Probability for Statistics

Kexing Ying

May 15, 2020

Contents

1 Measures

1.1 Probability Measures

Last year we saw briefly constructions and definitions relevant to working with probabilities such as σ-algebras, random variables and more. We will revisit them here with a more general (and more technical) approach.

Definition 1.1 (σ-algebra)**.** Let X be a set. A σ-algebra on X, $\mathcal{A}$ is a collection of subsets of X such that

- $\varnothing \in \mathcal{A}$
- for all $A \in \mathcal{A}$, $A^C \in \mathcal{A}$
- for all $(A_n)_{n=1}^{\infty} \subseteq \mathcal{A}$, $\bigcup_n A_n \in \mathcal{A}$.

Proposition 0.1. *Let X be a set and I a non-empty collection of σ-algebras on X. Then $\bigcap I$ is also a σ-algebra on X.*

This proposition is easy to check and thus, it makes sense to consider the σ-algebra generated by some set.

Definition 1.2 (Generator of σ-algebra)**.** Let X be a set and $S \subseteq \mathcal{P}(X)$ a collection of subsets of X. Then the σ-algebra generated by S is

$$\sigma(S) := \bigcap \{\mathcal{A} \supseteq S \mid \mathcal{A} \text{ is a } \sigma\text{-algebra on } X\}$$

By the fact that the power set of X is a σ-algebra containing S, we see that $\{\mathcal{A} \supseteq S \mid \mathcal{A} \text{ is a } \sigma\text{-algebra on } X\}$ is non-empty and so for all $S \subseteq \mathcal{P}(X)$, $\sigma(S)$ a (and the smallest) σ-algebra on X.

With this, we can construct a commonly seen σ-algebra, the Borel σ-algebra. Given some topological space X, the Borel σ-algebra on X is the σ-algebra generated by $\mathcal{T}_X$, i.e. $\mathcal{B}(X) = \sigma(\mathcal{T}_X)$. We will most commonly work with the Borel σ-algebra on the real numbers $\mathcal{B}(\mathbb{R})$.

We call the ordered pair $(X, \mathcal{A})$ where $\mathcal{A}$ is a σ-algebra n X a *measurable space*.

Definition 1.3 (Measure)**.** Given a measurable space $(X, \mathcal{A})$, a measure on this measurable space $\mu : \mathcal{A} \to [0, \infty]$ is a function such that

- $\mu(\varnothing) = 0$
- for all disjoint sequence $(A_n)_{n=1}^{\infty} \subseteq \mathcal{A}$, $\mu\left(\bigsqcup_n A_n\right) = \sum_n \mu(A_n)$

With measures defined, we can add an additional restriction to create a *probability space*.

Definition 1.4 (Probability Measure)**.** Let μ be a measure on the measurable space $(X, \mathcal{A})$, then μ is a probability measure if and only if $\mu(X) = 1$. We then call the order triplet $(X, \mathcal{A}, \mu)$ a probability space.

To distinguish probability space from normal measure spaces, we will often write $(\Omega, \mathcal{F}, \mathbb{P})$ to denote a probability space. We will call Ω the *sample space*, $\mathcal{F}$ the *events* and for all $A \in \mathcal{F}$, $\mathbb{P}(A)$ the *probability* of the event A.

1.1.1 Some Properties of the Probability Measure

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $(A_i)_{i=1}^{\infty}$ an increasing sequence in $\mathcal{F}$, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

Proof. Follows from additivity of the probability measure by writing $\bigcup_i A_i$ as the disjoint union $A_1 \sqcup \bigsqcup_i (A_{i+1} \setminus A_i)$. □

A corollary of the above is immediately deduced by considering the complement of a decreasing function.

Corollary 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $(A_i)_{i=1}^{\infty}$ a decreasing sequence in $\mathcal{F}$, then*

$$\mathbb{P}\left(\bigcap_i A_i\right) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

In fact the two above propositions apply to general measures with identical proofs.

Theorem 2. *Suppose $(\Omega, \mathcal{F})$ is a measurable space with the finitely additive function $\mathbb{P} : \mathcal{F} \to [0,1]$ such that theorem 1 holds, then $\mathbb{P}$ is a probability measure.*

Proof. Let $(A_i)_{i=1}^{\infty}$ be a sequence of disjoint sequence in $\mathcal{F}$, then, let us define $B_n = \bigcup_{i=1}^{n} A_i$. As σ-algebras are closed under unions, $B_n \in \mathcal{F}$ for all n. Now, by assumption, as (B_n) is increasing, $\mathbb{P}\left(\bigsqcup_i A_i\right) = \mathbb{P}\left(\bigcup_n B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ where the second to last equality if true by finite additivity. □

1.2 The Lebesgue Measure

As the point of measures in general is to assign sets (in the relevant σ-algebra) to some number, it might be useful to take a look at the most famous measure of them all – the Lebesgue measure.

In the easiest terms, the Lebesgue measure is a measure, that maps the interval $[a, b] \subseteq \mathbb{R}$ to the real number $b - a$. In probability, we can think of this as $\mathbb{P}([a, b])$, or the probability of $X \in [a, b]$ where X is a random variable with uniform distribution, (we will talk more about what this means in the next section).

In this course, we will assume the Lebesgue measure exists (and in fact, is unique which we shall prove from first principle in next term's measure theory course).

It turns out that a lot of sets are Lebesgue measurable, in fact, the set of sets that are Lebesgue measurable is greater than the Borel σ-algebra. However, unfortunately, not all sets a Lebesgue measurable. We will give an example of a non-Lebesgue measurable set here called the Vitali set.

Definition 1.5 (The Vitali Set)**.** Let $\Omega := [0, 2\pi)$, then we can some probability measure $\mathbb{P}$ such that $\mathbb{P}(s) = \frac{\beta-\alpha}{2\pi}$ corresponding to the Lebesgue measure. Now, let $\sim$ be the equivalence relation such that $x \sim y$ if and only if $x - y$ is a rational multiple of 2π. As $\sim$, is an equivalence relation, it partitions Ω, so there is a set of equivalence classes $\Omega/\sim$. Now, by using the axiom of choice, the Vitali set is defined to be the set A choosing one element from each equivalence classes in $\Omega/\sim$.

Theorem 3. *The Vitali Set is not measurable with respect to the measure in theorem 1.5.*

Proof. We suppose for contradiction that the Vitali set is measurable. As $\mathbb{Q}$ is countable, let $x_1, x_2, \cdots$ be the enumeration of all rational multiples of 2π in $[0, 2\pi)$. Now, define $A_i := A + x_i = \{a + x_i \mid a \in A\}$. We see that A_i, A_j are disjoint for all $i \neq j$ since if there exists some $a \in A + x_i \cap A + x_j$, so there exists $\alpha, \beta \in A$, $\alpha + x_i = a = \beta + x_j$, so $\alpha \sim \beta$ implying $\alpha = \beta$ by the construction of A and hence, $x_i = x_j$. Now, as $\Omega = \bigsqcup_{i=1}^{\infty} A_i$, we have $1 = \mathbb{P}(\Omega) = \mathbb{P}(\bigsqcup_{i=1}^{\infty} A_i) = \sum \mathbb{P}(A_i)$. However, as the Lebesgue measure is transitional invariant, for all i, j, $\mathbb{P}(A_i) = \mathbb{P}(A_j)$, so $1 = \sum \mathbb{P}(A_i) = \lim_{i\to\infty} i\mathbb{P}(A_1)$ which results in a contradiction by applying excluded middle on $\mathbb{P}(A_1) = 0$. □

2 Random variables

Now that we have the basic notion of a probability space, we would like to play around with it using *random variables*. In the most general sense, random variables are simply functions from the probability space to another measurable space, most commonly the real numbers equipped with $\mathcal{B}(\mathbb{R})$.

Definition 2.1 (Measurable Functions)**.** Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces and $f : X \to Y$ a mapping between the two. We call f measurable if and only if for all $A \in \mathcal{B}$, $f^{-1}(A) \in \mathcal{A}$.

Definition 2.2 (Random Variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{A})$ be a measurable space. Then an E-valued random variable is a measurable function $X : \Omega \to E$.

In general, we will only be working with real valued random variables, so the image measurable space is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Often, when we have a random variable $X : \Omega \to \mathbb{R}$, we might ask questions such as "what is the probability that $X \in A$" for some $A \subseteq \mathrm{Im}X$. We now see that this question is asking for exactly $\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$ (this makes sense as X is measurable).

Theorem 4. *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : \Omega \to \mathbb{R}$ is a function. Then X is a $\mathbb{R}$-valued random variable if and only if for all $x \in \mathbb{R}$,*

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}.$$

Proof. The forward direction is trivial so let us consider the reverse. Suppose for all $x \in \mathbb{R}$, $\{\omega \in \Omega \mid X(\omega) \leq x\} = X^{-1}((-\infty, x]) \in \mathcal{F}$. Then, for all $a, b \in \mathbb{R}$, $a < b$, $X^{-1}((-\infty, a]), X^{-1}((-\infty, b]) \in \mathcal{F}$, so $X^{-1}((-\infty, a])^c = X^{-1}((a, \infty)) \in \mathcal{F}$, and thus, $X^{-1}((a, \infty)) \cap X^{-1}((-\infty, b]) = X^{-1}((a, b]) \in \mathcal{F}$. □

Let us now consider some properties we can but on these random variables.

Definition 2.3 (Identically Distributed Random Variables)**.** Let X, Y be two real valued random variables. We say X and Y are identically distributed if for all $S \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X \in S) = \mathbb{P}(Y \in S).$$

We note that two random variables are identically distributed does not imply they are equal, that is they are not necessarily the same function. An easy example of this if to let X, Y be the number of heads and tails of n coin flips. We see that X, Y are identically distributed by symmetry but definitely not equal.

Another property that is useful for random variables is the notion of independence.

Definition 2.4 (Independence of Events)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n) \subseteq \mathcal{F}$ a sequence of events. Then (A_n) is said to be independent if and only if for all *finite* index set I,

$$\mathbb{P}\left(\bigcap_{n \in I} A_n\right) = \prod_{n \in I} \mathbb{P}(A_n).$$

Definition 2.5 (Independence of σ-algebras)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{A}_n)$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Then $(\mathcal{A}_n)$ is said to be independent if and only if for all $(A_n) \subseteq \mathcal{F}$ a sequence of events such that $A_i \in \mathcal{A}_i$, (A_n) is independent.

Equipped with these two notions of independence, it makes sense to create a notion of some σ-algebra induced by arbitrary measurable functions and with that the notion of independence of random variables is also induced.

Definition 2.6 (σ-algebra Generated by Functions)**.** Let E be a set and $\{f_i : E \to \mathbb{R} \mid i \in I\}$ be an indexed family of real-valued functions. Then the σ-algebra on E generated by these functions is

$$\sigma(\{f_i \mid i \in I\}) := \sigma(\{f_i^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R}), i \in I\}).$$

Note that with this definition, we created the smallest σ-algebra on E such that all f_i are measurable and for a single function f, $\sigma(\{f \mid i \in I\}) = \{f^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R})\}$.

Definition 2.7 (Independence of Random Variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (X_n) be a sequence of real-valued random variables. Then (X_n) is said to be independent if and only if the family of σ-algebras $\sigma(X_n)$ is independent.

We will check that this definition of independence of random variables behave as intended, that is $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

Theorem 5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y be real-valued random variables. Then X, Y are independent if and only if for all $A, B \in \mathcal{B}(\mathbb{R})$,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Proof. Recall that $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}((X \in A) \cap (Y \in B)) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B))$. Now, if $\sigma(X)$ and $\sigma(Y)$ are independent, as $X^{-1}(A) \in \sigma(X)$ and $Y^{-1}(B) \in \sigma(Y)$, by definition, we have $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

Similarly, if the equality in question is true for all $A, B \in \mathcal{B}(\mathbb{R})$, then the σ-algebras are independent by definition, and thus, so are the random variables. □

2.1 Cumulative Distribution Function

We would now like to take a look at the cumulative distribution function of a random variable X.

Definition 2.8 (Cumulative Distribution Function)**.** Given a random variable X, the cumulative distribution function, of simply the CDF of X is

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

This function is well defined as by our previous assertion, X is measurable on $\mathcal{B}(\mathbb{R})$ if and only if $\{\omega \in \Omega \mid X(\omega) \leq x\}$ is measurable for all x.

The CDF of a random variable is important as it characterised the random variable. Formally it can be stated as,

Theorem 6. *Let X, Y be real valued random variables. Then X, Y are identically distributed if and only if $F_X = F_Y$.*

Proof. The forward direction is trivial while the backwards direction follows from the fact that every open real set can be constructed using sets of the form $\{x \leq a | x \in \mathbb{R}\}$ from some $a \in \mathbb{R}$. □

The CDF of a random variable have some nice properties which we have used throughout our first year probability course.

Proposition 6.1. *Given a random variable X, its CDF, F_X, is non-decreasing.*

Proof. This follows from the fact for all $x, y \in \mathbb{R}$, if $x < y$, then we can write $\{\omega \in \Omega \mid X(\omega) \leq y\} = \{\omega \in \Omega \mid X(\omega) \leq x\} \sqcup \{\omega \in \Omega \mid x < X(\omega) \leq y\}$, and so

$$F_X(y) = F_X(x) + \mathbb{P}(x < X \leq y) \geq F_X(x).$$

□

Proposition 6.2. *Given a random variable X, $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.*

Proof. We recall that the axiom that $\mathbb{P}(\Omega) = 1$, so it suffices to prove that $X^{-1}(\lim_{x \to \infty}(-\infty, x])) = \Omega$. But this is trivial as every element of Ω is mapped to a real number so we are done. (This first claim is true by similar argument.)[1] □

Proposition 6.3. *Let X be a random variable, then $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.*

Proof. Similar proof to the previous proposition. □

2.2 Types of Random Variables

The most simple random variable we have is the point mass random variable.

Definition 2.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, then the point mass random variable X_a at a is the function $X_a : \Omega \to \mathbb{R} : \omega \mapsto a$.

We can easily see that the point mass random variable has the CDF $\delta_a(x) = \text{if } (x < a) \text{ then } 0, \text{ else } 1$. While the point mass random variable in itself is not very interesting, it is the building blocks for discrete random variables.

Definition 2.10 (Discrete Random Variable)**.** A random variable X is a discrete random variable if and only if there exist sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ such that $\sum b_i = 1$ and $F_X(x) = \sum b_i \delta_{a_i}(x)$.

Definition 2.11 (Continuous Random Variable)**.** A random variable X is a continuous random variable if and only if F_X is continuous on $\mathbb{R}$.

[1] Note that this proof is not technically true since we can't say $\lim_{x \to \infty}(-\infty, x] = \mathbb{R}$. But this can be fixed by considering any sequence (x_n) that it increasing to ∞.

Definition 2.12 (Absolutely Continuous Random Variable)**.** A random variable X is absolutely continuous if and only if there exists some $f_X : \mathbb{R} \to \mathbb{R}$ such that $F_X(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$.

We note that continuous random variables need not be absolutely continuous (see the Cantor distribution), however, for most purposes, we can assume interchangeability.

Proposition 6.4. *Let X be any random variable and let $x_n \uparrow x \in \mathbb{R}$, then $\mathbb{P}(X < x) = \lim_{x_n \uparrow x} \mathbb{P}(X \leq x_n)$.*

Proof. We define $A_n := \{\omega \in \Omega \mid X(\omega) \leq x_n\}$, then $A_n \uparrow A := \{\omega \in \Omega \mid X(\omega) < x\}$. So, by taking the probability of the limit of A_n, we have $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$. □

Proposition 6.5. *Let X be a continuous random variable, then $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.*

Proof. This follows as the probability measure is continuous. □

While these are the some nicely behaving random variables, often times, random variables appears to be neither discrete or continuous. An example of this is to consider the random variable X representing the units of beer an individual within the population had consumed today.

2.3 Transformations of Random Variables

Often times, we might want to work with transformed random variables. This can be done in many ways, but the most obvious way is to work with the transformed random variable straight away. While this can work in simple cases, we might find it is normally easier to work with the transformed CDF instead. But before we can discuss the consequences of transforming random variables, we should first consider when is a transformed random variable still a random variable.

Recall, by definition, a random variable is a measurable function from the measurable set Ω to some other measurable set, most often the reals. So, for a transformed random variable to also be a random variable, we require it to be measurable as well.

Theorem 7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a real random variable. Then, for all $g : \mathbb{R} \to \mathbb{R}$ where g is measurable with respect to $\mathcal{B}(\mathbb{R})$, $g(X) := g \circ X$ is a random variable.*

Proof. Follows directly from definitions. □

It is in general very hard to construct a non-$\mathcal{B}(\mathbb{R})$-measurable function (but one example of this is the indicator function of the Vitali set), we can regard most transformations of random variables to also be a random variable[2].

Working with transformed random variables is very simple. Say X is a real random variable and g is $\mathcal{B}(\mathbb{R})$-measurable. Then to get the CDF of $g(X)$ (recall that the CDF characterises the random variable) we simply consider $F_{g(X)}(x) = \mathbb{P}(g(X) \in (-\infty, x]) = \mathbb{P}(X \in g^{-1}(-\infty, x])$ which we can obtain using the CDF of X.

[2]For one, continuity implies $\mathcal{B}(\mathbb{R})$-measurable.

3 Multivariate Random Variables

Recall the definition regarding multivariate distributions from year one and we shall in this section consider some of their properties.

Theorem 8. *Let $X_1, \cdots, X_n$ be independent random variables and $f_1, \cdots, f_n$ are Borel measurable real-valued functions, then $f_1(X_1), \cdots, f_n(X_n)$ are also independent.*

Proof. Suppose $B_1, \cdots, B_n \in \mathcal{B}(\mathbb{R})$, then we need to show that $\mathbb{P}(\bigcap_{i=1}^n \{f_i(X_i) \in B_i\}) = \prod_{i=1}^n \mathbb{P}(f_i(X_i) \in B_i)$. By considering by definition $\{f_i(X_i) \in B_i\} = f_i(X_i)^{-1}(B_i) = X_i^{-1}(f_i^{-1}(B_i))$, we have $\mathbb{P}(\bigcap_{i=1}^n \{f_i(X_i) \in B_i\}) = \mathbb{P}(\bigcap_{i=1}^n X_i^{-1}(f_i^{-1}(B_i))) = \prod_{i=1}^n \mathbb{P}(X_i^{-1}(f_i^{-1}(B_i))) = \prod_{i=1}^n \mathbb{P}(f_i(X_i) \in B_i)$ where the third equality is due to the independence of X_i. □

3.1 Covariance and Correlation

While this is a nice theorem on independence of transformed random variables, it is also useful to develop some tools to help us determine whether or not two random variables are independent. Recall the definition of covariance.

Definition 3.1 (Covariance)**.** For random variables X, Y, with finite expectations μ_X, μ_Y respectively, the covariance between X and Y is defined to be

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

By expanding, we see that the covariance between X and Y is equivalently

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

Furthermore, we see that the covariance is zero for independent random variables, however, the reverse is not necessarily true. Indeed, not much can be interpreted from this value as $\mathrm{Cov}(X, Y)$ has the same dimension as XY, thus, it does not make sense the refer to the covariance as big or small, this is instead the role of the correlation.

Definition 3.2 (Correlation)**.** The correlation of the random variables X, Y is

$$\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

We recall from last year that the correlation is always between -1 and 1, so it does make sense to consider the size of the correlation.

From analysis, we recall the definition of a inner product space – that is a vector space equipped with an inner product. By checking the axioms, we find that the covariance of random variables forms an inner product over the space of random variables.

Remark. *The previous statement is not necessarily true since* $\mathrm{Cov}(X, X) = 0$ *for* $X = c$ *for some* $c \in \mathbb{R}$*; that is the covariance does not satisfy positive definiteness. To fix this, we quotient on the set of random variables with the equivalence relation* $X \sim Y$ *if and only if there exists* $c \in \mathbb{R}$, $\mathbb{P}(X = Y + c) = 1$.

3.2 Working with Multivariate Random Variables

3.2.1 Transforming Multivariate Random Variables

Let $D \subseteq \mathbb{R}^2$ and $T : D \to \mathbb{R}^2$ be a function with range $R \subseteq \mathbb{R}^2$. Suppose the partial derivatives of T exist and are continuous. We define the Jacobian of T is

$$J(u,v) = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

Then, if $(U,V) = T(X,Y)$is a function of the pair of random variables (X,Y) with joint probability density function f_{XY}, the joint pdf of (U,V) is

$$f_{UV}(u,v) = f_{XY}(x(u,v), y(u,v)) \left|J(u,v)\right|.$$

3.2.2 Conditioning on Multivariate Random Variables

We recall several definitions from last year. The Bayes' theorem for conditioning on events states

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

We can extend this to condition on any random variables with

$$F_{X|A}(x) = \frac{\mathbb{P}(\{X \leq x\} \cap A)}{\mathbb{P}(A)},$$

where X is a random variable. and $A \in \mathcal{F}$. If X is discrete, we see that

$$f_{X|A}(x) = \mathbb{P}(X = x \mid A),$$

by Bayes', while if X is absolutely continuous, we define

$$f_{X|A}(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{X|A}(x),$$

resulting

$$\mathbb{P}(X \in C \mid A) = \int_C f_{X|A}(x)\mathrm{d}x,$$

for any $C \in \mathcal{B}(\mathbb{R})$.

We need to be careful when working with conditioned probabilities such as $Y \mid X = x$ as $\mathbb{P}(X = x) = 0$ for absolutely continuous X and Y. To deal with this, we instead conditioning on $X \in (x, x+\delta)$ for some $\delta > 0$ and consider the limit of $Y \mid X \in (x, x+\delta)$ as $\delta \to 0$.

$$\mathbb{P}(Y \leq y \mid X = x) = \lim_{\delta \to 0} \mathbb{P}(Y \leq y \mid X \in (x, x+\delta)) = \lim_{\delta \to 0} \frac{\int_x^{x+\delta} \int_{-\infty}^{y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u}{\int_x^{x+\delta} f_X(u)\mathrm{d}u},$$

then by using l'Hopital's rule, we find

$$\mathbb{P}(Y \leq y \mid X = x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)\mathrm{d}v}{f_X(x)},$$

and so, by differentiating,

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

When interpreting this probability, we need to be careful as we should never by conditioning on events with probability 0 (see the Borel-Kolmogorov paradox in the problem sheet).

3.3 Multivariate Normal Distribution

We shall examine one of the most useful distributions – the multivariate normal distribution.

3.3.1 Bivariate Normal Distribution

We recall from last year the bivariate normal distribution with PDF given by

$$f(x, y \mid \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right),$$

where $(x, y) \in \mathbb{R}^2$ and the parameter $-1 < \rho < 1$.

Straight away, we see that whenever $\rho = 0$, the PDF, can be factored into two univariate functions. Therefore, by the factorisation theorem, if $\rho = 0$ we have X, Y are independent (this is **not** true in general).

By completing the square in the exponential, we can compute the marginal density of X and Y.

$$\begin{aligned} f_Y(y) &= \int_{x\in\mathbb{R}} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right) \mathrm{d}x \\ &= \int_{x\in\mathbb{R}} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}((x-\rho y)^2 + (1-\rho^2)y^2)\right) \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) \int_{x\in\mathbb{R}} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) \mathrm{d}x. \end{aligned}$$

By inspection, we see that the integral within the equation evaluates to 1 as its the integral over the support of a univariate normal random variable with mean ρy and variance $1 - \rho^2$. Thus, the marginal density of Y is simply,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2).$$

By, symmetry, we also obtain the marginal distribution of X,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

That is, $X, Y \sim N(0, 1)$.

While, the initial step of completing the square might seem like a cheap trick, it is equivalent to writing $f(x, y) = f(x \mid y)f(y)$. Thus, we find also that $X \mid Y = y \sim N(\rho y, 1 - \rho^2)$.

Let us now look at the covariance of the bivariate normal distribution.

$$\begin{aligned} E(XY) &= \iint_{\mathbb{R}^2} xy f_{X|Y}(x \mid y) f_Y(y) \mathrm{d}x\mathrm{d}y \\ &= \int_{\mathbb{R}} y f_Y(y) \int_{\mathbb{R}} x f_{X|Y}(x \mid y) \mathrm{d}x\mathrm{d}y \\ &= \int_{\mathbb{R}} y f_Y(y) \rho y \mathrm{d}y \\ &= \rho E(Y^2) = \rho, \end{aligned}$$

where the third equality because $X \mid Y = y \sim N(\rho y, 1 - \rho^2)$ as previously mentioned, and thus, has an expectation ρy. Thus, the covariance between X and Y is

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \rho - 0 = \rho,$$

so, in fact, the covariance between X and Y is 0 if and only if they are independent (**not** true in general).

In fact, we see this is simply the *law of iterated expectation*,

$$E(XY) = E_Y(E(X \mid Y = y)) = E_Y(\rho YY) = \rho E(Y^2) = \rho.$$

3.3.2 Multivariate Normal Distribution

We notice that what we have discussed above is the standard bivariate normal distribution – that is, the marginal distributions are standard normal distributions. Of course, in the real world, we rarely see just standard normal distributions but scaled normal distributions. We deal with this similar to how we deal with the 1 dimensional case – transforming it by some affine transformation such that we end up back with a standard bivariate normal distribution.

Definition 3.3 (Multivariate Normal Distribution)**.** The multivariate normal distribution is the probability density function of a vector of normal random variables $\mathbf{X} = (X_1, X_2, \cdots, X_d)$ where $X_i \sim N(\mu_i, \sigma_i^2)$ for all $i = 1, \cdots, d$. Writen out explicitly, we have

$$f_{\mathbf{X}}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \left|\mathbf{\Sigma}\right|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right),$$

where $\mathbf{\Sigma}$ is the covariance matrix with $\mathbf{\Sigma}_{i,j} = \text{Cov}(X_i, X_j)$ and $\mu = (\mu_1, \mu_2, \cdots, \mu_d)$ is the vector of means.

In the two dimension case, we have $\mu = (\mu_X, \mu_Y)$ and

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}.$$

Remark. *We see that*

$$f_{\mathbf{X}}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) \propto \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$

where the inner term $(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)$, is in positive definite quadratic form. In fact, we see that this quantity is the inner product $\langle \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu), (\mathbf{x} - \mu) \rangle$ and thus, represents some distance between $\mathbf{x}$ and μ. This is refereed to as Mahalanobis distance.

Proposition 8.1. *The covariance matrix is symmetric, positive definite and has diagonal $\mathbf{\Sigma}_{i,i} = \sigma_{X_i}$.*

Proof. The first and last property follows directly from the properties of covariance.

To show that the covariance matrix is positive definite, we consider that

$$\operatorname{Var}(a^T X) = \operatorname{Var}\left(\sum_{i=1}^{d} a_i X_i\right) = \sum_{i=1}^{d} a_i^2 \operatorname{Var}(X_i) + 2\sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j) = a^T \mathbf{\Sigma} a$$

for arbitrary $a \in \mathbb{R}^d$. So, as the variance is non-negative, we have $\mathbf{\Sigma}$ is positive definite. $\square$

Let us now consider the linear transformations of a multivariate normal distribution. Let $\mathbf{X} \sim MVN_d(\mu, \mathbf{\Sigma})$ and $\mathbf{Y} = A\mathbf{X}$ for some $A \in GL_d(\mathbb{R})$. So, $\mathbf{X} = A^{-1}\mathbf{Y}$ and $\frac{\partial X_i}{\partial Y_i} = (A^{-1})_{i,j}$ and thus, the Jacobian is simply A^{-1}. Hence,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= f_{\mathbf{X}}(A^{-1}\mathbf{y})\left|A^{-1}\right| \\ &\propto \exp\left(-\frac{1}{2}(A^{-1}\mathbf{y} - \mu)^T \mathbf{\Sigma}^{-1}(A^{-1}\mathbf{y} - \mu)\right) \\ &= \exp\left(-\frac{1}{2}(\mathbf{y} - A\mu)^T (A^{-1})^T \mathbf{\Sigma}^{-1} A^{-1}(\mathbf{y} - A\mu)\right), \end{aligned}$$

and thus $\mathbf{Y} \sim MVN_d(A\mu, A\mathbf{\Sigma}A^T)$.

3.4 Order Statistic

Given a random sample, we often would like to consider the ordering of the sample. This suggests we should investigate the joint distribution of ordered samples from a distribution. We call the random variables of such statistics the *order statistics*. A version of the *order statistic* is the random variable $Y := \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n$ is a random sample (i.e. independence and identically distributed) from an absolutely continuous distribution with CDF F_X and PDF f_X.

We see that Y is a random variable as, by previous results, the composition of measurable functions is measurable, so, as $\max\{X_1, \cdots, X_n\} = g \circ f : \Omega \to \mathbb{R}$ where $f : \Omega \to \mathbb{R}^n : \omega \mapsto (X_1(\omega), \cdots, X_n(\omega))$ and $g : \mathbb{R}^n \to \mathbb{R} = \max$, it suffices to show that g is measurable. Now, it is very easy to show that g is measurable. Let $a \in \mathbb{R}$, then, we see that $g^{-1}(-\infty, a) = \{(x_i) \mid x_i < a\} = (-\infty, a)^n \in \mathcal{B}(\mathbb{R})^n$ and we are done.

In general, given a random sample $X_1, \cdots, X_n$, we denote $X_{(k)}$ or Y_k the k-th smallest sampled variable and we call the the k-th order statistic.

To see the PDF of the k-th order statistic, we consider that $X_{(k)} \leq y$ for some $y \in \mathbb{R}$ if and only if there are at least k X_i which are less than y (since if otherwise $X_{(k)} > y$). So, by independence, we have,

$$F_{X_{(k)}} = \sum_{i=k}^{n} \binom{n}{i} F_X(X \leq y)^i (1 - F_X(X \leq y))^{n-i}.$$

Alternatively, by symmetry and independence, we notice that the joint density of the order statistics is

$$f_{X_{(i)}}(y_i) = n! \prod_i f_X(y_i).$$

4 Convergence of Random Variables

We would sometimes like to consider a sequence of random variables and how they behave with respect to some parameter. As random variables are functions, one might think to use the same notion of convergence we had learnt during last years analysis, that is pointwise convergence and uniform convergence of functions. Nonetheless, we shall look at a different notion of convergence – the convergence of probabilities (or more generally – the convergence of functions in measures).

Suppose we would like to estimate some disease' prevalence within some population. Say, we sample some individuals at random, then a simple model can describe this as

$$X_1, X_2, \cdots, X_n \sim \text{Ber}(p)$$

where X_i are i.i.d. for all i. From last year, we recall that the maximum likelihood estimator for the parameter p is $\hat{p} = \bar{\mathbf{x}}$ with $E(\hat{p}) = E\left(\frac{1}{n}\sum X_i\right) = p$ and $\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \text{Var}(X_i) = \frac{p(1-p)}{n}$. Thus $\hat{p}$ is unbiased and $\text{Var}(\hat{p}) \to 0$ as $n \to \infty$. Heuristically, we can interpret this as the estimator becoming more accurate as n becomes larger and indicates some sort of convergence for the estimator. We formalise the above notion with the converge of random variables.

4.1 Convergence in Measure

Definition 4.1 (Convergence in Measure)**.** Let $(X, \mathcal{A}, \mu)$ be a measure space and let $(f_i)_{i=1}^{\infty}$ be a sequence of measurable functions. Then, $f_i \to f$ in measure for some measurable function f if and only if

$$\mu(\{x \in X \mid |f_n(x) - f(x)| \geq \epsilon\}) \to 0,$$

as $n \to \infty$ for all $\epsilon > 0$.

If $(X, \mathcal{A}, \mu)$ is a probability space then we say f_i converges in probability. That is, given X_i is sequence of random variables, then $X_i \to X$ for some random variable X if and only if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$.

This is a weaker notion than pointwise convergence and in fact, if the sequence of function converges pointwise almost everywhere (i.e. $\mu(\{x \in X \mid f_n(x) \not\to f(x)\}) = 0$), then it converges in measure. The proof of this follows from the property that $\liminf \mu(A_n) \geq \mu(\liminf A_n)$ which definitions we shall encounter later.

4.1.1 Weak Law of Large Numbers

We recall the weak law of large numbers from last year. Now that we are equipped with the formal definition of what it means to converge in probability, we can finally prove it (semi-)formally.

We first recall Markov's and Chebychev's inequality.

Theorem 9 (Markov's Inequality)**.** *Let X be a random variable with $X(\Omega) \subseteq [0, \infty)$. Then for all $a \in \mathbb{R}$,*

$$\mathbb{P}(X \geq a) \leq \frac{E(X)}{a}.$$

Proof. Let $A = [a, \infty)$, then for all $\omega \in \Omega$, $X(\omega) \geq a\mathbb{I}_A(X(\omega))$ where $\mathbb{I}_A$ is the indicator of A; this is because if $X(\omega) < a$, the the right hand side is zero so we are done; on the other hand, if $X(\omega) \geq 0$, then the right hand side is a and the inequality is also true. So, by taking the expectation of this inequality on both sides, we have

$$E(X) \geq aE(\mathbb{I}_A(X)) = a\mathbb{P}(X \geq a).$$

□

Theorem 10 (Chebychev's Inequality)**.** *Let X be a random variable such that $E(X) = \mu, \mathrm{Var}(X) = \sigma^2 < \infty$. Then for any $\epsilon > 0$,*

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. Let $Y := (X - \mu)^2$ which is non-negative. So by Markov's inequality, by letting $a = \epsilon^2$, we are done! □

With that, the weak law of large number follows straight away.

Theorem 11 (Weak Law of Large Numbers)**.** *Let $(X_i)_{i=1}^{\infty}$ be a sequence of i.i.d. random variable such that $E(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2 < \infty$ for all i. Then $\bar{X}_n \to \mu$ is probability where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.*

Proof. By Chebychev's inequality, for all $\epsilon > 0$

$$\mathbb{P}(\left|\bar{X}_n - E(\bar{X}_n)\right| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2}.$$

By independence, we have $E(\bar{X}_n) = \mu$ and $\mathrm{Var}(E(\bar{X}_n)) = \frac{1}{n}\sigma^2$, so

$$\mathbb{P}(\left|\bar{X}_n - \mu\right| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Thus, by taking the limit as $n \to \infty$, we have $\bar{X}_n \to \mu$. □

4.2 Convergence in Distribution

Another notion of convergence of random variables is the convergence of random variables in distribution.

Definition 4.2 (Convergence in Distribution)**.** Let $(X_i)_{i=1}^{\infty}$ be a sequence of random variables with CDFs $(F_i)_{i=1}^{\infty}$ and let X be a random variable with CDF F. Then, we say X_n converges to X in distribution if and only if

$$\lim_{n\to\infty} F_n(x) = F_X(x),$$

at all points $x \in \mathbb{R}$ where F_x is continuous. We denote this by $X_n \xrightarrow{\mathcal{D}} X$.

This notion of convergence is weaker than convergence in probability.

Theorem 12. *Let $(X_i)_{i=1}^{\infty}$ be a sequence of random variables with CDFs $(F_i)_{i=1}^{\infty}$. Suppose X_i converges to X in probability where X is a random variable with CDF F_X, then $X_i \xrightarrow{\mathcal{D}} X$.*

Proof. Let $x \in \mathbb{R}$ such that F_X is continuous at x and let $\epsilon > 0$. By considering

$$\{X_n \leq x\} \subseteq \{X \leq x + \epsilon\} \cup \{|X_n - X| > \epsilon\},$$

by sub-additivity

$$\mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Similarly

$$\mathbb{P}(X \leq x - \epsilon) \leq \mathbb{P}(X_n \leq x) + \mathbb{P}(|X_n - X| > \epsilon).$$

and so,

$$\mathbb{P}(X \leq x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x - \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Thus, as $\epsilon \to 0$, we have by squeeze, $\mathbb{P}(X \leq x) = \mathbb{P}(X_n \leq x)$, and so $X_i \xrightarrow{\mathcal{D}} X$. □

4.3 Limiting Events

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(A_n)_{n=1}^{\infty}$ be a sequence of events in $\mathcal{F}$. We are often interested in considering whether or not A_n occurs *infinitely often*, denoted by $\{A_n \text{ i.o.}\}$, that is

$$\omega \in \{A_n \text{ i.o.}\} \iff \forall N \in \mathbb{N}, \exists n \geq N, \omega \in A_n.$$

So, by definition, for all $n \in \mathbb{N}$, $\omega \in \{A_n \text{ i.o.}\}$ if and only if $\omega \in \bigcup_{n=N}^{\infty} A_n$ and so

$$\{A_n \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n.$$

A closely related is the notion that A_n occurs *almost always*, that is all but finitely many of the A_n occur, we denote this by $\{A_n \text{ a.a}\}$. Formally, we define this as

$$\omega \in \{A_n \text{ a.a.}\} \iff \exists N \in \mathbb{N}, \forall n \geq N, \omega \in A_n.$$

Similarly, we can write this as

$$\{A_n \text{ a.a.}\} = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n.$$

Clearly, $\{A_n \text{ a.a.}\} \subseteq \{A_n \text{ i.o.}\}$, and furthermore, by De Morgan's, we see that $\{A_n \text{ i.o.}\}^c = \{A_n^c \text{ a.a.}\}$.

By recalling the definition of lim sup and lim inf of real sequences from last year, we see that this is an analogous construction of lim sup and lim inf of set with respect to the partial order $\subseteq$. That is, we define

$$\{A_n \text{ i.o.}\} = \limsup_{n \to \infty} A_n = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n;$$

$$\{A_n \text{ a.a.}\} = \liminf_{n\to\infty} A_n = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n.$$

Straight away, we see that $\limsup_{n\to\infty} A_n$ and $\liminf_{n\to\infty} A_n$ are in $\mathcal{F}$ since σ-algebras are closed under countable union and intersections.

Theorem 13 (Borel-Cantelli Lemmas)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose $(A_n)_{n=1}^{\infty}$ is a sequence of events in $\mathcal{F}$. Then,*

- *if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$;*
- *if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, and (A_n) is a independent sequence, then $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$.*

Proof. We provide a proof for the first part while the second part is left as exercise.

Suppose $\sum \mathbb{P}(A_n) < \infty$, then $\sum_{N=n}^{\infty} \mathbb{P}(A_N) \to 0$ as $n \to \infty$. Now, as

$$\mathbb{P}(\limsup_{n\to\infty} A_n) = \mathbb{P}(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n) \leq \mathbb{P}(\bigcup_{N=k}^{\infty} A_n) \leq \sum_{n=k}^{\infty} \mathbb{P}(A_n),$$

for all $k \in \mathbb{N}$, we have

$$0 \leq \mathbb{P}(\limsup_{n\to\infty} A_n) \leq \lim_{k\to\infty} \sum_{n=k}^{\infty} \mathbb{P}(A_n) = 0,$$

implying the first part of the theorem by the squeeze theorem. □

4.4 Convergence Almost Everywhere

Lastly, we have an even stronger notion of convergence than convergence in probability – convergence almost everywhere. But before, we can define this notion, we have to make sure that the event

$$\{X_n \to X\} := \{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}$$

is measurable.

Proposition 13.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose (X_n) is a sequence of random variables. Furthermore, let X also be a random variable. Then*

$$\{X_n \to X\} := \{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\} \in \mathcal{F}.$$

Proof. Consider for all $\omega \in \Omega$, $\omega \in \{X_n \to X\}$ if and only if, for all $m \in \mathbb{N}$, there exists some $N(m) \in \mathbb{N}$ such that for all $n \geq N(m)$,

$$|X_n(\omega) - X(\omega)| < \frac{1}{m}.$$

That is,

$$\omega \in \bigcap_{m=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega \mid |X_n(\omega) - X(\omega)| < 1/m\},$$

which is an event in $\mathcal{F}$ by previous arguments. □

With that, we can define the notion of convergence almost everywhere.

Definition 4.3 (Convergence Almost Everywhere)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose (X_n) is a sequence of random variables. Furthermore, let X also be a random variable. Then we say $X_n \to X$ almost everywhere (or almost surely) if and only if $\mathbb{P}(\{X_n \to X\}) = 1$. We denote this by $X_n \xrightarrow{a.s.} X$.

Proposition 13.2. *$X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{\mathcal{P}} X$, that is, convergence almost everywhere implies convergence in probability.*

Proof. For all $\epsilon > 0$, define the sequence of events,

$$A_N := \{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| < \epsilon, n \geq N\}.$$

We see that A_N is increasing and $\{X_n \to X\} \subseteq \bigcup A_N$, so,

$$1 = \mathbb{P}(\{X_n \to X\}) \leq \mathbb{P}\left(\bigcup A_N\right) \leq 1.$$

Hence, by continuity, $\lim_{N\to\infty} \mathbb{P}(A_N) = \mathbb{P}(\bigcup A_N) = 1$. Now as $A_N \subseteq \{|X_N - X| < \epsilon\}$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1,$$

which is equivalent to $X_n \xrightarrow{\mathcal{P}} X$. □

4.4.1 Strong Law of Large Numbers

Theorem 14 (Strong Law of Large Numbers)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables such that $E(X_n^4) < \infty$ and $E(X_n) = \mu$ for all n. Then*

$$\mathbb{P}(\bar{X}_n \to \mu) = 1.$$

Proof. Suppose we define $Z_n := X_n - \mu$, and $S_n := \sum_{i=1}^n X_i$ and consider

$$E((S_n - n\mu)^4) = E((\sum_{i=1}^n Z_i)^4) = nE(Z_1^4) + 3n(n-1)E(Z_1^2 Z_2^2),$$

where the second equality is true as $E(Z_i) = 0$ for all i and the all terms but the above two of the sum expansion contains it. Now, by defining $C := 4\max\{E(Z_1^4), E(Z_1^2)^2\}$,

$$nE(Z_1^4) + 3n(n-1)E(Z_1^2 Z_2^2) = n^2\left(\frac{C}{4n} + \frac{3C(n-1)}{4n}\right) \leq Cn^2,$$

and so,

$$E((\bar{X}_n - \mu)^4) = \frac{1}{n^4}E((S_n - n\mu)^4) \leq \frac{C}{n^2}.$$

Now, let $A_n := \{\left|\bar{X}_n - \mu\right| \geq n^{-1/8}\}$, then by Markov's inequality,

$$\mathbb{P}(\left|\bar{X}_n - \mu\right| \geq n^{-1/8}) \leq \frac{E(\bar{X}_n - \mu)^4}{n^{-2}} \leq Cn^{-3/2}.$$

Lastly, by taking the sum on both sides, we have $\sum \mathbb{P}(A_n) < \infty$, and so, by the Borel-Cantelli lemma, $\mathbb{P}(\{A_n^c \text{ a.a.}\}) = 1$, which implies straight away $\mathbb{P}(\bar{X}_n \to \mu) = 1$. □

5 Central Limit Theorem

5.1 Moment Generating Functions

We recall the definition of the moment generating functions of a random variable from last year.

Definition 5.1 (Moment Generating Function)**.** Let X be a random variable, then the moment generating function of X, is the function $M_X(t) = E(\exp(tX))$.

Also, recall some basic properties of the moment generating function.

Proposition 14.1. *Let X be a random variable and suppose there exists some $a, b \in \mathbb{R}$ such that $Y = aX + b$. Then $M_Y(t) = \exp(bt)M_X(at)$*

Proposition 14.2. *Let X, Y be independent random variables and let $Z = X + Y$. Then $M_Z = M_X M_Y$.*

Proposition 14.3. *Let X be a random variable and suppose there exists $t_0 > 0$ such that for all $t \in \mathbb{R}, |t| < t_0$, $M_X(t) < \infty$, then*

$$M_X(t) = \sum_{k=0}^{\infty} E(X^k)\frac{t^k}{k!},$$

and for all $k \geq 0$,

$$\frac{d^k}{dt^k}M_X(t) \mid_{t=0} = E(X^k).$$

Furthermore, we can show that this moment generating function is unique and continuous under certain conditions.

Proposition 14.4. *Suppose X and Y are random variables with common moment generating function M which is finite for all $t \in \mathbb{R}$, $|t| < t_0$ for some $t_0 > 0$. Then X and Y are identically distributed.*

Proposition 14.5. *Suppose X is a random variable with moment generating function M_X and $(X_n)_{n\geq 1}$ is a sequence of random variables, with respective moment generating function $M_{X_i}(t)$. If $M_{X_i}(t) \to M_X(t) < \infty$ as $n \to \infty$ for all $t \in \mathbb{R}$, $|t| < t_0$ for some $t_0 > 0$, then $X_n \xrightarrow{\mathcal{D}} X$.*

The above proposition provides us with another proof of the weak law of large numbers.

Proof. (Weak Law of Large Numbers again). Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables with moment generating function $M(t)$. Consider the moment generating function of $\bar{X}_n$,

$$M_{\bar{X}_n}(t) = E\left(\exp\left(\frac{t}{n}\sum_{i=1}^{n}(X_i)\right)\right) = \prod_{i=1}^{n} E\left(\exp\left(\frac{tX_i}{n}\right)\right) = M\left(\frac{t}{n}\right)^n.$$

Now, by considering the Taylor expansion of $M(t) = 1 + \mu t + o(t)$, we have

$$M_{\bar{X}_n}(t) = \left(1 + \frac{\mu t}{n} + o\left(\frac{t}{n}\right)\right)^n \to \exp(\mu t),$$

as $n \to \infty$. We see that this is the moment generating function of a constant random variable, and thus, by proposition 14.5, $\bar{X}_n \to \mu$ in distribution. Now, as μ is constant, we have $\bar{X}_n \to \mu$ in probability. □

We shall use a similar method when proving the central limit theorem in the next section.

5.2 Central Limit Theorem

Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2. We would like to consider the behaviour of $\bar{X}_n$ as n becomes large. As we have previously seen, the weak law of large numbers dictates that $\bar{X}_n \to \mu$ in probability as $n \to \infty$ but this is not sufficient for our analysis; instead, we would like to consider the behaviour of the region around μ as n becomes large. To achieve this, we shall consider the following transformation,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

in which we standardise both the mean and the variance. By empirical experimentation we find that as n becomes large, Z_n seems to become standard normal. This is not a coincidence and we shall formalise this observation with the central limit theorem.

Theorem 15 (The Central Limit Theorem)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Then, by defining*

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma},$$

then, $Z_n \xrightarrow{\mathcal{D}} Z \sim N(0,1)$.

Proof. Let $M(t)$ be the moment generating function of $X_i - \mu$ and furthermore, let $M_n(t)$ be the moment generating function of Z_n. By expanding the definitions, we have

$$M_n(t) = E\left(\exp\left(\frac{t}{\sigma\sqrt{n}}\sum(X_i - \mu)\right)\right) = \prod E\left(\exp\left(\frac{t}{\sigma\sqrt{n}}(X_i - \mu)\right)\right) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n.$$

Now, by considering the Taylor expansion of $M(t)$ at zero, that is

$$M(t) = 1 + tM'(0) + \frac{t^2}{2}M''(0) + o(t^2) = 1 + \frac{t^2}{2}\sigma^2 + o(t^2),$$

we have,

$$M_n(t) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to \exp\left(\frac{t^2}{2}\right),$$

as $n \to \infty$ resulting in the moment generating function of standard normal random variable. Hence, by proposition 14.5, we have $Z_n \xrightarrow{\mathcal{D}} Z \sim N(0,1)$. □

Remark. *We assumed within the proof that the moment generating function of X_i exists in some open interval around 0. This is in general not true, but the central limit theorem can be proved in the general case using characteristic functions.*

6 Stochastic Processes

6.1 Markov Chains

Definition 6.1 (Stochastic Process)**.** A stochastic process on the state space $\mathcal{E}$ is a collection of $\mathcal{E}$-valued random variables $(X_t)_{t\in\mathcal{T}}$ indexed by some set $\mathcal{T}$.

We note that we are no longer necessarily using real valued random variables and instead, considering random variables on some arbitrary measurable set $\mathcal{E}$. In general, we shall consider stochastic processes on discrete time, that is, the index set $\mathcal{T} = \mathbb{N}$.

Definition 6.2 (Markov Chain)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a discrete time stochastic process on the state space $\mathcal{E}$. Then, we say (X_n) is a Markov chain if and only if

$$\mathbb{P}(X_n = x_n \mid \bigcap_{i\leq n}\{X_i = x_i\}) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

That is, the stochastic process is memoryless.

A consequence of the Markov assumption results in

$$\mathbb{P}(X_n = z, X_{n-2} = x \mid X_{n-1} = y) = \mathbb{P}(X_n = z \mid X_{n-1=y}\mathbb{P}(X_{n-2} = x \mid X_{n-1} = y).$$

That is, X_n is conditionally independent of X_{n-2} given X_{n-1}.

Definition 6.3 (Time Homegenoeous)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain on some state space $\mathcal{E}$. Then, we say (X_n) is time homogeneous if and only if for all $n \in \mathbb{N}$, $i, j \in \mathcal{E}$,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_1 = j \mid X_0 = i).$$

A time homogeneous Markov chain results in the probability of a particular state to be independent of the indexing $\mathcal{T}$. We shall mostly look at time homogeneous Markov chains in this course.

Time homogeneous Markov chains can be represented nicely in matrix called the transition matrix.

Definition 6.4 (Transition Matrix)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a time homogeneous Markov chain and suppose $p_{ij} = \mathbb{P}(X_1 = j \mid X_0 = i)$ for all $i, j \in \mathcal{E}$. Then the transition matrix of (X_n) is the matrix P where $[P]_{ij} = p_{ij}$.

Since the entries of transition matrix are probabilities, we have $0 \leq p_{ij} \leq 1$. Furthermore, by total probability, we have the sum of each row $\sum_{j\in\mathcal{E}} p_{ij} = \sum_{j\in\mathcal{E}} \mathbb{P}(X_1 = j \mid X_o = i) = 1$.

The transition matrix tells us that, given a state, how the system will change in the next time frame. However, this is not sufficient to characterise a whole Markov chain as we still do not know how the system began. Thus, to fully specify the stochastic process, we will also provide the initial distribution, that is $\lambda = (\lambda_j)_{j\in\mathcal{E}} \in \mathbb{R}^n$ where $\lambda_j = \mathbb{P}(X_0 = j)$ and $n = |\mathcal{E}|$.

Sometimes, however, we are interested in the probability of reaching some state given some previous state.

Definition 6.5. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with state space $\mathcal{E}$. Then, the n-step transition matrix $P(n)$ is the matrix with entries

$$p_{ij}(n) = \mathbb{P}(X_n = j \mid X_0 = i).$$

Clearly, we have $P(1) = P$ the transition matrix and $P(0) = I$. In general however, the Chapman-Kolmogorov equations provides us explicitly the n-step transition probability.

Proposition 15.1 (The Chapman-Kolmogorov Equations)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with state space $\mathcal{E}$. Suppose now that, $m \geq 0$ and $n \geq 1$, then*

$$p_{ij}(n) = \sum_{l\in\mathcal{E}} p_{il}(m)p_{lj}(n).$$

As matrices,

$$P(m+n) = P(m)P(n).$$

From the second part of the proposition, we can deduce that $P(m) = P^m$.

Proof. By considering the law of total probability and the properties of time homogeneous Markov chains,

$$\begin{aligned} p_{ij}(n) = \mathbb{P}(X_{m+n} = k \mid X_0 = i) &= \sum_{l\in\mathcal{E}} \mathbb{P}(X_{m+n} = j, X_m = l \mid X_0 = i) \\ &= \sum_{l\in\mathcal{E}} \mathbb{P}(X_{m+n} = j \mid X_n = l, X_0 = i)\mathbb{P}(X_m = l \mid X_0 = i) \\ &= \sum_{l\in\mathcal{E}} \mathbb{P}(X_{m+n} = j \mid X_n = l)\mathbb{P}(X_m = l \mid X_0 = i) \\ &= \sum_{l\in\mathcal{E}} \mathbb{P}(X_n = j \mid X_0 = l)\mathbb{P}(X_m = l \mid X_0 = i) = \sum_{l\in\mathcal{E}} p_{il}(m)p_{lj}(n). \end{aligned}$$

□