STAT2003 Content Notes

Ryan White s4499039

Semester 1, 2022

Contents

1 Week 1

Random Experiments

A *random experiment*, or *trial*, is an experiment whose *outcome* is not known in advance. Three ingredients are needed to model a random experiments:

i. A ***Sample Space***

ii. A collection of ***Events***

iii. A way to assign ***Probability*** to events

Events

Sample Space: The set Ω of all possible outcomes is called the *sample space*. Outcomes are sometimes called the *sample points*.

Events: An *event* is any set of outcomes (any subset of Ω). Events are usually denoted by upper case letters, A, B, C, etc.
We say that an event A *occurs* if the outcome of the random experiment is one of the elements in A.

Combining Events: If A and B are two events, then the event that A or B occurs is represented by the ***union*** of A and B, denoted $A \cup B$, and consists of all outcomes in A *or* B (or in both).
If A and B are two events, then the event that A and B occur is represented by the ***intersection*** of A and B, denoted $A \cap B$, and consists of all outcomes in both A and B.

De Morgan's Laws (complements): If A is an event, then A^c (pronounced A complement) is the event that A does not occur, and consists of all outcomes in Ω which are *not* in A. De Morgan's Laws read

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c \quad \text{and} \quad \left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c$$

The left reads as: The complement of at least one A_i occurring results in none of the A_i occurring.
The right reads as: The complement of all A_i occurring results in at least one of the A_i not occurring.

Subsets: If A and B are two events, with all outcomes in A also being in B, then A is a *subset* of B, written $A \subset B$. i.e. if A occurs, then B must also occur (A *implies* B).

Empty Set: The empty set $\varnothing$ is called the ***impossible event*** because it never occurs, while its complement Ω is called the ***certain event*** because it always occurs.

Disjoint Events: Two events, A and B, are said to be *disjoint* if they have no outcomes in common. That is, $A \cap B = \varnothing$.

Mutual Exclusivity: More generally, $A_1, A_2, \ldots$ are said to be ***mutually exclusive*** if each pair of events (A_i, A_j), $i \neq j$, is disjoint.

Exhaustivity: Events $A_1, A_2, \ldots$ are said to be ***exhaustive*** if $A_1 \cup A_2 \cup \cdots = \Omega$. That is, at least one of the events must occur.

Partitions: A collection of mutually exclusive and exhaustive events, $\mathcal{P} = \{A_1, A_2, \ldots\}$ is called a *partition* of Ω.

Venn diagrams are an effective way of visualising the relationship between events.

Probability

In order to measure the relative likelihood of events, we define a function $\mathbb{P}$ in such a way that, for any event A, $\mathbb{P}(A)$ can be interpreted as the "probability of A" occurring.
Suppose that a random experiment is performed N times, and let A be some arbitrary event. Let $N(A)$ be the number of times that A occurs. The, as N becomes larger, the ratio of $N(A)/N$ *should* "settle down" to some constant value $\mathbb{P}(A)$. This can be interpreted as the ***Law of Large Numbers***.

The Probability Axioms

We settle on three basic properties (or *axioms*) from which all other properties can be deduced.

Probability Measures: The function $\mathbb{P}$ is called a *probability measure* if it satisfies:

a. $\mathbb{P}(A) \geq 0 \quad \forall A$

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

c. if $A_1, A_2, \ldots$ are *mutually exclusive* events, then $\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots$

From these axioms, one can find:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
$$\text{If } A \subset B, \text{ then} \quad \mathbb{P}(A) \leq \mathbb{P}(B)$$

Boole's Inequality: For any events $A_1, A_2, \ldots$

$$\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$$

The Equilikely Principle: If Ω is a finite set and all outcomes in Ω are equilikely, then for any event A,

$$\boxed{\mathbb{P}(A) = \frac{\text{the number of outcomes in } A}{\text{the number of outcomes in } \Omega} = \frac{|A|}{|\Omega|}}$$

This principle follows from the third axiom of probability. Since Ω is a finite set of n outcomes,

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

and since each outcome is equilikely, $\mathbb{P}(\{\omega_i\}) = 1/n$. Since A is an event with $k \leq n$ outcomes,

$$\mathbb{P}(A) = \mathbb{P}(\{\omega_1\}) + \cdots + \mathbb{P}(\{\omega_k\}) = k/n$$

Counting

Generally, if one chooses k objects from a collection of n ***without replacement***, then the number of ways of doing this is

$$n(n-1)(n-2)\dots(n-k+1)$$

We give this quantity the symbol ${}^nP_k$. Notice that

$$\boxed{{}^nP_k = \frac{n!}{(n-k)!}}$$

where $n!$, pronounced n factorial, is given by

$$n! = n \times (n-1) \times \cdots \times 2 \times 1$$

Generally, if N is the number of combinations of k objects taken from a collection of n ***without attention to order***, then

$$N \times^k P_k =^n P_k$$

The quantity N is given the symbol ${}^nC_k$, or $\binom{n}{k}$ and pronounced n choose k. Clearly, then,

$$\boxed{\binom{n}{k} = \frac{{}^nP_k}{{}^kP_k} = \frac{n!}{(n-k)!k!}}$$

In summary,

Replacement?	Order	
	Important	Unimportant
With	n^k	$\binom{n+k-1}{k}*$
Without	${}^nP_k$	$\binom{n}{k}$

Table 1: The Number of Samples of Size k from n Objects

2 Week 2

Conditional Probability

Given two events, A and B, how do we evaluate $\mathbb{P}(A|B)$, read as "the conditional probability that A occurs given that B occurs."
Suppose that we run an experiment N times, and let $N(E)$ be the number of times that a given event E occurs. Then, as N becomes large, we would expect the ratio $N(E)/N$ to settle down to some $\mathbb{P}(E)$. In the same way, we would expect $N(A \cap B)/N(B)$ to settle down to $\mathbb{P}(A|B)$. Moreover, since

$$\frac{N(A \cap B)}{N(B)} = \frac{N(A \cap B)/N}{N(B)/N}$$

we should evaluate $\mathbb{P}(A|B)$ as

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

Observe that $\mathbb{P}(B)$ must be greater than zero for this to have meaning.

Product Formula: Let D be the event $B \cap A$. Then, for three events,

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

Independence

Definition: Two events are said to be ***independent*** if

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

Definition: Suppose we have a collection of events, $A_1, A_2, \dots$ They are said to be *pairwise independent* if, for all $j \neq i$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$.
They are said to be *triplewise independent* if, for all distinct i, j and k, $\mathbb{P}(A_i \cap A_j \cap A_k) = \mathbb{P}(A_i)\mathbb{P}(A_j)\mathbb{P}(A_k)$.
Similar definitions hold for quadruplewise independent, etc.

More generally, they are ***mutually independent*** if they are pairwise independent *and* triplewise independent *and* quadruplewise independent, etc. Note that pairwise independent events are not necessarily mutually independent.

Suppose there are two events, A and B, such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. If A and B are mutually exclusive, that is $A \cap B = \varnothing$, they cannot also be independent. This is because $\mathbb{P}(A \cap B) = \mathbb{P}(\varnothing) = 0 \neq \mathbb{P}(A)\mathbb{P}(B) > 0$.
This is reasonable, for if we know that A has occurred, then B can't have occurred. That is, ***independent does not mean mutually exclusive***.

If A and B are independent events, then A^c and B^c are independent, i.e.

$$\mathbb{P}\left(A^c \cap B^c\right) = \mathbb{P}\left(A^c\right)\mathbb{P}\left(B^c\right)$$

The Law of Total Probability

Definition: If A and B are two arbitrary events such that $0 < \mathbb{P}(B) < 1$, then

$$\boxed{\mathbb{P}(A) = \mathbb{P}(A|B)\,\mathbb{P}(B) + \mathbb{P}(A|B^c)\,\mathbb{P}(B^c)}$$

A more general form of the law of total probability is:
If $B_1, B_2, \dots$ form a partition of Ω such that $\mathbb{P}(B_i) > 0$ for at least one value of i, then, for any event A,

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

with the interpretation that the i-th term is absent from the sum if $\mathbb{P}(B_i) = 0$.

An important corollary of the Law of Total Probability is:

Bayes' Theorem

Definition: If A and B are arbitrary events such that $\mathbb{P}(A) > 0$ and $0 < \mathbb{P}(B) < 1$, then

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

More generally, if $B_1, B_2, \ldots$ form a partition Ω, such that $\mathbb{P}(B_i) > 0$ for at least one value of i, and A is any event with $\mathbb{P}(A) > 0$, then, for each i,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}$$

Random Variables

Definition: A random variable X is a function that assigns a numerical value to each outcome of an experiment.

Definition: A random variable is said to be ***discrete*** if it takes values in a countable set S (S is the range of X). If X is a random variable that takes values in S, then the function given by

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

is called the ***probability (mass) function*** of X.

Properties of the Probability Function:

Let X be a discrete random variable taking values in S. Then,

i. $f_x(x) \geq 0$ for all real x, and

ii. $\sum_{x \in S} f_x(x) = 1$

where the last statement follows from the Law of Total Probability.

Definition: Let X be an arbitrary random variable. The ***(cumulative) distribution function*** F of X is defined by

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

Theorem: If X is a discrete random variable taking values in S, and with probability function f, then

$$F(x) = \sum_{y : y \leq x} f(y)$$

Properties of the Distribution Function

The distribution function F of a random variable X has the following properties:

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

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

iii. F is non-decreasing. That is, if $x < y$ then $F(x) \leq F(y)$

iv. F is continuous from the right. That is, for all x, $F(x + h) \to F(x)$ as $h \to 0$ (from above).

3 Week 3

Expectation

Definition: Let X be a discrete random variable taking values in S and with probability function f. Then, $\mathbb{E}(X)$ (or $\mathbb{E}X$), pronounced ***the expected value*** of X, is given by

$$\boxed{\mathbb{E}(X) = \sum_{x \in S} x f(x) = \sum_{x \in S} x \mathbb{P}(X = x)}$$

This is a natural idea, and $\mathbb{E}(X)$ is a weighted average of the values that X takes, weighted according to their probabilities.
The symbol μ_X is frequently used to denote $\mathbb{E}(X)$.

Theorem: (***Law of the Unconscious Statistician***). Let X be a discrete random variable taking values in S with probability mass function f. If $Y = h(X)$, where h is some function, then

$$\mathbb{E}(Y) = \mathbb{E}(h(X)) = \sum_{x \in S} h(x) f(x) = \sum_{x \in S} h(x) \mathbb{P}(X = x)$$

compare that with the definition of the expected value above, and you'll see they are very similar.

Properties of $\mathbb{E}$: If X and Y are two discrete random variables, then

i. $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ for all a and b

ii. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

iii. if $X \geq Y$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$

Definition: Let X be any random variable. Then, $\mathrm{Var}(X)$, pronounced the ***variance*** of X, is given by

$$\boxed{\mathrm{Var}(X) = \mathbb{E}(X - \mu_X)^2}$$

where $\mu_X = \mathbb{E}(X)$.
Frequently, σ_X^2 is used to denote $\mathrm{Var}(X)$. Furthermore, $\sigma_X = \sqrt{\sigma_X^2}$ is called the ***standard deviation*** of X.
These quantities measure ***variation*** about μ_X, or the ***spread*** of the distribution of X.
Alternatively,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X - \mu_X)^2 = \mathbb{E}\left(X^2\right) - 2\mu_X \mathbb{E}(X) + \mu_X^2 \\ &= \mathbb{E}\left(X^2\right) - \mu_X^2\end{aligned}$$

also note that

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

Continuous Random Variables

Definition: A random variable X is said to be ***continuous*** if it takes values, not in a countable set, but in an ***interval***.

Definition: Let X be a continuous random variable. Then, the ***probability density function*** (pdf) of X is a function f which satisfies $f(x) \geq 0$ for all x and

$$\boxed{\mathbb{P}(x \leq X \leq y) = \int_x^y f(u)du}$$

for all $x \leq y$, that is the probability that X lies in the interval $[x, y]$ is the area under the graph of the pdf from x to y.

Frequently, f_X is used to denote the pdf of X.

Note that

$$\int_{-\infty}^{\infty} f_X(u)du = 1$$

that is, the total area under the pdf is 1. Notice also that

$$F_X(X) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

for all x; here F_X is the distribution function of X.
It follows that if $f_X(x)$ exists, then $F'_X(x) = f_X(x)$.
Warning: In contrast to discrete random variables, $\mathbb{P}(X = x) = 0$ for all x. And, note carefully, $f(x)$ is *not* a probability. In particular, it is *not true* that $f(x) = \mathbb{P}(X = x)$ for all x.

Definition: Let X be a continuous random variable with pdf f. Then, the expected value of X is defined as:

$$\mathbb{E}(X) = \mu_X = \int_{-\infty}^{\infty} uf(u)du$$

In a way that is analogous to the definition of expectation for discrete random variables, $\mathbb{E}(X)$ is a weighted average of the values in the range of X, weighted according to the density f.
All of the properties of $\mathbb{E}$ mentioned in the last section hold for continuous random variables. Furthermore, the definition and properties of Var are the same.

Other measures of Centrality: Other measures of centrality are the ***mode*** of X; the value of x which maximises $f(x)$ (the mode may not be unique), the ***median*** of X; the *unique* number m with the property that $\mathbb{P}(X \leq m) = 1/2$.

Theorem: (Law of the Unconscious Statistician for cont. variables). Let X be a continuous random variable with pdf f and let h be any function. Then,

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(u)f(u)du$$

with

$$\mathbb{E}\left(X^2\right) = \int_{-\infty}^{\infty} u^2 f(u)du$$

Transforms

Definition: ***(Probability Generating Functions)*** Let X be a *non-negative discrete* random variable (with range being some subset of $\{0, 1, \dots\}$). The *probability generating function* (PGF) of X is defined to be

$$G(z) = G_X = \mathbb{E}\left(z^X\right) = \sum_{k=0}^{\infty} z^k \mathbb{P}(X = k), \quad |z| \leq 1$$

Mathematically, G is just a *power series* with coefficients $a_k = \mathbb{P}(X = k)$.
The series G converges for all $|z| \leq 1$ and often for $|z| \leq R$, with $R > 1$.
Furthermore, G can be differentiated or integrated termswise (on sets like $\{z : |z| \leq R_0 < R\}$).

Useful Properties of G:

i. G determines the distribution of X uniquely, with

$$\mathbb{P}(X = k) = \frac{G^{(k)}(0)}{k!}, \quad k \geq 0$$

ii. $\mathbb{E}(X) = G'(1)$

iii. More generally,

$$\mathbb{E}(X(X-1)\dots(X-k+1)) = G^{(k)}(1), \quad k \geq 1$$

iv. As a consequence, $\mathrm{Var}(X) = G''(1) + G'(1) - (G(1))^2 = \mathbb{E}\left(X^2\right) - (G(1))^2 = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$

Definition: ***(Moment Generating Function)*** The *moment generating function* (MGF) of a random variable X is the function M given by

$$M(t) = \mathbb{E}\left(e^{tx}\right)$$

provided the expectation exists on some open interval I containing 0 (here, t is usually taken to be real). We often write M_X to stress the role of X.
Thus, if X is a continuous random variable,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \quad t \in I$$

while if X is a discrete random variable,

$$M_X(t) = \sum_n e^{tx_n} f_X(x_n), \quad t \in I$$

where $\{x_1, x_2, \dots\}$ is the range of X. Notice that if X takes values in the non-negative integers, then $M_X(t) = G_X(e^t)$, $t \in I$, where G_X is the PGF of X, so we would not normally use MGFs in this case.

Useful Properties of M:

i. Two MGFs are the same if and only if their distribution function are the same.

ii. $\mathbb{E}(X^n) = M^{(n)}(0), \quad n \geq 1$

iii. Consequently, $\mathrm{Var}(X) = M''(0) - (M'(0))^2 = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$

4 Week 4

Bernoulli Distribution

Definition: A random experiment that has precisely two outcomes in called a ***Bernoulli Trial***. Traditionally, the outcomes are denoted S and F, and represent "success" and "failure" respectively.
The sample space is $\Omega = \{S, F\}$. Clearly, we must have $\mathbb{P}(\{S\}) = p$, and $\mathbb{P}(\{F\}) = 1 - p$ for some $0 < p < 1$.

The discrete random variable X defined by

$$X(S) = 1 \quad \text{and} \quad X(F) = 0$$

is called a ***Bernoulli random variable***. It "indicates" whether or not a success occurs, in that it takes the value 0 or 1 according to the experiment result.
Any event has a Bernoulli random variable that is naturally associated with it. Let $A \subseteq \Omega$ and define the random variable

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

Then I_A is a Bernoulli random variable. Note that

$$\mathbb{P}(I_A = 1) = \mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(I_A = 0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

The probability mass function and cumulative distribution function of such an X has the simple forms:

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

The expectation and variance of X are given by

$$\begin{aligned} \mathbb{E}(X) &= 0 \times (1 - p) + 1 \times p = p \\ \text{Var}(X) &= \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 \\ &= (0^2 \times (1 - p) + 1^2 \times p) - p^2 \\ &= p(1 - p) \end{aligned}$$

Suppose now that a ***sequence*** of n Bernoulli trials is performed. Let X be the number of successes, defined the sum of X_i — the Bernoulli random variable representing the outcome on the ith trial. Then, $X = X_1 + X_2 + \cdots + X_n$, and

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) = np$$

In this sequence of n trials, there are 2^n possible outcomes, with ${}^nC_x$ outcomes corresponding to exactly x successes. Each outcome corresponding to exactly x successes (and consequently, $n - x$ failures) has probability $p^x(1 - p)^{n-x}$, the probability that, in n trials, there are exactly x successes is

$$\mathbb{E}(X = x) =^n C_x\, p^x(1 - p)^{n-x}$$

Thus, the probability function of X (the number of successes) is given by

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) \\ &= \begin{cases} {}^nC_x\, p^x(1 - p)^{n-x} & \text{if } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

X is said to have a ***binomial distribution*** and we write $X \sim \text{Bin}(n, p)$. The expected value and the variance of X are given by

$$\mathbb{E}(X) = np \quad \text{and} \quad \text{Var}(X) = np(1 - p)$$

Poisson Distribution

Definition: A random variable X is said to have a ***Poisson Distribution*** with parameter $\lambda (> 0)$ if its probability function is given by

$$f(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$$

The Poisson distribution with parameter $\lambda = np$ approximates the binomial $\text{Bin}(n, p)$ distribution when n is large and p is small (which would be hard to calculate manually by the binomial distribution).

Proposition: If X has a Poisson distribution with parameter λ, then $\mathbb{E}(X) = \text{Var}(X) = \lambda$

Geometric Distribution

Definition: A random variable X is said to have a ***geometric distribution*** with parameter $p : (0 < p < 1)$ if its probability function is given by

$$f(x) = \begin{cases} p(1 - p)^{x-1} & \text{if } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Proposition: If X has a geometric distribution with parameter p, then $\mathbb{E}(X) = 1/p$ and $\text{Var}(X) = (1 - p)/p^2$

Uniform Distribution

Definition: A random variable X is said to have a ***uniform distribution*** on $[a, b]$ (written $X \sim U[a, b]$) if

$$\mathbb{P}(x \leq X \leq y) = \frac{y - x}{b - a}, \quad a \leq x \leq y \leq b$$

The distribution function of X is given by

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases}$$

and its pdf is given by

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

Furthermore,

$$\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{2}(a+b)$$

and $\mathrm{Var}(X) = \frac{1}{12}(b-a)^2$ (by computation of the above integral with x^2 and the usual variance formula).

The uniform distribution is characterized by each point in the domain being equilikely. This is evident in the expectation value being independent of x.

Exponential Distribution

Definition: A random variable X is said to have an ***exponential distribution*** with parameter $\lambda > 0$ if its distribution function F is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

Proposition: The pdf of X is given by

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

and

$$\mathbb{E}(X) = 1/\lambda \quad \text{and} \quad \mathrm{Var}(X) = 1/\lambda^2$$

Hypergeometric Distribution

Definition: We say that a random variable X has a ***hypergeometric distribution*** with parameters N, n and r if

$$\mathbb{P}(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

for $\max\{0,\, r+n-N\} \leq x \leq \min\{n, r\}$.
We write $X \sim \mathrm{Hyp}(n, r, N)$. The expectation and variance of the hypergeometric distribution are

$$\mathbb{E}(X) = n\frac{r}{N} \quad \text{and} \quad \mathrm{Var}(X) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1}$$

Notice that in the binomial distribution, $\mathbb{E}(Y) = np$.

Normal Distribution

Suppose we have a sequence of n (independent) Bernoulli trials, each with probability p of "success". Let X be the number of successes. With p fixed and n becoming large, X has an approximate normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

Definition: A continuous random variable X is said to have a ***Normal distribution*** (or ***Gaussian distribution***) with parameters μ and σ^2 (written $X \sim N(\mu, \sigma^2)$) if its pdf is given by

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

If $X \sim N(0,1)$, that is,

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

then X is said to have a ***standard*** normal distribution.

Suppose $X \sim N(\mu, \sigma^2)$ and let $Z = (X-\mu)/\sigma$. Using the properties of expectation and variance,

$$\mathbb{E}(Z) = \frac{\mathbb{E}(X)}{\sigma} - \frac{\mu}{\sigma} = 0$$
$$\mathrm{Var}(Z) = (1/\sigma)^2\,\mathrm{Var}(X) = 1$$

Proposition: $Z \sim N(0,1)$
Note: For any random variable with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, the random variable $Z = (X-\mu)/\sigma$ is called the ***standardized*** form of X because it satisfies $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

Gamma Distribution

Definition: X is said to have a ***gamma distribution*** with parameters n and λ if its pdf is given by

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{(\lambda x)^{n-1}}{(n-1)!}\lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

and, it can be shown that

$$\mathbb{E}(X) = \frac{n}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{n}{\lambda^2}$$

A more general version of the gamma distribution has the integer parameter n replaced by a positive real parameter α:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}\lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

where Γ is the ***gamma function*** given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx, \quad \alpha > 0$$

Its mean and variance are given by

$$\mathbb{E}(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}$$

The gamma distribution is denoted by $\mathrm{Gamma}(\alpha, \lambda)$.
Notice that the $\mathrm{Gamma}(1, \lambda)$ distribution is the exponential distribution. If $\lambda = 1/2$ and $\alpha = \nu/2$, for some positive integer ν, then X is said to have a ***chi-squared distribution with ν degrees of freedom***; this distribution is denoted by χ^2_ν.

5 Week 5

Simulations and Random Number Generation

Random experiments on a computer are called ***stochastic simulations***. In these, randomness is introduced via

uniform random numbers, which are used as building blocks to simulate more general stochastic systems.

Given positive integers a, c and m, and a ***seed*** X_0, generate $X_1, X_2, \ldots$ via the linear recurrence

$$X_{i+1} = (aX_i + c) \mod m$$

This means that aX_i+c is divided by m, and the remainder is taken as the value of X_{i+1}. Thus, each $X_i \in \{0, 1, \ldots, m-1\}$ and the quantities

$$U_i = \frac{X_i}{m}$$

called ***pseudorandom numbers***, constitute approximations to the true sequence of uniform random variables. Note that the sequence $\{X_i\}$ will repeat itself in at most m steps.

Inverse-Transform Method

The ***inverse-transform method*** is a general method for generating one-dimensional random variables from a prescribed distribution.
Let F be a cdf with inverse F^{-1}. If $U \sim U(0,1)$, then

$$X = F^{-1}(U)$$

has cdf F. Namely,

$$\mathbb{P}(X \le x) = \mathbb{P}\left(F^{-1}(U) \le x\right) = \mathbb{P}(U \le F(x)) = F(x)$$

Thus, to generate a random variable X with cdf F, draw $U \sim U(0,1)$ and set $X = F^{-1}(U)$.

The inverse-transform method required that the underlying cdf, F, has an inverse function F^{-1} that can be computed fast.
Even in the case where F^{-1} exists in an explicit form, the inverse-transform method may not necessarily be the most efficient random variable generation method.

Acceptance — Rejection

The ***acceptance-rejection method*** (ARM) is a general method for simulating random variables.
Suppose that the pdf from which we want to sample is bounded on some finite interval $[a, b]$, and zero outside this interval. Let

$$c = \max\{f(x) : x \in [a, b]\}$$

In this case, we can generate $Z \sim f$ in the following way:

1. Generate $X \sim U(a, b)$
2. Generate $Y \sim U(0, c)$
3. If $Y \le f(X)$ accepted the point (X, Y) and return $Z = X$. Otherwise, reject the point and go back to step 1.

Note that:

- Each random point (X, Y) is uniformly distributed over the rectangle $[a, b] \times [0, c]$.
- Therefore, the accepted pair (X, Y) is uniformly distributed under the graph of f.
- This implies that the distribution of the accepted values of X has the desired pdf f.

We can generalise this as follows. Let g be a pdf that

- is easy to sample from,
- for which there is a constant C such that

$$\phi(x) = Cg(x) \ge f(x) \quad \forall x$$

We call $g(x)$ the ***proposal*** pdf. The general acceptance-rejection algorithm can be written as:
Algorithm 1 (Acceptance-Rejection):

1. Generate $X \sim g$
2. Generate $Y \sim U(0, Cg(x))$
3. If $Y \le f(X)$, return $Z = X$. Otherwise, return to step 1.

The random variable X returned by the algorithm has pdf f. The efficiency of an ARM is defined as

$$\mathbb{P}((X, Y) \text{ is accepted}) = \frac{\text{Area under } f}{\text{Area under } Cg} = \frac{1}{C}$$

For an ARM to be of practical interest, the following criteria must be used in selecting the proposal density $g(x)$

1. It should be easy to generate a random variable from $g(x)$
2. The efficiency, $1/C$, of the procedure should be large. That is, C should be close to 1 (which occurs when $g(x)$ is close to $f(x)$)

Simulating From Some Known Distributions

Exponential Distribution

Algorithm 2 (Generation of $X \sim \exp(\lambda)$)

1. Generate $U \sim U(0, 1)$
2. Return $X = -1/\lambda \ln U$ as a random variable from $\exp(\lambda)$.

Normal Distribution

We can draw from $N(0, 1)$ based on the ARM. Note that in order to generate $Y \sim N(0, 1)$, one can first generate a nonnegative random variable X from the pdf

$$f(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}, \quad x \ge 0$$

and then assign to X a random sign. To generate a random variable X from the above pdf, we bound $f(x)$ by $Cg(x)$, where $g(x) = e^{-x}$ is the pdf of the $\exp(1)$. The smallest constant C such that $f(x) \le Cg(x)$ is $C = \sqrt{2e/\pi}$.

Bernoulli Distribution

If $X \sim \text{Ber}(p)$, its pmf is of the form

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

where p is the success probability. Applying the inverse-transform method, we obtain:

Algorithm 3 (Generation of $X \sim \text{Ber}(p)$)

1. Generate $U \sim U(0, 1)$
2. If $U \leq p$, return $X = 1$, otherwise return $X = 0$.

Binomial Distribution

If $X \sim \text{Bin}(n, p)$, then its pmf is of the form

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

(Recall that a binomial random variable X can be interpreted as the sum of n Bernoulli random variables).

Algorithm 4 (Generation of $X \sim \text{Bin}(n, p)$)

1. Generate iid random variables $X_1, \ldots, X_n$ from $\text{Ber}(p)$
2. Return $X = \sum_{i=1}^{n} X_i$ as a random variable from $\text{Bin}(n, p)$

Geometric Distribution

If $X \sim \text{Geom}(p)$, then its pmf is of the form

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

The random variable X can be interpreted as the number of trials required until the first success occurs, in a series of independent Bernoulli trials with success parameter p.

To generate a random variable from $\text{Geom}(p)$, we first generate a random variable from the exponential distribution with $\lambda = -\ln(1-p)$, truncate the obtained value to the nearest integer and add 1.

Algorithm 5 (Generation of $X \sim \text{Geom}(P)$)

1. Generate $Y \sim \exp(-\ln(1-p))$
2. Return $X = 1 + \lfloor Y \rfloor$ as a random variable from $\text{Geom}(p)$

Joint Distributions

Definition: The ***joint cdf*** F of two random variables, X and Y, is given by

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

To emphasize that this is the *joint* cdf of X and Y, we often write $F_{X,Y}$. It is clear that

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

(and similarly for F_Y). To emphasize their distinctness from $F_{X,Y}$, and the fact that they can be *derived* from $F_{X,Y}$, F_X and F_Y are referred to as ***marginal cdfs***.

Definition (Independent Random Variables): Two random variables X and Y are said to be *independent* if $F_{X,Y} = F_X F_Y$. That is,

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\,\mathbb{P}(Y \leq y) \quad \forall x, y \in \mathbb{R}$$

Definition (Jointly Discrete Random Variables): The ***joint probability mass function*** f of two discrete random variables X and Y is given by

$$f(x, y) = \mathbb{P}(X = x, Y = y), \quad x, y \in \mathbb{R}$$

To emphasise that f is the joint probability function of X and Y, $f_{X,Y}$ is often written.

If $S_X = \{x_1, x_2, \ldots\}$ is the range of X and $S_Y = \{y_1, y_2, \ldots\}$ is the range of Y, then $S_{X,Y} = S_X \times S_Y$ is the range of (X, Y)

f has the following properties:

- $f_X(x) = \sum_y f_{X,Y}(x, y)$
- $f_Y(y) = \sum_x f_{X,Y}(x, y)$
- $f_{X,Y}(x, y) > 0$ if and only if $(x, y) \in S_{X,Y}$
- $\sum_x \sum_y f_{X,Y}(x_i, y_j) = 1$
- $F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f_{X,Y}(u, v)$

Note that f_X and f_Y are called ***marginal probability functions***.

Theorem: The discrete random variables X and Y are independent if and only if $f_{X,Y} = f_X f_Y$

Definition (Jointly Continuous Random Variables): Two random variables X and Y are said to be ***jointly continuous*** with ***joint probability density function*** (joint pdf) $f_{X,Y}$ if

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv, \quad x, y \in \mathbb{R}$$

The joint pdf is not prescribed uniquely by this definition, but, if both of the partial derivatives of $F_{X,Y}$ exist at the point (x, y), then

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

The marginal cdfs F_X and F_Y can be expressed in terms of $f_{X,Y}$; for example (for x):

$$F_X(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v)\, du\, dv, \quad x \in \mathbb{R}$$

Thus, X and Y are (individually) continuous random variables with (marginal) pdfs f_X and f_Y; for example (again for x):

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy, \quad x \in \mathbb{R}$$

Further properties of f:

- $f_{X,Y}(x,y) \geq 0$ for all $x,\, y \in \mathbb{R}$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
- $\mathbb{P}\left(a \leq X \leq b,\, c \leq Y \leq d\right) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx$, where a or c can be $-\infty$ and b or d can be $+\infty$, and any of the inequalities can be replaced by strict ones.
- More generally, if A is a subset of $\mathbb{R}^2$, then

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

6 Week 6

Correlation

Correlation is a measure of ***linear*** dependence between random quantities.

Definition: The ***correlation*** (or correlation coefficient) of X and Y is defined by

$$\varrho(X,Y) = \frac{\mathrm{Cov}\left(X,Y\right)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}}$$

where $\mathrm{Cov}\left(X,Y\right)$, the ***covariance*** of X and Y, is given by

$$\begin{aligned}\mathrm{Cov}\left(X,Y\right) &= \mathbb{E}\left((X - \mathbb{E}\left(X\right))(Y - \mathbb{E}\left(Y\right))\right)\\ &= \mathbb{E}\left(XY\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\end{aligned}$$

X and Y are said to be positively or negatively correlated according to the sign of $\varrho(X,Y)$ respectively, otherwise they are uncorrelated. The larger the value of $|\varrho(X,Y)|$, the more strongly correlated X and Y are.

The following theorem follows from a version of the ***Cauchy-Schwarz Inequality***.

Theorem: $|\varrho(X,Y)| \leq 1$ with equality if and only if X and Y are (almost surely) linearly dependent. That is, $\mathbb{P}\left(Y = aX + b\right) = 1$ for some $a, b \in \mathbb{R}$ with $a \neq 0$.
We also have the following important identity:

$$\mathrm{Var}\left(X+Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\,\mathrm{Cov}\left(X,Y\right)$$

In order to calculate $\mathrm{Cov}\left(X,Y\right)$, we first need $\mathbb{E}\left(XY\right)$. More generally, how do we evaluate the expectation of a function of two (or more) random variables?
Suppose that $g : \mathbb{R}^2 \to \mathbb{R}$ is a function of two variables, and let Z be the random variable given by $Z(\omega) = g(X(\omega), Y(\omega))$.

Theorem (Law of the Unconscious Statistician): If X and Y are jointly discrete, then

$$\mathbb{E}\left(Z\right) = \sum_i \sum_j g(x_i, y_j) f_{X,Y}(x_i, y_j)$$

If X and Y are jointly continuous, then

$$\mathbb{E}\left(Z\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

Theorem: For any pair of *independent* random variables X and Y, we have $\mathbb{E}\left(XY\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$. That is, *independent random variables are uncorrelated.* The converse of this is not true.

Further properties of $\mathrm{Cov}\left(\right)$:

- $\mathrm{Cov}\left(X,X\right) = \mathrm{Var}\left(X\right)$
- For any $a, b, c, d \in \mathbb{R}$,

$$\mathrm{Cov}\left(aX = b, cY + d\right) = ac\,\mathrm{Cov}\left(X,Y\right)$$

  It follows that if either X or Y is constant, then $\mathrm{Cov}\left(X,Y\right) = 0$.

- For any collection of random variables, $X_1, X_2, \ldots, X_n$, we have

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}\left(X_i, X_j\right)\\ &= \sum_i \mathrm{Var}\left(X_i\right) + \sum_i\sum_{j\neq i} \mathrm{Cov}\left(X_i, X_j\right)\end{aligned}$$

- More generally, for any collections of random variables $X_1, \ldots$ and $Y_1, \ldots,$

$$\mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n\sum_{j=1}^m \mathrm{Cov}\left(X_i, Y_j\right)$$

Conditional Distributions

Assume that, given $X = x$, $Y \sim N(ax+b, \sigma^2)$. This is the idea of a conditional distribution: for a given value x of one of the random variables, we have a specified distribution of the other random variable *as a function of* x, the ***conditional distribution of*** Y ***given*** $X = x$.

As a consequence, we have

$$\mathbb{E}\left(Y|X=x\right) = ax + b \qquad \mathrm{Var}\left(Y|X=x\right) = \sigma^2$$

These are the ***conditional expected value*** and the ***conditional variance*** of Y given $X = x$.

Conditioning on a Discrete Random Variable

Let Y be a random variable with cdf F_Y and let X be a *discrete* random variable with pmf f_X.

Definition: The *conditional cdf of* Y *given* X is the function $F_{Y|X}$ given by

$$F_{Y|X}(y|x) = \mathbb{P}\left(Y \leq y\,|X = x\right), \quad y \in \mathbb{R}$$

Here, x must satisfy $\mathbb{P}\left(X = x\right) > 0$.
If Y is also a discrete random variable, then the *conditional pmf of* Y *given* X is the function $f_{Y|X}$ given by

$$f_{Y|X}(y|x) = \mathbb{P}\left(Y = y\,|X = x\right) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad y \in \mathbb{R}$$

Again, x must satisfy $\mathbb{P}(X = x) > 0$.
We use the notation $F_{Y|X}(y|x)$ and $f_{Y|X}(y|x)$ in order to emphasize that both $F_{Y|X}$ and $f_{Y|X}$ depend on x. It is sometimes better not to think of them as functions of two variables but rather that by varying x, we obtain a *family* of functions of one variable y.

Write $S_X = \{x_1, x_2, \dots\}$ for the range of X. Then, by the Law of Total Probability,

$$\mathbb{P}(Y \le y) = \sum_i \mathbb{P}(Y \le y | X = x_i)\, \mathbb{P}(X = x_i)$$

and, if Y is a discrete random variable,

$$\mathbb{P}(Y = y) = \sum_i \mathbb{P}(Y = y | X = x_i)\, \mathbb{P}(X = x_i)$$

That is,

$$f_Y(y) = \sum_i f_{Y|X}(y, x_i) f_X(x_i)$$

The above is true when X is a discrete random variable. When X is a continuous random variable,

$$f_Y(y) = \int_{\text{all } x} f_{Y|X}(y, x) f_X(x)\, dx$$

Conditioning on a Continuous Random Variable

Let Y be a random variable with cdf F_Y and let X be a *continuous* random variable with pdf f_X. Then, under mild conditions, there exists a function $F_{Y|X}$ with the property that, for all $y \in \mathbb{R}$,

$$F_Y(y) = \int_{-\infty}^{\infty} F_{Y|X}(y|x) f_X(x)\, dx$$

$F_{Y|X}$ is called the ***conditional cdf of*** Y ***given*** X, and it admits the following interpretation:

$$F_{Y|X} = \lim_{h \to 0} \mathbb{P}(Y \le y \,|x < X \le x + h)$$

If Y is a continuous random variable, then there exists a function $f_{Y|X}$ called the ***conditional pdf of*** Y ***given*** X, with the property that

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(u|x)\, du \quad y \in \mathbb{R}$$

This satisfies $f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)$, whenever $f_X(x) > 0$, where $f_{X,Y}(x, y)$ is the joint pdf of X and Y.
In particular,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\, dx$$

Conditional Expectation

Definition: Let X be a random variable, either continuous or discrete, with pdf/pmf f_X. If Y is a *discrete* random variable, then the expected value of Y given $X = x$ is defined to be

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

where $f_{Y|X}(\cdot|x)$ is the pmf of Y given X. Here, x must satisfy $f_X(x) > 0$.
If Y is a *continuous* random variable, then the expected value of Y given $X = x$ is

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy$$

where $f_{Y|X}(\cdot|x)$ is the pdf of Y given X. Again, x must satisfy $f_X(x) > 0$.

Definition: The variance of Y given $X = x$ is defined to be

$$\begin{aligned}\text{Var}(Y|X = x) &= \mathbb{E}\left((Y - \mathbb{E}(Y|X = x))^2 | X = x\right) \\ &= \mathbb{E}\left(Y^2 | X = x\right) - (\mathbb{E}(Y|X = x))^2\end{aligned}$$

The expected value and variance of Y given $X = x$ are just the mean and variance of the conditional distribution of Y given $X = x$.

If X is a discrete random variable, then

$$\mathbb{E}(Y) = \sum_x \mathbb{E}(Y|X = x)\, f_X(x)$$

while if X is a continuous random variable,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} \mathbb{E}(Y|X = x)\, f_X(x)\, dx$$

Definition (Conditional Expectation): Let X and Y be two random variables. The function ψ, given by

$$\psi(x) = \mathbb{E}(Y|X = x), \qquad x \in \mathbb{R}$$

is called the ***regression curve*** of Y on X. The conditional expected of Y given X, denoted $\mathbb{E}(Y|X)$, is given by $\mathbb{E}(Y|X) = \psi(X)$.
Note carefully: The conditional expectation of Y given X is a random variable, because it is a function of X. So, we could evaluate its expectation, done by the following theorem:

Theorem: $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$.

Some further properties of conditional expectation:

- $\mathbb{E}(Y|Y) = Y$
- $\mathbb{E}(aY|X) = a\mathbb{E}(Y|X)$, for all $a \in \mathbb{R}$
- $\mathbb{E}(X + Y|Z) = \mathbb{E}(X|Z) + \mathbb{E}(Y|Z)$
- $\mathbb{E}(Yg(Z)|Z) = g(Z)\mathbb{E}(Y|Z)$ — for "nice" functions g.
- If Y is non-negative, then $\mathbb{E}(Y|Z) \ge 0$. More generally, if $Y \ge X$, then $\mathbb{E}(Y|Z) \ge \mathbb{E}(X|Z)$
- ***Tower Property:***

$$\mathbb{E}(\mathbb{E}(Y|(X, Z))|Z) = \mathbb{E}(Y|Z)$$

- ***The Role of Independence:*** if X and Y are independent, then $\mathbb{E}(Y|X)$ is almost surely constant; indeed, $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ almost surely.
- $\text{Var}(Y) = \mathbb{E}(\text{Var}(Y|X)) + \text{Var}(\mathbb{E}(Y|X))$

7 Week 7

Transformations

If X is a random variable and $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ has an inverse g^{-1}, then we can easily obtain the distribution of Y from that of X. If g is *increasing*, then, for all $y \in \mathbb{R}$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y))$$

If g is *decreasing*, then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(X \geq g^{-1}(y)\right) = 1 - \lim_{x \uparrow y} F_X(g^{-1}(x))$$

So, if X is a *continuous* random variable,

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

For example, if Y is a linear function of X, that is, $Y = aX + b$, where $a, b \in \mathbb{R}$, $a \neq 0$, then, for all $y \in \mathbb{R}$,

$$F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & \text{if } a > 0 \\ 1 - \lim_{x \uparrow y} F_X\left(\frac{x-b}{a}\right) & \text{if } a < 0 \end{cases}$$

If X is a *continuous* random variable, we get

$$F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & \text{if } a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right) & \text{if } a < 0 \end{cases}$$

and, if F_X is differentiable at $x = (y-b)/a$,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

If X is a *discrete* random variable, then the argument is simple:

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}\left(X = \frac{y-b}{a}\right) = f_X\left(\frac{y-b}{a}\right)$$

Transformation Rule

Let X and $Y = g(X)$ be continuous random variables with pdfs f_X and f_Y. Suppose that g has inverse g^{-1} and is differentiable. We have

$$f_Y(y) = F_Y'(y) = \frac{f_X(x)}{|g'(x)|} = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

Random Vectors

Dealing with $n > 2$ random variables, that is, a ***random vector*** $\vec{X} = (X_1, X_2, \ldots, X_n)^T$, is conceptually no more difficult than dealing with two.
The distribution of $\vec{X}$ is completely determined by the joint cdf $F : \mathbb{R}^n \to [0, 1]$ given by

$$F(x) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

for $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

Definition: If $\vec{X} = (X_1, X_2, \ldots, X_n)^T$ is an n-dimensional random vector, then the ***covariance matrix***, written $\mathrm{Cov}\left(\vec{X}\right)$, is the $n \times n$ (symmetric) matrix $\vec{V} = (v_{ij})$ with elements $v_{ij} = \mathrm{Cov}(X_i, X_j)$; in particular, $v_{ii} = \mathrm{Var}(X_i)$.

Properties of $\mathrm{Cov}\left(\vec{X}\right)$:

- If $\vec{X} = (X_1, \ldots, X_n)^T$ is an n-dimensional random vector (column vector), with $\vec{\mu} = \mathbb{E}\left(\vec{X}\right)$ being the corresponding vector of expected values, then $\mathrm{Cov}\left(\vec{X}\right) = \mathbb{E}\left(\vec{X}\vec{X}^T\right) - \vec{\mu}\vec{\mu}^T$
- $\mathrm{Cov}\left(\vec{X}\right)$ is a ***symmetric matrix***.
- $\mathrm{Cov}\left(\vec{X}\right)$ is a ***positive semi-definite***, meaning

  $$\vec{x}^T \mathrm{Cov}\left(\vec{X}\right) \vec{x} \geq 0 \quad \text{for all } \vec{x} \in \mathbb{R}^n$$

- If $\vec{X} = (X_1, \ldots, X_n)^T$ and $\vec{Y} = (Y_1, \ldots, Y_n)^T$ are both n-dimensional random vectors, then

  $$\mathrm{Cov}\left(\vec{X} + \vec{Y}\right) = \mathrm{Cov}\left(\vec{X}\right) + \mathrm{Cov}\left(\vec{Y}\right)$$

  if and only if X_i and Y_j are uncorrelated for every i, j. In general, we have that

  $$\left(\mathrm{Cov}\left(\vec{X} + \vec{Y}\right)\right)_{ij} = \left(\mathrm{Cov}\left(\vec{X}\right)\right)_{ij} + \left(\mathrm{Cov}\left(\vec{Y}\right)\right)_{ij} + \mathrm{Cov}(X_i, Y_j) + \mathrm{Cov}(Y_i, X_j)$$

  In particular, if $\vec{b} = (b_1, \ldots, b_n)$ is a vector of constants, then $\mathrm{Cov}\left(\vec{X} + \vec{b}\right) = \mathrm{Cov}\left(\vec{X}\right)$
- Let $\vec{X} = (X_1, \ldots, X_n)^T$ be an n-dimensional random vector and let $\vec{A}$ be an $m \times n$ matrix.
  Define $\vec{Y} = (Y_1, \ldots, Y_m^T$ by $\vec{Y} = \vec{A}\vec{X}$. Then,

  $$\mathrm{Cov}\left(\vec{Y}\right) = \vec{A}\,\mathrm{Cov}\left(\vec{X}\right)\vec{A}^T$$

Sums of Random Variables

If X and Y are independent, non-negative discrete random variables, then so is $Z = X + Y$, and

$$f_Z(m) = \sum_{n=0}^{m} f_X(n) f_Y(m-n) \quad (m \in \mathbb{Z})$$

while if X and Y are independent, non-negative continuous random variables, then so is Z with

$$f_Z(z) = \int_0^z f_X(x) f_Y(z-x)\, dx \quad (z \geq 0)$$

Discrete Random Variables

Let X and Y be two discrete random variables and let $Z = X + Y$. Then, we have

$$f_Z(z) = \sum_x f_{Z|X}(z|x) f_X(x), \quad z \in \mathbb{R}$$

However, since $Z = X + Y$, we may write

$$f_{Z|X}(z|x) = \mathbb{P}(Z = z|X = x) = \mathbb{P}(Y = z - x|X = x)$$

and so, for $z \in \mathbb{R}$,

$$f_Z(z) = \sum_x f_{Y|X}(z - x|x) f_X(x) = \sum_x f_{X,Y}(x, z - x)$$

By conditioning on Y instead of X,

$$f_Z(z) = \sum_y f_{X|Y}(z - y|y) f_Y(y) = \sum_y f_{Y,X}(y, z - y)$$

If X and Y are independent, then

$$f_Z(z) = \sum_x f_X(x) f_Y(z - x) = \sum_y f_X(z - y) f_Y(y)$$

And notice that if X and Y are independent, the probability function of Z is the ***convolution*** of the probability functions of X and Y, often written as $f_Z = f_X \cdot f_Y (= f_Y \cdot f_X)$.
This implies that $f_{X_1+X_2+\cdots+X_n} = f_{X_1} \cdot f_{X_2} \cdot \ \cdots \ \cdot f_{X_n}$.

Continuous Random Variables

If X and Y are jointly continuous, the treatment is similar. If $Z = X + Y$, then,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z|X}(z|x) f_X(x)\, dx \quad z \in \mathbb{R}$$

Since $Z = X + Y$, we have that Z, conditional on $X = x$, has the same distribution as $Y + x$, conditional on $X = x$. That is,

$$f_{Z|X}(z|x) = f_{Y|X}(z - x|x)$$

Hence,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{Y|X}(z - x|x) f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} f_{X|Y}(x|z - x) f_X(x)\, dx \end{aligned}$$

By conditioning on Y instead of X,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{Y|X}(y|z - y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} f_{X|Y}(z - y|y) f_Y(y)\, dy \end{aligned}$$

If X and Y are independent, then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy$$

Thus, if X and Y are independent, the pdf of Z is the convolution of the pdfs of X and Y, as in the discrete variable case.

Often, sums are easier to handle via *transforms*.
For example, when X and Y are independent, non-negative discrete random variables, then

$$\mathbb{E}\left(z^{X+Y}\right) = \mathbb{E}\left(z^X z^Y\right) = \mathbb{E}\left(z^X\right) \mathbb{E}\left(z^Y\right)$$

8 Week 8

Functions of Multiple Random Variables

Consider the continuous case of multiple random variables. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be an n-dimensional random vector (taken to be a column vector) with pdf f_X, and let $\boldsymbol{A}$ be a non-singular $n \times n$ matrix.
Define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ by the ***linear transformation*** $\boldsymbol{Y} = \boldsymbol{AX}$.
We have already seen that $\operatorname{Cov}(\boldsymbol{Y}) = \boldsymbol{A} \operatorname{Cov}(\boldsymbol{X}) \boldsymbol{A}^T$.
We can also identify the distribution of $\boldsymbol{Y}$. Since $\boldsymbol{A}$ is invertible, the transformation is invertible with $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{Y}$, and the 1-D case:

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$$

generalizes to

$$f_Y(\boldsymbol{y}) = \frac{1}{|\det(\boldsymbol{A})|} f_X\left(\boldsymbol{A}^{-1}\boldsymbol{y}\right) \quad \boldsymbol{y} \in \mathbb{R}^n$$

Special care must be taken in specifying the region over which f_Y is positive.
Note that any n-dimensional rectangle with volume V is transformed into an n-dimensional parallelepiped with volume $V|\det(\boldsymbol{A})|$.

Now, let's consider the more general case where we're given an n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ whose pdf f_X is specified, as well as the invertible map $g : \mathbb{R}^n \to \mathbb{R}^n$. We wish to identify the distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, where $\boldsymbol{Z} = g(\boldsymbol{X})$.
Since g is bijective, there is, for each $\boldsymbol{z} \in \mathbb{R}^n$, a unique $\boldsymbol{x} \in \mathbb{R}^n$ with $g(\boldsymbol{x}) = \boldsymbol{z}$ (written $\boldsymbol{x} = g^{-1}(\boldsymbol{z})$).
The ***Jacobi matrix*** of g, that is the matrix of first partial derivatives, has a fundamental role to play. First, we have the transformation

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_n(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

Then the Jacobi of g at $\boldsymbol{x}$ is

$$J_{\boldsymbol{X}}(g) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{bmatrix} \qquad \text{notation: } \frac{\partial g}{\partial \boldsymbol{x}} \text{ or } \frac{\partial z}{\partial \boldsymbol{x}}$$

The determinant of this matrix is called the ***Jacobian***.
The transformation formula for the linear case,

$$f_Z(\boldsymbol{z}) = \frac{f_X(\boldsymbol{A}^{-1}\boldsymbol{z})}{|\det(\boldsymbol{A})|} \qquad \boldsymbol{z} \in \mathbb{R}^n$$

generalizes to

$$f_Z(\boldsymbol{z}) = \frac{f_X(g^{-1}(\boldsymbol{z}))}{|\det(\partial \boldsymbol{z}/\partial \boldsymbol{z})|} \qquad \boldsymbol{z} \in \mathbb{R}^n$$

The Jacobi matrix of g^{-1} is the inverse of the Jacobi matrix of g. In elegant notation:

$$\left(\frac{\partial \boldsymbol{x}}{\partial z}\right) = \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right)^{-1}$$

So we also have:

$$f_Z(\boldsymbol{z}) = f_X(g^{-1}(\boldsymbol{z}))\left|\det\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}\right)\right|, \quad \boldsymbol{z} \in \mathbb{R}^n$$

In any given application, one should take care to determine the domain and the range of the map, and then check carefully that it is a bijection.

Suppose that $Z_1 = X_1X_2$ and $Z_2 = X_1$. Clearly, $X_1 = Z_2$ and $X_2 = Z_1/Z_2$. Then,

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \begin{bmatrix} x_2 & x_1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{bmatrix}$$

with determinant $-x_1$ $(= -z_2)$. It follows that for $(z_1,\, z_2) \in \mathbb{R}^2$,

$$f_{Z_1,Z_2}(z_1, z_2) = \frac{f_{X_1,X_2}(z_2,\, z_1/z_2)}{|z_2|}$$

The probability density function of $Z_1 = X_1X_2$ is obtained by "integrating out" z_2.

9 Week 9

Bivariate Normal Distribution

We have already seen a special bivariate normal pdf:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2}Q(x,y)\right) \qquad (1)$$

where $x, y \in \mathbb{R}$ and Q is the quadratic function

$$Q(x,y) = \frac{1}{(1-\varrho^2)}(x^2 - 2\varrho xy + y^2)$$

and ϱ is a constant satisfying $-1 < \varrho < 1$.
On integrating over x and y, we find that the marginal distributions are both $N(0,1)$:

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \qquad x \in \mathbb{R}$$

For $\varrho = 0$, we have the *standard* bivariate normal distribution, with joint pdf

$$f(x,y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2+y^2)\right) \qquad x,\, y \in \mathbb{R}$$

and note that this is the joint distribution of two independent standard normal random variables: $X,\, Y \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. Thus, for the bivariate normal pdf in equation (1), X and Y are independent ($f_{X,Y}$ factorises as $f_X f_Y$) if and only if $\varrho = 0$.

The general bivariate normal distribution is slightly more complicated: X and Y are said to have a ***bivariate normal distribution*** if

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2}Q(x,y)\right), \quad x,\, y \in \mathbb{R}$$

where $\sigma_X,\, \sigma_Y > 0$, $|\varrho| < 1$, and Q is the quadratic function

$$Q(x,y) = \frac{1}{1-\varrho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\varrho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]$$

The bivariate normal random vector $Z = (Z_1, Z_2)^T$ (column vector) can be described in a more transparent way. See it as an ***affine transformation*** $Z = \boldsymbol{\mu} + B\boldsymbol{X}$ of a ***standard normal random vector*** $\boldsymbol{X}$.

Idea: in the 1-dimensional case, if $Z \sim N(\mu, \sigma^2)$, then $Z = \mu + \sigma X$ where $X \sim N(0,1)$.
In the 2-dimensional case, start with $X_1, X_2 \overset{\text{iid}}{\sim} N(0,1)$. Let $\boldsymbol{X} = (X_1, X_2)^T$, and define $Z = \boldsymbol{\mu} + B\boldsymbol{X}$, where $\boldsymbol{\mu}$ is a vector and B is a matrix. Then, Z is said to have a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^T$.
Assuming B is invertible, we have by the transformation formula:

$$f_Z(z) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{|B|} = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-1\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\Sigma = BB^T$, and so $|B| = \sqrt{|\Sigma|}$. If we take

$$B = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2\varrho & \sigma_2\sqrt{1-\varrho^2} \end{pmatrix}$$

then the covariance matrix is ($\Sigma = BB^T$):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

This gives the general bivariate pdf.

Multivariate Normal Distribution

Definition: $Z = (Z_1, \ldots, Z_n)^T$ has a ***multivariate normal distribution*** with parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\Sigma = (\sigma_{ij})$, written $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if its pdf is given by

$$f(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right), \quad \boldsymbol{z} \in \mathbb{R}^n$$

where Σ is a positive-definite symmetric matrix with determinant $|\Sigma|$.
Important: Think of Z as a transformation $\boldsymbol{\mu} + B\boldsymbol{X}$ of a standard multivariate normal random vector, where $BB^T = \Sigma$.

Theorem: Let $Z \sim N(\boldsymbol{\mu}, \Sigma)$. Then $\mathbb{E}(Z) = \boldsymbol{\mu}$, that is, $\mathbb{E}(Z_i) = \mu_i$ with $i = 1, \ldots, n$ and $\Sigma = \mathrm{Cov}(Z)$.
Theorem: Let A be an $m \times n$ matrix with $m \le n$, and define

$\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ by $\boldsymbol{Y} = A\boldsymbol{X}$, where $\boldsymbol{X} = (X_1, \ldots, X_n)^T$. If $\boldsymbol{X} \sim N(0, \Sigma)$, then $\boldsymbol{Y} \sim N(0, A\Sigma A^T)$.

Thus, normality is preserved under linear transformations.
We can take A to be the row vector $\boldsymbol{a}^T = (a_1, \ldots, a_n)$ and see that any linear combination $X = a_1X_1 + \cdots + a_nX_n$ $(= \boldsymbol{a}^T\boldsymbol{X})$ has a $N(\mu, \sigma^2)$ distribution, where $\mu = \boldsymbol{a}^T\boldsymbol{\mu} = \sum_i a_i\mu_i$, and

$$\sigma^2 = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} = \sum_{i,j} a_i a_j \sigma_{ij} = \sum_i a_i^2\sigma_i^2 + 2\sum_{i<j} a_i a_j \sigma_{ij}$$

10 Week 10

Convergence

Let $X_1, X_2 \ldots$ be a sequence of random variables and let X be another random variable. We will examine the ***convergence*** of $\{X_n\}$ to X. There are several ways to give meaning to this statement.

For example, suppose that the $X_1, X_2, \ldots$ are independent and identically distributed (iid) with common mean μ and variance σ^2, and let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ (the sample mean of $X_1, \ldots, X_n$). We have already seen that $\mathbb{E}\left(\overline{X}_n\right) = \mu$ (the same for all n), and $\mathrm{Var}\left(\overline{X}_n\right) = \sigma^2/n$. Therefore, $\mathrm{Var}\left(\overline{X}_n\right) \to 0$ as $n \to \infty$. So, as n gets large, $\overline{X}_n$ becomes ***less random***.

Chebyshev's Inequality

Let a random variable I_A denote the indicator of event A; it takes the value 1 if A occurs and 0 otherwise. Then,
$\mathbb{E}(I_A) = 1 \times \mathbb{P}(A) + 0 \times \mathbb{P}\left(A^C\right) = \mathbb{P}(A)$.
Let X be a random variable and let $h(x)$ be a non-negative function. Fix $a > 0$ and let $A = \{h(X) \geq a\}$. Then, clearly $h(X) \geq aI_A$. Therefore,

$$\mathbb{E}(h(X)) \leq \mathbb{E}(aI_A) = a\mathbb{P}(A) = a\mathbb{P}(h(X) \geq a)$$

This proves the following theorem:

Theorem: If $h : \mathbb{R} \to [0, \infty)$, then

$$\mathbb{P}(h(X) \geq a) \leq \frac{\mathbb{E}(h(X))}{a} \qquad \forall\ a > 0$$

Setting $h(x) = |x|$, we get ***Markov's Inequality:***

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a} \qquad \forall\ a > 0$$

and setting $h(x) = x^2$ and replacing a by a^2, we get ***Chebyshev's Inequality***:

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}\left(X^2\right)}{a^2} \qquad \forall\ a > 0$$

Now, replace X by $X - \mu_X$, where $\mu_X = \mathbb{E}(X)$:

$$\mathbb{P}(|X - \mu_X| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2} \qquad \forall\ a > 0$$

Definition: A sequence $\{X_n\}$ of random variables ***converges in probability*** to a random variable X (written $X_n \xrightarrow{P} X$) if, for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

Theorem (Weak Law of Large Numbers): Suppose $\{X_n\}$ are iid with common finite mean μ and finite variance σ^2. Then,

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mu$$

Note: by "$\xrightarrow{P} \mu$" we mean convergence in probability to a random variable that takes the value μ with probability 1.
In fact, there is a ***Strong Law of Large Numbers***, giving the strongest statement imaginable (woah, Ross!). It says that the event $\{\omega : 1/n\sum_{i=1}^n X_i(\omega) \to \mu\}$ has probability 1.

Theorem: If $\{X_n\}$ are iid, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \to \mu\right) = 1$$

for some constant μ *if and only if* $\mathbb{E}(|X_i|) < \infty$, in which case $\mu = \mathbb{E}(X_i)$.

Definition: Let $\{X_n\}$ be a sequence of random variables with distribution functions $F_1, F_2, \ldots$ and let X be another random variable with cumulative distribution function F. Then, $\{X_n\}$ ***converges in distribution*** to X (written $X_n \xrightarrow{D} X$) if $F_n(x) \to F(x)$, that is $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$, at each point x where F is continuous.
(It doesn't matter what happens at points x where F is not continuous).

The Central Limit Theorem (CLT)

Theorem: Let $\{X_n\}$ be a sequence of iid random variables with common *finite* expectation μ and *strictly positive* and *finite* variance σ^2. Let $Z \sim N(0, 1)$. Then,

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \left(= \frac{\overline{X}_n - \mu}{(\sigma/\sqrt{n})}\right) \xrightarrow{D} Z \qquad \text{as } n \to \infty$$

The Normal Approximation to the Binomial Distribution

If $X_1, X_2, \ldots$ are iid $\mathrm{Bin}(1, p)$ random variables, then $S_n := \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, p)$.
The CLT implies that

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z \quad \text{as } n \to \infty$$

where $Z \sim N(0, 1)$.

The Normal Approximation to the Poisson Distribution

If $X_1, X_2, \ldots$ are iid $\mathrm{Poi}(1, p)$ random variables, then $S_n := \sum_{i=1}^n X_i \sim \mathrm{Poi}(n, p)$.

The CLT implies that

$$\frac{S_n - n}{\sqrt{n}} \xrightarrow{D} Z \quad \text{as } n \to \infty$$

where $Z \sim N(0,1)$.

Markov Chains

Markov chains are important tools for modelling random phenomena. A defining property of Markov chains is that "the future is conditionally independent of the past, given the present". Essentially, the probability of a variable having some value at step n of the Markov chain is only dependent on it's value at step $n-1$. That is, to predict X_n, we need only know X_{n-1}.

Definition: A sequence $\{X_n,\, n = 0, 1, \dots\}$ of random variables is called a ***discrete-time stochastic process***; X_n usually represents the state of the process at *time* n. If $\{X_n\}$ takes values in a discrete space S, then it is called a ***Markov chain*** if

$$\begin{aligned}\mathbb{P}\left(X_{m+1} = j \mid X_m = i,\, X_{m-1} = i_{m-1}, \dots, X_0 = i_0\right) \\ = \mathbb{P}\left(X_{m+1} = j \mid X_m = i\right)\end{aligned}$$

for all time points m and all states $i_0, \dots, i_{m-1}$, with $i, j \in S$.

We restrict ourselves to Markov chains for which

$$P_{ij} = \mathbb{P}\left(X_{m+1} = j \mid X_m = i\right) \quad i, j \in S$$

does not depend on time m for which the state space S is discrete (countable).
We can arrange these one-step transition probabilities in a (one-step) ***transition matrix*** of X, usually denoted by P. For example, when $S = \{0, 1, 2, \dots\}$, the transition matrix P has the form

$$\begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Note that each row sums to one, e.g.
$p_{00} + p_{01} + p_{02} + \cdots + p_{0n} = 1$.

Can we calculate the chance of being in *each* of the various states after n steps? (Yes!)
As we will only consider time-homogeneous chains, we write

$$\begin{aligned}p_{ij}^{(n)} &= \mathbb{P}\left(X_{m+n} = j \mid X_m = i\right) \\ &= \mathbb{P}\left(X_n = j \mid X_0 = i\right)\end{aligned}$$

which are the ***n-step transition probabilities***. Recall that the *1-step* transition probabilities (or simply ***transition probabilities*** are

$$\begin{aligned}p_{ij} \coloneqq p_{ij}^{(1)} &= \mathbb{P}\left(X_{m+1} = j \mid X_m = i\right) \\ &= \mathbb{P}\left(X_1 = j \mid X_0 = i\right)\end{aligned}$$

Note that

$$\sum_{j \in S} p_{ij}^{(n)} = \sum_{j \in S} \mathbb{P}\left(X_n = j \mid X_0 = i\right) = 1$$

and in particular that $\sum_{j \in S} p_{ij} = 1$.
The matrix $P^{(n)} = (p_{ij}^{(n)},\ i, j \in S)$ is called the ***n-step transition matrix***.
By the law of total probability, we have

$$\begin{aligned}p_{ij}^{(n+m)} &= \mathbb{P}\left(X_{n+m} = j \mid X_0 = i\right) \\ &= \sum_{k \in S} \mathbb{P}\left(X_{n+m} = j \mid X_n = k,\ X_0 = i\right) \mathbb{P}\left(X_n = k \mid X_0 = i\right)\end{aligned}$$

But,
$$\begin{aligned}&\mathbb{P}\left(X_{n+m} = j \mid X_n = k,\ X_0 = i\right) \\ &\quad = \mathbb{P}\left(X_{n+m} = j \mid X_n = k\right) && \text{(Markov Property)} \\ &\quad = \mathbb{P}\left(X_m = j \mid X_0 = k\right) && \text{(Time homogeneous)} \\ &\quad = p_{kj}^{(m)}\end{aligned}$$

and so, for all $m,\, n \geq 1$

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)} \qquad i, j \in S$$

or equivalently, in terms of transition matrices: $P^{(n+m)} = P^{(n)} P^{(m)}$. Thus, in particular, we have $P^{(n)} = P^{(n-1)} P$ (remembering that $P \coloneqq P^{(1)}$). Therefore,

$$\boxed{P^{(n)} = P^n, \quad n \geq 1}$$

Note that since $P^{(0)} = I = P^0$, this expression is also valid when $n = 0$.

If a process ends with some condition, the probability that the process lasts more than n steps is $\mathbb{P}\left(X_n \neq 0\right) = 1 - \mathbb{P}\left(X_n = 0\right)$.

Arbitrary Initial Conditions

What if we are unsure where the process starts?
Let $\pi_j^{(n)} = \mathbb{P}\left(X_n = j\right)$ and define a row vector

$$\boldsymbol{\pi}^{(n)}(\pi_j^{(n)},\ j \in S)$$

giving the distribution of the state at time n.
Suppose that we know the *initial distribution* $\boldsymbol{\pi}^{(0)}$, that is, the distribution of X_0. By the law of total probability, we have

$$\begin{aligned}\pi_j^{(n)} = \mathbb{P}\left(X_n = j\right) &= \sum_{i \in S} \mathbb{P}\left(X_n = j \mid X_0 = i\right) \mathbb{P}\left(X_0 = i\right) \\ &= \sum_{i \in S} p_{ij}^{(n)} \pi_i^{(0)}\end{aligned}$$

and so

$$\boxed{\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} P^n \qquad n \geq 0}$$

Note that the join distribution of the $\{X_n\}$ are completely specified by the initial distribution and the one-step transition probabilities. Namely, by the product rule and the Markov property:

$$\begin{aligned}
&\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) \\
&= \mathbb{P}(X_0 = x_0) \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \ldots \\
&\qquad \cdot \mathbb{P}(X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\
&= \mathbb{P}(X_0 = x_0) \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \ldots \\
&\qquad \cdot \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})
\end{aligned}$$

A convenient way to describe a Markov chain $\{X_n\}$ is through its ***transition graph***. States are indicated by the nodes of the graph, and a strictly positive (> 0) transition probability p_{ij} from state i to j is indicated by an arrow from i to j with weight p_{ij}.

As n grows, $P^n \to P^\infty$ for some matrix P^∞ that has equal rows. The interpretation of $P^\infty(i, j)$ is the probability of being in j "far away in the future", starting from i. This probability does not depend on i; if we run a Markov chain for a very large number of iterations, it doesn't really matter where we started from.

11 Week 11

Simulating Markov Chains

Here is how to simulate a Markov chain $X_0, X_1, \ldots, X_n$ with initial distribution $\boldsymbol{\pi}^{(0)}$ and transition matrix P:

Algorithm 1 (Simulating a Markov Chain):

1. Draw X_0 from the initial distribution $\boldsymbol{\pi}^{(0)}$. Set $n = 0$.
2. Draw X_{n+1} from the distribution corresponding to the X_n-th row of P.
3. Set $n = n + 1$ and go to step 2.

Simulating a Random Walk on the Integers

Let p be a number between 0 and 1. A random walk on $\mathbb{Z}$, with parameter p, is the Markov chain $\{X_n\}$ with state space $\mathbb{Z}$ and transition matrix P, given by

$$p_{i,i+1} = p; \quad p_{i,i-1} = q = 1 - p \quad \forall\, i \in \mathbb{Z}$$

The transition graph of this is given by

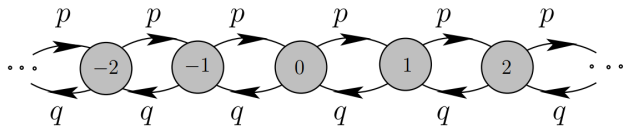

Figure 1: Transition Graph for the Random Walk

Limiting Distribution

A Markov chain exhibits "limiting behaviour" if the n-step transition matrix P^n converges to a matrix P^∞ as $n \to \infty$. Moreover, this P^∞ has all rows equal to some row vector $\boldsymbol{\pi}$. For general Markov chains (satisfying some mild conditions) it holds that

$$\lim_{n\to\infty} P^n(i, j) = \pi_j$$

for some number $0 \le \pi_j \le 1$. When the $\{\pi_j\}$ sum up to 1, they form the ***limiting distribution*** of the Markov chain.

Limiting Distribution (Two States)

Let's look at an example of a two-state chain. Let $S = \{0, 1\}$, and let

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

where $p, q \in (0, 1)$. It can be shown that

$$P = \frac{1}{p+q} \begin{pmatrix} 1 & p \\ 1 & -q \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix} \begin{pmatrix} q & p \\ 1 & -1 \end{pmatrix}$$

where $r = 1 - p - q$. This is of the from $P = VDV^{-1}$ in a procedure called ***diagonalization***. This is good news because

$$\begin{aligned}
P^2 &= (VDV^{-1})(VDV^{-1}) = VD(V^{-1}V)DV^{-1} \\
&= V(DID)V^{-1} = VD^2V^{-1}
\end{aligned}$$

Similarly, $P^n = VD^nV^{-1}$ for all $n \ge 1$. Hence,

$$\begin{aligned}
P^n &= \frac{1}{p+q} \begin{pmatrix} 1 & p \\ 1 & -q \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & r^n \end{pmatrix} \begin{pmatrix} q & p \\ 1 & -1 \end{pmatrix} \\
&= \frac{1}{p+q} \begin{pmatrix} q + pr^n & p - pr^n \\ q - qr^n & p + qr^n \end{pmatrix}
\end{aligned}$$

For a general Markov chain $\{X_n\}$, let $\pi_j^{(n)} = \mathbb{P}(X_n = j)$, and let $\boldsymbol{\pi}^{(n)} = (\pi_j^{(n)},\ j \in S) = \boldsymbol{\pi}^{(0)} P^n$.
If we can find a vector $\boldsymbol{\pi} = (\pi_j,\ j \in S)$ that satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}P$, then $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$ will imply that $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$ for all $n \ge 0$. Such a distribution is called a ***stationary distribution***.

Also, if $\lim_{n\to\infty} \boldsymbol{\pi}^{(n)}$ exists and equals $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ is called a ***limiting distribution***. Since we also have $\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)} P$ (from $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} P^n$), then letting $n \to \infty$ shows that $\boldsymbol{\pi} = \boldsymbol{\pi}P$. That is, if a limiting distribution exists, it is a stationary distribution.

If, for a general Markov chain, a limiting distribution $\boldsymbol{\pi}$ exists, then it is a stationary distribution ($\boldsymbol{\pi}P = \boldsymbol{\pi}$).

Limiting Distribution (General)

Theorem: The limiting distribution $\boldsymbol{\pi}$, if it exists, is uniquely determined by the solution of

$$\boldsymbol{\pi} = \boldsymbol{\pi}P \tag{2}$$

with $\pi_{ij} \ge 0$ and $\sum_j \pi_j = 1$. Conversely, if there exists a unique positive row vector $\boldsymbol{\pi}$ satisfying equation (2) and

summing up to 1, then $\boldsymbol{\pi}$ is the limiting distribution of the Markov chain.

Equation (2) can be rewritten as a system of equations:

$$\sum_j \pi_i p_{ij} = \sum_j \pi_j p_{ij} \qquad \forall\ i \in \mathcal{E} \tag{3}$$

where $\mathcal{E} = \{0, 1, 2, \dots\}$. These are called the ***global balance equations***.

To find the limiting distribution $\boldsymbol{\pi}$, we need to solve the equation (2), or equivalently

$$\boldsymbol{\pi}(P - I) = 0$$

which in turn is equivalent to

$$(P^T - I)\boldsymbol{\pi}^T = 0^T$$

where ^{T} denotes transposition. In other words, the column vector $\boldsymbol{\pi}^T$ lies in the null-space of the matrix $P^T - I$.

Random Walk on the Positive Integers

Let X be a random walk on $\mathcal{E} = \{0, 1, 2, \dots\}$ with transition matrix

$$P = \begin{pmatrix} q & p & 0 & \dots & & \\ q & 0 & p & 0 & \dots & \\ 0 & q & 0 & p & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

where $0 < p < 1$ and $q = 1 - p$.
Note that all states can be reaches from each other. The equation $\boldsymbol{\pi} = \boldsymbol{\pi}P$ becomes

$$\begin{aligned} \pi_0 &= q\pi_0 + q\pi_1 \\ \pi_1 &= p\pi_0 + q\pi_2 \\ \pi_2 &= p\pi_1 + q\pi_3 \\ \pi_3 &= p\pi_2 + q\pi_4 \end{aligned}$$

we can solve this equation sequentially. If we let $r = p/q$, we can express $\pi_1, \pi_2, \dots$ in terms of π_0 and r as

$$\pi_j = r^j \pi_0, \qquad j = 0, 1, 2, \dots$$

If $p < q$, then $r < 1$ and $\sum_{j=0}^{\infty} \pi_j = \frac{1}{1-r}\pi_0$, and by choosing $\pi_0 = 1 - r$, we can make the sum $\sum \pi_j = 1$.
Hence, for $r < 1$, we have found the limiting distribution $\boldsymbol{\pi} = (1 - r)(1, r, r^2, \dots)$ for this Markov chain.
On the other hand, when $p \geq q$, then $\sum \pi_j$ is either 0 or infinite, and hence the limiting distribution does not exist.