

## 2

# Probability Theory

Before even mentioning networks, I need to lay the groundwork for a basic understanding of a few concepts necessary to make you a good network analyst. These concepts are part of four broad subjects: probability theory, statistics, linear algebra, and machine learning. We start with probability theory here, because that is the foundation on top of which this pyramid is built. Then, we deal with more specialized statistical concepts in Chapter 3, with machine learning in Chapter 4, and linear algebra in Chapter 5.

These chapters deal with entire fields of human knowledge that are much more vast and complicated than the extremely simplified picture I present here. As with many other chapters, my coverage of the subject is the bare minimum I can get away with. The wisest course of action for you would be to skip this book part entirely and take entire courses on these subjects and then come back to this book and start directly with Part II. Alas, that's not an option for everybody, and that is why this book part covers the basics you need to know to enjoy the rest of the book.

If you want to dive deep into probability theory, there are good books on the subject you should check out<sup>1,2</sup>. I will attempt, where possible, to give forward references to later topics in this book, to let you know why understanding probability theory is important for a network scientist.

<sup>1</sup> William Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 1968

<sup>2</sup> Rick Durrett. *Probability: theory and examples*, volume 49. Duxbury Press, 1996

### 2.1 Frequentism and Bayesianism

Probability theory is the branch of mathematics that allows you to work with uncertain events. It gives you the tools to make inferences in cases of uncertainty.

Probability theory is grounded in mathematical axioms. However, there are different ways to interpret what we really mean with the term “probability”. With a very broad brush, we can divide the main interpretations into two camps: the frequentist and the Bayesian.

There are more subtleties to this, but since these are the two main approaches we will see in this book, there is no reason to make this picture more complex than it needs to be.

To understand the difference, let's suppose you have Mrs. Frequent and Mr. Bayes experimenting with coin tosses. They toss a coin ten times and six out of ten times it turns heads up. Now they ask themselves the question: what is the probability that, if we toss the coin, it will turn heads up again?

Mrs. Frequent reasons as follows: "An event's probability is the relative frequency after many trials. We had six heads after ten tosses, thus my best guess about the probability it'll come out as heads is 60%". Note that Mrs. Frequent doesn't really believe that ten tosses gave him a perfect understanding of that coin's odds of landing on heads. Mrs. Frequent knows that he will get the answer wrong a certain number of times, that is what confidence intervals are for, but for the sake of this example we need not to go there.

"Hold on a second," Mr. Bayes says, "Before we tossed it, I examined the coin with my Coin Examiner<sup>TM</sup> and it said it was a fair coin. Of course my Coin Examiner<sup>TM</sup> might have malfunctioned, but that rarely happens. We haven't performed enough experiments to say it did, but I admit that the data shows it might have. So I think the probability we'll get heads again is 51%". Just like Mrs. Frequent, also Mr. Bayes is uncertain, and he has a different procedure to estimate such uncertainty – in this case dubbed "credible intervals" – which again we leave out for simplicity.

Herein lies the difference between a frequentist and a Bayesian. For a frequentist only the outcome of the physical experiment matters. If you toss the coin an infinite number of times, eventually you'll find out what the true probability of it landing on heads is. For a Bayesian it's all about degrees of beliefs. The Bayesian has a set of opinions about how the world works, which they call "priors". Performing enough new experiments can change these priors, using a standard set of procedures to integrate new data. However a Bayesian will never take a new surprising event at face value if it is wildly off its priors, because those priors were carefully obtained knowledge coherent with how the world worked thus far.

Figure 2.1 shows the difference between the mental processes between a frequentist and a Bayesian. The default mode for this book is taking a frequentist approach. However, here and there, Bayesian interpretations are going to pop up, thus you have to know why we're doing things that way.

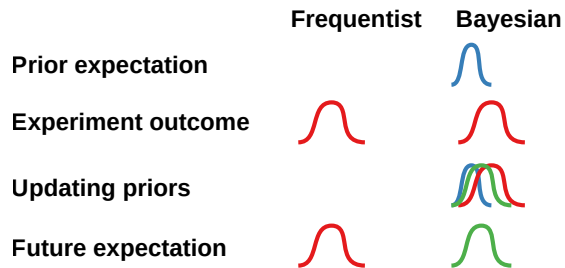


Figure 2.1: Schematics of the mental processes used by a frequentist and a Bayesian when presented with the results of an experiment.

## 2.2 Notation

Probability theory is useful because it gives us the instruments to talk about uncertain processes. For instance, a process could be tossing a die. The first important thing is to understand the difference between *outcome* and *event*. An *outcome* is a *single* possible result of the experiment. A die landing on 2 is an outcome. An *event* is a *set* of possible outcomes on which we're focusing. In our convention, we use  $X$  to refer to outcomes.  $X$  is a random variable and it can take many values and forms, and we don't know which of them it will be before actually running the process. As for events, they are the focus of all questions in probability theory: you can sum up probability theory as the set of instruments that allow you to ask and answer questions about events (sets of  $X$ ) such as: "What is the probability that  $X$ , the outcome of the process, is this and/or this but not that and/or that?"

Mathematically one writes such a question as  $P(X \in S)$ , where  $S$  is a set of the values that  $X$  takes in our question.  $X$  is an outcome,  $X \in S$  is an event. For instance, if we were asking about the event "will the die land on an even number?",  $S = \{2, 4, 6\}$ . So,  $P(X \in S)$  asks what's the probability of the "die lands on an even number" event – or for  $X$  to take either of the 2, 4, 6 values. Note that elements in  $S$  are all possible alternatives: if we write  $P(X \in \{2, 4, 6\})$ , we're asking about the probability of landing on 2 *or* 4 *or* 6. If you want to have the probability of two events happening simultaneously, you have to explicitly specify it with set notation:  $P(X \in \{2, 4, 6\}) \cap P(X \in \{1\})$  asks the probability of landing on an even side and on 1 at the same time.

We also need to consider special questions. For instance, there is the case in which no event happens:  $P(X \in \emptyset)$  (here  $\emptyset$  refers to the empty set, a set containing no elements). The converse is also important: the probability of any event happening. In the case of the die, there are a total of six possible outcomes. Notation-wise, we define the set of all possible outcomes as  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . So this is represented as  $P(X \in \{1, 2, 3, 4, 5, 6\})$ , or  $P(X \in \Omega)$ . Figure 2.2 shows how the mathematical notation corresponds to our visual

intuition.

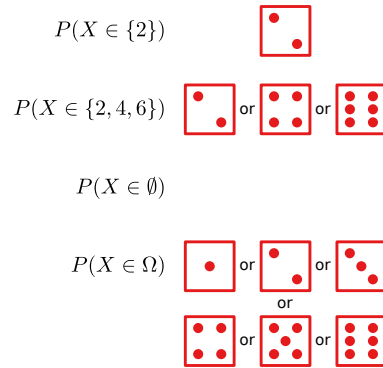


Figure 2.2: A visual shorthand for understanding the mathematical notation of probabilities (left) and the possible outcomes of the “tossing a die” event.

To be more concise, we can skip the explicit reference to the variable  $X$ . For instance, we can codify the outcome “the die lands on 3” with the symbol 3. In this way, we can write  $P(3)$  to refer to the probability of the die landing on 3,  $P(\{2, 4, 6\})$  for the probability of landing on an even number,  $P(\{2, 4, 6\}) \cap P(1)$  for landing on an even number and on 1,  $P(\emptyset)$  for the probability of nothing happening, and  $P(\Omega)$  for the probability of anything possible happening.

### 2.3 Axioms

When building probability theory we need to establish a set of axioms: unprovable and – hopefully – self-evident statements that allow you to derive all other statements of the theory. Probability theory rests on three of such axioms.

First, the probability of an event is a non-negative number. Or: talking about a “negative probability” doesn’t make any sense. Worst case scenario, an event  $A$  is impossible, therefore  $P(A) = 0$  – for instance, this is the “nothing happens” case from the previous section when  $A = \emptyset$ . If  $A$  is possible,  $P(A) > 0$ . In a borderless coin toss, there are only two possible outcomes: heads ( $H$ ) or tails ( $T$ ). The coin cannot land on the non-existing rim. Thus, the probability of landing on the rim is zero. It cannot be negative.

Second, certain events occur with probability equal to one. That is, if  $A$  is an absolutely certain event,  $P(A) = 1$ . Using the notation from the previous section:  $P(\Omega) = 1$ , with  $\Omega = \{H, T\}$  for a coin toss. Note that there isn’t anything magical about the number 1, we could have said that the maximum probability is equal to 42,  $\pi$ , or “meh”. It’s just a convenient convention to define your units.

Third, the probability of happening for mutually exclusive events is the sum of their probabilities, or  $P(\{H, T\}) = P(H) + P(T)$ . A coin cannot land on heads and tails at the same time<sup>3</sup>, thus the probability

<sup>3</sup> Get those Schrödinger coins out of my classroom!

that it lands on heads or tails is the sum of the probability of landing on heads and the probability of landing on tails.

As a corollary, you can also multiply probabilities. If  $A$  and  $B$  are *independent* events, then  $P(A)P(B)$  – their multiplication – tells you the probability of *both* events happening. Independent events are events that have no relation to each other, such as you getting a promotion and the appearance of a new spot on the sun. For *dependent* events, you need to take into account this dependence before applying the multiplication. In a fair die,  $P(1) = P(2) = 1/6$ , but we know that you can't get a 1 if you are getting a 2, so  $P(1)P(2)$  is actually zero, not  $1/36$ . How to perform this check leads us to the world of conditional probabilities.

## 2.4 Conditional Probability

Events do not usually happen in isolation. Things that have happened in the past might influence what will happen in the future. There is a certain probability that the coin will land on heads:  $P(H)$ . But if I know something happened to the coin before the toss – maybe I put some weights in it, event  $W$  – then the probability of heads will change. To handle this scenario, we introduce the concept of “conditional probability”. In our scenario, the notation is  $P(H|W)$ .  $P(H|W)$  is the probability of the coin landing on heads –  $H$  – given that event  $W$  happened.

This view of probability is particularly in line with the Bayesian interpretation, as what you call “prior” is really a synthesis of everything that happened in the past. That is not to say that a frequentist cannot understand conditional probabilities: they can, they just take the usual approach of simply observing what happen before/after something and be done with it.

Conditional probabilities enable you to make a nice set of inferences. Figure 2.3 shows the most basic ones. If you measure  $P(H|W)$ ,

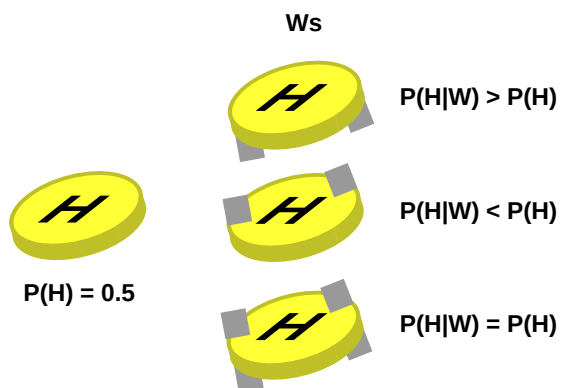


Figure 2.3: The baseline probability of  $H$  is 0.5. When you add feet to the coin ( $W$ ) the coin is more likely to land on the opposite side. Thus,  $P(H|W) \neq P(H)$  and the two events are not independent – unless you add feet on both sides as in the bottom example.

you can figure out what event  $W$  did to the coin. If  $P(H|W) > P(H)$ , it means that adding the weight to the coin made it more likely to land on heads.  $P(H|W) < P(H)$  means the opposite: your coin is loaded towards tails. The  $P(H|W) = P(H)$  case is equally interesting: it means that you added the weight uniformly and the odds of the coin to land on either side didn't change.

This is a big deal: if you have two events and this equation, then you can conclude that the events are independent – the occurrence of one has no effect on the occurrence of the other<sup>4</sup>. This should be your starting point when testing a hypothesis: the null assumption is that there is no relation between an outcome (landing on heads) and an intervention (adding a weight). “Unless,” Mr. Bayes says, “You have a strong prior for that to be the case.”

Reasoning with conditional probabilities is trickier than you might expect. The source of the problem is that, typically,  $P(H|W) \neq P(W|H)$ , and often dramatically so. Suppose we're tossing a coin to settle a dispute. However, I brought the coin and you think I might be cheating. You know that, if I loaded the coin, the probability of it landing on heads is  $P(H|W) = 0.9$ . However, you can't see nor feel the weights: the only thing you can do is tossing it and – presto! – it lands on heads. Did I cheat?

Naively you might rush and say yes, there's a 90% chance I cheated. But that'd be wrong, because the coin already had a 50% chance of landing on heads without any cheating. Thus  $P(H|W) \neq P(W|H)$ , and what you really want to estimate is the probability I cheated given that the coin landed on heads:  $P(W|H)$ . How to do so, using what you know about coins ( $P(H)$ ) and what you know about my integrity ( $P(W)$ ), is the specialty of Bayes' Theorem.

<sup>4</sup> Note that here I'm talking about *statistical* independence, which is not the same as *causal* independence. Two events could be statistically dependent without being causally dependent. For instance, the number of US computer science doctorates is statistically dependent with the total revenue of arcades (<http://www.tylervigen.com/spurious-correlations>). This is what the mantra “correlation does not imply causation” means: correlation is mere statistical dependence, causation is causal dependence, and you shouldn't confuse one with the other. You should check [Pearl and Mackenzie, 2018] to delve deeper into this.

## 2.5 Bayes' Theorem

Bayes' Theorem is an almost magical formula that allows you to estimate the probability of an event based on your priors. Keeping the example of cheating on a coin toss, we want to estimate the probability I cheated and rigged the coin so it lands on heads after we tossed it and it indeed landed on heads – in mathematical notation:  $P(W|H)$ . To do so, you need to have priors. You need to know: what's the probability of heads for all coins in the world (whether they are rigged or not,  $P(H)$ ), what's the probability I rigged the coin ( $P(W)$ ), and what is the probability of obtaining heads on a rigged coin ( $P(H|W)$ ). Without further ado, here's one of the most important formulas in human history:

$$P(W|H) = \frac{P(H|W)P(W)}{P(H)}.$$

Figure 2.4 shows a graphical proof of the theorem. When trying to derive  $P(W|H)P(H)$ , we realize that's identical to  $P(H|W)P(W)$ , from which Bayes' theorem follows.

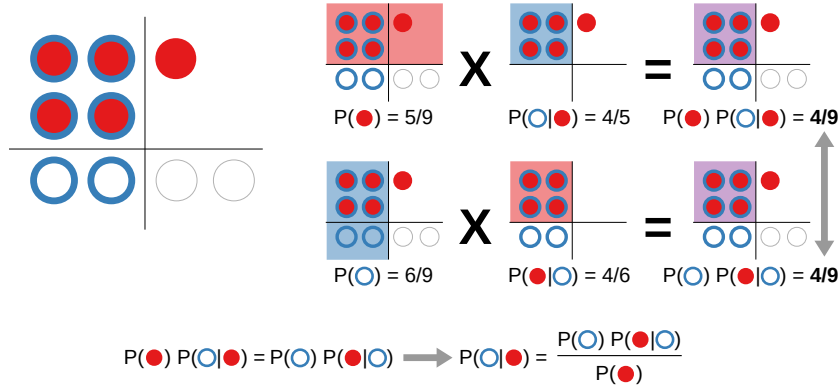


Figure 2.4: The table on the left shows the occurrence of all possible events: red circles (5), blue borders (6), red circles with blue borders (4) and neither (2).

I already told you that I'm a pretty good coin rigger ( $P(H|W) = 0.9$ ). For the sake of the argument, let's assume I'm a very honest person: the probability I cheat is fairly low ( $P(W) = 0.3$ ).

Now, what's the probability of landing on heads ( $P(H)$ )?  $P(H)$  is trickier than it appears, because we're in a world where people might cheat. Thus we can't be naive and saying  $P(H) = 0.5$ .  $P(H)$  is 0.5 if rigging coins is impossible. It's more correct to say  $P(H|-W) = 0.5$ : a non rigged coin (if  $W$  didn't happen, which we refer to as  $-W$ ) is fair and lands on heads 50% of the times. The real  $P(H)$  is  $P(H|-W)P(-W) + P(H|W)P(W)$ . In other words: the probability of the coin landing on heads is the non rigged heads probability if I didn't rig it ( $P(H|-W)P(-W)$ ) plus the rigged heads probability if I rigged it ( $P(H|W)P(W)$ ).

The probability of not cheating  $P(-W)$  is equal to  $1 - P(W)$ . This is because cheating and non cheating are mutually exclusive and either of the two *must* happen. Thus we have  $\Omega = \{W, -W\}$ . Since  $P(\Omega) = 1$  and  $P(W) = 0.3$ , the only way for  $P(W, -W)$  to be equal to 1 is if  $P(-W) = 0.7$ .

This leads us to:  $P(H) = P(H|-W)P(-W) + P(H|W)P(W) = 0.5 \times 0.7 + 0.9 \times 0.3 = 0.62$ . Shocking.

The aim of Bayes' theorem is to update your prior about me cheating ( $P(W)$ ) given that, suspiciously, the toss went in my favor ( $P(W) \rightarrow P(W|H)$ ). Plugging in the numbers in the formula:

$$P(W|H) = \frac{0.9 \times 0.3}{0.62} = 0.43.$$



A couple of interesting things happened here. First, since the event went in my favor, your prior about me possibly cheating got updated. Specifically, the event became more likely: from 0.3 to 0.43. Second, even if my success probability after cheating is very high, it is still more likely that I didn't cheat, because your prior about my lack of integrity was low to begin with.

This second aspect is absolutely crucial and it's easy to get it wrong in everyday reasoning. The textbook example is the cancer diagnosing machine. Let's say that 0.1% of people develop a cancer, and we have this fantastic diagnostic machine with an accuracy of 99.9%: the vast majority of people will be diagnosed correctly (positive result for people with cancer and negative for people without). You test yourself and the test is positive. What's your chance of having cancer? 99.9% accuracy is pretty damning, but before working on your last will, you apply Bayes' Theorem:

$$P(C|+) = \frac{0.999 \times 0.001}{0.999 \times 0.001 + 0.001 \times 0.999} = 0.5.$$

The probability you have cancer is *not* 99.9%: it's a coin toss! (Still bad, but not *that* bad).<sup>5</sup>

The real world is a large and scary environment. Many different things can alter your priors and have different effects on different events. The way a Bayesian models the world is by means of a Bayesian network: a special type of network connecting events that influence each other. Exploring a Bayesian network allows you to make your inferences by moving from event to event. I talk more about Bayesian networks in Section 6.4.

## 2.6 Stochasticity

Colloquially, a stochastic process is one or more random variables that change their values over time. The quintessential stochastic process is Brownian motion. Brown observed very light pollen particles on water changing directions, following a *stochastic* path that seemed governed purely by randomness. Interestingly, this problem was later solved by Einstein in one of his first contributions to science<sup>6</sup>, working off important prior work<sup>7</sup>. He explained the seemingly random changes of direction as the result of collision between the pollen and water molecules jiggling in the liquid.

When you have a stochastic process, there is an almost infinite set of results. The pollen can follow potentially infinite different paths. When you observe an actual grain, you obtain only one of those paths. The observed path is called a realization of the process. Figure 2.5 shows three of such realizations, which should help you visualize

<sup>5</sup> Of course, in the real world, if you took the test it means you thought you might have cancer. Thus you were not drawn randomly from the population, meaning that you have a higher prior that you had cancer. Therefore, the test is more likely right than not. Bayes' theorem doesn't endorse carelessness when receiving a bad news from a very accurate medical test.

<sup>6</sup> Albert Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der physik*, 4, 1905

<sup>7</sup> Louis Bachelier. Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86, 1900



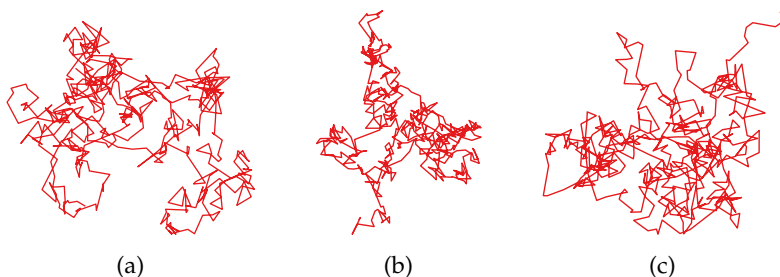


Figure 2.5: Three realizations of a Brownian stochastic motion on a two dimensional plane.

the intrinsic randomness of the change of direction.

Whenever you encounter the word “stochastic” in this book or in a paper, we’re referring to a process governed by these dynamics. For instance, a stochastic matrix is a matrix whose rows and/or columns sum up to one. We call it stochastic, because such matrices are routinely used to describe stochastic processes. By having their rows to sum to one, you can interpret each entry of the row as the *probability* of its corresponding event. The row in which you are tells you the current state of the process, the column tells you the next possible state, and the cell value tells you the probability of transitioning to each of the next possible states (column) given the current state (row). In other words, it is the probability of one possible realization of a single step in a stochastic process. In network science, you normally have stochastic adjacency matrices, which are the topic of Section 8.2.

0.3	0.2	0.08	0	0.07	0.1	0	0.07	0.2
0.1	0.3	0.04	0.04	0.05	0.2	0.04	0.09	0.1
0.08	0.06	0.3	0.06	0	0.2	0.06	0.06	0.1
0	0.08	0.08	0.2	0	0.2	0.08	0.2	0.08
0.07	0.07	0	0	0.3	0.1	0.08	0.2	0.2
0.07	0.1	0.1	0.06	0.07	0.3	0.04	0.09	0.08
0	0.08	0.08	0.08	0.1	0.1	0.2	0.08	0.2
0.05	0.09	0.04	0.1	0.1	0.1	0.04	0.3	0.1
0.1	0.1	0.08	0.03	0.1	0.1	0.08	0.1	0.3

Figure 2.6: A right stochastic matrix.

Figure 2.6 is a stochastic matrix<sup>8</sup>. The rows tell you your current state and the columns tell you your next state. If you are in the first row, you have a 30% probability of remaining in that state (the value of the cell in the first row and first column is 0.3). You have a 20% probability of transitioning to state two (first row, second column), 8% probability of transitioning to state three, and so on.

## 2.7 Markov Processes

It should be clear now that, even if the next state is decided by a random draw, a stochastic process isn’t necessarily uniformly random. In Brownian motion, the next position is determined by your

<sup>8</sup> Specifically, it is a right stochastic matrix: the rows sum to one, although there’s a bit of rounding going on. In a left stochastic matrix, the columns sum to one.

previous position as well as a random kick. This observation is at the basis of a fundamental distinction between three flavors of stochastic processes, which are the most relevant for network science. The three flavors are: Markov processes, non-Markov processes, and higher-order Markov processes.

In a Markov process, the next state is exclusively dependent on the current state and nothing else. No information from the past is used: only the present state matters. That is why a Markov process is usually called “memoryless”. The stochastic process I described when discussing Figure 2.6 is a typical Markov process. The only thing we needed to know to determine the next state was the current state: in which row are we?

The classical Markov process in network science is the random walk. A random walker simply chooses the next node it wants to occupy, and its options are determined solely by the node it is currently occupying. Rather surprisingly, random walks are one of the most powerful tools in network science and have been applied to practically everything. I’m going to introduce them properly in Chapter 11, but they will pop up throughout the book – for instance, in community discovery (Part X) and in network sampling (Chapter 29). Figure 2.7 shows an example of a random walk. As you can see, we start from the leftmost node. From that state, reaching the two rightmost ones is impossible because the nodes are not connected. Only when you transition to another state, new states become available.

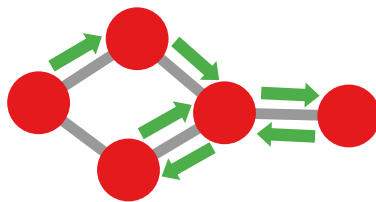


Figure 2.7: A random walk. The green arrows show the state transitions.

A bit more formally, let’s assume you indicate your state at time  $t$  with  $X_t$ . You want to know the probability of this state to be a specific one, let’s say  $x$ .  $x$  could be the id of the node you visit at the  $t$ -th step of your random walk. If your process is a Markov process, the only thing you need to know is the value of  $X_{t-1}$  – i.e. the id of the node you visited at  $t - 1$ . In other words, the probability of  $X_t = x$  is  $P(X_t = x | X_{t-1} = x_{t-1})$ . Note how  $X_{t-2}, X_{t-3}, \dots, X_1$  aren’t part of this estimation. You don’t need to know them: all you care about is  $X_{t-1}$ .

On the other hand, a non-Markov process is a process for which knowing the current state doesn’t tell you anything about the next possible transitions. For instance, a coin toss is a non-Markov process. The fact that you toss the coin and it lands on heads tells you nothing

about the result of the next toss – under the absolute certainty that the coin is fair. The probability of  $X_t = x$  is simply  $P(X_t = x)$ : there's no information you can gather from your previous state.

Finally, we have higher-order Markov processes. Higher-order means that the Markov process now has a memory. A Markov process of order 2 can remember one step further in the past. This means that, now,  $P(X_t = x | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2})$ : to know the probability of  $X_t = x$ , you need to know the state value of  $X_{t-2}$  as well as of  $X_{t-1}$ . More generally,  $P(X_t = x | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots, X_{t-m} = x_{t-m})$ , with  $m \leq t$ .

The classical network examples of a higher order Markov process is the non-backtracking random walk (Figure 2.8). In a non-backtracking random walk, once you move from node  $u$  to node  $v$ , you are forbidden to move back from  $v$  to  $u$ . This means that, once you are in  $v$ , you also have to remember that you came from  $u$ . Higher order Markov processes are the bread and butter of higher order network problems, which is the topic of Chapter 34.

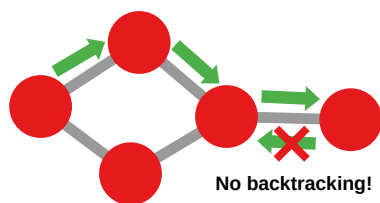


Figure 2.8: A non-backtracking random walk. The green arrows show the state transitions.

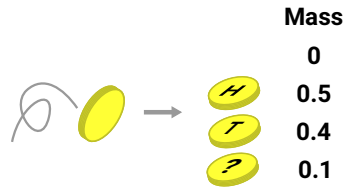
## 2.8 Alternatives to Probability Theory

When it comes to dealing with uncertainty, probability theory is not the only game in town. Here I briefly present two alternatives. These will be handy when we focus specifically on probabilistic networks in Chapter 28.

### *Dempster-Shafer's Theory of Evidence*

In Dempster-Shafer's theory of evidence (DST) we take the key insight from Bayes and we turn it up to eleven. One thing that underlies Bayesian thought is that probabilities are subjective. If two people have different priors, say  $A$  and  $B$  such that  $P(A) \neq P(B)$ , then they will disagree on the probability of event  $C$ , which will depend on the priors. This doesn't happen with a frequentist framework, because there are no priors and  $P(C)$  is based on objective data available to everyone. However, besides this subjectivity, Bayes still uses the axioms and the rules of probability theory.

DST is a generalization of probability theory, which moves from exact probabilities to probability intervals. Its central parts are *beliefs* and *plausibilities*, and these two things don't necessarily behave like probabilities<sup>9,10</sup>.



Starting with beliefs, the first thing you need is a *degree of belief*, which estimates your ability to prove a set of beliefs<sup>11</sup>. This degree of belief is quantified by a function which is conventionally called its Mass function. Figure 2.9 shows a relatively simple example when tossing a coin – slightly loaded on heads. The distinction between Mass in DST and classical probability is that it considers the case “we don’t know whether heads or tail” as distinct from “heads” and “tail”. In probability theory, you wouldn’t make this distinction, because no other outcome than heads or tails can happen, even if for some reason you don’t know the outcome. But in DST you want to model this, because we’re talking about the ability of proving our statement, so we need to specifically take into account the situation in which we don’t actually know the result – e.g., if the coin rolled under the sofa and we can’t see it. In that case, we don’t have any evidence to say that the coin landed on heads or tail.

In summary, if  $\Omega = \{H, T\}$ , then  $p(\Omega) = p(H) + p(T) = 1$ , but  $Mass(\Omega) \neq Mass(H) + Mass(T)$ :  $Mass(\Omega)$  is less trivial and actually informative – it is the amount of uncertainty we have about the outcome of the event given the imperfection of our evidence. So the Mass function is basically giving all the available evidence a probability and obeys the following two rules:

1.  $Mass(\emptyset) = 0$ , and
2.  $\sum_{X \in 2^\Omega} Mass(X) = 1$ .

Here,  $2^\Omega$  means all possible subsets of  $\Omega$ , which in my simple case are:  $\emptyset$ ,  $\{H\}$ ,  $\{T\}$ , and  $\{H, T\}$ . The first property means that it’s impossible to prove that nothing happened – we know we tossed the coin, it must have landed on something. That’s why the ugly coin toss drawing on the left of Figure 2.9 is necessary: to show we performed the tossing. The second property means that  $\Omega$  contains all possible answers to our question – so what’s actually true must be a subset of  $\Omega$ .

<sup>9</sup> Arthur Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38, 1967

<sup>10</sup> Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976

Figure 2.9: An illustration of Mass in DST. After tossing a coin (left) each subset of  $\Omega$  obtains a value.


<sup>11</sup> Judea Pearl. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5-6):363–389, 1990



	Mass	Hyp.	Belief
	0		0
	0.5		0.5
	0.4		0.4
	0.1		1.0

Figure 2.10: An illustration of Belief in DST, based on the Mass from Figure 2.9. Note how Belief of a subset of size 1 is equal to its Mass.

Now that you know the probabilities of all possible outcomes, you must make a hypothesis which is a set of potential outcomes. To estimate your ability to prove your hypothesis, you sum all the Mass values of all the subsets of your hypothesis. This is the Belief function, and you can see how it works in Figure 2.10. If your hypothesis has no support in the gathered evidence then its Belief value is zero, while if it is absolutely certain then Belief evaluates to one. So Belief tells you how likely your hypothesis – or any of its subsets – is to be proven given the available evidence.









	Mass	Hyp.	Belief	Plaus.
	0		0	0
	0.5		0.5	0.6
	0.4		0.4	0.5
	0.1		1.0	1.0

Figure 2.11: An illustration of Plausibility in DST, based on the Mass and Belief Figure 2.10.

DST also allows to compute the Plausibility function, which is an upper bound of Belief. In practice, you can tally up all the evidence of your hypothesis not containing the truth and take the inverse of it. Figure 2.11 shows how it works. Plausibility is basically estimating how much your hypothesis can survive an attempt to prove it false. A handy rule to remember is that  $Plausibility(X) = 1 - Belief(\bar{X})$ , where  $\bar{X}$  is the complementary set of  $X$  – the plausibility of something is the opposite of your belief of that something being proven false.

We go through the trouble of defining these things because DST has some advantages. For instance, there are some operations you can do with the Belief and Plausibility functions – which we are not going to see here – but allow you to work with different hypotheses in conflict. Classical probability theory is ill suited to handle these cases. For instance, suppose that the ice cream shop has three flavors: chocolate, strawberry, and vanilla. We want to share an ice cream and my preference is 99% chocolate and 1% vanilla, while yours are 99% strawberry and 1% vanilla. We should obviously go for vanilla, and DST agrees, but if we took a probabilistic approach ignoring DST, you might instead say that vanilla is the least likely solution, and we would end up with either chocolate or strawberry, much

to the distress of either of us. For instance by naively aggregating probabilities as 49.5% for each strawberry and chocolate. To be fair, DST also ends up saying weird things occasionally, which has led researchers to formulate ways to turn it into computable functions that are different from the ones I explained<sup>12</sup>.

Moreover, probability theory must assign a probability to an event, even if there is no evidence for it, while DST can simply give it zero Mass. For instance, if we have a potentially loaded die, in probability theory using a Bayesian approach we must start by assigning a prior probability of  $1/6$  to all outcomes *even if we have zero evidence for it*. In DST, you'd give them Mass zero instead, and Mass one to  $\Omega$  and then start gathering evidence.

### *Fuzzy Logic*

In probability theory, you only deal with boolean events, whose truth values can either be zero or one. Either something is false or it is true. The coin either landed on heads or on tails. In fuzzy logic, you work with something different. Things can have degrees of truthiness, which is to say we assign them a truth value between zero and one. If it is zero, we're certain that a statement is false, if it is one we're certain it is true, and if it is a value in between then there is some vagueness about whether it is true or false<sup>13</sup>.

For instance, at the moment of writing this paragraph I am 39 years old. Is that young or old? Well, you could line up 100 people and ask them this question. Maybe 60 will say that I'm young, 39 will say that I'm old (I'm so insecure I am disrespected even in my thought experiments), and 1 will say something else. In probability theory you could model this as something like: there's a 39% chance a random person will call me old (hey!). But in fuzzy logic you'd do something different. You could say that I belong to both the sets of young people and old people, with different strengths. I'm 60% young and 39% old – I frankly don't know if that's an improvement over the alternative.

The consequences of this difference lead to different outcomes when working with fuzzy logic. We'll see a basic common example<sup>14</sup>, but know that there are alternative ways to implement fuzzy logic<sup>15</sup>. Let's assume that the degree to which a person belongs to an age class depends on their age, following the function I draw in Figure 2.12.

For the age I highlight, probability theory can say the following things:

- The probability of picking somebody who could call me both young and old (assuming I ask multiple time and people can

<sup>12</sup> Kari Sentz and Scott Ferson. *Combination of evidence in dempster-shafer theory*. 2002

<sup>13</sup> Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 2013

<sup>14</sup> Ebrahim H Mamdani. Application of fuzzy algorithms for control of simple dynamic plant. In *Proceedings of the institution of electrical engineers*, volume 121, pages 1585–1588. IET, 1974

<sup>15</sup> Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, (1):116–132, 1985

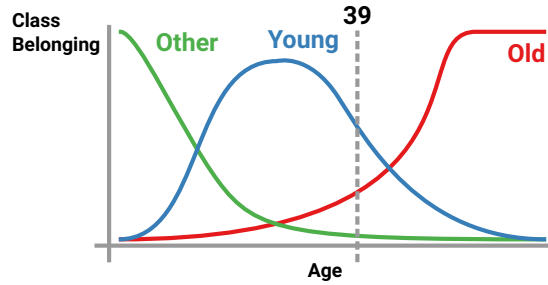


Figure 2.12: An illustration of fuzzy logic. The degrees of belonging (y axis) to different age classes (line color) for a given age (x axis).

change their mind independently from what they said the first time) is  $P(Y \cap O) = P(Y)P(O) = 0.234$ ;

- The probability of picking somebody who will call me either young or old is  $P(Y \cup O) = P(Y) + P(O) = 0.99$ ;
- The probability of somebody not calling me young is  $P(\bar{Y}) = 1 - P(Y) = 0.61$ .

But in fuzzy logic we have:

- My belonging to the class of people who are both young and old is  $P(Y \cap O) = \min(P(Y), P(O)) = 0.39$  – it can't be any higher than my minimum belonging, because to fully belong to the young-old class I must be fully young and fully old;
- My belonging to the class of people who are either young or old is  $P(Y \cup O) = \max(P(Y), P(O)) = 0.6$  – it can't be any lower than my maximum belonging, because if I am fully old then I am also fully-young-or-fully-old;
- My belonging to the class of people who are not young is  $P(\bar{Y}) = 1 - P(Y) = 0.61$ .

## 2.9 Summary

1. Probability theory gives you the tools to make inferences about uncertain events. We often use a frequentist approach, the idea that an event's probability is approximated by the aggregate past tests of that event. Another important approach is the Bayesian one, which introduces the concept of priors: additional information that you should use to adjust your inferences.
2. Probabilities are non-negative estimates. The set of all possible outcomes has a probability sum of one. Summing two probabilities tells you the probability of either of two independent outcomes to happen.



3. The conditional probability  $P(A|B)$  tells you the probability of an outcome  $A$  given that you know another outcome  $B$  happened. If  $P(A|B) \neq P(A)$  then the two outcomes are not independent. Bayes' Theorem allows you to infer  $P(A|B)$  from  $P(B|A)$ .
4. When we track the change over time of one or more random variables, we're observing a stochastic process. Markov processes are stochastic processes whose status exclusively depends on the status of the system in the previous time step.
5. There are alternative to probability theory when working with uncertainty. Dempster-Shafer's theory of evidence allows to work with the degrees of beliefs in specific hypotheses, while fuzzy logic allows for multiple things to be a little bit true at the same time.

### 2.10 Exercises

1. Suppose you're tossing two coins at the same time. They're loaded in different ways, according to the table below. Calculate the probability of getting all possible outcomes:

$p_1(H)$	$p_2(H)$	H-H	H-T	T-H	T-T
0.5	0.5				
0.6	0.7				
0.4	0.8				
0.1	0.2				
0.3	0.4				

2. 60% of the emails hitting my inbox is spam. You design a phenomenal spam filter which is able to tell me, with 98% accuracy, whether an email is spam or not: if an email is not spam, the system has a 98% probability of saying so. The filter knows 60% of emails are spam and so it will flag 60% of my emails. Suppose that, at the end of the week, I look in my spam box and see 963 emails. Use Bayes' Theorem to calculate how many of those 963 emails in my spam box I should suspect to be non-spam.
3. You're given the string: "OCZ XJMMZXO VINRZM". Each letter follows a stochastic Markov process with the rules expressed by the table at <http://www.networkatlas.eu/exercises/2/3/data.txt>. Follow the process for three steps and reconstruct the correct answer. (Note, this is a Caesar cipher<sup>16</sup> with shift 7 applied three times, because the Caesar cipher is a Markov process).
4. Suppose that we are examining a painting and we're trying to date it with the century when it was produced. Find out the Belief and Plausibility values for all hypotheses given the following Mass

<sup>16</sup> [https://en.wikipedia.org/wiki/Caesar\\_cipher](https://en.wikipedia.org/wiki/Caesar_cipher)

estimation (note that, by definition  $\Omega = \{\text{XIV}, \text{XV}, \text{XVI}\}$  must have Belief and Plausibility equal to one):

Hypothesis	Mass	Belief	Plausibility
$\emptyset$	0.00		
XIV	0.16		
XV	0.04		
XVI	0.21		
{XIV, XV}	0.34		
{XV, XVI}	0.16		
{XIV, XVI}	0.08		
$\Omega$	0.01	1	1

# 3

## Statistics

In Chapter 2 I explained the basic concepts of probability theory you need to understand to be a good network scientist. This chapter focuses on basic statistical concepts that are also necessary to analyze your networks. The disclaimer I put at the beginning of Chapter 2 also holds here: statistics is much more vast and complicated than what I present here. This chapter is emphatically *not* a substitute for a proper statistics textbook<sup>1,2</sup>, which you could use to study this stuff further. You'll get a sense of how powerful statistics is<sup>3</sup>, in that it can allow you to support any point. As the saying goes: there are lies, damn lies, and statistics.

The main difference between probability and statistics is that you can do a lot of work in probability theory without actually looking at any data. When data takes the center stage, you enter in the world of statistics. Statistics covers more than simply describing your data: you should think in statistical terms also when collecting, cleaning, validating your data. But here we ignore all that and we focus on the tools that allow you to say something interesting about your data, assuming you did a good job collecting, cleaning, and validating it.

<sup>1</sup> Charles Wheelan. *Naked statistics: Stripping the dread from the data*. WW Norton & Company, 2013

<sup>2</sup> Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018

<sup>3</sup> Darrell Huff. *How to lie with statistics*. Penguin UK, 1954

### 3.1 Summary Statistics

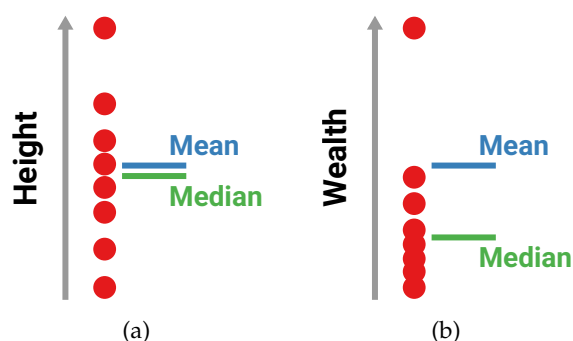
#### *Mean & Median*

One common use of statistics is to give a quick description of what is in the data. The most classical task is to try and figure out what are the values you'd expect to find if you were to look directly at the data. For instance, if you want to know the height of the average human, you might want to calculate the mean height:  $\mu(H) = \sum_i H_i / |H|$ , where  $H_i$  is the height of one human. The mean in this case would tell you what you'd expect to see if you were to measure the height of a random person.

However, that works for height because height is normally dis-

tributed – meaning that the average value is actually the mean and values farther from the mean are progressively more rare. We’ll see what a normal distribution looks like in Section 3.2. The same section will also tell you that not all variables distribute like that: in some cases the mean is actually not a great approximation of your average (or “typical”) case. Wealth is like that: a tiny fraction of people own vastly more than the majority. In this case, the median gives you a better idea<sup>4</sup>. The median tells you the value that splits the data in two equally populated halves: 50% of the points are below the median and 50% of the points are above.

Figure 3.1 shows that the mean and the median can be quite different. In network science, we use the mean extensively – even though maybe we shouldn’t. When we’ll talk about the number of connections a node has in a network (Chapter 9), we’ll see we routinely take its “average” by calculating its mean. But connection counts in real networks typically do not follow a neat normal distribution. These distributions tend to look more like Figure 3.1(b) than Figure 3.1(a). So, perhaps the mean count of connections is not the most meaningful thing you can calculate.



<sup>4</sup> David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman and hall/CRC, 2003

Figure 3.1: The mean (blue) and median (green) for two different variables (y axis): (a) normally distributed height; (b) skewed distribution of wealth.

This disconnect between the mean and the typical case is true for the arithmetic mean I show here, but there are other types of means – such as geometric or harmonic – which can take into account some special properties of the data<sup>5</sup>.

### Variance & Standard Deviation

When we deal with an average observation, we might want to know not only its expected value, but also how much we expect it to differ from the actual average value. Even if the average human height is, let’s say, 1.75 meters, we could think of two radically different populations. In the first, almost everyone is more or less 1.75 and heights don’t vary much. In the other, the opposite is true: the average is still 1.75, but people could be anything between 1 meter and 2.5 meters. So the heights in this second population vary much more. We

<sup>5</sup> Philip J Fleming and John J Wallace. How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, 1986

need to have a tool allowing us to distinguish these two populations. Since the difference is all about how much the heights *vary*, we call this measure *variance*. Variance (and standard deviation) helps you quantify how dispersed your values are away from the mean.

Variance is almost literally the mean difference from the mean. The only tweak is that we take the square of this difference:  $\text{var}(H) = \mu((H - \mu(H))^2)$ . We take the square because we don't want values below the mean to cancel out values above the mean. The standard deviation is simply the square root of the variance:  $\sigma(H) = \sqrt{\text{var}(H)}$ . The advantage of the standard deviation is that it ends up having the same units as the original variable. For example, for heights measured in cm, the variance will be in  $\text{cm}^2$ , but the standard deviation will be in cm again. Figure 3.2 shows what it looks like to have different variances for variables with the same mean.

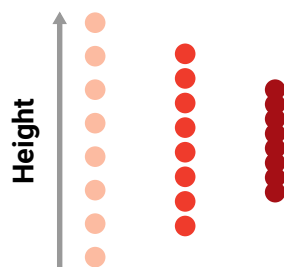


Figure 3.2: The variances of three different populations, going from left (high variance, bright red) to right (low variance, dark red).

The concept of variance will be important when we talk about data dimensionality reduction (Section 5.6) and degree distributions (Section 9.3). We will even see how to modify its definition to create a notion of network variance (Section 47.5)

### *Distribution & Skewness*

Knowing how skewed your data is can be quite important – in wealth distribution, how skewed the data is equals how screwed people are. Variance and standard deviation can help you quantify this. There are different formulas and different terms to talk about skewness<sup>6</sup>, but for our purposes we limit ourselves to a bit of terminology.

First: what actually is a distribution? I've used this term in an intuitive way without really defining it. Let's do it here. When you perform an experiment, or observe a stochastic process, you have many possible outcomes. For instance, you're measuring the heights of all people in a country and so your possible outcomes are all the possible heights a person can have. Sometimes, you're not interested on the frequency of measuring a specific outcome – say 175 cm. Sometimes, you want to study all possible outcomes together, to determine which is more likely, what you could expect when you

<sup>6</sup> Paul T Von Hippel. Mean, median, and skew: Correcting a textbook rule. *Journal of statistics Education*, 13(2), 2005

measure more people, if there are maximums and minimums you don't expect to ever exceed, and so on. This is the task of a distribution. A distribution is a function that, for each outcome in the set of all possible ones (called the “sample space”), tells you how many times you measured a given outcome. Figure 3.3 shows a vignette on how to interpret a plot showing you a distribution.

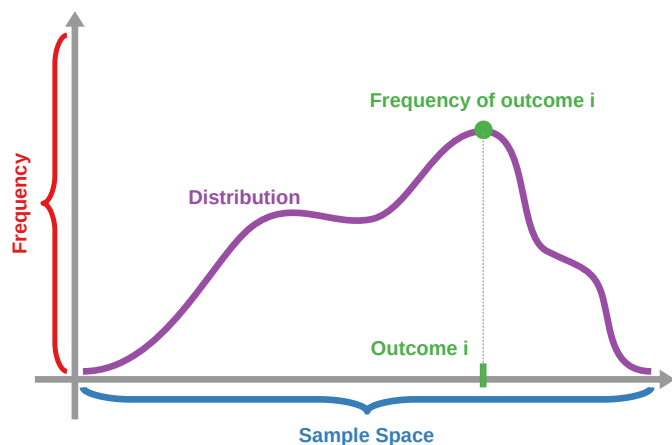


Figure 3.3: A distribution, connecting every possible outcome in the sample space (x axis) to a frequency (y axis).

Skewness is a property of a distribution, it measures its symmetry. A symmetric distribution has no skewness, and any asymmetry will create a non-zero skewness value: positive if the skewedness is on the right and negative if it is on the left – see Figure 3.4.

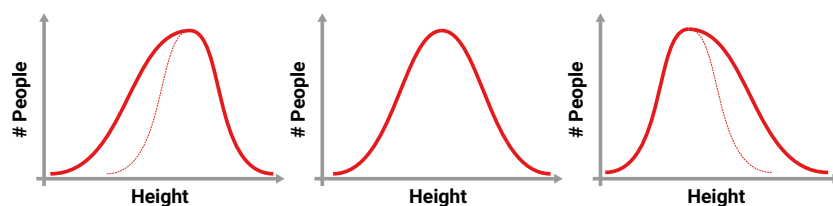


Figure 3.4: Three distributions with different skewness: (a) the values lower than the average are more likely, (b) no skewness, (c) the values higher than the average are more likely.

An important related concept to skewness is the heavy-tailed distribution. There are two types of heavy-tailed distributions that interest network scientists, because they're often observed in real world networks. They are the long tail and the fat tail<sup>7</sup>. In a long tail, you can find arbitrarily large outliers: that means the very highest value can be many times larger than the second-highest one. You might know these from the popular concept of the black swan<sup>8</sup>. In these kinds of distributions, observations can happen that are much more extreme than anything we have seen so far. You'll see a network example in Chapter 9, when we'll see it is common for nodes in real network to have a long tail in the number of connections attached to them. With a fat tail, you still have outliers that can be many times

<sup>7</sup> In case you were wondering: yes, statisticians body-shame distributions.

<sup>8</sup> Nassim Nicholas Taleb. *The black swan: The impact of the highly improbable*, volume 2. Random house, 2007

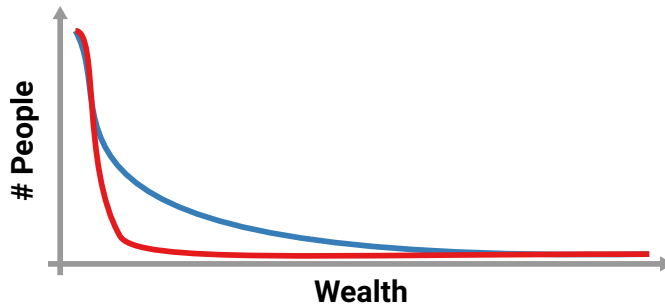


Figure 3.5: A long tail (red) and a fat tail (blue).

over the average value. However, these outliers are more common and less extreme. Figure 3.5 shows a graphical example.

If we're talking wealth, a long tail world is a world with a single Jeff Bezos and everybody else works in an Amazon warehouse. In a fat tail world, Jeff might not be quite as rich, but there are a few billionaire friends to keep him company.

### 3.2 Important Distributions

Chapter 9 will drill in your head how important distributions are for network science, so it pays off to become familiar with a few of them. First, let's make an important distinction. There are two kinds of distributions, depending on the kinds of values that their underlying variables can take. There are discrete distributions – for instance, the distribution of the number of ice cream cones different people ate on a given day. And there are continuous distributions – for instance the distances you rode on your bike on different days. The difference is that the former has specific values that the underlying variable can take (you may have eaten two or three ice cream cones, but 2.5 is not an option), the latter can take any real value as an outcome. In the first discrete case, we call the distribution a “mass function”. In the second case, we call it a “density function”.

Figure 3.6 shows some stylized representations of the most important distributions you should pay attention to, which are:

- **Uniform:** in this distribution each event is equally likely. This distribution can be both discrete or continuous. In the discrete case, if you have  $n$  possible events, each occurs with probability  $p = 1/n$ . You get a discrete uniform distribution if you look at the number on a ball extracted from an urn, where all balls in the urn are identified by distinct, progressive numbers without gaps. A continuous uniform distribution could be the amount of time you have to wait for the next drop to come from a leaky faucet that drips once a minute: if you haven't seen the last drop falling, the wait time could be any time between 0 and 60 seconds.



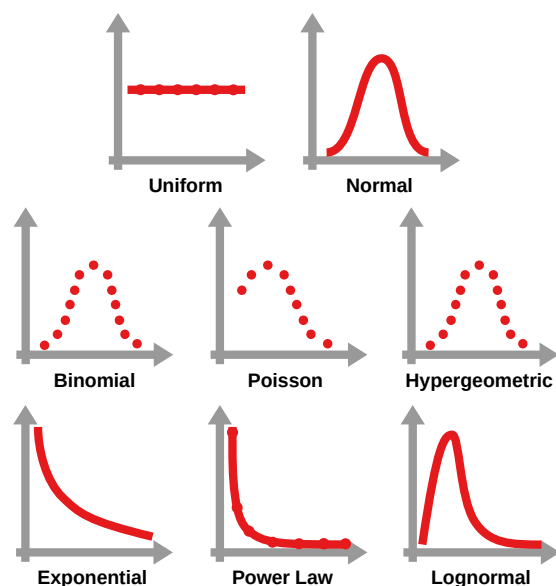


Figure 3.6: A stylized representation of the most common distributions you'll encounter as a network scientist. Solid lines show continuous distributions, while dots show discrete ones – note that some distributions can be both (dotted lines).

- **Normal (or Gaussian):** this is a very common distribution for continuous variables. The classical example is the distribution of people's heights: most people are of average height, and larger and larger deviations from the average get steadily less likely.
- **Binomial:** this is a discrete distribution, in which you do  $n$  experiments, each with success probability  $p$ , and you calculate the probability of having  $n'$  successes. For instance, suppose you take five balls from an urn containing 50 red and 50 green balls – each time putting the ball you extracted back into the urn. Let's say we count getting a green ball as a "success" here. The number of times you got 0, 1, 2, 3, 4 or 5 green balls over a bunch of trials would follow a binomial distribution.
- **Hypergeometric:** this is yet another discrete probability function. It is very similar to a binomial distribution. Where the binomial described the number of "successes" in an extraction-with-replacement urn game, the hypergeometric describes the more common case of extraction-without-replacement. When you extract a ball from the urn, you don't put it back. It is mathematically less tractable, but much more useful. This is used especially for the task of network backboning (Chapter 27).
- **Poisson:** this is another discrete distribution, which is the number of successes in a given time interval, assuming that each success arrives independently from the previous ones. For instance, the number of meteorites impacting on the moon each year will have a Poisson distribution. Interestingly, many examples commonly

mentioned for explaining a Poisson distribution (number of admittances to a hospital in an hour, number of emails written in an hour, and so on) aren't actually Poisson distributions, because of the "burstiness" of human behavior<sup>9</sup>.

- **Exponential:** the exponential distribution is a continuous distribution modeling cases in which the probability of something happening is not dependent of how much time has passed since you starting observing the phenomenon. For instance, if there's an epidemics out, the amount of time you have been infection-free bears no weight in determining your probability of being infected, if exposed. This is the reason why this distribution is sometimes called "memoryless" or that it "doesn't age".
- **Power law:** a power law can be both a discrete or a continuous distribution. It describes the relationship between two quantities, the second quantity changes as a power of the first. One practical consequence is that, if you were given a power law plot without axis labels, you would not be able to tell where you are in the distribution, because the slope of the line always looks the same no matter how much you zoom in or out, or whether you're on the head or the tail. An example of discrete power law is Zipf's law<sup>10</sup> recording the frequency of words in a document against their frequency rank. We'll see more than you want to know about power laws when talking about fitting degree distributions in Section 9.3.
- **Lognormal:** a lognormal distribution is the distribution of a continuous random variable whose logarithm follows a normal distribution – meaning the logarithm of the random variable, not of the distribution. This is the typical distribution resulting from the multiplication of two independent random positive variables. If you throw a dozen 20-sided dice and multiply the values of their faces up, you'd get a lognormal distribution. It's very tricky to tell this distribution apart from a power law, as we'll see.

Sometimes, rather than looking at the mass/density functions, it's more useful to look at their cumulative versions. In practice, you want to ask yourself what is the number of – say –  $x$  or *fewer* successes. Each distribution changes in predictable ways, as Figure 3.7 shows.

For instance, a cumulative uniform distribution is a line that goes straight up, because each event adds the same value to the cumulative sum. A cumulative normal distribution has an the shape of a flattened "S". In the power law case, as we'll see, we actually want to see the complement of the cumulative distribution (1 - CDF). This is, interestingly, also a power law.

<sup>9</sup> Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207, 2005

<sup>10</sup> Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005b

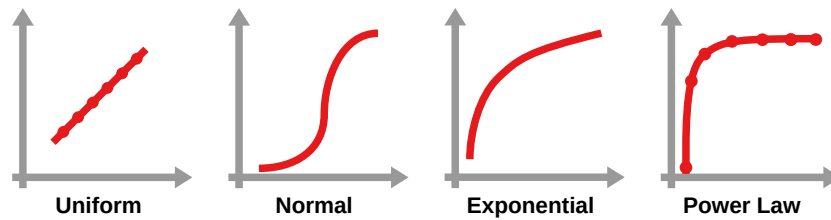


Figure 3.7: A stylized representation of a few cumulative distributions. Same legend as Figure 3.6.

### 3.3 *p*-Values

One of the key tasks of statistics is figuring out whether what you’re observing – a natural phenomena or the result of an experiment – can tell us something bigger about how the world works. The way this is normally done is making an hypothesis, for instance that a specific drug will cause weight loss. To figure out whether it is true, we need to prove that taking the drug actually does something rather than nothing. “The drug does nothing” is what we call the *null hypothesis*, which is what we’d expect – after all, most drugs don’t cause weight loss. This is what we colloquially call “burden of proof”: the person making the claim that something exists needs to prove that it does, because if we haven’t proven that something exists yet there is no reason to believe it does.

We call what you want to prove – “the drug causes weight loss” – the *alternative hypothesis*, because it’s the alternative to the null hypothesis.

p-values are among of the most commonly used tools to deal with this problem<sup>11</sup>. The interpretation of p-values is tricky and it is easy to get it wrong. The “p” stands for “probability”. Suppose that you give your drug to a bunch of people and, after a few weeks, you see that their weight decreased by 5kg. The p-value tells you the probability that you would be observing an effect this strong – a loss of 5kg – if the null hypothesis was true – i.e. if the drug actually did nothing. Lower p-values mean there is stronger evidence against the null hypothesis. What the p-value does not tell you (but might trick you into thinking it does) is:

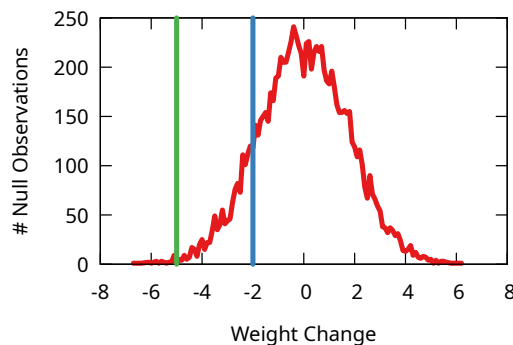
- The p-value does **NOT** tell you how likely you are to be right;
- The p-value does **NOT** tell you how strong the effect is;
- The p-value does **NOT** tell you how much evidence you have against your hypothesis.

Etch these bullet points into your brain, because it is so easy to fool yourself. The last of them means that a high p-value does not

<sup>11</sup> Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016

mean that the null hypothesis is true. A high p-value just means that the observations we have are compatible with a world where the null hypothesis is true. But it could also mean that our sample is not big enough to draw firm conclusions. Given how tricky it is to get them right, some researchers have called for not using p-values altogether<sup>12</sup>.

Figure 3.8 shows a graphical way to understand the p-value. You have a distribution of values that would be produced by the null hypothesis, you pit your measurements against those values, and the more unusual your observations look compared to those that would be produced by the null hypothesis, the lower the p-value. Exactly how to produce this null hypothesis distribution is not something we'll cover here, because it is not so close to the core of network science – although we'll see something similar in Section 19.1.



<sup>12</sup> Raymond Hubbard and R Murray Lindsay. Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1):69–88, 2008

Figure 3.8: In red we have the number of observations (y axis) with a given weight change (x axis) under the null hypothesis. In blue we have an observation of 2kg weight loss after taking the drug ( $p = 0.14$ ). In green we have an observation of 5kg weight loss after taking the drug ( $p = 0.003$ ).

One thing is worth mentioning, though. You will often see some magical p-value thresholds that people use as standards – the most common being  $p < 0.05$  and  $p < 0.01$ . These p-value thresholds say something about the strength of the evidence we want to see before we are willing to reject the null hypothesis of no effect. Beware of these. Not only because – as I said before – they don't mean what you think they mean, but also because of Goodhart's Law<sup>13</sup>. If we say  $p < 0.01$  is the gold standard we need to achieve to publish a paper, the natural tendency would be to try and repeat/modify experiments until we get  $p < 0.01$ . This is known in the literature as p-hacking<sup>14</sup> and has led researchers to publish a flurry of false results.

If the p-value tells you the probability of observing a given result under the null hypothesis, then if you repeat your experiments 100 times the probability that at least one of them will leave to a p-value  $\leq 0.01$  is actually 63%!<sup>15</sup> That is because the probability of getting a  $p > 0.01$  is 99% – with this standard of evidence, one percent of the time, you will reject the null hypothesis by accident. You try 100 times, so the formula to know how likely a  $p \leq 0.01$  is

<sup>13</sup> "When a measure becomes a target, it ceases to be a good measure."

<sup>14</sup> Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106, 2015

<sup>15</sup> <https://xkcd.com/882/>

becomes  $1 - (0.99^{100})$ . Sometimes, of course, you do need to run more than one test, and look at more than one p-value. A couple of common options to deal with this are applying the Bonferroni<sup>16,17</sup> or the Holm-Bonferroni<sup>18</sup> corrections to your p-values, systematically lowering the significance threshold to make up for “cheating” by running multiple tests.

p-values are important for network science because we will often have to create null models to figure out whether some property we observe in a network is actually interesting (Section 19.1). Another case is figuring out whether an edge has a weight significantly different from zero (i.e. it actually exists) or not (Chapter 27).

### 3.4 Correlation Coefficients

So far we have worked with a single variable and we have done our best to describe how it distributes. However, more often than not, you have more than one variable and you want to know something interesting about how one relates to the other. For instance, you might want to describe how the weight of a person tends to be related to their height.

In general, the taller a person is, the more we expect them to weigh. In this sense the two variables **vary together**. Therefore, we call this concept “**covariance**”<sup>19</sup>. The formula is pretty simple, it is the mean of the product of the deviations from the mean of both variables, so:  $cov(H, W) = \mu((H - \mu(H))(W - \mu(W)))$ .

If a person is both a lot taller than the mean and a lot heavier than the mean, the product of their height and weight deviations will be big, and they will contribute a large positive value to the covariance calculation. If a person is both a lot shorter than the mean and a lot lighter than the mean, the product of their height and weight deviations, both negative, will again be a large positive. So, they will also make the covariance turn out bigger. If someone is both tall and light, their positive height deviation and negative weight deviation will be multiplied to become a large negative number, pulling the covariance down. The covariance will be large in total if we observe many tall/heavy and short/light people, and not so many tall/light or short/heavy people (see Figure 3.9, the covariance values are in the figure’s caption, first row under  $cov(H, W)$ ).

One problem is that the covariance depends on the units of your variables. That means the size of the covariance can be difficult to interpret – is a covariance of 5.5 high or low? To make sense of it, it helps to normalize it. This is what the Pearson correlation coefficient tries to do<sup>20</sup>. It divides the covariance by the variances of both variables. That means, for example, that if start measuring

<sup>16</sup> Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936

<sup>17</sup> Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293): 52–64, 1961

<sup>18</sup> Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979

<sup>19</sup> John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006

<sup>20</sup> Francis Galton. Typical laws of heredity. Royal Institution of Great Britain, 1877

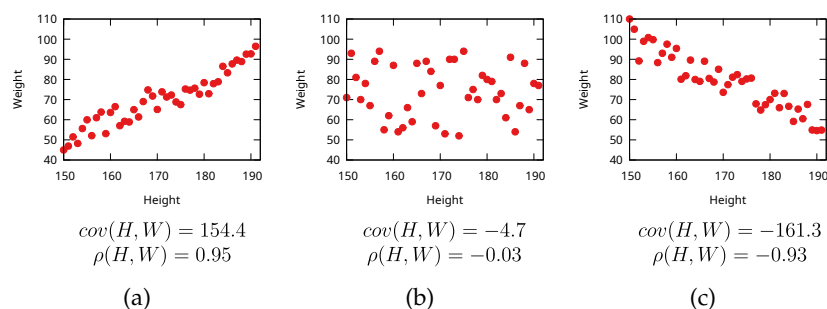


Figure 3.9: Three examples of datasets with (a) positive, (b) no, and (c) negative covariance between height and weight. Covariance ( $cov(H, W)$ ) and correlation ( $\rho(H, W)$ ) values below the scatter plots.

height in meters rather than centimeters, its covariance with weight will change, but its Pearson correlation with weight will stay the same. The Pearson correlation can only take values between  $-1$  (perfect anticorrelation) and  $+1$  (perfect correlation).

The Pearson correlation coefficient is nice and it is used for many things – including in network science to predict links (Section 23.7), project bipartite networks (Section 26.2), estimate assortativity (Section 31.1), ... you get the idea. There are a couple of problems with the Pearson correlation. The first is that it only estimates how monotone a relationship is: this means that it wants variables to always vary together in the same general direction. Figure 3.10 shows a couple of non-monotone relationships – the classic “U” and “inverted U” shapes. In this case, even if there clearly is a relationship, Pearson will return zero and there isn’t much you can do about it.

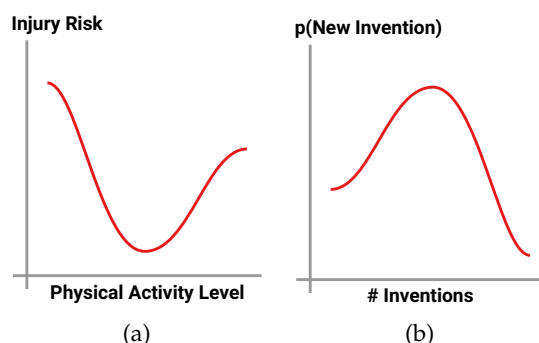


Figure 3.10: Two non monotone relationships (a) U-shaped: if you never exercise you’re frail and prone to injury, and if you exercise too much you have more chances to injury yourself. (b) Inverted U-shaped (or A-shaped): if there are no inventions innovation is hard, if there are too many innovations there’s nothing left to innovate.

The second issue is that, even if the variables have a roughly monotone relationship, Pearson only measures it accurately if this relationship is *linear*. Pearson will give us the wrong idea if increases in one variable are associated with changes in another variable, but less and less so. For example, we visit our friends less often if they live further away from us, but moving 20 kilometers away if they lived next door before could make a big difference, whereas moving 20 extra kilometers away if they already lived in another country is not so meaningful.

One neat trick you can often do to fix this is to do a log-transformation

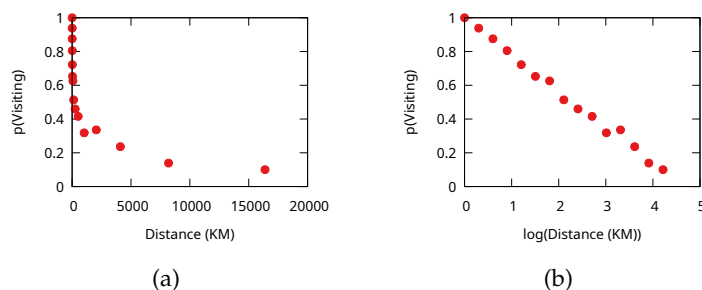


Figure 3.11: Data with skewed distributions across various orders of magnitude. (a) Linear plot. (b) Same data, with the x axis log-transformed.

of your data before you calculate a correlation. In Figure 3.11(a) you can see a pair of variables with a non-linear relationship – the probability of deciding to visit a friend living at a given distance. Once you take the logarithm of the distance (Figure 3.11(b)), a linear relationship with the visit probability comes to the surface. Often, you can log both variables.

Most of the times you can stop here, but when this fails you still have one tool that allows to calculate correlations of data related in just about any way. As long as a change in one variable monotonically correspond to a change in the other (so, no U- or A-shapes where the direction of the relationship changes midway), you can summarize this with the Spearman rank correlation<sup>21</sup>. You do not assume anything at all about *how much* each variable changes as the other changes. The Spearman rank correlation is still normalized to take values between  $-1$  and  $+1$  with the exact same meaning as Pearson.

How can you achieve this? It's actually rather simple. You still calculate a Pearson correlation. But rather than calculating it directly on the values of your variable, you do it on their *ranks*. The observation with the highest value becomes a 1, the second largest becomes a 2, and so on. Then, you calculate the Pearson correlation of these ranks. That is why this is called the Spearman *rank* correlation.

Figure 3.12 provides some examples. You can see what a monotonic but not linear relationship looks like (Figure 3.12(a)), that for data fitting the Pearson's assumptions Spearman returns comparable values (Figure 3.12(b)), and the robustness of Spearman to outliers (Figure 3.12(c)).

### 3.5 Mutual Information

Mutual Information (MI) is a key concept in information theory. It is another measure of how related two variables are. You can see it as a sort of a special correlation. Formally, MI quantifies how much information you obtain about one variable if you know the value of the other variable. For example, if we know how tall someone is,

<sup>21</sup> Charles Spearman. The proof and measurement of association between two things. *Am J Psychol*, 15:72–101, 1904



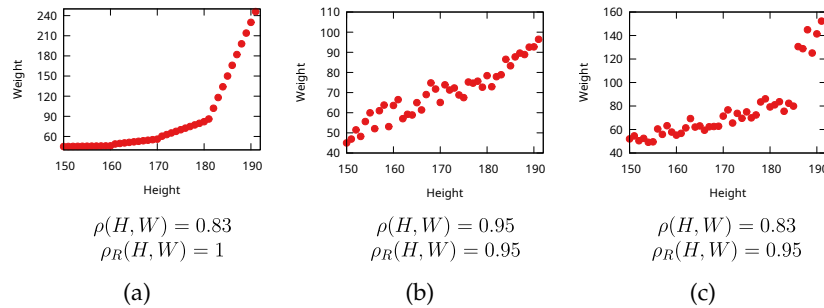
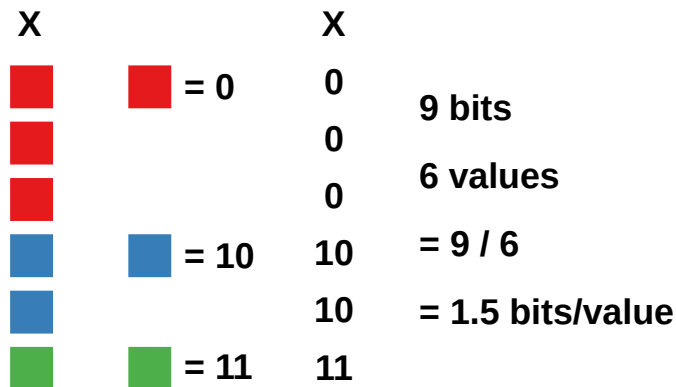


Figure 3.12: Three pairs of variables showing the difference between the Pearson ( $\rho(H, W)$ ) and the Spearman ( $\rho_R(H, W)$ ) correlation coefficients.

does this allow us to make a better guess at their weight? How much better? This “amount of information” is usually measured in bits.

To understand MI, we need to take a quick crash course on information theory<sup>22,23</sup>, which starts with the definition of information entropy. It is a lot to take in, but we will extensively use these concepts when it comes to link prediction and community discovery in Parts VII and X, thus it is a worthwhile effort.

Consider Figure 3.13. The figure contains a representation of a vector of six elements that can take three different values. The first thing we want to know is how many bits of information we need to encode its content.



We can be smart and use the shortest codes for the elements that appear most commonly, in this case the red square. Every time we see a red square, we encode it with a zero. If we don't see a red square, we write a one, which means that we need to look at a second bit to know whether we saw a blue or a green square. If it was a blue square, we write a zero, if it was green we write another one. With these rules, we can encode the original vector using nine bits, i.e. we use 1.5 bits per element.

This is close to – but not exactly – the definition of information entropy. In information entropy, we weigh the probability of an event by its logarithm<sup>24,25</sup>.

<sup>22</sup> Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999

<sup>23</sup> David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003

Figure 3.13: A simple example to understand information entropy. From left to right: the vector  $x$  has six elements taking three different values. We can encode each value with a sequence of zeros and ones. Doing so allows us to transmit  $x$ 's six elements using nine bits of information. This means that the number of bits per value is 1.5.

<sup>24</sup> Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948

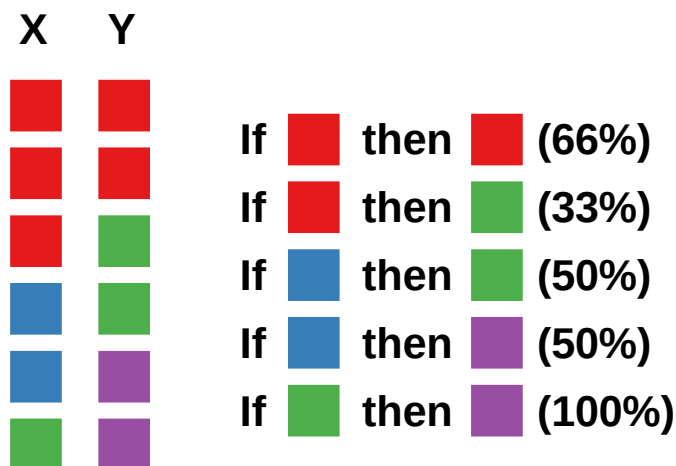
<sup>25</sup> Andrei N Kolmogorov. Three approaches to the quantitative definition of information'. *Problems of information transmission*, 1(1):1–7, 1965

Consider flipping a coin. Once you know the result, you obtain one bit of information. That is because there are two possible events, equally likely with a probability  $p$  of 50%. Generalizing to all possible cases, every time an event with probability  $p$  occurs, it gives you  $-\log_2(p)$  bits of information for... reasons<sup>26</sup>. So, the total information of an event is the amount of information you get per occurrence times the probability of occurrence:  $-p \log_2(p)$ . Summed over all possible events  $i$  in  $x$ :  $H_x = -\sum_i p_i \log_2(p_i)$ , which is Shannon's information entropy – how many bits you need to encode the occurrence of all events.

Mutual information is defined for two variables. As I said, it is the amount of information you gain about one by knowing the other, or how much entropy knowing one saves you about the other. Consider Figure 3.14. It shows the relationship between two vectors,  $x$  and  $y$ . Note how  $y$  has equally likely outcomes: each color appears three times. However, if we observe a green square in  $x$ , we know with 100% confidence that the corresponding square in  $y$  is going to be purple. This means that, knowing  $x$ 's values gives us information about  $y$ 's value. Mathematically speaking, mutual information is the amount of information entropy shared by the two vectors.

It would take  $-\log_2(1/3) \sim 1.58$  bits to encode  $y$  on its own (it is a random coin with three sides). However, knowing  $x$ 's values makes you able to use the inference rules we see in Figure 3.14. Those rules are helpful: note how our confidence is almost always higher than 33%, which is the probability of getting  $y$ 's color right without any further information. The rules will save you around 0.79 bits, which is  $x$  and  $y$ 's mutual information.

The exact formulation of mutual information is similar to the formula of entropy:



<sup>26</sup> The amount of information of an event is a function that only depends on the probability  $p$  of the event to happen, e.g.  $i_a = f(p_a)$  for event  $a$ . If we have two events,  $a$  and  $b$ , happening with probability  $p_a$  and  $p_b$ , the event  $c$  defined as  $a$  and  $b$  happening has probability  $p_c = p_a p_b$ . Now, each event also gives you an amount of information, namely  $i_a$  and  $i_b$ . When  $c$  happens, it means that both  $a$  and  $b$  happened, thus you got both pieces of information, or  $i_c = i_a + i_b$ . What we just said can be rewritten as  $f(p_c) = f(p_a) + f(p_b)$ , given the equation at the beginning. Since  $p_c = p_a p_b$ , then we can also rewrite the equation as  $f(p_a p_b) = f(p_a) + f(p_b)$ . The only function  $f$  that we can possibly plug into this equation maintaining it true is the logarithm. Since probabilities are lower than 1, the logarithm would be lower than zero, which would be nonsense – you cannot get negative information. Thus we take the negated logarithm:  $i_a = -\log(p_a)$ .

Figure 3.14: An illustration of what mutual information means for two vectors. Vector  $y$  has equal occurrences for its values (there is one third probability of any colored square). However, if we know the value of  $x$  we can usually infer the corresponding  $y$  value with a higher than chance confidence.

$$MI_{xy} = \sum_{j \in y} \sum_{i \in x} p_{ij} \log \left( \frac{p_{ij}}{p_i p_j} \right),$$

where  $p_{ij}$  is the joint probability of  $i$  and  $j$ . The meat of this equation is in comparing the joint probability of  $i$  and  $j$  happening with what you would expect if  $i$  and  $j$  were completely independent. If they are, then  $p_{ij} = p_i p_j$ , which means we take the logarithm of one, which is zero. But any time the happening of  $i$  and  $j$  is not independent, we add something to the mutual information. That something is the number of bits we save.

### 3.6 Summary

1. If you want to know what an average observation looks like in a variable, you might want to calculate the mean or the median.
2. The average observation will deviate from the mean: to estimate how far from the mean it will be on average, you can calculate the variance or the standard deviation.
3. Data can be skewed, meaning it distributes asymmetrically around the mean. There could be long or fat tails, depending how extreme and uncommon outliers are. The more skewed your data is, the more mean and median will disagree.
4. There are important distributions one should know: uniform, where all values are equally likely; normal, where values cluster around the mean; and various skewed distributions such as exponential, lognormal, and power law.
5. A cumulative distribution tells you the probability of observing a values equal to or smaller than a given threshold.
6. The p-value tells you the probability of making a given observation under the null hypothesis (no change, no effect, nothing interesting is happening). A low p-value can prove the null hypothesis wrong but a high p-value does not prove the null hypothesis is true.
7. If you do multiple experiments, you need to correct the p-values you get otherwise you are going to misinterpret them.
8. When you have two variables, covariance tells you how much the two change together. Correlation coefficients are normalized covariances that do not change depending on the scale of the data. If your variables have a non-linear relationship, you need to use a correlation coefficient that can handle it.

9. Another approach to measure of how related two random variables are is mutual information. It tells you how many bits of information you gain about the status of one variable by knowing the other.

### 3.7 Exercises

1. Calculate the mean, median, and standard deviation of the two variables at <http://www.networkatlas.eu/exercises/3/1/data.txt> (one variable per column).
2. Make a scatter plot of the variables used in the previous exercise – with one variable on the x axis and the other on the y axis. Do you think that they are skewed or not? Calculate their skewness to motivate your answer.
3. Draw the mass function and the cumulative distribution of the following outcome probabilities:

Outcome	$p$
1	0.1
2	0.15
3	0.2
4	0.21
5	0.17
6	0.09
7	0.06
8	0.02

4. Which correlation coefficient should you use to calculate the correlation between the variables used in the exercise 2? Motivate your answer by calculating covariance, and the Pearson and Spearman correlation coefficients (and their p-values). Does the Spearman correlation coefficient agree with the Pearson correlation calculated on log-transformed values?
5. How many bits do we need to independently encode  $v_1$  and  $v_2$  from <http://www.networkatlas.eu/exercises/3/5/data.txt>? How much would we save in encoding  $v_1$  if we knew  $v_2$ ?