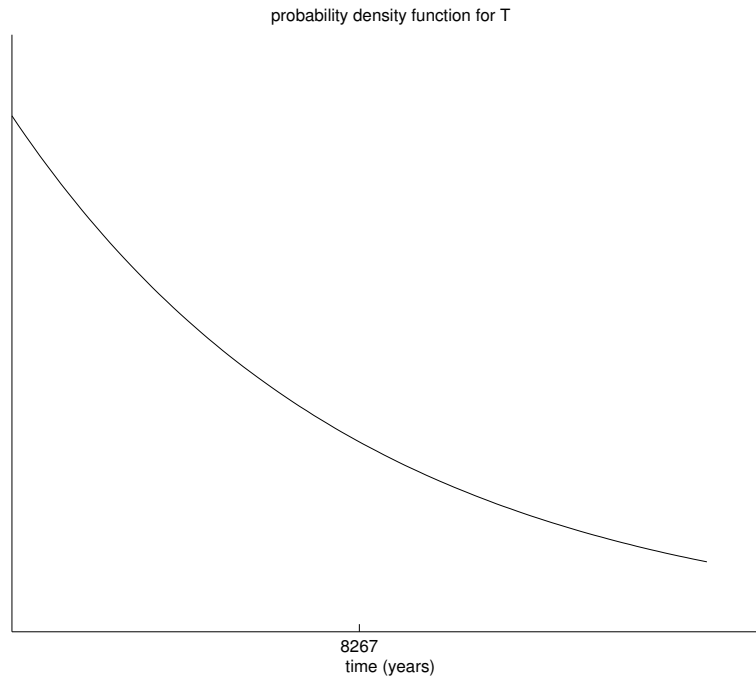


# 1 Exponential Distribution

We have seen two types of continuous random variables thus far, the uniform and the normal. The exponential distribution is a continuous distribution that is often used to model times like lifetimes, time until failure, time until decay, or time between events. Time is a continuous quantity because it can occur any amount - like 4 seconds or 3.999 seconds or 3.99999 seconds.

As an example of an exponential random variable, we consider radioactive decay. Carbon-14 atoms decay to Nitrogen-14 with mean time 8267 years. Let  $T$  denote the time, in years, that a C-14 atom takes to decay to N-14. Below is the graph of the probability density function for  $T$ .



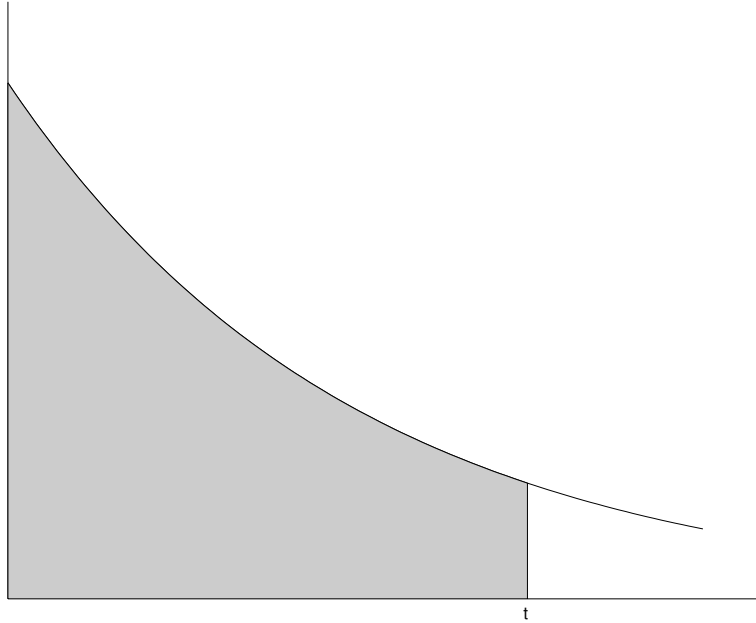
The key point to label here is the mean time. The mean is located where the density function falls to roughly 1/3 its initial value. We do not care about what exactly the initial value is. Since times are always positive, the possible values of  $T$  start at 0.

As with the other continuous random variables, the probability that the random variable is between two values  $a$  and  $b$  is the area under the graph of the probability density function between  $a$  and  $b$ . For uniform random variables, regions under the pdf are always rectangles, so their areas are easily computed by multiplying their length by their width. For normal random variables, we first need to standardize the distribution and then we can look up areas in a table. We could only look up areas of regions that are bounded on one side by 0. For exponential random variables, areas are given by a formula. We can only use this formula to find areas of regions that are bounded on the left by 0.

In this particular example, the area of the region below represents  $P(T < t)$  and is given by the formula

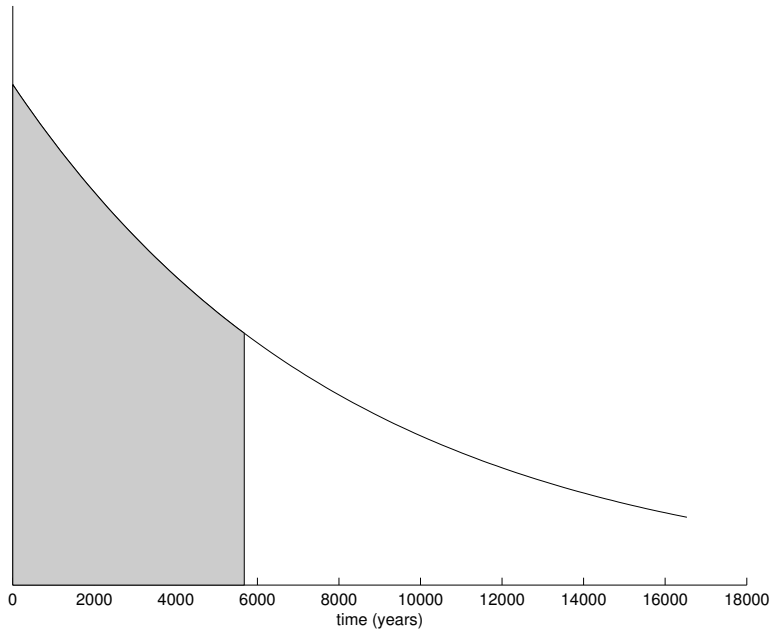
$$P(T < t) = 1 - e^{-t/8267},$$

where  $e \approx 2.718$  is a special constant. Notice that the mean 8267 appears in the exponent.



Some calculators have a button that raises  $e$  to whatever power you want, and saves you the effort of typing in 2.718 every time you want to do a computation. This button may be labelled `exp` or  $e^x$  or  $e$ .

As an example, find the probability that a C-14 atom decays to N-14 within 5730 years. That is, we want to find  $P(T < 5730)$ . This probability is precisely the area of the shaded region below.

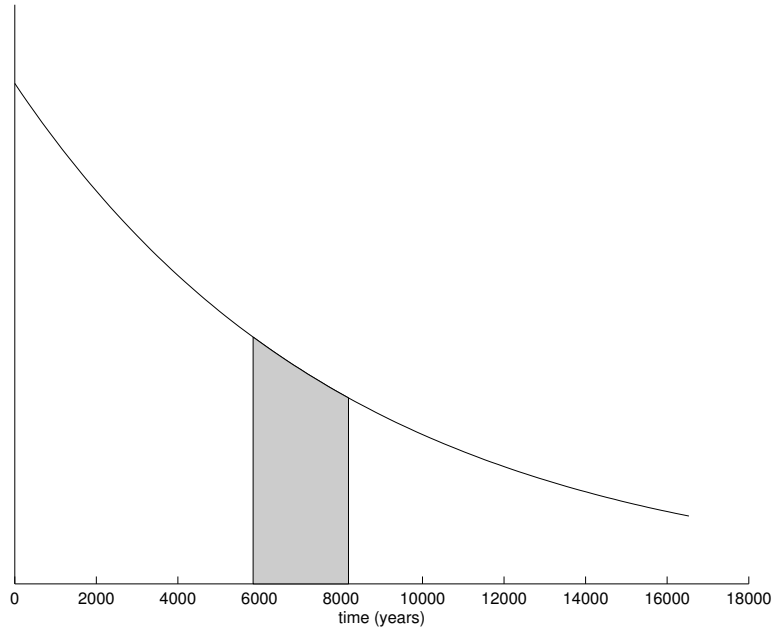


To find this, we use our formula.

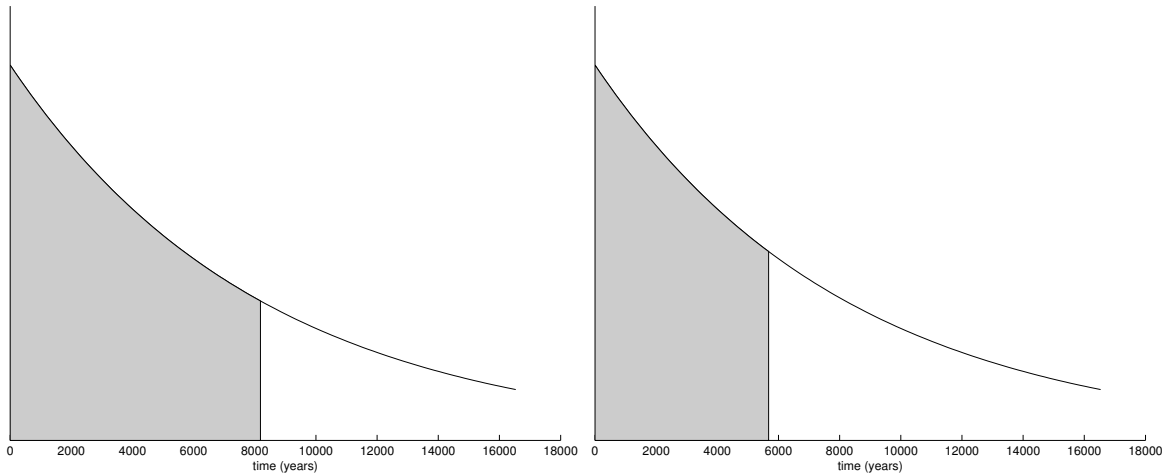
$$\begin{aligned}
 P(T < 5730) &= 1 - e^{-\frac{5730}{8267}} \\
 &= 1 - 2.718^{-0.693} \\
 &= 1 - 0.5 \\
 &= 0.5
 \end{aligned}$$

That is, the probability that the decay happens within 5730 years is  $1/2$ . It is for this reason that we call 5730 years the half-life of C-14. Notice that the half-life is not the same as the mean decay time.

As another example, find the probability that an atom of C-14 survives for at least 5730 years but decays within 8263 years. That is, we want to find  $P(5730 < T < 8267)$ . This probability is precisely the area of the shaded region below.



We can not directly apply our formula because this region is not bounded on the left by 0. We have handled issues similar to this when working with the normal distribution. The trick is to observe that the area is the difference of the two areas below:



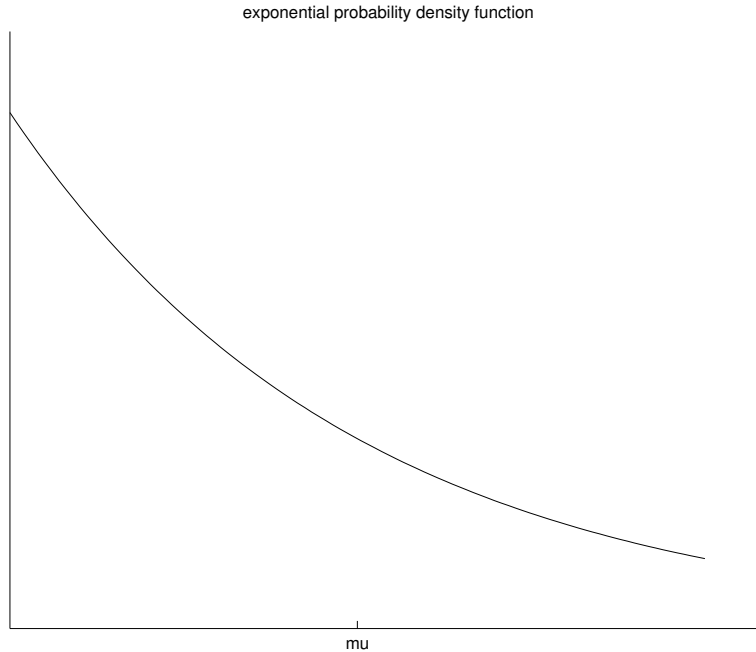
We have already found the area of the shaded region on the right to be  $1/2$ . To find the area of the shaded region on the left, we use our formula.

$$\begin{aligned}
 P(T < 8267) &= 1 - e^{-\frac{8267}{8267}} \\
 &= 1 - 2.718^{-1} \\
 &= 1 - 0.3679 \\
 &= 0.632.
 \end{aligned}$$

Now we argue that

$$\begin{aligned} P(5730 < T < 8267) &= P(T < 8267) - P(T < 5730) \\ &= 0.632 - 0.5 \\ &= 0.132. \end{aligned}$$

In general, the graph of the pdf for an exponentially distributed random variable with mean  $\mu$  looks like this:



Again, we label the mean  $\mu$  at a point where the pdf has decreased to approximately 1/3 its original value.

To compute probabilities for an exponentially distributed random variable, we use the fact that the area under the graph of the pdf between 0 and a point  $t$  is

$$1 - e^{-t/\mu},$$

where  $e \approx 2.718$ . The standard deviation of an exponentially distributed random variable is the same as the mean. That is,

$$\sigma = \mu.$$

## 2 Sampling

We now begin our study of inferential statistics. Suppose we have a huge population (people in the world, C-14 atoms in the universe, vehicles on the road, employees of a company). We want to obtain some information about the population (heights of people, times to decay of atoms, colours of vehicles, salaries of employees). It is unrealistic to collect data from the entire population. Instead, we just collect data from a small sample. Using our knowledge of probability, we use the sample data to make inferences about the population data.

We study four methods of obtaining samples: simple random, systematic random, clustered, and stratified.

In a simple random sample, all individuals in the population have the same independent chance of being selected for the sample. A reliable way to obtain a simple random sample is to number all the individuals in the population and then use a computer to choose numbers randomly. Humans are not good at choosing random numbers. The simple random sample is the method which generates a truly random sample. Below is an illustration of a simple random sample.

[coming soon]

The large box represents the population and the dots represent the sampled individuals. The sampled individuals are completely random.

Simple random samples can allow or forbid replacement. By replacement, we mean that, once an individual has been selected for the sample, they are placed back in the pool so that they may possibly be chosen again. When replacement is forbidden, an individual may be selected at most one time for the sample.

Allowing replacement is more convenient for several reasons. First, drawing each individual for the sample is independent, because it does not depend on the results of previous draws. We have observed, in our study of probability, that independent events are easier to handle than dependent events. Second, when we have a computer generate random numbers, it is easiest to allow the computer to pick the same number more than once. Forbidding this requires extra programming.

If the population size is substantially larger than the sample size, allowing replacement is essentially indistinguishable from forbidding replacement, because the chance that the same individual is chosen twice is very small.

In a systematic random sample, you randomly select the first element and then select subsequent elements in a systematic way. That is, the randomness only happens once, at the beginning, and then other individuals are chosen according to some pattern. The advantage of this is that systematic selection is easy for humans to do. Below is an illustration of a systematic random sample.

[coming soon]

The sampled individuals follow a pattern.

There is some danger in using the systematic random sample. Suppose that a rotating piece of equipment in a factory is malfunctioning in a way that it causes every fifth product to be defective. If you conduct a systematic random sample where you are sampling every fifth product, your sample will be a bad representation. There are two possibilities. First, your selection could line up exactly with the malfunctioning equipment, and then you would only see defective products. If your selection and the malfunctioning do not line up exactly, you will never see any of the defective products. In both of these cases, you will draw inaccurate conclusions about the true fraction (1 in 5) of products that are defective.

In a clustered sample, the population is split up into groups. You sample only a few of the groups, but you sample them in their entirety. This method is advantageous if physically nearby individuals are grouped together and collecting data requires travelling to the locations of the individuals. Below is an illustration of a clustered random sample.

[coming soon]

The large box represents the population and the smaller boxes contained inside it represent the groups. Some groups are entirely shaded in, indicating that all the individuals in those groups are in the sample.

In a stratified sample, the population is split up into groups. You then obtain a simple random or systematic random sample from each of the groups. Your sample is then formed by combining all the samples from all the groups. This ensures that all groups get representation in the sample. Below is an illustration of a systematic random sample.

[coming soon]