Chapter 4: Continuous Random Variables

4.1 Introduction

When Mozart performed his opera *Die Entführung aus dem Serail*, the Emperor Joseph II responded wryly, 'Too many notes, Mozart!'

In this chapter we meet a different problem: *too many numbers!*

We have met ***discrete*** random variables, for which we can *list all the values and their probabilities, even if the list is infinite:*

e.g. for $X \sim \text{Geometric}(p)$,

x	0	1	2	...
$f_X(x) = \mathbb{P}(X = x)$	p	pq	pq^2	...

But suppose that X takes values in a *continuous set, e.g.* $[0, \infty)$ *or* $(0, 1)$.

We can't even begin to list all the values that X can take. For example, how would you list all the numbers in the interval $[0,\ 1]$?

- the smallest number is 0, but what is the next smallest? 0.01? 0.0001? 0.0000000001? We just end up talking *nonsense.*

In fact, there are so many numbers in any continuous set that *each of them must have probability 0.*

If there was a probability > 0 for all the numbers in a continuous set, however 'small', there simply wouldn't be enough probability to go round.

> *A continuous random variable takes values in a continuous interval* $(a,\ b)$.
>
> *It describes a continuously varying quantity such as time or height.*
>
> *When* X *is continuous,* $\mathbb{P}(X = x) = 0$ *for ALL* x.
>
> *The probability function is meaningless.*

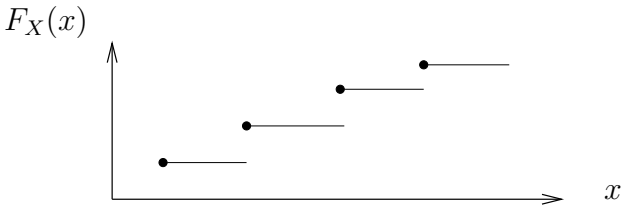
Although we cannot assign a probability to any *value* of X, we *are* able to assign probabilities to *intervals:*
eg. $\mathbb{P}(X = 1) = 0$, *but* $\mathbb{P}(0.999 \leq X \leq 1.001)$ *can be* > 0.

This means we should use *the distribution function,* $F_X(x) = \mathbb{P}(X \leq x)$.

The cumulative distribution function, $F_X(x)$

Recall that for ***discrete*** random variables:

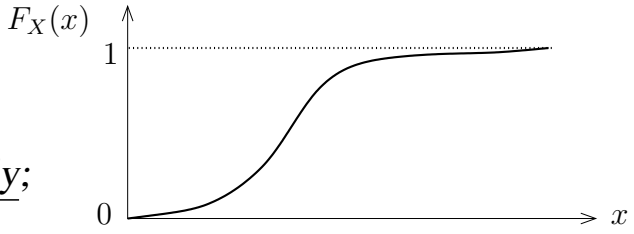
- $F_X(x) = \mathbb{P}(X \leq x)$;
- $F_X(x)$ *is a step function: probability accumulates in discrete steps;*



- $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F(b) - F(a)$.

For a ***continuous random variable:***

- $F_X(x) = \mathbb{P}(X \leq x)$;
- $F_X(x)$ *is a continuous function: probability accumulates continuously;*



- *As before,* $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F(b) - F(a)$.

However, for a *continuous* random variable,

$$\mathbb{P}(X = a) = 0.$$

So it makes *no difference* whether we say $\mathbb{P}(a < X \leq b)$ *or* $\mathbb{P}(a \leq X \leq b)$.

For a continuous random variable,

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a).$$

This is ***not*** true for a discrete random variable: in fact,

For a discrete random variable with values $0, 1, 2, \ldots$,

$$\mathbb{P}(a < X < b) = \mathbb{P}(a + 1 \leq X \leq b - 1) = F_X(b - 1) - F_X(a).$$

Endpoints are not important for continuous r.v.s.
Endpoints are very important for discrete r.v.s.

4.2 The probability density function

Although the cumulative distribution function gives us an interval-based tool for dealing with continuous random variables, it is not very good at telling us what the distribution *looks like.*
For this we use a different tool called the *probability density function.*

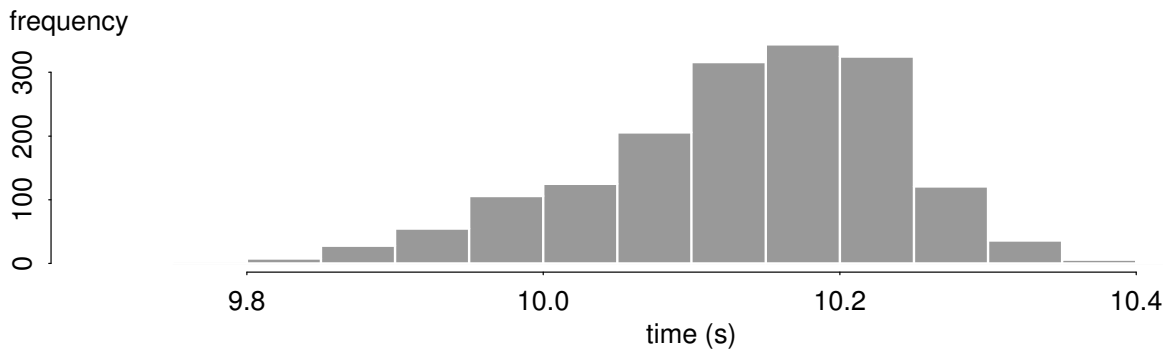
The probability density function (p.d.f.) is the best way to describe and recognise a continuous random variable. We use it all the time to calculate probabilities and to gain an intuitive feel for the shape and nature of the distribution. Using the p.d.f. is like recognising your friends by their faces. You can chat on the phone, write emails or send txts to each other all day, but you never really know a person until you've seen their face.

Just like a cell-phone for keeping in touch, the cumulative distribution function is a tool for facilitating our interactions with the continuous random variable. However, we never really understand the random variable until we've seen its 'face' — the probability density function. Surprisingly, it is quite difficult to describe exactly what the probability density function *is*. In this section we take some time to motivate and describe this fundamental idea.

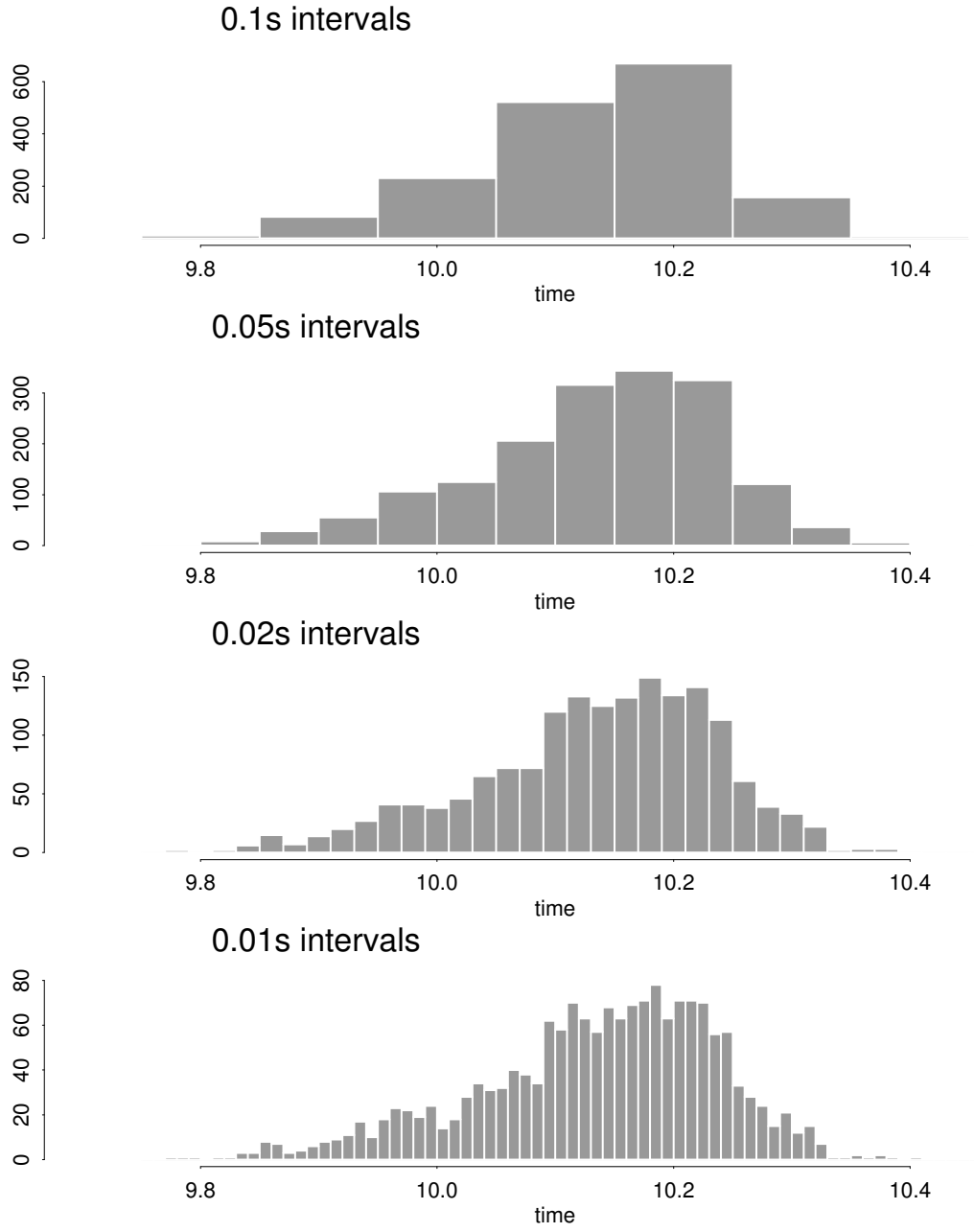
All-time top-ten 100m sprint times

The histogram below shows the best 10 sprint times from the 168 all-time top male 100m sprinters. There are 1680 times in total, representing the top 10 times up to 2002 from each of the 168 sprinters. Out of interest, here are the summary statistics:

```
Min. 1st Qu. Median  Mean 3rd Qu.  Max.
9.78 10.08   10.15  10.14 10.21   10.41
```



We could plot this histogram using different time intervals:



We see that *each histogram has broadly the same shape, although the heights of the bars and the interval widths are different.*

The histograms tell us the most intuitive thing we wish to know about the distribution: its *shape*:

- the *most probable* times are *close to 10.2 seconds;*
- the distribution of times has a *long left tail (left skew);*
- times below 10.0s and above 10.3 seconds have *low probability.*

We could fit a curve over any of these histograms to show the desired shape, but the problem is that *the histograms are not standardized:*

• every time we change the interval width, *the heights of the bars change.*

How can we derive a curve or function that captures the common shape of the histograms, but keeps a constant height? What should that height be?

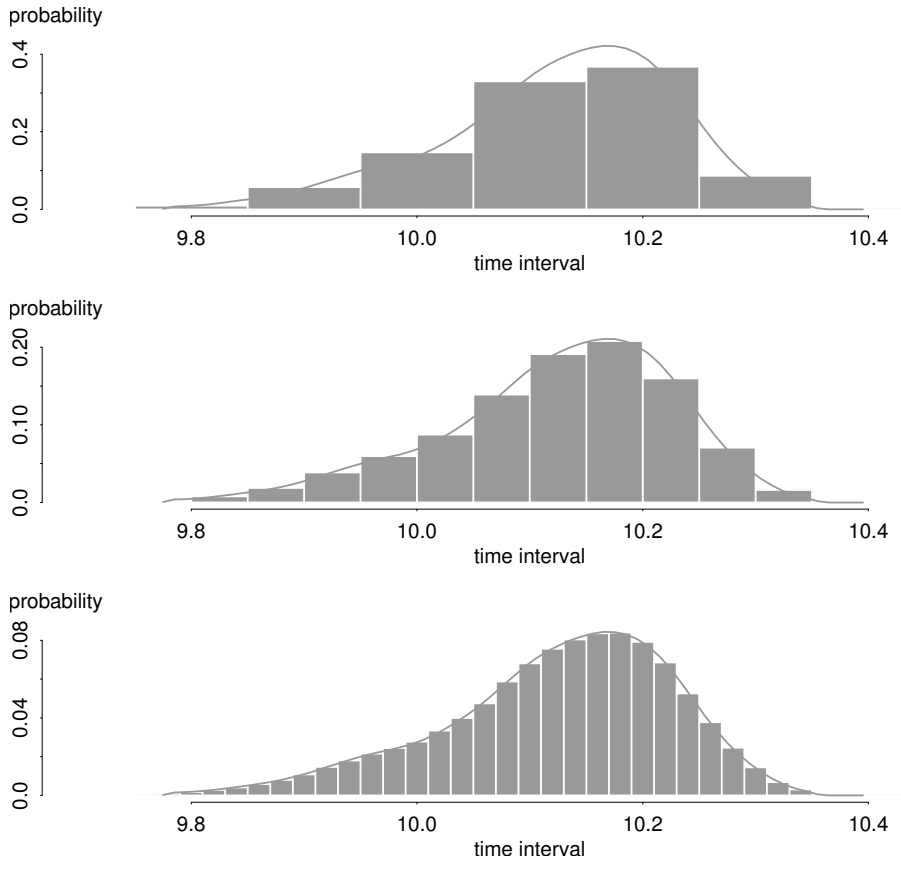
The standardized histogram

We now focus on an *idealized* (smooth) version of the sprint times distribution, rather than using the exact 1680 sprint times observed.

We are aiming to derive a curve, or function, that captures the shape of the histograms, but will keep the same height for any choice of histogram bar width.

First idea: plot the probabilities instead of the frequencies.

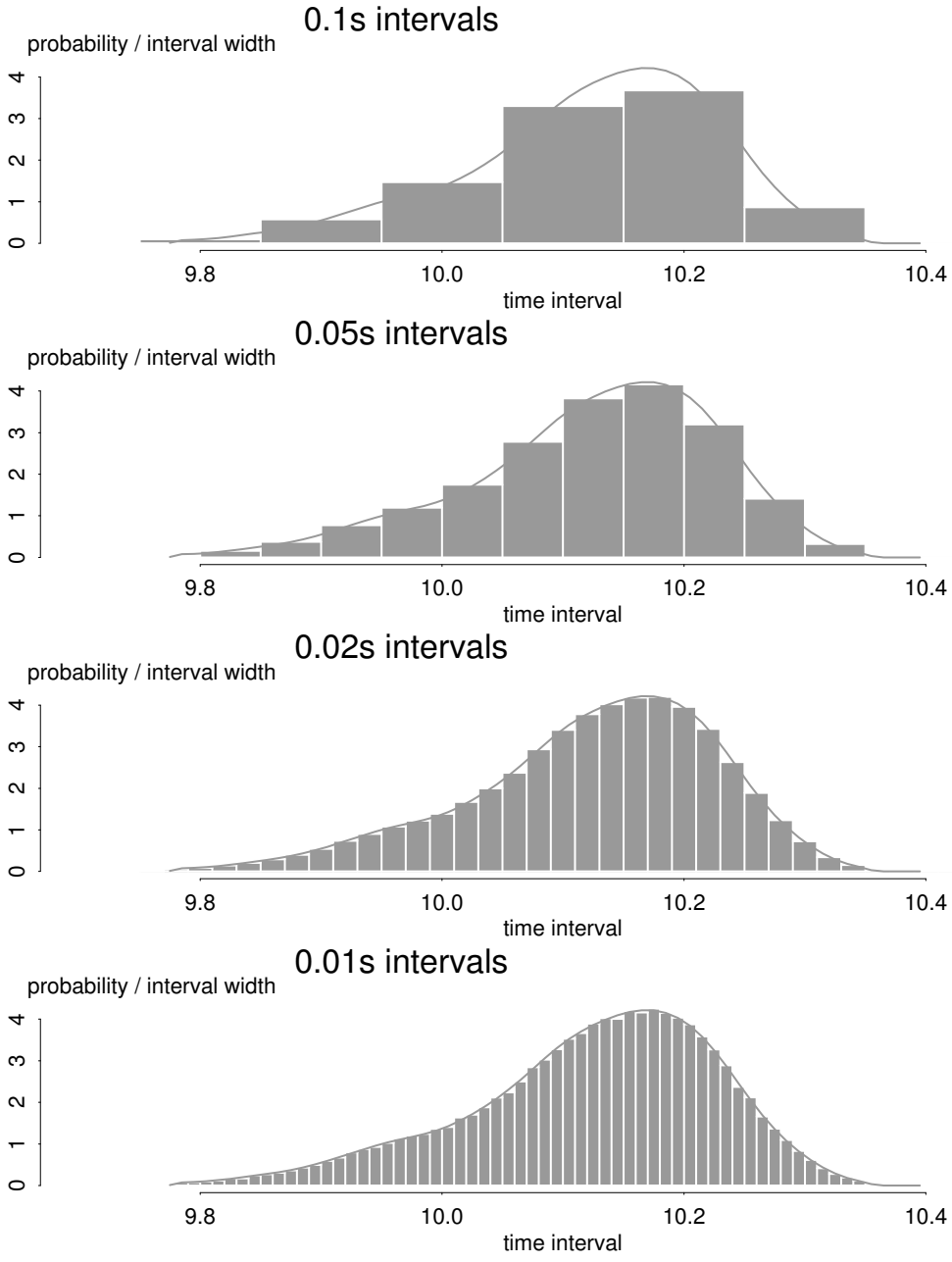
The height of each histogram bar now represents the probability of getting an observation in that bar.



This doesn't work, because *the height (probability) still depends upon the bar width. Wider bars have higher probabilities.*

Second idea: plot the probabilities divided by bar width.

The height of each histogram bar now represents the probability of getting an observation in that bar, divided by the width of the bar.



This seems to be exactly what we need! *The same curve fits nicely over all the histograms and keeps the same height regardless of the bar width.*

These histograms are called *standardized histograms.*

The nice-fitting curve is *the probability density function.*

But... what *is* it?!

The probability density function

We have seen that there is a single curve that fits nicely over any standardized histogram from a given distribution.

This curve is called the *probability density function (p.d.f.).*

We will write the p.d.f. of a continuous random variable X as

The p.d.f. $f_X(x)$ is

However, as the histogram bars of the standardized histogram get narrower, the bars get closer and closer to the p.d.f. curve. The p.d.f. is in fact the

What is the height of the standardized histogram bar?

For an interval from x to $x+t$, the standardized histogram plots *the probability of an observation falling between x and $x+t$, divided by the width of the interval, t.*

Thus the height of the standardized histogram bar over the interval from x to $x+t$ is:

Now consider the limit as the histogram bar width (t) goes to 0:

This expression should look familiar: it is

The probability density function (p.d.f.) is therefore

Formal definition of the probability density function

Definition: Let X be a continuous random variable with distribution function $F_X(x)$. The **probability density function (p.d.f.)** of X is defined as

It gives:

-
-

Using the probability density function to calculate probabilities

As well as showing us the shape of the distribution of X, the probability density function has another major use:

-

Suppose we want to calculate

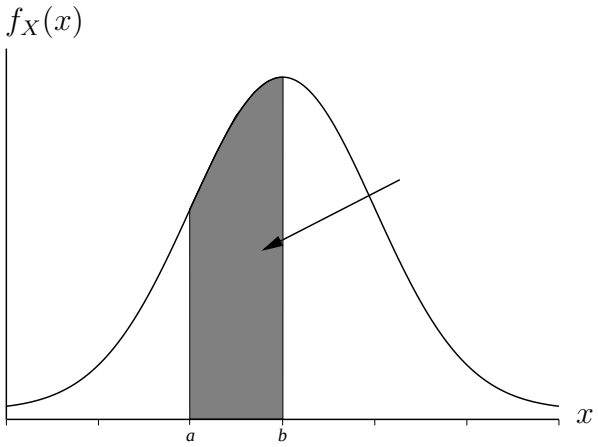
We already know that:

But we also know that:

This is a very important result:

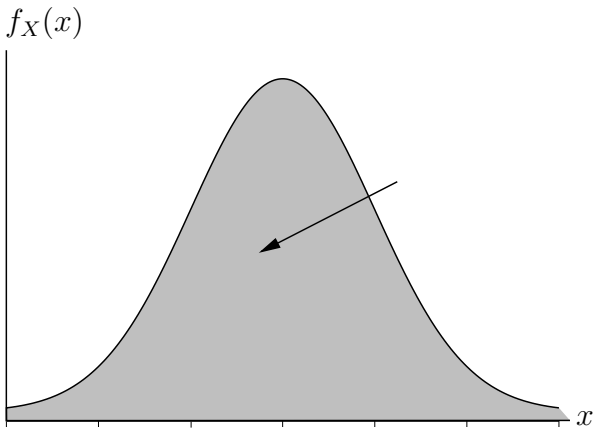
Let X be a continuous random variable with probability density function $f_X(x)$. Then

This means that



The **total area under the p.d.f. curve** is:

This says that the total area under the p.d.f. curve is equal to the total probability that X takes a value between $-\infty$ and $+\infty$, which is



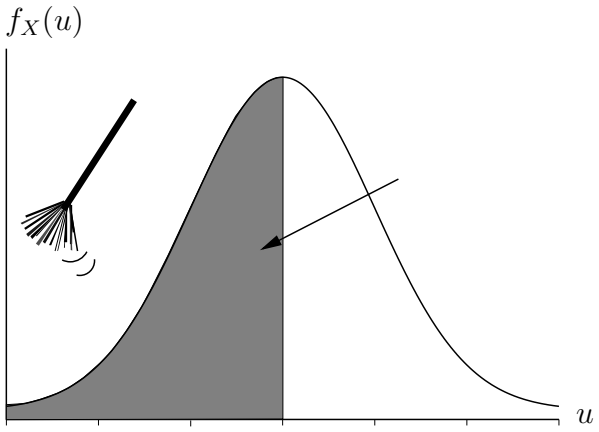
Using the p.d.f. to calculate the distribution function, $F_X(x)$

Suppose we know the probability density function, $f_X(x)$, and wish to calculate the distribution function, $F_X(x)$. We use the following formula:

Proof:

Using the dummy variable, u:

Writing $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$ means:



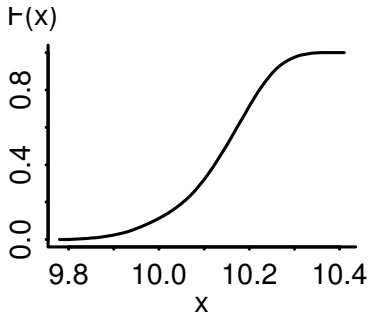
Writing $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$ is

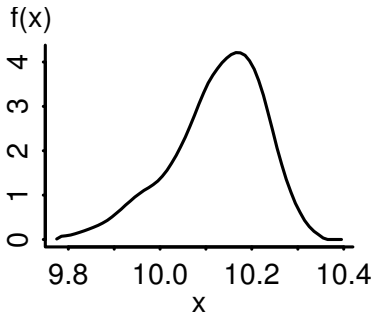
In words, $\int_{-\infty}^{x} f_X(x)\,dx$ means: *integrate $f_X(x)$ as x ranges from $-\infty$ to x. It's nonsense!*

How can x range from $-\infty$ to x?!

Why do we need $f_X(x)$? Why not stick with $F_X(x)$?

These graphs show $F_X(x)$ and $f_X(x)$ from the men's 100m sprint times (X is a random top ten 100m sprint time).



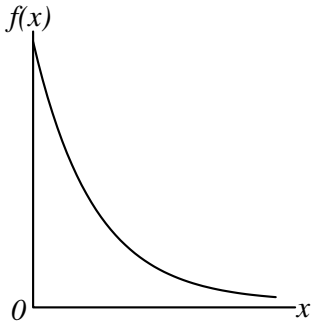


Just using $F_X(x)$ gives us very little intuition about the problem. For example, which is the region of highest probability?

Using the p.d.f., $f_X(x)$, we can see that it is about

Using the c.d.f., $F_X(x)$, we would have to *inspect the part of the curve with the steepest gradient: very difficult to see.*

Example of calculations with the p.d.f.



Let $f_X(x) = \begin{cases} k\,e^{-2x} & \text{for } 0 < x < \infty, \\ 0 & \text{otherwise.} \end{cases}$

(i) Find the constant k.

(ii) Find $\mathbb{P}(1 < X \leq 3)$.

(iii) Find the cumulative distribution function, $F_X(x)$, for *all* x.

(i)

Total area under the p.d.f. curve is 1:

The p.d.f. is NOT a probability:

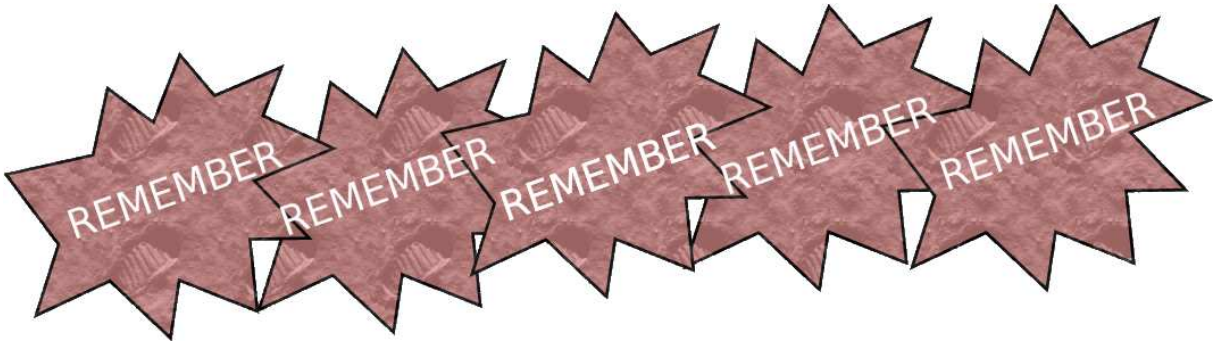
Calculating probabilities:

1. If you only need to calculate *one* probability $\mathbb{P}(a \leq X \leq b)$:

2. If you will need to calculate *several* probabilities, it is easiest to

   Then use:

Endpoints:



4.3 The Exponential distribution

When will the next volcano erupt in Auckland? We never quite answered this question in Chapter 3. The Poisson distribution was used to count the

We have not said *how long* we have to wait for the next volcano: this is a

Auckland Volcanoes

About 50 volcanic eruptions have occurred in Auckland over the last 100,000 years or so. The first two eruptions occurred in the Auckland Domain and Albert Park — right underneath us! The most recent, and biggest, eruption was Rangitoto, about 600 years ago. There have been about 20 eruptions in the last 20,000 years, which has led the Auckland Council to assess current volcanic risk by assuming that volcanic eruptions in Auckland follow a Poisson process with rate $\lambda = \frac{1}{1000}$ volcanoes per year. For background information, see: `www.aucklandcouncil.govt.nz` and search for 'volcanic hazard'.

Distribution of the waiting time in the Poisson process

The length of time between events in the Poisson process is called the

To find the distribution of a continuous random variable, we often work with the

This is because $F_X(x) =$ gives us a *probability*, unlike the p.d.f. $f_X(x)$. We are comfortable with handling and manipulating probabilities.

Suppose that $\{N_t : t > 0\}$ forms a Poisson process with rate

N_t is the

We know that

Let X be a continuous random variable giving the

We will derive an expression for $F_X(x)$.

(i) When $x < 0$:

(ii) When $x \geq 0$:

Overall:

The distribution of the waiting time X is called the
because of the exponential formula for $F_X(x)$.

Example: What is the probability that there will be a volcanic eruption in Auckland within the next 50 years?

There is about a that there will be a volcanic eruption in Auckland over the next 50 years. This is the figure given by the Auckland Council at the above web link.

The Exponential Distribution

We have defined the Exponential(λ) distribution to be the distribution of

We write

However, just like the Poisson distribution, the Exponential distribution has many other applications: it does not always have to arise from a Poisson process.

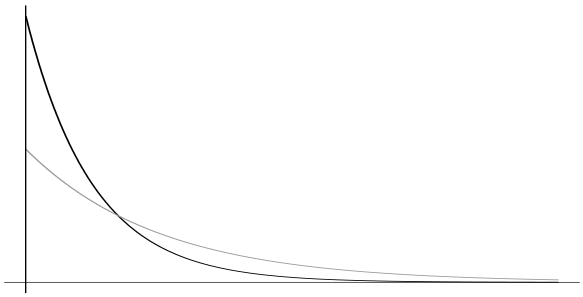
Let $X \sim$ Exponential(λ).

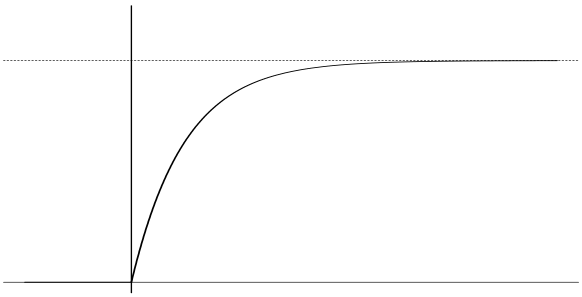
Distribution function:

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Probability density function:

$$f_X(x) = F'_X(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{for } x \geq 0, \\ 0 & \textit{for } x < 0. \end{cases}$$





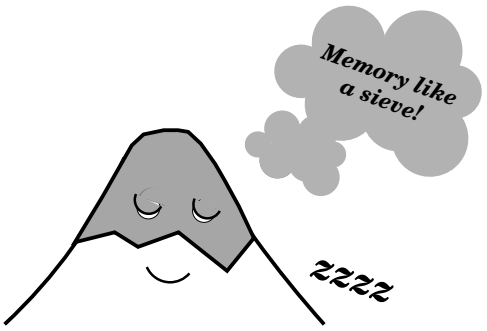
Link with the Poisson process

Let $\{N_t : t > 0\}$ be a Poisson process with rate λ. Then:

- N_t is the number of events to occur by time t;
- $N_t \sim \text{Poisson}(\lambda t)$; so $\mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$;
- Define X to be *either* the time till the first event, *or* the time from now until the next event, *or* the time between any two events.

Then
X is called the

Memorylessness



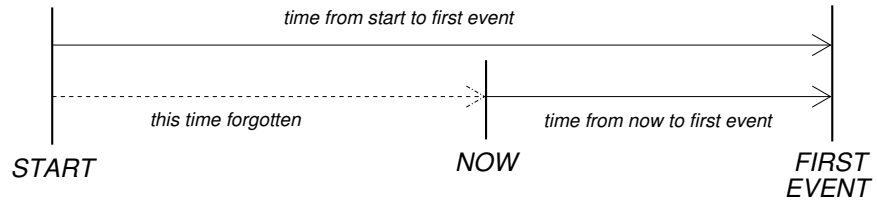
We have said that the waiting time of the Poisson process can be defined *either* as the time from the start to the first event, *or* the time from now until the next event, *or* the time between any two events.

All of these quantities have the ***same distribution:*** $X \sim$ ***Exponential***(λ).

The derivation of the Exponential distribution was valid for all of them, because events occur at a constant average rate in the Poisson process.

This property of the Exponential distribution is called ***memorylessness:***

- the distribution of the time from *now* until the first event is the same as the distribution of the time from *the start* until the first event: ***the time from the start till now has been forgotten!***



The Exponential distribution is famous for this memoryless property: it is the *only* continuous memoryless distribution.

For volcanoes, memorylessness means that ***the 600 years we have waited since Rangitoto erupted have counted for nothing.***

The chance that we still have 1000 years to wait for the next eruption is the same today as it was 600 years ago when Rangitoto erupted.

Memorylessness applies to *any* Poisson process. It is not always a desirable property: you don't want a memoryless waiting time for your bus!

The Exponential distribution is often used to model *failure times* of components: for example $X \sim$ Exponential(λ) is the amount of time before a light bulb fails. In this case, memorylessness means that 'old is as good as new' — or, put another way, 'new is as bad as old'! A memoryless light bulb is quite likely to fail almost immediately.

For private reading: proof of memorylessness

Let $X \sim \text{Exponential}(\lambda)$ be the *total* time waited for an event.

Let Y be the amount of *extra* time waited for the event, given that we have *already* waited time t (say).

We wish to prove that Y has the same distribution as X, i.e. that the time t already waited has been 'forgotten'. This means we need to prove that $Y \sim \text{Exponential}(\lambda)$.

Proof: We will work with $F_Y(y)$ and prove that it is equal to $1 - e^{-\lambda y}$. This proves that Y is Exponential(λ) like X.

First note that $X = t + Y$, because X is the *total* time waited, and Y is the time waited after time t. Also, we must condition on the event $\{X > t\}$, because we know that we have already waited time t. So $\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq t + y \,|\, X > t)$.

$$
\begin{aligned}
F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(X \leq t + y \,|\, X > t) \\
&= \frac{\mathbb{P}(X \leq t + y \quad \text{AND} \quad X > t)}{\mathbb{P}(X > t)} \\
&\qquad\qquad \text{(definition of conditional probability)} \\
&= \frac{\mathbb{P}(t < X \leq t + y)}{1 - \mathbb{P}(X \leq t)} \\
&= \frac{F_X(t + y) - F_X(t)}{1 - F_X(t)} \\
&= \frac{(1 - e^{-\lambda(t+y)}) - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})} \\
&= \frac{e^{-\lambda t} - e^{-\lambda(t+y)}}{e^{-\lambda t}} \\
&= \frac{e^{-\lambda t}(1 - e^{-\lambda y})}{e^{-\lambda t}} \\
&= 1 - e^{-\lambda y}. \qquad \text{So } Y \sim \text{Exponential}(\lambda) \text{ as required.}
\end{aligned}
$$

Thus the *conditional* probability of waiting time y *extra*, given that we have already waited time t, is the same as the probability of waiting time y in total. The time t already waited is forgotten. □

4.4 Likelihood and estimation for continuous random variables

- For discrete random variables, we found the likelihood using the
- For continuous random variables, we find the likelihood using the
- Although the notation $f_X(x)$ *means something different for continuous and discrete random variables, it is used in exactly the same way for likelihood and estimation.*

Note: Both discrete and continuous r.v.s have the same definition for the cumulative distribution function:

Example: Exponential likelihood

Suppose that:

-
-
-

Then the likelihood function is:

We estimate λ by

Two or more independent observations

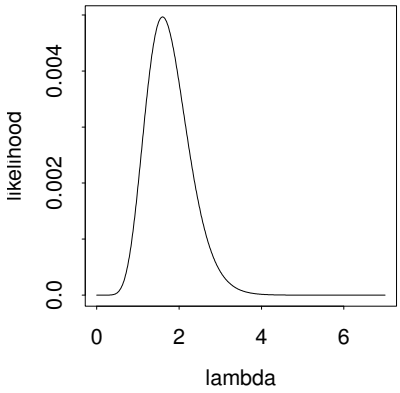
Suppose that $X_1, \ldots, X_n$ are continuous random variables such that:

-
-

then the likelihood is

Example: Suppose that $X_1, X_2, \ldots, X_n$ are independent, and $X_i \sim$ Exponential(λ) for all i. Find the maximum likelihood estimate of λ.

Likelihood graph shown for $\lambda = 2$ and $n = 10$. $x_1, \ldots, x_{10}$ generated by R command



4.5 Hypothesis tests

Hypothesis tests for continuous random variables are just like hypothesis tests for discrete random variables. The only difference is:

•

Example: discrete. Suppose $H_0 : X \sim \text{Binomial}(n = 10, p = 0.5)$, and we have observed the value $x = 7$. Then the *upper-tail* p-value is

Example: continuous. Suppose $H_0 : X \sim \text{Exponential}(2)$, and we have observed the value $x = 7$. Then the *upper-tail* p-value is

Other than this trap, the procedure for hypothesis testing is the same:

- Use H_0 to specify the distribution of X completely, and offer a one-tailed or two-tailed alternative hypothesis H_1.
- Make observation x.
- Find the one-tailed or two-tailed p-value as the probability of seeing an observation *at least as weird* as what we have seen, if H_0 is true.
- That is, find the probability under the distribution specified by H_0 of seeing an observation *further out in the tails* than the value x that we have seen.

Example with the Exponential distribution

A very very old person observes that the waiting time from Rangitoto to the next volcanic eruption in Auckland is 1500 years. Test the hypothesis that $\lambda = \frac{1}{1000}$ against the one-sided alternative that $\lambda < \frac{1}{1000}$.

Note: If $\lambda < \frac{1}{1000}$, we would expect to see
This is because X is the time between volcanoes, and λ is the rate at which volcanoes occur. A smaller value of λ means

Hypotheses:

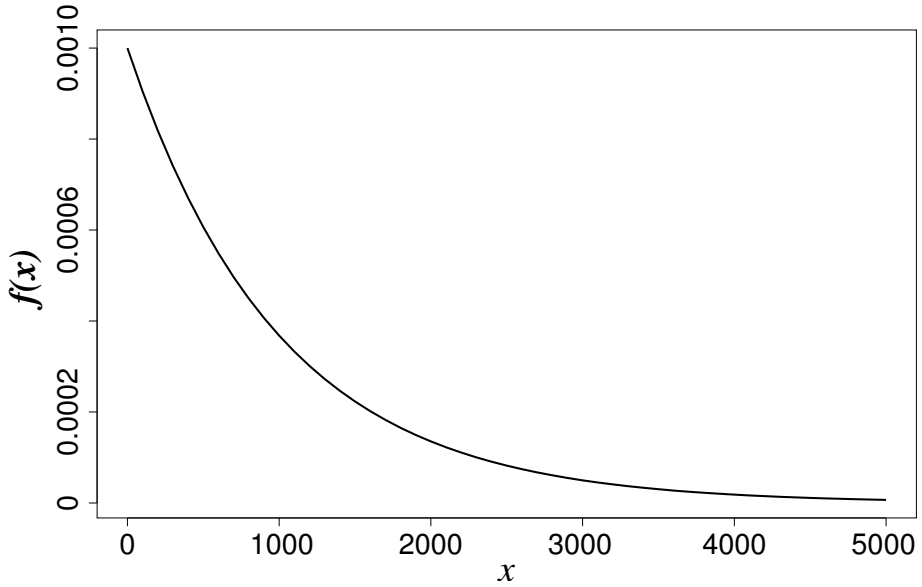
Observation:

Values weirder than $x = 1500$ years:

***p*-value:**

***R* command:**

Interpretation:



4.6 Expectation and variance

Remember the expectation of a ***discrete*** random variable is *the long-term average:*

$$\mu_X = \mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x x f_X(x).$$

(For each value x, we add in the value and multiply by the proportion of times we would expect to see that value: $\mathbb{P}(X = x)$.)

For a ***continuous*** random variable, *replace the probability function with the probability density function, and replace* $\sum_x$ *by* $\int_{-\infty}^{\infty}$*:*

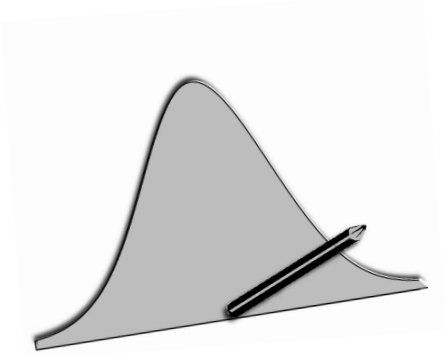
$$\mu_X = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

where $f_X(x) = F'_X(x)$ *is the probability density function.*

Note: There exists no concept of a 'probability function' $f_X(x) = \mathbb{P}(X = x)$ for continuous random variables. In fact, if X is continuous, then $\mathbb{P}(X = x) = 0$ *for all* x.

The idea behind expectation is the same for both discrete and continuous random variables. $\mathbb{E}(X)$ is:

- the long-term average of X;
- a 'sum' of values multiplied by how common they are:
  $\sum x f(x)$ *or* $\int x f(x)\, dx$.



Expectation is also the balance point of $f_X(x)$ for both continuous and discrete X.
Imagine $f_X(x)$ cut out of cardboard and balanced on a pencil.

Discrete:	**Continuous:**
$\mathbb{E}(X) = \sum_x x f_X(x)$	$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$
$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x)$	$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$
Transform the values, leave the probabilities alone;	Transform the values, leave the probability density alone.
$f_X(x) = \mathbb{P}(X = x)$	$f_X(x) = F_X'(x)$ **(p.d.f.)**

Variance

If X is continuous, its variance is defined in exactly the same way as a discrete random variable:

$$\textbf{Var}(X) = \sigma_X^2 = \mathbb{E}\left((X - \mu_X)^2 \right) = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

For a continuous random variable, we can either compute the variance using

$$\textbf{Var}(X) = \mathbb{E}\left((X - \mu_X)^2 \right) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx,$$

or

$$\textbf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - (\mathbb{E}X)^2.$$

The second expression is usually easier (although not always).

Properties of expectation and variance

All properties of expectation and variance are *exactly the same for continuous and discrete random variables.*

For *any* random variables, X, Y, and $X_1, \ldots, X_n$, continuous or discrete, and for constants a and b:

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
- $\mathbb{E}(ag(X) + b) = a\mathbb{E}(g(X)) + b$.
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
- $\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)$.
- $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.
- $\mathrm{Var}(ag(X) + b) = a^2\,\mathrm{Var}(g(X))$.

The following statements are generally true *only when X and Y are INDEPENDENT:*

- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ *when X, Y independent.*
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ *when X, Y independent.*

4.7 Exponential distribution mean and variance

When $X \sim \text{Exponential}(\lambda)$, then:

$$\boxed{\mathbb{E}(X) = \tfrac{1}{\lambda} \qquad \mathrm{Var}(X) = \tfrac{1}{\lambda^2}\,.}$$

Note: If X is the waiting time for a Poisson process with rate λ events per year (say), it makes sense that $\mathbb{E}(X) = \frac{1}{\lambda}$. For example, if $\lambda = 4$ events per hour, the average time waited between events is $\frac{1}{4}$ hour.

Proof: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx.$

Integration by parts: recall that $\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx.$

Let $u = x$, *so* $\frac{du}{dx} = 1$, *and let* $\frac{dv}{dx} = \lambda e^{-\lambda x}$, *so* $v = -e^{-\lambda x}$.

Then
$$\begin{aligned}
\mathbb{E}(X) = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx &= \int_0^{\infty} u\frac{dv}{dx}\,dx \\
&= \Big[uv\Big]_0^{\infty} - \int_0^{\infty} v\frac{du}{dx}\,dx \\
&= \Big[-xe^{-\lambda x}\Big]_0^{\infty} - \int_0^{\infty}(-e^{-\lambda x})\,dx \\
&= 0 + \Big[\tfrac{-1}{\lambda}e^{-\lambda x}\Big]_0^{\infty} \\
&= \tfrac{-1}{\lambda}\times 0 - \big(\tfrac{-1}{\lambda}\times e^0\big) \\
\therefore\quad \mathbb{E}(X) &= \tfrac{1}{\lambda}\,.
\end{aligned}$$

Variance: $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \frac{1}{\lambda^2}.$

Now $\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^{\infty} x^2\lambda e^{-\lambda x}\,dx.$

Let $u = x^2$, so $\frac{du}{dx} = 2x$, and let $\frac{dv}{dx} = \lambda e^{-\lambda x}$, so $v = -e^{-\lambda x}$.

Then
$$\begin{aligned}
\mathbb{E}(X^2) = \Big[uv\Big]_0^{\infty} - \int_0^{\infty} v\frac{du}{dx}\,dx &= \Big[-x^2e^{-\lambda x}\Big]_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x}\,dx \\
&= 0 + \frac{2}{\lambda}\int_0^{\infty}\lambda x e^{-\lambda x}\,dx \\
&= \frac{2}{\lambda}\times\mathbb{E}(X) = \frac{2}{\lambda^2}\,.
\end{aligned}$$

So
$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 &= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 \\
\mathrm{Var}(X) &= \frac{1}{\lambda^2}\,. \qquad \square
\end{aligned}$$

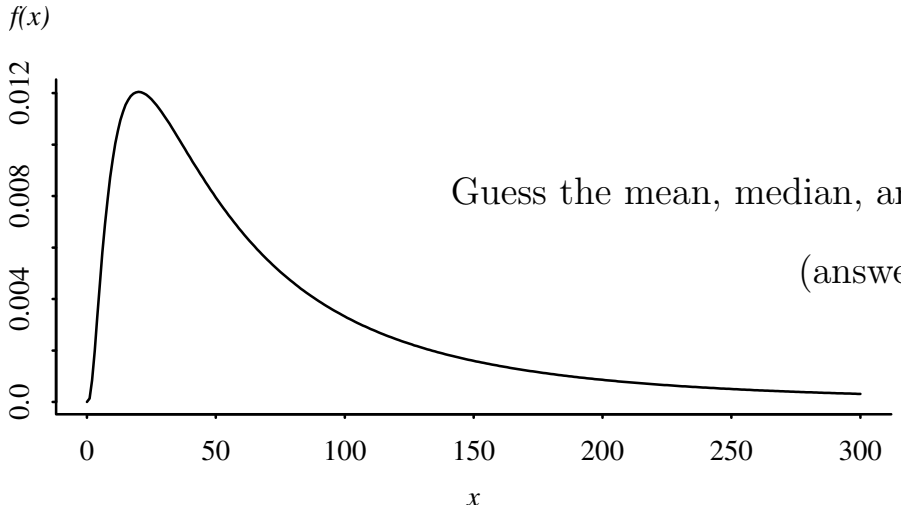
Interlude: Guess the Mean, Median, and Variance

For any distribution:

- the ***mean*** is the ***average*** that would be obtained if a large number of observations were drawn from the distribution;
- the ***median*** is the ***half-way point*** of the distribution: every observation has a 50-50 chance of being above the median or below the median;
- the ***variance*** is the ***average squared distance*** of an observation from the mean.

Given the probability density function of a distribution, we should be able to guess roughly the distribution mean, median, and variance . . . but it isn't easy! Have a go at the examples below. As a hint:

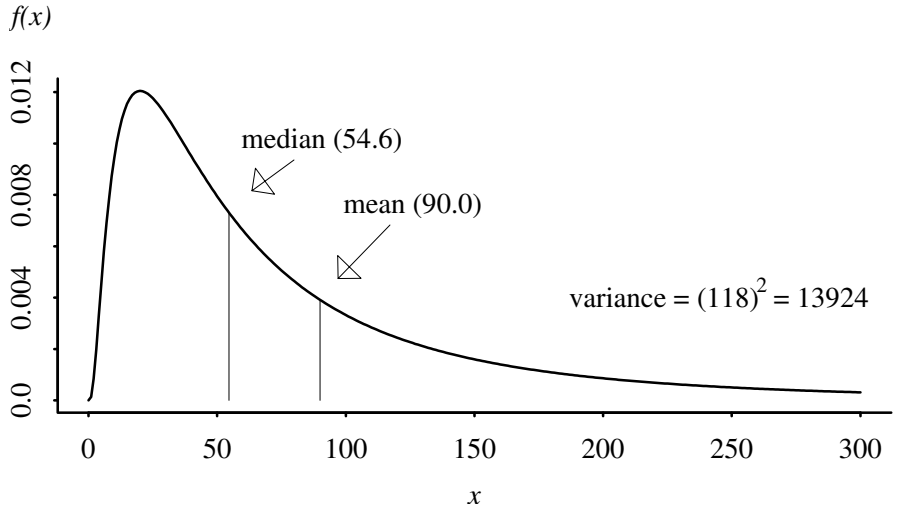
- the ***mean*** is the ***balance-point*** of the distribution. Imagine that the p.d.f. is made of cardboard and balanced on a rod. The mean is the point where the rod would have to be placed for the cardboard to balance.
- the ***median*** is the half-way point, so it divides the p.d.f. into two equal areas of 0.5 each.
- the ***variance*** is the average ***squared*** distance of observations from the mean; so to get a ***rough*** guess (not exact), it is easiest to guess an average distance from the mean and square it.



Guess the mean, median, and variance.

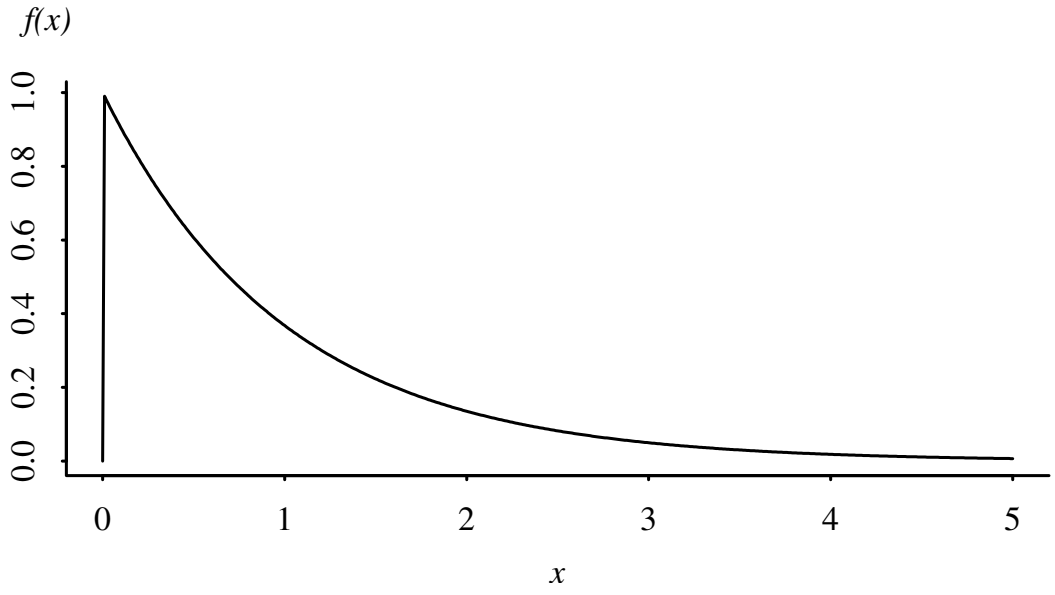
(answers overleaf)

Answers:



Notes: The mean is larger than the median. This always happens when the distribution has a long right tail (positive skew) like this one.
The variance is ***huge*** . . . but when you look at the numbers along the horizontal axis, it is quite believable that the average squared distance of an observation from the mean is 118^2. Out of interest, the distribution shown is a Lognormal distribution.

Example 2: Try the same again with the example below. Answers are written below the graph.



Answers: Median = 0.693; Mean=1.0; Variance=1.0.

4.8 The Uniform distribution

X has a **Uniform distribution on the interval $[a, b]$** if

We write

Probability density function, $f_X(x)$

If $X \sim U[a, b]$, *then*

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Distribution function, $F_X(x)$

Mean and variance:

Proof:

$$\begin{aligned}\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b x\left(\frac{1}{b-a}\right)\,dx &= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ &= \left(\frac{1}{b-a}\right)\cdot\frac{1}{2}(b^2-a^2) \\ &= \left(\frac{1}{b-a}\right)\frac{1}{2}(b-a)(b+a) \\ &= \frac{a+b}{2}\,.\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}[(X-\mu_X)^2] = \int_a^b \frac{(x-\mu_X)^2}{b-a}\,dx &= \frac{1}{b-a}\left[\frac{(x-\mu_X)^3}{3}\right]_a^b \\ &= \left(\frac{1}{b-a}\right)\left\{\frac{(b-\mu_X)^3-(a-\mu_X)^3}{3}\right\}\end{aligned}$$

But $\mu_X = \mathbb{E}X = \frac{a+b}{2}$, so $b-\mu_X = \frac{b-a}{2}$ and $a-\mu_X = \frac{a-b}{2}$.
So,

$$\begin{aligned}\mathrm{Var}(X) = \left(\frac{1}{b-a}\right)\left\{\frac{(b-a)^3-(a-b)^3}{2^3\times 3}\right\} &= \frac{(b-a)^3+(b-a)^3}{(b-a)\times 24} \\ &= \frac{(b-a)^2}{12}\,. \qquad \square\end{aligned}$$

Example: let $X \sim \mathrm{Uniform}[0,1]$. Then

4.9 The Change of Variable Technique: finding the distribution of $g(X)$

Let X be a continuous random variable. Suppose

-
-
-

We use the

Example: Let $X \sim \text{Uniform}(0, 1)$, and let

The p.d.f. of X is

What is the p.d.f. of Y,

Change of variable technique for monotone functions

Suppose that $g(X)$ is

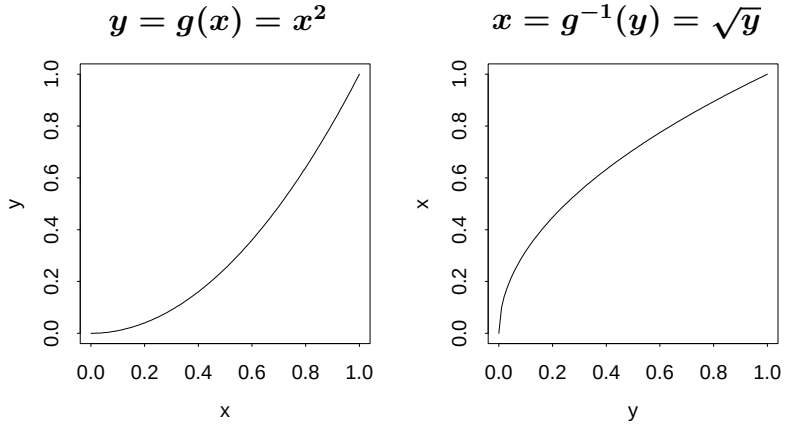
This means that

When g is monotone,

That is,

This means that the inverse function, $g^{-1}(y)$, is well-defined as a function for a certain range of y.

When $g : \mathbb{R} \to \mathbb{R}$, as it is here, then g can *only* be (1–1) if it is monotone.



Change of Variable formula

Let $g : \mathbb{R} \to \mathbb{R}$ be a monotone function and let $Y = g(X)$. Then *the p.d.f. of* $Y = g(X)$ *is*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| .$$

Easy way to remember

Working for change of variable questions

1) Show you have checked

2)

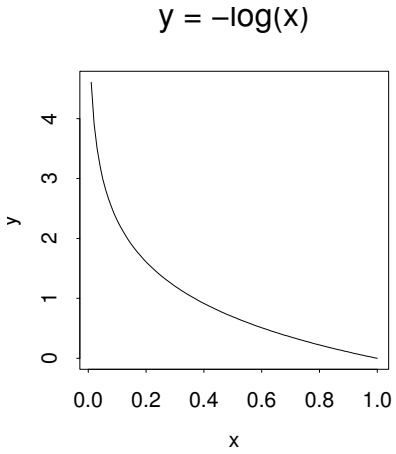
3)

4)

5)

Refer back to the question to find $f_X(x)$*: you often have to deduce this from information like*
Or it may be given explicitly.

Note: *There should be no x's left in the answer!*



Example 1: Let $X \sim \text{Uniform}(0, 1)$, and let $Y = -\log(X)$. Find the p.d.f. of Y. Hence **name** the distribution of Y, with parameters.

1)

2)

3)

4)

5)

Note: In change of variable questions, *you lose a mark for:*

1. *not stating $g(x)$ is monotone over the required range of x;*
2. *not giving the range of y for which the result holds, as part of the FINAL answer. (eg. $f_Y(y) = \ldots$ for $0 < y < \infty$).*

Example 2: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} \frac{1}{4}x^3 & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = 1/X$. Find the probability density function of Y, $f_Y(y)$.

For mathematicians: proof of the change of variable formula

Separate into cases where g is increasing and where g is decreasing.

i) g increasing

g is increasing if $u < w \Leftrightarrow g(u) < g(w)$. $\quad \circledast$
Note that putting $u = g^{-1}(x)$, and $w = g^{-1}(y)$, we obtain

$$\begin{aligned} g^{-1}(x) < g^{-1}(y) &\Leftrightarrow g(g^{-1}(x)) < g(g^{-1}(y)) \\ &\Leftrightarrow x < y, \end{aligned}$$

so g^{-1} is also an increasing function.

Now

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) &= \mathbb{P}(X \le g^{-1}(y)) \quad \text{put} \begin{cases} u = X, \\ w = g^{-1}(y) \end{cases} \text{ in } \circledast \text{ to see this.} \\ &= F_X(g^{-1}(y)). \end{aligned}$$

So the p.d.f. of Y is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= F_X'(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) \quad \text{(Chain Rule)} \\ &= f_X(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) \end{aligned}$$

Now g is increasing, so g^{-1} is also increasing (by overleaf), so $\frac{d}{dy}(g^{-1}(y)) > 0$, and thus $f_Y(y) = f_X(g^{-1}(y)) |\frac{d}{dy}(g^{-1}(y))|$ as required.

ii) g decreasing, i.e. $\quad u > w \iff g(u) < g(w). \quad (\star)$

(Putting $u = g^{-1}(x)$ and $w = g^{-1}(y)$ gives $g^{-1}(x) > g^{-1}(y) \iff x < y$,
so g^{-1} is also decreasing.)

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \le y) &= \mathbb{P}(g(X) \le y) \\ &= \mathbb{P}(X \ge g^{-1}(y)) \qquad (\text{put } u = X,\ w = g^{-1}(y) \text{ in } (\star)) \\ &= 1 - F_X(g^{-1}(y)). \end{aligned}$$

Thus the p.d.f. of Y is

$$f_Y(y) = \frac{d}{dy}\Big(1 - F_X(g^{-1}(y))\Big) = -f_X\Big(g^{-1}(y)\Big) \frac{d}{dy}\Big(g^{-1}(y)\Big).$$

This time, g is decreasing, so g^{-1} is also decreasing, and thus

$$-\frac{d}{dy}\Big(g^{-1}(y)\Big) = \left|\frac{d}{dy}\Big(g^{-1}(y)\Big)\right|.$$

So once again,

$$f_Y(y) = f_X\Big(g^{-1}(y)\Big)\left|\frac{d}{dy}\Big(g^{-1}(y)\Big)\right|. \qquad \square$$

4.10 Change of variable for non-monotone functions: non-examinable

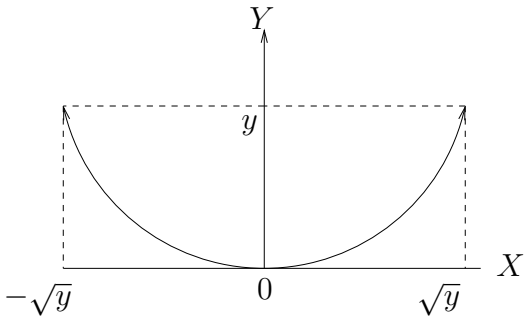
Suppose that $Y = g(X)$ and g is ***not*** monotone. We wish to find the p.d.f. of Y. We can sometimes do this *by using the distribution function directly.*

Example: Let X have ***any*** distribution, with distribution function $F_X(x)$. Let $Y = X^2$. Find the p.d.f. of Y.

Clearly, $Y \geq 0$, *so* $F_Y(y) = 0$ *if* $y < 0$.

For $y \geq 0$*:*

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y})\,. \end{aligned}$$



So

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0. \end{cases}$$

So the p.d.f. of Y is

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y &= \frac{d}{dy}(F_X(\sqrt{y})) - \frac{d}{dy}(F_X(-\sqrt{y})) \\ &= \tfrac{1}{2}y^{-\frac{1}{2}}F'_X(\sqrt{y}) + \tfrac{1}{2}y^{-\frac{1}{2}}F'_X(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}}\Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big) \quad \textit{for } y \geq 0. \end{aligned}$$

$$\therefore \qquad f_Y(y) = \frac{1}{2\sqrt{y}}\Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big) \textit{ for } y \geq 0, \textit{ whenever } Y = X^2.$$

Example: Let $X \sim \text{Normal}(0, 1)$. This is the familiar bell-shaped distribution (see later). The p.d.f. of X is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}.$$

Find the p.d.f. of $Y = X^2$.

By the result above, $Y = X^2$ has p.d.f.

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \cdot \frac{1}{\sqrt{2\pi}}(e^{-y/2} + e^{-y/2}) \\ &= \frac{1}{\sqrt{2\pi}}y^{-1/2}e^{-y/2} \quad \textit{for } \; y \geq 0. \end{aligned}$$

This is in fact the Chi-squared distribution with $\nu = 1$ degrees of freedom.

The Chi-squared distribution is a special case of the Gamma distribution (see next section). This example has shown that if $X \sim \text{Normal}(0, 1)$, then $Y = X^2 \sim$Chi-squared(df=1).

4.11 The Gamma distribution

The Gamma(k, λ) distribution is a very flexible family of distributions.

It is defined as the *sum of k independent Exponential r.v.s:*

if $X_1, \ldots, X_k \sim$ Exponential(λ) and $X_1, \ldots, X_k$ are independent, then $X_1 + X_2 + \ldots + X_k \sim$ Gamma(k, λ).

Special Case: When $k = 1$, *Gamma$(1, \lambda)$ = Exponential(λ)*
(the sum of a single Exponential r.v.)

Probability density function, $f_X(x)$

For $X \sim \text{Gamma}(k, \lambda)$,

$$f_X(x) = \begin{cases} \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\Gamma(k)$, called the Gamma function of k, is a constant that ensures $f_X(x)$ integrates to 1, i.e. $\int_0^\infty f_X(x)dx = 1$. It is defined as $\Gamma(k) = \int_0^\infty y^{k-1}e^{-y}\,dy$. When k is an integer, $\Gamma(k) = (k-1)!$

Mean and variance of the Gamma distribution:

For $X \sim \text{Gamma}(k, \lambda)$,

$$\mathbb{E}(X) = \tfrac{k}{\lambda} \text{ and } \text{Var}(X) = \tfrac{k}{\lambda^2}$$

Relationship with the Chi-squared distribution

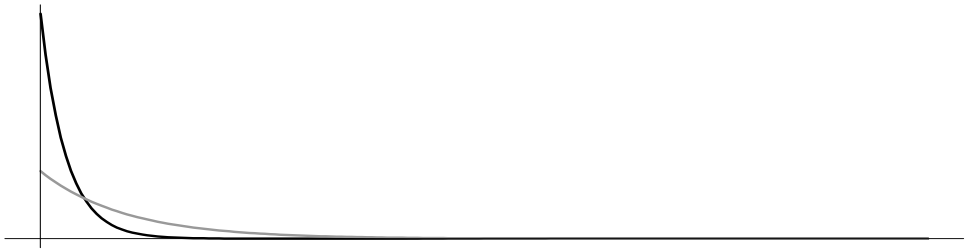
The Chi-squared distribution with ν degrees of freedom, χ^2_ν, is a special case of the Gamma distribution.

$$\chi^2_\nu = \text{Gamma}(k = \tfrac{\nu}{2}, \lambda = \tfrac{1}{2}).$$

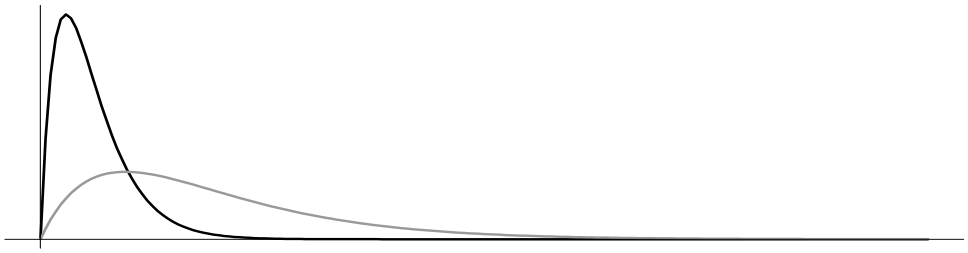
So if $Y \sim \chi^2_\nu$, then $\mathbb{E}(Y) = \frac{k}{\lambda} = \nu$, and $\text{Var}(Y) = \frac{k}{\lambda^2} = 2\nu$.

Gamma p.d.f.s

$k = 1$

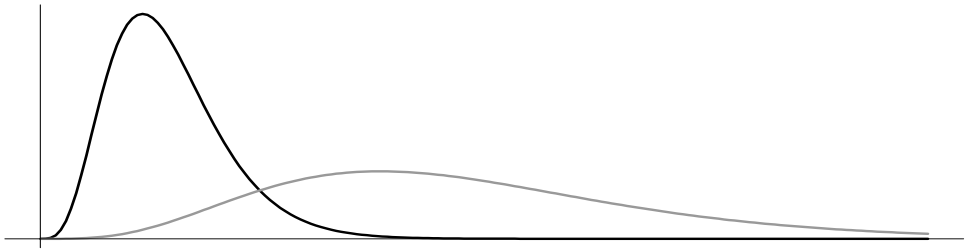


$k = 2$



Notice: right skew (long right tail); flexibility in shape controlled by the 2 parameters

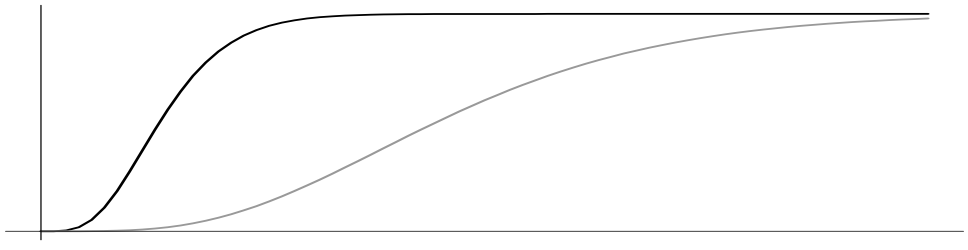
$k = 5$



Distribution function, $F_X(x)$

There is no closed form for the distribution function of the Gamma distribution. If $X \sim \text{Gamma}(k, \lambda)$, then $F_X(x)$ can can only be calculated *by computer.*

$k = 5$



Proof that $\mathbb{E}(X) = \frac{k}{\lambda}$ and $\mathrm{Var}(X) = \frac{k}{\lambda^2}$ (non-examinable)

$$\begin{aligned}
\mathbb{E}X = \int_0^\infty x f_X(x)\,dx &= \int_0^\infty x \cdot \frac{\lambda^k x^{k-1}}{\Gamma(k)} e^{-\lambda x}\,dx \\
&= \frac{\int_0^\infty (\lambda x)^k e^{-\lambda x}\,dx}{\Gamma(k)} \\
&= \frac{\int_0^\infty y^k e^{-y}(\frac{1}{\lambda})\,dy}{\Gamma(k)} \qquad (\text{letting } y = \lambda x,\ \tfrac{dx}{dy} = \tfrac{1}{\lambda}) \\
&= \frac{1}{\lambda} \cdot \frac{\Gamma(k+1)}{\Gamma(k)} \\
&= \frac{1}{\lambda} \cdot \frac{k\,\Gamma(k)}{\Gamma(k)} \qquad (\text{property of the Gamma function}), \\
&= \frac{k}{\lambda}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 &= \int_0^\infty x^2 f_X(x)\,dx - \frac{k^2}{\lambda^2} \\
&= \int_0^\infty \frac{x^2 \lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}\,dx - \frac{k^2}{\lambda^2} \\
&= \frac{\int_0^\infty (\frac{1}{\lambda})(\lambda x)^{k+1} e^{-\lambda x}\,dx}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{1}{\lambda^2} \cdot \frac{\int_0^\infty y^{k+1} e^{-y}\,dy}{\Gamma(k)} - \frac{k^2}{\lambda^2} \qquad \left[\text{where } y = \lambda x, \frac{dx}{dy} = \frac{1}{\lambda}\right] \\
&= \frac{1}{\lambda^2} \cdot \frac{\Gamma(k+2)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{1}{\lambda^2} \frac{(k+1)k\,\Gamma(k)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{k}{\lambda^2}. \qquad \square
\end{aligned}$$

Gamma distribution arising from the Poisson process

Recall that the waiting time between events in a Poisson process with rate λ has the *Exponential*(λ) *distribution.*

That is, if X_i =time waited between event $i-1$ and event i, then $X_i \sim$ *Exp*(λ).

The time waited from time 0 to the time of the kth event is

$$X_1 + X_2 + \ldots + X_k, \text{ the sum of } k \text{ independent Exponential}(\lambda) \text{ r.v.s.}$$

Thus the time waited until the kth event in a Poisson process with rate λ has the *Gamma*(k, λ) *distribution.*

Note: There are some similarities between the Exponential(λ) distribution and the (discrete) Geometric(p) distribution. Both distributions describe the 'waiting time' before an event. In the same way, the Gamma(k, λ) distribution is similar to the (discrete) Negative Binomial(k, p) distribution, as they both describe the 'waiting time' before the kth event.

4.12 The Beta Distribution: non-examinable

The Beta distribution has two parameters, α and β. We write $X \sim \text{Beta}(\alpha, \beta)$.

P.d.f.

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function $B(\alpha, \beta)$ is the *Beta function* and is defined by the integral

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx, \quad \text{for } \alpha > 0,\ \beta > 0.$$

It can be shown that $$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$
