

## 1 Continuous random variables

We distinguish between two types of random variables: discrete and continuous.

**Definition 1.1** (discrete random variable). A random variable  $X$  is discrete if its possible values can be listed  $x_1, x_2, x_3, \dots$ .

The binomial model is an example of a discrete random variable. Obviously, we could always get by just using discrete random variables as all measurement scales are ultimately discrete. It often is more useful to model a measured quantity as if all real number measurements are possible. We have already used this device when we modeled finite datasets using the normal model.

**Definition 1.2** (continuous random variable). A random variable  $X$  is continuous if its possible values are all  $x$  in some interval of real numbers.

A discrete random variable is determined by its probability mass function which is the function  $f(x) = P(X = x)$ . To model a continuous random variable, we again make use of density functions. The important thing to realize is that we are not now interested in  $P(X = x)$  but rather in events such as  $P(a \leq X \leq b)$ . For example, if  $X$  is the height of a randomly chosen Calvin student measured in inches, then the event  $71.5 \leq X \leq 72.5$  is interesting (the student is six feet tall) but the event  $X = 72$  is not. The analogue to a probability mass function for a continuous variable is a probability density function. We have already met density functions before as (continuous) models of discrete data distributions. Now we use them as models for the distribution of probability of a random variable.

**Definition 1.3** (density function). A density function is a function  $f$  such that

- $f(x) \geq 0$  for all real numbers  $x$ , and
- $\int_{-\infty}^{\infty} f(x) dx = 1$ .

Just as we used an integral of the density function to model the proportion of data in a given interval, we use integrals to compute the probability that a random variable has its value in a given interval.

**Definition 1.4** (probability density function). The density function  $f$  is a probability density function (pdf) for the random variable  $X$  if for all real numbers  $a \leq b$ ,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

The following simple lemma demonstrates one way in which continuous random variables are very different from discrete random variables.

**Lemma 1.5.** Let  $X$  be a continuous random variable with pdf  $f$ . Then for any real number  $a$ ,

1.  $P(X = a) = 0$ ,
2.  $P(X < a) = P(X \leq a)$ , and
3.  $P(X > a) = P(X \geq a)$ .

*Proof.*  $\int_a^a f(x) dx = 0$ . And  $P(X \leq a) = P(X < a) + P(X = a) = P(X < a)$ . □

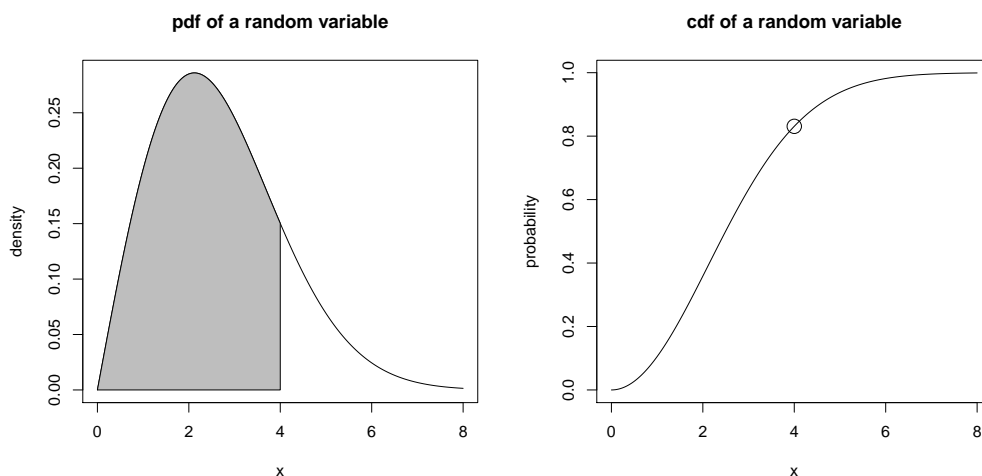
It is important to note that even though  $P(X = a) = 0$ , this does not make the event  $X = a$  impossible. After all, something has to happen. What is true however is that the limiting relative frequency of occurrence of the event  $X = a$  is 0.

**Definition 1.6** (Cumulative distribution function). The cumulative distribution function (cdf) of the random variable  $X$  is the function  $F$  defined by  $F(x) = P(X \leq x)$ .

To find the cdf of a discrete random variable we add. To find the cdf of a continuous random variable we integrate. If  $X$  is a continuous random variable with pdf  $f$ , then the cumulative distribution function (cdf) for  $X$  is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt .$$

R has a function to compute the cdf for each of the standard families of random variables. The pdf and cdf of a typical random variable are illustrated below with the event  $X \leq 4$  illustrated appropriately on each graph.



Notice that the cdf  $F$  of the random variable  $X$  is an antiderivative of the pdf  $f$  of  $X$ . This follows immediately from the Fundamental Theorem of Calculus. Notice also that  $P(a \leq X \leq b) = F(b) - F(a)$ .

**Lemma 1.7.** Let  $F$  be the cdf of a continuous random variable  $X$ . Then the pdf  $f$  satisfies

$$f(x) = \frac{d}{dx} F(x) . \quad \square$$

Just as there are several common discrete probability models, there are also important (families of) continuous probability models. Our favorite density function is that for the normal model. We will use the normal model to model random variables as well.

**Definition 1.8** (normal model). The normal model has two parameters  $\mu$  and  $\sigma >$  and pdf given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty .$$

We write  $\text{Norm}(\mu, \sigma)$  for the normal model.

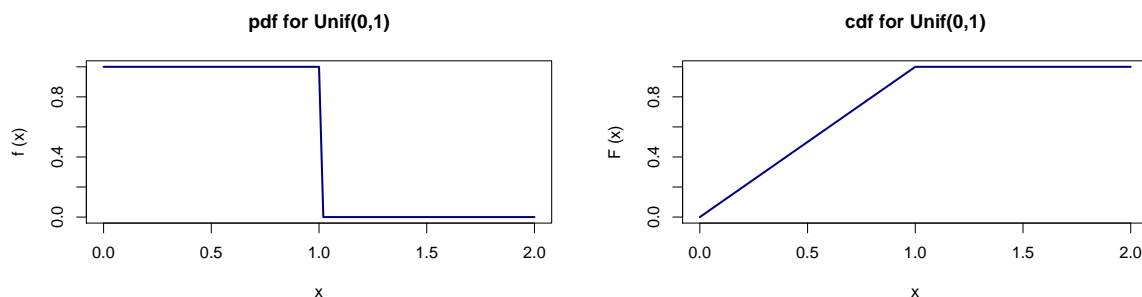
A much simpler model (but still important) is the uniform model.

**Definition 1.9** (uniform random variable). The uniform model has two parameters  $(a, b)$  and pdf given by

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

We write  $\text{Unif}(a, b)$  for the uniform model with parameters  $a, b$ .

It is easy to confirm that this function is indeed a pdf. We could integrate, but of course we should simply use geometry. The region under the graph of the uniform pdf is a rectangle with width  $b - a$  and height  $\frac{1}{b-a}$ , so the area is 1. Graphs of the pdf and cdf of a uniform random variable (with parameters  $a = 0$  and  $b = 1$ ) are given below.



There are R functions for computing the pdf and cdf of a uniform random variable as well as a function to return random numbers. An additional function computes the quantiles of the uniform distribution. If  $X \sim \text{Unif}(\text{min}, \text{max})$  the following functions can be used.

<u>function (&amp; parameters)</u>	<u>explanation</u>
<code>runif(n,min,max)</code>	makes <code>n</code> random draws of the random variable $X$ and returns them in a vector.
<code>dunif(x,min,max)</code>	returns $f(x)$ , (the pdf).
<code>punif(q,min,max)</code>	returns $P(X \leq q)$ (the cdf).
<code>qunif(p,min,max)</code>	returns $x$ such that $P(X \leq x) = p$ .

Here are examples of computations for  $\text{Unif}(0, 10)$ .

```
> runif(6,0,10)    # 6 random values on [0,10]
[1] 5.449745 4.124461 3.029500 5.384229 7.771744 8.571396
```

```

> dunif(5,0,10)    # pdf is 1/10
[1] 0.1
> punif(5,0,10)    # half the distribution is below 5
[1] 0.5
> qunif(.25,0,10)  # 1/4 of the distribution is below 2.5
[1] 2.5

```

Another important family of continuous random variables are the exponential random variables. These often provide a model for a waiting time. Waiting times are important random variables in reliability studies. For example, a common characteristic of a manufactured object is MTF or mean time to failure. Note that a waiting time can be any  $x$  in the range  $0 \leq x < \infty$ .

**Definition 1.10** (The exponential model). The exponential model has one parameter  $\lambda > 0$  (called the rate) and pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

We write  $\text{Exp}(\lambda)$  for the exponential model with parameter  $\lambda$ .

It is easy to see that the function  $f$  of the previous definition is a pdf for any positive value of  $\lambda$ . R refers to the value of  $\lambda$  as the **rate** so the appropriate functions in R are **rexp(n,rate)**, **dexp(x,rate)**, **pexp(q,rate)**, and **qexp(p,rate)**. We will see later that **rate** is an apt name for  $\lambda$  as  $\lambda$  will be the rate per unit time if  $X$  is a waiting time random variable.

**Example 1.11.**

Suppose that a random variable  $T$  measures the time until the next radioactive event is recorded at a Geiger counter (time measured since the last event). For a particular radioactive material, a plausible models for  $T$  is  $T \sim \text{Exp}(0.1)$  where time is measured in seconds. Then the following R session computes some important values related to  $T$ .

```

> pexp(q=0.1,rate=.1)  # probability waiting time less than .1
[1] 0.009950166
> pexp(q=1,rate=.1)    # probability waiting time less than 1
[1] 0.09516258
> pexp(q=10,rate=.1)
[1] 0.6321206
> pexp(q=20,rate=.1)
[1] 0.8646647
> pexp(100,rate=.1)
[1] 0.9999546
> pexp(30,rate=.1)-pexp(5,rate=.1)  # probability waiting time between 5 and 30
[1] 0.5567436
> qexp(p=.5,rate=.1)   # probability is .5 that T is less than 6.93
[1] 6.931472

```

The graphs in Figure 1 are graphs of the pdf and cdf of this random variable. All exponential distributions look the same except for the scale. The rate of 0.1 here means that we can expect that in the long run this process will average 0.1 counts per second.

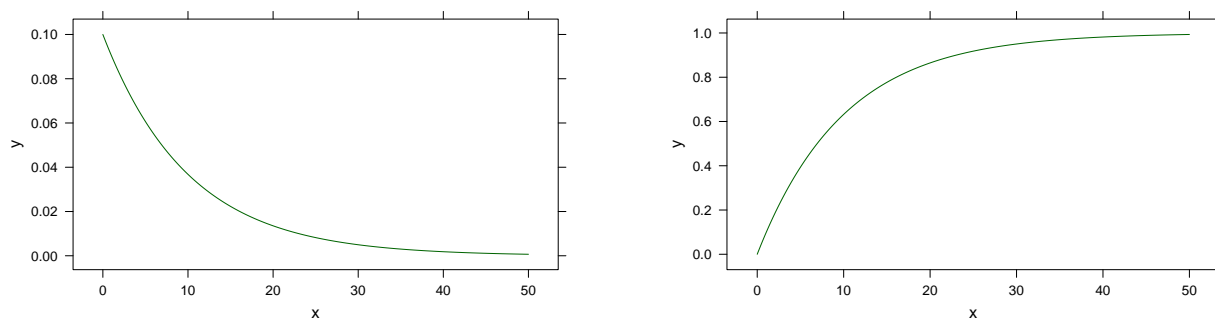


Figure 1: The pdf and cdf of the random variable  $T \sim \text{Exp}(0.1)$ .

Of course when given a random variable such as the waiting time to a geiger counter event, we are not handed its pdf as well. The pdf is a model of the situation. In the case of an example such as this, we really are faced with two decisions.

1. Which family (e.g., uniform, exponential, etc.) of models best fits the situation?
2. What particular values of the parameters should we use for the pdf?

Sometimes we can begin to answer question 1 even before we collect data. Each of the distributions that we have met has certain properties which we check against our process. For example, it is often apparent whether the properties of a binomial process should apply to a certain process we are examining. Of course it is always useful to check our answer to question 1 by collecting data and verifying that the shape of the distribution of the data collected is consistent with the distribution we are using. The only reasonable way to answer the second question however is to collect data. In the last example, for instance, we saw that if  $X \sim \text{Exp}(0.1)$  that  $P(X \leq 6.93) = .5$ . Therefore if about half of our data are less than 6.93, we would say that the data are consistent with the hypothesis that  $X \sim \text{Exp}(0.1)$  but if almost all the data are less than 5, we would probably doubt that  $X$  has this distribution.

**Definition 1.12** (mean and variance). Let  $X$  be a continuous random variable with pdf  $f$ . The mean of  $X$  is defined by

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx .$$

The variance  $\sigma^2$  of  $X$  is defined by

$$\sigma^2 = \text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx .$$

We compute the mean of two of our favorite continuous random variables in the next Theorem.

**Theorem 1.13.**

1. If  $X \sim \text{Unif}(a, b)$  the  $E(X) = \frac{a+b}{2}$ .
2. If  $X \sim \text{Exp}(\lambda)$  is  $E(X) = 1/\lambda$ .

Our intuition tells us that in a large sequence of trials of the random process described by  $X$ , the sample mean of the observations should be usually be close the mean of  $X$ . This is in fact true and is known as the Law of Large Numbers. We will not state that law precisely here but we will illustrate it using several simulations in R.

```
> r=rexp(100000,rate=1)
> mean(r)                # should be 1
[1] 0.9959467
> r=runif(100000,min=0,max=10)
> mean(r)
[1] 5.003549             # should be 5
> r=rbinom(100000,size=100,p=.1)
> mean(r)
[1] 9.99755             # should be 10
```

The following lemma records the variance of several of our favorite random variables.

**Lemma 1.14.** 1. If  $X \sim \text{Unif}(a, b)$  then  $\text{Var}(X) = (b - a)^2/12$ .

2. If  $X \sim \text{Exp}(\lambda)$  then  $\text{Var}(X) = 1/\lambda^2$ .

## Problems

**1.1** A random variable  $X$  has the following pdf:

$$f(x) = \begin{cases} 2x & x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

1. Show that  $f$  is indeed a pdf. (Hint: Draw a picture.)
2. Compute  $P(0 \leq X \leq 1/2)$ . (Hint: Draw a picture.)
3. Find the number  $m$  such that  $P(0 \leq X \leq m) = 1/2$ . It is natural to call  $m$  the median of the random variable. (Hint: Draw a picture.)
4. Compute the mean of this random variable.

**1.2** The file <http://www.calvin.edu/~stob/data/scores.csv> contains a dataset that records the time in seconds between scores in a basketball game played between Kalamazoo College and Calvin College on February 7, 2003. The dataset is available in the class package and also by `get243()`.

1. This waiting time data might be modeled by an exponential distribution. Make some sort of graphical representation of the data and use it to explain why the exponential distribution might be a good candidate for this data.
2. If we use the exponential distribution to model this data, which  $\lambda$  should we use? (A good choice would be to make the sample mean equal to the expected value of the random variable.)
3. Your model of part (b) makes a prediction about the proportion of times that the next score will be within 10, 20, 30 and 40 seconds of the previous score. Test that prediction against what actually happened in this game.