of $\mathcal{F}$ is called an **event**.

(iii) The **probability measure** $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0,1]$. It assigns a **probability** to each event in $\mathcal{F}$.

We can think of the probability space as modelling an experiment. The sample space Ω represents the set of all possible outcomes of the experiment.

The set $\mathcal{F}$ of events should satisfy certain natural conditions:

(1) $\Omega \in \mathcal{F}$.

(2) If $\mathcal{F}$ contains a set A, then it also contains the complement A^c (i.e. $\Omega \setminus A$).

(3) If $(A_i, i \in \mathcal{I})$ is a finite or countably infinite collection of events in $\mathcal{F}$, then their union $\bigcup_{i \in \mathcal{I}} A_i$ is also in $\mathcal{F}$.

By combining (2) and (3), we can also obtain finite or countable intersections, as well as unions.

Finally, the probability measure $\mathbb{P}$ should satisfy the following conditions (the **probability axioms**):

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

(2) If $(A_i, i \in \mathcal{I})$ is a finite or countably infinite collection of **disjoint** events, then

$$\mathbb{P}\left(\bigcup_{i\in\mathcal{I}} A_i\right) = \sum_{i\in\mathcal{I}} \mathbb{P}(A_i). \tag{1.1}$$

A **random variable** is a function defined on Ω. We will consider real-valued random variables, i.e. functions from Ω to $\mathbb{R}$.

A random variable represents an **observable** in our experiment; something we can "measure".

Formally, for a function X from Ω to $\mathbb{R}$ to be a random variable, we require that the subset

$$\{\omega \in \Omega \colon X(\omega) \le x\}$$

of Ω is an event in $\mathcal{F}$, for every $x \in \mathbb{R}$. (Then, by taking complements, unions and intersections, we will in fact have that the set $\{\omega \in \Omega \colon X(\omega) \in B\}$ is in $\mathcal{F}$ for a very large class of sets B).

We will usually write X rather than $X(\omega)$ for the value taken by a random variable. Thus if X is a random variable we can talk about the probability of the event

$$\{X \in B\} = \{\omega \in \Omega \colon X(\omega) \in B\},$$

which we will write as $\mathbb{P}(X \in B)$.

Within one experiment, there will be many observables! That is, on the same probability space we can consider many different random variables.

Remarks:

(a) For very simple models, there may be a natural way to set up the sample space Ω (e.g. to represent the set of possible outcomes of the throw of a die or a coin). For more complicated models, this quickly becomes less straightforward. In practice, we hardly ever want to consider Ω directly; instead we work directly with the "events" and "random variables" (the "observables") in the experiment.

(b) In contrast, there are settings in probability theory where we care a lot about the collection of events $\mathcal{F}$, and its structure. (For example, modelling a process evolving in time, we might have a family of different collections $\mathcal{F}_t$, $t \ge 0$, where $\mathcal{F}_t$ represents the set of events which can be observed by watching the evolution of the process up to time t). However, for the purposes of this course we will hardly ever worry about $\mathcal{F}$ directly; we will be safe to assume that $\mathcal{F}$ will always contain any event that we wish to consider.

1.1.1 Examples

Here are some examples of systems (or "experiments") that we might model using a probability space, and, for each one, some examples of random variables that we might want to consider within our model:

- We throw two dice, one red and one blue. Random variables: the score on the red die; the score on the blue die; the sum of the two; the maximum of the two; the indicator function of the event that the blue score exceeds the red score....

- A Geiger counter detecting particles emitted by a radioactive source. Random variables: the time of the kth particle detected, for $k = 1, 2, \ldots$; the number of particles detected in the time interval $[0, t]$, for $t \in [0, \infty)$....

- A model for the evolution of a financial market. Random variables: the prices of various stocks at various times; interest rates at various times; exchange rates at various times....

- The growth of a colony of bacteria. Random variables: the number of bacteria present at a given time; the diameter of the colonised region at given times; the number of generations observed in a given time interval....

- A call-centre. The time of arrival of the kth call; the length of service required by the kth caller; the wait-time of the kth caller in the queue before receiving service....

1.2 Probability distributions

We consider the distribution of a random variable X. This can be summarised by the **distribution function** (or **cumulative distribution function**) of X, defined by

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

for $x \in \mathbb{R}$. (Once we know F, we can derive the probabilities $\mathbb{P}(X \in B)$ for a very wide class of sets B by taking complements, intersections and unions. Formally, the **distribution** of a random variable X is the map $B \mapsto \mathbb{P}(X \in B)$, considered on a suitable collection of subsets $B \subseteq \mathbb{R}$. In practice, we identify distributions by identifying the cumulative distribution function or any other associated function that uniquely determines a distribution.)

Any distribution function F must obey the following properties:

(1) F is non-decreasing.

(2) F is right-continuous.

(3) $F(x) \to 0$ as $x \to -\infty$.

(4) $F(x) \to 1$ as $x \to \infty$.

Remark 1.1. Note that two different random variables can have the same distribution! For example, consider the model of two dice mentioned above. If the dice are "fair", then the distribution of the score on the blue die might be the same as the distribution of the score on the red die. However, that does not mean that the two scores are always the same! They are two different "observables" within the same experiment.

If two random variables X and Y have the same distribution, we write $X \overset{d}{=} Y$.

We single out two important classes of random variables: **discrete** and **continuous**.

1.2.1 Discrete random variables

A random variable X is **discrete** if there is a finite or countably infinite set B such that $\mathbb{P}(X \in B) = 1$.

We can represent the distribution of a discrete random variable X by its probability mass function

$$p_X(x) = \mathbb{P}(X = x)$$

for $x \in \mathbb{R}$. This function is zero except at a finite or countably infinite set of points. We have

- $\sum_x p_X(x) = 1$.
- $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$ for any set $A \subseteq \mathbb{R}$.

The points x where $\mathbb{P}(X = x) > 0$ are sometimes called the **atoms** of the distribution of X. In many examples these will be a set of integers such as $\{1, 2, \dots, n\}$ or $\{0, 1, 2, \dots\}$ or $\{1, 2, 3, \dots\}$.

The cumulative distribution function of X has jumps at the location of the atoms, and is constant on any interval that does not contain an atom.

1.2.2 Continuous random variables

A random variable X is called (absolutely) **continuous** if its distribution function F can be written as an integral. That is, there is a function f such that

$$\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du.$$

f is called the **density function** (or **probability density function**) of X.

This certainly implies that F is a continuous function (although note that not all continuous F can be written in this way). In particular, $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y) = 0$ for any x. The density function is not unique; for example, we can change the value of f at any single point without affecting the integral of f. At points where F is differentiable, it is natural to take $f(x) = F'(x)$. For any $a < b$, we have

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f(u)du.$$

1.2.3 Median

The median of a distribution with cumulative distribution function F is $m \in \mathbb{R}$ such that $F(m) = 1/2$, if such m is unique, or the midpoint $m = (a+b)/2$ of the interval (a, b), where where $F(x) = 1/2$, $x \in (a, b)$.

1.3 Expectation and variance

Let X be a discrete random variable with probability mass function $p_X(x) = \mathbb{P}(X = x)$. The **expectation** (or **mean**) of X is defined by

$$\mathbb{E}\,X = \mathbb{E}\,(X) = \sum_x x p_X(x), \tag{1.2}$$

when this sum converges.

If instead X is a continuous random variable with density function f, then its expectation is given by

$$\mathbb{E}\, X = \mathbb{E}\,(X) = \int_{-\infty}^{\infty} x f(x) dx, \tag{1.3}$$

when this integral converges.

We often want to express the expectation of a function of a random variable X in terms of the density function or the mass function of X. We have

$$\mathbb{E}\, g(X) = \sum_x g(x) p_X(x)$$

in the discrete case, and

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

in the continuous case, always provided that the **expectation exists**, i.e. that the sum or integral converges.

It is rather unsatisfactory that we have two different definitions of expectation for two different cases, and no definition at all for random variables which are neither continuous nor discrete. In fact it is not difficult to unify the definitions. A very natural way is to consider approximations of a general random variable by discrete random variables. This is analogous to the construction of the integral of a general function by defining the integral of a step function using sums, and then defining the integral of a general function using an approximation by step functions, which you saw in last year's analysis course.

This unifies the two definitions above, and extends the definition to all types of random variable, whether discrete, continuous or neither. We will not pursue this here – but we will collect together basic properties of expectation which we will use constantly:

(1) For any event A, write $\mathbf{1}_A$ for the indicator function of A. Then $\mathbb{E}\,\mathbf{1}_A = \mathbb{P}(A)$.

(2) If $\mathbb{P}(X \geq 0) = 1$, then $\mathbb{E}\, X \geq 0$.

(3) **(Linearity 1):** $\mathbb{E}\,(aX) = a\mathbb{E}\, X$ for any constant a.

(4) **(Linearity 2):** $\mathbb{E}\,(X + Y) = \mathbb{E}\, X + \mathbb{E}\, Y$.

Formally, $\mathbb{E}$ is a linear operator on the vector space of random variables $X \colon \Omega \to \mathbb{R}$ whose expectations exist.

1.3.1 Variance and covariance

The **variance** of a random variable X is defined by

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

provided that the expectations exist. This can then alternatively be expressed as

$$\mathrm{Var}(X) = \mathbb{E}\,(X^2) - (\mathbb{E}\, X)^2.$$

The **covariance** of two random variables X and Y is defined by

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

if the expectations exist, which can then alternatively be expressed as

$$\operatorname{Cov}(X,Y) = \mathbb{E}\,(XY) - (\mathbb{E}\,X)(\mathbb{E}\,Y).$$

Note that $\operatorname{Var}(X) = \operatorname{Cov}(X,X)$. From the linearity of expectation, we get a bi-linearity property for covariance:

$$\operatorname{Cov}(aX+b, cY+d) = ac\operatorname{Cov}(X,Y).$$

As a special case we can obtain

$$\operatorname{Var}(aX+b) = a^2\operatorname{Var}(X).$$

We also have the property

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$$

and more generally

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{1\le i<j\le n}\operatorname{Cov}(X_i,X_j),$$

always provided that all expectations exist.

1.4 Independence

Events A and B are **independent** if

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

More generally, a family of events $(A_i, i\in\mathcal{I})$, possibly infinite, even uncountable, is called independent if for all finite subsets J of $\mathcal{I}$,

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

Remark 1.2. Remember that this is a stronger condition than pairwise independence! Even for three events, it is possible that A_1, A_2 are independent, A_2, A_3 are independent and A_1, A_3 are independent but that A_1, A_2, A_3 are **not** independent.

Random variables $X_1, X_2, \ldots, X_n$ are independent if for all $B_1, B_2, \ldots, B_n \subset \mathbb{R}$, the events $\{X_1\in B_1\}$, $\{X_2\in B_2\}, \ldots, \{X_n\in B_n\}$ are independent.

In fact, it turns out to be enough to check that for all $x_1, x_2, \ldots, x_n$,

$$\begin{aligned}\mathbb{P}(X_1\le x_1,\ldots,X_n\le x_n) &= \mathbb{P}(X_1\le x_1)\cdots\mathbb{P}(X_n\le x_n)\\ &= F_{X_1}(x_1)\cdots F_{X_n}(x_n).\end{aligned}$$

If the random variables are all discrete, another equivalent condition is that

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

When X and Y are independent random variables whose expectations exist, we have $\mathbb{E}\,(XY) = \mathbb{E}\,(X)\mathbb{E}\,(Y)$, or equivalently $\mathrm{Cov}(X, Y) = 0$. That is, X and Y are uncorrelated. The converse is *not* true; uncorrelated does not imply independent!

Various of the properties above can be summarised by the phrase "**independence means multiply**".

1.5 Examples of probability distributions

We review some of the families of probability distributions which are of particular importance in applications and in theory.

1.5.1 Continuous distributions

Uniform distribution

X has the uniform distribution on an interval $[a, b]$ if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim U[a, b]$.

Exponential distribution

X has exponential distribution with parameter (or *rate*) λ if its distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

Its density function is

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \ge 0 \end{cases}.$$

We write $X \sim \mathrm{Exp}(\lambda)$. We have $\mathbb{E}\,X = 1/\lambda$ and $\mathrm{Var}\,X = 1/\lambda^2$. If $X \sim \mathrm{Exp}(\lambda)$ and $a > 0$, then $aX \sim \mathrm{Exp}(\lambda/a)$. An important property of the distribution is the **memoryless property**: $\mathbb{P}(X > x + t \,|\, X > t)$ does not depend on t.

Normal distribution

X has the normal (or Gaussian) distribution with mean μ and variance σ^2 if its density function is given by

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

We write $X \sim N(\mu, \sigma^2)$. The **standard normal** distribution is $N(0,1)$.

If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. In particular, $(X - \mu)/\sigma$ has standard normal distribution.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

The normal distribution has an extremely important role in probability theory, exemplified by the fact that it appears as the limit in the Central Limit Theorem.

We often write Φ for the distribution function of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

Gamma distribution

The family of gamma distributions generalises the family of exponential distributions. The gamma distribution with *rate* λ and *shape* r has density

$$f(x) = \begin{cases} \dfrac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Here $\Gamma(r)$ is the gamma function, defined by $\Gamma(r) = \int_0^\infty z^{r-1} e^{-z} dz$. It is the analytic continuation of the factorial function, in that $\Gamma(r) = (r-1)!$ when r is an integer.

A gamma distribution with shape $r = 1$ is an exponential distribution.

If $X \sim \text{Gamma}(r_X, \lambda)$ and $Y \sim \text{Gamma}(r_Y, \lambda)$ are independent, then we have $X + Y \sim \text{Gamma}(r_X + r_Y, \lambda)$. As a special case, if $X_1, X_2, \ldots, X_n$ are i.i.d. with $\text{Exp}(\lambda)$ distribution, then $X_1 + X_2 + \cdots + X_n$ has $\text{Gamma}(n, \lambda)$ distribution.

1.5.2 Discrete distributions

Discrete uniform distribution

X has the discrete uniform distribution on a set B of size n (for example the set $\{1, 2, \ldots, n\}$) if

$$p_X(x) = \begin{cases} 1/n, & x \in B \\ 0, & x \notin B \end{cases}.$$

Bernoulli distribution

X has Bernoulli distribution with parameter p if

$$p_X(1) = p, \;\; p_X(0) = 1 - p$$

(and so of course $p_X(x) = 0$ for other values of x).

We have $\mathbb{E}\, X = p$ and $\operatorname{Var} X = p(1-p)$.

If A is an event with $\mathbb{P}(A) = p$, then its indicator function $\mathbf{1}_A$ has Bernoulli distribution with parameter p.

Binomial distribution

If $X_1, X_2, \dots, X_n$ are i.i.d. Bernoulli random variables with the same parameter p, then their sum $X_1 + \dots + X_n$ has Binomial distribution with parameters n and p.

Equivalently, if $A_1, \dots, A_n$ are independent events, each with probability p, then the total number of those events which occur has Binomial(n, p) distribution.

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

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, \dots, n\}.$$

$\mathbb{E}\, X = np$ and $\operatorname{Var} X = np(1-p)$.

Geometric distribution

Let $p \in (0, 1)$ and let X have mass function

$$p_X(k) = (1-p)^{k-1} p \quad \text{for } k \in \{1, 2, 3, \dots\}.$$

Let $Y = X - 1$; then Y has mass function

$$p_Y(k) = (1-p)^k p \quad \text{for } k \in \{0, 1, 2, \dots\}.$$

The terminology is not consistent; either X or Y might be said to have a geometric distribution with parameter p. (Or even sometimes with parameter $1-p$).

If we have a sequence of independent trials, with probability p of success at each trial, then X could represent the number of trials needed for the first success to occur, while Y could represent the number of failures needed before the first success occurs.

We have

$$\mathbb{P}(X > k) = \mathbb{P}(Y \geq k) = (1-p)^k \quad \text{for } k = 0, 1, 2, \dots.$$

The geometric distribution can be thought of as a discrete anaologue of the exponential distribution. It too has a memoryless property; for $k, m \in \{0, 1, 2, \dots\}$, the conditional probability $\mathbb{P}(X > k + m \,|\, X > k)$ does not depend on k.

Poisson distribution

X has Poisson distribution with mean λ if

$$\mathbb{P}(X = r) = \frac{e^{-\lambda} \lambda^r}{r!} \quad \text{for } r = 0, 1, 2, \dots.$$

$\mathbb{E}\, X = \lambda$ and $\operatorname{Var} X = \lambda$.

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

The Poisson distribution arises in many applications; it is a good model for the total number of events occurring when there are a large number of possible events which each occur with small probability. There are close connections between the Poisson distribution and the exponential distribution, which we will see in detail when we study Poisson processes at the end of the course.

2

Convergence of random variables, and limit theorems

Let X and Y be random variables. What might we mean by saying "X and Y are close"?

(1) We might be describing a particular realisation. For example, we made an observation of X and Y, and on this occasion $|X - Y| < \epsilon$.

(2) We might be making a statement about the joint distribution of X and Y, for example

$$\mathbb{P}(|X - Y| < \epsilon) > 1 - \epsilon,$$

or

$$\mathbb{E}\big(|X - Y|\big) < \epsilon.$$

(3) We might be comparing the distribution of X with the distribution of Y, for example

$$|F_X(x) - F_Y(x)| < \epsilon \text{ for all } x.$$

Correspondingly, there are several different things we might mean when we say that a sequence of random variables converges to a limit.

2.1 Modes of convergence

Let $X_1, X_2, \ldots$ and X be random variables.

Note that $\{X_n \to X \text{ as } n \to \infty\}$ is an **event**. More formally we could write

$$\{\omega \in \Omega \colon X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$$

to emphasise the dependence on ω; the event might hold for some ω but not for others.

Definition. *$X_n \to X$ **almost surely** (or "with probability 1") if*

$$\boxed{\mathbb{P}\left(X_n \to X \textit{ as } n \to \infty\right) = 1.} \tag{2.1}$$

We often abbreviate to "$X_n \to X$ a.s.".

Definition. $X_n \to X$ ***in probability*** *(written* $X_n \xrightarrow{P} X$*) if for every* $\epsilon > 0$*,*

$$\boxed{\mathbb{P}\Big(\big|X_n - X\big| < \epsilon\Big) \to 1 \text{ as } n \to \infty.} \tag{2.2}$$

Let $F_1, F_2, \ldots$ and F be the distribution functions of $X_1, X_2, \ldots$ and X respectively.

Definition. $X_n \to X$ ***in distribution*** *(or* ***weakly****), written* $X_n \xrightarrow{d} X$*, if, for every* x *such that* F *is continuous at* x*,*

$$\boxed{F_n(x) \to F(x) \text{ as } n \to \infty.} \tag{2.3}$$

We will see later that these formulations are in decreasing order of strength.

2.2 Convergence in distribution

Notice that in the definition of convergence in distribution in (2.3), the random variables involved appear only through their distributions. Hence we do not even need all the random variables to be defined on the same probability space. This is really a definition about convergence of *distributions*, not about convergence of random variables. The *joint* distribution of the random variables does not need to be defined. This is in contrast to the definitions of almost sure convergence in (2.1) and of convergence in probability in (2.2), where we genuinely do need all the random variables to be defined on the same space.

As a result, we might sometimes vary the notation by writing a distribution rather than a random variable on the right-hand side; e.g. "$X_n \xrightarrow{d} N(0,1)$" if the limit in distribution is the standard normal, or "$X_n \xrightarrow{d} U[0,1]$" if the limit in distribution is the uniform distribution on $[0,1]$.

In many cases the limit will be deterministic; e.g. if the limit is a distribution which puts all its mass at the value 0, then we will write $X_n \xrightarrow{d} 0$.

In (2.3), why did we ask for the limit to hold only for x which are continuity points of F, rather than at all x? The first couple of examples (which are almost trivial) make this clear.

Example 2.1. Let X_n have the uniform distribution on the interval $[-1/n, 1/n]$. Then $F_n(x) \to 0$ for all $x < 0$, and $F_n(x) \to 1$ for all $x > 0$.

So we have $X_n \xrightarrow{d} 0$, i.e. the distribution of X_n converges to that of a deterministic random variable which is equal to 0 with probability 1. Such a random variable has distribution function given by $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x \geq 0$.

Note that $F_n(0) = 1/2$ for all n, while $F(0) = 1$. So convergence does not hold at the point 0 itself (but this is OK, since 0 is not a continuity point of F).

Example 2.2. Let X_n be a deterministic random variable taking the value $1/n$ with probability 1. Let X be a deterministic random variable taking the value 0 with probability 1 (as above). Then once again, $X_n \xrightarrow{d} X$, (even though $\mathbb{P}(X_n \leq 0) = 0$ for all n while $\mathbb{P}(X \leq 0) = 1$).

There are many situations in which a sequence of discrete random variables converges to a continuous limit. Here is one example, showing that a geometric distribution with a small parameter is well approximated by an exponential distribution:

Example 2.3. Let X_n have geometric distribution on the positive integers, with parameter p_n, i.e. $\mathbb{P}(X_n = k) = (1-p_n)^{k-1} p_n$ for $k = 1, 2, \ldots$. Show that if $p_n \to 0$ as $n \to \infty$, then $p_n X_n$ converges in distribution to the exponential distribution with mean 1.
Solution: We have $\mathbb{P}(X_n > k) = (1-p_n)^k$ for $k = 0, 1, 2, \ldots$. For $x \geq 0$, we have

$$\begin{aligned}
\mathbb{P}(p_n X_n > x) &= \mathbb{P}\left(X_n > \frac{x}{p_n}\right) \\
&= \mathbb{P}\left(X_n > \left\lfloor \frac{x}{p_n} \right\rfloor\right) \\
&= (1-p_n)^{\lfloor x/p_n \rfloor} \\
&\to e^{-x} \text{ as } n \to \infty
\end{aligned}$$

because $p_n \to 0$; here we use the fact that $(1-\epsilon)^{x/\epsilon} \to e^{-x}$ as $\epsilon \to 0$, and also that $\lfloor x/p_n \rfloor - x/p_n$ is bounded.

Hence if F_n is the distribution function of $p_n X_n$, then $1 - F_n(x) \to e^{-x}$ as $n \to \infty$. So $F_n(x) \to 1 - e^{-x}$ for all $x > 0$, while $F_n(x) = 0$ for all $x \leq 0$ and all n.

So indeed $F_n(x) \to F(x)$ for all x, where F is the distribution function of a random variable with Exp(1) distribution.

There are several more examples on the problem sheets.

2.3 Comparison of different modes of convergence

Theorem 2.4. *The following implications hold:*

$$\boxed{X_n \to X \textit{ almost surely}} \quad \Rightarrow \quad \boxed{X_n \to X \textit{ in probability}} \quad \Rightarrow \quad \boxed{X_n \to X \textit{ in distribution}}$$

The reverse implications do not hold in general.

Before starting the proof we note a useful fact, which is a simple consequence of the countable additivity axiom for unions of disjoint sets (1.1).

Lemma 2.5. *Let $A_n, n \geq 1$, be an increasing sequence of events; that is, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$. Then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\geq 1} A_n\right). \tag{2.4}$$

Proof. Because the sequence A_n is increasing, it is easy to rewrite the union as a disjoint union:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) &= \mathbb{P}\Big(A_1 \cup \bigcup_{n\geq 1} (A_{n+1} \setminus A_n)\Big) \\
&= \mathbb{P}(A_1) + \sum_{i=1}^{\infty} \mathbb{P}\big(A_{i+1} \setminus A_i\big) \qquad \text{(using countable additivity)} \\
&= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n-1} \mathbb{P}\left(A_{i+1} \setminus A_i\right)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \Big(\mathbb{P}(A_1) + \sum_{i=1}^{n-1} \mathbb{P}\left(A_{i+1} \setminus A_i\right)\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\Big(A_1 \cup \bigcup_{1\leq i\leq n-1} (A_{i+1} \setminus A_i)\Big) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}
$$

□

Proof of Theorem 2.4.
(1) First we will show that convergence in probability implies convergence in distribution. Let F_n be the distribution function of X_n, and F the distribution function of X. Fix any x such that F is continuous at x, and fix any $\epsilon > 0$.

Observe that if $X_n \leq x$, then either $X \leq x + \epsilon$ or $|X_n - X| > \epsilon$. Hence

$$
\begin{aligned}
F_n(x) &= \mathbb{P}(X_n \leq x) \\
&\leq \mathbb{P}\big(X \leq x + \epsilon \text{ or } |X_n - X| > \epsilon\big) \\
&\leq \mathbb{P}\big(X \leq x + \epsilon\big) + \mathbb{P}\big(|X_n - X| > \epsilon\big) \\
&\to F(x+\epsilon) \text{ as } n \to \infty,
\end{aligned}
$$

using the convergence in probability. So $F_n(x) < F(x + \epsilon) + \epsilon$ for all large enough n.

Similarly by looking at $1 - F_n(x) = \mathbb{P}(X_n > x)$, we can obtain that $F_n(x) > F(x - \epsilon) - \epsilon$ for all large enough n.

Since $\epsilon > 0$ is arbitrary, and since F is continuous at x, this implies that $F_n(x) \to F(x)$ as $n \to \infty$.

---

(2) For convergence in distribution, we do not need the random variables to be defined on the same probability space. But even if they are, convergence in distribution does not imply convergence in probability. For example, suppose that X and Y are random variables with the same distribution but with $\mathbb{P}(X = Y) < 1$. Then the sequence $X, X, X, \ldots$ converges to Y in distribution, but not in probability.

---

(3) Now we will show that almost sure convergence implies convergence in probability. Fix $\epsilon > 0$ and for $N \in \mathbb{N}$, define the event A_N by

$$A_N = \{|X_n - X| < \epsilon \text{ for all } n \geq N\}.$$

Suppose that $X_n \to X$ almost surely. If the event $\{X_n \to X\}$ occurs, then the event A_N must occur for some N, so we have $\mathbb{P}\big(\bigcup A_N\big) = 1$. A_N is an increasing sequence of events, so (2.4) then gives $\lim_{N\to\infty} \mathbb{P}(A_N) = 1$.

But A_N implies $\big|X_N - X\big| < \epsilon$, giving $\mathbb{P}\Big(\big|X_N - X\big| < \epsilon\Big) \to 1$. Since ϵ is arbitrary, this means that $X_n \to X$ in probability, as desired.

---

(4) Finally we want to show that convergence in probability does not imply almost sure convergence.

Consider a sequence of independent random variables X_n where $\mathbb{P}(X_n = 1) = 1/n$ and $\mathbb{P}(X_n = 0) = (n-1)/n$.

We have $X_n \to 0$ in probability as $n \to \infty$ because for any $\epsilon > 0$, $\mathbb{P}(|X_n - 0| < \epsilon) \geq \mathbb{P}(X = 0) \to 1$.

Since X_n only take the values 0 and 1, the event $\{X_n \to 0\}$ is the same as the event $\{X_n = 0 \text{ eventually}\}$. This is $\bigcup_{N\geq 1} B_N$ where $B_N = \{X_n = 0 \text{ for all } n \geq N\}$.

But for any N and K,

$$\begin{aligned}\mathbb{P}(B_N) \leq \mathbb{P}(X_n = 0 \text{ for all } n = N, \dots, N+K) &= \frac{N-1}{N}\frac{N}{N+1}\frac{N+1}{N+2}\cdots\frac{N+K-1}{N+K} \\ &= \frac{N-1}{N+K}.\end{aligned}$$

As $K \geq 1$ is arbitrary, we obtain that $\mathbb{P}(B_N) = 0$. Hence also by Lemma 2.5, $\mathbb{P}(\bigcup_{N\geq 1} B_N) = 0$, and so $\mathbb{P}(X_n \to 0) = 0$. Hence it is not the case that X_n converges to 0 almost surely. □

Although convergence in distribution is weaker than convergence in probability, there is a partial converse, for the case when the limit is deterministic:

Theorem 2.6. *Let $X_1, X_2, \dots$ be a sequence of random variables defined on the same probability space. If $X_n \to c$ in distribution where c is some constant, then also $X_n \to c$ in probability.*

Proof. Exercise (see problem sheet). □

2.4 Review: Weak law of large numbers

This was covered at the end of last year's course (without explicitly introducing the notion of convergence in probability).

Let $S_n = X_1 + X_2 + \cdots + X_n$, where X_i are i.i.d. with mean μ. The law of large numbers tells us that, roughly speaking, S_n behaves to first order like $n\mu$ as $n \to \infty$. The weak law phrases this in terms of convergence in probability. (Later we will see a stronger result in terms of almost sure convergence).

Theorem (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with finite mean μ. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \xrightarrow{P} \mu \text{ as } n \to \infty.$$

That is, for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1 \text{ as } n \to \infty. \tag{2.5}$$

Given Theorem 2.6, we could equivalently write $\frac{S_n}{n} \xrightarrow{d} \mu$.

We will give an extremely simple proof of the weak law of large numbers, under an additional condition (that the X_i have finite variance). To do this, we need some results which give probability bounds on the tail of a distribution in terms of its mean and variance.

Theorem (Markov's inequality)**.** *Let X be random variable taking non-negative values (i.e. $\mathbb{P}(X \geq 0) = 1$). Then for any $z > 0$,*

$$\mathbb{P}(X \geq z) \leq \frac{\mathbb{E}\, X}{z}. \tag{2.6}$$

Proof. We consider a random variable $X_z = z\mathbf{1}\{X \geq z\}$. So X_z takes the value 0 whenever X is in $[0, z)$ and the value z whenever X is in $[z, \infty)$. So $X \geq X_z$ always (here we use the fact that X is non-negative).

Then $\mathbb{E}\, X \geq \mathbb{E}\, X_z = z\mathbb{E}\, \mathbf{1}\{X \geq z\} = z\mathbb{P}(X \geq z)$. Rearranging gives the result. □

Theorem (Chebyshev's inequality)**.** *Let Y be a random variable with finite mean and variance. Then for any $\epsilon > 0$,*

$$\mathbb{P}\big(|Y - \mathbb{E}\, Y| \geq \epsilon\big) \leq \frac{\operatorname{Var} Y}{\epsilon^2}.$$

Proof.

$$\begin{aligned}\mathbb{P}\big(|Y - \mathbb{E}\, Y| \geq \epsilon\big) &= \mathbb{P}\big([Y - \mathbb{E}\, Y]^2 \geq \epsilon^2\big) \\ &\leq \frac{\mathbb{E}\left([Y - \mathbb{E}\, Y]^2\right)}{\epsilon^2}\end{aligned}$$

(by applying Markov's inequality (2.6) with $X = [Y - \mathbb{E}\, Y]^2$ and $z = \epsilon^2$)

$$= \frac{\operatorname{Var} Y}{\epsilon^2}.$$

□

Proof of the weak law of large numbers in the case of random variables with finite variance. Let X_i be i.i.d. with mean μ and variance σ^2. Recall $S_n = X_1 + \cdots + X_n$. We want to show that $S_n/n \xrightarrow{P} \mu$ as $n \to \infty$.

We have $\mathbb{E}\,(S_n/n) = \mu$, and (using the independence of the X_i),

$$\begin{aligned}\operatorname{Var}\left(\frac{S_n}{n}\right) &= \frac{\operatorname{Var} S_n}{n^2} \\ &= \frac{\operatorname{Var} X_1 + \cdots + \operatorname{Var} X_n}{n^2} \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Fix any $\epsilon > 0$. Using Chebyshev's inequality applied to the random variable S_n/n, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) &\leq \frac{\operatorname{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

So indeed (2.5) holds, as required. □

Remark. Observe that we could relax considerably the assumptions in the weak law of large numbers, and still get the same result using almost the same proof. We never used at all the assumption that the X_i all had the same distribution. We could also relax the assumption that the X_i are independent, as long as the variance of S_n grows more slowly than n^2. For example, if we have an upper bound on the variance of each X_i, and a bound which is $o(n^2)$ on the sum $\sum_{1\leq i<j\leq n} \text{Cov}(X_i, X_j)$, then exactly the same idea applies to show that $(S_n - \mathbb{E}\, S_n)/n$ converges to 0 in distribution.

2.5 Strong law of large numbers

In the weak law of large numbers, we proved convergence in distribution of the average of i.i.d. random variables to the mean. The strong law says more: the convergence occurs with probability 1.

Theorem (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with mean μ. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \text{ almost surely as } n \to \infty.$$

2.5.1 Proof of the strong law of large numbers (non-examinable)

A proof of the strong law of large numbers in full generality is somewhat involved. A nice proof uses martingales and is part of next year's course on Probability, Measure and Martingales.

However, if we assume an extra condition, namely that the distribution has a finite fourth moment, then a relatively straightforward proof is possible. [NB the proof is not examinable.]

Proof of Strong Law of Large Numbers, under the additional condition $\mathbb{E}\, X_n^4 < \infty$.
Let us centre the X_n, writing $W_n = X_n - \mu$.

Then $\mathbb{E}\, W_n = 0$, and we have $\mathbb{E}\, X_n^4 < \infty \Rightarrow \mathbb{E}\, W_n^4 < \infty$ (exercise).

Note also that

$$\begin{aligned}(\mathbb{E}\, W_n^2)^2 &= \mathbb{E}\,(W_n^4) - \text{Var}(W_n^2) \\ &\leq \mathbb{E}\,(W_n^4).\end{aligned}$$

We will consider $\mathbb{E}\, \left[(S_n - n\mu)^4\right]$. Expanding the fourth power and using linearity of expectation, we obtain

$$\begin{aligned}\mathbb{E}\, \left[(S_n - n\mu)^4\right] &= \mathbb{E}\, \left[(W_1 + W_2 + \cdots + W_n)^4\right] \\ &= \sum_{1\leq i\leq n} \mathbb{E}\, W_i^4 + 4 \sum_{\substack{1\leq i,j\leq n \\ i\neq j}} \mathbb{E}\, W_i^3 W_j + 3 \sum_{\substack{1\leq i,j\leq n \\ i\neq j}} \mathbb{E}\, W_i^2 W_j^2 \\ &\qquad + 6 \sum_{\substack{1\leq i,j,k\leq n \\ i,j,k \text{ distinct}}} \mathbb{E}\, W_i^2 W_j W_k + \sum_{\substack{1\leq i,j,k,l\leq n \\ i,j,k,l \text{ distinct}}} \mathbb{E}\, W_i W_j W_k W_l.\end{aligned}$$

(The exact constants in front of the sums are not too important!) Using independence and $\mathbb{E}\, W_i = 0$, most of these terms vanish. For example, $\mathbb{E}\, W_i^3 W_j = \mathbb{E}\, W_i^3 \mathbb{E}\, W_j = 0$. We are left with only

$$\begin{aligned}\mathbb{E}\left[(S_n-n\mu)^4\right] &= n\mathbb{E}\,W_1^4+3n(n-1)\left(\mathbb{E}\left[W_1^2\right]\right)^2\\ &\le 3n^2\mathbb{E}\,W_1^4.\end{aligned}$$

From this we have

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}-\mu\right)^4\right] &= \sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n}-\mu\right)^4\right]\\ &= \sum_{n=1}^{\infty}\frac{1}{n^4}\mathbb{E}\left[(S_n-n\mu)^4\right]\\ &\le \sum_{n=1}^{\infty}\frac{3\mathbb{E}\,W_1^4}{n^2}\\ &< \infty.\end{aligned}$$

Formally, interchanging the infinite series and expectation in the first line requires a justification refining the notion of absolute convergence to a framework involving general expectations, which is beyond the scope of this course in the present generality. This is not so hard for discrete random variables (changing the order of terms in absolutely convergent series). The theory for such interchanging of series and integrals is developed in next term's course on Integration and transferred to a general setting of expectations on probability spaces at the very beginning of next year's course on Probability, Measure and Martingales.

But if Z is a random variable with $\mathbb{E}\,Z<\infty$, then certainly $\mathbb{P}(Z<\infty)=1$. Applying this with $Z=\left(\frac{S_n}{n}-\mu\right)^4$, we get

$$\mathbb{P}\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}-\mu\right)^4<\infty\right)=1.$$

Finally, if $\sum(a_n-\mu)^4$ is finite, then certainly $a_n\to\mu$ as $n\to\infty$. So we can conclude that

$$\mathbb{P}\left(\frac{S_n}{n}\to\mu \text{ as } n\to\infty\right)=1,$$

as required. □

2.6 Central limit theorem

The weak law of large numbers tells us that the distribution of S_n/n concentrates around μ as n becomes large. The **central limit theorem** (CLT) goes much further, telling us that (if the random variables X_i have finite variance) the "fluctuations" of S_n around $n\mu$ are of order $\sqrt{n}$. Moreover, the behaviour of these fluctuations is *universal*; whatever the distribution of the X_i, if we scale $S_n-n\mu$ by $\sqrt{n}$, we obtain a normal distribution as the limit as $n\to\infty$.

Theorem (Central Limit Theorem). *Let $X_1,X_2,\ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2\in(0,\infty)$. Let $S_n=X_1+X_2+\cdots+X_n$. Then*

$$\frac{S_n-n\mu}{\sigma\sqrt{n}}\xrightarrow{d}N(0,1)\ \textit{as}\ n\to\infty. \tag{2.7}$$

We will prove the CLT later using generating functions.

Remark 2.7. We can summarise the CLT in three stages:

(1) The distribution of S_n concentrates around $n\mu$;

(2) The fluctuations of S_n around $n\mu$ are of order $\sqrt{n}$;

(3) The asymptotic distribution of these fluctuations is normal.

These are somehow in increasing order of refinement. Some students take in the third of these, but not the first two; they remember that the RHS in (2.7) is a normal distribution, but are hazy about what is going on on the LHS. This is a bit perverse; without knowing the scale of the fluctuations, or what they fluctuate around, knowing their distribution is not so useful!

Example 2.8. An insurance company sells $10,000$ similar car insurance policies. They estimate that the amount paid out in claims on a typical policy has mean £240 and standard deviation £800. Estimate how much they need to put aside in reserves to be 99% sure that the reserve will exceed the total amount claimed.

Solution: Let $\mu = £240$, $\sigma = £800$, $n = 10,000$, and note $\Phi^{-1}(0.99) = 2.326$ where Φ is the distribution function of the standard normal.

Let S_n be the total amount claimed. For large n, the Central Limit Theorem tells us that

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \Phi^{-1}(0.99)\right) \approx 0.99,$$

i.e.

$$\mathbb{P}\left(S_n < \Phi^{-1}(0.99)\sigma\sqrt{n} + n\mu\right) \approx 0.99.$$

So the amount needed in reserves is approximately $\Phi^{-1}(0.99)\sigma\sqrt{n} + n\mu$, which in this case is £$2,586,080$.

Notice that the reserve required per customer is about £258, which is £18 higher than μ. We can see from the calculation above that this surplus is proportional to $n^{-1/2}$. If we had 100 customers rather than 10,000, we would need a surplus 10 times bigger, while with 1,000,000 customers it would be 10 times smaller.

The fact that the amount per customer needed to cover the fluctuations around the mean is decreasing in the number of customers is an example of **risk pooling**.

Of course, the example of identical customers is a bit simplistic, but the effect of risk pooling that we observe is a very real one. Our analysis also assumed that the different customers are independent – is that realistic? For car insurance, it is not such a bad assumption. Similarly for life insurance. In the case of property insurance, it could be a very bad assumption (for example, floods can damage many properties simultaneously). In that situation, the effect of risk pooling is a lot smaller (which explains why obtaining insurance for a property subject to a risk of flooding can be problematic, even if the risk is not that great).

Example 2.9 (Binomial distribution: CLT and Poisson approximation)**.** Let $p \in (0, 1)$ and let Y_n have Binomial(n, p) distribution. Then we can write $Y_n = X_1 + \cdots + X_n$ where the

X_i are i.i.d. Bernoulli(p) random variables. (We can think of the X_i as indicator functions of independent events, all with the same probability, e.g. arising from random sampling).

The X_i each have mean p and variance $p(1-p)$. So we can apply the CLT to obtain

$$\frac{Y_n - np}{\sqrt{n}} \overset{d}{\to} N\big(0, p(1-p)\big) \qquad \text{as } n \to \infty.$$

Now instead of considering fixed p, consider random variables W_n with Binomial(n, p_n) distribution, where $p_n \to 0$ as $n \to \infty$. Now a very different limit applies, describing a situation in which we have a very large number of trials but each one has a very small probability of success. Let $\lambda_n = np_n$, the mean of W_n. Suppose that λ_n converges to a limit λ as $n \to \infty$, so that the expected total number of successes stays approximately constant. Then we will show that W_n converges in distribution to Poisson(λ).

It is enough to show (check!) that for each fixed $k = 0, 1, \dots,$

$$\mathbb{P}(W_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

(since the RHS is the probability that a Poisson(λ) random variable takes the value k).

We have

$$\begin{aligned}
\mathbb{P}(W_n = k) &= \lim \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \binom{n}{k} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda_n^k}{k!} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-k} \\
&= 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1
\end{aligned}$$

as desired.