Probability theory basics

Michael Franke

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

1 Probability

1.1 Relative degrees of credence

What is the weather going to be like tomorrow: sunny, misty or rainy?[1] We do not know, but we may have different beliefs. Our beliefs might or might not agree in as much as that all three states of the weather are possible, i.e., could actually happen tomorrow. But even if they do, our beliefs might differ (drammaticallly) regarding the question of how likely each possibility is.

[1] Let's assume that there are only these three options and that these are mutually exclusive.

Probabilities are one way of formally representing such beliefs. Probability distributions assign numbers to possible states of affairs (think: sets of possible worlds). What matters most about these numbers are the *relative degrees of credence* they express. Concretely, the information we really care about is that "sunny" is deemed to be twice as likely as "rainy." Take an example.

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 *relative degrees of credence* alone, suffices to conclude that, according to Jones, the probability of the three weather conditions are identical to the probability of outcome when turing the wheel of fortune on the left-hand side in Figure 1. The area covered by the option "sunny" in Jones' wheel of fortune is exactly 60%, while that of both "misty" and "rainy" is exactly 20% each.

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." Smith's wheel of fortune is shown on the right-hand side in Figure 1.

Smith, the pessimist, believes that it is surely not going to be sunny. Probability theory offers a formal representation of uncertainty.

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

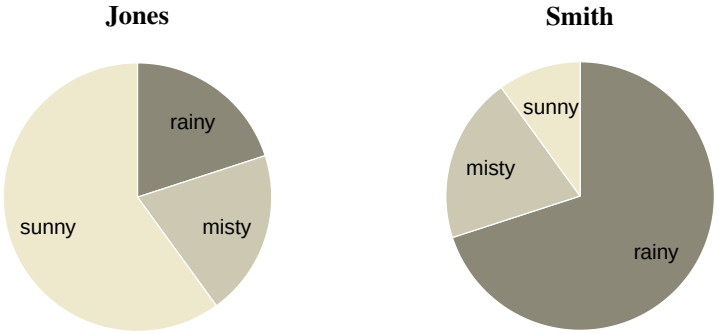

Figure 1: Two beliefs about the weather, represented as wheels of fortune.

1.2 *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.[2]

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 $\Omega_{\text{standard die}} = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.[3]

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.

1.3 *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. $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$ whenever $A_1, A_2, A_3, \ldots$ are mutually exclusive[4]

Occasionally we encounter notation $P \in \Delta(\Omega)$ to express that P is a probability distribution over Ω.[5] If $\omega \in \Omega$ is an elementary event, we often write

[2] For simplicity of exposure, we gloss over subtleties arising when dealing with infinite sets Ω. We make up for this when we define probability *density* functions for continuous random variables, which is all the uncountable infinity that we will usually be concerned with in applied statistics.

[3] Think of Ω as a partition of the space of all possible 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 aspects we currently do not care about.

[4] 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.

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

$P(\omega)$ as a shorthand for $P(\{\omega\})$. In fact, if Ω is finite, it suffices to assign probabilities to elementary outcomes.

A number of rules follow immediately from of this definition:[6]

[6] Prove this!

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 $A, B \subseteq \Omega$

1.4 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 a rational agent who is uncertain about the outcome of a random event.
3. Realist: Probabilities are a property of an intrinsically random world.

1.5 Urns and frequencies

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: $\Omega_{\text{our urn}} = \{\text{white}, \text{black}\}$. If we imagine an infinite sequence of single draws from our urn, putting whichever ball we drew back in after every draw, the limiting proportion with which we draw a black ball is $\frac{k}{N}$.[7] This statement about frequency is what motivates saying that the probability of drawing a black ball on a single trial is (or should be[8]) $P(\text{black}) = \frac{k}{N}$.

[7] If in doubt, execute this experiment. By hand or by computer.

[8] If probabilities are subjective beliefs, a rational agent is, in a sense, normatively required to assign exactly this probability.

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.

An elementary outcome of this two-step process of flip-&-draw is a pair $\langle \text{outcome-flip}, \text{outcome-draw} \rangle$. 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 $\langle\text{heads, black}\rangle$ 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 1.[9]

[9] 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 1: Probabilities of elementary outcomes (pairs of $\langle$outcome-flip, outcome-draw$\rangle$) in the flip-&-draw example.

2.2 *Structured events and joint-probability distributions*

Table 1 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 $\Omega_{flip-\&-draw} = \Omega_{flip} \times \Omega_{draw}$,[10] we can use notation $P(\text{heads, black})$ as shorthand for $P(\langle\text{heads, black}\rangle)$. 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.

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

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:[11]

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

[11] 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 1 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 1 has $P(\text{black}) = 0.3$ and $P(\text{black}) = 0.7$.[12]

[12] 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.[13] It is defined like so:

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

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

Conditional probabilities are only defined when $P(B) > 0$.[14]

[14] 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 2. 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 $A = \{⚂, ⚃, ⚅\}$ 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 $P(A \cap B) = P(\{⚂, ⚅\}) = \frac{1}{3}$. 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 1, 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 is straightforward corollary of 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*,[15] 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:

[15] 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.

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

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.

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 3. 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 4. 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 2 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 3 gives the corresponding cumulative distributions.

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.

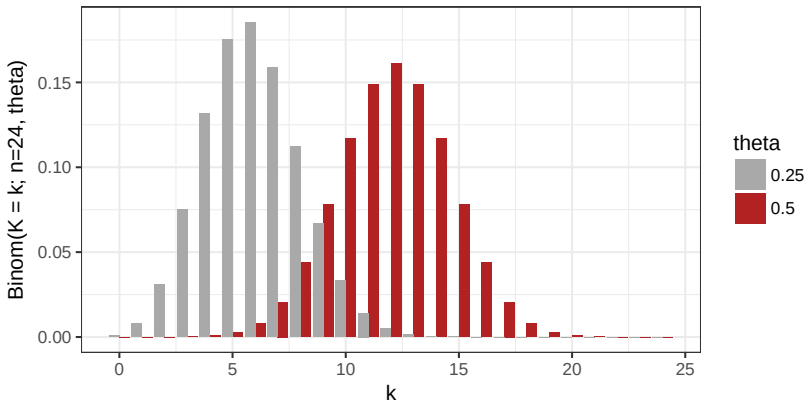

Figure 2: Examples of the Binomial distribution. The y-axis give the probability of seeing k heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

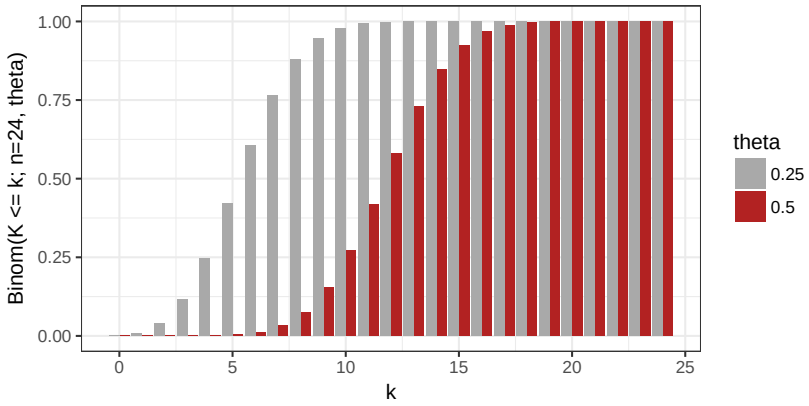

Figure 3: Examples of the cumulative distribution of the Binomial. The y-axis gives the probability of seeing k or less outcomes of heads when flipping a coin $n = 24$ times with a bias of either $\theta = 0.25$ or $\theta = 0.5$.

Example 5. 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 4 gives examples of the probability density function of two normal distributions. Figure 5 gives the corresponding cumulative distribution functions.

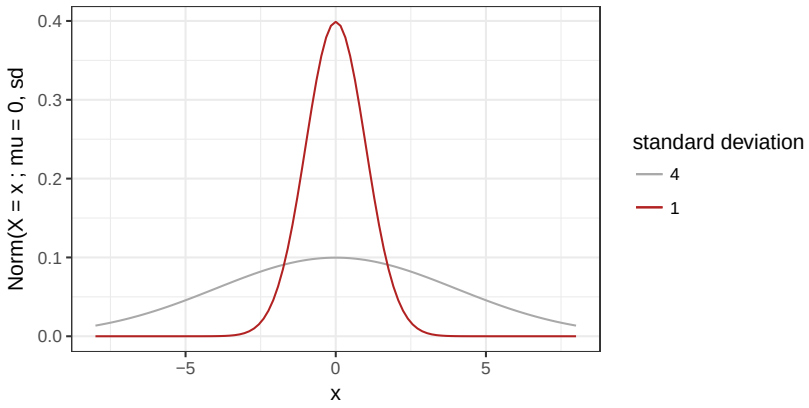

Figure 4: Examples of the Normal distribution. In both cases $\mu = 0$, once with $\sigma = 1$ and once with $\sigma = 4$.

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:[16]

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

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:

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

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

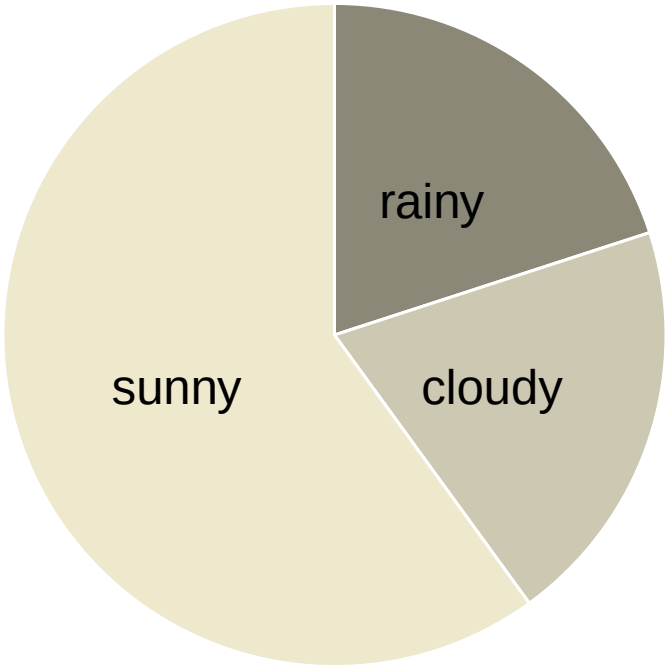

Figure 5: Examples of the cumulative normal distribution corresponding to the previous probability density functions.

Example 6. 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.[17] 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.

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