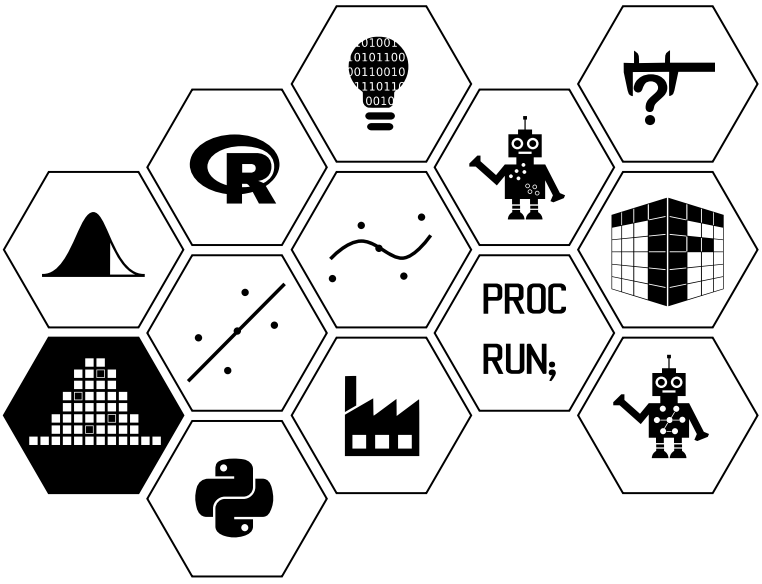
Probability and Sampling Fundamentals

Week 6: Continuous Random Variables II



DATA ANALYTICS
GLASGOW

Continuous random variables II

In Week 5 we learned about **continuous** random variables, which differ from **discrete** random variables in that they describe experiments which result in a real (rather than integer) value being recorded. For example, time taken to run a mile or measurement of height. This week's material introduces some of the most commonly used continuous random variables and will define properties of these distributions with a focus on real-world examples.

Week 6 learning material aims

The material in Week 6 covers:

- the normal distribution;
- defining the standard normal distribution;
- properties of the normal distribution;
- the uniform distribution;
- the exponential distribution;
- some other continuous distributions (in less detail).

The normal distribution

The first distribution we will discuss is the most important of all, and is used widely throughout statistics and probability. It is the **normal** or **Gaussian** distribution, and can be used to model a wide variety of data sets as well as being the basis for numerous statistical methods, many of which you will learn about in your future courses. The normal distribution can also be used to find approximations to probabilities associated with other distributions which you have already met, such as the binomial and Poisson, as we shall see in Week 8.

Note: this distribution is referred to interchangeably as either the normal or the Gaussian distribution after Carl Friedrich Gauss, one of the greatest mathematicians in history, who discovered its utility as a model for astronomical measurement errors. Statistician Karl Pearson coined the name 'normal distribution' in 1920, although later admitted that it 'had the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another *abnormal*'.

Example 1

Motivating example

Marathon time

The marathon times of 300 runners from a recent race were collected and are summarised in the table below in hours.

Time	Frequency	Relative frequency
2.0 - 2.5	1	0.003
2.6 - 3.0	16	0.053
3.1 - 3.5	50	0.167
3.6 - 4.0	76	0.253
4.1 - 4.5	99	0.330
4.6 - 5.0	41	0.137
5.1 - 5.5	13	0.043
5.6 - 6.0	4	0.013

These data should be modelled by a continuous random variable because the sample space includes any positive number, e.g. 3.1 hours, 3.12 hours, 3.127 hours, etc. These data are plotted in various histograms below, which illustrate the shape of the normal distribution.

Figure (i) is a histogram drawn for the data consisting of the finishing times for the 300 marathon runners. The bar heights represent raw counts, i.e. we are plotting absolute frequencies.

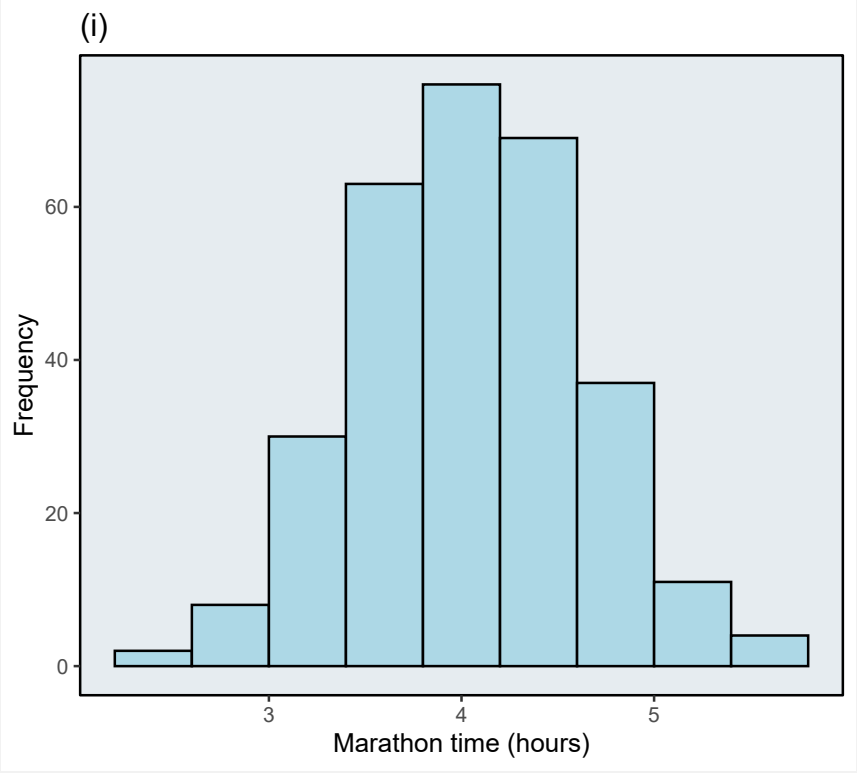


Figure 1

Figure (ii) is the same shape as (i) but has now been scaled so that the total area under the histogram is 1. This implies that the relative frequency is now given by the area of the histogram in that interval. The difference between the plots is the y-axis.

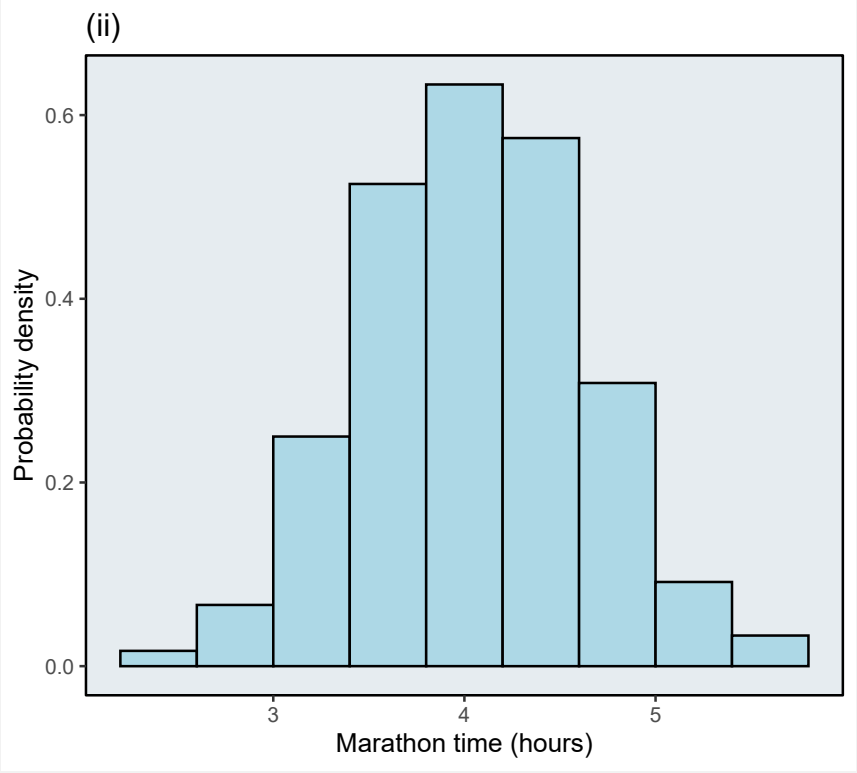


Figure 2

In Figure (iii) we now decrease the interval width. This produces the same basic shape as (ii) but smoother. If we take this process far enough, i.e. reducing interval width and collecting more data, then in the limit we get something very smooth as shown by the solid line.

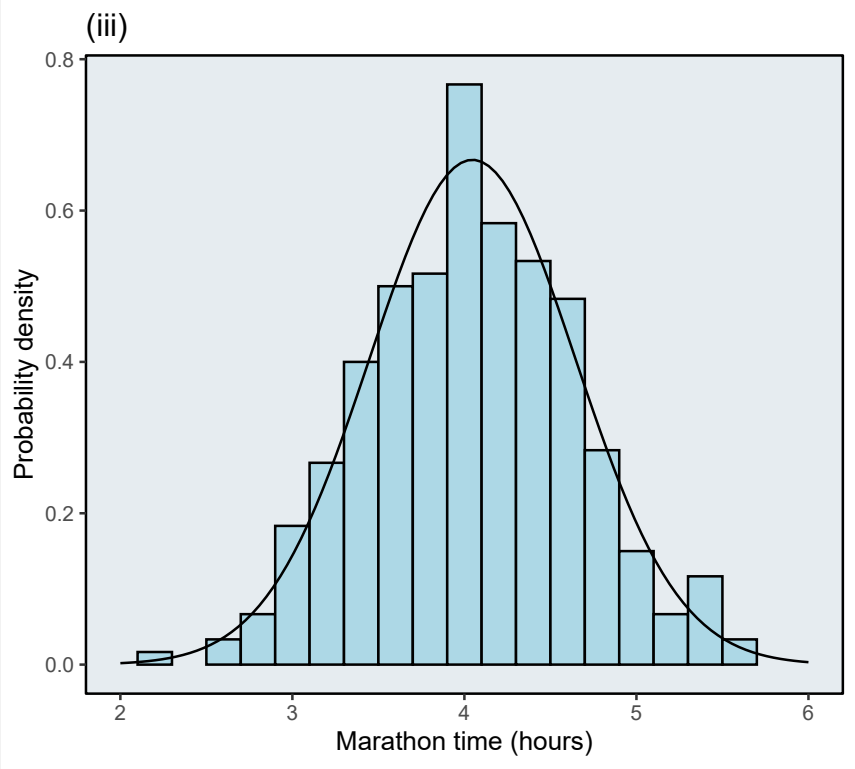


Figure 3

Definition 1

Probability density function of the normal distribution

Suppose that the random variable X can take any real value and that X has p.d.f.

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

for all $x \in \mathbb{R}$, then X is said to have a **normal** distribution, with parameters μ and σ^2, written

$$X \sim \mathrm{N}(\mu, \sigma^2).$$

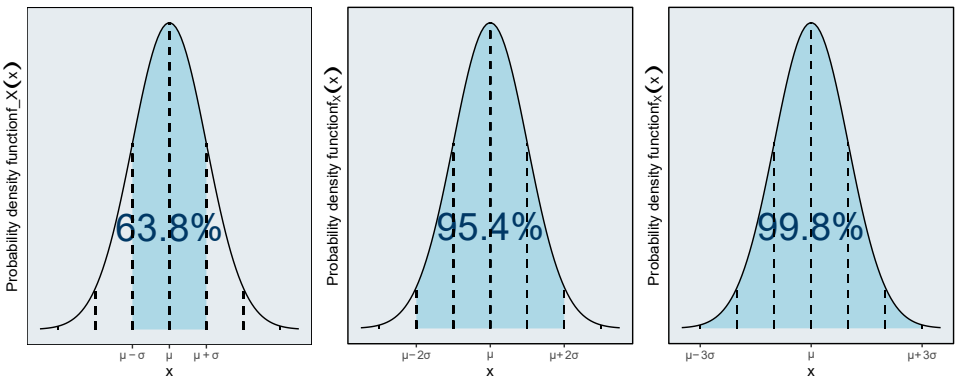


Figure 4

Should this not be 68%

The normal distribution is symmetrical and `bell-shaped' and is completely defined if we know its mean μ and variance σ^2. The plot above shows that the curve is symmetrical about the mean μ. The vertical lines on this graph show the mean μ along with values 1, 2, and 3 standard deviations above and below the mean. The normal distribution has the property that 63.8% of distribution lies within 1 standard deviation of the mean, 95.4% of distribution lies within 2 standard deviations of the mean and 99.8% of distribution lies within 3 standard deviations of the mean.

It can be shown that

$$\mathrm{E}(X) = \mu \quad \text{and} \quad \mathrm{Var}(X) = \sigma^2.$$

The standard deviation is thus equal to σ.

The normal distribution is used to model a number of real world phenomena including:

- **Population characteristics** - such as height, weight, etc of a given population.
- **Scientific experiments** - measurements made from experiments that are subject to error.
- **Statistical methods** - a number of statistical methods assume that the data has a normal distribution.

The four figures below illustrate how the location and spread change depending on the values for the mean (μ) and variance (σ^2).

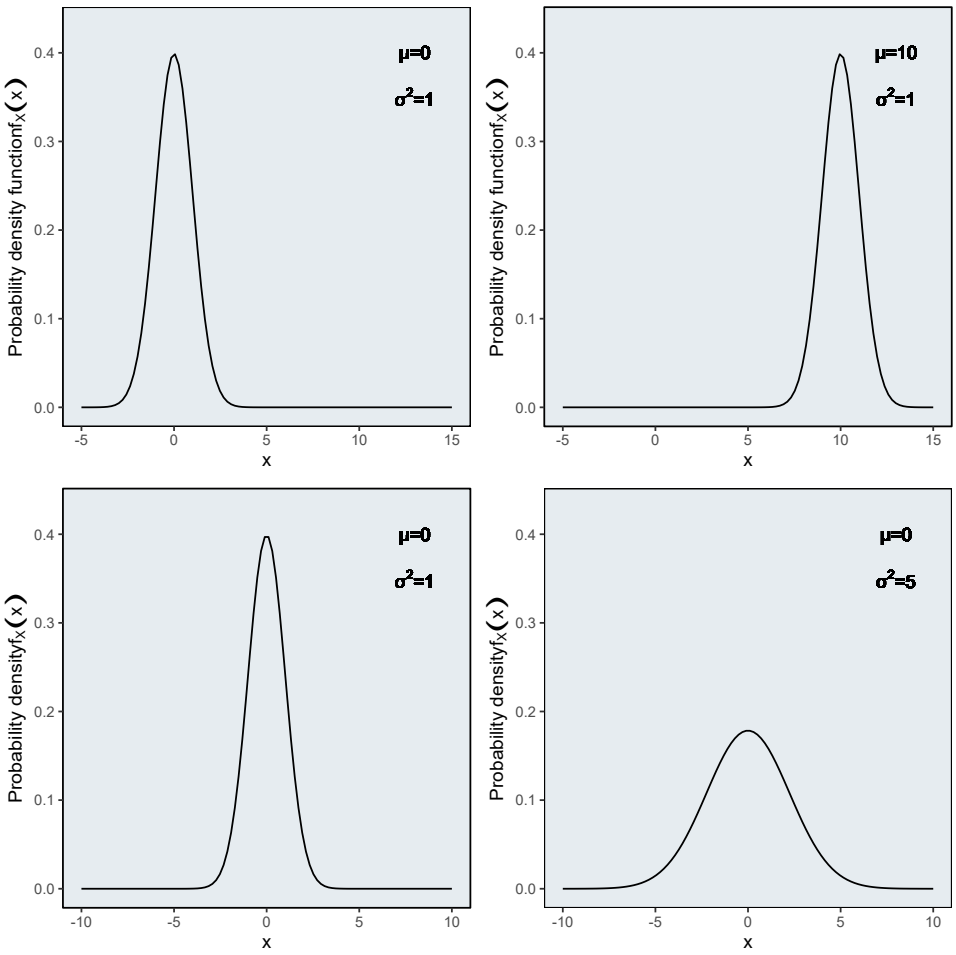


Figure 5

Most notably, the curve is always centered around the mean and the larger the variance, the flatter the probability density function is.

This video gives an overview of the normal distribution, including illustrating some of it's key properties and how it can be used to model real data.

Video

The normal distribution

Duration 3:54

Properties of the normal distribution

- If $X \sim \mathrm{N}(\mu, \sigma^2)$ and a, b are constants, then the linear transformation $aX + b$ also has a normal distribution:

$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2).$$

- If $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ and X_1, X_2 are independent, then their sum and their difference also has a normal distribution:

$$X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$
$$X_1 - X_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

The result generalises to any number of independent random variables. The next step is to determine how to calculate probabilities from the normal distribution.

Calculating probabilities

We want to be able to calculate probabilities of the form $\mathrm{P}(X \leq x)$ like we did in Week 5. To calculate these probabilities using what we learned in Week 5 we need to evaluate integrals of the form

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

However, this integral cannot be evaluated analytically. Instead, this integral needs to be evaluated numerically. In practice this means that we would either need to use a computer to calculate these probabilities or obtain them from a table.

Fortunately, we do not need to create separate tables for each combination of the parameters μ and σ^2. Instead we can relate probabilities of the form $\mathrm{P}(X \leq x)$ for the $\mathrm{N}(\mu, \sigma^2)$ distribution to probabilities for the $N(0, 1)$ distribution, which we will call the **standard normal distribution**. This means that we can calculate probabilities of the form $\mathrm{P}(X \leq x)$ in two stages: in the first stage we relate the probability of interest to the standard normal distribution, a process we will call **standardisation**. We can then in the second step obtain the required probability from a table.

The standard normal distribution is a special case of the normal distribution which has a mean of 0 and a standard deviation (and hence a variance) of 1.

Definition 2

p.d.f. of standard normal distribution

A random variable Z is said to have a **standard normal distribution** if $Z \sim \mathrm{N}(0,1)$, i.e. if it has the p.d.f.

$$\phi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Note: by convention Z is used to denote a standard normal distribution and X is used to denote any other normal distribution. The lower-case Greek letter phi (ϕ) is often used to denote the p.d.f. of the normal distribution.

We can now use this distribution to allow us to calculate probabilities of the form $\mathrm{P}(X \leq x)$ using the following two stages.

1. Transform $\mathrm{P}(X \leq x)$ into a probability of the form $\mathrm{P}(Z \leq z)$, where Z is a standard normal random variable with mean 0 and variance 1.
2. Use statistical tables of the c.d.f. of the standard normal distribution to calculate the probability (or calculate them in `R`).

The c.d.f. of the standard normal distribution is so important in practice that it is given a special symbol, the upper-case Greek letter Phi (Φ). It is defined and plotted below.

If $Z \sim \mathrm{N}(0,1)$, then the c.d.f. of Z is given by

$$\Phi(z) = \mathrm{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt.$$

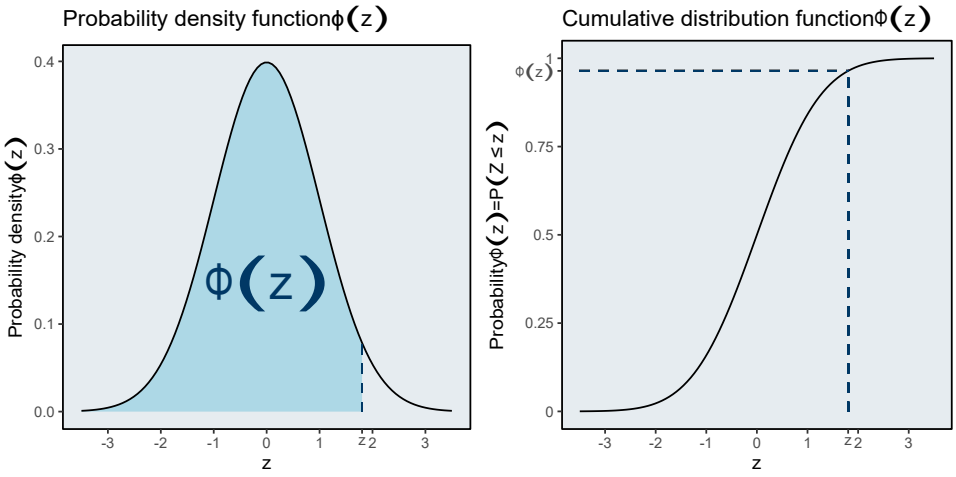


Figure 6

Each stage of the two-step process to calculate probabilities of the form $\mathrm{P}(X \leq x)$ for the normal distribution is now described in further detail.

Step 1 - standardisation

Standardisation is the process by which a probability of the form $\mathrm{P}(X \leq x)$ is transformed into one of the form $\mathrm{P}(Z \leq z)$, where Z is a standard normal random variable with mean 0 and variance 1. Let $X \sim \mathrm{N}(\mu, \sigma^2)$, then the transformation

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1),$$

is a standard normal random variable. Note that we divide by the **standard deviation and not the variance**. If we assume that the transformation Z is a normal random variable, then the mean and variance are easy to obtain.

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X - \mu)}{\sigma} = \frac{E(X) - \mu}{\sigma} = 0.$$

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{\mathrm{Var}(X - \mu)}{\sigma^2} = \frac{\mathrm{Var}(X)}{\sigma^2} = 1.$$

Therefore any problem involving a probability of the form $\mathrm{P}(X \leq x)$ can be transformed into one of the form $\mathrm{P}(Z \leq z)$ by applying the above transformation to x, as shown below.

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Example 2

Standardisation

Let $X \sim \mathrm{N}(1, 4)$, so that $\mu = 1$ and $\sigma = 2$. Transform $\mathrm{P}(X \leq 5)$ into a probability involving Z.

Answer:

$$\mathrm{P}(X \leq 5) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq \frac{5-\mu}{\sigma}\right) = \mathrm{P}\left(Z \leq \frac{5-1}{2}\right) = \mathrm{P}(Z \leq 2) = \Phi(2).$$

Step 2 - calculate probability using standard normal

As previously described the c.d.f. of a standard normal distribution (denoted by $\Phi(z)$) is given by

$$\Phi(z) = \mathrm{P}(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt,$$

and cannot be obtained analytically as the integral does not exist. Instead of doing the integration, we obtain probabilities of the form $\mathrm{P}(Z \leq z)$ either in `R` using `pnorm()` or by using the statistical tables for the standard normal distribution:

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9999		

Figure 7

The rows provide the units and the first decimal place of z and the columns the second decimal place. So to find $\Phi(1.96)$, we need to find the row corresponding to '1.9' and then the column corresponding to the digit '6', giving 0.975, i.e. $\Phi(1.96) = 0.975$ (highlighted in purple).

You might have noticed that the table above only covers positive z, and only shows probabilities of at least 0.5. The reason for this that we can exploit the symmetry of the standard normal distribution to obtain $\Phi(z)$ for negative z.

The probability density function $\phi(z)$ of the standard normal distribution is symmetric about 0, i.e.

$$\phi(-z) = \phi(z).$$

One can show that this implies for the cumulative distribution function that

$$\Phi(-z) = 1 - \Phi(z).$$

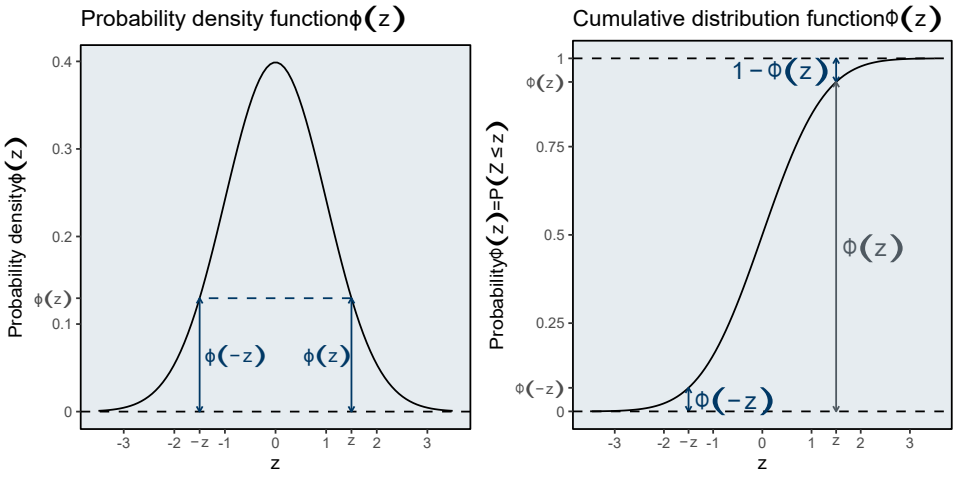


Figure 8

Using the tables

Example 3

Let $Z \sim \mathrm{N}(0,1)$, calculate the following.

1. $\mathrm{P}(Z < 1.52) = \Phi(1.52)$.

To find the value of $\Phi(z)$ for $z = 1.52$, go to the tables and find the value '1.5' in the left most column. Then follow that row across until you get to '2' and the probability you want is in that row and column. This gives 0.9357 (highlighted in blue).

Or in `R`

```
pnorm(1.52)
```

```
R Console
[1] 0.9357445
```

2. $\mathrm{P}(Z > 1.52)$.

This can be re-written as

$$\mathrm{P}(Z > 1.52) = 1 - \mathrm{P}(Z < 1.52) = 1 - \Phi(1.52) = 1 - 0.9357 = 0.0643.$$

Or in `R`

```
1-pnorm(1.52)
```

```
R Console
[1] 0.06425549
```

3. $\mathrm{P}(Z < -1.52)$.

By symmetry this is equal to

$$\mathrm{P}(Z < -1.52) = \mathrm{P}(Z > 1.52) = 0.0643.$$

Or in `R`

```
pnorm(-1.52)
```

```
R Console
[1] 0.06425549
```

4. Find the value of c such that $\mathrm{P}(Z \leq c) = 0.975$.

To find c look for 0.975 in the middle of the table and c is the number in the left most column and top most row. In this case $c = 1.96$. Note: $\Phi^{-1}(0.975)$ has the meaning of the value of the standard normal random variable that has probability 0.975 'to the left'. So $\Phi^{-1}(0.975) = c$ is equivalent to writing $\mathrm{P}(Z \leq c) = 0.975$. In other words the 0.975-quantile (or equivalently, the 97.5% percentile) of the standard normal distribution is 1.96.

Task 1

Find:

1. $\mathrm{P}(Z < -1)$
2. $\mathrm{P}(-2 < Z < -1)$
3. $\mathrm{P}(-1.65 < Z < 1.65)$

Example 4

Motivating example continued

For the marathon runners in Example .marathon we know that their running time is approximately normally distributed with mean 4 hours and standard deviation 0.6 hours.

1. What is the distribution for the finishing time for a marathon runner?
2. Calculate the following:

(a) the probability that a runner finishes in less than 4.5 hours.

(b) the probability that a runner finishes in more than 3 hours.

(c) the probability that a runner finishes between 3.5 and 4.5 hours.

(d) the time a runner must finish in order to finish within the top 70% of all runners.

Answer:

1. Let $X = \{\text{finishing time for marathon runner}\}$. We are told in the question the mean finishing time, $\mu = 4$ and the standard deviation, $\sigma = 0.6$. Therefore:

$$X \sim \mathrm{N}(4, 0.6^2) = \mathrm{N}(4, 0.36).$$

   Note: we need to square the standard deviation since a normal random variable X with mean μ and variance σ^2 is denoted by $X \sim \mathrm{N}(\mu, \sigma^2)$. Below is a plot of the distribution.

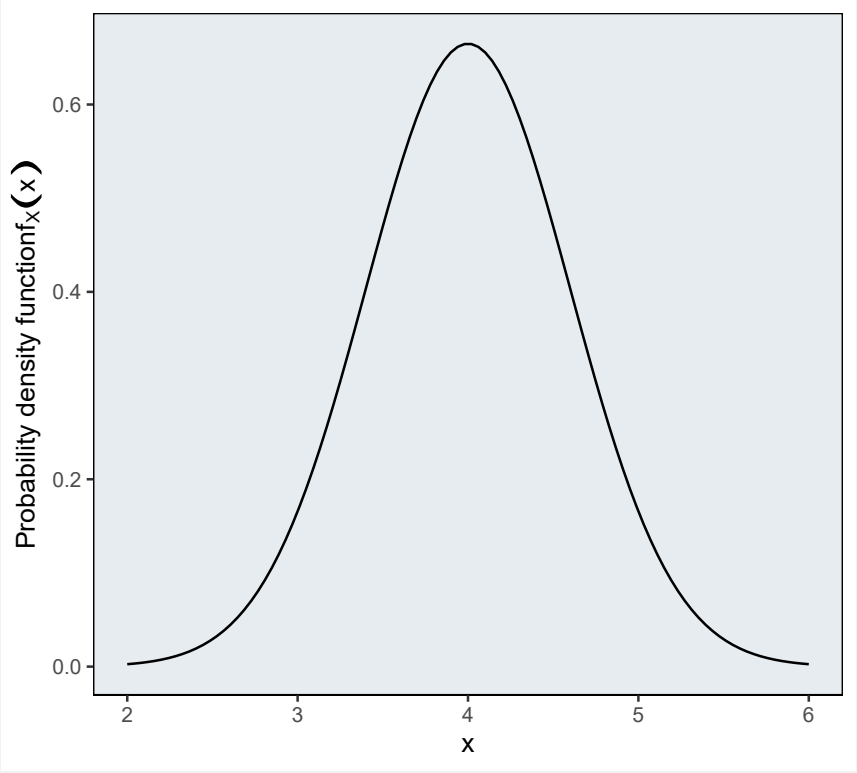


Figure 9

2.

(a) The probability we want to calculate is the shaded area in the plot below.

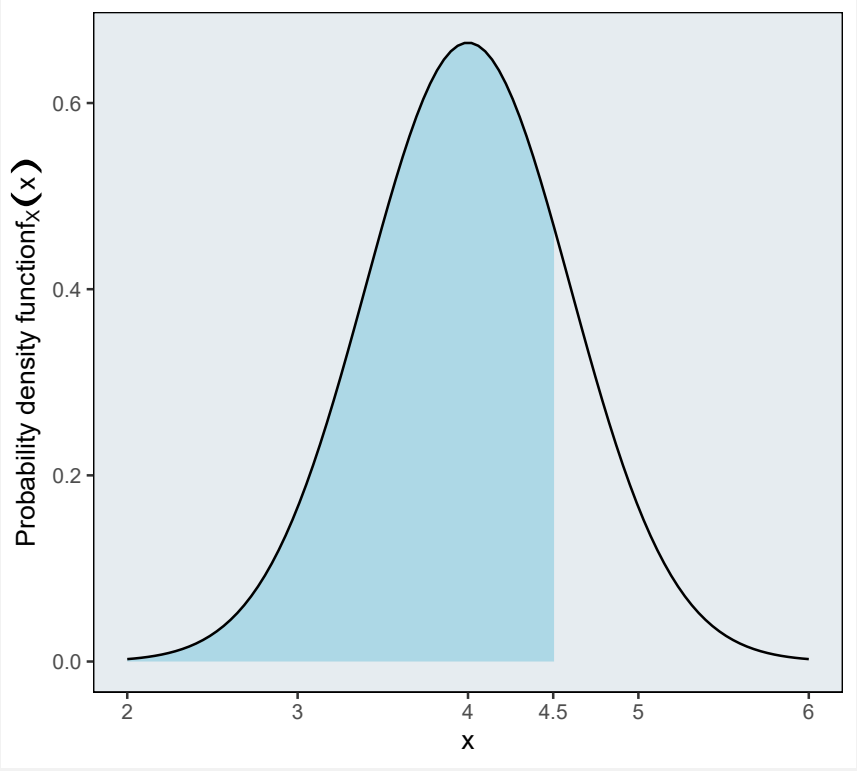


Figure 10

$$
\begin{aligned}
\mathrm{P}(X < 4.5) &= \mathrm{P}\left(\frac{X-4}{0.6} < \frac{4.5-4}{0.6}\right) \\
&= \mathrm{P}(Z < 0.833) \quad \text{where } Z \text{ has the standard normal distribution} \\
&= \Phi(0.833) = 0.7967 \quad \text{from the standard normal table.}
\end{aligned}
$$

So the probability that a runner finishes the marathon in less than 4.5 hours is 0.80.

We could have also turned to R and used

```
pnorm(4.5, mean=4, sd=0.6)
```

R Console

```
[1] 0.7976716
```

or

```
pnorm((4.5-4)/0.6)                    # Standardising manually
```

R Console

```
[1] 0.7976716
```

Warning: when you use the functions for the normal distribution , you need to specify not σ^2 but σ in the argument, a recurring source of mistakes.

(b) The probability we want to calculate is the shaded area in the plot below.

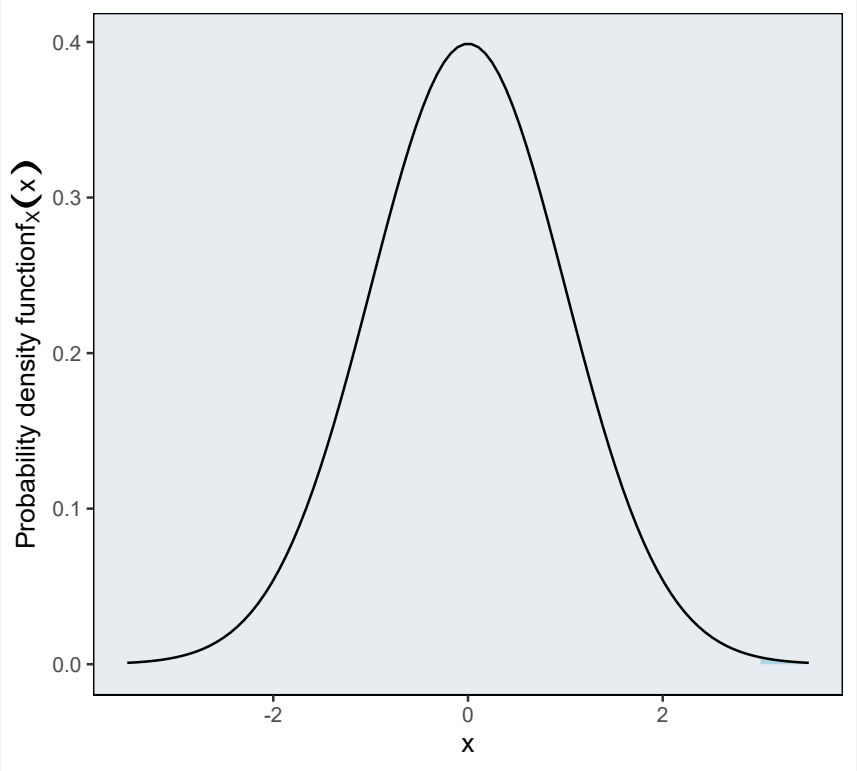


Figure 11

$$
\begin{aligned}
\mathrm{P}(X > 3) &= 1 - \mathrm{P}(X < 3) \\
&= 1 - \mathrm{P}\left(\frac{X-4}{0.6} < \frac{3-4}{0.6}\right) \\
&= 1 - \mathrm{P}(Z < -1.67) \\
&= 1 - \{1 - \mathrm{P}(Z < 1.67)\} \quad \text{where } Z \text{ has the standard normal distribution} \\
&= 1 - \{1 - \Phi(1.67)\} \\
&= 1 - \{1 - 0.9525\} \\
&= 0.9525 \quad \text{from the standard normal table.}
\end{aligned}
$$

So the probability that a runner finishes the marathon in more than 3 hours is 0.95.

In R, we can use

```
1 - pnorm(3, mean=4, sd=0.6)
```

R Console

```
[1] 0.9522096
```

or

```
pnorm(3, mean=4, sd=0.6, lower.tail=FALSE)
```

R Console

```
[1] 0.9522096
```

For the latter command we have exploited that setting `lower.tail=FALSE` calculates probabilities of the form $P(X > x)$.

(c) The probability we want to calculate is the shaded area in the plot below.

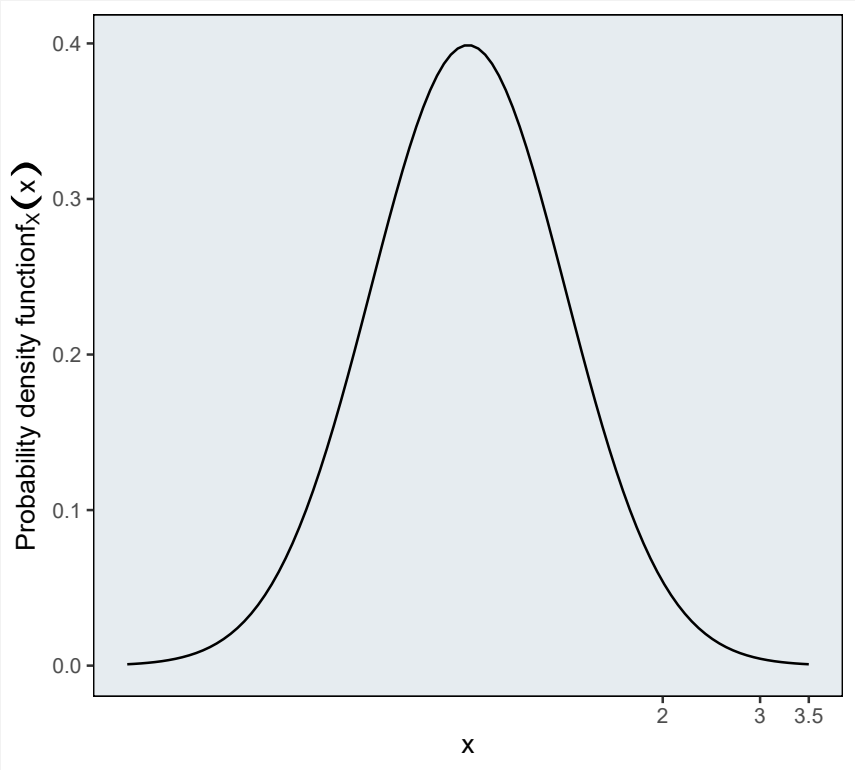


Figure 12

$$\begin{aligned} \mathrm{P}(3.5 < X < 4.5) &= \mathrm{P}\left(\frac{3.5-4}{0.6} < \frac{X-4}{0.6} < \frac{4.5-4}{0.6}\right) \\ &= \mathrm{P}(0.83 < Z < 0.83) \\ &= \Phi(0.83) - \Phi(-0.83) \\ &= 0.7967 - \{1 - 0.7967\} = 0.5934. \end{aligned}$$

So the probability that a runner finishes the marathon between 3.5 and 4.5 hours is 0.59.

In R, we can use

```
pnorm(4.5, mean=4, sd=0.6) - pnorm(3.5, mean=4, sd=0.6)
```

R Console

```
[1] 0.5953432
```

(d)

$$0.7 = \mathrm{P}(X < c) = \mathrm{P}\left(\frac{X-4}{0.6} < \frac{c-4}{0.6}\right) = \mathrm{P}\left(Z < \frac{c-4}{0.6}\right) = \Phi((c-4)/0.6).$$

From the table on page 14, $\Phi(0.52) = 0.7$, so that

$$\frac{c-4}{0.6} = 0.52 \quad \text{and} \quad c = 4.312.$$

So in order to finish within the top 70% of the other runners, the runner must finish in around 4.31 hours or less.

In `R` we can use the function `qnorm` to evaluate the inverse c.d.f. and calculate quantiles. This gives a more accurate answer.

```
qnorm(0.7, mean=4, sd=0.6)
```

```
R Console
[1] 4.31464
```

Note: The solutions may be slightly different when using `pnorm` (or `qnorm`) to calculate these probabilities using `R`. This is because the statistical tables are only tabulated up to a specified number of decimal points and so often we have to round to the nearest tabulated number, whereas `R` can compute these probabilities numerically.

This video explains how to calculate probabilities of the form $\mathrm{P}(X \leq x)$ for a normal distribution using the standard normal distribution.

Video

The standard normal distribution

Duration 5:03

Task 2

Let $X \sim \mathrm{N}(7, 16)$. Calculate the following:

1. $\mathrm{P}(X < 9)$.
2. $\mathrm{P}(X > 9)$.
3. $\mathrm{P}(4 < X < 9)$.
4. Find c such that $\mathrm{P}(X < c)$=0.95.

The uniform distribution

The uniform distribution is one of the simplest continuous distributions, and gives each point in the specified interval equal probability density.

Example 5

Motivating example

Let's look back at Example 2 from Week 5.

Suppose a very reliable train line runs exactly every 20 mins between 6am and 6pm. A person walks into the train station at a random time between 6am and 6pm with no idea when the last train left the station or when the next train is due. How long can the person expect to wait on their train?

The person could be very lucky with the train arriving immediately after they arrive such that their waiting time is 0 mins or the person could be very unlucky and just miss a train with a waiting time of 20 mins. The problem is that they have no way of knowing.

We could say that their waiting time could lie between 0 and 20 mins with every value equally likely. The figure below illustrates this information

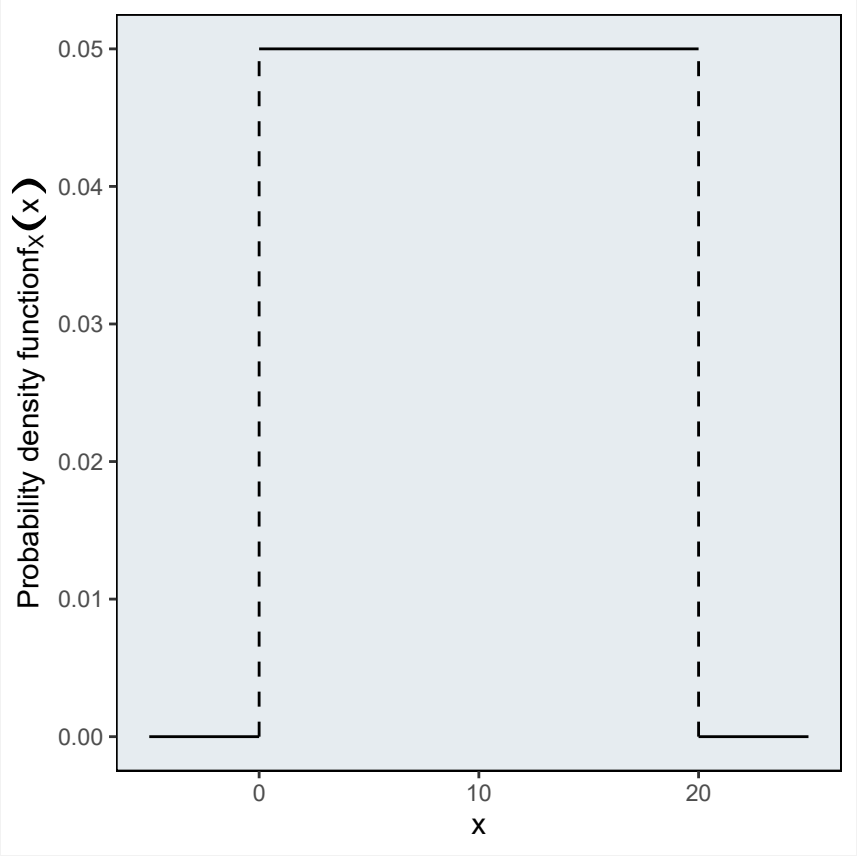


Figure 13

Let $X = \{\text{time spent waiting on train}\}$, then X can be modelled using the **uniform** distribution.

A **uniform** random variable gives each outcome in the sample space the same probability density.

Definition 3

Uniform distribution

The continuous random variable X is said to have the **uniform** distribution on the interval $[a, b]$, written $X \sim \mathrm{U}(a, b)$, if X has the following p.d.f. and c.d.f.:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise,} \end{cases} \qquad F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & \text{otherwise.} \end{cases}$$

Note that the `R` functions `dunif()` and `punif()` compute values of the p.d.f. and the c.d.f. of the uniform distribution.

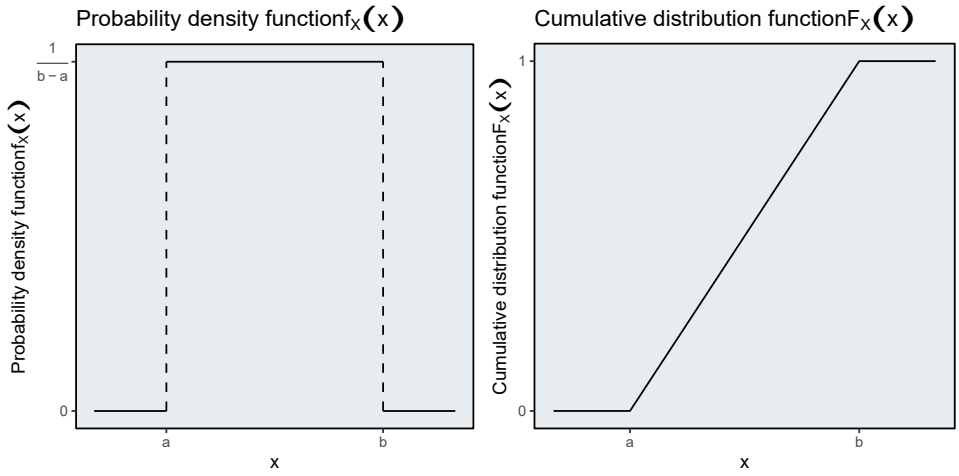


Figure 14

Supplement 1

Derivation of the cumulative distribution function

Suppose that $x \geq a$ and $x \leq b$, then

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt = \int_{a}^{x} \frac{1}{b-a}\, dt$$
$$= \left[\frac{t}{b-a}\right]_{t=a}^{x} = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a}$$

Thus the c.d.f. of the uniform distribution is given by

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & \text{otherwise} \end{cases}.$$

Example 6

Motivating example continued

In Example 5 we are told that the train runs every 20 minutes and so a random person who walks into the train station could wait from 0 minutes to 20 minutes for their train. So, X, the length of time spent waiting on the train, follows a uniform distribution with $a = 0$ and $b = 20$:

$$X \sim \mathrm{U}(0, 20).$$

It follows that the p.d.f. is equal to

$$f_X(x) = \begin{cases} \frac{1}{20}, & 0 \leq x \leq 20 \\ 0, & \text{otherwise} \end{cases}$$

We can then calculate probabilities that we are interested in using the c.d.f. as provided above or in the usual way for a continuous random variable. For example, what is the probability that a person will wait less that 5 minutes? Both methods are shown below, however you will notice that one requires far less work.

$$\begin{aligned} \mathrm{P}(X < 5) &= \int_{-\infty}^{x} f_x(x) dx \\ &= \int_{0}^{5} \frac{1}{20} dx \\ &= \left[\frac{x}{20} \right]_{x=0}^{5} \\ &= \frac{1}{4}. \end{aligned}$$

Or using the fact that

$$F_X(x) = \begin{cases} \frac{x-a}{b-a}, & a < x \leq b \\ 1, & \text{otherwise} \end{cases}$$

so $\mathrm{P}(X < 5) = F_X(5) = \frac{5-0}{20-0} = \frac{1}{4}$.

The probability that we have calculated corresponds to the area shaded in blue in the figure below.

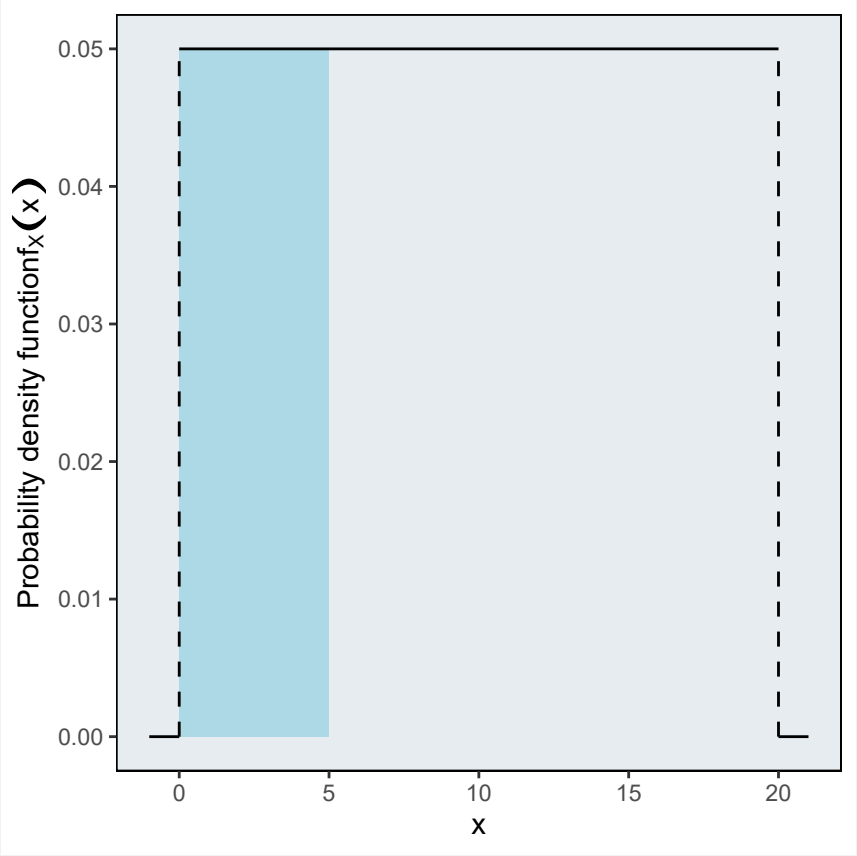


Figure 15

Task 3

The world's most famous geyser, Old Faithful in Yellowstone national park, erupts every 91 minutes. You arrive there at random and wait for 20 minutes.

1. What is the distribution for the time until Old Faithful erupts?
2. What is the p.d.f. for the time until Old Faithful erupts?
3. What is the probability that you will see it erupt?

Example 7

Motivating example continued

In week 5, we calculated the expectation and variance for Example 5. Here is a reminder:

$$\begin{aligned}\mathrm{E}(X) &= \int_{x\in R_X} x f_X(x)dx \\ &= \frac{1}{20}\int_0^{20} x dx \\ &= \frac{1}{20}\left[\frac{1}{2}x^2\right]_0^{20} \\ &= \frac{1}{20}\frac{1}{2}\left[(20)^2-(0)^2\right] \\ &= \frac{1}{40}[400] \\ &= 10.\end{aligned} \qquad \begin{aligned}\mathrm{E}(X^2) &= \int_{x\in R_X} x^2 f_X(x)dx \\ &= \int_0^{20} x^2\frac{1}{20} dx \\ &= \frac{1}{20}\left[\frac{1}{3}x^3\right]_0^{20} \\ &= \frac{1}{20}\frac{1}{3}\left[(20)^3-(0)^3\right] \\ &= \frac{1}{60}[8000] \\ &= \frac{400}{3}.\end{aligned}$$

Therefore, using the result $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$,

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \\ &= \frac{400}{3} - 10^2 \\ &= 33.33.\end{aligned}$$

More generally, it can be shown that

$$\mathrm{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

Example 8

Motivating example continued

Using the formulae above to calculate the expectation and variance in Example 5 requires much less work.

$$\mathrm{E}(X) = \frac{20}{2} = 10,$$

$$\mathrm{Var}(X) = \frac{20^2}{12} = 33.33.$$

Task 4 (Optional)

If $X \sim (a, b)$, show that

1. $\mathrm{E}(X) = \frac{a+b}{2}$,
2. $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Task 5

The worlds most famous geyser, Old Faithful in Yellowstone national park, erupts every 91 minutes. You arrive there at random and wait until it erupts.

1. How long would you expect to wait on Old Faithful erupting?
2. Find the variance and standard deviation of the time until Old Faithful erupts.

The exponential distribution

The **exponential distribution** is another common distribution, both in terms of its practical and theoretical use. It is often used to model the length of time until a specific event occurs, although it can be used in other applications. Examples include:

- the length of time a bee spends gathering nectar at a flower;
- the length of time until your next text message arrives;
- the length of time that a light bulb lasts;
- the size of a raindrop.

Generally, the exponential distribution is a skewed distribution, so there are more small values than large values for an exponential distribution random variable. For example, if you were interested in the amount of money customers spent in one trip to a department store, there would be more people who spend a small amount of money and fewer who spend a large amount of money.

Definition 4

The exponential distribution

The continuous random variable X has an **exponential** distribution with parameter $\lambda > 0$, written $X \sim \text{Exp}(\lambda)$, if X has the following p.d.f. and c.d.f.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Note: λ is often referred to as the rate of the distribution.

Supplement 2

Derivation of the cumulative distribution function

Suppose that $x > 0$, then

$$\begin{aligned} F_X(x) = P(X \leq x) &= \int_0^x f_X(t)\, dt = \int_0^t \lambda e^{\lambda t}\, dt \\ &= \left[-e^{-\lambda t}\right]_{t=0}^{x} = -e^{-\lambda x} - (-e^{-\lambda \times 0}) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Note that the `R` functions `dexp()` and `pexp()` compute values of the p.d.f. and the c.d.f. of the exponential distribution.

It can be shown that the mean and variance of the exponential distribution are

$$\text{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

Supplement 3

Derivation of expected value and variance

Using integration by parts,

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dt = \int_{0}^{+\infty} x\lambda e^{-\lambda x}\, dx = \underbrace{\left[-xe^{-\lambda x}\right]_{x=0}^{+\infty}}_{=0-0=0} + \frac{1}{\lambda}\underbrace{\int_{0}^{+\infty} \lambda e^{-\lambda x}}_{=1} = \frac{1}{\lambda}.$$

The second integral is one because it is the integral over the probability density function of the exponential distribution.

Using again integration by parts,

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{+\infty} x^2 f(x)\, dx = \int_{0}^{+\infty} x^2 \lambda e^{-\lambda x}\, dx \\ &= \underbrace{\left[-x^2 e^{-\lambda x}\right]_{x=0}^{+\infty}}_{=0-0=0} + \frac{2}{\lambda}\underbrace{\int_{0}^{+\infty} x\lambda e^{-\lambda x}\, dx}_{=E(X)=\frac{1}{\lambda}} = \frac{2}{\lambda^2} \end{aligned}$$

Finally,

$$\mathrm{Var}(X) = E(X) - (E(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Example 9

Motivating example

A busy lecturer at the University of Glasgow receives emails throughout the day. The time (in minutes) between emails is modelled with an exponential distribution with mean of 10. An email just arrived. What is the probability another email will not be received in the next 15 minutes?

Answer:

In this example let $X = \{\text{time between emails}\}$, then

$$X \sim \mathrm{Exp}(\lambda)$$

with the parameter $\lambda = \frac{1}{10}$ since we are told that the mean, $\mathrm{E}(X) = 10$.

The figure below shows the p.d.f. for X in this example.

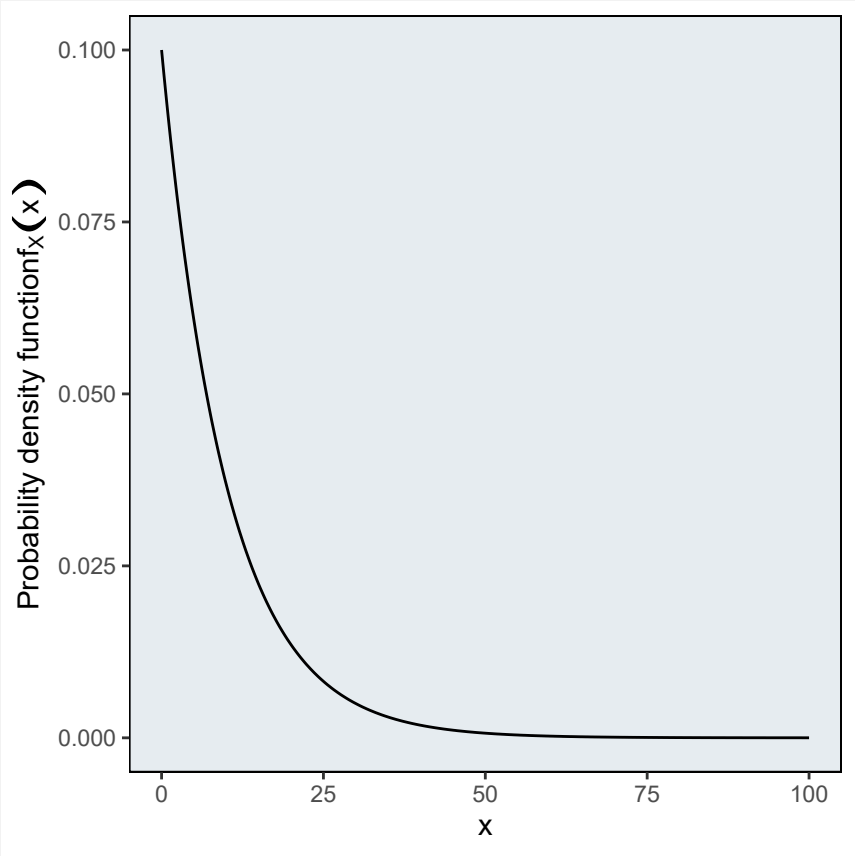


Figure 16

Notice that the graph is a decreasing curve and when $X = 0$,

$$f_X(x) = \lambda e^{(-\lambda \cdot 0)} = \lambda \cdot 1 = \lambda = \frac{1}{10}.$$

If no email is received in the next 15 minutes, then the time of the next email is greater than 15. The desired probability is

$$\mathrm{P}(X > 15) = 1 - \mathrm{P}(X \leq 15) = 1 - F_X(15) = e^{-15/10} = 0.223.$$

```
1 - pexp(15, rate=1/10)
```

R Console

```
[1] 0.2231302
```

The probability that we have calculated is shown in the figure below.

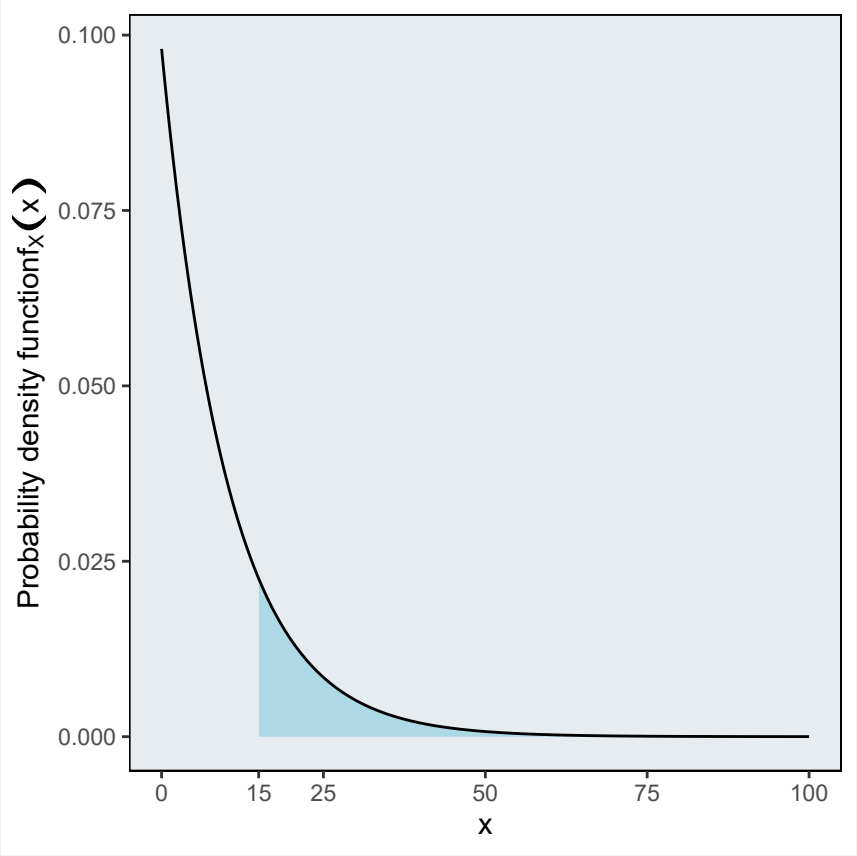


Figure 17

Task 6

The number of days ahead travellers purchase their airline tickets can be modelled by an exponential distribution with the average amount of time equal to 15 days.

1. Write down the distribution.
2. Find the probability that a traveller will book a ticket fewer than 10 days in advance.
3. Find the probability that a traveller will book their ticket between 10 and 15 days in advance.
4. How many days in advance will half of the travellers have booked their tickets by?

Supplement 4

The exponential distribution has no memory

An interesting property of the exponential distribution is that it has "no memory". Imagine that in Example 5, the person has already spent 10 minutes waiting for their train. Would this make the probability of them having to wait for at least another 5 minutes any lower? So in more general, if we know that we have already waited for x_0 minutes, what is the probability that we have to wait for at least another x minutes (i.e. wait for at least $x + x_0$ minutes in total)?

$$\begin{aligned} P(X > x + x_0 | X > x_0) &= \frac{P(X > x + x_0 \text{ and } X > x_0)}{P(X > x_0)} = \frac{P(X > x + x_0)}{P(X > x_0)} \\ &= \frac{1 - F_X(x + x_0)}{1 - F_X(x_0)} = \frac{e^{-\lambda(x+x_0)}}{e^{-\lambda x_0}} = \frac{e^{-\lambda x_0} e^{-\lambda x}}{e^{-\lambda x_0}} \\ &= e^{-\lambda x} = 1 - F_X(x) = P(X > x) \end{aligned}$$

So, the probability of having to wait for at least another x minutes is independent of how long we have already been waiting (x_0 minutes).

This is not always a realistic assumption. If we were to assume that the life time of a component is exponentially distributed, then the memorylessness means that we assume that its reliability is not related to its age, which is often not realistic.

Supplement 5

Relationship to Poisson distribution

Imagine we observe a process in which the number of events occurring during any time interval of length t has a Poisson distribution with rate λt (with counts in non-overlapping intervals being independent). In this so-called Poisson process both the waiting time between two events and the waiting time until the next event occurs can be shown to have an exponential distribution with rate λ.

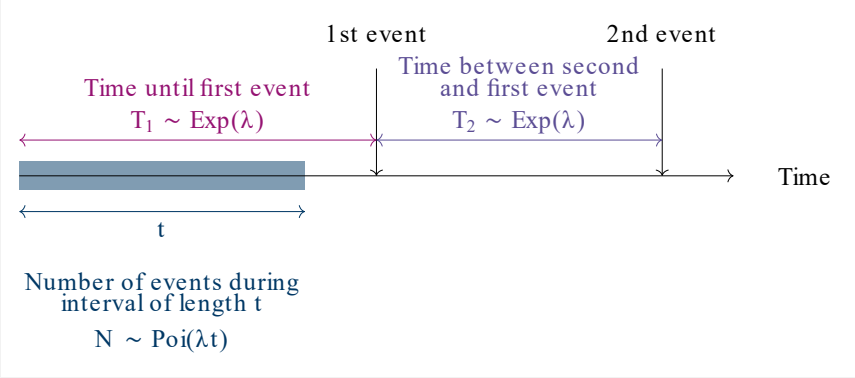


Figure 21

To illustrate this link, let's compute the probability that no events occur during an interval of time t.

- Using the fact that the number of events N occurring during this interval has a Poisson distribution with rate λt,

$$P(N = 0) = \frac{e^{-\lambda t}\lambda^0}{0!} = e^{-\lambda t}$$

- No event occurring during an interval of length t is the same as having to wait for at least t until the first event occurs. Using the fact that this waiting time has an exponential distribution with rate λ we obtain:

$$P(T_1 > t) = 1 - F_{T_1}(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$$

Both ways of calculating this probability give the same answer, illustrating this link between the Poisson distribution and the exponential distribution.

Other distributions

The t-distribution

Earlier in this week we introduced the normal distribution which is used to model a wide variety of data sets. We learned that this distribution is completely defined if we know its mean μ and variance (or standard deviation) σ^2 (σ). These parameters are often referred to as the **population** mean and variance. In practice we rarely have access to data for an entire population and so μ and σ^2 are **unknown** and have to be estimated from a **sample** of data.

The t-distribution is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown. Compared to the normal distribution, the t-distribution has heaver tails Which means the peak is lower

The t-distribution is defined by it's **degrees of freedom** ν, and is denoted $t(\nu)$. The degrees of freedom (df) of an estimate is the number of independent pieces of information that go into the estimate. In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated when calculating the estimate in question. For example, to estimate the population variance, one must first estimate the population mean. Therefore, if the estimate of variance is based on n observations, there are $n-1$ degrees of freedom. The t-distribution is very similar to the standard normal distribution when the estimate of the variance is based on many degrees of freedom. The figure below shows a comparison of t-distributions with 2, 4, and 10 df and the standard normal distribution. You can see that the distribution with the lowest peak and the heaviest tails is $t(2)$ and as the degrees of freedom increase, the t-distribution approaches the standard normal ($\mathrm{N}(0,1)$).

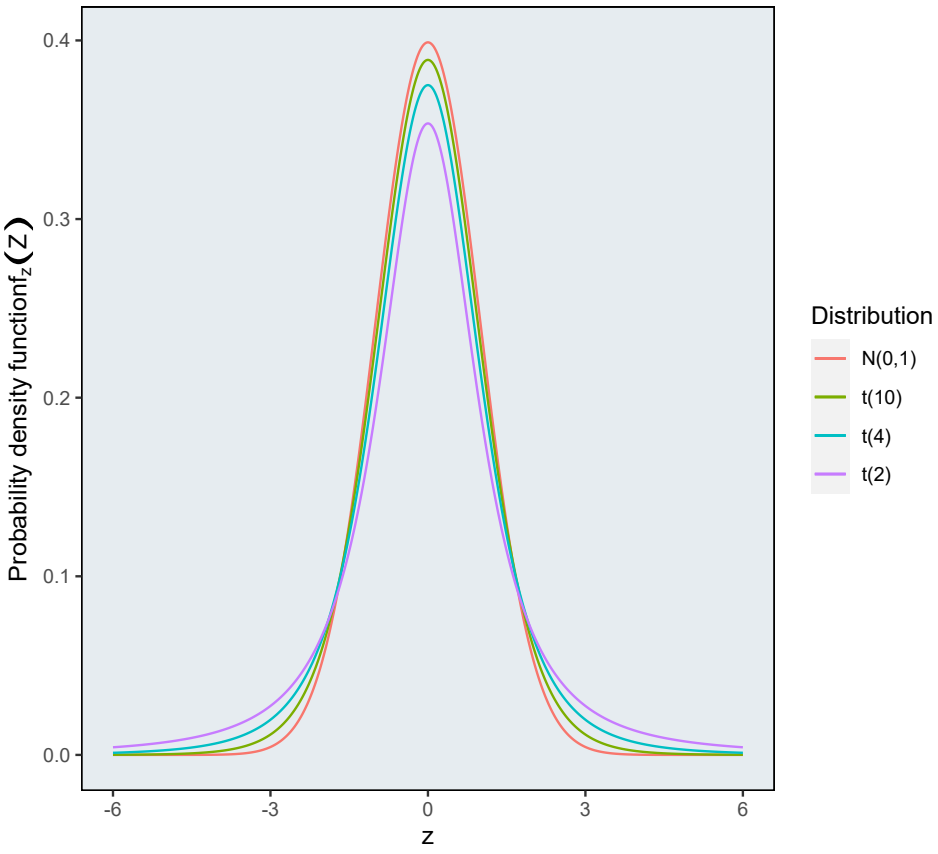


Figure 22

The χ^2 distribution

The χ^2 distribution with k degrees of freedom is the distribution of a sum of squares of k **independent standard normal** random variables. In other words, the degrees of freedom of the distribution is equal to the number of standard normal random variables being summed. Therefore, a χ^2 with one degree of freedom, $\chi^2(1)$, is the distribution of a single normal random variable squared.

The χ^2 distribution is important as many test statistics (some which you will learn about in Learning from Data/Data Science Foundations) are approximately distributed as χ^2.

The mean of a χ^2 distribution is its degrees of freedom and χ^2 distributions are positively skewed with the skewness decreasing as the degrees of freedom increase. Similarly to the t-distribution, as the degrees of freedom increase, the χ^2 distribution approaches a normal distribution. The figure below shows a comparison of χ^2 distributions with 2, 4, and 10 df.

Both the χ^2 distribution and the exponential distribution are special cases of a more general distribution, the gamma distribution.

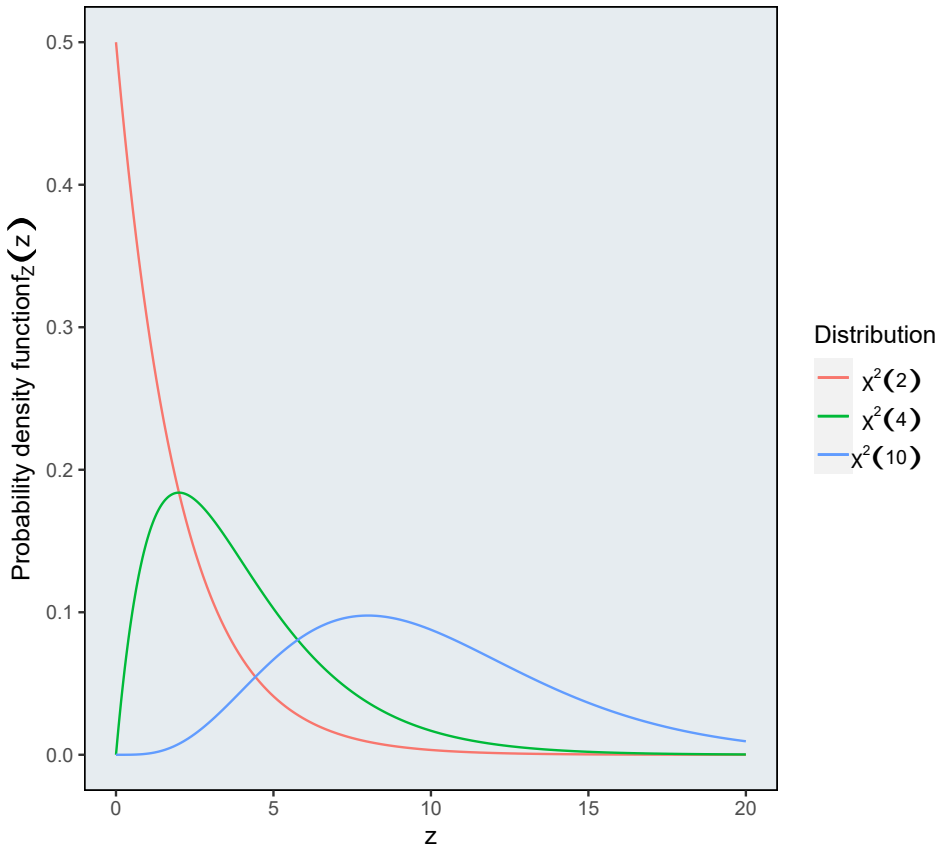


Figure 23

Learning outcomes for week 6

By the end of week 6, you should be able to:

- recognise some of the standard continuous probability distributions in a context;
- obtain probabilities and percentiles for a normal distribution;
- obtain probabilities and percentiles and calculate the expectation and variance of a uniform and exponential distribution;
- know how the t and χ^2 distributions relate to the normal distribution.

A summary of the most important concepts, selected video solutions and written answers to all tasks are provided overleaf.

Week 6 summary

The normal distribution

Probability density function

Suppose that the random variable X can take any real value and that X has p.d.f.

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

then X is said to have a **normal** distribution, with parameters μ and σ^2, written

$$X \sim \mathrm{N}(\mu, \sigma^2).$$

Important properties

The normal distribution is symmetrical and `bell-shaped' and is completely defined if we know its mean μ and variance σ^2.

It can be shown that

$$\mathrm{E}(X) = \mu \quad \text{and} \quad \mathrm{Var}(X) = \sigma^2.$$

Calculating probabilities

We calculate probabilities of the form $\mathrm{P}(X \leq x)$ in two stages using the **standard normal distribution** which is a special case of the normal distribution which has a mean of 0 and a standard deviation (and hence a variance) of 1.

1. Transform $\mathrm{P}(X \leq x)$ into a probability of the form $\mathrm{P}(Z \leq z)$, where Z is a standard normal random variable with mean 0 and variance 1.
2. Use statistical tables of the c.d.f. of the standard normal distribution to calculate the probability.

The uniform distribution

The continuous random variable X is said to have the **uniform** distribution on the interval $[a, b]$, written $X \sim \mathrm{U}(a, b)$, if X has the following p.d.f. and c.d.f.:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}, \qquad F_X(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a}, & a < x \le b \\ 1, & \text{otherwise} \end{cases}.$$

It can be shown that the mean and variance of the uniform distribution are

$$\mathrm{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

The exponential distribution

The continuous random variable X has an **exponential** distribution with parameter $\lambda > 0$, written $X \sim \mathrm{Exp}(\lambda)$, if X has the following p.d.f. and c.d.f.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}.$$

It can be shown that the mean and variance of the exponential distribution are

$$\mathrm{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Answer 1

1. $\mathrm{P}(Z < -1) = \Phi(-1) = 1 - \Phi(1) \approx 1 - 0.8413 = 0.1587.$

```
pnorm(-1)
```

```
R Console
[1] 0.1586553
```

2. $\mathrm{P}(-2 < Z < -1) = \Phi(-1) - \Phi(-2) = 1 - \Phi(1) - [1 - \Phi(2)] = \Phi(2) - \Phi(1) \approx 0.9772 - 0.841$

```
pnorm(-1) - pnorm(-2)
```

```
R Console
```

```
[1] 0.1359051
```

3.

$$\mathrm{P}(-1.65 < Z < 1.65) = \Phi(1.65) - \Phi(-1.65) = \Phi(1.65) - [1 - \Phi(1.65)] = 2\Phi(1.65) - 1 \approx 2 \times 0.950$$

```
pnorm(1.65) - pnorm(-1.65)
```

R Console

```
[1] 0.9010571
```

Here is a video worked solution.

Video

Week 6 - Task 1

Duration 6:49

Answer 2

For all examples $\mu = 7$ and $\sigma = \sqrt{16} = 4$.

1.

$$\begin{aligned}\mathrm{P}(X < 9) &= \mathrm{P}\left(\frac{X-7}{4} < \frac{9-7}{4}\right) \\ &= \mathrm{P}(Z < 0.5) \quad \text{where } Z \text{ has the standard normal distribution} \\ &= \Phi(0.5) \approx 0.6915 \text{ from Table.}\end{aligned}$$

```
pnorm(9, mean =7, sd=4)
```

R Console

```
[1] 0.6914625
```

2.

$\mathrm{P}(X > 9) = 1 - \mathrm{P}(X < 9) = 0.3085.$

```
1 - pnorm(9, mean =7, sd=4)
```

R Console

```
[1] 0.3085375
```

3.

$$\begin{aligned}\mathrm{P}(4 < X < 9) &= \mathrm{P}\left(\frac{4-7}{4} < \frac{X-7}{4} < \frac{9-7}{4}\right)\\ &= \mathrm{P}(-3/4 < Z < 1/2)\\ &= \Phi(1/2) - \Phi(-3/4)\\ &= 0.6915 - 0.2266 = 0.4649.\end{aligned}$$

```
pnorm(9, mean =7, sd=4) - pnorm(4, mean =7, sd=4)
```

R Console

```
[1] 0.4648351
```

4.

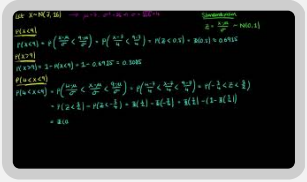
$$0.95 = \mathrm{P}(X < c) = \mathrm{P}\left(\frac{X-7}{4} < \frac{c-7}{4}\right) = \mathrm{P}\left(Z < \frac{c-7}{4}\right) = \Phi((c-7)/4).$$

From the table, $\Phi(1.645) = 0.95$, so that

$$\frac{c-7}{4} = 1.645 \quad \text{and} \quad c = 13.58.$$

Here is a video worked solution.

Video



Week 6 - Task 2

Duration 9:11

Answer 3

1. Let $X = \{\text{time until Old Faithful erupts}\}$. Since the time until the next eruption could lie between 0 and 91 mins, with every value equally likely, then X can be modelled using the uniform distribution with parameters, $a = 0, b = 91$, i.e.

$$X \sim \mathrm{U}(0, 91).$$

2.

$$f_X(x) = \begin{cases} \frac{1}{91}, & 0 \leq x \leq 91 \\ 0, & \text{otherwise} \end{cases}$$

3.

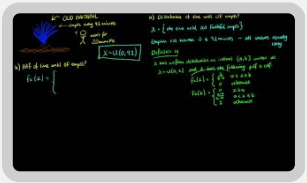
Using the fact that

$$F_X(x) = \begin{cases} \frac{x-a}{b-a}, & a < x \leq b \\ 1, & \text{otherwise} \end{cases}$$

so $\mathrm{P}(X < 20) = F_X(20) = \frac{20-0}{91-0} = 0.22$.

Here is a video worked solution.

Video



Week 6 - Task 3

Duration 4:56

Answer 4

1.

$$\begin{aligned}\mathrm{E}(X) &= \int_a^b \frac{x}{b-a}\mathbf{d}x \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{b+a}{2}.\end{aligned}$$

2.

$$\begin{aligned}\mathrm{E}(X^2) &= \int_a^b \frac{x^2}{b-a}\mathbf{d}x \\ &= \left[\frac{x^3}{3(b-a)}\right]_a^b \\ &= \frac{b^3-a^3}{3(b-a)} \\ &= \frac{a^2+ab+b^2}{3}.\end{aligned}$$

Then

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{(b+a)^2}{4} \\ &= \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12} \\ &= \frac{(a-b)^2}{12}.\end{aligned}$$

Answer 5

From Task 3 we know that X can be modelled using the uniform distribution with parameters, $a = 0, b = 91$, i.e.

$$X \sim \mathrm{U}(0, 91).$$

1. Use

$$\mathrm{E}(X) = \frac{a+b}{2}$$

to get

$$\mathrm{E}(X) = \frac{91}{2} = 45.5.$$

2. Use

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

to get

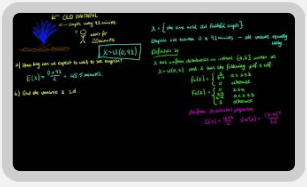
$$\mathrm{Var}(X) = \frac{91^2}{12} = 690.08.$$

and

$$\mathrm{sd}(X) = \sqrt{690.08} = 26.27.$$

Here is a video worked solution.

Video



Week 6 - Task 3 (continued)

Duration 2:05

Answer 6

1. Let $X = \{\text{number of days ahead tickets purchased}\}$. Then

$$X \sim \mathrm{Exp}\left(\frac{1}{15}\right).$$

2.

$$
\begin{aligned}
\mathrm{P}(X < 10) &= F_X(10) \\
&= 1 - e^{-(\frac{1}{15} \cdot 10)} \\
&= 0.487
\end{aligned}
$$

So the probability that a traveller books their flight less than 10 days in advance is 0.487.

```
pexp(10, 1/15)
```

R Console

```
[1] 0.4865829
```

The probability that we have calculated is shown in the figure below.

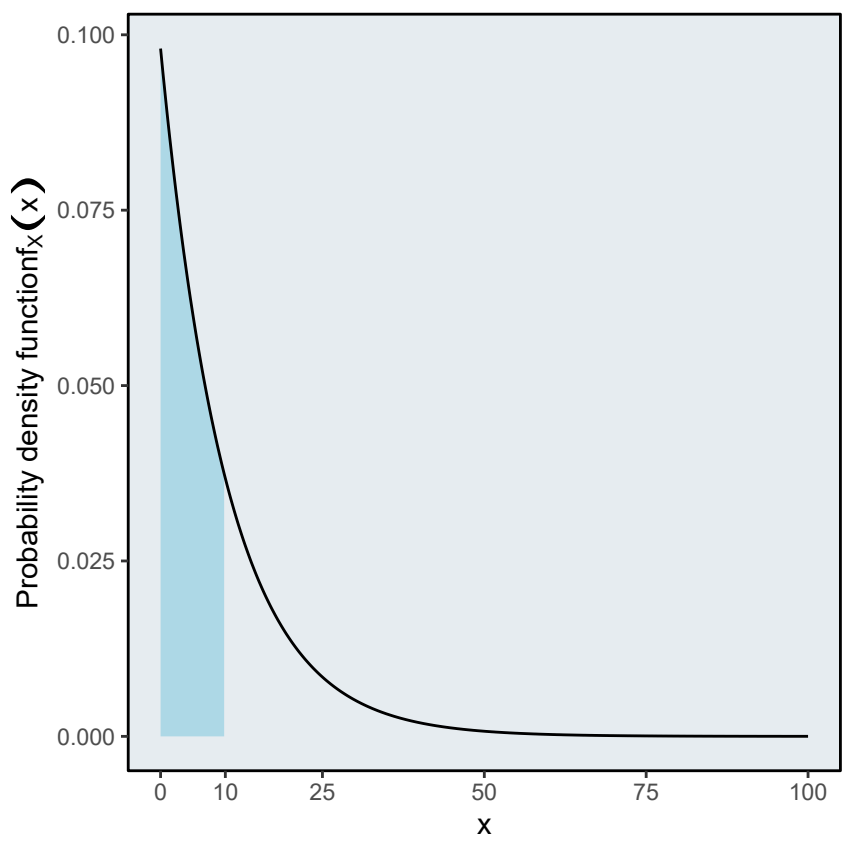


Figure 18

3.

$$
\begin{aligned}
\mathrm{P}(10 < X < 15) &= \mathrm{P}(X < 15) - \mathrm{P}(X < 10) \\
&= F_X(15) - F_X(10) \\
&= (1 - e^{-\frac{1}{15} \cdot 15}) - (1 - e^{-\frac{1}{15} \cdot 10}) \\
&= 0.632 - 0.487 \\
&= 0.145
\end{aligned}
$$

So the probability that a traveller books their flight between 10 and 15 days in advance is 0.145.

```
pexp(15, 1/15) - pexp(10, 1/15)
```

```
R Console

[1] 0.1455377
```

The probability that we have calculated is shown in the figure below.

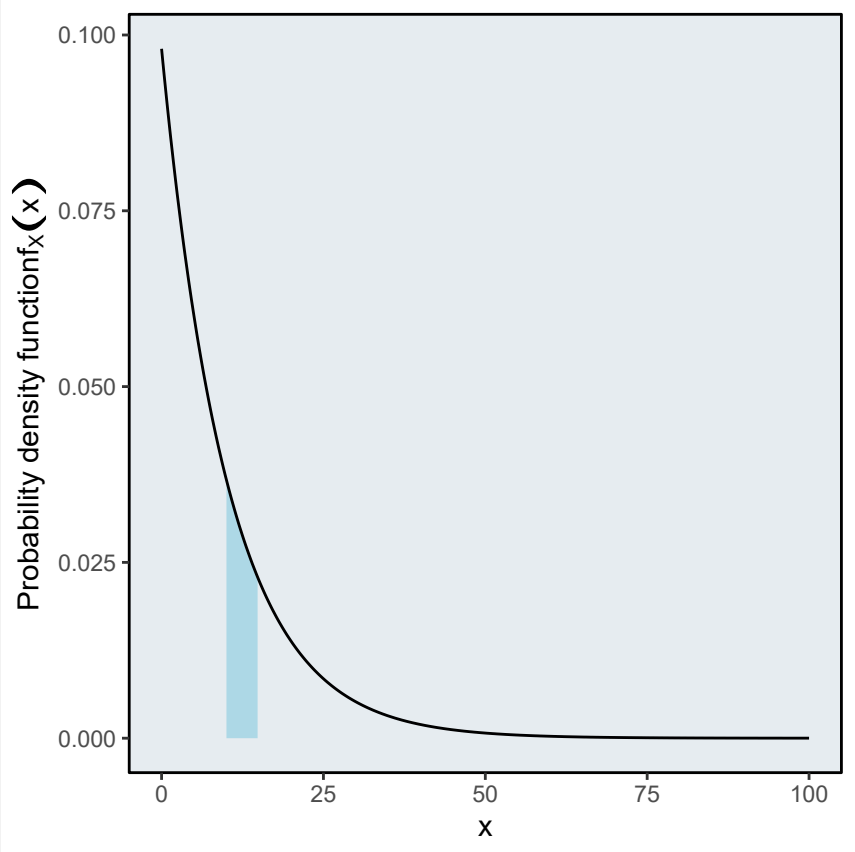


Figure 19

4.

To answer this question we need to find the 50^{th} percentile, i.e.

$$\mathrm{P}(X < c) = 0.50 \quad \text{and} \quad \mathrm{P}(X < c) = 1 - e^{-\frac{1}{15}c}.$$

Note: Since we are looking for the 50^{th} percentile, $\mathrm{P}(X < c) = \mathrm{P}(X > c) = 0.5$.

$$\begin{aligned} 0.50 &= 1 - e^{-\frac{1}{15}c} \\ e^{-\frac{1}{15}c} &= 1 - 0.5 = 0.5 \\ \ln(e^{-\frac{1}{15}c}) &= \ln(0.5) \\ -\frac{1}{15}c &= \ln(0.5) \\ c &= \frac{\ln(0.5)}{-1/15} \\ c &= 10.397. \end{aligned}$$

So 50% of travellers will have booked their tickets by 10.397 days in advance.

```
qexp(0.5, 1/15)
```

R Console

```
[1] 10.39721
```

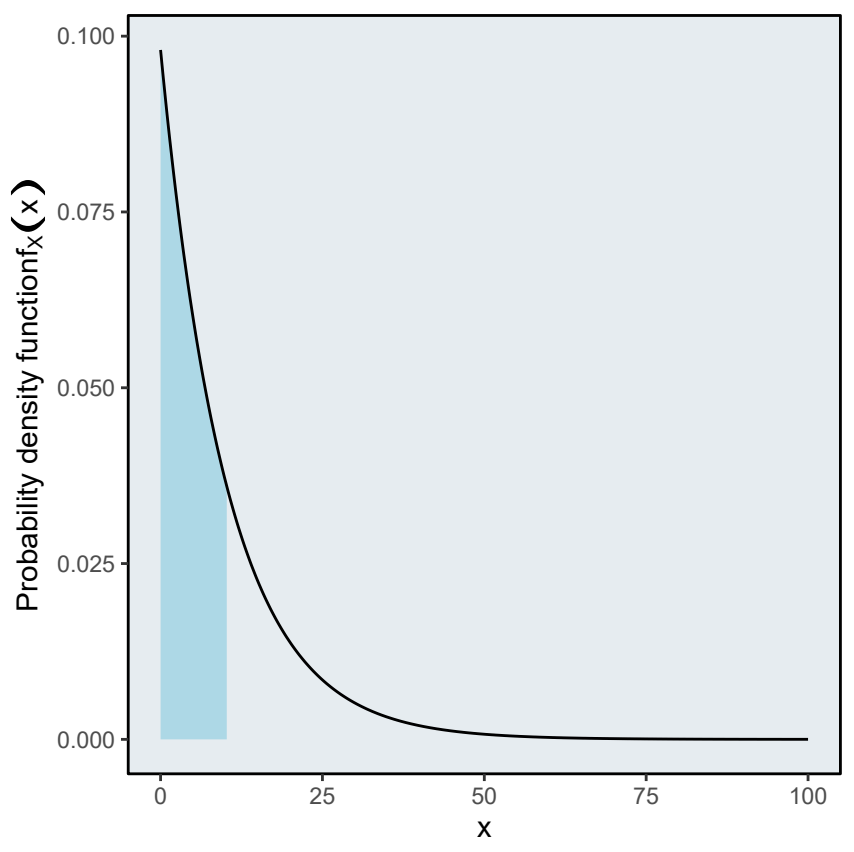
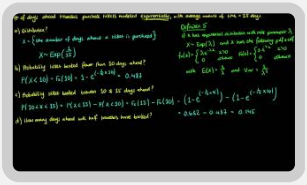


Figure 20

Note here that the 50^{th} percentile, or the median, is 10.397 whereas the mean is 15 days. So here the mean is larger than the median. We would expect the mean and median to be different here since the exponential distribution is not **symmetric**.

Here is a video worked solution.

Video



Week 6 - Task 5

Duration 6:20