Supporting Australian Mathematics Project

A guide for teachers – Years 11 and 12

Probability and statistics: Module 22

Exponential and normal distributions

Education Services Australia

AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Exponential and normal distributions – A guide for teachers (Years 11–12)

Professor Ian Gordon, University of Melbourne

Editor: Dr Jane Pitkethly, La Trobe University

Illustrations and web design: Catherine Tan, Michael Shaw

Full bibliographic details are available from Education Services Australia.

Published by Education Services Australia
PO Box 177
Carlton South Vic 3053
Australia

Tel: (03) 9207 9600
Fax: (03) 9910 9800
Email: info@esa.edu.au
Website: www.esa.edu.au

This publication is funded by the Australian Government Department of Education, Employment and Workplace Relations.

Supporting Australian Mathematics Project

Australian Mathematical Sciences Institute
Building 161
The University of Melbourne
VIC 3010
Email: enquiries@amsi.org.au
Website: www.amsi.org.au

Exponential and Normal distributions

Assumed knowledge

The content of the modules:

- *Probability*
- *Discrete probability distributions*
- *Binomial distribution*
- *Continuous probability distributions*.

Motivation

The module *Continuous probability distributions* covered the basic ideas involved in continuous distributions. In this module, we meet two of the more important continuous distributions: the exponential distribution and the Normal distribution.

The exponential distribution is used for the waiting time until the first event in a random process where events are occurring at a given rate. It is a relatively simple distribution; a random variable having this distribution is necessarily positive, and it is one of the more important distributions among those used for positive random variables.

There is good reason to say that the Normal distribution is the most important distribution of all, principally because of a result known as the *central limit theorem*, which is covered in the module *Inference for means*. This distribution is characterised by the well-known 'bell curve'.

In this module, we cover the calculation of probabilities and quantiles associated with the exponential distribution and the Normal distribution. Through examples, we will see how these distributions can be applied to solve practical problems.

Content

Continuous random variables: brief review

Random variables are introduced in the module *Discrete probability distributions*. Recall that a random variable is a variable whose value is determined by the outcome of a random procedure.

There are two main types of random variables: discrete and continuous. The modules *Discrete probability distributions* and *Binomial distribution* deal with discrete random variables, and the module *Continuous probability distributions* introduces continuous random variables and their distributions.

In this module, we study two specific continuous distributions, so we will be applying much of the theory developed in the module *Continuous probability distributions*.

A **continuous random variable** X can take any real value within a specified range. It has a probability density function (pdf) denoted by $f_X(x)$ and a cumulative distribution function (cdf) denoted by $F_X(x)$. Recall that

$$\Pr(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt,$$

for each real number x.

Exponential distribution

The **exponential distribution** is defined as follows. Suppose that the continuous random variable T has an exponential distribution with rate $\alpha > 0$, which we write as $T \stackrel{\mathrm{d}}{=} \exp(\alpha)$. Then T has the following pdf:

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Four exponential pdfs are shown in figure 1 on the same scale. Note that they all have the same shape. The greater the rate α, the more likely it is that the corresponding exponential random variable takes a small value. This makes sense: if the events are occurring at a high rate, it will tend to be a short time until the first event, and vice versa.

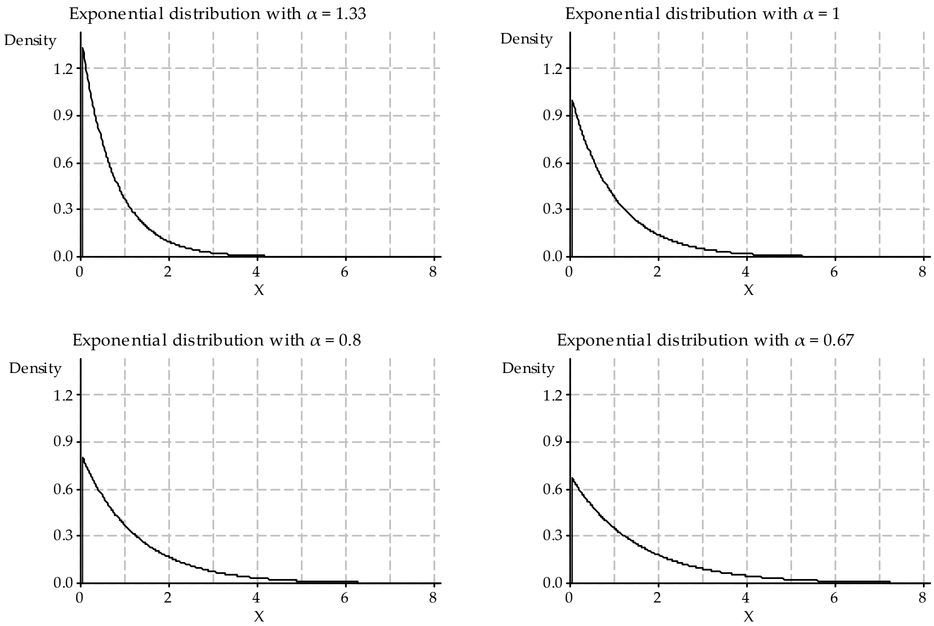

Figure 1: Four exponential probability density functions with different values of α.

An exponential random variable can be regarded as the waiting time until the first event in a Poisson process with rate α. The curriculum does not cover Poisson processes, so we need to describe them briefly here. It is appropriate to think of a 'random process' in which events occur in time, independently of each other, at a rate per unit time. This means that processes that are systematic (such as train timetables) or approximately regular (the arrival of waves on a beach) are not Poisson processes. Examples of phenomena that might be suitably modelled with this distribution include:

- radioactive decay
- the occurrences of a rare disease in a large population
- arrival of a packet of information on the internet.

Example: Country hospital

Let T be the interval between births at a country hospital, for which the average time between births is seven days. We assume the distribution of the time between births follows an exponential distribution. Clearly, multiple births (twins, triplets, ...) will violate the assumption of independence; we deal with this by defining a 'birth' to be a birth event for one mother, regardless of the number of babies born. The unit of time is 'day', and the corresponding average rate of events is one birth every seven days, so that $\alpha = \frac{1}{7}$. Hence, $T \stackrel{d}{=} \exp(\frac{1}{7})$ and

$$f_T(t) = \begin{cases} \frac{1}{7} e^{-\frac{1}{7} t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We now consider some of the characteristics of the exponential distribution.

Mean of the exponential distribution

If $T \stackrel{\mathrm{d}}{=} \exp(\alpha)$, *then* $\mu_T = \mathrm{E}(T) = \frac{1}{\alpha}$.

Proof

We have

$$\begin{aligned}
\mathrm{E}(T) &= \int_0^\infty t\, f_T(t)\, dt \\
&= \int_0^\infty t\alpha e^{-\alpha t}\, dt \\
&= \big[-te^{-\alpha t}\big]_0^\infty + \int_0^\infty e^{-\alpha t}\, dt \qquad \text{(using integration by parts)} \\
&= 0 - 0 + \big[-\tfrac{1}{\alpha} e^{-\alpha t}\big]_0^\infty \\
&= 0 + \tfrac{1}{\alpha} \\
&= \tfrac{1}{\alpha}.
\end{aligned}$$

□

The proof for the variance also uses integration by parts; it is not provided here.

Variance of the exponential distribution

If $T \stackrel{\mathrm{d}}{=} \exp(\alpha)$, *then* $\mathrm{var}(T) = \frac{1}{\alpha^2}$.

The cdf of T is given by

$$\Pr(T \le t) = F_T(t) = \int_{-\infty}^t f_T(u)\, du.$$

So, for $t \le 0$, we have $F_T(t) = 0$, and for $t > 0$, we have

$$F_T(t) = \int_0^t \alpha e^{-\alpha u}\, du = \big[-e^{-\alpha u}\big]_0^t = 1 - e^{-\alpha t}.$$

Hence, we can write the cdf of T as

$$\Pr(T \le t) = F_T(t) = \begin{cases} 0 & \text{if } t \le 0, \\ 1 - e^{-\alpha t} & \text{if } t > 0. \end{cases}$$

An obvious consequence of the cdf having this form is that the probability of the waiting time T exceeding t is

$$\Pr(T > t) = e^{-\alpha t}, \qquad \text{for } t > 0.$$

Example: Country hospital, continued

We return to the example of births at a country hospital, in which we assume that the time between successive births, T, has an exponential distribution with rate $\frac{1}{7}$; that is, $T \stackrel{d}{=} \exp(\frac{1}{7})$. Hence, we have

- $\mu_T = \mathrm{E}(T) = 7$ days
- $\sigma_T^2 = \mathrm{var}(T) = 49$
- $\sigma_T = \mathrm{sd}(T) = 7$ days.

What is the chance that there is a birth in the next 10 days? 10 hours? 10 minutes?

In general, the chance of a birth in the next t days is

$$F_T(t) = 1 - e^{-\frac{1}{7}t}, \qquad \text{for } t > 0.$$

The unit of time being used here is days. The probability that the waiting time until the next birth is less than or equal to 10 days is therefore

$$\begin{aligned} F_T(10) &= 1 - \exp\big(-\tfrac{10}{7}\big) \\ &= 1 - e^{-1.429} = 1 - 0.240 = 0.760. \end{aligned}$$

Ten hours equals $\frac{5}{12}$ days. The probability that the waiting time until the next birth is less than or equal to 10 hours is therefore

$$\begin{aligned} F_T\big(\tfrac{5}{12}\big) &= 1 - \exp\big(-\tfrac{1}{7} \times \tfrac{5}{12}\big) \\ &= 1 - e^{-0.060} = 0.058. \end{aligned}$$

Finally, note that 10 minutes equals $\frac{1}{144}$ days, so the probability that the waiting time until the next birth is less than or equal to 10 minutes is

$$\begin{aligned} F_T\big(\tfrac{1}{144}\big) &= 1 - \exp\big(-\tfrac{1}{7} \times \tfrac{1}{144}\big) \\ &= 1 - e^{-0.00099} = 0.00099. \end{aligned}$$

Note that, for small values of k, we have $1 - e^{-k} \approx 1 - (1 - k) = k$. Hence, if αt is small, then the chance of an event before time αt is approximately equal to αt. This approximation applies to the second and third cases here.

An intriguing feature of the exponential distribution is its **lack of memory property**. Roughly speaking, it is as the name suggests: the process 'does not remember what has happened up until now' and the distribution of the waiting time, given that it has already exceeded some amount of time t_0, has the same exponential-distribution form, just translated by t_0.

The lack of memory property is quite readily established. For $t_0 > 0$ and $t > t_0$:

$$\begin{aligned}
\Pr(T > t \mid T > t_0) &= \frac{\Pr(T > t \text{ and } T > t_0)}{\Pr(T > t_0)} && \text{(rule for conditional probability)} \\
&= \frac{\Pr(T > t)}{\Pr(T > t_0)} && \text{(since "}T > t\text{"} \subseteq \text{"}T > t_0\text{")} \\
&= e^{-\alpha(t - t_0)}.
\end{aligned}$$

Exercise 1

Suppose that the time between emergency calls to a small suburban fire station follows an exponential distribution with an average rate of 1.8 calls per day.

- **a** Phil the fireman has just clocked on. What is the chance of a call in the next 15 minutes?
- **b** Phil has nearly finished his shift: 15 minutes to go. There has been no call during his shift so far. What is the chance of a call in the next 15 minutes?
- **c** Judy works a 10-hour shift, Mondays to Thursdays. What is the probability that she has no calls in a shift?
- **d** What is the probability that she has no calls in four successive days?
- **e** Judy is talking about her job: 'In 10% of shifts, there's a call in the first x hours of the shift.' What is x, to one decimal place?

Normal distribution

The Normal distribution is arguably the most important continuous distribution. It is used throughout the sciences, because of a remarkable result known as the *central limit theorem*, which is covered in the module *Inference for means*. Due to the phenomenon behind the central limit theorem, many variables tend to show an empirical distribution that is close to the Normal distribution.

If X has a **Normal distribution** with mean μ and standard deviation σ, then we write that $X \stackrel{d}{=} \mathrm{N}(\mu, \sigma^2)$; the probability density function of X is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big), \qquad \text{for } x \in \mathbb{R}.$$

This distribution is so important that it is well known in general culture, where it is often referred to as the **bell curve** — for example, in the controversial 1994 book by R. J. Herrnstein entitled *The Bell Curve: Intelligence and Class Structure in American Life*.

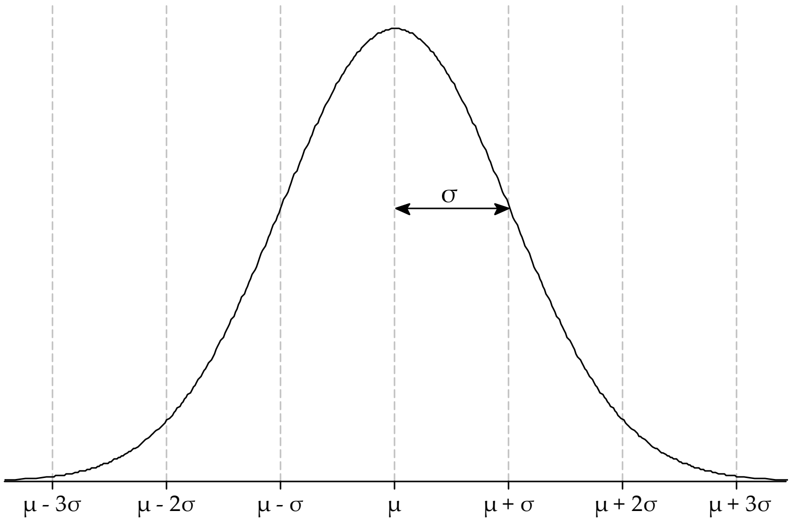

Figure 2: The pdf of a Normal random variable with mean μ and standard deviation σ.

Several properties of the Normal distribution are worth noting:

- It is easy to see from the formula for $f_X(x)$ that the distribution is symmetric around $x = \mu$. By the properties of the mean, this confirms that $\mu_X = \mathrm{E}(X) = \mu$.
- The pdf has one peak, which is at $x = \mu$.
- The pdf has two points of inflexion, where the second derivative of the pdf changes sign. They are at $x = \mu - \sigma$ and $x = \mu + \sigma$; see figure 2. This is very useful when we need to graph a Normal pdf with a given μ and σ: we can use μ to position the curve correctly, and σ to get the scale right.

  When drawing the bell-shaped curve, it can sometimes be easier to write the value of μ on the x-axis, annotate the pdf with the value of σ for the distance between μ and $\mu + \sigma$, and then fill in the scale on the x-axis. At the very least, it is helpful to show the actual value of σ on the plot, when thinking about a practical application.
- For the Normal distribution:

$$\begin{aligned}
\Pr(\mu - \sigma \le X \le \mu + \sigma) &= 0.6827 \\
\Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) &= 0.9545 \\
\Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) &= 0.9973.
\end{aligned}$$

  These probabilities are often thought of more approximately as 68.3%, 95.4% and 99.7%, or even as 68%, 95% and 99.7%; they are illustrated in figure 3.

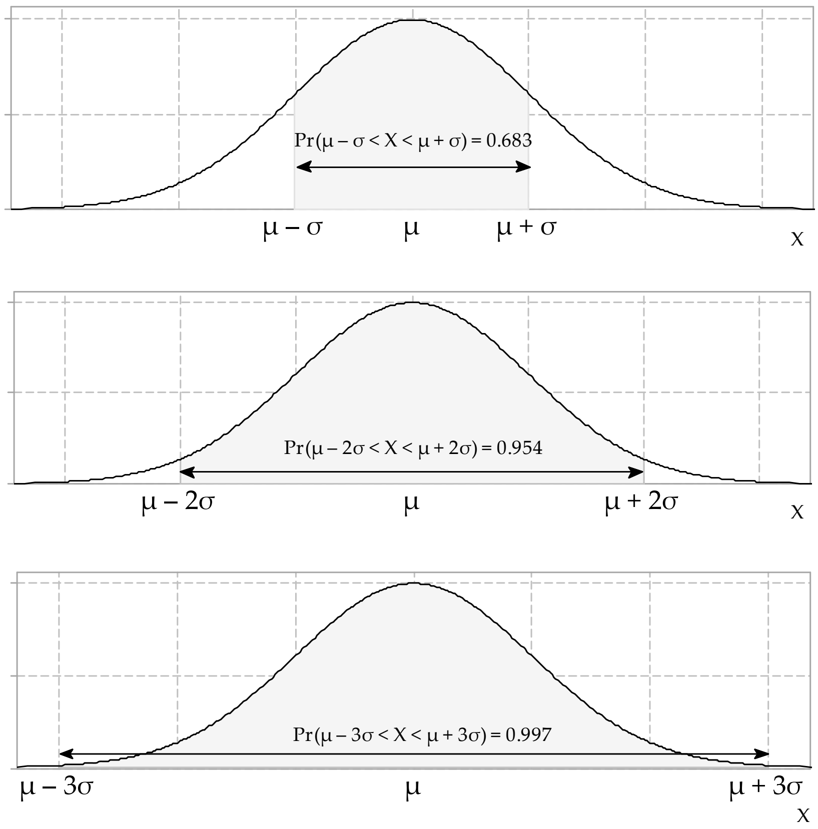

Figure 3: Probabilities of three intervals for the Normal distribution.

Exercise 2

Suppose that $f_X(x)$ is the pdf of a Normal random variable with mean μ and standard deviation σ.

- **a** What is the value of $f_X(\mu)$?
- **b** Show that $f_X(\mu)$ is the maximum value of f_X.
- **c** Show that f_X has points of inflexion at $x = \mu \pm \sigma$.
- **d** Find $f_X(\mu + k\sigma)$ for $k = 0, 1, 2, 3, 4, 5$ and interpret the result.

Recall that, for continuous random variables, it is the cumulative distribution function (cdf) and not the pdf that is used to find probabilities, because we are always concerned with the probability of the random variable being in an interval.

Before considering the cdf of $X \stackrel{d}{=} \mathrm{N}(\mu, \sigma^2)$, we explore a very useful feature of the Normal distribution.

A random variable with the **standard Normal distribution**, commonly denoted by Z, has mean zero and standard deviation one. That is, $Z \stackrel{\mathrm{d}}{=} \mathrm{N}(0,1)$. The pdf for the standard Normal distribution is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}z^2\right), \qquad \text{for } z \in \mathbb{R}.$$

The probabilities for any Normal distribution can be reduced to probabilities for the standard Normal distribution, using the device of standardisation. Therefore probability calculations for any Normal distribution can be reduced to calculations for the standard Normal distribution, as shown by the following result.

Standardisation of a Normal distribution

If $X \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu,\sigma^2)$ *and* $X_s = \dfrac{X-\mu}{\sigma}$, *then* $X_s \stackrel{\mathrm{d}}{=} \mathrm{N}(0,1)$.

Proof

The result is established by first considering the cdf of X_s. We have

$$\begin{aligned} F_{X_s}(z) &= \Pr(X_s \le z) \\ &= \Pr\left(\frac{X-\mu}{\sigma} \le z\right) \\ &= \Pr(X \le \sigma z + \mu) \\ &= F_X(\sigma z + \mu). \end{aligned}$$

Hence,

$$\begin{aligned} f_{X_s}(z) &= \frac{d}{dz} F_{X_s}(z) \\ &= \frac{d}{dz} F_X(\sigma z + \mu) \\ &= \sigma f_X(\sigma z + \mu) \qquad \text{(by the chain rule)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}z^2\right). \end{aligned}$$

It follows that $X_s \stackrel{\mathrm{d}}{=} \mathrm{N}(0,1)$. $\square$

Finding probabilities for the standard Normal distributions requires technology: the cdf of $Z \stackrel{d}{=} N(0,1)$ is

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}t^2\right) dt.$$

This integral does not have a closed form, and must be evaluated using numerical integration. It is available in statistical software, on many calculators, in Matlab and in Excel; here we describe the Excel function. It is NORM.S.DIST, which requires two arguments:

1. the value of z for which the cdf is required
2. a true/false (or equivalently, 1/0) argument that controls whether the cdf (argument equals TRUE or 1) or pdf (argument equals FALSE or 0) is returned.

For example, to use Excel to find the value of $F_Z(1.5)$, the cdf of the standard Normal distribution when $z = 1.5$, enter

```
=NORM.S.DIST(1.5, 1)
```

in a cell and hit return. You should obtain the value 0.9332.

Example: Crowd size

Suppose that crowd size at home games for a particular football club follows a Normal distribution with mean 26 000 and standard deviation 5000. What percentage of crowds are between 31 000 and 36 000?

We standardise to solve this. Let $X \stackrel{d}{=} N(26\,000, 5000^2)$. Then $X_s = \dfrac{X - 26\,000}{5000} \stackrel{d}{=} N(0,1)$, and therefore

$$\begin{aligned}
\Pr(31\,000 < X < 36\,000) &= \Pr\Big(\frac{31\,000 - 26\,000}{5000} < \frac{X - 26\,000}{5000} < \frac{36\,000 - 26\,000}{5000}\Big) \\
&= \Pr(1 < X_s < 2) \\
&= F_{X_s}(2) - F_{X_s}(1) \\
&= 0.9772 - 0.8413 \\
&= 0.1359.
\end{aligned}$$

Note that, in this example, 31 000 = $\mu + \sigma$ and 36 000 = $\mu + 2\sigma$. If the mean and the standard deviation were different from these, but we still sought the probability of being between one and two standard deviations greater than the mean, then the same probability would be obtained. This is illustrated in figure 4, in which the same probability as that obtained in the example (0.1359) is found in all four cases.

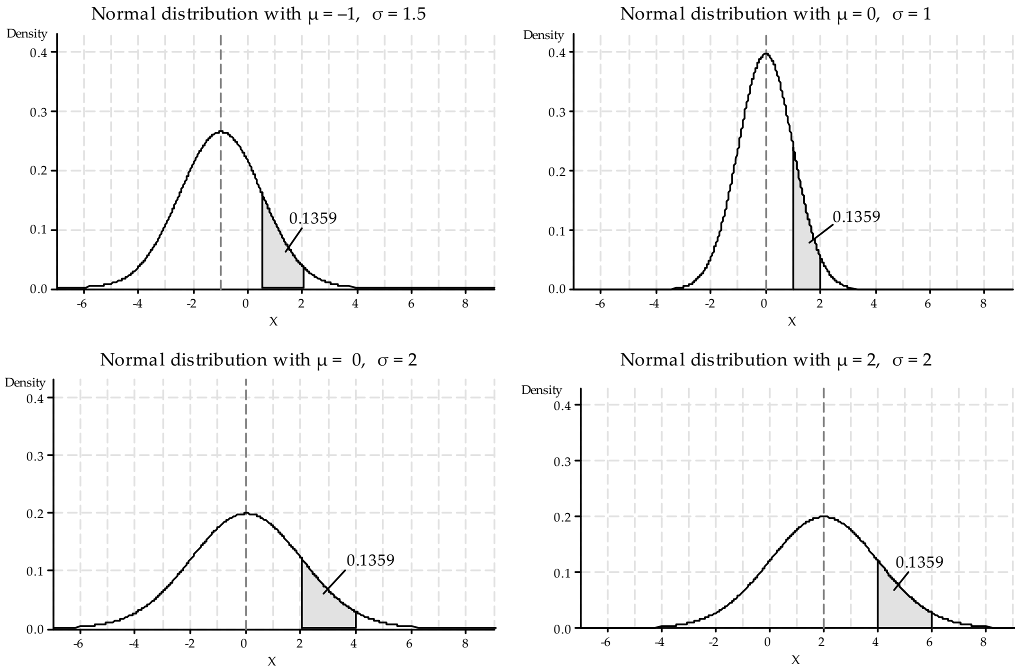

Figure 4: Four Normal probability density functions.

The cdf of any Normal distribution can also be found, using technology, without first standardising. If $X \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma^2)$, then the cdf of X is given by

$$\Pr(X \le x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\frac{-(t-\mu)^2}{2\sigma^2}\Big)\, dt, \qquad \text{for } x \in \mathbb{R}.$$

One way to obtain this is in Excel using the function NORM.DIST. This function requires four arguments:

1. the value of x for which the cdf should be evaluated
2. the mean μ
3. the standard deviation σ
4. a true/false (or equivalently, 1/0) argument that controls whether the cdf (argument equals TRUE or 1) or pdf (argument equals FALSE or 0) is returned.

We can use this function to find the required probabilities in the crowd-size example directly. For example, you should find that typing

```
=NORM.DIST(36000, 26000, 5000, 1)
```

returns the value 0.9772.

Sometimes we need to find a quantile of the Normal distribution. Let q be a number between 0 and 1. Then the q **quantile**, c_q, of the Normal distribution with cdf F_X is defined by the equation

$$F_X(c_q) = q.$$

To obtain the value of c_q, we can use technology. In Excel, for example, the function is NORM.INV. It requires three arguments:

1. the value of q for which the inverse cdf should be evaluated
2. the mean μ
3. the standard deviation σ.

Exercise 3

Suppose that the difference between the forecast maximum temperature and the actual maximum temperature (in degrees Celsius) in a city is Normally distributed with mean 0 and standard deviation 1.2.

a. Find the probability that the actual maximum is within 1.0 degrees of the forecast maximum.
b. Which is more likely: an underestimate of 0.5 degrees or more, or a forecast within 0.5 degrees of the actual maximum?
c. A reporter is writing up this information for an article about weather forecasts, and wants a sensationalist angle, so she asks: 'How bad can it get? Let's say, on the low side, the most extreme 1% of differences are in what range? And what about the worst 1% on the high side?'

Exercise 4

Animals of a given weight are operated on in a veterinary hospital. The dose of anaesthetic A (in mg) required to render the animals suitably unconscious for the operation is Normally distributed with mean 120 and standard deviation 20. The lethal dose L (in mg) of the same anaesthetic for these animals is also Normally distributed, with mean 400 and standard deviation 50.

a. Sketch the pdfs of the random variables A and L on the same axes.
b. Find the dose d^* that the vet should administer, in order that 99.9% of animals will be suitably unconscious for the operation.
c. If d^* mg of anaesthetic is administered, what percentage of animals die?

We shall study confidence intervals in the two modules *Inference for proportions* and *Inference for means*. In that context, we want to know the bounds of the central 95% of the distribution for $Z \stackrel{d}{=} \mathrm{N}(0,1)$. That is, we want z such that

$$\Pr(-z < Z < z) = 0.95.$$

We can find this z using the same techniques as for quantiles.

Since the standard Normal distribution is symmetric about 0, we require

$$\Pr(Z \le -z) = \tfrac{1}{2}(1-0.95) = 0.025 \quad \text{and} \quad \Pr(Z \ge z) = \tfrac{1}{2}(1-0.95) = 0.025.$$

So we want

$$F_Z(z) = \Pr(Z \le z) = 1 - 0.025 = 0.975.$$

We can now find z in Excel using

=NORM.INV(0.975, 0, 1),

which gives 1.96. This is illustrated in the following figure.

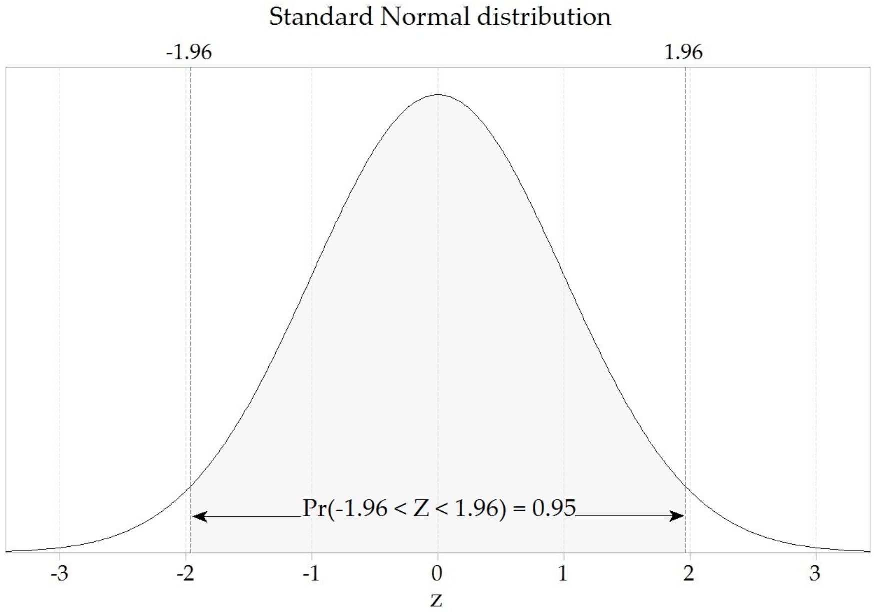

Figure 5: The standard Normal distribution, $Z \stackrel{d}{=} \mathrm{N}(0,1)$.

More generally, if we are given a probability p and we want z with $\Pr(-z < Z < z) = p$, then we find z such that

$$F_Z(z) = \frac{p+1}{2}.$$

Answers to exercises

Exercise 1

Let T be the waiting time (in days) until the first call. Then $T \stackrel{d}{=} \exp(1.8)$, and therefore $F_T(t) = 1 - \exp(-1.8t)$, for $t > 0$. We need to be careful about the units of time used here.

a 15 minutes equals $\frac{15}{24\times 60}$ days, or 0.01042 days. Hence, the chance of a call in the first 15 minutes equals $F_T(0.01042) = 1 - \exp(-1.8 \times 0.01042) = 0.0186$.

b Due to the lack of memory property, the probability is the same as that in part a, namely 0.0186.

c 10 hours equals $\frac{10}{24}$ days, or 0.41667 days. So the probability of no calls during a shift is $\Pr(T > 0.41667) = \exp(-1.8 \times 0.41667) = 0.4724$.

d Assuming independence between days, the probability of no calls in four successive days equals $0.4724^4 = 0.0498$.

e Solving for the time y in days:

$$\begin{aligned} & \Pr(T \le y) = 0.1 \\ \Longrightarrow \quad & F_T(y) = 0.1 \\ \Longrightarrow \quad & 1 - \exp(-1.8y) = 0.1 \\ \Longrightarrow \quad & -1.8y = \ln(0.9) \\ \Longrightarrow \quad & y = 0.0585 \text{ days.} \end{aligned}$$

Hence, x is 1.4 hours, which is 1 hour 24 minutes.

Exercise 2

a $f_X(\mu) = \dfrac{1}{\sigma\sqrt{2\pi}} \approx \dfrac{0.40}{\sigma}$.

b We have $f_X(x) = f_X(\mu)\exp(k(x))$, where $k(x) \le 0$ for all values of x. So $\exp(k(x)) \le 1$, and the result follows.

c To find points of inflexion, we need the second derivative of $f_X(x)$. Using the chain rule, we have

$$\begin{aligned} f_X'(x) &= \frac{d}{dx}\Big[\frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big)\Big] \\ &= \frac{1}{\sigma\sqrt{2\pi}}\Big(\frac{-(x-\mu)}{\sigma^2}\Big) \exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big). \end{aligned}$$

Now, using the product rule, we have

$$f_X''(x) = \frac{1}{\sigma\sqrt{2\pi}}\Big[\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\Big] \exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big).$$

At a point of inflexion, $f_X''(x) = 0$. This gives

$$\begin{aligned} f_X''(x) &= 0 \\ \Longrightarrow \quad \frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} &= 0 \\ \Longrightarrow \quad (x-\mu)^2 &= \sigma^2 \\ \Longrightarrow \quad x &= \mu \pm \sigma. \end{aligned}$$

Hence, the points of inflexion are at $x = \mu - \sigma$ and $x = \mu + \sigma$; these are the points either side of μ at which the curve changes from convex to concave.

d $f_X(\mu + k\sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}k^2)$.

k	0	1	2	3	4	5
$\exp(-\frac{1}{2}k^2)$	1	0.607	0.135	0.011	0.0003	0.000004

Note that $f_X(\mu - k\sigma) = f_X(\mu + k\sigma)$, by symmetry.

There are a couple of useful interpretations:

- Firstly, when sketching a Normal pdf, the height of the curve at $\mu \pm \sigma$ is 61% of the height of the central peak, and so on.
- Second, recall that the height of a pdf reflects *relative* probabilities, so that if $f_X(b) = 2 f_X(a)$, then the chance of an observation near b is approximately twice as likely as an observation near a. This means, for example, that observations near μ are approximately 250 000 times more likely that observations near $\mu + 5\sigma$, since $\frac{1}{0.000004} = 250\ 0000$.

Exercise 3

Let D be the difference between the forecast maximum temperature and the actual maximum temperature (in degrees Celsius). Then $D \stackrel{\mathrm{d}}{=} \mathrm{N}(0, 1.2^2)$.

a $\Pr(-1.0 < D < 1.0) = 0.595$.

b $\Pr(D < -0.5) = 0.338$ and $\Pr(-0.5 < D < 0.5) = 0.323$, so it is very slightly more probable that there is an underestimate of 0.5 degrees or more.

c We want to find the 0.01 quantile of the distribution; that is, we want $c_{0.01}$ satisfying $F_D(c_{0.01}) = 0.01$. We find that $c_{0.01} = -2.79$ degrees. So 1% of forecast maximums are 2.79 degrees or more lower than the actual maximum. By symmetry, 1% of forecast maximums are 2.79 degrees or more higher than the actual maximum.

Exercise 4

a The pdfs of the random variables A and L are shown on the same axes in figure 6. The green distribution, on the left, is for the dose of anaesthetic required to render the animal unconscious. The average dose is 120 mg, and most values are in the range from about 60 mg to 180 mg. The red distribution, on the right, is for the lethal dose. The mean is 400 mg — much higher than the mean of the green distribution. There is little overlap of the two distributions. (Which is how we want things to be!) In fact, you might think that they do not overlap at all, based on a visual assessment of figure 6.

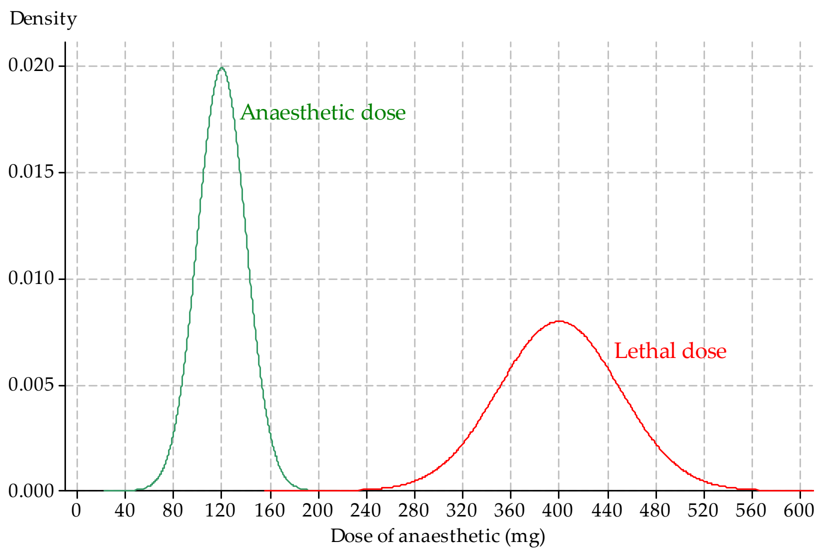

Figure 6: The pdfs of anaesthetic and lethal doses.

b This question is about the anaesthetic dose administered, so we need to consider the distribution that renders animals suitably unconscious (the green distribution in figure 6). We need to find the value d^* that corresponds to 99.9% of the animals being rendered suitably unconscious; this means a cumulative probability of 0.999. This is shown in figure 7.

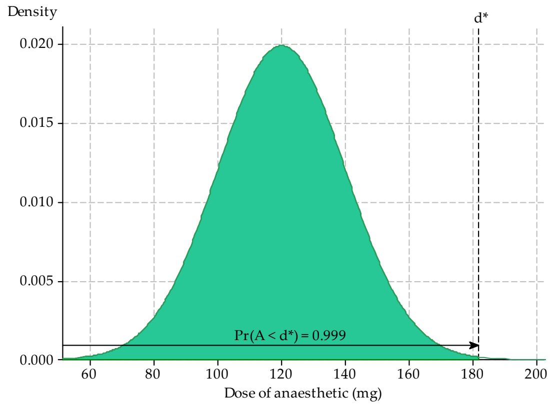

Figure 7: The pdf of anaesthetic dose, showing d^*.

We want to find d^* such that $\Pr(A \leq d^*) = 0.999$; this is the 0.999 quantile of the distribution. This can be achieved in Excel using the function NORM.INV. If you enter =NORM.INV(0.999, 120, 20) in a cell, you should find that the dose required to render 99.9% of animals unconscious is $d^* = 181.80$ mg.

c We now consider the distribution of the lethal dose, and what happens if a dose of 181.80 mg is administered. If a dose of 181.80 mg is used, there will be a small proportion of animals for whom this dose is lethal: those for whom the lethal dose is less than or equal to 181.80 mg. We are considering the pdf of L (the red distribution in figure 6), and need to find the area under the curve corresponding to a dose of 181.80 or less. This is shown on the left in the following figure.

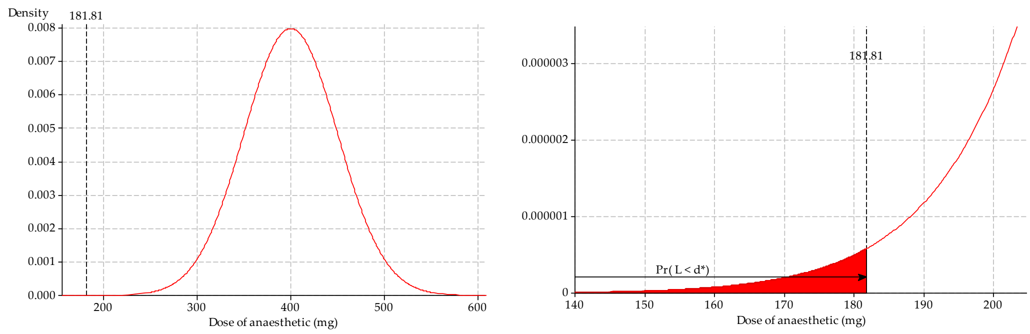

Figure 8: The pdf of lethal dose, showing d^*.

The tail area is extremely small and hard to see, so the graph on the right is zoomed in to show the detail of the pdf of L near d^*.

To find the left-tail area in this distribution, we can use the cdf function in Excel; we enter =NORM.DIST(181.80, 400, 50, 1) in a cell. The value returned, and hence the proportion dying, is 0.0000064, or 0.00064%. This corresponds to 6 in a million, which is very small, as we suspect from the diagrams: a good result.

0 1 2 3 4 5 6 7 8 9 10 11 12