ASHWIN REDDY

INFORMATION THEORY

Contents

1 *Introduction* 9

1.1 *Philosophy* 9

1.2 *Prerequisites* 9

1.3 *Why study Information Theory?* 9

I Theory 11

2 *Stochastic Variables* 13

2.1 *Probability Measure* 13

2.2 *Bayes' Theorem* 14

2.3 *Probability Mass Functions* 14

2.4 *Expectation and Variance* 15

2.5 *Probability Density Functions* 17

3 *Entropy* 19

3.1 *Surprisal* 19

3.2 *Bit Representations* 21

3.3 *Jensen's Inequality* 22

3.4 *Joint Entropy* 23

3.5 *Differential Entropy* 23

4 *Source Coding* 25

4.1 Encoding the English Alphabet 25

4.2 Huffman Coding 25

Appendix 29

A *Discrete Stochastic Variables* 29

A.1 Bernoulli Distribution 29

A.2 Binomial Distribution 29

A.3 Geometric Distribution 29

A.4 Poisson Distribution 30

B *Continuous Stochastic Variables* 31

B.1 Uniform 31

B.2 Exponential 32

B.3 Gaussian Distribution 33

C *Extra Formalism* 37

C.1 Kolmogorov Axioms 37

List of Tables

2.2 Expectation and Variance of Common Distributions 17

List of Figures

3.1 Bernoulli Entropy 21

B.1 Example of Uniform stochastic variable 31
B.2 Simple Gaussian 33

1
Introduction

Often, lecture notes like these jump into the material without establishing the basics. This chapter will cover why this book exists and how to use it.

1.1 Philosophy

Two other sets of notes for this course are available at https://github.com/mananshah99/infotheory and http://tiny.cc/infotheory1 . My goal in preparing *this* book was simply to ensure that I understand the concepts. As a result, I've spared no expense in filling these notes with as much content as needed to develop the theory from first principles.

It's also worth noting that the stylistic choices I've made here are very opinionated. The margins are very wide and are filled with important asides and footnotes. I've tried to stick to the typical definition-theorem-proof style with explanatory prose in between. Speaking of color, you've probably noticed that URLs are in blue.

1.2 Prerequisites

This book assumes a proficiency with single-variable calculus and some familiarity with multivariable calculus. Important concepts of probability are developed from scratch, however.

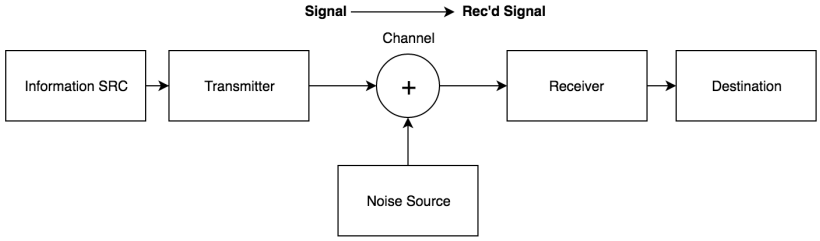
1.3 Why study Information Theory?

Information Theory is a mathematical framework for thinking about what it means to have efficient communication. Examples of information include:

- Email
- Telegraph

- Images
- Speech
- Video

Much of today's digital world revolves around transmitting information: we zip files, email them across the internet, download MP3s, etc. The ideas of information theory give us a rigorous way of characterizing streams of information. The cornerstone of our model will start with the following pipeline of sorts



This model gives us a general way of abstracting the transmission of information. On the left we have a source of information (where a signal originates) that is sent to a transmitter. A channel then allows that signal to flow to a receiver, although noise may be added at this stage. Finally, the receiver sends the signal to the destination. Right now, this model may not be very elucidating. However, we will see that it gives us a structure within which we can consider various mathematical characterizations of information. However, before we can begin to consider information in depth, we must first start our foundation with a solid understanding of probability as that is the underlying theory below Information Theory.

Part I

Theory

2
Stochastic Variables

Definition 2.1 (Stochastic Variable)**.** A **stochastic variable** is a real-valued function of an outcome of an experiment.

Remark. Stochastic variables are also called random variables, but this term seems to imply all possibilities are equally likely or cannot be determined. The word stochastic generally means something like "depending upon probabilities."

Example 2.1. Toss a coin 7 times. The number of heads in the sequence could be a stochastic variable.

Remark. The 7-long sequence itself is not a stochastic variable, since it is not a real value. We should always have a clear way of assigning a real number to the outcome.

Example 2.2. Sum two rolls of a die. This value could be a stochastic variable. The number of 5's rolled could also be a stochastic variable.

A stochastic variable can either be discrete, taking on values from a countable set or continuous, taking on values on an interval from the real number line.

2.1 Probability Measure

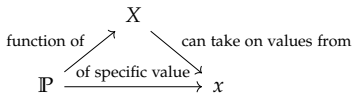
Definition 2.2 (Sample Space)**.** The sample space is the set of all possible outcomes for an experiment.

The sample space is typically denoted as Ω.

Definition 2.3 (Event)**.** An **event** is a subset of the sample space Ω.

Remark. Any event belongs to the power set of Ω.[1] Thus, events range from the empty set to a singleton set (i.e. a set with size unity) to Ω itself and anything in between.

[1] The Wikipedia article on events has good examples.



Definition 2.4 (Probability Measure)**.** The probability measure $\mathbb{P}$ is a real-valued function that assigns events probabilities obeying the Kolmogorov axioms.

Our notions of probability are fairly intuitive, so a careful treatment of the Kolmogorov axioms is not needed. They are available in the appendix, however.

Notation. In general, if X is a stochastic variable, the probability mass function can be written in two equivalent ways:

$$p_X(x) = \mathbb{P}(X = x) \tag{2.1}$$

2.2 *Bayes' Theorem*

When events A and B are not independent, knowing what A is gives us information about what B might be (and vice versa). The notation $\mathbb{P}(A|B)$ denotes the probability of A given that or conditioned upon B occuring. Consider that for A and B to happen that B must first happen and A must happen under those circumstances.

$$\mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}\left(A|B\right)$$

From this definition, Thomas Bayes determined how to compute the support B provides for A given *priors* and *posteriors*.

Theorem 2.1 (Bayes' Theorem)**.**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{2.2}$$

2.3 *Probability Mass Functions*

A **probability mass function** (pmf) characterizes a discrete stochastic variable by returning the probability measure of some x in Ω occuring.

Example 2.3. Consider 2 tosses of a fair coin. What is the probability mass function (pmf) of the number of heads given this experiment?

Solution. It is useful to construct a table of all the possibilities.

	Heads	Tails
Heads	2	1
Tails	1	0

From the table, we can conclude that

$$\mathbb{P}(X = x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Example 2.4. Consider a 4-sided die rolled twice. What is the probability mass function for the maximum value of 2 rolls?

Solution. As before, let us think about the various possibilities. There are $4 \times 4 = 16$ total possibilities, and the maximum value can take on values 1, 2, 3, and 4. To take on a value of 1, both rolls must have been a 1; the probability of this happening is $1/16$. To take on a value of 2, one of the rolls must have been a 2 and the other one must have been a 1 or a 2. The possibilites are enumerated as (2,1), (2,2), (1,2). Thus, its chance is $3/16$. For a max value of 3, the possibilities are (3,1), (3,2), (3,3), (1,3), (2,3). Thus,

$$p_X(x) = \begin{cases} 1/16 & x = 1 \\ 3/16 & x = 2 \\ 5/16 & x = 3 \\ 7/16 & x = 4 \\ 0 & \text{otherwise} \end{cases}$$

There are a few common discrete stochastic variables that are worth discussing separately. The chapter on Discrete Stochastic Variables covers 4 important ones. The following sections will assume familiarity with these variables.

2.4 *Expectation and Variance*

As is often the case in mathematics, we like to define general operators on objects to understand their properties. Firstly, note that we can use functions of stochastic variables to build other stochastic variables.

Example 2.5 (Simple Functions of a Stochastic Variable)**.** Consider the following pmf

$$p_X(x) = \begin{cases} 1/9 & x \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\} \\ 0 & \text{otherwise} \end{cases}$$

Here are two functions of X and their probability mass functions:

a) $Y = |X| \implies p_Y(y) = \begin{cases} 2/9 & y \in \{1,2,3,4\} \\ 1/9 & y = 0 \end{cases}$

b) $Z = X^2 \implies p_Z(z) = \begin{cases} 2/9 & z \in \{1,4,9,16\} \\ 1/9 & z = 0 \end{cases}$

A special class of functions, known as moments, include two operators, **expectation** and **variance**.

Definition 2.5 (Expectation)**.** The expectation of a function g of a stochastic variable X is

$$\mathbb{E}[g(X)] \equiv \sum_{x \in \Omega} g(x) \cdot \mathbb{P}(X = x)$$

Theorem 2.2 (Linearity of Expectation)**.**

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

The nth moment about x_0 is defined as $\mathbb{E}[(x - x_0)^n]$.

Definition 2.6 (Variance)**.** The variance is the 2nd moment about the mean.

$$\text{Var}[X] \equiv \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{2.3}$$

Unfortunately, the definition given above in Equation (2.3) tends to be difficult to compute by hand. A little algebra results in a computationally simpler alternative.

Lemma 2.1 (Determinism of Expectation of Expectation)**.**

$$\mathbb{E}\Big[\mathbb{E}[X]\Big] = \mathbb{E}[X]$$

Theorem 2.3 (Computationally Simpler Alternative for Variance)**.**

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}\left[X\right]^2 \tag{2.4}$$

Proof.

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}\Big[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\Big] \\ &= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2. \end{aligned}$$

□

As an aside, the indicator function $\mathbb{1}_A$ indicates whether certain events are in A or not.

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

As a result,

$$\mathbb{E}[\mathbb{1}_A(\omega)] = \mathbb{P}(A)$$
$$\text{Var}[\mathbb{1}_A(\omega)] = P(A)(1 - P(A))$$

Table 2.2: Expectation and Variance of Common Distributions

Distribution	Expectation	Variance
Binomial	np	$np(1-p)$
Geometric	$1/p$	$(1-p)/p^2$
Poisson	λ	λ

2.5 *Probability Density Functions*

Now we turn to stochastic variables that return values on an interval of the real number line. We can use our knowledge of discrete stochastic variables to find analagous versions for continuous stochastic variables. Instead of a probability mass function, we call the function for a continuous stochastic variables **probability density function** (pdf).

Notation. Probability density functions are typically denoted with lowercase English letters, most commonly f.

The requirement that the total probability must add up to unity is still in place:

$$\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$$

However, asking whether $X = x$ is no longer a well-formed question, since X has a continuum. The probability that X takes on the exact value x is nil. Therefore, probabilities of a continuous stochastic variable may only be queried in the following form.

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

By extension, the expectation of a pdf can be computed as

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$$

The chapter on Continuous Stochastic Variables discusses 3 important stochastic variables: uniform, exponential, and Gaussian. These variables will come up later, but for the sake of cleanliness, the proofs and computations have been moved into the appendix.

3
Entropy

Why did we spend the time reviewing probability for information? The answer lies in Proposition 3.1 below:

Proposition 3.1 (Stochasticity of Information)**.** *Information can be modelled as samples from a stochastic source.*

Consider the fact that many emails you send will have a typical structure: a greeting, a body of text, a conclusion. Or that the changes in the frames of a video tend to be rather small. Intuitively, we have a sense that the repetitive elements of data are not information-dense, and therefore, when we transmit this information, we should really only focus on what is novel about each message.

Shannon's original example was to show that text can modelled probabilistically in this way. Let's imagine that we want to generate some text that looks like English. We can first start by creating a sample space Ω that includes letters and spaces. Then, we sample randomly. This is callled a zero-order approximation. Next, we can refine the approximation by making characters more or less frequent. Adapting the probability based on the frequency with which that character appears in a corpus (e.g. make *e* show up about 12% of the time) is called a first-order approximation. An even more refined approach would consider digram (2-character sequences) and their frequencies.

3.1 Surprisal

Now that you're convinced that information can be modelled stochastically, we can consider what information means.

Example 3.1 (The Q-U Question)**.** Consider a game where you predict the next letter of a piece of English text given all the previous letters. You are given the phrase `elephants are q`. You know that

the letter q is nearly always followed by a u. You would then predict that the next letter is a u, and you would be confident in your guess. If, by some odd reason, that next letter is not a u, you would be "surprised".

What we've captured in this example is a working conception of information.

Definition 3.1 (Information)**.** New information (which is really the only kind of information we care about) **is** the "surprisal" of an information source. The more surprised you are, the more information you gain, since you didn't expect to see that result.

We now need a way of quantifying surprisal/information. The core of our theory is that surprisal should be inversely correlated with the probability of occurence. So naturally, we gravitate towards picking something like $1/p$ as our information function. Consider the limit cases, however, of literally using $1/p$. When $p = 0$, we have infinite surprisal, and when $p = 1$, we have 1 unit of surprisal. If something is guaranteed to happen, 1 unit is an odd baseline to use. As a result, we pick $\log(1/p)$, which is more attractive for a few reasons.

1. Continuity
2. Monotonically decreasing in p
3. Never negative
4. With $p = 1$, information becomes 0
5. Information due to independent events is additive

Which base we pick is entirely arbitrary as long as it makes sense (sensible options include 2, 10, and e). The value of entropy changes predictably with a change of base, so it really doesn't pose much of a problem. In most cases, we will pick base 2, since we like to consider binary digits (abbreviated bits).

To each event, we now attach a surprisal value. To characterize a stochastic variable as a whole, we now define entropy.

Definition 3.2 (Entropy)**.** The entropy H of a stochastiv variable X is the expectation of surprisal of X.

$$\begin{aligned} H(X) &\equiv \mathbb{E}\left[\log \frac{1}{\mathbb{P}(X = x)}\right] \\ &= \sum_i p_i \log(1/p_i) \\ &= -\sum_i p_i \log p_i \end{aligned}$$

Shannon's formula for entropy actually has a similarity to thermodynamic entropy:

$$S = -k_B \sum p_i \ln p_i$$

Example 3.2 (Entropy of Bernoulli Distribution)**.**

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

Plotting for every possible value for p, we yield a nice graph:

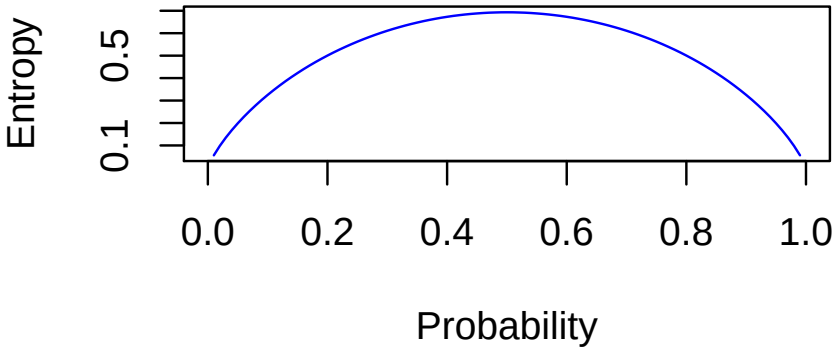


Figure 3.1: Bernoulli Entropy

When p is 0 or 1, we need 0 bits of information, which makes sense because the result was guaranteed. As we go more towards complete randomness (which colloquially, we might also call "entropy" from a physics standpoint), we need more bits to represent the possibilites (a maximum of 1 in this case).

But what does it mean to have 0.47 bits, which we might have if $p(\text{heads}) = 0.9$? Imagine that we had a 100-long sequence of coin flips and we transmitted the information. For the purely random case (i.e. using a fair coin), we would need 100 bits. However, for this extremely unfair case, we could get away with 47 bits without losing any information (on average).

3.2 *Bit Representations*

While we will be overloading the word bit in different contexts in this book, it is useful to understand what it represents. As noted before, bit is an abbreviation of "binary digit." When we talk about a bit in computer science, we typically mean 0 or 1, low voltage or high voltage, etc. Here, we take a bit to mean something like the answer to a single yes or no question with yes and no equally likely. In other words, a coin toss. That is, one bit captures the information of a Bernoulli distribution with $p = 0.5$. From there, we can meaningfully interpret values of entropy as telling us roughly how many of these yes/no questions or coin flips or sequence of binary digits are needed to transmit the information on average.

Example 3.3 (Entropy of a Fair Dice Roll)**.** Find the entropy of a fair dice roll.

Solution. Note that entropy does not care about the actual values of X. Therefore, the entropy is computed as

$$\begin{aligned} H(x) &= \sum_x p(x)\log(1/p(x)) \\ &= \sum_{x=1}^{6} \frac{1}{6}\log(6) \\ &= 6 \cdot \frac{1}{6}\log(6) \\ &= \log 6 \approx 2.585 \end{aligned}$$

Exercise 3.1 (Double the Possibilities)**.** Repeat Example 3.3 except that there are double the number of possible values (i.e. a 12-sided die).

Solution. Intuitively, we just need to add one more bit to flip between the first 6 and last 6 values. Mathematically, we consider $\log(12)$, which by log properties is $\log(2) + \log(6) = 1 + \log(6)$.

3.3 *Jensen's Inequality*

Definition 3.3 (Convex Function)**.** A function $f(x)$ is convex on the interval (a, b) if it is concave up on that interval (i.e. second derivative is positive). Alternatively, it obeys the property that for all (x_1, x_2) within the interval (a, b) and for all λ normalized between 0 and 1,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Theorem 3.1 (Jensen's Inequality)**.** *For a stochastic variable X and a function f,*

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

Theorem 3.2. *If X assumes real values $\{x_1, \ldots, x_n\}$ and $0 \leq H(X) \leq \log r$. Then,*

$$\forall 1 \leq i \leq n, p_i = \frac{1}{r} \iff H(X) = \log r$$

3.4 Joint Entropy

Definition 3.4 (Joint Entropy)**.**

$$H(X,Y) = -\sum_x \sum_y \mathbb{P}(x,y) \log(p(x,y))$$

Definition 3.5 (Conditional Entropy)**.**

$$H(Y|X) = \sum_x p(x) H(Y|X=x)$$

Theorem 3.3.

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

Proof. □

3.5 Differential Entropy

Differential entropy is the continuous version of discrete entropy.

Proof. The probability that X^Δ is in the ith bin is $p(x_i)\Delta x$. Then,

$$\begin{aligned} H(X^\Delta) &= -\sum_i p(x_i)\Delta x \log\left(p(x_i)\Delta x\right) \\ &= \sum_i \left[p(x_i)\Delta x \log \frac{1}{p(x_i)} + p(x_i)\Delta x \log \frac{1}{\Delta x} \right] \end{aligned}$$

□

$$h(x) \equiv -\int_{\mathbb{R}} f(x) \log f(x) \, \mathrm{d}x$$

Example 3.4 (Differential Entropy of Uniform Stochastic Variable)**.** On the interval $[0, a]$,

$$h(x) = \int_0^a \frac{1}{a} \log a \, \mathrm{d}x = \log a$$

Notice that if $a \leq 1$, we have 0 and negative values of entropy, so differential entropy really isn't like discrete entropy.

4
Source Coding

4.1 Encoding the English Alphabet

Here's a practical problem we would like to solve: Given the 26 letters of the English alphabet and assuming letters are coming independently, design an encoder (a schema that converts text into a binary message) to minimize the expected number of bits used per letter. Essentially, can we find an encoding of the English alphabet using a zero-order approximation?

We'll start with a simple solution:

1. Compute how many bits it would take if each letter had the same number of bits. The number of bits needed is given by $\lceil \log_2(26) \rceil$[1]
2. Then, A becomes `00000`, B becomes `00001`, and so on and so forth.
3. Store the characters and their numbers in a matrix. The matrix serves as both the encoding and decoding scheme.

[1] The upper brackets denote the ceiling function or greatest integer function. The number of bits must be a natural number.

Solution #1 is actually the best approach if each character had an equal probability (i.e. $1/26$) of appearing. However, certain characters tend to appear more than others. Therefore, using the same number of bits to represent a commonly occuring character like e and a infrequent character like z is not making the best use of each bit.

Our next solution will take into account frequency. The three most common letters in the English alphabet are E, T, and A. Assign E the value `0`, T the value `1`, and A the value `10`. However, it now becomes impossible to determine whether `10` is a "TA" message or an "E" message. We need to avoid such prefix-collisions to interpret messages without ambiguity.

4.2 Huffman Coding

The solution devised by David Huffman has optimality given certain conditions. First, we'll need to describe the algorithm of Huff-

man coding.

Data: A map $m : \Sigma \to [0, 1]$
Result: A binary tree representing an encoding scheme
while $|\text{preimage}(m)| > 1$ **do**
  $a \leftarrow \arg\min m$;
  Remove a from m;
  $b \leftarrow \arg\min m$;
  Remove b from m;
  Insert symbol ab with frequency $m(a) + m(b)$;
end

Algorithm 1: Huffman Coding

Theorem 4.1 (Huffman Coding Optimality)**.** *If X is a random variable, and L is the expected number of bits per letter using Huffman coding,*

$$H(X) \leq L \leq H(X) + 1$$

An intuitive video explaining Huffman coding can be found here: https://www.youtube.com/watch?v=JsTptu56GM8 . A Python program is available here: https://gist.github.com/ashwinreddy/8b8eb194bc3bf264a81affb5b6cdff06 .

Appendix

A
Discrete Stochastic Variables

A.1 Bernoulli Distribution

A Bernoulli distribution represents the number of heads in tossing a potentially unfair coin once. The unfairness is characterized by a probability p that the coin lands heads. Therefore, the probability that the coin lands tails is $1 - p$.

Definition A.1 (Bernoulli Distribution)**.**

$$\text{Bern}(x) \equiv \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases} \tag{A.1}$$

A.2 Binomial Distribution

Next, the binomial distribution can be imagined as tossing the same unfair coin N times and counting the number of heads. The probability that the number of heads is nil is $(1 - p)^N$ since the implication is that the coin came up tails every single time. The probability that the number of heads is exactly one is $Np(1 - p)^{N-1}$. This is because it must have come up heads once with probability p and tails $N - 1$ times with probability $1 - p$. Additionally, the one heads could come up at any point in the sequence, which introduces a factor of $\binom{N}{k}$.

Definition A.2 (Binomial Distribution)**.**

$$\mathbb{P}(X = k; n, p) \equiv \binom{n}{k} p^k (1 - p)^{n-k} \tag{A.2}$$

You can check that $\mathbb{P}(X = 1; 1, p) = \text{Bern}(p)$

A.3 Geometric Distribution

In a geometric distribution, we keep tossing the coin until there is one heads.

Definition A.3 (Geometric Distribution)**.**

$$p_X(x) = (1-p)^{x-1}p \tag{A.3}$$

A.4 Poisson Distribution

Definition A.4 (Poisson Distribution)**.**

$$p_X(x=k) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

Similar to a binomial distribution, the Poisson distribution can be thought of as the number of replacements needed for a biased lightbulb (rather than a coin) in a given amount of time. In this case, we are typically dealing with a large n and a small p, which leads to a "moderate" np.[1] If $\lambda \equiv np$ (can be thought of as the expected number of times the bulb will burn out),

[1] I have no idea what this means.

Proof. Start with binomial distribution, using λ instead of p:

$$\begin{aligned} p_X(x=k) &= \lim_{n\to\infty}\lim_{p\to 0}\frac{n!}{(n-k)!k!}\binom{\lambda}{n}^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty}\lim_{p\to 0}\frac{n(n-1)\dots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k)} \\ &= e^{-\lambda}\lambda^k/k! \end{aligned}$$

□

What can be modelled with a Poisson distribution?

- Number of customers entering a bank in a given period of time
- Number of misprints on a page
- Number of alpha particles discharged from a radioactive substance.

B
Continuous Stochastic Variables

B.1 Uniform

Definition B.1 (Uniform Stochastic Variable)**.** For the interval $[a, b]$, the uniform stochastic variable assigns all x in the interval the same probability, so that

$$f_X(x) = \frac{1}{b-a}$$

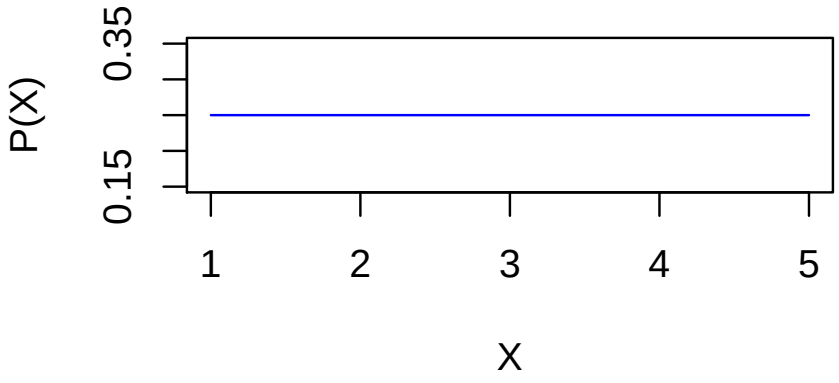


Figure B.1: Example of Uniform stochastic variable

Intuitively, the expectation for the uniform stochastic variable on the interval $[a, b]$ is $(a + b)/2$.

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}} x \frac{1}{b-a}\,\mathrm{d}x \\
&= \frac{1}{b-a}\int_a^b x\,\mathrm{d}x \\
&= \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] \\
&= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{b+a}{2}
\end{aligned}$$

□

The variance may also be computed using the formula in Equation (2.4).

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \left(\frac{a+b}{2}\right)^2 \\
&= \int_a^b x^2 f_X(x)\,\mathrm{d}x - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right] - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{1}{12}(a-b)^2
\end{aligned}$$

B.2 Exponential

Definition B.2 (Exponential Stochastic Variable)**.**

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

Expectation of Exponential Stochastic Variable.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \lambda e^{-\lambda x}\,\mathrm{d}x \\
&= \lambda \int_0^\infty e^{-\lambda x}\,\mathrm{d}x
\end{aligned}$$

The integral can be evaluated by using integration by parts with $u = x$ and $\mathrm{d}v = e^{-\lambda x}\,\mathrm{d}x$. We record $\mathrm{d}u = \mathrm{d}x$ and $v = -\frac{e^{-\lambda x}}{\lambda}$.

$$\begin{aligned}\int_0^\infty e^{-\lambda x}\,\mathrm{d}x &= \left[-\frac{xe^{-\lambda x}}{\lambda}\right]_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\,\mathrm{d}x \\ &= \frac{1}{\lambda}\left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty \\ &= \frac{1}{\lambda^2}\end{aligned}$$

Finally, multiply the λ from the original expression to obtain $1/\lambda$. □

Definition B.3 (Memoryless Property)**.** A stochastic variable is memoryless iff

$$p(x > s + t \,|\, x > t) = p(x > s) \qquad s, t \geq 0$$

The exponential stochastic variable is memoryless.

B.3 Gaussian Distribution

We now give a separate treatment for the very common Gaussian distribution (aka normal distribution). A Gaussian or normal distribution is essentially what people imagine when they are talking about a "bell curve." The base function is $\exp(-x^2/2)$, whose graph is given in Figure B.2.

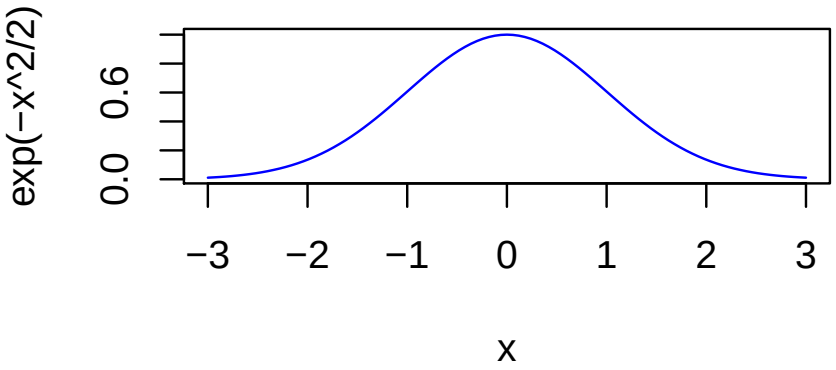


Figure B.2: Simple Gaussian

This function is known as the **standard normal**. However, we have not yet checked that it integrates to unity.

Integral of Gaussian. We want to find $I = \int_{\mathbb{R}} f(x)\,\mathrm{d}x$. However, a simplified I is not possible as is. The solution is a bit tricky.

$$
\begin{aligned}
I &= \sqrt{\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)} \\
&= \sqrt{\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)\left(\int_{\mathbb{R}} f(y)\,\mathrm{d}y\right)} \\
&= \sqrt{\iint_{\mathbb{R}\times\mathbb{R}} \left[f(x)f(y)\right]\,\mathrm{d}x\,\mathrm{d}y} \\
&= \sqrt{\iint_{\mathbb{R}\times\mathbb{R}} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y} \\
&= \sqrt{\int_0^{\infty}\int_0^{2\pi} e^{-r^2/2} r\,\mathrm{d}\theta\,\mathrm{d}r} \\
&= \sqrt{\left(\int_0^{2\pi}\mathrm{d}\theta\right)\left(\int_0^{\infty}\left[e^{-r^2/r} r\right]\mathrm{d}r\right)} \\
&= \sqrt{2\pi\left(e^{-r}\left(-1-r\right)\right)\Big|_0^{\infty}} \\
&= \sqrt{2\pi}
\end{aligned}
$$

□

The general form of the Gaussian includes a factor of $1/\sqrt{2\pi}$ for normalization:

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

Properties:

1. $\mathbb{E}[X] = \mu$
2. $\mathrm{Var}[X] = \sigma^2$

Mean of Gaussian Stochastic Variable. To show that the mean is μ, we first construct a new stochastic variable that is a function of X.

$$Z = \frac{X-\mu}{\sigma}$$

$$\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2}\,\mathrm{d}z$$

If we look at a plot of the function $\exp(-x^2/2)$ as in Figure B.2, it becomes apparent that this function is even. Multiplied by an odd function z, the integrand is odd with endpoints $[-a, a]$ (in this case, $a \to \infty$). Thus, $\mathbb{E}[Z] = 0$. Performing the appropriate shift, we find that this implies that $\mathbb{E}[X] = \mu$. □

Similarly, we can compute the variance of Z.

$$\text{Var}[Z] = \sigma^2$$

C
Extra Formalism

C.1 Kolmogorov Axioms

There are three Kolmogorov axioms:

1. $\nexists A \in \Omega, \mathbb{P}(A) < 0$ (There are no events with a negative probability of happening)
2. $\mathbb{P}(\Omega) = 1$ (The probability of *something* in the sample space happening must be 100%)
3. For a sequence of disjoint[1] sets $A_1, A_2, \ldots, \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$ (The probability of mutually exclusive events happening is the total probability of any one happening)

[1] Disjoint sets have no elements in common. If A and B are disjoint, $A \cap B = \emptyset$