Chapter 5

Random Variables and Probability Distributions

When we consider outcomes of an experiment, we often want to consider numerical summaries of the experiment. Random variables are useful for this.

Example 5.1. Suppose we toss a coin 3 times and are interested in the number of heads thrown. The sample space, which lists all of the possible outcomes of the experiment, looks like this:

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

but the possible values of the variable we are interested in are $X = \{0, 1, 2, 3\}$.

The above is an example of a random variable. A formal definition is as follows:

Definition 5.1. A **random variable** is a function $X : S \to \mathbf{R}$ that associates a numerical value, $X(s)$, with every outcome $s \in S$.

Example 5.2. Suppose we count the number of trains to arrive at Exeter St David's station from 9am-10am tomorrow. Then

$$S = \{0, 1, 2, \ldots\}.$$

The number of trains to arrive during that period is a random variable $X(s) = s$ for every $s \in S$.

Example 5.3. Suppose, in example 5.2 we are interested in the time (in seconds after 9am), X, that the first train arrives. Here

$$S = \{t : t \in [0, \infty)\}$$

and $X(t) = t$, but now $X(t)$ could be any value on the positive real line (it is continuous).

Definition 5.2. The **range space**, R, of a random variable is the set of all possible values a random variable can take.

Definition 5.3. A **discrete random variable**, has the property that its range space is a countable set. A random variable is **continuous** if its range space is an uncountable set.

In example 5.2 the number of trains was countable. In fact it was *countably infinite* as there was a one to one mapping of $X(s)$ with the natural numbers $\mathbf{N}$. The number of heads in 3 tosses is an example of a finite countable set.

In example 5.3, the time taken for the first train to arrive could have been any non-negative real number and so was uncountable.

There are slight differences in how we must treat discrete and continuous random variables, but the ideas and reasoning are the same and so we develop our account in parallel as far as possible, to fix ideas. We start with the notion of the **probability distribution**.

5.1 Probability distributions

We know that, effectively, X is a random function of the outcomes of an experiment. I.e., we don't know what value X will take when we perform the experiment. A probability distribution represents all of the probability statements that we can make about a random variable, simultaneously.

Definition 5.4. The **cumulative distribution function** (cdf) of a random variable X is

$$F(x) = P(X \leq x).$$

- Note that this definition applies to either a discrete or continuous random variable.
- Note we will always use a captial letter (e.g. X) to denote the random variable itself and the corresponding lower case letter (e.g. x) to denote the particular value X took after the stochastic experiment.

The cdf has the following properties:

1. $F(-\infty) = P(X \leq -\infty) = 0$,
2. $F(\infty) = P(X \leq \infty) = 1$,
3. If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$ (F is non-decreasing).

Example 5.4. Let X have CDF $F(x)$. Express $P(a < X \leq b)$ and the CDF of X^2 in terms of F.

Solution. First, we have

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a). \end{aligned}$$

For the second piece, suppose we write $F_{X^2}(x)$ to be the CDF of X^2, then

$$\begin{aligned} F_{X^2}(x) = P(X^2 \leq x) &= P(-\sqrt{x} \leq X \leq \sqrt{x}) \\ &= P(X \leq \sqrt{x}) - P(X \leq -\sqrt{x}) \\ &= F(\sqrt{x}) - F(-\sqrt{x}). \end{aligned}$$

□

Note: that knowing the cdf of a random variable for all possible values of x gives any probability of the form $P(a \leq X \leq b)$ for real a and b via

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Definition 5.5. For a discrete random variable, X, the **probability mass function** (pmf) is $p(x) = P(X = x)$ for all $x \in R$.

Example 5.5. Suppose we roll two dice and let the random variable X be the total score of the two rolls. We can tabulate the **probability distribution** for this random variable by giving $P(X = x)$ for all $x \in R$.

$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
x	2	3	4	5	6	7	8	9	10	11	12

Note, that we can find $P(X = x)$ for a discrete random variable X from the cdf via

$$P(X = x) = F(x) - F(x - 1)$$

.

For continuous random variables, the quantity $P(X = x)$ is unhelpful as

$$\begin{aligned} P(X = x) &= \lim_{h \to 0} P(x - h < X < x + h) \\ &= \lim_{h \to 0} (F(x + h) - F(x - h)) \\ &= 0. \end{aligned}$$

So, for any $x \in R$, if X is a continuous random variable $P(X = x) = 0$, though some value of x must occur. This essentially says that 0 probability events are not impossible!

The pmf is very useful for discrete random variables as it enables us to write down a probability distribution for the variable as the set $\{(x, p(x)), x \in R\}$. Though the cdf gives this information too, and is also sufficient for a continuous variable, there is a continuous analogue to the pmf that we can derive from the cdf of a continuous random variable called the probability density function (pdf).

Definition 5.6. The **probability density function** (pdf) of a random variable X, is $f(x)$ where

$$f(x) = F'(x),$$

or, written another way,

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

The pdf, $f(x)$ is NOT the probability that $X = x$ (that is 0). To see what it is, note that for $h > 0$,

$$F'(x) = \lim_{h \to 0} \frac{F(x + h) - F(x)}{h},$$

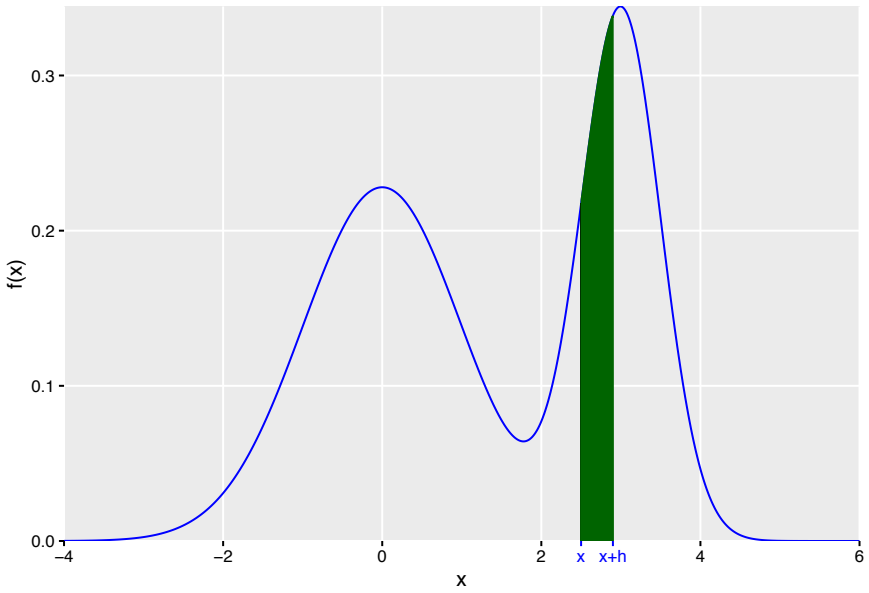
so that, for small h,

$$P(x < X < x + h) = F(x + h) - F(x) = hf(x),$$

i.e. the probability of being within a small interval of x is approximately the density times the width of the interval. We can get there from the integral expression via

$$\begin{aligned} P(x < X < x+h) = F(x+h) - F(x) &= \int_{-\infty}^{x+h} f(t)dt - \int_{-\infty}^{x} f(t)dt \\ &= \int_{x}^{x+h} f(t)dt \\ &\approx hf(x). \end{aligned}$$

```
a_pdf <- function(x){(1/sqrt(2*pi))*((4/7)*(1/1)*exp(-(1/(2*1^2))*(x-0)^2)
                                    + (3/7)*(1/0.5)*exp(-(1/(2*0.5^2))*(x-3)^2))}
xs <- seq(from=-4,to=6,len=1000)
pdf_data <- data.frame(x=xs, f = a_pdf(xs))
myplot <- ggplot(data=pdf_data) +
  geom_line(mapping = aes(x=x,y=f),colour="blue") +
  scale_x_continuous(breaks=c(-4,-2,0,2,2.5,2.9,4,6),
                     labels = c("-4","-2","0","2","x","x+h","4","6"),
                     expand=c(0,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme(axis.text.x = element_text(color = c(rep("black", 4), "blue", "blue", "black", "black")),
        axis.ticks.x = element_line(color=c(rep("black" ,4), "blue", "blue", "black", "black"))) +
  theme(axis.text.y = element_text(color="black"),
        axis.ticks.y = element_line(color="black")) +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major.x = element_line(color=c(rep("white", 4), NA, NA, "white", "white"))) +
  labs(y="f(x)")
shade <- rbind(c(2.5,0), subset(pdf_data, (x>2.5)&(x<2.9)), c(2.9,0))
myplot2 <- myplot + geom_segment(aes(x=2.5,y=0,xend=2.5,yend=a_pdf(2.5))) +
  geom_polygon(data=shade,aes(x,f),fill="darkgreen",colour="darkgreen")
myplot2
```



The probability distribution for a continuous random variable is completely characterised by its pdf or cdf.

5.1.1 Properties

Though the pdf $f(x)$ is not a probability for any value of x, unlike the discrete analogue the pmf $p(x)$, the two nevertheless have corresponding properties. For discrete X we have

1. $0 \leq p(x) \leq 1$ (as each $p(x)$ is a probability it must obey the axioms)
2. $\sum_{x \in R} p(x) = 1$, as, this is the probability that an outcome in S occurs.

For continuous X,

1. $f(x) \geq 0$ for every x ($F(x)$ is non-decreasing so cannot have a negative first derivative).
2. $\int_{-\infty}^{\infty} f(x)dx = F(\infty) = 1$.

Note that $f(x) > 1$ is possible (consider a range space consisting only of a tiny interval h, then for h small enough, our previous argument plus property 2 gives $hf(x) \approx 1$).

Example 5.6. Let $X \in [0, \frac{1}{2}]$ continuous with constant pdf, $f(x) = c$ (we will recognise this as a uniform random variable later). Find c.

Solution. Using property 2 we have

$$\int_{-\infty}^{\infty} f(x)dx = 1 = \int_{0}^{\frac{1}{2}} cdx = [cx]_0^{\frac{1}{2}} = \frac{c}{2}.$$

So $c = 2\ (> 1)$.

□

Example 5.7. Suppose a continuous random variable $X \in [0, 1]$ has pdf $f(x)$ that is proportional to x. Find $f(x)$.

Solution. To solve this problem, we first note that X is on $[0, 1]$ and so, by property 2,

$$\int_{0}^{1} f(x)dx = 1.$$

$f(x) = kx$ for some k so we have

$$\int_{0}^{1} kxdx = k\left[\frac{x^2}{2}\right]_0^1 = \frac{k}{2} = 1$$

which implies that $f(x) = 2x$ for $x \in [0, 1]$ and 0 otherwise.

□

Example 5.8. Suppose you have a battery powered toy which will run out of charge in T hours. Suppose, further, that the pdf of time T is

$$f(t) = \begin{cases} C(10 - t) & 0 < t < 10 \\ 0 & \text{otherwise.} \end{cases}$$

a. Find C.
b. Find $P(T = 7)$.
c. Find the probability that the toy lasts more than 8 hours.

Solution. Firstly,

a.

$$\begin{aligned} F(\infty) &= \int_{-\infty}^{\infty} f(t)dt = 1 \\ &= \int_{0}^{10} C(10-t)dt \\ &= C\int_{0}^{10} (10-t)dt = C\left[10t - \frac{t^2}{2}\right]_0^{10} \\ &= C(100-50) = 50C. \end{aligned}$$

So $C = 1/50$.

b. Of course, as T is continuous $P(T = 7) = 0$.

c. This is written $P(T > 8)$ and

$$\begin{aligned} P(T > 8) &= 1 - P(T \leq 8) \\ &= 1 - \int_0^8 \frac{1}{50}(10-t)dt \\ &= 1 - \frac{1}{50}\left[10t - \frac{t^2}{2}\right]_0^8 \\ &= 1 - \frac{1}{50}(80-32) = 1 - \frac{48}{50} \\ &= \frac{1}{25}. \end{aligned}$$

□

Example 5.9. Suppose we conduct a sequence of independent coin tosses with probability p of tossing heads, stopping when we first toss a head. Let X be the random variable that takes the number of tosses required to toss a head and stop the sequence. Then

$$P(X = n) = (1-p)^{n-1}p,$$

as it would take $n-1$ tails and then one head in sequence to stop the experiment at the nth toss. Clearly, $p(x)$ satisfies property 1 for a discrete random variable, so to prove it is a valid probability mass function, we need to sure that

$$\sum_{n=1}^{\infty} p(n) = 1.$$

5.1.2 Digression: Arithmeric and Geometric series

When doing calculatons with discrete probability distributions, we must often evaluate and expand series. We present a few key results that will be used throughout the rest of this course in probability calculations. More complete derivations will be presented to you in other courses.

Suppose we have an *arithmetic progression* $a_1, a_2, \ldots$ with $a_k = a_1 + (k-1)d$ for some fixed difference d. Then,

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n \implies$$

$$2\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n + a_n + a_{n-1} + \cdots + a_1$$

$$= (a_1 + a_n) + (a_2 + a_{n-1}) + \cdots + (a_{n-1} + a_2) + (a_n + a_1)$$

Each bracketed element is the same value as, e.g.

$$\begin{aligned}(a_i + a_{n-i+1}) &= a_1 + (i-1)d + a_1 + (n-i+1)d \\ &= 2a_1 + nd \\ &= a_1 + a_n.\end{aligned}$$

So

$$\sum_{k=1}^{n} a_k = \frac{n(a_1 + a_n)}{2}.$$

For example, the sum of the first n integers is the sum of an arithmetic progression with common difference 1, so that

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

A *geometric progression* is a sequence $a_1, a_2, \ldots$ where each member of the sequence is multiplied by a fixed number r. So

$$a_k = r a_{k-1}, \qquad a_k = r^{k-1} a,$$

where $a = a_1$. To work out the sum of a geometric sequence, note that

$$\sum_{k=1}^{n} a_k = \sum_{k=0}^{n-1} r^k a = a + ra + r^2 a + \cdots + r^{n-1} a,$$

so that

$$\sum_{k=1}^{n} a_k - r\sum_{k=1}^{n} a_k = a - r^n a \implies$$

$$\sum_{k=1}^{n} a_k = \frac{a - r^n a}{1 - r}.$$

As $n \to \infty$, this has a finite limit for $|r| < 1$ ($r^n \to 0$). Hence, for $|r| < 1$ we have

$$\sum_{k=0}^{\infty} r^k a = \frac{a}{1-r}.$$

Note then we have

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$$

for $|x| < 1$ and we can differentiate to get

$$\begin{aligned}\frac{1}{(1-x)^2} &= \sum_{k=0}^{\infty} k x^{k-1} \\ &= \sum_{k=1}^{\infty} k x^{k-1} = \sum_{n=0}^{\infty} (n+1) x^n.\end{aligned}$$

In Example 5.9 we saw the pmf of a random variable (number of tosses until the first head) was $p(n) = P(X = n) = (1-p)^{n-1}p$. To check this satisfies property 2, we have

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p = \sum_{n=0}^{\infty}(1-p)^n p$$
$$= \frac{p}{1-(1-p)} = \frac{p}{p} = 1.$$

Note $P(X > n) = (1-p)^n$ as the first n tosses must be tails for this to happen. Hence the cdf of this distribution (which is known as the **geometric** distribution) is

$$P(X \leq x) = 1 - (1-p)^n$$

by property 2.

5.2 Common distributions and their applications

We have just seen one example distribution the **geometric**. We will now examine some more and some of the applications they are used for.

5.2.1 The Uniform Distribution

Definition 5.7. Let X be a discrete random variable with range space $R = \{1, \ldots, n\}$. X has the **discrete uniform distribution** if its probability mass function is

$$P(X = x) = \begin{cases} \frac{1}{n} & x \in R \\ 0 & \text{otherwise} \end{cases}$$

Here is a plot of the pmf for n=10

This is the probability distribution describing the case that "all outcomes are equally likely". It's close cousin is the continuous version, claiming that all intervals on a predefined domain are equally likely. This idea requires a constant pdf (try to prove this as an exercise).

Definition 5.8. Let X be a continuous random variable on $[a, b]$. X has the **continuous uniform distribution** (normally shortened to the uniform distribution) if its probability density function is constant on $[a, b]$ and 0 otherwise.

We can derive the pdf of any particular uniform distribution (finding the constant) by ensuring that the pdf integrates to 1.

$$\int_a^b f(x)dx = \int_a^b k dx = 1 \implies$$
$$k = \frac{1}{b-a}$$

Note that the cdf is a straight line. The pdf and cdf are shown in here

To write that a random variable has a uniform distribution we write $X \sim \text{Unif}(a, b)$, and for the standard distribution we write $X \sim \text{Unif}(0, 1)$.

The $\sim$ is read as 'is distributed as' and can be used for any distribution. For instance we can write $X \sim \text{Geo}(p)$ to show that X has a geometric distribution with parameter p.

At first glance the uniform looks to be a fairly uninteresting distribution but it is the basis of a lot of statistics. If we want to simulate from any distribution using a computer, we must start with a **random number**. This is a draw from the standard uniform distribution, i.e. a number between 0 and 1 where any outcome is equally likely.

The importance of such numbers rests on the following theorem.

Theorem 5.1 (The Probability Integral Transform)**.** *Let F be a cdf which is a continuous function, strictly increasing on the support of the distribution. This means that the inverse F^{-1} exists and is a function from [0,1] to the support of the distribution (often the real line). Then we have*

1. *Let $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$. Then X is a random variable with cdf F.*
2. *Let X be a random variable with cdf F. Then $F(X) \sim \text{Unif}(0, 1)$.*

Proof. 1. Let $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$. $\forall x \in \mathbf{R}$.

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

since $P(U \leq u) = u$ for $u \in (0, 1)$,

so the cdf of X is F, as claimed.

2. Let X have the cdf F. We want to find the cdf of $Y = F(X)$. From the definition of the cdf Y takes values in $[0, 1]$, $P(Y \leq y)$ equals 0 for $y \leq 0$ and equals 1 for $y \geq 1$. For $y \in [0, 1]$

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$$

Thus Y has a $\text{Unif}(0, 1)$ distribution.

□

The probability integral transform means that if we can simulate from a standard uniform we can simulate from any randon variable with a continuous cdf. There are more general versions that include discrete random variables. The implications of this are enormous. Although we may not use the probability integral transform itself to simulate from a given random variable the fact of its existence allows us to search for faster algorithms knowing that the simulation can be done.

Example 5.10. Supose the pdf of a distribution is given by

$$f(x) = \begin{cases} \exp(-x) & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(this is an example of an exponential distribution as we will see later). First we need to compute the cdf by integrating the pdf.

$$F(x) = \int_0^x \exp(-t)dt = 1 - \exp(-x).$$

We now set $F(X) = U$ which gives

$$U = 1 - \exp(-X).$$

Rearranging gives

$$1 - U = \exp(-X)$$

Since U is a uniform random number in $(0, 1)$ then $1 - U$ is also a uniform random number in $(0, 1)$ so we can write

$$U = 1 - U$$

which gives

$$X = -\ln(U).$$

So if I take the natural log of uniform random numbers on [0,1], I will obtain numbers that are draws from the distribution with pdf as above.

5.2.2 Random Numbers

In order to use the probability integral transform we need to produce random numbers from a standard uniform distribution. There are few ways to generate truely random numbers; measuring radioactive decay may be one. The original Electronic Random Number Indicating Equipment (ERNIE) used for premium bonds in the 1960's counted cosmic rays.

However do we actually want 'truly random' numbers, or numbers that appear to be random but are in fact from a predicatble series? Such numbers are known as *pseudo-random numbers*.

There are a number of good reasons to use pseudo-random numbers rather than 'truly random' numbers. For example

- They are much easier to generate inside the computer, counting cosmic rays or radioactive decay would require specialist hardware.
- We can generate the same set of numbers again if we want to repeat a calculation.

The disadvantage is, of course, that the numbers aren't random! Pseudo-random number generation involves highly complex and advanced number theory and the algorithms are well studied and although they all fail some tests of randomness, for most purposes they can be considered random.

As an example we will look at the *linear congruential* random number generator. Consider a sequence of integers X_n such that

$$X_{n+1} = aX_n + c \qquad \text{mod } m.$$

We start the sequence for some random number seed, X_0. This is multiplied by the constant a, is added to the constant c and the remainder is taken from the integer m. The pseudo-random numbers are given by X_n/m.

The values used for a,c and m vary but, for example, in the book *Numerical Recipes* they use $a = 1664525$, $c = 1013904223$, $m = 2^{32}$.

Although the use of linear congruential generators is still widespread other more advanced algorithms are taking their place. For example the default pseudo-random number generator in R is the *Mersenne-Twister* which uses a rather more complex algorithm. In all cases though the pseudo-random numbers are a sequence which is started with a *random number seed*. If we start the sequence with the same seed we get the same sequence of pseudo-random numbers. This can be an advantage as it allows you to repeat the sequence if you want to. Most computer systems (such as R) will use some function of the time to initiate the sequence if you don't explicitly supply a random number seed.

5.2.3 Random Numbers in R

To generate random numbers in R we use the command

```
runif(n)
```

We can specify the number, n, to be generated. If we want to generate randon numbers from a non-standard uniform distribution we can add the values of the limits (a,b) with

```
runif(n, min=a, max=b)
```

If we want to set the randon number seed we use the command

```
set.seed(seed)
```

where *seed* is an integer. Starting with the same seed will generate the same series of pseudo-random numbers. If you don't specify a seed R generates one from a combination of the time and process id in the computer.

By default R uses the *Mersenne-Twister* to generate the random numbers. However, this can be changed in the call to *set.seed*.

5.2.4 The Bernoulli Distribution

Definition 5.9. A random variable X is said to have bernoulli distribution with probability p, written $X \sim \text{Ber}(p)$ if it can only take the values 0 or 1 and $P(X = 1) = p$.

An example might be a game where there is a probability of p of winning (e.g. tossing an unfair coin where winning means you land heads), and otherwise you lose. Then X is the number of wins (and in 1 game that can only be 1 or 0).

A "game" such as this is known as a *Bernoulli Trial*. The idea of a Bernoulli trial is an important one in statistics and we often refer to *Independent Repeated Bernoulli Trials*. These are a series of trials where the probability of success is constant and each trial is independent of all the others (i.e. the result of one does not affect the results of any of the others).

5.2.5 The Binomial Distribution

Consider an set of n independent repeated Bernoulli trials, what is the probability of getting $(0, 1, \ldots, n)$ successes in our n trials?

If we have x successes we must have $n - x$ failures. The number of ways of getting x successes and $n - x$ failures is given by the binomial coefficient. Since we have a series of independent Bernoulli trials the probability of each success is p and the probability of each failure is $(1 - p)$ the probability of x successes is given by

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Definition 5.10. A discrete random variable X with range space $\{0, 1, 2, \ldots, n\}$ has a *binomial distribution*, with probability p, written $X \sim \text{Bin}(n, p)$, if it can be represented as the sum of n independent Bernoulli trials with probability p.

Note that its probability mass function is therefore as given above.

If we plot the Binomial pmf for n=10 and p=0.2 we get

```
n <- 10
p <- 0.2
exp_data <- tibble(x=seq(0,n,1))
exp_data <- mutate(exp_data, y=choose(n,x)*p^x*(1-p)^(n-x))

ggplot(exp_data, aes(x, y)) +
  geom_segment(aes(xend = x, yend = 0), size = 10, lineend = "butt")
```

For $p = 0.5$ the pmf is symmetric

```
n <- 10
p <- 0.5
bino_data <- tibble(x=seq(0,n,1))
bino_data <- mutate(bino_data, y=choose(n,x)*p^x*(1-p)^(n-x))

ggplot(bino_data, aes(x, y)) +
  geom_segment(aes(xend = x, yend = 0), size = 10, lineend = "butt")
```

And for $p > 0.5$ it is a mirror image of $p < 0.5$. (Why?)

The binomial distribution has a large number of practical applications. It is the correct distribution to use if we are drawing n things from an infinite population each of whom has the same probability of being a success, or if there is a finite population and we are sampling with replacement.

Here is an example from ecology. We wish to know the number of small furry animals in a wood. We capture a number of then, m say, and we put tags on their ears. We let them go and the next day we capture n. If the total number of animals is M then the probability of capturing the animal is m/M (assuming that the tags don't make it easier or harder to catch it). The number of marked animals follows a binomial distribution and if we estimate p $(= m/M)$ from the data we can then estimate the population of small furry animals. We will do estimation of the parameters next term. In ecology this method of estimating animal populations is known as mark-release or capture-recapture.

Another example application is in product testing. Say we have a machine that is producing widgets. If the production process is stable so that the probability of a bad widget (a failure) is constant, then the number of failures (bad widgets) in a sample of n taken from the output of the machine is binomial.

The Bernoulli distribution is of course just the binomial with $n = 1$.

In R we use the commands *dbinom*, *pbinom* and *rbinom* to give the pmf, the cdf and the sample from the binomial distribution. The function *qbinom* gives the **quantile function** - the value of x such that $P(X < x) = q$, where q is a specified probability.

5.2.6 The Poisson Distribution

Definition 5.11. Let $X = 0, 1, 2, \ldots$ be a discrete random variable. X has a **Poisson distribution** with parameter λ if its probability mass function is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

To see where this distribution comes from, consider a sequence of independent repeated Bernoulli trials with n such trials per unit time. If the probability of success is λ/n we can write the rate of successes as λ per unit time. Now we let $n \to \infty$ but keep λ constant.

The number of successes, X, has a Binomial distribution so we have

$$\begin{aligned} P(X = x) &= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)\dots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\ &\to \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{as } n \to \infty, \end{aligned}$$

since, as $n \to \infty$

$$\frac{n(n-1)\dots(n-x+1)}{n^x} \to 1$$

$$\left(1-\frac{\lambda}{n}\right)^n \to \exp(-\lambda)$$

and

$$\left(1-\frac{\lambda}{n}\right)^{-x} \to 1.$$

In R *dpois*, *ppois*, *qpois* and *rpois* give the pmf, the cdf, the quantile function and a random variable for the Poisson disitribution.

Here is a plot of the Poisson pmf for $\lambda = 1$

```
n <- 10
theta=1
poi_data <- tibble(x=seq(0,n,1))
poi_data <- mutate(poi_data, y= dpois(x,lambda=theta))

ggplot(poi_data, aes(x, y)) +
  geom_segment(aes(xend = x, yend = 0), size = 10, lineend = "butt")
```

The Poisson distribution is used for the number of events that occur randomly over time.

In this context, from the above derivation we see that the number of trials has to be large and the probability of success has to be small. One of the early applications of the Poisson distribution, in the 1880's, was to the number of deaths of Prussian cavalry officers from being kicked by their horses. There were a lot of officers and it was quite rare from them to die from being kicked.

Anywhere where we have events occurring randomly that are relatively rare (compared to the number of trials) we can expect to use the Poisson distribution. For example in ecology to count plants a quadrant (a square of wire) is thrown and the number of plant of interest is counted. This is repeated a large number of times. If the distribution of plants in space is random, as opposed to being clustered or spaced on a more regular grid we would expect the number of plants the quadrants to follow a Poisson distribution if they are growing at random locations. If they don't follow the Poisson distribution there is some non-random process at work, either clustering the plants or some process is inhibiting the growth of plants next to each other so that they are more spaced apart than would be expected at random.

5.2.7 The Poisson Process

One of the major applications of the the Poisson distribution is in the study of *stochastic processes*. A stochastic process is a sequence of events where the occurance of the events is governed by some stochastic (random) law. The study of stochastic processes is a little advanced for a first year statistics course (there is a good third year option), but we will mention one stochastic process (without proof): the Poisson process. If events happen at random with a constant rate then the resulting stochastic process is known as a Poisson process. The parameter of the Poisson process, θ, gives the rate at which events happen.

5.2.8 The Exponential Distribution

Definition 5.12. A continuous random variable $X > 0$ has an exponential distribution if its pdf is

$$f(x) = \lambda \exp\{-\lambda x\} \qquad \lambda > 0.$$

Example 5.11. Show that the exponential distribution is a valid probability distribution.

Solution. To do this we need to show that the pdf is a valid pdf. I.e. that $f(x) \geq 0$ for all x and that it integrates to 1. The first part is trivially true for all $\lambda > 0$. To see the second, and noting that the range space is $(0, \infty)$, we have

$$\begin{aligned}\int_0^\infty f(x)dx &= \int_0^\infty \lambda e^{-\lambda x} dx \\ &= \left[-e^{-\lambda x}\right]_0^\infty \\ &= -(0-1) = 1\end{aligned}$$

□

We can derive this as a limiting distribution of a real process, a bit like we did for the Poisson distribution.

Consider some electronic component, these components break at random. The age of the component has no influence on the probability of failure which is constant. As we have seen the number of components breaking over a period of time is given by the Poisson distribution, but consider now the length of time before a failure.

Because there is no 'aging' of the component we can write

$$P(X \leq x_0 + x | X > x_0) = P(X \leq x)$$

where X is the time until the next failure (a random variable).

This says that getting to the time x_0 doesn't affect the time to the next failure, it is simply a restating of our assumption of failures happening at random.

Let $f(x)$ be the pdf of X.

The pdf of $X|X > x_0$ is given by

$$\frac{f(x)}{1 - F(x_0)} \qquad x > x_0.$$

This is the *conditional* pdf and we will define these formally later this term. This is the same as the unconditional pdf but for the normalising constant $(1 - F(x_0))$. This is required to make the pdf integrate to 1 when we remove everything $< x_0$.

(The probability of being less than x_0 is $F(x_0)$. So the integral of $f(x)$ conditional on $x > x_0$ is $1 - F(x_0)$)

From our definition of no aging we have

$$\frac{f(x + x_0)}{1 - F(x_0)} = f(x).$$

Putting x=0 gives

$$\frac{f(x_0)}{1 - F(x_0)} = f(0) = \lambda.$$

This means that $F(X)$ has to satisfy the differential equation

$$\frac{dF(x)}{dx} = \lambda(1 - F(x)),$$

the solution to which is

$$1 - F(x) \propto e^{-\lambda x}.$$

Since x has to be positive we have

$$\lim_{x \to 0} F(x) = 0,$$

giving

$$F(x) = 1 - e^{-\lambda x}.$$

So

$$f(x) = \lambda \exp(-\lambda x), \qquad x > 0; \quad \lambda > 0.$$

The pdf of the standard exponential (with $\lambda = 1$) is plotted here.

Sometimes we add an additional parameter which replaces 0 as the minimum value.

In R the functions *dexp*, *pexp*, *qexp* and *rexp* give the pdf, the cdf, the quantile function and random draws for the exponential disitribution.

We have seen there is an intimate relationship between the Poisson and exponential distributions. If events occur as a Poisson process with rate λ the the inter-event times will have an exponential distribution with the parameter equal to λ.

There is also a relationship to the geometric distribution. Remember that the geometric distribtuion is the distribution of the number of failures before the first success in a series of independent repeated Bernoulli trials. The pmf is given by

$$p(x) = p(1 - p)^x$$

where p is the probability of a success.

Let the number of Bernoulli trials per unit time be m and let λ/m be the probability of a success. N is the number of failures until a success and let T be the time to the first success. Then

$$P\left(t < T \leq t + \frac{1}{m}\right) = P(N = mt) = \frac{\lambda}{m}\left(1 - \frac{\lambda}{m}\right)^{mt}$$

So the limit as m tends to infinity is

$$f(t) = \lim_{m \to \infty} \frac{P\left(t < T \leq t + \frac{1}{m}\right)}{1/m} = \lambda e^{-\lambda t}$$

which is an exponential distribution.

5.3 Distribution summaries

5.3.1 Expectation

For a random variable, the probability distribution gives any probability associated with a particular value (for a discrete variable) or the probability that the variable takes a value within any interval (for a continuous variable). Often we do not require all of that information, and instead look to summaries of the random variable.

The most important of these is **expectation**, which can be thought of as a typical value (the "expected value") or a long-run average of a random variable over many repeated trials of the stochastic experiment. In this course, we will give the classical definition of expectation from a probability distribution. In fact, probability can be defined *from* expectation in some theories (as those who take MTH3041 will see).

Definition 5.13. Let X be a discrete random variable with range space R and probability mass function $p(x)$, then it's **expectation**, written $\mathrm{E}[X]$ is

$$\mathrm{E}[X] = \sum_{x \in R} xp(x).$$

The analogous definition for continuous random variables is as follows:

Definition 5.14. Let X be a continuous random variable with probability denisty function $f(x)$, then it's **expectation**, written $\mathrm{E}[X]$ is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

Example 5.12. Consider the battery powered toy lasting time T from example 5.8. What is the expected time the toy will last, $\mathrm{E}[T]$?

Solution.

$$\begin{aligned}
\mathrm{E}[T] &= \int_{-\infty}^{\infty} tf(t)dt \\
&= \frac{1}{50}\int_{0}^{10} 10t - t^2 dt \\
&= \frac{1}{50}\left[5t^2 - \frac{t^3}{3}\right]_0^{10} \\
&= \frac{1}{50}(500 - \frac{1000}{3}) = \frac{500}{3 \times 50} \\
&= \frac{10}{3} = 3.33.
\end{aligned}$$

□

Example 5.13. Show that the expectation of a geometric random variable with probability of success p is $1/p$.

Solution. As we saw earlier, for geometric random variables $P(X = n) = (1-p)^{n-1}p$. So

$$\begin{aligned}\mathrm{E}[X] &= \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty} n(1-p)^{n-1}p \\ &= p\sum_{n=1}^{\infty} n(1-p)^{n-1} \\ &= \frac{p}{(1-(1-p))^2} = \frac{p}{p^2} = \frac{1}{p}.\end{aligned}$$

□

The expectation of a real-valued function $g(X)$ is given by

$$\mathrm{E}[g(X)] = \sum_{x\in R} g(x)p(x), \qquad \mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx,$$

for discrete and continuous random variables, respectively. For example for continuous X, $\mathrm{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f(x)dx$.

5.3.2 Linearity

Expectation is a *linear* operator i.e., in both the discrete and continuous cases if X and Y are random variables, with a, b, c real constants then,

$$\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c.$$

This can be proved later, when we tackle joint distributons.

As an example of the approach to proving this in general, consider the continuous case for $\mathrm{E}[aX + b]$.

$$\begin{aligned}\mathrm{E}[aX+b] &= \int_{-\infty}^{\infty} (ax+b)f(x)dx \\ &= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \\ &= a\mathrm{E}[X] + b.\end{aligned}$$

Example 5.14. Let $X \sim \mathrm{Bin(n, p)}$. Find $\mathrm{E}[X]$.

Solution. Though it is possibe to solve this by finding

$$\mathrm{E}[X] = \sum_{x=1}^{n} xP(X=x)$$

directly, it is in fact easier to consider X as the sum of n independent Bernoulli random variables directly. Let $X_i \sim \mathrm{Ber}(p)$ for $i = 1, \ldots, n$. Then

$$\mathrm{E}[X_i] = \sum_{k=0}^{1} kP(X=k) = 0\cdot(1-p) + 1\cdot p = p,$$

and we have

$$\begin{aligned} X = X_1 + X_2 + \cdots + X_n \implies \\ \mathrm{E}[X] &= \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n] \\ &= np, \end{aligned}$$

using the linearity of expectation.

□

5.3.3 Moments

Definition 5.15. The **rth moment** of a random variable X is $\mathrm{E}[X^r]$, for $r = 1, 2, \ldots$.

So, for example, $\mathrm{E}[X]$ is the first moment of X and $\mathrm{E}\left[X^2\right]$ is the second moment.

Definition 5.16. The **rth central moment** of a random variable X is $\mathrm{E}[(X-\mu)^r]$, for $r = 1, 2, \ldots$, where $\mu = \mathrm{E}[X]$.

Note that the 1st central moment of X is always 0 as

$$\mathrm{E}[X - \mu] = \mathrm{E}[X - \mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[X] = 0.$$

Definition 5.17. The **variance** of a random variable X, $\mathrm{Var}[X]$, is the 2nd central moment of X.

Note, using the linearity of expectation, we have the identity

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}\left[(X-\mu)^2\right] \\ &= \mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\mu\mathrm{E}[X] + \mu^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2, \end{aligned}$$

which is usually much easier to use for finding $\mathrm{Var}[X]$ in practice. The variance measures the "long run average" of the squared distance between the random variable and its expectation, hence it is a measure of the expected "spread" of the variable.

Example 5.15. Let $X \sim \mathrm{Unif}(0, 2)$. Find $\mathrm{E}[X]$ and $\mathrm{Var}[X]$.

Solution. The pdf of the uniform distribution on $[0, 2]$ is $f(x) = \frac{1}{2}$.

$$\begin{aligned} \mathrm{E}[X] = \int_0^2 x f(x) dx = \int_0^2 \frac{x}{2} dx = \left[\frac{x^2}{4}\right]_0^2 \\ &= 1, \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}\left[X^2\right] = \int_0^2 x^2 f(x) dx = \int_0^2 \frac{x^2}{2} dx = \left[\frac{x^3}{6}\right]_0^2 \\ &= \frac{4}{3}. \end{aligned}$$

So

$$\operatorname{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{1}{3}.$$

□

Example 5.16. For $X \sim \mathrm{Ber}(p)$,

$$\begin{aligned}\operatorname{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\ &= p - p^2 \\ &= p(1-p).\end{aligned}$$

Example 5.17. Let $X \sim \mathrm{Exp}(\lambda)$ with

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\operatorname{Var}[X]$.

Solution. We first require $\mathrm{E}[X]$, which is

$$\begin{aligned}\mathrm{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \left[x e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= \left[\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty \\ &= \frac{1}{\lambda}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{E}\left[X^2\right] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \left[x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx \\ &= \left[-\frac{2x}{\lambda} e^{-\lambda x}\right]_0^\infty + 2\int_0^\infty \frac{1}{\lambda} e^{-\lambda x} dx \\ &= 2\left[-\frac{1}{\lambda^2} e^{-\lambda x}\right]_0^\infty \\ &= \frac{2}{\lambda^2}.\end{aligned}$$

So, we have

$$\begin{aligned}\operatorname{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2}\end{aligned}$$

□

For $X_1, \ldots, X_n$ independent and $X = \sum_{i=1}^{n} X_i$, we have

$$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i]$$

(this will be proved later in the course, when we meet covariance). For now we can use it when we have independent random variables.

Example 5.18. Let $X \sim \text{Bin}(n, p)$. As seen previously, $X = \sum_{i=1}^{n} X_i$ and $X_i \sim \text{Ber}(p)$, so

$$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

Instead of variance, we often work with the standard deviation of a random variable.

Definition 5.18. The **standard deviation** of a random variable, X, is defined to be

$$\text{sd}[X] = \sqrt{\text{Var}[X]}.$$

Definition 5.19. The **rth standardised central moment** of a random variable X is $\text{E}\left[\frac{(X-\mu)^r}{\sigma^r}\right]$, where $\sigma = \text{sd}[X]$ is the standard deviation of X.

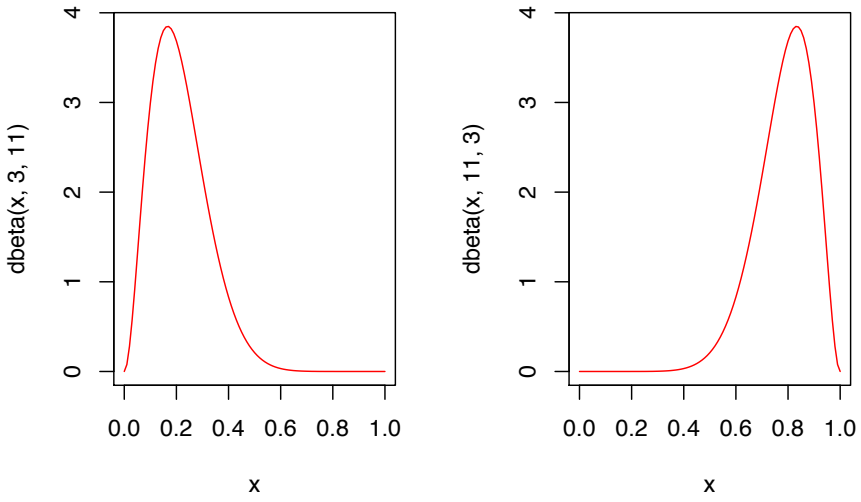
Definition 5.20. The **skewness** of a random variable X, also known as the skew of X, is the 3rd standardised central moment of X.

The skewness measures the asymmetry of a distribution. If a random variable has a symmetric distribution, average cubic departures from its mean will cancel out and the skew will be zero. Negatively and positively skewed distributions typically arise from asymteric tails, where negative skew, typically has a longer left tail, and positive skew has a typically longer right tail.

```
skewBeta <- function(a,b){(2*(b-a)*sqrt(a+b+1))/((a+b+2)*sqrt(a*b))}
skewBeta(9,3)
```

```
## [1] -0.5947617
```

```
par(mfrow=c(1,2))
x <- seq(from=0,to=1,by=0.01)
plot(x,dbeta(x,3,11),xlim=c(0,1),col=2,type='l')
plot(x,dbeta(x,11,3),xlim=c(0,1),col=2,type='l')
```



5.3.4 Quantiles

There are other important distribution summaries away from moments. Quantiles are perhaps the most useful of these as they report key probabilities.

Quantiles are cut-points that divide the range of a random variable with a probability distribution into intervals of equal probability. There is one less quantile than the number of intervals created.

Let the cdf of a random variable be $F(x)$. The **q-quantiles** are the set $\{F^{-1}(\frac{1}{q}), F^{-1}(\frac{2}{q}), \ldots, F^{-1}(\frac{q-1}{q})\}$.

Some quantiles have special names, alluding to the number of intervals $(q+1)$ they determine.

- The 2-quantile is called the **median**, m, so that $P(X \leq m) = 1/2$.
- The 3-quantiles are called **terciles**, so that $P(X \leq w_1) = 1/3$, and $P(X \leq w_2) = 2/3$, wher w_1 and w_2 are the upper and lower terciles.
- The 4-quantiles are called **quartiles**. The first quartile is w_1 with $P(X \leq w_1) = 1/4$ and is called the lower quartile. The first quartile is w_3 with $P(X \leq w_3) = 3/4$ and is called the upper quartile. $w_3 - w_1$ is known as the **inter-quartile range** and is another measure of spread. Note that the 2nd quartile is the median.
- There are also **deciles**, (10-quantiles) and **percentiles**, (100-quantiles).

Example 5.19. $X \sim \text{Unif}(0, 3)$. Find the median and $\text{E}[X]$.

Solution. First,

$$\begin{aligned}\text{E}[X] &= \int_0^3 x f(x) dx = \int_0^3 \frac{x}{3} dx \\ &= \left[\frac{x^2}{6}\right]_0^3 = \frac{3}{2}.\end{aligned}$$

Next, the median, M is found via

$$\int_0^M f(x)dx = \frac{1}{2} = \left[\frac{x}{3}\right]_0^M = \frac{M}{3}$$
$$\implies M = \frac{3}{2}.$$

□

Example 5.20. Let $X \sim \text{Exp}(\lambda)$. Find the Median and inter quartile range of X.

Solution. The pdf of X is

$$f(x) = \lambda e^{-\lambda x}; \qquad x > 0.$$

The median, M, is such that

$$\int_0^M \lambda e^{-\lambda x}dx = \frac{1}{2} \iff$$
$$\left[-e^{-\lambda x}\right]_0^M = 1 - e^{-\lambda M} = \frac{1}{2}$$
$$\implies e^{-\lambda M} = \frac{1}{2}$$
$$\implies M = \frac{1}{\lambda}\log(2).$$

$IQR = w_3 - w_1$. w_3 is such that

$$\int_0^{w_3} \lambda e^{-\lambda x}dx = \frac{3}{4} = 1 - e^{-\lambda w_3} \iff$$
$$w_3 = \frac{1}{\lambda}\log(4).$$

w_1 is such that

$$\int_0^{w_1} \lambda e^{-\lambda x}dx = \frac{1}{4} = 1 - e^{-\lambda w_1} \iff$$
$$w_1 = \frac{1}{\lambda}\log(\frac{4}{3}).$$

So

$$IQR = \frac{1}{\lambda}(\log(4) - \log(\frac{4}{3})) = \frac{1}{\lambda}\log(3).$$

□

5.3.5 The Gamma distribution and Gamma function

A random variable has a **Gamma distribution** if it has a pdf, $f(x)$, with

$$f(x) = \frac{1}{C}\beta^\alpha x^{\alpha-1}e^{-\beta x}, \qquad \text{for } x > 0.$$

for known parameters $\alpha, \beta > 0$ that control the shape of the distribution. The value C is chosen to ensure that $f(x)$ integrates to 1. So

$$C = \int_0^\infty \beta^\alpha x^{\alpha-1}e^{-\beta x}dx.$$

Using the change of variables $u = \beta x$ we have

$$C = \int_0^\infty u^{\alpha-1} e^{-u} du,$$

which means that C only depends on α. C is a function of α known as the **gamma function**, usually written

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du.$$

Though this is not usually analytically tractable, $\Gamma(\cdot)$ has some useful properties. Firstly, note $\Gamma(1) = 1$. Also, by integrating by parts we have

$$\begin{aligned}\Gamma(\alpha+1) &= \left[-\frac{1}{\alpha+1} u^{\alpha+1} e^{-u}\right]_0^\infty + \alpha \int_0^\infty u^{\alpha-1} e^{-u} du \\ &= \alpha\Gamma(\alpha).\end{aligned}$$

This is a general property for any $\alpha > 0$. Note, if α is a positive integer, $\Gamma(\alpha) = (\alpha-1)!$ (though it can often help avoid mistakes to work with gamma functions directly, rather than converting to factorials).

The expectation of a gamma distribution is found using a useful trick for solving integrals for this type of distribution.

$$\begin{aligned}\mathrm{E}[X] &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\beta x} dx \\ &= \frac{\beta^\alpha \Gamma(\alpha+1)}{\Gamma(\alpha)\beta^{\alpha+1}} \int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\beta x} dx \\ &= \frac{\beta^\alpha \Gamma(\alpha+1)}{\Gamma(\alpha)\beta^{\alpha+1}} \\ &= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} \\ &= \frac{\alpha}{\beta},\end{aligned}$$

where the integral disappears because we use the fact that the Gamma pdf integrates to 1.

5.4 The Normal Distribution

5.4.1 A VERY brief introduction to multiple integration

Multiple integration is a fundamental requirement for mathematics, and is very important within statistics and probability. It is typically required as soon as we cover joint distributions (in a few lectures), arising from the natural extension of the cdf of a random variable, $F(X)$ to multiple related random variables, $F(X, Y, ...)$. (E.g. random variables rainfall and temperature cannot be independent and we might want to forecast them at the same time).

We will see joint distributions and joint cdfs as multiple integrals soon, but it makes sense to briefly look at multiple integration first so we can use it for the Normal distribution.

You will see multiple integration formally in Methods next term. All we need is the very basics. If we have a function in n variables, $f(x_1, x_2, , \ldots, x_n)$, we can integrate it over a domain, D, within those n dimensions to find the volume of that domain under the surface defined by f.

We would write this as

$$\int_{D_{x_1}} \int_{D_{x_2}} \cdots \int_{D_{x_n}} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

where the D_{x_i} represent expressions for the domains in each variable. For example, we might have $x \in [-1, 1]$, $y \in [0, 2]$ and $f(x, y) = x^4 - 2y$ and the integral is

$$\int_0^2 \int_{-1}^1 (x^4 - 2y) dx dy.$$

The notation means that you first do the interior integral, then the exterior. So in the above example,

$$\begin{aligned}
\int_0^2 \int_{-1}^1 (x^4 - 2y) dx dy &= \int_0^2 \left(\int_{-1}^1 (x^4 - 2y) dx \right) dy \\
&= \int_0^2 \left[\frac{x^5}{5} - 2yx \right]_{-1}^1 dy \\
&= \int_0^2 (\frac{2}{5} - 2y + 2y) dy = \frac{4}{5}.
\end{aligned}$$

Often the order you do things can make life easier or harder. Above, the terms in y cancelled out. When f is well behaved (and it always will be in this course), the order of integration can be swapped and so finding the right way to do things is important.

Limits can be more complicated. E.g., in the above problem, suppose $x \in [-1, 1]$ as before, but $y \in [-x^2, x^2]$. Then, it might make sense to swap the order of integration to work with

$$\int_{-1}^1 \int_{-x^2}^{x^2} (x^4 - 2y) dy dx.$$

Sometimes, integrals are very hard to compute but can be simplified by transforming variables. Consider, for example, the function $f(x, y) = x$ over the unit circle $x^2 + y^2 \leq 1$. This particular integral is tricky because the limits are hard to work with. Changing variables (like substitution) can make these integrals very easy. The general idea in 2 dimensions is as follows (the extension to n dimensions is immediate and will be covered in Methods: Let

$$x = x(u, v), \qquad y = y(u, v)$$

functions of new variables u and v (so $(x, y) \to (u, v)$) Then

$$\int \int_D f(x, y) dx dx = \int \int_\Gamma f(x(u, v), y(u, v)) |J(u, v)| du dv,$$

where Γ is the image of D in the new coordinates and J is the Jacobian given by

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u}.$$

A particularly useful transformation, which can be used for our circle example, is the transformation to *polar coordinates*.

$$x = r \cos \theta, \qquad y = r \sin \theta,$$

where r represents a radius and θ an angle in the transformed coordinates. The Jacobian of this transformation is just r and, for our problem above we have $x^2 + y^2 = r^2 \leq 1$ and $0 \leq \theta \leq 2\pi$. So the integral becomes

$$\int_0^{2\pi} \int_0^1 r^2 \cos \theta dr d\theta = \int_0^{2\pi} \frac{1}{3} \cos \theta d\theta = 0$$

You will learn a great deal more about multiple integration in Methods.

5.4.2 The Normal distribution

Definition 5.21. Let X be a random variable with $X \in (-\infty, \infty)$. X has a *Normal* or *Gaussian* distribution with mean μ and variance σ^2, written $X \sim \mathrm{N}(\mu, \sigma^2)$ if its pdf is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

The Normal distribution was first used by the German mathematician Gauss to discribe the errors in the measured positions of stars in 1816 (although Laplace had used it as an approximation to the Binomial in the eighteenth century).

The *standard* Normal is given by setting $\mu = 0$ and $\sigma = 1$:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \qquad -\infty < x < \infty,$$

and is often written $\phi(x)$.

It has the well known bell-shape shown here.

It isn't possible to evaluate the definite integral

$$G(x) = \int_{-\infty}^{x} \exp\left(-\frac{1}{2}t^2\right) dt \propto F(x),$$

but we can evaluate $G(\infty)$.

Proposition 5.1.

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx = \sqrt{2\pi}$$

Proof. To prove the proposition we use an amazing trick:- rather than try to integrate

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx,$$

we multiply the integral by itself and integrate

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx.$$

Here, x is a dummy variable, so we can replace it by y (or anything we like) in the second integral, giving

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy.$$

Rearranging gives

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy.$$

This works because when integrating wrt y, the integral in x is a constant (so can come inside the integral). We now change to polar variables, i.e. to $r^2 = x^2 + y^2$ and $\theta = tan(x/y)$. This gives

$$\int_0^{2\pi}\int_0^{\infty}\exp\left(-\frac{r^2}{2}\right)rdrd\theta$$

Note the extra r that comes from the transformation (again, if the reason why this appears is not familiar now, it will be after transformations are tackled in Methods and later in Term 2).

Making a second substitution $u = r^2/2$, $du = rdr$ gives

$$\int_0^{2\pi}\int_0^{\infty}\exp\left(-u\right)dud\theta$$

which is

$$\int_0^{2\pi}1d\theta = 2\pi$$

Therefore

$$\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}x^2\right)dx = \sqrt{2\pi}$$

□

This shows that the Normal pdf defines a valid probability distribution.

The cdf cannot be written down analytically but can be evaluated numerically. Tables of the Normal cdf are widely available. The cdf of the standard Normal is very often written as $\Phi(x)$.

In R we have the following functions for the pdf, the cdf, the quantile function and to simulate Normal variables: *dnorm*, *pnorm*, *qnorm* and *rnorm*.

5.4.3 Moments

Consider the standard Normal distribution. The r^{th} moment of this distribution is given by

$$\mathrm{E}\left[X^r\right] = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}x^r\exp\left(-\frac{1}{2}x^2\right)dx.$$

We can write this as

$$\mathrm{E}\left[X^r\right] = \int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}x^r\exp\left(-\frac{1}{2}x^2\right)dx + \int_0^{\infty}\frac{1}{\sqrt{2\pi}}x^r\exp\left(-\frac{1}{2}x^2\right)dx.$$

Now consider r odd. By symmetry we have

$$\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}x^r\exp\left(-\frac{1}{2}x^2\right)dx = -\int_0^{\infty}\frac{1}{\sqrt{2\pi}}x^r\exp\left(-\frac{1}{2}x^2\right)dx$$

So all the odd moments are zero.

Consider now the even moments, again by symmetry we have

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} x^r \exp\left(-\frac{1}{2}x^2\right) dx = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} x^r \exp\left(-\frac{1}{2}x^2\right) dx.$$

So the r^{th} moment is given by

$$E[X^r] = \int_{0}^{\infty} \frac{2}{\sqrt{2\pi}} x^r \exp\left(-\frac{1}{2}x^2\right) dx.$$

Changing variables $t = x^2/2$ gives

$$E[X^r] = \sqrt{\frac{2}{\pi}} 2^{\frac{r+1}{2}} \int_{0}^{\infty} t^{\frac{r-1}{2}} e^{-t} dt == \frac{2^{\frac{r}{2}}}{\sqrt{\pi}} \Gamma\left(\frac{r+1}{2}\right),$$

as the integral is a Γ function. Now

$$\Gamma(n/2) = \frac{(n-2)!!\sqrt{\pi}}{2^{(n-1)/2}}$$

where $n!! = n(n-2)\ldots 1$. So

$$E[X^r] = (r-1)(r-3)\cdots 3 \cdot 1,$$

and since $E[X] = 0$, we have that the variance (for the standard Normal) is 1.

Using the linearity of expectation we can show that for the general Normal distribution, $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. Often, for example in the R rountines, μ is refered to as the mean, σ^2 as the variance and σ as the standard deviation.

The parameters μ and σ^2 are *location* and *scale* parameters. Changing μ moves the location of the centre of the distribution while changing σ^2 broadens or narrows the pdf. This is illustrated here.

5.4.4 Applications of the Normal Distribution

The Normal distribution is the most widely used distribution in statistics and many statisticians will not need to use another distribution during their entire career. For example the distribution of the height of female students in this room will almost certainly be Normal.

One reason for the ubiquity of the Normal is the Central Limit Theorem.

Theorem 5.2. *The Central Limit Theorem states that, under certain conditions, if $\bar{X}$ is the mean of a sample of size n, then the distribution of*

$$\frac{\bar{X} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}(X)/n}}$$

in the limit as $n \to \infty$, is the standard Normal regardless of the original distribution of X.

We will prove the Central Limit Theorem next term but the important thing at the moment is that regardless of the underlying distribution the mean of a large number of observations will have a Normal distribution.

5.4.5 Relationship to other distributions

The Normal distribution is often used as an approximation to other distributions. We have seen that if n is large and p small then the binomial can be approximated by the Poisson. Remember we used the limit to derive the Poisson. If np is large the Normal can be used as an approximation to the binomial. Similarly for the Poisson distribution with parameter λ if λ is large the Normal is a good approximation.

5.5 Sampling and Samples

When we collect data we are taking a *sample* from a population. For example consider the heights of young women in the UK. we might be interested in knowing the mean height and the variance or the form of the distribution. This information could be important for a clothing manufacturer. It would be impractical to measure the height of all the young women in the UK, the population, so instead we would take a sample (for example the women in this class) and use the sample to make inferences about the population.

We will use n to denote the size of the sample and write the random variables in the sample as $X_1, X_2, \ldots, X_n$, or X_i, $i = 1, \ldots, n$; and the actual sampled values as $x_1, x_2, \ldots x_n$.

If the population has a distribution with pdf $f(x)$ then each X_i has pdf $f(x)$. Since each member of the sample has the same distribution and since they are independent we refer to the sample as being independent, identically distributed. This is abbreviated to *iid*. So if we write

$$X_i \sim_{iid} N(0, \sigma^2) \qquad i = 1, \ldots, n,$$

we are saying we have a sample size n each with an independent Normal distribution having zero mean and variance σ^2.

5.5.1 Sample moments

In the same way as we have moments for a distribution we can define (and calculate) *sample moments*.

For example the sample mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the sample variance by

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

(the mean and variance of the actual sampled values). We will now derive the expected value of $\bar{X}$. By definition

$$\mathrm{E}\left[\bar{X}\right] = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right].$$

By linearity of expectation

$$\mathrm{E}\left[\bar{X}\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[X_i\right].$$