Probability theory

Michael Franke

Basics of probability theory: axiomatic definition, interpretation, joint distributions, marginalization, conditional probability, Bayes rule, stochastic independence. Random variables & expected values.

Classical logic explores which conclusions follow from a set of premises. The conclusion must follow necessarily from the premises, based on the logical structure of the premises and the conclusions, not based on additional knowledge about the actual world, not on things that could (logically) have been different.

In contrast, much of human knowledge and reasoning revolves around *statistical knowledge*: since most birds can fly, if I learn that Tweety is a bird, it is reasonable to conclude that Tweety is *likely* to fly (unless I have more information about Tweety that might provide evidence against this uncertain inference). Probability theory is a formal framework to capture such *reasoning under uncertainty*.

Just like there are several logics, there are also several formalizations for reasoning with uncertainty, some of which are simpler and some of which are way more complex than standard probability theory. Some alternative systems are argued to be more empirically adequate for capturing human reasoning than probability theory. However, what singles out probability theory is that it strikes a good balance between simplicity, adequacy and applicability. As such, it lies at the heart of much of modern statistics and machine learning, with a plethora of mathematical results and algorithms supporting its wide-spread applications in all areas of science.

1 Probability

The most central concept of probability theory is that of a probability distribution. A probability distribution captures a state of uncertainty regarding which of a number of relevant events might hold or will occur. It assigns a number to each relevant event. This number indicates how likely the event is supposed to be (relative to others).

1.1 Outcomes, events, observations

We are interested in the space Ω of all *elementary outcomes* $\omega_1, \omega_2, \ldots$ of a process or event whose execution is (partially) random or unknown. Elementary outcomes are mutually exclusive. The set Ω exhausts all possibilities.[1]

Example 1. The set of elementary outcomes of a single coin flip is $\Omega_{\text{coin flip}} = \{\text{heads}, \text{tails}\}$. The elementary outcomes of tossing a six-sided die is $\Omega_{\text{standard die}} = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.[2]

[1] For simplicity of exposure, we gloss over subtleties arising when dealing with infinite sets Ω.

[2] Think of Ω as a partition of the space of all possible worlds, i.e., ways in which the world could be, where we lump together into one partition cell all ways in which the world could be that are equivalent regarding those aspects of reality that we are interested in. We do not care whether the coin lands in the mud or in the sand. It only matters whether it came up heads or tails. Each elementary event can be realized in myriad ways. Ω is our, the modellers', first crude simplification of nature, abstracting away

An *event* A is a subset of Ω. Think of an event as a (possibly partial) observation. We might observe, for instance, not the full outcome of tossing a die, but only that there is a dot in the middle. This would correspond to the event $A = \{\boxdot, \boxdot, \boxdot\} \subseteq \Omega_{\text{standard die}}$, i.e., observing an odd numbered outcome. The trivial observation $A = \Omega$ and the impossible observation $A = \emptyset$ are counted as events, too. The latter is included for technical reasons.

For any two events $A, B \subseteq \Omega$, standard set operations correspond to logical connections in the usual way. For example, the conjunction $A \cap B$ is the observation of both A and B; the disjunction $A \cup B$ is the observation that it is either A or B; the negation of A, $\overline{A} = \{\omega \in \Omega \mid \omega \notin A\}$, is the observation that it is not A.

1.2 *Probability distribution*

A *probability distribution* P over Ω is a function $P : \mathcal{P}(\Omega) \to \mathbb{R}$ that assigns to all events $A \subseteq \Omega$ a real number (from the unit interval, see A1), such that the following (so-called Kolmogorov axioms) are satisfied:

A1. $0 \leq P(A) \leq 1$

A2. $P(\Omega) = 1$

A3. $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$
whenever $A_1, A_2, A_3, \dots$ are mutually exclusive[3]

[3] A3 is the axiom of *countable additivity*. Finite additivity may be enough for finite or countable sets Ω, but infinite additivity is necessary for full generality in the uncountable case.

Occasionally we encounter notation $P \in \Delta(\Omega)$ to express that P is a probability distribution over Ω.[4] If $\omega \in \Omega$ is an elementary event, we often write $P(\omega)$ as a shorthand for $P(\{\omega\})$. In fact, if Ω is finite, it suffices to assign probabilities to elementary outcomes.

[4] E.g., in physics, theoretical economics or game theory. Less so in psychology or statistics.

A number of useful rules follows immediately from of this definition. Here we prove one (see exercises for more).

Claim 2. If $\Omega = \{\omega_1, \dots, \omega_n\}$ is a finite set of elementary outcomes and $P \in \Delta(\Omega)$ a probability distribution over Ω, then the sum of the probabilities of all elementary outcomes is equal to 1:

$$\sum_{i=1}^{n} P(\omega_i) = 1$$

Proof. From A3 we know that:

$$\sum_{i=1}^{n} P(\omega_i) = P(\{w_1\} \cup \dots \cup \{w_n\})$$

Since $\Omega = \{w_1\} \cup \dots \cup \{w_n\}$, it follows from A2 that $\sum_{i=1}^{n} P(\omega_i) = 1$. □

It follows from Claim 2 that, in order to fully determine a probability distribution $P \in \Delta(\Omega)$ over a finite Ω with n elements, we only need to specify $n - 1$ probabilities, since the nth probability can be retrieved as "one minus the sum of all others."

1.3 Interpretations of probability

It is reasonably safe, at least preliminarily, to think of probability, as defined above, as a handy mathematical primitive which is useful for certain applications. There are at least three ways of thinking about where this primitive probability might come from, roughly paraphrasable like so:

1. *Frequentist:* Probabilities are generalizations of intuitions/facts about frequencies of events in repeated executions of a random event.
2. *Subjectivist:* Probabilities are subjective beliefs by an agent who is uncertain about the outcome of a random event.
3. *Realist:* Probabilities are a property of an intrinsically random world.

While trying to stay away from philosophical quibbles, we will adopt a subjectivist interpretation of probabilities, since this interpretation is most encompassing and —arguably— intuitive. But note that frequentist considerations should affect what a rational agent should believe (see the urns scenario in Section 1.6).

Example 3 (Subjective beliefs about the weather). Consider the set of elementary outcomes $\Omega_{\text{weather}} = \{\text{sunny}, \text{misty}, \text{rainy}\}$ of potential weather condition for tomorrow at noon.[5] Jones, the optimist, does not know what the weather will bring, but believes that it is most likely to be sunny. In fact, Jones believes —for whatever reason— that it is three times as likely to be sunny than that it is going to be misty. Jones also believes that being misty and being rainy is equallly likely. This information about Jones' *relative degrees of credence* alone, is enough to know that, according to Jones, the probability of the three weather conditions are given in the first line of Table 1.

[5] We assume that these states are independent and that these are all the states the weather might be in (for simplicity of an example).

Smith is more pessimistic. Smith's believes that "misty" is twice as likely as "sunny" and that "rainy" is seven times more likely than "sunny." This information about relative probabilities is enough to know that Smith's beliefs are those represented in the second line of Table 1.

Notice that, in the case of beliefs about the weather, it is fairly unproblematic to imagine that two agents, even rational ones, might have quite different (subjective) beliefs about the same set of elementary outcomes. This is not always the case (see Section 1.6).

	sunny	cloudy	rainy
Jones' beliefs	0.6	0.2	0.2
Smith's beliefs	0.1	0.2	0.7

Table 1: Subjective beliefs about the weather.

1.4 Odds & wheels of fortune

The previous example demonstrated how the probabilities of a space with three elementary outcomes, like $\Omega_{\text{weather}} = \{\text{sunny}, \text{misty}, \text{rainy}\}$, is com-

pletely determined by two numbers describing *relative probabilities*, so-called *odds*, e.g.:

$$\frac{P(\text{sunny})}{P(\text{rainy})} = o_1 \qquad \frac{P(\text{misty})}{P(\text{rainy})} = o_2$$

Indeed, odds are actually much more meaningful than absolute numbers for probabilities. A nice way to see this is to think of probabilities like those in Table 1 as the probabilities of outcomes on a wheel of fortune (see Figure 1). A wheel of fortune is spun and the outcome is determined by whatever area is on top (usually indicated by a marker or needle). The wheel of fortune that corresponds to Jones' beliefs, shown on the left-hand side in Figure 1, has areas that correspond to the three elementary outcomes. Notice that neither the absolute size of any area, nor the absolute length of the circumference which is covered by that area is important. In order for the wheel of fortunes in Figure 1 to match the probabilities in Table 1, what matters is only that the areas have the right proportion. In other words, relative areas (odds) are what matters most.

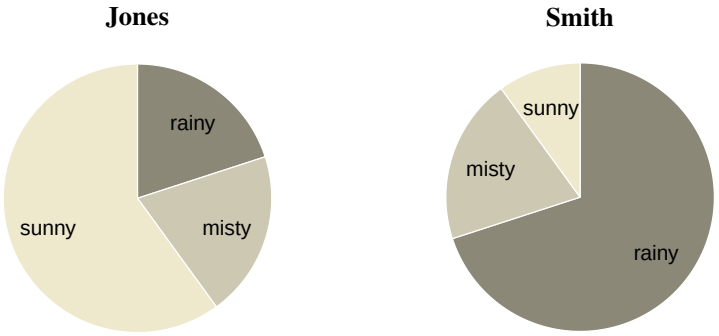

Figure 1: Two beliefs about the weather, represented as wheels of fortune.

1.5 Non-normalized probabilities

Since what matters most are odds, not absolute numbers, we can also specify probability distributions in terms of non-normalized probabilities. Concretely, the following is a complete and sufficient alternative way of specifying Jones' beliefs as in Table 1:

$$P(\text{sunny}) \propto 120 \qquad P(\text{misty}) \propto 40 \qquad P(\text{rainy}) \propto 40$$

Here, the operator $\propto$ (read: "proportional to") is used. More generally, if $f\colon \Omega \to \mathbb{R}^{\geq 0}$ is a function that maps each elementary outcome of Ω onto a non-negative real number, then writing $P(\omega) \propto f(\omega)$ fully and unambiguously defines a probability distribution in terms of the non-normalized probabilities assigned to each ω by f, namely:

$$P(\omega) = \frac{f(\omega)}{\sum_{\omega'} f(\omega')}$$

1.6 *Urns and frequencies*

Another way of thinking about probabilities for discrete sets Ω, is in terms of urns. Think of an urn as a container which contains a number of $N > 1$ balls. Balls can be of different color. For example, let us suppose that our urn has $k > 0$ black balls and $N - k$ white balls. (There is at least one black and one white ball.) For a single random draw from our urn we have: $\Omega_{\text{urn}} = \{\text{white}, \text{black}\}$. Figure 2 shows such an urn with $k = 7$ and $N = 10$.

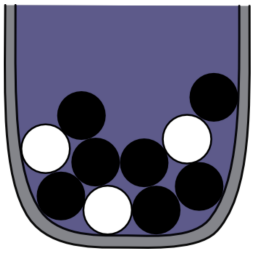

Figure 2: Urn with $k = 7$ black balls out of $N = 10$ balls in total.

Imagine a long sequence of single draws from the urn in Figure 2, putting whichever ball we drew back in after every draw. We keep a record of how many times we drew a black ball, and divide this number by the number of times we drew a ball. Figure 3 shows the results from a computer simulation of this process. In general, the limiting proportion with which we draw a black ball is $\frac{k}{N}$. Another way of saying this is that the *objective* probability is $P(\text{black}) = \frac{k}{N}$. Consequently, a rational agent's subjective beliefs should conform to the objective probability $P(\text{black}) = \frac{k}{N}$, unlike in other cases like the weather.

Figure 3: Temporal development of the proportion of drawing a black ball from the urn.

Exercise 1. Using the rules of probability theory, prove that the following claims hold.

C1. $P(\emptyset) = 0$

C2. $P(\overline{A}) = 1 - P(A)$

C3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any $A, B \subseteq \Omega$

Exercise 2. Write down the probability distributions over $\Omega_{\text{weather}} = \{\text{sunny}, \text{misty}, \text{rainy}\}$ that are defined in terms of the following pieces of information

(i) Miller believes that "rainy" is impossible, and that "sunny" is three times as likely as "misty."

(ii) Ford has beliefs given by the following non-normalized probabilities:

$$P(\text{sunny}) \propto 3 \qquad P(\text{misty}) \propto 9 \qquad P(\text{rainy}) \propto 27$$

(iii) Johnson believes that "sunny" is as likely as not and that the odds in favor of "rainy" over "misty" are 3 to 2.

2 *Structured events & marginal distributions*

2.1 *Probability table for a flip-&-draw scenario*

Suppose we have two urns. Both have $N = 10$ balls. Urn 1 has $k_1 = 2$ black and $N - k_1 = 8$ white balls. Urn 2 has $k_2 = 4$ black and $N - k_2 = 6$ white balls. We sometimes draw from urn 1, sometimes from urn 2. To decide, we flip a fair coin. If it comes up heads, we draw from urn 1; if it comes up tails, we draw from urn 2. A schematic representation of this *flip-&-draw scenario* is shown in Figure 4.

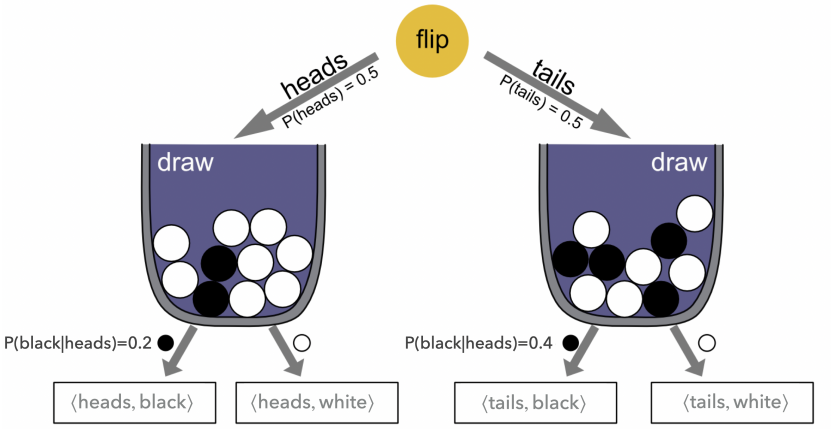

Figure 4: The flip-&-draw scenario, illustrating structured event spaces and conditional probabilities.

An elementary outcome of this two-step process of flip-&-draw is a pair ⟨outcome-flip, outcome-draw⟩. The set of all possible such outcomes is

$$\Omega_{\text{flip-\&-draw}} = \{\langle\text{heads, black}\rangle, \langle\text{heads, white}\rangle, \langle\text{tails, black}\rangle, \langle\text{tails, white}\rangle\}$$

The probability of event ⟨heads, black⟩ is given by multiplying the probability of seeing "heads" on the first flip, which happens with probability 0.5, and then drawing a black ball, which happens with probability 0.2, so that $P(\langle\text{heads, black}\rangle) = 0.5 \cdot 0.2 = 0.1$. The probability distribution over $\Omega_{\text{flip-draw}}$ is consequently as in Table 2.[6]

[6] If in doubt, start flipping & drawing and count your outcomes.

	black	white
heads	$0.5 \cdot 0.2 = 0.1$	$0.5 \cdot 0.8 = 0.4$
tails	$0.5 \cdot 0.4 = 0.2$	$0.5 \cdot 0.6 = 0.3$

Table 2: Probabilities of elementary outcomes (pairs of ⟨outcome-flip, outcome-draw⟩) in the flip-&-draw example.

2.2 *Structured events and joint-probability distributions*

Table 2 is an example of a *joint probability distribution* over a structured event space, which here has two dimensions. Since our space of outcomes is the Cartesian product of two simpler outcome spaces, namely $\Omega_{flip-\&-draw} = \Omega_{flip} \times \Omega_{draw}$,[7] we can use notation $P(\text{heads, black})$ as shorthand for $P(\langle\text{heads, black}\rangle)$.

[7] With $\Omega_{\text{flip}} = \{\text{heads, tails}\}$ and $\Omega_{\text{draw}} = \{\text{black, white}\}$.

More generally, if $\Omega = \Omega_1 \times \ldots \Omega_n$, we can think of $P \in \Delta(\Omega)$ as a joint probability distribution over n subspaces.

2.3 *Marginalization*

If P is a joint-probability distribution over event space $\Omega = \Omega_1 \times \ldots \Omega_n$, the *marginal distribution* over subspace Ω_i, $1 \leq i \leq n$ is the probability distribution that assigns to all $A_i \subseteq \Omega_i$ the probability:[8]

$$P(A_i) = \sum_{A_1 \subseteq \Omega_1, \ldots, A_{i-1} \subseteq \Omega_{i-1}, A_{i+1} \subseteq \Omega_{i+1}, \ldots, A_n \subseteq \Omega_n} P(A_1, \ldots, A_{i-1}, A_i, A_{i+1}, \ldots A_n)$$

[8] This notation, using $\sum$, assumes that subspaces are countable. In other cases, a parallel definition with integrals can be used.

For example, the marginal distribution over coin flips derivable from the joint probability distribution in Table 2 gives $P(\text{heads}) = P(\text{tails}) = 0.5$, since the sum of each row is exactly 0.5. The marginal distribution over flips derivable from Table 2 has $P(\text{black}) = 0.3$ and $P(\text{white}) = 0.7$.[9]

[9] The term "marginal distribution" derives from such probability tables, where traditionally the sum of each row/column was written in the margins.

3 *Conditional probability*

Fix probability distribution $P \in \Delta(\Omega)$ and events $A, B \subseteq \Omega$. The conditional probability of A given B, written as $P(A \mid B)$, gives the probability of A on the assumption that B is true.[10] It is defined like so:

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

[10] We also verbalize this as "the conditional probability of A conditioned on B."

Conditional probabilities are only defined when $P(B) > 0$.[11]

[11] Updating with events which have probability zero entails far more severe adjustments of the underlying belief system than just ruling out information hitherto considered possible. Formal systems that capture such *belief revision* are studied in formal epistemology.

Example 4. If a dice is unbiased, each of its six faces has equal probability to come up after a toss. The probability of event $B = \{\boxdot, \boxdot, \boxdot\}$ that the tossed number is odd has probability $P(B) = \frac{1}{2}$. The probability of event $A = \{\boxdot, \boxdot, \boxdot, \boxdot\}$ that the tossed number is bigger than two is $P(A) = \frac{2}{3}$. The probability that the tossed number is bigger than two *and* odd is $P(A \cap B) = P(\{\boxdot, \boxdot\}) = \frac{1}{3}$. The conditional probability of tossing a number that is bigger than two, when we know that the toss is even, is $P(A \mid B) = \frac{1/3}{1/2} = \frac{2}{3}$.

Algorithmically, conditional probability first rules out all events in which B is not true and then simply renormalizes the probabilities assigned to the remaining events in such a way that the relative probabilities of surviving events remains unchanged. Given this, another way of interpreting conditional probability is that $P(A \mid B)$ is what a rational agent *should* believe about A after observing that B is in fact true and nothing more. The agent rules out, possibly hypothetically, that B is false, but otherwise does not change opinion about the relative probabilities of anything that is compatible with B.

3.1 Bayes rule

Looking back at the joint-probability distribution in Table 2, the conditional probability $P(\text{black} \mid \text{heads})$ of drawing a black ball, given that the initial coin flip showed heads, can be calculated as follows:

$$P(\text{black} \mid \text{heads}) = \frac{P(\text{black}, \text{heads})}{P(\text{heads})} = \frac{0.1}{0.5} = 0.2$$

This calculation, however, is quite spurious. We knew that already from the way the flip-&-draw scenario was set up. After flipping heads, we draw from urn 1, which has $k = 2$ out of $N = 10$ black balls, so clearly: if the flip is heads, then the probability of a black ball is 0.2. Indeed, in a step-wise random generation process like the flip-&-draw scenario, some conditional probabilities are very clear, and sometimes given by definition. These are, usually, the conditional probabilities that define how the process unfolds forward in time, so to speak.

Bayes rule is a way of expressing, in a manner of speaking, conditional probabilities in terms of the "reversed" conditional probabilities:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

Bayes rule follows directly from the definition of conditional probabilities, according to which $P(A \cap B) = P(A \mid B) \cdot P(B)$, so that:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

Bayes rule allows for reasoning backwards from observed causes to likely underlying effects. When we have a feed-forward model of how unobservable effects probabilistically constrain observable outcomes, Bayes rule allows us to draw inferences about *latent/unobservable variables* based on the observation of their downstream effects.

Consider yet again the flip-&-draw scenario. But now assume that Jones flipped the coin and drew a ball. We see that it is black. What is the probability that it was drawn from urn 1, equivalently, that the coin landed heads? It is not $P(\text{heads}) = 0.5$, the so-called *prior probability* of the coin landing heads. It is a conditional probability, also called the *posterior probability*,[12] namely $P(\text{heads} \mid \text{black})$, but one that is not as easy and straightforward to write down as the reverse $P(\text{black} \mid \text{heads})$ of which we said above that it is an almost trivial part of the set up of the flip-&-draw scenario. It is here that Bayes rule has its purpose:

$$P(\text{heads} \mid \text{black}) = \frac{P(\text{black} \mid \text{heads}) \cdot P(\text{heads})}{P(\text{black})} = \frac{0.2 \cdot 0.5}{0.3} = \frac{1}{3}$$

This result is quite intuitive. Drawing a black ball from urn 2 (i.e., after seeing tails) is twice as likely as drawing a black ball from urn 1 (i.e., after seeing heads). Consequently, after seeing a black ball drawn, with equal probabilities of heads and tails, the probability that the coin landed tails is also twice as large as that it landed heads.

[12] The terms *prior* and *posterior* make sense when we think about an agent's belief state before (prior to) and after (posterior to) an observation.

3.2 Stochastic (in-)dependence

Event A is *stochastically independent* of B if, intuitively speaking, learning B does not change one's beliefs about A, i.e., $P(A \mid B) = P(A)$.

Claim 5. If A is stochastically independent of B, then B is stochastically independent of A.

Proof.

$$\begin{aligned} P(B \mid A) &= \frac{P(A \mid B)\, P(B)}{P(A)} && \text{[Bayes rule]} \\ &= \frac{P(A)\, P(B)}{P(A)} && \text{[by ass. of independence]} \\ &= P(B) && \text{[cancellation]} \end{aligned}$$

□

For example, imagine a flip-and-draw scenario like in Figure 4 where the initial coin flip has a bias of 0.8 towards heads, but each of the two urns has the same number of black balls, namely 3 black and 7 white balls. Intuitively and formally, the probability of drawing a black ball is then *independent* of the outcome of the coin flip; learning that the coin landed heads, does not change our beliefs about how likely the subsequent draw will result in a black ball. The probability table for this example is in Table 3.

	heads	tails	$\sum$ rows
black	$0.8 \times 0.3 = 0.24$	$0.2 \times 0.3 = 0.06$	0.3
white	$0.8 \times 0.7 = 0.56$	$0.2 \times 0.7 = 0.14$	0.7
$\sum$ columns	0.8	0.2	

Table 3: Joint probability table for a flip-and-draw scenario where the coin has a bias of 0.8 towards heads and where each of the two urns holds 3 black and 7 white balls.

Independence shows in Table 3 in the fact that the probability in each cell is the product of the two marginal probabilities. This is a direct consequence of stochastic independence:

Claim 6 (Probability of conjunction of stochastically independent events). For any pair of events A and B with non-zero probability:

$$P(A \cap B) = P(A)\, P(B) \quad \text{[if } A \text{ and } B \text{ are stoch. independent]}$$

Proof. By assumption of independence, it holds that $P(A \mid B) = P(A)$. But then:

$$\begin{aligned} P(A \cap B) &= P(A \mid B)\, P(B) && \text{[def. of conditional probability]} \\ &= P(A)\, P(B) && \text{[by ass. of independence]} \end{aligned}$$

□

Exercise 3. Consider a flip-&-draw scenario like the process in Figure 4, but with a biased coin that lands heads with probability 0.7. Assume further that the "heads-urn" (the urn to draw from after a "heads" outcome) has 25 balls in total out of which 10 are black, and that the "tails-urn" has 30 balls in total out of which 20 are black.

(i) Calculate the joint probability of all elementary outcomes (pairs of flips and draws).

(ii) Compute the marginal probabilities of "heads" and "tails."

(iii) Compute the conditional probability of "heads" given "black."

Exercise 4. Consider the following (fictitious) joint-probability table of hair and eye color.

	black hair	brown hair	blonde hair	red hair
brown eyes	0.4	0.22	0.05	0.03
blue eyes	0.05	0.12	0.08	0.004
green eyes	0.001	0.01	0.005	0.03

(i) Calculate the following marginal probabilities. (NB: This exercise uses informal notation like "non-green eyes" instead of cumbersome set-theoretic notation.)

a. P(black hair)

b. P(black hair or red hair)

c. P(non-green eyes)

(ii) Calculate the following conditional probabilities:

a. P(blue eyes | black hair)

b. P(brown hair | non-brown eyes)

c. P(non-green eyes | brown eyes)

d. P(black hair | blue eyes)

(iii) Which of the following events are stochastically independent?

a. black hair and green eyes

b. black hair and brown hair

Exercise 5. Alex and Bo are Jones' children. Each child is allowed to take two sweets each day and they always do take exactly two sweets each. However, there is also a rule that says that the two sweets have to be different. Today Jones observes that three candy bars are missing and one lollipop. That means that one of the children must have taken the same sweet twice. Ha!

Jones decides to test them. The first child to enter the living room will have to show the content of their right pocket. It turns out that this is Alex who shows Jones a chocolate bar. Should Jones conclude from this that Alex is more or less likely to be the culprit than Bo?

Assume that it is, according to Jones, *a priori* equally likely that either child would enter the living room first (whether they are the culprit or not). Assume also that the kids always have one sweet in one pocket, the other in

4 Random variables

We have so far define a probability distribution as a function that assigns a probability to each subset of the space Ω of elementary outcomes. A special case occurs when we are interested in a space of numeric outcomes.

A *random variable* is a function $X : \Omega \to \mathbb{R}$ that assigns to each elementary outcome a numerical value.

Example 7. For a single flip of a coin we have $\Omega_{\text{coin flip}} = \{\text{heads}, \text{tails}\}$. A usual way of mapping this onto numerical outcomes is to define $X_{\text{coin flip}} : \text{heads} \mapsto 1; \text{tails} \mapsto 0$. Less trivially, consider flipping a coin two times. Elementary outcomes should be individuated by the outcome of the first flip and the outcome of the second flip, so that we get:

$$\Omega_{\text{two flips}} = \{\langle\text{heads}, \text{heads}\rangle, \langle\text{heads}, \text{tails}\rangle, \langle\text{tails}, \text{heads}\rangle, \langle\text{tails}, \text{tails}\rangle\}$$

Consider the random variable $X_{\text{two flips}}$ that counts the total number of heads. Crucially, $X_{\text{two flips}}(\langle\text{heads}, \text{tails}\rangle) = 1 = X_{\text{two flips}}(\langle\text{tails}, \text{heads}\rangle)$. We assign the same numerical value to different elementary outcomes.

4.1 Notation & terminology

Traditionally random variables are represented by capital letters, like X. Variables for the numeric values they take on are written as small letters, like x.

We write $P(X = x)$ as a shorthand for the probability $P(\{\omega \in \Omega \mid X(\omega) = 2\})$ that an event occurs that is mapped onto x by random variable X. For example, if our coin is fair, then $P(X_{\text{two flips}} = x) = 0.5$ for $x = 1$ and 0.25 otherwise. Similarly, we can also write $P(X \leq x)$ for the probability of observing an event that X maps to a number not bigger than x.

If the range of X is countable, we say that X is *discrete*. For ease of exposition, we may say that if the range of X is an interval of real numbers, X is called *continuous*.

4.2 Cumulative distribution functions, mass and density

For a discrete random variable X, the *cumulative distribution function* F_X associated with X is defined as:

$$F_X(x) = P(X \leq x) = \sum_{x' \in \{\text{Rng}(X) | x' \leq x\}} P(X = x)$$

The *probability mass function* f_x associated with X is defined as:

$$f_X(x) = P(X = x)$$

Example 8. Suppose we flip a coin with a bias of θ n times. What is the probability that we will see heads k times? If we map the outcome of heads

to 1 and tails to 0, this probability is given by the *Binomial distribution*, as follows:

$$\text{Binom}(K = k; n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. It gives the number possibilities of drawing an unordered set with k elements from a set with a total of n elements. Figure 5 gives an example of the Binomial distribution, concretely its probability mass function, for two values of the coin's bias, $\theta = 0.25$ or $\theta = 0.5$, when flipping the coin $n = 24$ times. Figure 6 gives the corresponding cumulative distributions.

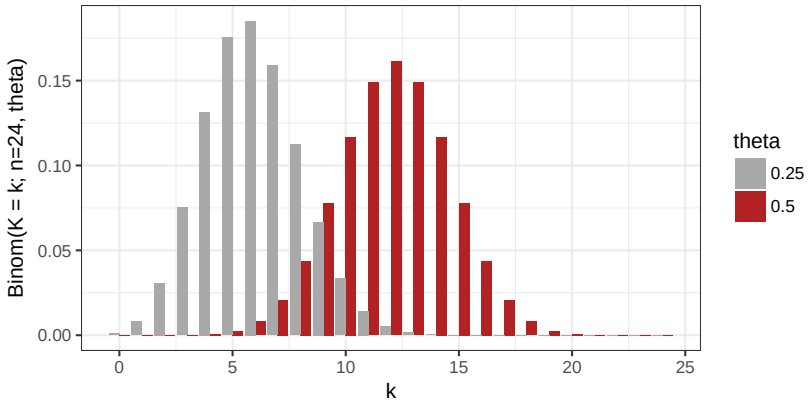

Figure 5: Examples of the Binomial distribution. The y-axis give the probability of seeing k heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

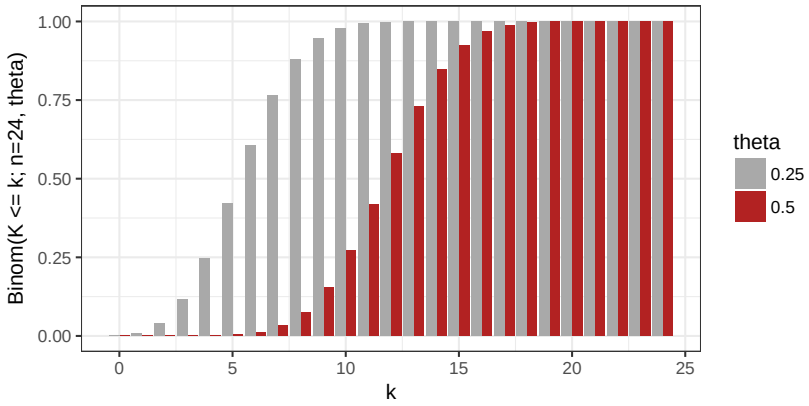

Figure 6: Examples of the cumulative distribution of the Binomial. The y-axis gives the probability of seeing k or less outcomes of heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

For a continuous random variable X, the probability $P(X = x)$ will usually be zero: it is virtually impossible that we will see precisely the value x realized in a random event that can realize uncountably many numerical values of X. However, $P(X \leq x)$ does take workable values and so we define the

cumulative distribution function F_X associated with X as:

$$F_X(x) = P(X \leq x)$$

Instead of a probability *mass* function, we derive a *probability density function* from the cumulative function as:

$$f_X(x) = F'(x)$$

A probability density function can take values greater than one, unlike a probability mass function.

Example 9. The *Gaussian or Normal distribution* characterizes many natural distributions of measurements which are symmetrically spread around a central tendency. It is defined as:

$$\mathcal{N}(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where parameter μ is the *mean*, the central tendency, and parameter σ is the *standard deviation*. Figure 7 gives examples of the probability density function of two normal distributions. Figure 8 gives the corresponding cumulative distribution functions.

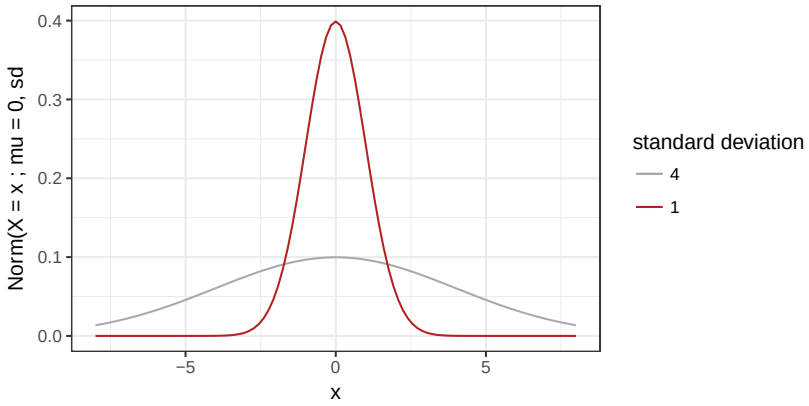

Figure 7: Examples of the Normal distribution. In both cases $\mu = 0$, once with $\sigma = 1$ and once with $\sigma = 4$.

4.3 *Expected value & variance*

The *expected value* of a random variable X is a measure of central tendency. It tells us, like the name suggests, which average value of X we can expect when repeatedly sampling from X. If X is continuous, the expected value is:[13]

$$\mathbb{E}_X = \sum_x x \cdot f_X(x)$$

[13] The expected value is also frequently called the *mean*.

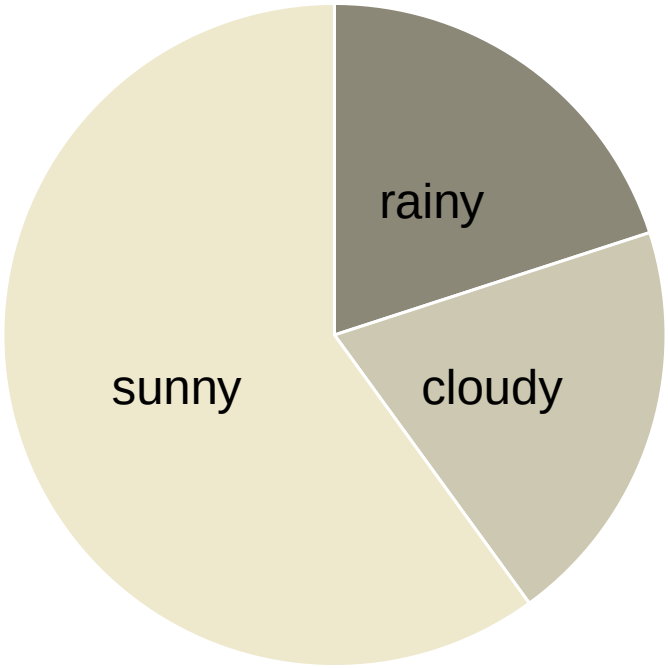

Figure 8: Examples of the cumulative normal distribution corresponding to the previous probability density functions.

If X is continuous, it is:

$$\mathbb{E}_X = \int x \cdot f_X(x)\,\mathrm{d}x$$

The *variance* of a random variable X is a measure of how much likely values of X are spread or clustered around the expected value. If X is discrete, the variance is:

$$\mathrm{Var}(X) = \sum_x (\mathbb{E}_X - x)^2 \cdot f_X(x)$$

If X is continuous, it is:

$$\mathrm{Var}(X) = \int (\mathbb{E}_X - x)^2 \cdot f_X(x)\,\mathrm{d}x$$

Example 10. If we flip a coin with bias $\theta = 0.25$ a total of $n = 24$, we expect on average to see $n \cdot \theta = 24 \cdot 0.25 = 6$ outcomes showing heads.[14] The variance is $n \cdot \theta \cdot (1 - \theta) = 24 \cdot 0.25 \cdot 0.75 = \frac{24 \cdot 3}{16} = \frac{18}{4} = 4.5$.

The expected value of a normal distribution is just its mean μ and its variance is σ^2.

[14] This is not immediately obvious from our definition, but it is intuitive and you can derive it.