Appendix 3

Malcolm

Started 5 Nov 2024

Contents

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 $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

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

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:

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

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:

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

which can be shown be 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

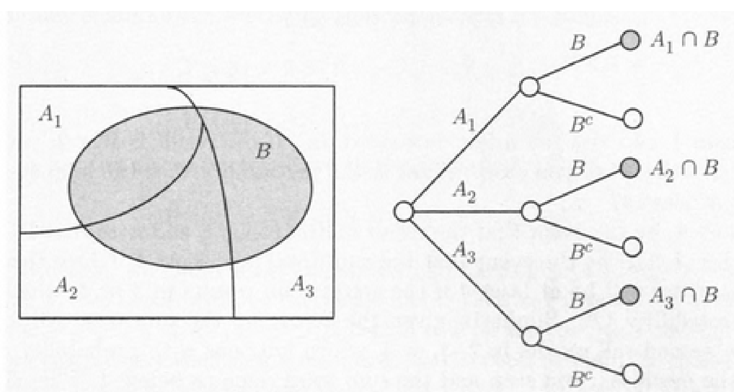
$$\mathbb{P}(\cap_{i=1}^{n} A_i) = \mathbb{P}(A_1) \cdot \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \cdot \frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(\cap_{i=1}^{n} A_i)}{\mathbb{P}(\cap_{i=1}^{n-1} A_i)}$$

A.1.5 Total probability theorem

Let $A_1, \ldots, A_n$ 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 $A_1, A_2, \ldots, A_n$ 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

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

A.1.7 Independence

Definition
Two events $\boldsymbol{A}$ 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 ca 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:

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

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 probility 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.1.10 Expectation and Variance

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

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

Variance and Standard Deviation
We define the *variance* associated with a random variable X as

$$\boxed{\operatorname{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_x (X - \mathbb{E}[X])^2 p_X(x)}$$

(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:

$$\boxed{\sigma_X = \sqrt{\operatorname{var}(X)}}$$

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

A.1.11 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

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

This can be shown, since

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

we have

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

Variance
Using this we can write the variance of X as

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

A.1.12 Expectation and variance of linear functions

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

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

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:

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

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.1.13 Variance in terms of Moments Expression

We show

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

see that

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

A.1.14 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:

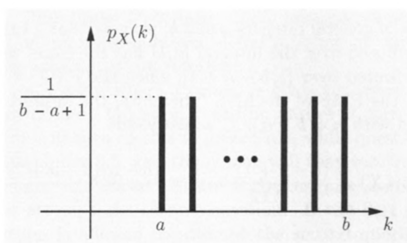
$$\begin{aligned} \mathbb{E}[X] &= 1 \cdot p + 0 \cdot (1-p) = p \\ \mathbb{E}[X^2] &= 1^2 \cdot p + 0 \cdot (1-p) = p \\ \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p) \end{aligned}$$

A.1.15 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.1.16 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

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

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

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

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

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

A.2 Limit Theorems

A.2.1 Sample mean

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

$$S_n = X_1 + \cdots + X_n$$

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

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

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*

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

Expectation and Variance
We have the expectation as

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

and the variance as

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

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

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

A.2.2 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:

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

(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

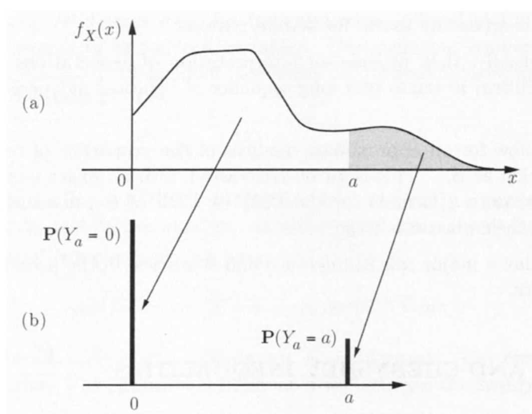
$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

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.2.3 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,

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

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:

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

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 justifcation

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.2.4 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 $X_1, X_2, \ldots$ of independent identically distributed random variables with expectation μ and variance σ^2, recall the sample mean is defined as

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

We had the expectation as

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

and the variance as

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

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 $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ, for every $\epsilon > 0$ we have

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

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 $n \to \infty$, this probability converges to 1.

A.2.5

Appendix B

Supplementary Notes

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

We have that

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

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

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

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

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

See that this follows from (B.3):

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