

Appendix 3

Malcolm

Started 5 Nov 2024

Contents

A	Probability	3
A.1	Fundamental concepts	3
A.1.1	Probability Axioms	3
A.1.2	Discrete probability law	4
A.1.3	Some properties of probability laws	4
A.1.4	Properties of conditional probability	5
A.1.5	Total probability theorem	5
A.1.6	Bayes' rule	6
A.1.7	Independence	6
A.1.8	Independence of a collection of events	7
A.1.9	Permutations and Combinations, Binomial Coefficient	8
A.2	Discrete random variables	10
A.2.1	Functions of random variables	10
A.2.2	Expectation and Variance	10
A.2.3	Expected value of a function of a RV	11
A.2.4	Expectation and variance of linear functions	12
A.2.5	Variance in terms of Moments Expression	12
A.2.6	Expectation and Variance of Bernoulli	13
A.2.7	Expectation of Discrete Uniform	13
A.2.8	Variance of Discrete Uniform	15
A.2.9	Joint PMFs of multiple random variables	16
A.2.10	Conditioning	19
A.2.11	Conditional Expectation	22
A.2.12	Independence	24
A.2.13	Expectation of independent variables	25
A.2.14	For independent X, Y , $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$	25
A.2.15	Summary for independent random variables	26
A.2.16	Independence of several random variables	26
A.2.17	Variance of sum of independent random variables	27
A.2.18	Mean and variance of the Poisson	27
A.2.19	Mean and variance of the Binomial	28
A.2.20	Mean and variance of the Geometric	29
A.2.21	Summary of results for some PMFs	30
A.3	General Random Variables	31

A.3.1	Continuous random variables and PDFs	31
A.3.2	Continuous Uniform Random Variable	32
A.3.3	Summary of PDF properties	32
A.3.4	Expectation	33
A.3.5	Mean and variance of the Uniform Random Variable . . .	34
A.3.6	Exponential Random Variable	35
A.3.7	Cumulative distribution functions	37
A.3.8	The geometric and exponential CDFs	39
A.3.9	Normal random variables	40
A.3.10	The Standard Normal RV	41
A.4	Limit Theorems	42
A.4.1	Sample mean	42
A.4.2	Markov Inequality	43
A.4.3	Chebyshev Inequality	45
A.4.4	Weak law of large numbers	46
A.4.5	47
B	Supplementary Notes	48
B.1	The sum of the first n natural numbers is $n(n+1)/2$	48
B.2	Telescoping series	49
B.3	Sum of series of products of consecutive integers	50
B.4	Sum of sequence of squares	51

Appendix A

Probability

A.1 Fundamental concepts

A.1.1 Probability Axioms

Nonnegativity

$$\mathbb{P}(A) \geq 0, \text{ for every event } A.$$

Additivity

If A and B are two disjoint (mutually exclusive) events, then the probability of their union satisfies

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

More generally, if the sample space has an infinite number of elements and A_1, A_2, \dots is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \dots$$

Normalisation

The probability of the entire sample space Ω is equal to 1, that is, $\mathbb{P}(\Omega) = 1$.

A.1.2 Discrete probability law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. That is, the probability of any event $\{s_1, s_2, \dots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbb{P}(\{s_1, s_2, \dots, s_n\}) = \mathbb{P}(s_1) + \mathbb{P}(s_2) + \dots + \mathbb{P}(s_n)$$

Discrete uniform probability law

If the sample space consists of n possible outcomes which are equally likely (all single-element events have the same given probability), then the probability of any event A is given by

$$\mathbb{P}(A) = \frac{\text{number of elements of } A}{n}$$

A.1.3 Some properties of probability laws

Consider a probability law, and let A, B , and C be events.

1. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
4. $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A^c \cap B^c \cap C)$.

Note that the third property can be generalised as follows:

$$\mathbb{P}(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n \mathbb{P}(A_i)$$

which can be shown by recursively applying the property for each element.

A.1.4 Properties of conditional probability

The conditional probability of an event A , given an event B with $\mathbb{P}(B) > 0$, is defined by

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

If the possible outcomes are finitely many and equally likely, then

$$\mathbb{P}(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}$$

Multiplication rule

We have

$$\mathbb{P}(\cap_{i=1}^n A_i) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|\cap_{i=1}^{n-1} A_i)$$

This can be verified by

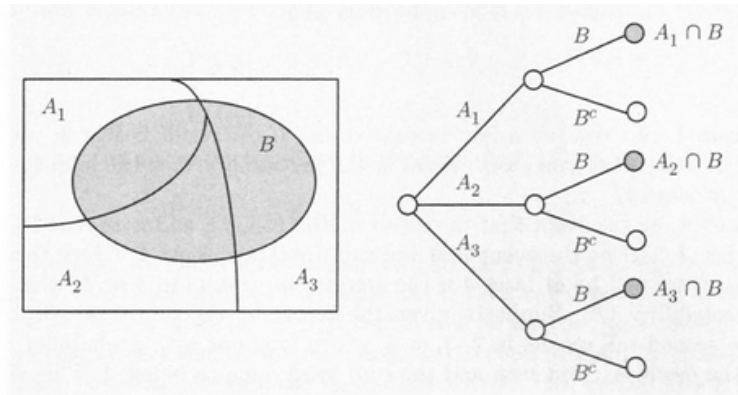
$$\mathbb{P}(\cap_{i=1}^n A_i) = \mathbb{P}(A_1) \cdot \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \cdot \frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(\cap_{i=1}^n A_i)}{\mathbb{P}(\cap_{i=1}^{n-1} A_i)}$$

A.1.5 Total probability theorem

Let A_1, \dots, A_n be disjoint events that form a partition of the sample space and assume that $\mathbb{P}(A_i) > 0$ for all i . Then, for any event B , we have

$$\mathbb{P}(B) = \mathbb{P}(A_1 \cap B) + \cdots + \mathbb{P}(A_n \cap B)$$

Visualised:



A.1.6 Bayes' rule

Let A_1, A_2, \dots, A_n be disjoint events that form a partition of the sample space, and assume that $\mathbb{P}(A_i) > 0$ for all i . Then, for any event B such that $\mathbb{P}(B) > 0$ we have

$$\begin{aligned}\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(A_1)\mathbb{P}(B|A_1) + \dots + \mathbb{P}(A_n)\mathbb{P}(B|A_n)}\end{aligned}$$

A.1.7 Independence

Definition

Two events A and B are said to be independent if

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

If in addition $\mathbb{P}(B) > 0$, independence is equivalent to the condition

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

Complement is also independent

If A and B are independent, so are A and B^c . Intuitively, if $\mathbb{P}(B|A) = \mathbb{P}(B)$:

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

to show the final equality, see that

$$\begin{aligned}\mathbb{P}(B^c|A) + \mathbb{P}(B|A) &= \frac{\mathbb{P}(B^c \cap A)}{\mathbb{P}(A)} + \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(B^c \cap A) + \mathbb{P}(B \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1\end{aligned}$$

(next page)

Conditional independence

Two events A and B are said to be conditionally independent, given another event C with $\mathbb{P}(C) > 0$, if

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

If in addition, $\mathbb{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

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

To derive this alternative characterisation, see

$$\begin{aligned}\mathbb{P}(A \cap B|C) &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} \\ &= \frac{\mathbb{P}(C)\mathbb{P}(B|C)\mathbb{P}(A|B \cap C)}{\mathbb{P}(C)} \\ &= \mathbb{P}(B|C)\mathbb{P}(A|B \cap C)\end{aligned}$$

Compare this with the initial definition and eliminate the common factor $\mathbb{P}(B|C)$ to get what we want.

Note that independence of two events A and B unconditioned does not imply conditional independence, and vice versa.

A.1.8 Independence of a collection of events

We say that the events A_1, A_2, \dots, A_n are independent if

$$\mathbb{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbb{P}(A_i), \quad \text{for every subset } S \text{ of } \{1, 2, \dots, n\}$$

Take the case of three events A_1, A_2 , and A_3 , independence amounts to satisfying the four conditions

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1)\mathbb{P}(A_2), \\ \mathbb{P}(A_1 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_3), \\ \mathbb{P}(A_2 \cap A_3) &= \mathbb{P}(A_2)\mathbb{P}(A_3), \\ \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)\end{aligned}$$

The first three conditions simply assert that any two events are independent; this property is called *pairwise independence*. The fourth condition is also a requirement for independence. Note it is not implied by the first three and vice versa—pairwise independence does not imply independence.

A.1.9 Permutations and Combinations, Binomial Coefficient

***k*-permutations**

Starting with n distinct objects, and letting k be some positive integer where $k \leq n$, consider counting the number of different ways that we can pick k out of these n objects and arrange them into a sequence—the number of distinct k -object sequences.

We first have n choices for the first object. Having chosen the first, there are only $n - 1$ possible choices for the second, $n - 2$ for the third, and so on. This continues until we have chosen $k - 1$ objects, leaving us with $n - (k - 1)$ choices for the last one. The number of possible sequences, called *k-permutations*, can be written as

$$n(n - 1) \cdots (n - k + 1)$$

This can be rewritten, giving us

$$\begin{aligned} n(n - 1) \cdots (n - k + 1) &= \frac{n(n - 1) \cdots (n - k + 1)(n - k) \cdots 2 \cdot 1}{(n - k) \cdots 2 \cdot 1} \\ &= \frac{n!}{(n - k)!} \end{aligned}$$

See that in the special case where $k = n$ we have

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

(This can also be seen from substituting $k = n$ into the formula and recalling the convention $0! = 1$.)

(next page)

Reordering a set

Starting with k objects, consider trying to find how many ways can we order them in a set of k elements. This follows a fairly similar principle to permutation; think of having k ‘slots’ to order k elements in: the first ‘slot’ has k possible inputs, the second $k - 1$ and so on. See that this just gives us $k!$.

Combinations

Combinations can be viewed as counting the number of k -element subsets of a given n -element set. Combinations are different from permutations in that *there is no ordering of selected elements*. For instance, where the 2-permutations of the letters A, B, C, and D are

AB, BA, AC, CA, AD, DA, BC, CB, BD, DB, CD, DC

the *combinations* of two out of these four letters are

AB, AC, AD, BC, BD, CD

See that the ‘duplicates’ are grouped together; for instance AB and BA are not viewed as distinct.

This reasoning can be generalised: each combination is associated with $k!$ ‘duplicate’ k -permutations—all ‘duplicate’ permutations of any given combination is just that permutation reordered for the maximum number of times:

(any single combination of length k) $\cdot k! =$ (permutations of that combination)

The number $n!/(n-k)!$ of k -permutations is equal to the number of combinations times $k!$. Hence the number of possible combinations is equal to

$$\frac{n!}{k!(n-k)!}$$

Binomial Coefficient

Consider a bernoulli process with probability p . We want the probability of k ‘successes’ in n trials. See that the probability of one *specific* sequence of n trials yielding k ‘successes’ would be

$$p^k(1-p)^{n-k}$$

We obtain the desired probability by multiplying this by the number of *combinations* of k ‘successes’ we can obtain in n trials:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

(think tossing a coin three times and obtaining two heads—the heads might occur on the first and third tosses, or other *combinations* of trials).

A.2 Discrete random variables

A.2.1 Functions of random variables

If $Y = g(X)$ is a function of a random variable X , then Y is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value of x for X , and hence also a numerical value $y = g(x)$ for Y .

If X is discrete with PMF p_X , then Y is also discrete, and its PMF p_Y can be calculated using the PMF of X . In particular, to obtain $p_Y(y)$ for any y , we add the probabilities of all values of x such that $g(x) = y$:

$$p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x)$$

A.2.2 Expectation and Variance

Expectation

We define the *expected value* of a random variable X with a PMF p_X by

$$\mathbb{E}[X] = \sum_x xp_X(x)$$

Variance and Standard Deviation

We define the *variance* associated with a random variable X as

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x (X - \mathbb{E}[X])^2 p_X(x)$$

(See that the because of the square the variance is always nonnegative). The variance provides a measure of dispersion of X around the mean. Another measure of dispersion is the *Standard deviation* of X , which is defined as the square root of the variance and is denoted by σ_X :

$$\sigma_X = \sqrt{\text{var}(X)}$$

The standard deviation is often easier to interpret because it has the same units as X .

A.2.3 Expected value of a function of a RV

Expectation of a function

Let X be a RV with PMF p_X , and let $g(X)$ be a function of X . Then the expected value of the random variable $g(X)$ is given by

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

This can be shown, since

$$p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x)$$

we have

$$\begin{aligned}\mathbb{E}[g(X)] &= \mathbb{E}[Y] \\ &= \sum_y yp_Y(y) \\ &= \sum_y y \sum_{\{x|g(x)=y\}} p_X(x) \\ &= \sum_y \sum_{\{x|g(x)=y\}} yp_X(x) \\ &= \sum_y \sum_{\{x|g(x)=y\}} g(x)p_X(x) \\ &= \sum_x g(x)p_X(x)\end{aligned}$$

Variance

Using this we can write the variance of X as

$$\text{var}(X) = \mathbb{E}[(x - \mathbb{E}[X])^2] = \sum_x (x - \mathbb{E}[X])^2 p_X(x)$$

A.2.4 Expectation and variance of linear functions

We show for a random variable X , and letting $Y = aX + b$:

$$\boxed{\mathbb{E}[Y] = a\mathbb{E}[X] + b, \quad \text{var}(Y) = a^2\text{var}(X)}$$

Linearity of Expectations:

$$\mathbb{E}[Y] = \sum_x (ax + b)p_X(x) = a \underbrace{\sum_x xp_x(x)}_{=\mathbb{E}[X]} + b \underbrace{\sum_x p_x(x)}_{=1} = a\mathbb{E}[X] + b$$

Variance:

$$\begin{aligned} \text{var}(Y) &= \sum_x (ax + b - \mathbb{E}[aX + b])^2 p_X(x) \\ &= \sum_x (ax + b - a\mathbb{E}[X] + b)^2 p_X(x) \\ &= a^2 \sum_x (x - \mathbb{E}[X])^2 p_X(x) \\ &= a^2 \text{var}(X) \end{aligned}$$

Note that unless $g(X)$ is a linear function, it is not generally true that $\mathbb{E}[g(X)]$ is equal to $g(\mathbb{E}[X])$.

A.2.5 Variance in terms of Moments Expression

We show

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

see that

$$\begin{aligned} \text{var}(X) &= \sum_x (x - \mathbb{E}[X])^2 p_X(x) \\ &= \sum_x (x^2 - 2x\mathbb{E}[X] + (\mathbb{E}[X])^2) p_X(x) \\ &= \sum_x x^2 p_X(x) - 2\mathbb{E}[X] \sum_x x p_X(x) + (\mathbb{E}[X])^2 \sum_x p_X(x) \\ &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

A.2.6 Expectation and Variance of Bernoulli

Consider a Bernoulli RV X with PMF

$$p_X(k) = \begin{cases} p, & \text{if } k = 1. \\ 1 - p, & \text{if } k = 0. \end{cases}$$

The mean, second moment, and variance of X are as follows:

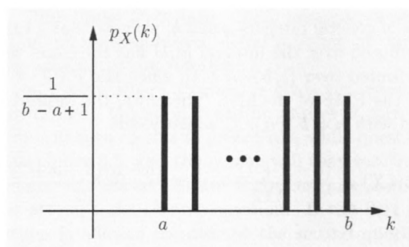
$$\begin{aligned} \mathbb{E}[X] &= 1 \cdot p + 0 \cdot (1 - p) = p \\ \mathbb{E}[X^2] &= 1^2 \cdot p + 0 \cdot (1 - p) = p \\ \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p) \end{aligned}$$

A.2.7 Expectation of Discrete Uniform

Consider a Discrete Uniform RV X with PMF, for $k \in [a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b \\ 0, & \text{otherwise.} \end{cases}$$

An illustration is useful here:



Expectation

Upon inspection one might suppose that the expectation is

$$\mathbb{E}[X] = \frac{a+b}{2}$$

(next page)

Expectation (cont.)

The formula can be elucidated from the definition of the expectation. First see that a sequence $\sum_{k=a}^b k$ can be written as

$$\begin{aligned}\sum_{k=a}^b k &= \sum_{k=1}^b k - \sum_{k=1}^{a-1} k \\ &= \frac{(b)(b+1)}{2} - \frac{(a-1)(a)}{2} \quad (\text{see B.1}) \\ &= \frac{b^2 + b - a^2 + a}{2} = \frac{(b-a+1)(a+b)}{2}\end{aligned}$$

The last step isn't easy to factor, but working back from our 'hypothesis' for the expectation it coincides.

so now we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=a}^b k \left(\frac{1}{b-a+1} \right) \\ &= \frac{1}{b-a+1} \sum_{k=a}^b k \\ &= \frac{1}{b-a+1} \cdot \frac{(b-a+1)(a+b)}{2} \\ \mathbb{E}[X] &= \frac{(a+b)}{2}\end{aligned}$$

A.2.8 Variance of Discrete Uniform

Case for $k \in [1, n]$:

We can obtain the second moment for a discrete uniform distributed over $k \in [1, n]$ as

$$\begin{aligned}
 \mathbb{E}[X^2] &= \sum_{k=1}^n k^2 \left(\frac{1}{n} \right) \\
 &= \frac{1}{n} \sum_{k=1}^n k^2 \\
 &= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} \quad (\text{see B.4}) \\
 &= \frac{(n+1)(2n+1)}{6}
 \end{aligned}$$

We then use the formula for variance in terms of moments expression:

$$\begin{aligned}
 \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2} \right)^2 \\
 &= \frac{1}{12} (n+1)(4n+2-3n-3) \\
 &= \frac{n^2-1}{12}
 \end{aligned}$$

General case $k \in [a, b]$:

For the general case, note that a RV uniformly distributed over an interval $[a, b]$ has the *same variance* as one which is uniformly distributed over $[1, b-a+1]$ —the PMF of the second is just a shifted version of the PMF of the first.

Therefore, the desired variance is given by the first case, but instead with $n = b-a+1$, yielding

$$\boxed{\text{var}(X) = \frac{(b-a+1)^2-1}{12} = \frac{(b-a)(b-a+2)}{12}}$$

A.2.9 Joint PMFs of multiple random variables

We extend the concepts of PMF to multiple variables. Consider two discrete random variables X and Y associated with the same experiment. The probabilities of the values that X and Y can take are captured by the *joint PMF* of X and Y , denoted p_{XY} .

If (x, y) is a pair of possible values of X and Y , the probability mass of (x, y) is the probability of the event $\{X = x, Y = y\}$:

$$p_{XY}(x, y) = \mathbf{P}(X = x, Y = y)$$

We use the abbreviated notation $\mathbf{P}(X = x, Y = y)$ instead of the more precise notations $\mathbf{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbf{P}(X = x \text{ and } Y = y)$.

If A is the set of all pairs (x, y) that have a certain property, then

$$\mathbf{P}((X, Y) \in A) = \sum_{(x, y) \in A} p_{X, Y}(x, y)$$

In fact we can calculate the PMFs of X and Y using the formulas

$$p_X(x) = \sum_y p_{X, Y}(x, y), \quad p_Y(y) = \sum_x p_{X, Y}(x, y)$$

This comes from the total probability theorem

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_y \mathbf{P}(X = x, Y = y) \\ &= \sum_y p_{X, Y}(x, y) \end{aligned}$$

Where the second equality follows by noting that the event $\{X = x\}$ is the union of the disjoint events $\{X = x, Y = y\}$ as y ranges over all the different values of Y . The formula for $p_Y(y)$ is verified similarly. We may refer to p_X and p_Y as the *marginal* PMFs. Illustration on next page.
(next page)

Cont.

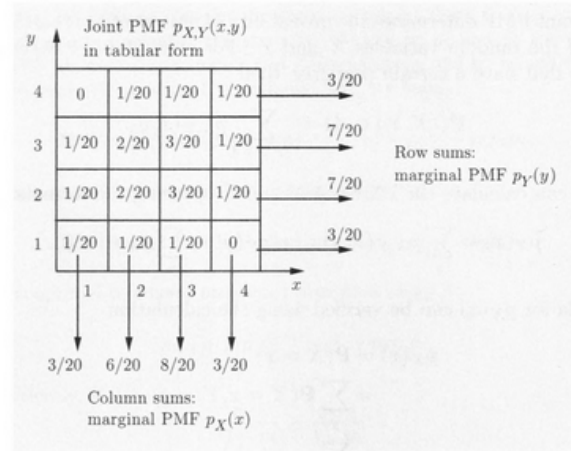


Figure 2.10: Illustration of the tabular method for calculating the marginal PMFs from the joint PMF in Example 2.9. The joint PMF is represented by the table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x , we add the numbers in the column corresponding to x . For example $p_X(2) = 6/20$. Similarly, to calculate the marginal PMF $p_Y(y)$ for a given value of y , we add the numbers in the row corresponding to y . For example $p_Y(2) = 7/20$.

Functions of Multiple Random Variables

A function $Z = g(X, Y)$ of the random variables X and Y defines another random variable. Its PMF can be calculated from the joint PMF $p_{X,Y}$ according to

$$p_Z(z) = \sum_{\{(x,y)|g(x,y)=z\}} p_{X,Y}(x,y)$$

Furthermore, the expected value rule for functions naturally extends and takes the form

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

Where the verification for this is very similar to the earlier case for a function of a single random variable. In the special case where g is linear and of the form $aX + bY + c$, where a, b , and c are given scalars, we have

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

(next page)

More random variables

The joint PMF of three random variables X, Y, Z is defined analogously as

$$p_{X,Y,Z}(x, y, z) = \mathbf{P}(X = x, Y = y, Z = z)$$

for all possible triplets of (x, y, z) . The corresponding marginal PMFs are analogously obtained using

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z)$$

and

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z)$$

The expected value rule for functions is given by

$$\mathbf{E}[g(X, Y, Z)] = \sum_x \sum_y \sum_z g(x, y, z) p_{X,Y,Z}(x, y, z)$$

where if g is linear and has the form $aX + bY + cZ + d$, then

$$\mathbf{E}[aX + bY + cZ + d] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c\mathbf{E}[Z] + d$$

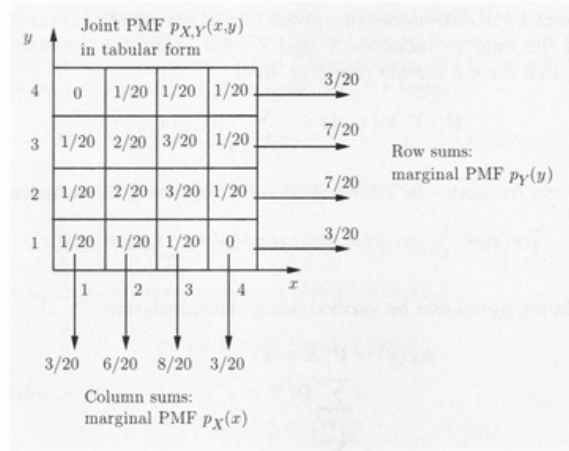


Figure 2.10: Illustration of the tabular method for calculating the marginal PMFs from the joint PMF in Example 2.9. The joint PMF is represented by the table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x , we add the numbers in the column corresponding to x . For example $p_X(2) = 6/20$. Similarly, to calculate the marginal PMF $p_Y(y)$ for a given value of y , we add the numbers in the row corresponding to y . For example $p_Y(2) = 7/20$.

A.2.10 Conditioning

A *conditional PMF* of a random variable X , conditioned on a particular event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x|A) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}$$

By the total probability theorem we have

$$\mathbf{P}(A) = \sum_x \mathbf{P}(\{X = x\} \cap A)$$

Combining the two formulas, see that

$$\sum_x p_{X|A}(x) = 1$$

which reinforces the idea that $p_{X|A}(x)$ is a PMF.

Conditioning one random variable on another

Let X and Y be two random variables associated with the same experiment. If we know that Y is some particular y with some nonzero probability, this provides partial knowledge about the value of X .

This knowledge is captured by the *conditional* PMF $p_{X|Y}$ of X given Y , which is defined by specialising the definition of $p_{X|A}$ to events A of the form $\{Y = y\}$:

$$p_{X|Y}(x|y) = \mathbf{P}(X = x|Y = y)$$

By the definition of conditional probabilities, we have

$$p_{X|Y}(x|y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Intuitively, let us fix some y with $p_Y(y) > 0$, and consider $p_{X|Y}(x|y)$ as a function of x . This gives us a PMF for that specific y , with

$$\sum_x p_{X|Y}(x|y) = 1$$

which can be verified in a similar manner to earlier. Visualisation on next page. (next page)

Cont.

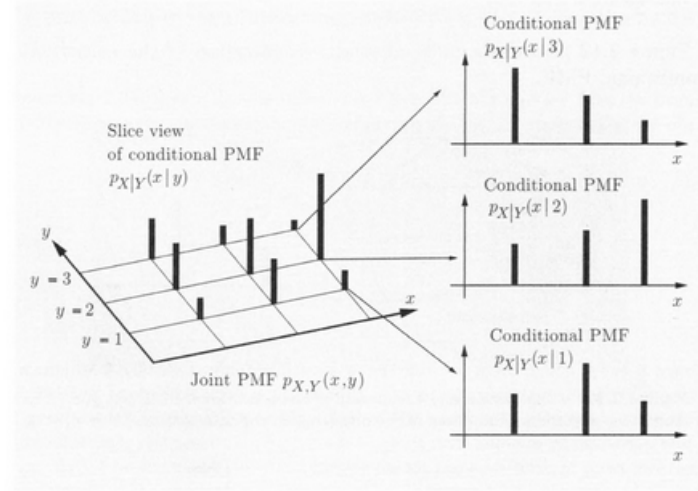


Figure 2.13: Visualization of the conditional PMF $p_{X|Y}(x|y)$. For each y , we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x|y) = 1.$$

The conditional PMF is often convenient for the calculation of the joint PMF, using a sequential approach and the formulas

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y)$$

or

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)$$

See that the conditional PMF can then also be used to calculate the marginal PMFs:

$$p_X(x) = \sum_y p_{X,Y}(x,y) = \sum_y p_Y(y)p_{X|Y}(x|y)$$

We finally note that one can define conditional PMFs involving more than two random variables such as $p_{X,Y|Z}(x,y|z)$ or $p_{X|Y,Z}(x|y,z)$. The concepts and methods described above generalise easily.

(next page)

Cont.

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but pertain to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x | A)$$

and satisfies

$$\sum_x p_{X|A}(x) = 1.$$

- If A_1, \dots, A_n are disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i , then

$$p_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) p_{X|A_i}(x).$$

(This is a special case of the total probability theorem.) Furthermore, for any event B , with $\mathbf{P}(A_i \cap B) > 0$ for all i , we have

$$p_{X|B}(x) = \sum_{i=1}^n \mathbf{P}(A_i | B) p_{X|A_i \cap B}(x).$$

- The conditional PMF of X given $Y = y$ is related to the joint PMF by

$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x | y).$$

- The conditional PMF of X given Y can be used to calculate the marginal PMF of X through the formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x | y).$$

- There are natural extensions of the above involving more than two random variables.

To justify the second part of the third point:

$$\begin{aligned} p_{X|B}(x) &= \frac{\mathbf{P}(\{X = x\} \cap B)}{\mathbf{P}(B)} \\ &= \frac{\sum_{i=1}^n \mathbf{P}(\{X = x\} \cap B \cap A_i)}{\mathbf{P}(B)} \\ &= \frac{\sum_{i=1}^n \mathbf{P}(A_i \cap B) \mathbf{P}(\{X = x\} | A_i \cap B)}{\mathbf{P}(B)} \\ &= \frac{\sum_{i=1}^n \mathbf{P}(B) \mathbf{P}(A_i | B) \mathbf{P}(\{X = x\} | A_i \cap B)}{\mathbf{P}(B)} \end{aligned}$$

A.2.11 Conditional Expectation

We define the conditional expectation and describe a few of its properties:

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$\mathbf{E}[X | A] = \sum_x x p_{X|A}(x).$$

For a function $g(X)$, we have

$$\mathbf{E}[g(X) | A] = \sum_x g(x) p_{X|A}(x).$$

- The conditional expectation of X given a value y of Y is defined by

$$\mathbf{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

- If A_1, \dots, A_n be disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i , then

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i].$$

Furthermore, for any event B with $\mathbf{P}(A_i \cap B) > 0$ for all i , we have

$$\mathbf{E}[X | B] = \sum_{i=1}^n \mathbf{P}(A_i | B) \mathbf{E}[X | A_i \cap B].$$

- We have

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X | Y = y].$$

(next page)

Cont

The last three equalities apply in different contexts, but are essentially equivalent; we refer to them as the *total expectation theorem*. To verify the first of the three, since we have

$$p_X(x) = \sum_{i=1}^n p_{X,A}(x, A_i) = \sum_{i=1}^n P(A_i) p_{X|A_i}(x|A_i)$$

We can express the expectation as

$$\begin{aligned} E[X] &= \sum_x x p_X(x) \\ &= \sum_x x \sum_{i=1}^n P(A_i) p_{X|A_i}(x|A_i) \\ &= \sum_{i=1}^n P(A_i) \sum_x x p_{X|A_i}(x|A_i) \\ &= \sum_{i=1}^n P(A_i) E[X|A_i] \end{aligned}$$

The remaining two equalities are verified similarly.

A.2.12 Independence

We say that the random variable X is independent of the event A if

$$P(X = x \text{ and } A) = P(X = x)P(A) = p_X(x)P(A) \quad \text{for all } x$$

From the definition of the conditional PMF, we have

$$P(X = x \text{ and } A) = p_{X|A}(x)P(A)$$

So that as long as $P(A) > 0$, independence is the same as the condition

$$p_{X|A}(x) = p_X(x) \quad \text{for all } x$$

Two random variables

We say that two random variables X and Y are independent if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \text{for all } x, y$$

This is the same as requiring that the two events $\{X = x\}$ and $\{Y = y\}$ be independent for every x and y . The formula

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$$

Leads us to

$$p_{X|Y}(x|y) = p_X(x) \quad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x$$

Conditional independence

Random variables X and Y are said to be conditionally independent given a positive probability event A , if

$$P(X = x, Y = y|A) = P(X = x|A)P(Y = y|A) \quad \text{for all } x, y$$

or in a more compact notation

$$p_{X,Y|A}(x, y) = p_{X|A}(x)p_{Y|A}(y) \quad \text{for all } x, y$$

in a manner analogous to the previous ideas, this is equivalent to

$$p_{X|Y,A}(x|y) = p_{X|A}(x) \quad \text{for all } x, y \text{ such that } p_{Y|A}(y) > 0$$

As mentioned previously, note that conditional independence may not imply unconditional independence and vice versa.

A.2.13 Expectation of independent variables

If X and Y are independent random variables, then

$$E[XY] = E[X]E[Y]$$

this can be shown:

$$\begin{aligned} E[XY] &= \sum_x \sum_y xy p_{XY}(x, y) \\ &= \sum_x \sum_y xy p_X(x) p_Y(y) \quad (\text{independence}) \\ &= \sum_x x p_X(x) \sum_y y p_Y(y) \\ &= E[X]E[Y] \end{aligned}$$

It is proven that if X and Y are independent, then the same is true for $g(X)$ and $h(Y)$ (not proven here). Given this, it is clear that if X and Y are independent, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

A.2.14 For independent X, Y , $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

Consider now the sum $X + Y$ of two independent variables X and Y , and let us calculate its variance. Since the variance of a random variable is unchanged when it is shifted by a constant, it is convenient to work with the zero-mean random variables $\tilde{X} = X - E[X]$ and $\tilde{Y} = Y - E[Y]$. We have

$$\begin{aligned} \text{var}(X + Y) &= \text{var}(\tilde{X} + \tilde{Y}) \\ &= E[(\tilde{X} + \tilde{Y})^2] \\ &= E[\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2] \\ &= E[\tilde{X}^2] + 2E[\tilde{X}\tilde{Y}] + E[\tilde{Y}^2] \\ &= E[\tilde{X}^2] + E[\tilde{Y}^2] \\ &= \text{var}(\tilde{X}) + \text{var}(\tilde{Y}) \\ &= \text{var}(X) + \text{var}(Y) \end{aligned}$$

Where the property $E[\tilde{X}\tilde{Y}] = 0$ is justified by the idea that functions of independent random variables are independent, so $\tilde{X} = X - E[X]$ and $\tilde{Y} = Y - E[Y]$ are independent. Since they have zero-mean, we obtain

$$E[\tilde{X}\tilde{Y}] = E[\tilde{X}]E[\tilde{Y}] = 0$$

The variance of the sum of two independent random variables is equal to the sum of their variances. This is different from the mean, where the mean of the sum of two random variables is always equal to the sum of their means even if they are not independent.

A.2.15 Summary for independent random variables

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A) > 0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x,$$

that is, if for all x , the events $\{X = x\}$ and A are independent.

- X and Y are independent if for all pairs (x, y) , the events $\{X = x\}$ and $\{Y = y\}$ are independent, or equivalently

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad \text{for all } x, y.$$

- If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X] \mathbf{E}[Y].$$

Furthermore, for any functions g and h , the random variables $g(X)$ and $h(Y)$ are independent, and we have

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \mathbf{E}[h(Y)].$$

- If X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

A.2.16 Independence of several random variables

The preceding discussion extends naturally to the case of more than two random variables; three random variables X, Y and Z are said to be independent if

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z), \quad \text{for all } x, y, z$$

As before (although we didn't prove this), if X, Y, Z are independent random variables, then any three random variables of the form $f(X)$, $g(Y)$, and $h(Z)$ are also independent.

Similarly, any two random variables of the form $g(X, Y)$ and $h(Z)$ are independent. Note however that random variables of the form $g(X, Y)$ and $h(Y, Z)$ are usually not independent because they are both affected by Y .

Properties such as these are intuitively clear; they can be formally verified but this is sometimes tedious. Fortunately there is a general agreement between intuition and what is mathematically correct.

A.2.17 Variance of sum of independent random variables

If X_1, X_2, \dots, X_n are independent random variables, then

$$\text{var}(X_1 + X_2 + \dots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n)$$

This can be verified by repeated use of the formula $\text{var}(X + Y)$ for two independent random variables X and Y .

A.2.18 Mean and variance of the Poisson

The mean of the Poisson PMF

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

can be calculated as follows:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \quad (\text{the } k=0 \text{ term is zero}) \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} \quad (\text{let } m = k-1) \\ &= \lambda \end{aligned}$$

where the last equality is obtained by noting that

$$\sum_{m=0}^{\infty} e^{-\lambda} \frac{\lambda^m}{m!} = \sum_{k=0}^{\infty} p_X(m) = 1$$

Which is the normalisation property for the Poisson PMF. (next page)

Variance of the Poisson

We have

$$\begin{aligned} E[X^2] &= \sum_{k=1}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} \quad (k=0 \text{ term is } 0) \\ &= \lambda \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} (m+1) e^{-\lambda} \frac{\lambda^m}{m!} \\ &= \lambda(E[X] + 1) \\ &= \lambda(\lambda + 1) \end{aligned}$$

from which

$$\text{var}(X) = E[X^2] - (E[X])^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

A.2.19 Mean and variance of the Binomial

For each i , we let X_i be an independent bernoulli random variable. Then $X = X_1 + X_2 + \dots + X_n$ is a binomial random variable. Its mean is $E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = np$.

The variance of single bernoulli random variable is $E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$. By independence of the bernoullis, the variance of the binomial is

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = np(1 - p)$$

A.2.20 Mean and variance of the Geometric

A geometric random variable X has the PMF

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$

The mean and variance are given by

$$E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p, \quad \text{var}(X) = \sum_{k=1}^{\infty} (k - E[X])^2(1-p)^{k-1}p$$

evaluating these sums is somewhat tedious. As an alternative here we will apply the total expectation theorem, with $A_1 = \{X = 1\} = \{\text{first try is a success}\}$, and $A_2 = \{X > 1\} = \{\text{first try is a failure}\}$, and end up with a much simpler calculation.

If the first try is successful, we have $X = 1$, and

$$E[X|X = 1] = 1$$

Intuitively, if the first try fails then ($X > 1$); we have wasted one try and are back to where we started. So the expected number of remaining tries is $E[X]$, and

$$E[X|X > 1] = 1 + E[X]$$

Thus,

$$\begin{aligned} E[X] &= P(X = 1)E[X|X = 1] + P(X > 1)E[X|X > 1] \\ &= p + (1-p)(1 + E[X]) \end{aligned}$$

From which we obtain

$$E[X] = \frac{1}{p}$$

With similar reasoning we also have

$$E[X^2|X = 1] = 1, \quad E[X^2|X > 1] = E[(1 + X)^2] = 1 + 2E[X] + E[X^2]$$

so that

$$E[X^2] = p \cdot 1 + (1-p)(1 + 2E[X] + E[X^2])$$

from which we obtain

$$E[X^2] = \frac{1 + 2(1-p)E[X]}{p}$$

and using our formula $E[X] = 1/p$,

$$E[X^2] = \frac{2}{p^2} - \frac{1}{p}$$

and we conclude that

$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

A.2.21 Summary of results for some PMFs

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)(b-a+1)}{12}.$$

Bernoulli with Parameter p : (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \quad \text{var}(X) = p(1-p).$$

Binomial with Parameters p and n : (Describes the number of successes in n independent Bernoulli trials.)

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

$$\mathbf{E}[X] = np, \quad \text{var}(X) = np(1-p).$$

Geometric with Parameter p : (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ : (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots,$$

$$\mathbf{E}[X] = \lambda, \quad \text{var}(X) = \lambda.$$

A.3 General Random Variables

A.3.1 Continuous random variables and PDFs

A random variable X is *continuous* if there is a nonnegative function f_X called the probability density function of X , or PDF for short, such that

$$P(X \in B) = \int_B f_X(x) dx$$

for every subset B of the real line. In particular, the probability that the value of X falls within an interval is

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

and can be interpreted as the area under a graph of the PDF:

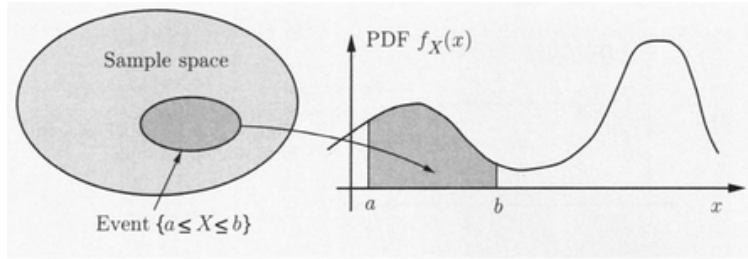


Figure 3.1: Illustration of a PDF. The probability that X takes a value in an interval $[a, b]$ is $\int_a^b f_X(x) dx$, which is the shaded area in the figure.

For any single value a , we have $\int_a^a f_X(x) dx = 0$. For this reason, including or excluding the endpoints of an interval has no effect on its probability:

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

Note that to qualify as a PDF, a function f_X must be nonnegative, meaning $f_X(x) \geq 0$ for every x , and must also have the normalisation property

$$\int_{-\infty}^{\infty} f_X(x) dx = P(-\infty < X < \infty) = 1$$

The entire area under the graph of the PDF must sum to 1.

A.3.2 Continuous Uniform Random Variable

Consider a random variable X that takes values in an interval $[a, b]$, and that any two subintervals of the same length have the same probability. We refer to this type of random variable as *uniform*. Its PDF has the form

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

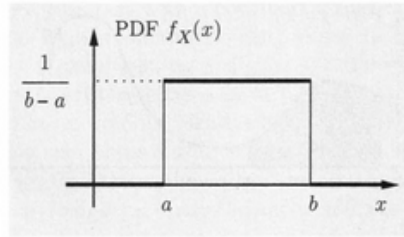


Figure 3.3: The PDF of a uniform random variable.

The constant value of the PDF within $[a, b]$ is determined from the normalisation property. We have

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_a^b \frac{1}{b-a} dx$$

A.3.3 Summary of PDF properties

Summary of PDF Properties

Let X be a continuous random variable with PDF f_X .

- $f_X(x) \geq 0$ for all x .
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
- If δ is very small, then $\mathbf{P}([x, x + \delta]) \approx f_X(x) \cdot \delta$.
- For any subset B of the real line,

$$\mathbf{P}(X \in B) = \int_B f_X(x) dx.$$

A.3.4 Expectation

The *expected value/expectation/mean* of a continuous random variable X is defined by

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

If X is a continuous random variable with a given PDF, any real-valued function $Y = g(X)$ of X is also a random variable. In which case the mean of $g(X)$ satisfies the *expected value rule*

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

In complete analogy with the discrete case. The idea of moments of expectation and variance are also the same as that for the discrete case.

Expectation of a Continuous Random Variable and its Properties

Let X be a continuous random variable with PDF f_X .

- The expectation of X is defined by

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

- The expected value rule for a function $g(X)$ has the form

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- The variance of X is defined by

$$\text{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbf{E}[X])^2 f_X(x) dx.$$

- We have

$$0 \leq \text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- If $Y = aX + b$, where a and b are given scalars, then

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2 \text{var}(X).$$

A.3.5 Mean and variance of the Uniform Random Variable

Consider a uniform PDF over an interval $[a, b]$. We have

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b x \cdot \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \cdot \frac{1}{2} x^2 \Big|_a^b \\ &= \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

To obtain the variance we calculate the second moment. We have

$$\begin{aligned} E[X^2] &= \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \int_a^b x^2 dx \\ &= \frac{1}{b-a} \cdot \frac{1}{3} x^3 \Big|_a^b = \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

Thus the variance is obtained as

$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

A.3.6 Exponential Random Variable

An exponential random variable has a PDF of the form

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

where λ is a positive parameter characterising the PDF.

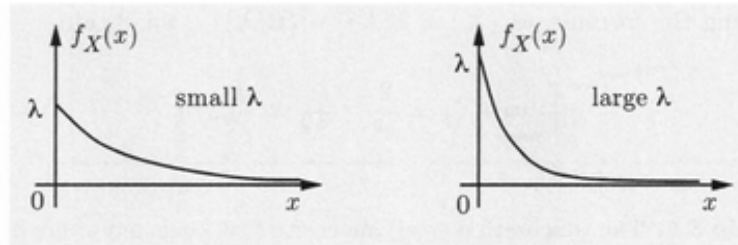


Figure 3.5: The PDF $\lambda e^{-\lambda x}$ of an exponential random variable.

This is a legitimate PDF because

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^{\infty} = 1$$

The probability that X exceeds a certain value decreases exponentially. For any $a \geq 0$, we have

$$P(X \geq a) = \int_a^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_a^{\infty} = e^{-\lambda a}$$

The mean and variance can be calculated as

$$E[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}$$

We verify the mean using integration by parts,

$$\begin{aligned} E[X] &= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\ &= (-x e^{-\lambda x}) \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= 0 - \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

(next page)

Cont.

Again using integration by parts, we have the second moment as

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= (-x^2 e^{-\lambda x}) \Big|_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx \\ &= 0 + \frac{2}{\lambda} E[X] \\ &= \frac{2}{\lambda^2} \end{aligned}$$

Finally, using $\text{var}(X) = E[X^2] - (E[X])^2$, we obtain

$$\text{var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

A.3.7 Cumulative distribution functions

The CDF of a random variable X is denoted F_X and provides the probability $P(X \leq x)$. In particular, for every x we have

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^x f_X(t) dt, & \text{if } X \text{ is continuous.} \end{cases}$$

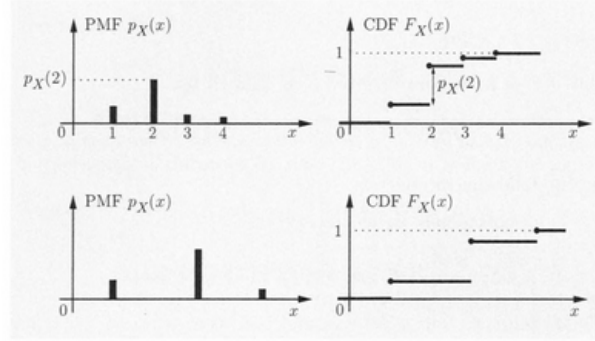


Figure 3.6: CDFs of some discrete random variables. The CDF is related to the PMF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{k \leq x} p_X(k)$$

and has a staircase form, with jumps occurring at the values of positive probability mass. Note that at the points where a jump occurs, the value of F_X is the larger of the two corresponding values (i.e., F_X is continuous from the right).

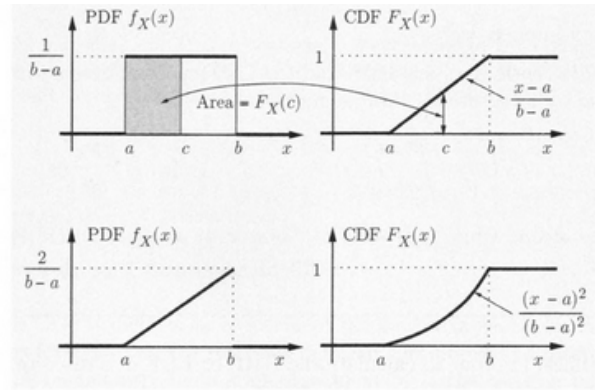


Figure 3.7: CDFs of some continuous random variables. The CDF is related to the PDF through the formula

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Thus, the PDF f_X can be obtained from the CDF by differentiation:

$$f_X(x) = \frac{dF_X}{dx}(x).$$

For a continuous random variable, the CDF has no jumps, i.e., it is continuous.

(next page)

Cont.

Properties of a CDF

The CDF F_X of a random variable X is defined by

$$F_X(x) = \mathbf{P}(X \leq x), \quad \text{for all } x,$$

and has the following properties.

- F_X is monotonically nondecreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y).$$

- $F_X(x)$ tends to 0 as $x \rightarrow -\infty$, and to 1 as $x \rightarrow \infty$.
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x .
- If X is continuous, then $F_X(x)$ is a continuous function of x .
- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^k p_X(i),$$

$$p_X(k) = \mathbf{P}(X \leq k) - \mathbf{P}(X \leq k-1) = F_X(k) - F_X(k-1),$$

for all integers k .

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad f_X(x) = \frac{dF_X}{dx}(x).$$

(The second equality is valid for those x at which the PDF is continuous.)

A.3.8 The geometric and exponential CDFs

Let X be a geometric random variable with parameter p ; that is, X is the number of trials until the first success in a sequence of independent bernoulli trials, where the probability of success at each trial is p . For $k = 1, 2, \dots$, we have $P(X = k) = p(1 - p)^{k-1}$ and the CDF is given by

$$F_{\text{geo}}(n) = \sum_{k=1}^n p(1 - p)^{k-1} = p \frac{1 - (1 - p)^n}{1 - (1 - p)} = 1 - (1 - p)^n, \quad \text{for } n = 1, 2, \dots$$

Suppose now that X is an exponential random variable with parameter $\lambda > 0$. Its CDF is given by

$$F_{\text{exp}}(x) = P(X \leq x) = 0, \quad \text{for } x \leq 0$$

and

$$F_{\text{exp}}(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}, \quad \text{for } x > 0$$

To compare the two CDFs, we can define $\delta = -\ln(1 - p)/\lambda$, so that

$$e^{-\lambda \delta} = 1 - p$$

Then we see that the values of the exponential and the geometric CDFs are equal whenever $x = n\delta$, with $n = 1, 2, \dots$

$$F_{\text{exp}}(n\delta) = F_{\text{geo}}(n) \quad n = 1, 2, \dots$$

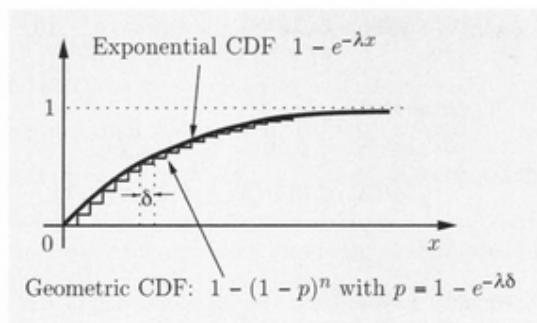


Figure 3.8: Relation of the geometric and the exponential CDFs. We have

$$F_{\text{exp}}(n\delta) = F_{\text{geo}}(n), \quad n = 1, 2, \dots,$$

where δ is chosen so that $e^{-\lambda \delta} = 1 - p$. As δ approaches 0, the exponential random variable can be interpreted as a “limit” of the geometric.

A.3.9 Normal random variables

A continuous random variable X is said to be *normal/gaussian* if it has a PDF of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

where μ and σ are two scalar parameters characterising the PDF, with σ assumed positive. It can be verified that the normalisation property holds:

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1$$

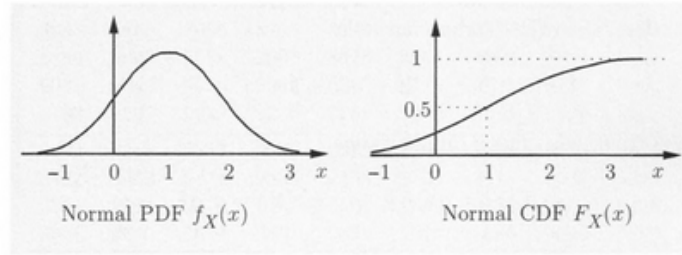


Figure 3.9: A normal PDF and CDF, with $\mu = 1$ and $\sigma^2 = 1$. We observe that the PDF is symmetric around its mean μ , and has a characteristic bell shape. As x gets further from μ , the term $e^{-(x-\mu)^2/2\sigma^2}$ decreases very rapidly. In this figure, the PDF is very close to zero outside the interval $[-1, 3]$.

The mean and variance can be calculated to be

$$E[X] = \mu, \quad \text{var}(X) = \sigma^2$$

To see this, note that the PDF is symmetric around μ , so intuitively the mean can only be μ . The variance is given by

$$\text{var}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx$$

Changing variables $y = (x - \mu)/\sigma$ and integration by parts using $u = y$ and $dv = ye^{-y^2/2}$, we have

$$\begin{aligned} \text{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \sigma^2 \end{aligned}$$

(next page)

Cont.

The last equality was obtained using the fact that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = 1$$

which is just the normalisation property of the normal PDF in the case where $\mu = 0$ and $\sigma = 1$.

Another important property of the normal PDF (proved later):

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2 , and if $a \neq 0$, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbf{E}[Y] = a\mu + b, \quad \text{var}(Y) = a^2\sigma^2.$$

A.3.10 The Standard Normal RV

A.4 Limit Theorems

A.4.1 Sample mean

Definition

Here we discuss asymptomatic behaviour of sequences of random variables. The principal context involves a sequence X_1, X_2, \dots of independent identically distributed random variables with expectation μ and variance σ^2 . We denote

$$S_n = X_1 + \dots + X_n$$

to be the sum of the first n of them. Since they are independent we also have

$$\text{var}(S_n) = \text{var}(X_1) + \dots + \text{var}(X_n) = n\sigma^2$$

See that the distribution of S_n spreads out (it's variance increases) as n increases and doesn't have a meaningful limit. Consider instead the *sample mean*

$$M_n = \frac{X_1 + \dots + X_n}{n} = \frac{S_n}{n}$$

Expectation and Variance

We have the expectation as

$$\begin{aligned}\mathbb{E}[M_n] &= \frac{\mathbb{E}[X_1 + \dots + X_n]}{n} \\ &= \frac{\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]}{n} \\ &= \frac{n\mu}{n} = \mu\end{aligned}$$

and the variance as

$$\text{var}(M_n) = \frac{1}{n^2} \text{var}(S_n) = \frac{\sigma^2}{n}$$

See that the variance of M_n decreases to 0 as n increases.

With this consider a new random variable, that we modify based off M_n and S_n :

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

This has the properties

$$\mathbb{E}[Z_n] = 0, \quad \text{var}(Z_n) = \frac{\text{var}(S_n - n\mu)}{\sigma^2 n} = 1$$

A.4.2 Markov Inequality

Definition

Here we consider the *Markov inequality*. Loosely speaking it asserts that if a *nonnegative* random variable has a small mean, then the probability that it takes a large value must also be small:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}, \quad \text{if } X \geq 0 \text{ and } a > 0.$$

(intuitively, as a increases, the probability that X is greater than it decreases)

Justification

Consider fixing a positive number a and considering the random variable Y_a defined by

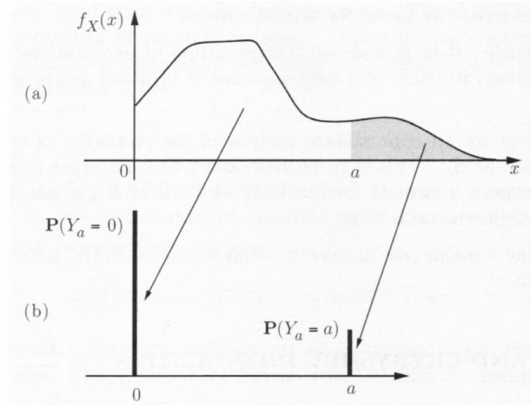
$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

See that the relation

$$Y_a \leq X$$

always holds and therefore

$$\mathbb{E}[Y_a] \leq \mathbb{E}[X]$$



See that all of the probability mass in the PDF of X between 0 and a is assigned to 0, and that above a assigned to a . Since mass is shifted to the left, the expectation can only decrease:

$$\mathbb{E}[X] \geq \mathbb{E}[Y_a] = a\mathbb{P}(Y_a = a) = a\mathbb{P}(X \geq a)$$

from which we obtain

$$a\mathbb{P}(X \geq a) \leq \mathbb{E}[X]$$

(next page)

Another justification

See that if $X \geq 0$ and $a > 0$:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f_X(x) \, dx \geq \int_a^\infty x f_X(x) \, dx \\ &\geq \int_a^\infty a f_X(x) \, dx \\ &= a \mathbb{P}(X \geq a)\end{aligned}$$

so

$$\mathbb{E}[X] \geq a \mathbb{P}(X \geq a)$$

and

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

A.4.3 Chebyshev Inequality

Definition

The *Chebyshev inequality*, loosely speaking, asserts that if a random variable has small variance, then the probability that it takes a value far from its mean is also small: Given a random variable X with mean μ and variance σ^2 ,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}, \quad \text{for all } c > 0$$

Note that the Chebyshev inequality does not require the random variable to be negative.

Justification

Consider the nonnegative random variable $(X - \mu)^2$ and apply the Markov inequality with $a = c^2$ to obtain:

$$\mathbb{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

Now observe that since the event $(X - \mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$, so that

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

The Chebyshev inequality tends to be more powerful than the Markov inequality since it also uses information on the variance of X . An alternative form can also be obtained by letting $c = k\sigma$, $k > 0$, which yields

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

(the probability that a random variable takes a value more than k standard deviations away from its mean is at most $1/k^2$)

Another justification

For a derivation that doesn't use the Markov inequality, introducing the function

$$g(x) = \begin{cases} 0, & \text{if } |x - \mu| < c, \\ c^2, & \text{if } |x - \mu| \geq c \end{cases}$$

since $(x - \mu)^2 \geq g(x)$ for all x we can write

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \geq \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &= c^2 \left(\int_{-\infty}^{\mu-c} f_X(x) dx + \int_{\mu+c}^{\infty} f_X(x) dx \right) \\ &= c^2 \mathbb{P}(|X - \mu| \geq c) \end{aligned}$$

which can be arranged into the desired inequality.

A.4.4 Weak law of large numbers

Justification

The weak law of large numbers asserts that the *sample mean* of a large number of independent identically distributed random variables is very close to the expectation with high probability.

Considering a sequence of X_1, X_2, \dots of independent identically distributed random variables with expectation μ and variance σ^2 , recall the sample mean is defined as

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

We had the expectation as

$$\mathbb{E}[M_n] = \frac{\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu$$

and the variance as

$$\text{var}(M_n) = \frac{1}{n^2} \text{var}(X_1 + \dots + X_n) = \frac{n\text{var}(X)}{n^2} = \frac{\sigma^2}{n}$$

Applying the Chebyshev inequality gives us

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \quad \text{for any } \epsilon > 0$$

We observe that for any fixed $\epsilon > 0$, the right hand side of this equation goes to 0 as n increases.

Definition

This is called the *weak law of large numbers*: Letting X_1, X_2, \dots be independent identically distributed random variables with mean μ , for every $\epsilon > 0$ we have

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) = \mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Intuitively, this means that for large n , the bulk of the distribution of M_n is concentrated near μ . That is, if we consider an interval $[\mu - \epsilon, \mu + \epsilon]$ around μ , then there is a high probability that M_n falls in that interval; as $n \rightarrow \infty$, this probability converges to 1.

A.4.5

Appendix B

Supplementary Notes

B.1 The sum of the first n natural numbers is $n(n+1)/2$

We have that

$$\sum_{i=1}^n i = 1 + 2 + \cdots + n$$

Now consider $2 \sum_{i=1}^n i$:

$$\begin{aligned} 2 \sum_{i=1}^n i &= 2(1 + 2 + \cdots + (n-1) + n) \\ &= (1 + 2 + \cdots + (n-1) + n) + (n + (n-1) + \cdots + 2 + 1) \\ &= (1 + n) + (2 + (n-1)) + \cdots + ((n-1) + 2) + (n + 1) \\ &= (n+1)_1 + (n+1)_2 + \cdots + (n+1)_n \\ &= n(n+1) \end{aligned}$$

so

$$\begin{aligned} 2 \sum_{i=1}^n i &= n(n+1) \\ \sum_{i=1}^n i &= \frac{n(n+1)}{2} \end{aligned}$$

B.2 Telescoping series

Let $\langle b_n \rangle$ be a sequence in \mathbb{R} . Let $\langle a_n \rangle$ be a sequence defined as

$$a_k = b_k - b_{k-1}$$

we show

$$\boxed{\sum_{k=m}^n a_k = b_n - b_{m-1}}$$

See that

$$\begin{aligned} \sum_{k=m}^n a_k &= \sum_{k=m}^n (b_k - b_{k-1}) \\ &= \sum_{k=m}^n b_k - \sum_{k=m}^n b_{k-1} \\ &= \sum_{k=m}^n b_k - \sum_{k=m-1}^{n-1} b_k \\ &= \left(\sum_{k=m}^{n-1} b_k + b_n \right) - \left(b_{m-1} + \sum_{k=m}^{n-1} b_k \right) \\ &= b_n - b_{m-1} \end{aligned}$$

B.3 Sum of series of products of consecutive integers

We show

$$\sum_{j=1}^n j(j+1) = 1 \cdot 2 + 2 \cdot 3 + \cdots + n(n+1) = \frac{n(n+1)(n+2)}{3}$$

See that

$$\begin{aligned} 3i(i+1) &= i(i+1)(i+2) - i(i+1)(i-1) \\ &= (i+1)((i+1)+1)((i+1)-1) - i(i+1)(i-1) \end{aligned}$$

Thus we have the basis of a telescoping series (see (B.2)):

$$3i(i+1) = b(i+1) - b(i)$$

where

$$b(i) = i(i+1)(i-1)$$

So we have

$$\begin{aligned} \sum_{j=1}^n 3j(j+1) &= \sum_{j=1}^n (j+1)((j+1)+1)((j+1)-1) - j(j+1)(j-1) \\ &= n(n+1)(n+2) - 0(0+1)(0-1) \\ &= n(n+1)(n+2) \end{aligned}$$

Thus

$$\sum_{j=1}^n j(j+1) = \frac{n(n+1)(n+2)}{3}$$

B.4 Sum of sequence of squares

We show

$$\forall n \in \mathbb{N} : \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

See that this follows from (B.3):

$$\begin{aligned} \sum_{i=1}^n 3i(i+1) &= n(n+1)(n+2) \\ \sum_{i=1}^n 3i^2 + \sum_{i=1}^n 3i &= n(n+1)(n+2) \\ \sum_{i=1}^n 3i^2 &= n(n+1)(n+2) - 3\frac{n(n+1)}{2} \quad \text{see (B.1)} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(n+2)}{3} - \frac{n(n+1)}{2} \\ &= \frac{2n(n+1)(n+2) - 3n(n+1)}{6} \\ &= \frac{n(n+1)(2n+1)}{6} \end{aligned}$$