Probability Theory & Statistics

Random Variables & Distributions

Random Variables

Random variables. A *random variable* (r.v.) is a function assigning a real number to every possible outcome of an experiment.

There are two main types of random variables used in practice: *discrete* r.v.s and *continuous* r.v.s. Discrete r.v.s. are conceptually easier to grasp so we will start with them. At the end of the text, we briefly introduce continuous r.v.s. but most of the intuition needed to understand random variables and their distributions can be gained from working with discrete random variables.

Discrete random variables. A random variable X is *discrete* if there is a finite list of values a_1, a_2, ..., a_n or an infinite list of values a_1, a_2, ... If X is a discrete random variable, then the set of values such that $P(X = x) > 0$ is called the *support* of X.

Coin tosses. Consider an experiment where we toss a fair coin twice. The sample space, S, consists of four possible outcomes, $S = \{HH, HT, TH, TT\}$. We list three possible random variables on this space, where each r.v. is a numerical summary of some aspect of the experiment:

- Let X be the number of heads. The r.v. X has possible values 0, 1, and 2. X can be seen as a function mapping the outcomes to a numerical value, i.e., $X(HH) = 2$, $X(HT) = X(TH) = 1$, and $X(TT) = 0$
- Let Y be the number of tails. We can express Y in terms of X, where $Y = 2 - X$.
- Let I be 1 if the first toss lands heads and 0 otherwise. Then I assigns the value 1 to the outcomes HH and HT and a value 0 to the outcomes TH and TT. The r.v. I is an example of an *indicator random variable.*

Distributions (Discrete)

Given a random variable, we would like to be able to describe its behaviour using the language of probability. For example, we might want to answer questions about the probability that the r.v. will fall into a given range: if L is the lifetime earnings of a randomly chosen company employee, what is the probability that L exceeds a million pounds? If M is the number of major earthquakes in Japan in the next five years, what is the probability that M equals 0?

The *distribution* of a random variable provides the answers to these questions; it specifies the probabilities of all events associated with the r.v., such as the probability of it being equal to 3 and the probability of it being at least 110.

The distribution of a discrete r.v. can be defined in several equivalent ways: using a *probability mass function*, a *cumulative distribution function*, or a *story.* For a discrete

random r.v. the most natural way to define its distribution is with a *probability mass function* (PMF) which we define now.

Probability mass function. The PMF of a discrete r.v. X is the function

$$p_X(x) = P(\{X = x\}) = P(X = x) \tag{1}$$

for $x \in \mathbb{R}$. The PMF is positive if x is in the support of X, 0 otherwise.

Coin tosses continued. We will now find the PMFs of each of the r.v.s from the previous section using the naïve definition of probability:[1]

- If X is the number of heads, then.
  - $p_X(0) = P(X = 0) = 1/4$
  - $p_X(1) = P(X = 1) = 1/2$
  - $p_X(2) = P(X = 2) = 1/4$
- If $Y = 2 - X$ then $P(Y = y) = P(2 - X = y) = P(X = 2 - y) = p_X(2 - y)$
  - $p_Y(0) = P(Y = 0) = 1/4$
  - $p_Y(1) = P(Y = 1) = 1/2$
  - $p_Y(2) = P(Y = 2) = 1/4$
- The indicator r.v. has the PMF.
  - $p_I(0) = P(I = 0) = 1/2$
  - $p_I(1) = P(I = 1) = 1/2$

Children in UK households. Let X be the number of children in a randomly chosen household in the UK. Since X can only be integer values, it is a *discrete r.v.* We can approximate the proportion of households with 0, 1, 2, ... children, and hence approximate the PMF of X by sampling the population.

Remark. We can think of the distribution of an r.v. as a map or blueprint describing the r.v. Just as different houses can share the same blueprint, different r.v.s can have the same distribution, even if the experiments they summarize, and the sample spaces they map from, are not the same. *The word is not the thing, the map is not the territory...*

Named Distributions

Some distributions are so ubiquitous in probability and statistics they have their own names. Starting with a very simple but useful case: an r.v. that can take on only two possible values, 0 and 1.

[1] $P(A) = |A|/|S|$, i.e., the probability of an event A is the number of favourable outcomes in A divided by the total number of outcomes in the sample space S.

Bernoulli & Binomial Distributions

Bernoulli distribution. An r.v. X is said to have the *Bernoulli distribution* with parameter p and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write this as $X \sim Bern(p)$ where the symbol $\sim$ is read as "is distributed as."

Bernoulli trial. An experiment that can result in either a "success" or a "failure" (but not both) is called a *Bernoulli trial.* A Bernoulli r.v. can be thought of as the *indicator of success* in a Bernoulli trial: it equals 1 if successful and 0 if it fails.

Binomial distribution. Suppose that n *independent* Bernoulli trails are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the *Binomial distribution* with parameters n and p. We write $X \sim Bin(n, p)$ to mean that X has the Binomial distribution with parameters n and p, where n is a positive integer and $0 < p < 1$.

Binomial PMF. If $X \sim Bin(n, p)$, then the PMF of X is,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{2}$$

Proof. An experiment consisting of n independent Bernoulli trails produces a sequence of successes and failures 0, 0, 1, 0, 1, 1, ... The probability of any specific sequence of k successes and $n - k$ failures is $p^k(1 - p)^{n-k}$. There are $\binom{n}{k}$ such sequences since we just need to select the successes in any given sequence, e.g., 0, 1, 0; is the same as 1, 0, 0. (The number of ways to choose k out of n things). Therefore, letting X be the number of successes,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{3}$$

Hypergeometric Distribution

Hypergeometric distribution. Consider a bag filled with black b and white w marbles. Draw n marbles out of the bag at random without replacement such that all $\binom{w+b}{n}$ samples are equally likely. Let X be the number of white marbles in the sample. Then X is said to have the *Hypergeometric distribution* with parameters w, b, and n; we denote this by $X \sim HGeom(w, b, n)$.

Hypergeometric PMF. If $X \sim HGeom(w, b, n)$, then the PMF of X is,

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} \tag{4}$$

for integers k satisfying $0 \le k \le w$ and $0 \le n - k \le b$, and $P(X = k) = 0$.

Proof. First count the number of ways to pick k white marbles and $n - k$ black marbles from the bag without considering ordering, i.e., distinguishing between getting the same

sets of marbles so B, B, W is the same as W, B, B. If $k > w$ or $n - k > b$ then the draw is impossible. Otherwise, there are $\binom{w}{k}\binom{b}{n-k}$ ways to draw k white marbles and $n - k$ black marbles by the multiplication rule. The total number of ways to sample $n \leq w + b$ marbles is $\binom{w+b}{n}$, hence by the naïve definition of probability,

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}. \tag{5}$$

A useful way to think about the Hypergeometric distribution is to imagine items in a population that are classified twice – with *two tags*. Each marble is first labelled either *black* or *white* (first tag) and then either sampled or not sampled (second tag). At least one of the tags must be randomly assigned, e.g., the marbles are randomly sampled with all sets of the correct size equally likely, i.e., if there are 5 black marbles and 3 white marbles, the probability a black marble is sampled is 1 in 5. After tagging twice, $X \sim HGeom(w, b, n)$ represents the number of twice-tagged items, e.g., the number of marbles that are *both* white and sampled.

Marbles ($w = 4$, $b = 2$)	Sample ($n = 3$)
W	
~~B~~	B
W	
W	
~~W~~	W
B	B

Table 1: Hypergeometric distribution of marbles in a bag, sampling 3 marbles without replacement.

Aces in a poker hand. In a five-card hand drawn at random from a well-shuffled standard deck, the number of aces in the hand has the $HGeom(4, 48, 5)$ distribution. This can be seen by thinking of the aces as white marbles and the non-aces as black marbles. Using the Hypergeometric PMF we can calculate the probability any given hand has exactly 3 aces,

$$P(X = 3) = \frac{\binom{4}{3}\binom{48}{5-3=2}}{\binom{48+4=52}{5}} \approx 0.0017 \tag{6}$$

So roughly once in every 500 hands!

Cumulative Distribution Functions (CDFs)

Another function that describes the distribution of an r.v. is the *cumulative distribution function* (CDF). The CDF is defined for all r.v.s.

Cumulative distribution function. The CDF of an r.v. X is the function F_X given by

$$F_X(x) = P(X \leq x). \tag{7}$$

For discrete r.v.s we can freely convert from the PMF to the CDF. Consider the PMF of a Binomial r.v. $X \sim Bin(4,\ 1/2)$; to find $P(X \leq 1.5)$, we sum the PMF over all values of the support that are less than or equal to 1.5:

$$P(X \leq 1.5) = P(X = 0) + P(X = 1) = \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^4 = \frac{5}{16} \tag{8}$$

We have seen three equivalent ways of expressing the distribution of a random variable. Two of these are the PMF and the CDF: we know these two functions contain the same information, since we can always figure out the CDF from the PMF and vice versa. Generally, the PMF is easier to work with for discrete r.v.s, since evaluating the CDF requires a summation.

A third way to describe a distribution is with a story that explains (in a precise way) how the distribution can arise. We used the stories of the Binomial and Hypergeometric distributions to derive the corresponding PMFs. Thus, the story and the PMF also contain the same information, though we can often achieve more intuitive proofs with the story than with PMF calculations.

Independence

Independent R.V.s

Independence of random variables. The random variables X and Y are said to be *independent* if,

$$P(X = x,\ Y = y) = P(X = x)P(Y = y) \tag{9}$$

for all x, y with x in the support of X and y in the support of Y.

Roll of two fair dice. In a roll of two fair dice, if X is the number on the first die and Y is the number on the second die, then $X + Y$ is *not independent* of $X - Y$ since,

$$0 = P(X + Y = 12, X - Y = 1) \neq P(X + Y = 12)P(X - Y = 1) = \frac{1}{36} \cdot \frac{5}{36} \tag{10}$$

Knowing the total is 12 tells us the difference must be zero, so the r.v.s provide information about each other (and are therefore not independent). Note, if X and Y are independent, any function of X is independent of any function of Y. Proof is omitted.

Independent and identically distributed. Random variables that are independent and have the same distribution are called *independent and identically distributed* or IID for

short. Note, "independent" and "identically distributed" are two often confused but completely different concepts. Random variables are independent if they provide no information about each other; they are identically distributed if they have the same PMF (or equivalently, the same CDF). Whether two r.v.s are independent has nothing to do with whether they have the same distribution.

We can have r.v.s that are:

- *independent and identically distributed.* Let X be the result of a die roll and let Y be the result of a second, independent die roll. Then X and Y are IID.
- *independent and not identically distributed.* Let X be the result of a die roll and let Y be the closing price of the Dow Jones (a stock market index) a month from now. Then X and Y provide no information about each other (one would hope), and X and Y do not have the same distribution.
- *dependent and identically distributed.* Let X be the number of heads in n independent fair coin tosses and let Y be the number of tails in those same n tosses. Then X and Y are both distributed $Bin(n,\ 1/2)$, but they are highly dependent: if we know X, then we know Y perfectly.
- *dependent and not identically distributed.* Let X be the indicator of whether the majority party retains control of parliament after the next election and let Y be the average likeability rating of the majority party in polls taken within a month of the election. Then X and Y are dependent, and X and Y do not have the same distribution.

Conditionally Independent of R.V.s

Conditional independence. Random variables X and Y are *conditionally independent* given an r.v. Z if for all x, y in the support of X and Y, and all z in the support of Z,

$$P\left(X = x, Y = y | Z = z\right) = P\left(X = x | Z = z\right) P(Y = y | Z = z) \tag{11}$$

Conditional PMFs

Conditional PMFs. For any discrete r.v.s X and Z, the function $P(X = x|Z = z)$, when considered as a function of x for fixed z, is called the *conditional PMF of X given $Z = z$.*

Examples

Matching pennies. Consider the game of matching pennies. Each of two players, A and B, has a fair penny. They flip their pennies independently. If the pennies match, A wins; otherwise, B wins. Let X be 1 if A's penny lands heads and -1 otherwise and define Y similarly for B (r.v.s X and Y are called *random signs*).

A	B	Winner	X	Y	Z
H	T	B	1	-1	-1
T	H	B	-1	1	-1
T	T	A	-1	-1	1
H	H	A	1	1	1

Table 2: Matching pennies game with outcomes and winning represented by random variables.

Let $Z = XY$, which is 1 if A wins and -1 if B wins. Then X and Y are unconditionally independent, but given $Z = 1$, we know that $X = Y$ (the pennies match).[2] So X and Y are *conditionally dependent* given .

Two friends. "I have only two friends who ever call me" – Alice and Bob. Let X be the indicator of Alice calling me next Friday, let Y be the indicator of Bob calling me next Friday, and let Z be the indicator of exactly one of them calling me next Friday.

Then X and Y are independent (by assumption). But given $Z = 1$, we have that X and Y are completely dependent: given that $Z = 1$, we have $Y = 1 - X$, i.e., X and Y are conditionally dependent given Z.

Mystery opponent. Imagine you are going to play two games of tennis against one of two identical twins. Against one of the twins, you are evenly matched, and against the other you have a 3/4 chance of winning. Suppose that you can't tell which twin you are playing against until after the two games. Let Z be the indicator of playing against the twin with whom you're evenly matched, and let X and Y be the indicators of victory in the first and second games, respectively.

Conditional on $Z = 1$, X and Y are IID $Bern(1/2)$, and conditional on $Z = 0$, X and Y are IID $Bern(3/4)$. So, X and Y are *conditionally independent* given Z. (If we know who we're playing against, the probability of winning each game conditional on knowing if given by the Bernoulli distribution).

Unconditionally, X and Y are dependent because observing $X = 1$ makes it more likely

[2] If we have no information about the outcomes $P(X = Y) = P(X = z, Y = z) = P(X = z)P(Y = z) = 1/2$, however knowing $Z = 1$, we know that $P(X = Y) = 1 \neq P(X = z)P(Y = z) = 1/2$.

that we are playing the twin who is worse. That is,

$$P(Y = 1|X = 1) > P(Y = 1). \tag{12}$$

Past games give us information which helps us infer who our opponent is, which in turn helps us predict future games! Note, this problem is the same as the biased coin problem from the Bayesian inference example!

Binomial-Hypergeometric Connection

The Binomial and Hypergeometric distributions are connected in two important ways. As we will see, we can get from the Binomial distribution to the Hypergeometric distribution by *conditioning*, and we can get from the Hypergeometric to the Binomial by *taking a limit*.

Fisher exact test. A scientist wishes to study whether women or men are more likely to have a certain disease, or whether they are equally likely. A random sample of n women and m men is gathered, and each person is tested for the disease (assume for this problem that the test is completely accurate). The numbers of women and men in the sample who have the disease are X and Y respectively, with $X \sim Bin(n, p_1)$ and $Y \sim Bin(m, p_2)$, independently.

Here p_1 and p_2 are unknown, and we are interested in testing whether $p_1 = p_2$ (this is known as a *null hypothesis* in statistics).

Consider a 2-by-2 table with rows corresponding to disease status and columns corresponding to gender. Each entry is the count of how many people have that disease status and gender, so $n + m$ is the sum of all four entries. Suppose that it is observed that $X + Y = r$.

	Women	Men	Total
Disease	x	$r - x$	r
No Disease	$n - x$	$m - r + x$	$n + m - r$
Total	n	m	$n + m$

Table 3: Rows correspond to disease status and columns correspond to gender.

The *Fisher exact test* is based on conditioning on both the row and column sums, so n, m, r are all treated as fixed, and then seeing if the observed value of X is "extreme" compared to this conditional distribution. Thus, assuming the null hypothesis, find the conditional PMF of X given $X + Y = r$.

Solution. Treating n, m, and r as fixed, compute the conditional PMF

$$P\left(X = x|X + Y = r\right) \tag{13}$$

using Bayes' rule

$$P\left(A|B\right) = \frac{P\left(B|A\right)P(A)}{P(B)}. \tag{14}$$

From Bayes' rule we get

$$P\left(X = x|X + Y = r\right) = \frac{P\left(X + Y = r|X = x\right)P(X = x)}{P(X + Y = r)} \tag{15}$$

$$= \frac{P(Y = r - x)P(X - x)}{P(X + Y = r)} \tag{16}$$

where the second step, $P\left(X + Y = r|X = x\right) = P(Y = r - x)$ is justified by the independence of X and Y, i.e., conditioning on X doesn't change the probability of Y, so $P\left(X + Y = r|X = x\right) = P(x + Y = r)$.

Assuming the null hypothesis and letting $p = p_1 = p_2$, we have $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ independently so $X + Y \sim Bin(n + m, p)$. Feeding this into the above equation we get

$$P\left(X = x|X + Y = r\right) = \frac{P(Y = r - x)P(X - x)}{P(X + Y = r)} \tag{17}$$

$$= \frac{\binom{m}{r-x}p^{r-x}(1-p)^{m-r+x}\binom{n}{x}p^{x}(1-p)^{n-x}}{\binom{n+m}{r}p^{r}(1-p)^{n+m-r}} \tag{18}$$

$$= \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} \tag{19}$$

To understand why the Hypergeometric appeared seemingly out of nowhere, think about the double tagging system from the Hypergeometric story. Think of women as tagged and men as untagged. If we infect $X + Y = r$ people with the disease; under the null hypothesis, the set of diseased people is equally likely to be any set of r people. Thus, conditional on $X + Y = r$, X represents the number of women among the r diseased individuals. This is exactly analogous to the number of black or white marbles sampled or not sampled, which is distributed $HGeom(b, w,\ n)$.

Making this explicit for clarity. When $p_1 = p_2$ there is no preferred set of r diseased people. So, for 2 diseased people out of 3, $\{Alice,\ Bob,\ Carl\}$, the probability Alice and Bob are assigned the disease is the same probability that Alice and Carl are assigned the disease (and all other pairwise combinations). This is the same as randomly sampling n marbles from a bag of $b + w$ marbles, i.e., $r = n$. Finally, assigning the label "Man" or "Woman" to the population corresponds to assigning a second tag black or white to the "population" of marbles, hence the problems are *isomorphic*.

Gene set analysis. The Fisher exact test can be used to quantify the significance of overlapping lists, e.g., lists of genes. Suppose $G_1 \subset G$ and $G_2 \subset G$ are two sets of genes

in the universe of all possible genes G. We are interested in a measure of significance of their overlap, i.e., is their overlap random or is there some underlying process that makes their intersection more likely. Let X be the number of genes in both G_1 and G_2, fixing n as the number of genes in G_1, m as the number of genes that aren't in G_1 and r as the number of genes in G_2 we get the following conditional outcomes equivalent to both previous examples.

	In G_1	Not in G_1	Total
In G_2	x	$r - x$	r
Not in G_2	$n - x$	$m - r + x$	$n + m - r$
Total	n	m	$n + m$

Table 4: 2-by-2 gene contingency table.

Again, we see that overlapping lists can be interpreted as double tagging objects, and are thus modelled by the Hypergeometric distribution.

Question. Review the assumptions in the three described examples. What are the assumptions and are they valid?

A useful fact that is often used for calculations says the conditional distribution of X does not depend on p: unconditionally, $X \sim Bin(n, p)$, but p disappears from the parameters of the conditional distribution! This makes sense upon reflection since once we know $X + Y = r$, we can work directly with the fact that we have a population with r diseased and $n + m - r$ healthy people, without worrying about the value of p that originally generated the population.

Question. We can get from the Binomial to the Hypergeometric by conditioning. How do we get from the Hypergeometric back to the Binomial?

Counting Distributions

The Binomial and Hypergeometric distributions are two of four counting distributions with different sampling schemes. Table **??** summarizes the sampling schemes: with and without replacement, and the stopping rule – a fixed number of draws or a fixed number of successes.

	With Replacement	Without Replacement
Fixed Number of Trials	Binomial	Hypergeometric
Fixed Number of Successes	Negative Binomial	Negative Hypergeometric

Table 5: Table of discrete counting distributions with different sampling schemes.

Continuous Random Variables

So far, we have been working with discrete random variables, whose possible values can be written down as a list. Continuous r.v.s can take on any real value in an interval (possibly of infinite length, such as $(0, \infty)$ or the entire real line).

Continuous random variables. An r.v. has a *continuous distribution* if its CDF is *differentiable*. (With some technical caveats that are omitted here).

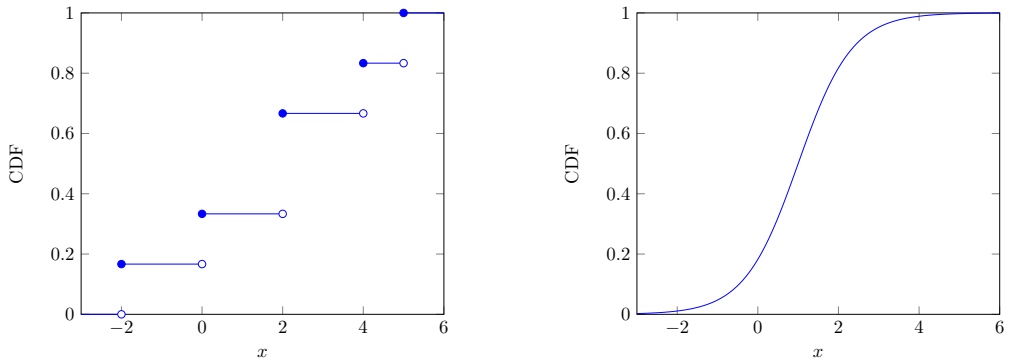

Figure 1: (Left) Cumulative Distribution Function (CDF) of a discrete r.v.; (Right) CDF of a continuous r.v.

For discrete r.v.s, the CDF is awkward to work with because of its jumpiness, and its derivative is almost useless since it's undefined at the jumps and 0 everywhere else. But for continuous r.v.s, the CDF is often more convenient to work with, and its derivative is a very useful function, called the *probability density function*.

Probability density function. For a continuous r.v. X with CDF F, the probability density function (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The support of X, and of its distribution, is the set of all x where $f(x) > 0$.

An important way in which continuous r.v.s differ from discrete r.v.s is that for a continuous r.v. X, $P(X = x) = 0$ for all x. This is because $P(X = x)$ is the height of a jump in the CDF at x, but the CDF of X has no jumps! Since the PMF of a continuous r.v. would just be 0 everywhere, we work with a PDF instead. (This can also be seen by calculating the naïve probability of a sample space $S = \mathbb{R}$.)

The PDF is analogous to the PMF in many ways, but there is a key difference: for a PDF f, the quantity $f(x)$ is *not a probability*, and in fact it is possible to have $f(x) > 1$ for some values of x. To obtain a probability, we need to *integrate* the PDF. The fundamental theorem of calculus tells us how to get from the PDF back to the CDF.

PDF to CDF. Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{20}$$

The proof is omitted.

Intuition

Units

Let F be the CDF and f be the PDF of a continuous r.v. X. As mentioned earlier, $f(x)$ is not a probability; for example, we could have $f(3) > 1$, and we know $P(X = 3) = 0$. But thinking about the probability of X being very close to 3 gives us a way to interpret $f(3)$. Specifically, the probability of X being in a tiny interval of length ε, centred at 3, will essentially be $f(3) \cdot \varepsilon$. This is because,

$$P(3 - \varepsilon/2 < X < 3 + \varepsilon/2) = \int_{3-\varepsilon/2}^{3+\varepsilon/2} f(x)dx \approx f(3) \cdot \varepsilon \tag{21}$$

if the interval is so tiny that f is approximately the constant $f(3)$ on that interval. In general, we can think of $f(x)dx$ as the probability of X being in an infinitesimally small interval containing x, of length dx.

In practice, X often has units in some system of measurement, such as units of distance, time, area, or mass. Thinking about the units is not only important in applied problems, but also it often helps in checking that answers make sense.

Suppose for concreteness that X is a length, measured in centimetres (cm). Then $f(x) = dF(x)/dx$ is the probability per cm at x, which explains why $f(x)$ is a probability density. Probability is a dimensionless quantity (a number without physical units), so the units of $f(x)$ are per-cm. Therefore, to get a probability again, we need to multiply $f(x)$ by a length. When we do an integral such as $f(x) \cdot dx$, this is achieved by the dx.

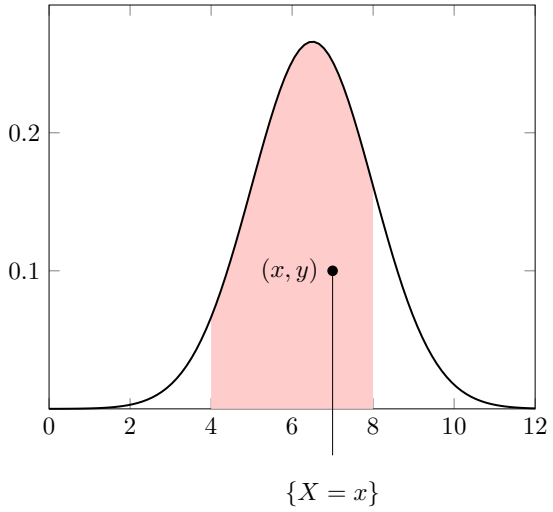

Figure 2: Uniformly sampling a random point (x, y) under the curve and letting $X = x$ generates an r.v., X, with this PDF.

Simulation

Another way to think about PDFs is to consider them as generators of the process X. To generate X, choose a uniformly random point under the PDF curve; this means that the probability of any region under the curve is the area of that region. Then let X be the x-coordinate of the random point. X then has the desired distribution since, by construction, $P(a \leq X \leq b)$ is the area under the PDF curve between the lines $x = a$ and $x = b$. Thinking about this method helps build intuition for PDFs, how random variables are sampled according to a particular PDF curve.

Question. Use the simulation intuition to describe the probability of falling in an interval $[a, b]$ given a constant PDF (that of continuous uniform random variable).

Probability Spaces

In the previous section we showed that intuitively there are several ways to think about continuous random variables and their distributions. The paradoxes that arise when calculating probabilities in a naïve sense were addressed by the Soviet mathematician Andrey Kolmogorov in the 1930s by introducing the notion of a *probability space*, together with other axioms of probability (see the session notes on the axioms of probability). For most (but not all) practical applications, the standard probability space is assumed (and omitted from calculations).

Discussion

What's the point of learning this? When will we ever come across these examples in practice – they aren't applicable in the "real world".

This text provides *definitions* of the concepts with examples that illustrate – *in the simplest possible terms* – what each concept looks like in practice. As an analogy, when learning how to add and subtract we don't start by balancing the accounting books of a multinational corporation. We start with simple examples that build an intuitive foundation and build on those foundations until we are ready to perform sophisticated accounting tasks. (Note, reducing accounting to addition and subtraction is equivalent to reducing statistics to calculating conditional probabilities – it's the contextual understanding around the calculations that makes it hard!)

Appendix

Longhand Notation

In writing $P(X = x)$, we are using $X = x$ to denote an *event*, consisting of all outcomes s to which X assigns the number x. This event is also written as $\{X = x\}$; formally, $\{X = x\}$ is *defined* as $\{s \in S : X(s) = x\}$, read as " ...the set where each element (outcome) s in the set (of all possible outcomes) S, satisfies the property (mapping) $X(s) = x$." But writing $\{X = x\}$ is shorter and more intuitive.

If X is the number of heads in two fair coin tosses, then $\{X = 1\}$ consists of the sample outcomes HT and TH, which are the two outcomes to which X assigns the number 1. Since $\{HT, TH\}$ is a subset of the sample space, it is an event. So, it makes sense to talk about $P(X = 1)$, or more generally, $P(X = x)$. If $\{X = x\}$ were anything other than an event, it would make no sense to calculate its probability! It does not make sense to write "$P(X)$"; we can only take the probability of an event, not of a random variable.

The Law of the Unconscious Statistician

The Law of the Unconscious Statistician. If X is a discrete r.v. and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then,

$$E(g(X)) = \sum_x g(X)P(X = x) \tag{22}$$

where $E(\cdot)$ is the *mathematical expectation* and the sum is taken over all possible values of X. This means that we can get the expected value of $g(X)$ knowing only $P(X = x)$, the PMF of X; we don't need to know the PMF of $g(X)$. The name comes from the fact that in going from $E(X)$ to $E(g(X))$ it is tempting just to change x to $g(x)$ in the definition, which can be done very easily and mechanically, perhaps in a state of unconsciousness.

On second thought, it may sound too good to be true that finding the distribution of $g(X)$ is not needed for this calculation, but LOTUS says it is true! Before proving LOTUS in general, let's see why it is true in some special cases.

Let X have support 0, 1, 2, ... with probabilities p_0, p_1, p_2, ..., so the PMF is $P(X = n) = p_n$. Then X^3 has support 0, 1, 2^3, 2^4, ... with probabilities p_0, p_1, p_2, ..., so

$$E(X) = \sum_{n=0}^{\infty} np_n, \tag{23}$$

and

$$E\left(X^3\right) = \sum_{n=0}^{\infty} n^3 p_n. \tag{24}$$

As claimed by LOTUS, to edit the expression for $E(X)$ into an expression for $E(X^3)$, we can just change the n in front of the p_n to an n^3. This was an easy example since the function $g(x) = x$ is one-to-one. But LOTUS holds much more generally. The key insight needed for the proof of LOTUS for general g is the same as the proof of linearity: the expectation of $g(X)$ can be written in ungrouped form as

$$E(g(X)) = \sum_{s} g(X(s))P(\{s\}), \tag{25}$$

where the sum is over all the elements in the sample space, but we can also group the elements into groups of elements according to the value that X assigns to them. Within group $X = x$, $g(X)$ always takes on the value $g(x)$. Therefore, $E(g(X)) = X$. In the last step, we used the fact that $\sum_{s:X(s)=x} P(\{s\})$ is the weight of the group $X = x$.