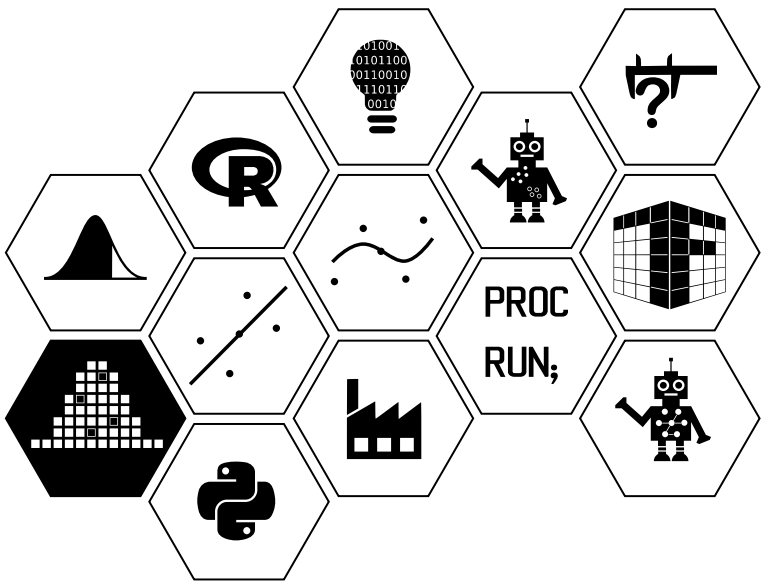
Probability and Sampling Fundamentals

Week 5: Continuous random variables I



DATA ANALYTICS
GLASGOW

Week 5 learning material aims

The material in week 5 covers:

- the definition of continuous random variables;
- how to calculated probabilities associated with continuous random variables;
- the definition of probability density functions;
- related probability density functions to cumulative distribution functions;
- the calculation of expectation and variance of continuous random variables;
- the median, percentiles and quantiles

Continuous random variables

Over the last two weeks we have introduced discrete random variables. Recall the definition of a discrete random variable. A random variable that has a finite or countable range space is a discrete random variable. This week we will consider random variables with uncountable range spaces. In particular, random variables with uncountable range spaces are continuous random variables. The range spaces of continuous random variables are typically the entire real line, the positive half of it, or intervals on the real line. We will begin with a few examples.

Example 1

Time to run a mile

Suppose I asked everyone in this class how long it would take them to run a mile. We can write down a random variable to describe this experiment.

The outcome of this experiment is the time it takes to run a mile and has to be a positive real number

$$S = \mathbb{R}_{>0}.$$

We can define a random variable X such that $X(s) = s$ for all $s \in S$. Here X corresponds to the time taken to run a mile. Notice the range space is $\mathbb{R}_{>0}$. This means that X can take **any** positive real number. For instance, the first 10 observations may look like

R Console

```
   Student Time Taken
1    student 1    8.16583
2    student 2   9.813256
3    student 3   9.297321
4    student 4   8.572493
5    student 5    7.18624
6    student 6   7.193078
7    student 7   6.122763
8    student 8   6.927451
9    student 9   9.134974
10 student 10   7.161841
```

Supposing we have observations from 500 students and plot the data

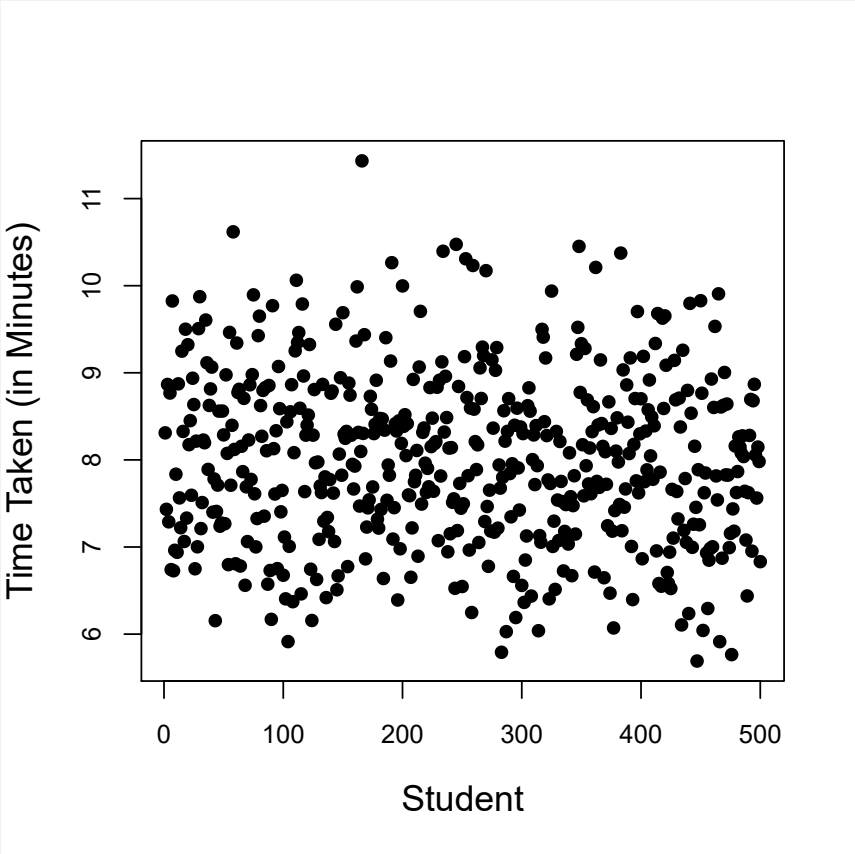


Figure 1

It appears on average, students take around 8 minutes to run a mile (roughly shown in the centre of the y-axis) with some variation around this value. It is unlikely that a student would run a sub-6 minute mile or take more that 10 minutes. Notice, as with discrete random variables, we can summarise this distribution with a mean value and spread.

Video

Probability and Sampling Fundamentals - Week 5

Duration 4:45

Example 2

Randomly waiting on a train

Suppose a very reliable train line runs exactly every 20 mins between 6am and 6pm. A man 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 man expect to wait on his train?

The man could be very lucky with the train arriving immediately after he arrives such that his waiting time is 0 mins or the man could be very unlucky and just miss a train with a waiting time of 20 mins. The problem is that he has no way of knowing.

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

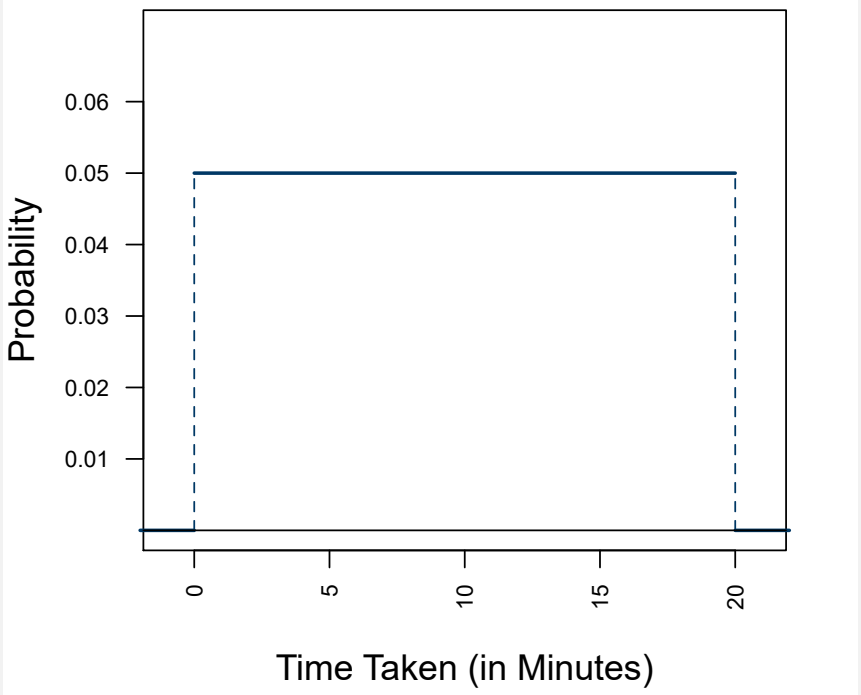


Figure 2

Considering this example in more detail, we know that the waiting time needs to lie between 0 and 20. We can define a random variable X such that $X =$waiting time of the person with range space $[0, 20]$. Therefore, we know that $P(0 \leq X \leq 20) = 1$. This constraint determines the height of the box in the above plot.

In order to see that $P(0 \leq X \leq 20) = 1$ we need to consider the area under the curve. We can see that the height of the line when the waiting time equals 0 or 20 is 0.05 (i.e. $\frac{1}{20}$). We just need to work out the area of a rectangle of height 0.05 and length 20

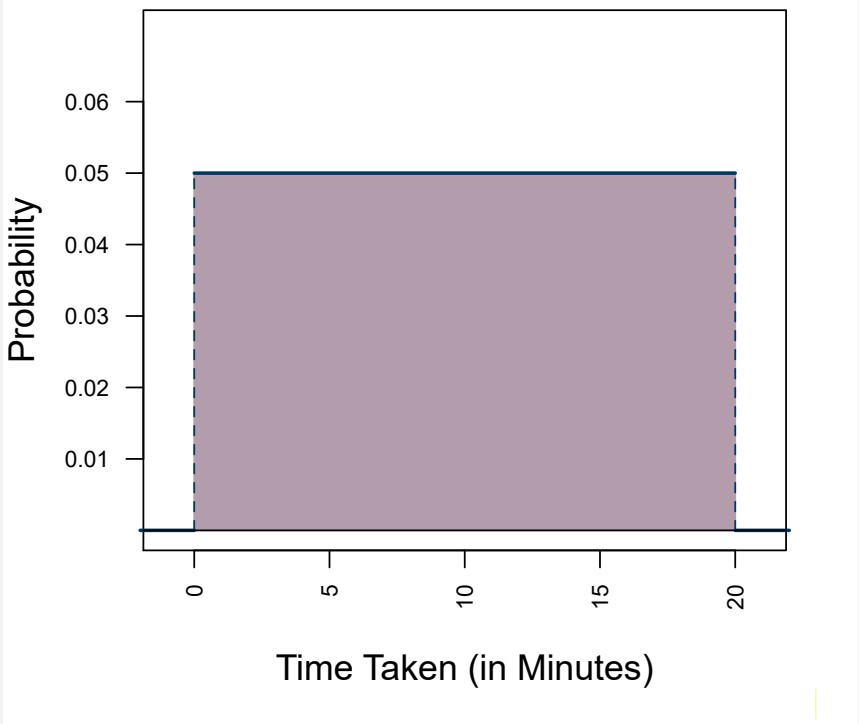


Figure 3

$$\begin{aligned}\text{Area under the curve} &= \text{length} \times \text{height} \\ &= 20 \times 0.05 \\ &= 1 \\ &= P(0 \leq X \leq 20)\end{aligned}$$

Using the same logic, what is $P(5 \leq x \leq 10)$?

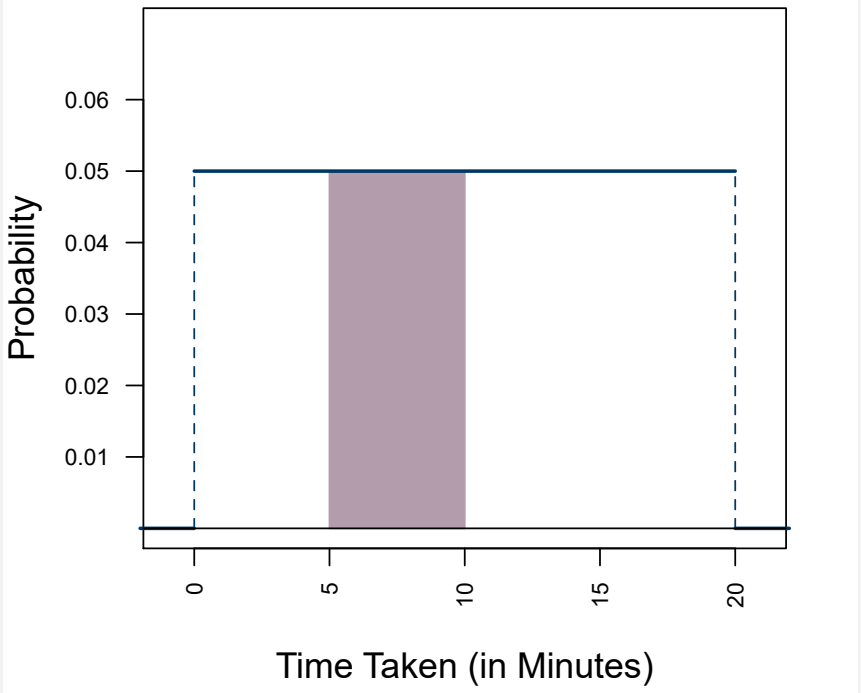


Figure 4

$$\begin{aligned}\text{Area under the curve} &= \text{length} \times \text{height} \\ &= 5 \times 0.05 \\ &= 0.25\end{aligned}$$

What is $P(5 \leq x \leq 6)$?

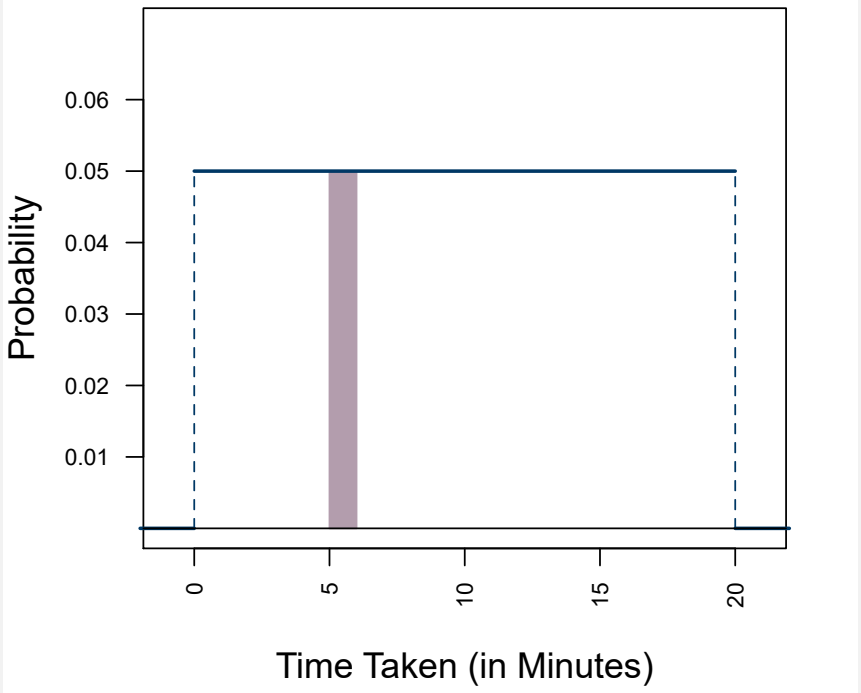


Figure 5

$$\begin{aligned}\text{Area under the curve} &= \text{length} \times \text{height} \\ &= 1 \times 0.05 \\ &= 0.05\end{aligned}$$

What is $P(5 \leq x \leq 5.001)$?

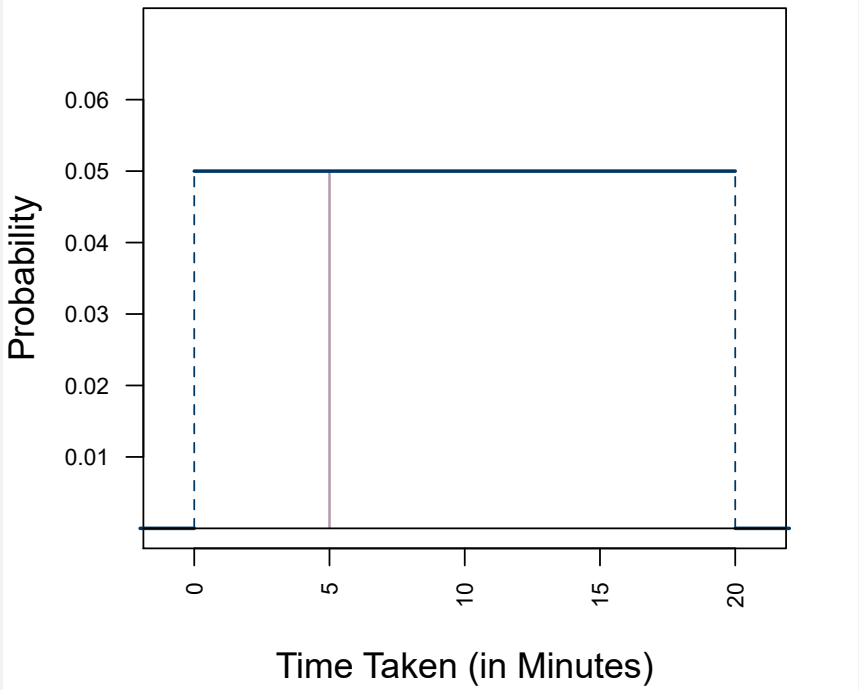


Figure 6

$$\begin{aligned}\text{Area under the curve} &= \text{length} \times \text{height} \\ &= 0.001 \times 0.05 \\ &= 0.00005\end{aligned}$$

We can see that as the length of the interval decreases, the probability of that event also decreases. Taking it one step further

What is the $P(5 \leq x \leq 5.000001)$?

$$\begin{aligned}\text{Area under the curve} &= \text{length} \times \text{height} \\ &= 0.000001 \times 0.05 \\ &= 0.00000005\end{aligned}$$

Also notice that since we are in an uncountable range space, we can continue to make the upper limit `5.000001' closer to 5 without equaling 5 (in other words, we can infinitely add 0's after the decimal point). As the upper limit tends to the lower limit, the probability of X lying in that interval goes to 0. We will soon see that this is an important property of continuous random variables.

We start by formally stating what a continuous random variable is.

Definition 1

Continuous Random Variable

A random variable with an uncountable range space is a continuous random variable.

In Example 2 we have looked at a plot of the distribution and we treated the area under that curve as a measure of probability. This function was an example of a probability density function (pdf), which we will now define.

Definition 2

Probability Density Function

Let X be a continuous random variable then the probability density function (pdf) $f_X(x)$ is the function for which for all $a \leq b$

$$P(a < X \leq b) = \int_a^b f_X(x)\, dx.$$

The integral in the definition says nothing other than that we can determine probabilities by measuring the area under the probability density function, as we have seen in Example 2. This is illustrated by the figure below.

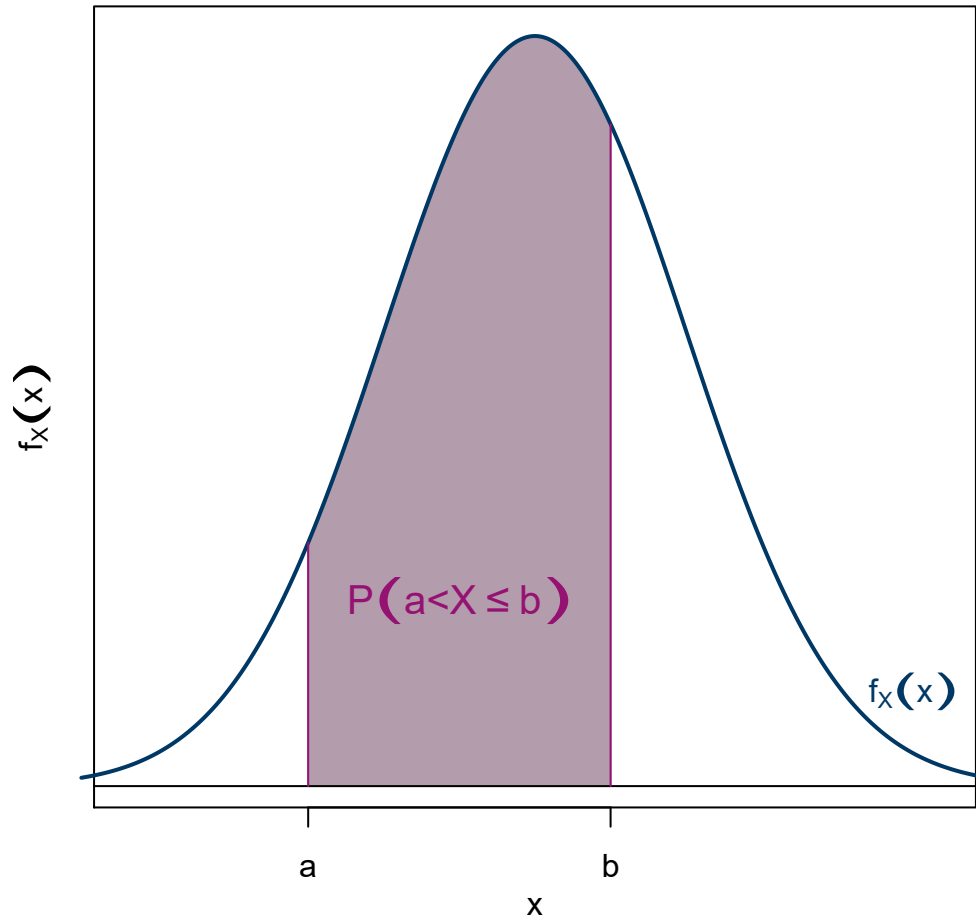


Figure 7

From the definition we can see that probability density functions need to satisfy two conditions in order to be valid.

- The probability density function cannot be negative, i.e.

$$f_X(x) \geq 0 \qquad \text{for all } x.$$

  If this wasn't the case, we could obtain negative probabilities.

- The area under the probability density function needs to be 1, i.e.

$$\int_{x \in R_X} f_X(x)\, dx.$$

  We need to enforce this condition to make sure we do not have probabilities exceeding 1 or have a mass of probability that is unaccounted for.

The notation $\int_{x \in R_X}$ indicates that we need to integrate over the entire range space of X. If the range space of the random variable is the entire real line then this just corresponds to $\int_{-\infty}^{+\infty}$. If the range space is an interval $[a, b]$ (or the interval (a, b), which is equivalent), then this corresponds to $\int_a^b$.

Note that in contrast to the probability mass function for discrete random variables (for which $p_X(x) = P(X = x)$), the probability density function for continuous random variable does not give the probability $P(X = x)$ that X is exactly x. For continuous random variables this probability is exactly zero, i.e.

$$P(X = x) = 0 \qquad \text{for all } x.$$

Being based on uncountable range spaces, continuous random variables spread the probability so thinly, that there is a probability of zero of obtaining exactly one given number. We have seen this at the end of **Example 2**. What this means in practice is that under the model of a continuous random variable, we would be very surprised if we observed the same value more than once.

Because $P(X = x) = 0$ it does not matter whether we use $<$ or $\leq$ (and similarly $>$ or $\geq$) when computing probabilities for continuous random variables, like for example the one used in **Definition 2**. Note that for discrete random variables, there can be a difference between $P(X < x)$ and $P(X \leq x)$, whereas these would be the same for continuous random variables.

Example 3

Let X be a random variable with pdf

$$f_X(x) = \begin{cases} cx^2 & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant c.

1. Find the value of c such that this is a valid pdf.

2. Find $P(X \leq 0.5)$.

In order for this to be a valid pdf, we need to make sure it is non-negative, which will be the case if c is non-negative. We also need to make sure that the area under the probability density function is 1, i.e.

$$\int_{x \in R_X} f_X(x)\, dx = 1$$

The range space of X is the interval $[-1, 1]$, so this integral corresponds to

$$\int_{-1}^{1} cx^2 = 1.$$

Therefore,

$$\begin{aligned}\int_{-1}^{1} cx^2 &= \frac{c}{3}[x^3]_{x=-1}^{1} \\ &= \frac{c}{3}[1^3-(-1)^3] \\ &= \frac{c}{3}[2] \\ &= 1 \text{ when} \\ c &= \frac{3}{2}\end{aligned}$$

integration calculator: 2c/3

so c= 3/2

Hence the pdf is

$$f_X(x) = \begin{cases} \frac{3}{2}x^2 & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

In order to find $P(X \leq 0.5)$, we need to find the area under the curve shown in purple.

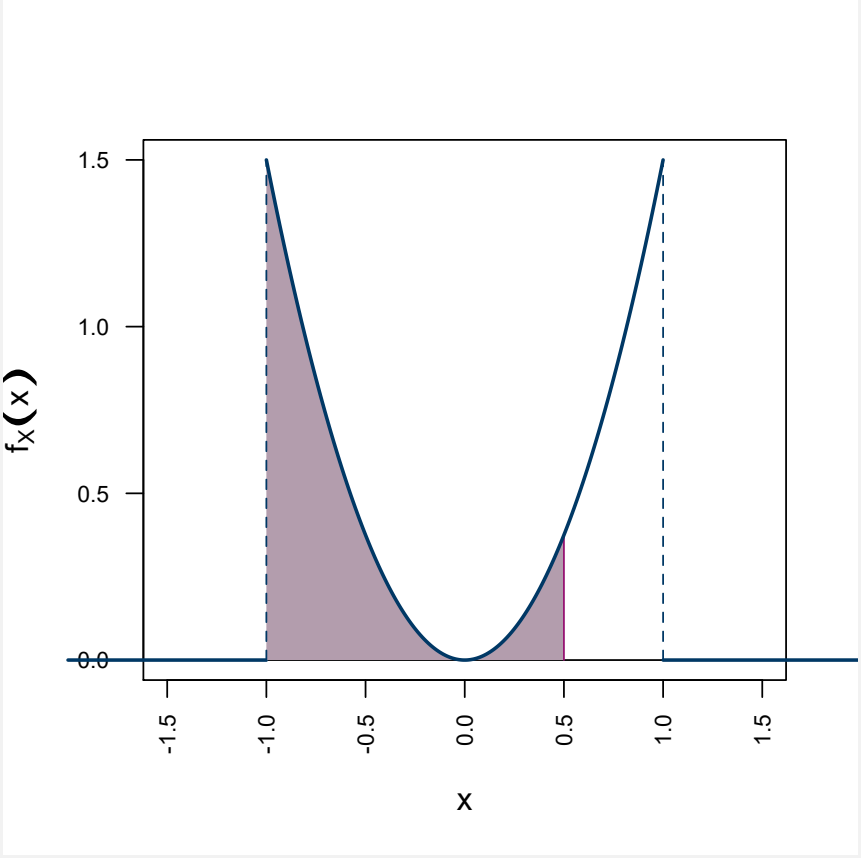


Figure 8

$$
\begin{aligned}
P(X \leq 0.5) &= \frac{3}{2} \int_{-1}^{0.5} x^2 \, dx \\
&= \frac{3}{2} \left[\frac{1}{3} x^3 \right]_{x=-1}^{0.5} \\
&= \frac{1}{2} \left[(0.5)^3 - (-1)^3 \right] \\
&= \frac{1}{2} \left[\frac{1}{8} + 1 \right] \\
&= \frac{9}{16}
\end{aligned}
$$

We do not need to consider a separate definition of the cumulative distribution function (cdf) for continuous random variables, we can still define it as

$$F_X(x) = P(X \leq x).$$

Using the definition of the probability density function and the fact that $F_X(x) = P(X \leq x) = P(-\infty < X \leq x)$ we obtain the following proposition linking the probability density function (pdf) and the cumulative distribution function (cdf).

Proposition 1

Let X be a continuous random variable with probability density function (pdf) $f_X(x)$ and cumulative distribution function (cdf) $F_X(x)$. Then for all real numbers x

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

and

$$f_X(x) = \frac{d}{dx} F_X(x).$$

The figure below illustrates these properties.

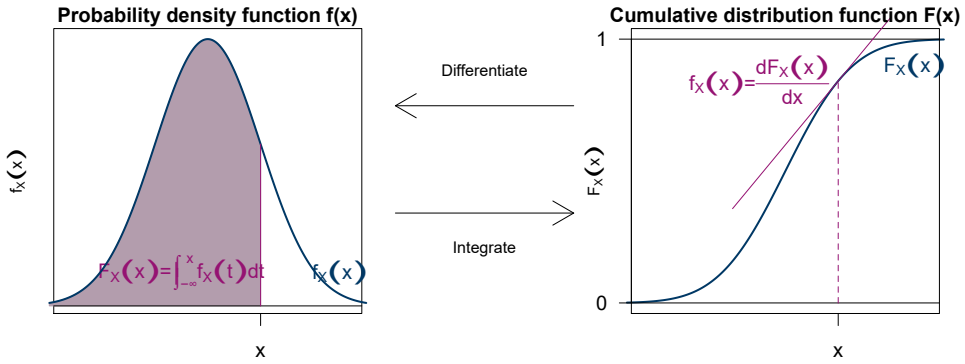


Figure 9

In contrast to the discrete case, in which the cumulative distribution function is a step function, the cumulative distribution function of a continuous random variable is a non-decreasing continuous function which "starts with being 0" ($F(-\infty) = \lim_{x \to -\infty} F_X(x) = 0$) and "ends with being 1" ($F(+\infty) = \lim_{x \to +\infty} F_X(x) = 1$).

We can use both the probability density function and the cumulative distribution function to determine probabilities of interest.

Using the probability density function (pdf)	**Using the cumulative distribution function (cdf)**
$P(X \leq b) = \int_{-\infty}^{b} f_X(x)\,dx$	$P(X \leq b) = F_X(b)$
$P(X > a) = \int_{a}^{+\infty} f_X(x)\,dx$	$P(X > a) = 1 - F_X(a)$
$P(a < X \leq b) = \int_{a}^{b} f_X(x)\,dx$	$P(a < X \leq b) = F_X(b) - F_X(a)$

As discussed earlier, we can freely swap $<$ and $\leq$ (as well as $>$ and $\geq$) for continuous random variables, as the probability of obtaining exactly one given value is zero.

The figures below illustrate how these probabilities can be obtained from the probability density function (pdf) and the cumulative distribution function (cdf).

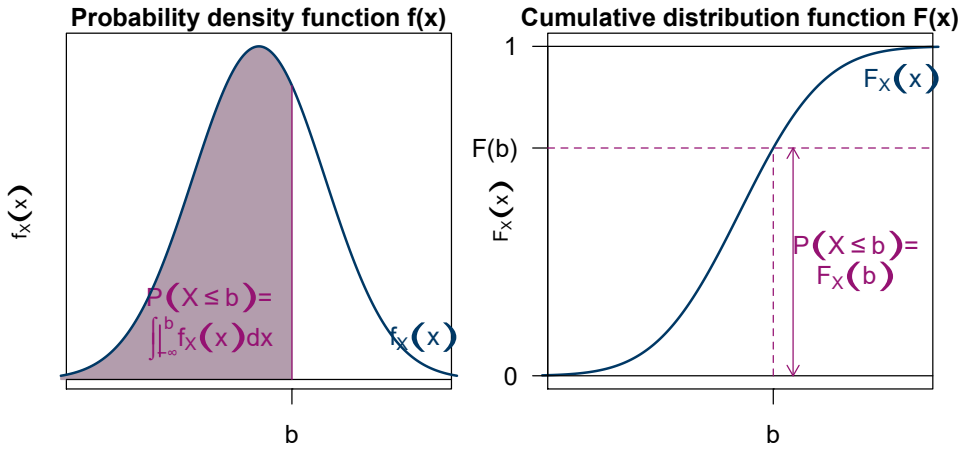


Figure 10

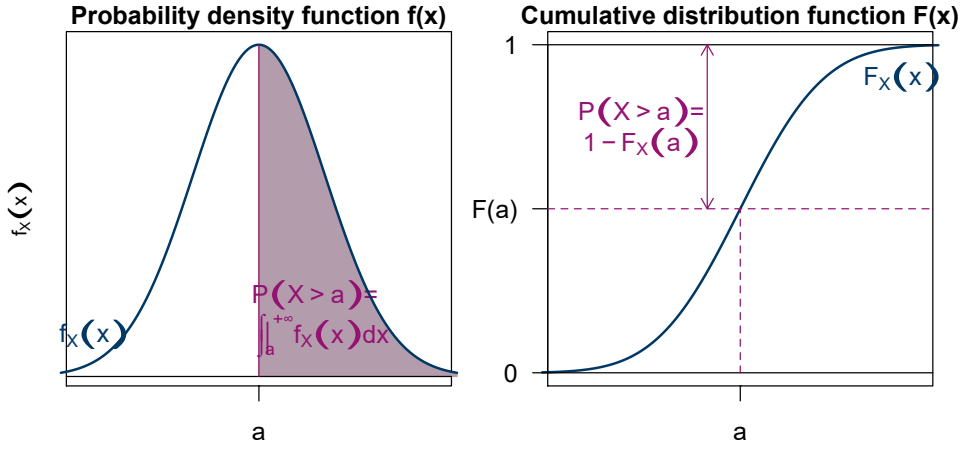


Figure 11

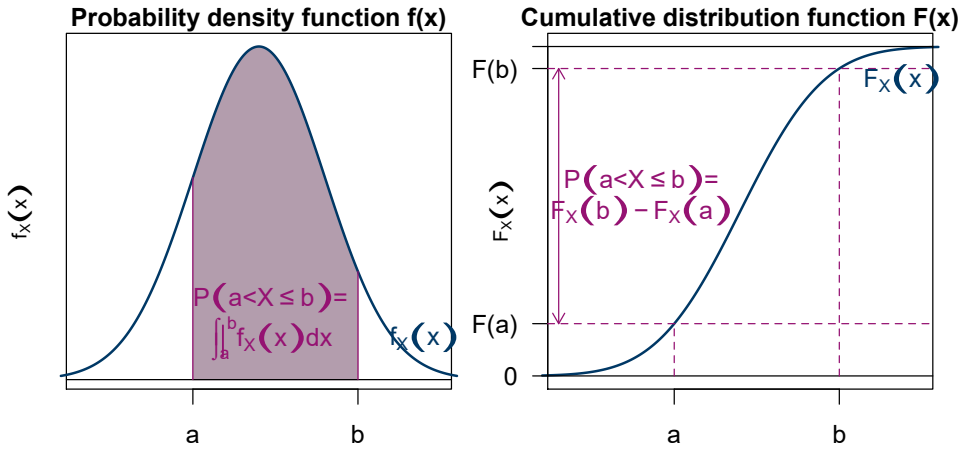


Figure 12

Example 4

In Example 3 we have found the probability $P(X \leq \frac{1}{2})$ by integrating the probability density function. We will now derive the cumulative distribution function and find the probability from the cumulative distribution function.

Because the range space of X is $[-1, 1]$ we know for $x < -1$ that $F_X(x) = P(X \leq x) = 0$. Similarly, for $x > 1$ $F_X(x) = P(X \leq x) = 1$. So we can now focus on the case that $-1 \leq x \leq 1$. In that case,

$$\begin{aligned} F_X(x) = P(X \leq x) &= \frac{3}{2} \int_{-1}^{x} t^2 \, dt \\ &= \frac{3}{2} \left[\frac{1}{3} t^3 \right]_{t=-1}^{x} \\ &= \frac{1}{2} \left[x^3 - (-1)^3 \right] \\ &= \frac{1}{2} \left[x^3 + 1 \right] \\ &= \frac{x^3 + 1}{2} \end{aligned}$$

The cumulative distribution function is thus

$$F_X(x) = \begin{cases} 0 & \text{for } x < -1 \\ \frac{x^3+1}{2} & \text{for } -1 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

It is shown in the figure below.

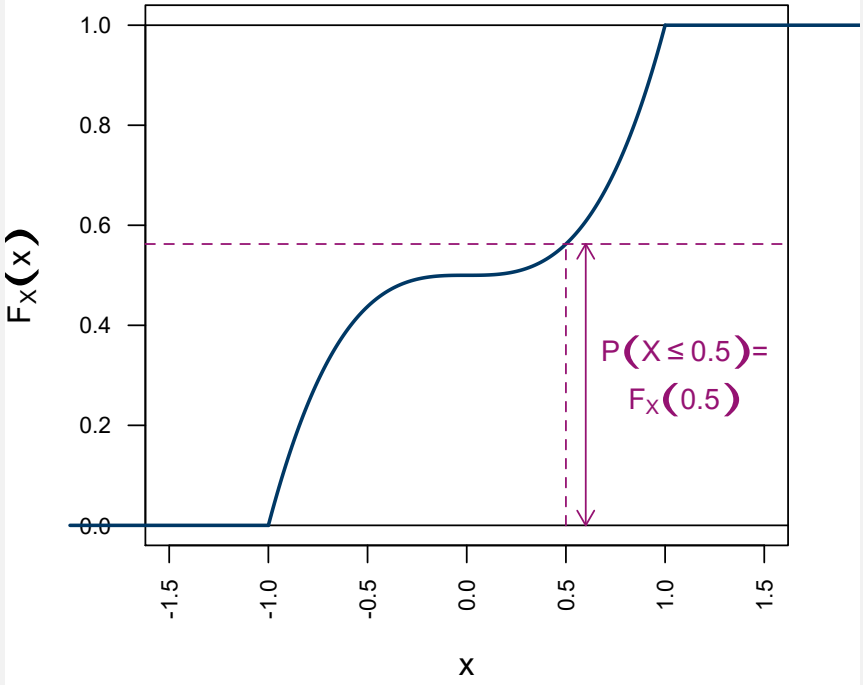


Figure 13

We can double check our result by differentiating the cumulative distribution function. This should give us the probability density function we started out with. For $-1 \leq x \leq 1$,

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}\frac{x^3+1}{2} = \frac{3}{2}x^2,$$

which is the probability density function we were given.

Finding the probability $P(X \leq 0.5)$ becomes a lot easier when we can use the cumulative distribution function.

$$P(X \leq 0.5) = F_X(0.5) = \frac{0.5^3+1}{2} = \frac{1.125}{2} = \frac{9}{16}.$$

Task 1

Let X be a continuous random variable with $R_X = (0,1)$ and

$$f_X(x) = \begin{cases} cx^n & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant c and integer n.

1. Find the value of c such that this is a valid pdf.

2. Find $P(X > x)$ for some x in R_X.

Expectation of continuous random variables and its properties

When we looked at discrete random variables, we defined the expected value ("mean") and used it as a measure of the central location of a distribution. We will now do the same for continuous distributions.

Definition 3

Expectation of a continuous random variable

Let X be a continuous random variable, then

$$E(X) = \int_{x \in R_X} x f_X(x) dx$$

(provided this integral converges)

Because $f(x)$ is zero outside the range space, it is enough to compute the integral from the lower end of the range space to its upper end.

The properties of expectation of continuous random variables are the same as those for discrete random variables covered in week 3 and week 4.

1. For continuous random variable X and constants a and b in $\mathbb{R}$,

$$E(aX + b) = aE(X) + b.$$

2. Suppose we have two continuous random variables X and Y, then

$$E(X + Y) = E(X) + E(Y).$$

3. Putting these two properties together, for continuous random variables X and Y and constants a, b and c in $\mathbb{R}$

$$E(aX + bY + c) = aE(X) + bE(Y) + c.$$

4. More generally, for any function of continuous random variable X, $g(X)$ say, then

$$E(g(X)) = \int_{-\infty}^{+\infty} g(X) f_X(x) dx.$$

Example 5

Let X be a random variable with probability density function

$$f_X(x) = \begin{cases} x(x+1) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find

1. $E(X)$
2. $E(X^2)$
3. $E(5X + 2)$

$$E(X) = \int_{x \in R_X} x f_X(x) dx$$

The range space of X is the interval $[0, 1]$, so we need to calculate

$$
\begin{aligned}
E(X) &= \int_0^1 xx(x+1)dx \\
&= \int_0^1 x^3 + x^2 dx \\
&= \left[\frac{1}{4}x^4 + \frac{1}{3}x^3\right]_{x=0}^1 \\
&= \left[\frac{1}{4}(1)^4 + \frac{1}{3}(1)^3\right] - \left[\frac{1}{4}(0)^4 + \frac{1}{3}(0)^3\right] \\
&= \frac{1}{4} + \frac{1}{3} \\
&= \frac{7}{12}
\end{aligned}
$$

$$
\begin{aligned}
E(X^2) &= \int_{x \in R_X} x^2 f_X(x) dx \\
&= \int_0^1 x^2 x(x+1)dx \\
&= \int_0^1 x^4 + x^3 dx \\
&= \left[\frac{1}{5}x^5 + \frac{1}{4}x^4\right]_{x=0}^1 \\
&= \left[\frac{1}{5}(1)^5 + \frac{1}{4}(1)^4\right] - \left[\frac{1}{5}(0)^5 + \frac{1}{4}(0)^4\right] \\
&= \frac{1}{5} + \frac{1}{4} \\
&= \frac{9}{20}
\end{aligned}
$$

$$
\begin{aligned}
E(5X+2) &= 5E(X) + 2 \\
&= 5\left(\frac{7}{12}\right) + 2 \\
&= \frac{35}{12} + 2 \\
&= \frac{59}{12}
\end{aligned}
$$

Video

Integration with R

Duration 3:05

Example 6

Randomly waiting on a train

Suppose a very reliable train line runs exactly every 20 mins between 6am and 6pm. A man 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 man expect to wait on his train?

We may assume

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

Therefore,

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

Therefore the man can expect to wait 10 mins for his train.

Video

Waiting on a Train (part 2)

Duration 1:58

Variance of continuous random variables and its properties

We will re-state the definition of the mean and variance we have introduced in week 3. There is no need to introduce a separate definition for continuous random variables.

Definition 4

Variance of a continuous random variable

The variance of a random variable X, with $E(X) = \mu$, is defined as

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= E(X^2) - \mu^2 \end{aligned}$$

Definition 5

Standard deviation of a continuous random variable

The standard deviation of a random variable X is defined as

$$\begin{aligned} \text{sd}(X) &= \sigma_X \\ &= \sqrt{\text{Var}(X)} \end{aligned}$$

Notice the definition of variance and standard deviation are exactly the same as in the discrete case. Whereas, to find the expectation, we just need to integrate $xf_X(x)$ over the range space as opposed to summing in the discrete case.

The properties of variance are also the same as in the discrete case

1. For continuous random variable X and constants a and b in $\mathbb{R}$,

$$\text{Var}(aX+b) = a^2\text{Var}(X)$$

2. Suppose we have two **independent** continuous random variables X and Y, then

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

3. Putting these two properties together, for **independent** continuous random variables X and Y and constants a, b and c in $\mathbb{R}$

$$\text{Var}(aX+bY+c) = a^2\text{Var}(X) + b^2\text{Var}(Y).$$

Example 7

Let X be a random variable with pdf

$$f_X(x) = \begin{cases} x(x+1) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathrm{Var}(X)$ and $\mathrm{sd}(X)$

We have already calculated

$$E(X) = \frac{7}{12}$$
$$E(X^2) = \frac{9}{20}$$

Therefore,

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{9}{20} - \left(\frac{7}{12}\right)^2 \\ &= \frac{9}{20} - \frac{49}{144} \\ &= 0.1097 \text{ (4 decimal places).} \end{aligned}$$

Hence $\mathrm{sd}(X) = \sqrt{0.1097} = 0.3312$

Task 2

Randomly waiting on a train

Suppose like in Example 2 a very reliable train line runs exactly every 20 mins between 6am and 6pm. A man 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. Assuming

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

find $\mathrm{Var}(X)$ and $\mathrm{sd}(X)$.

Example 8

If it is not rainy, it takes Jane anywhere between 10 to 20 minutes to drive to work and if it is rainy it takes Jane anywhere between 25 to 35 minutes to drive to work. You can assume that within each time interval, travel times are equally likely.

On average, how long does it take Jane to drive to work?

Let's begin by defining a random variable X corresponding to Jane's travel time. We want to find $E(X)$.

$$X = \text{Jane's travel time.}$$

We can define the pdf of X since we know $R_x = \{x \in \mathbb{R} | 10 \leq x \leq 20 \text{ or } 25 \leq x \leq 35\}$ and within each of these two intervals, travel times are equally likely.

$$f_X(x) = \begin{cases} t_1 & \text{for } 10 \leq x \leq 20 \\ t_2 & \text{for } 25 \leq x \leq 35 \\ 0 & \text{otherwise} \end{cases}$$

For some unknown values t_1 and t_2 that we need to estimate. In addition, we know that it rains in Glasgow with probability $\frac{2}{3}$ and so it doesn't rain with probability $\frac{1}{3}$. Therefore,

$$P(10 \leq x \leq 20) = \frac{1}{3} \qquad \text{and} \qquad P(25 \leq x \leq 35) = \frac{2}{3}$$

With this in mind, we can calculate the values of t_1 and t_2.

$$\begin{aligned} \frac{1}{3} &= \int_{10}^{20} t_1 dx \\ &= [t_1 x]_{x=10}^{20} \\ &= [20t_1 - 10t_1] \\ &= 10t_1 \end{aligned}$$

$$t_1 = \frac{1}{30}$$

$$\begin{aligned}\frac{2}{3} &= \int_{25}^{35} t_2 dx \\ &= [t_2 x]_{x=25}^{35} \\ &= [35t_2 - 25t_2] \\ &= 10t_2\end{aligned}$$

$$t_2 = \frac{1}{15}$$

Hence

$$f_X(x) = \begin{cases} \frac{1}{30} & 10 \leq x \leq 20 \\ \frac{1}{15} & 25 \leq x \leq 35 \\ 0 & \text{otherwise} \end{cases}$$

To check this is a valid pdf, we can see if $\int_{-\infty}^{+\infty} f_X(x)dx = 1$

$$\begin{aligned}\int_{-\infty}^{+\infty} f_X(x)dx &= \int_{10}^{20} \frac{1}{30}dx + \int_{25}^{35} \frac{1}{15}dx \\ &= [\frac{1}{30}x]_{x=10}^{20} + [\frac{1}{15}x]_{x=25}^{35} \\ &= [\frac{20}{30} - \frac{10}{30}] + [\frac{35}{15} - \frac{25}{15}] \\ &= [\frac{10}{30} + \frac{10}{15}] \\ &= 1\end{aligned}$$

Now we can find $E(X)$.

$$\begin{aligned}\int_{-\infty}^{+\infty} x f_X(x)dx &= \int_{10}^{20} \frac{1}{30}xdx + \int_{25}^{35} \frac{1}{15}xdx \\ &= [\frac{1}{30}\frac{1}{2}x^2]_{x=10}^{20} + [\frac{1}{15}\frac{1}{2}x^2]_{x=25}^{35} \\ &= \frac{1}{60}[20^2 - 10^2] + \frac{1}{30}[25^2 - 35^2] \\ &= \frac{1}{60}[300] + \frac{1}{30}[600] \\ &= 5 + 20 \\ &= 25\end{aligned}$$

Therefore, on average, Jane takes 25 minutes to drive to work.

Median, Quantiles and Percentiles

Median

We have already looked at the expected value as a measure of the central location of a distribution. The expected value is the theoretical equivalent of the sample mean (and thus often referred to as the mean). The expected value is not guaranteed to exist. For some heavy-tailed distributions, the expected value does not exist as the integral used in its definition is not convergent in these cases.

The median is another, more robust, take on providing a measure of central location. The median lies exactly in the "middle" of a distribution in the sense that the probability of observing a value that is at most as large as the median is 50%. This implies that the probability that X is at least as large as the median would also be 50%. If we were to drawn random realisations from the distribution of X, then, on average, half of these realisations would be less (or equal) than the median and the other half would be greater (or equal) than the median.

Definition 6

Median

The median $\xi_{50\%}$ of a continuous distribution is the value $\xi_{50\%}$, for which

$$P(X \leq \xi_{50\%}) = \frac{1}{2}$$
$$P(X \geq \xi_{50\%}) = \frac{1}{2}$$

The figures below illustrate how the median can be obtained from both the probability density function (pdf) and the cumulative distribution function (cdf).

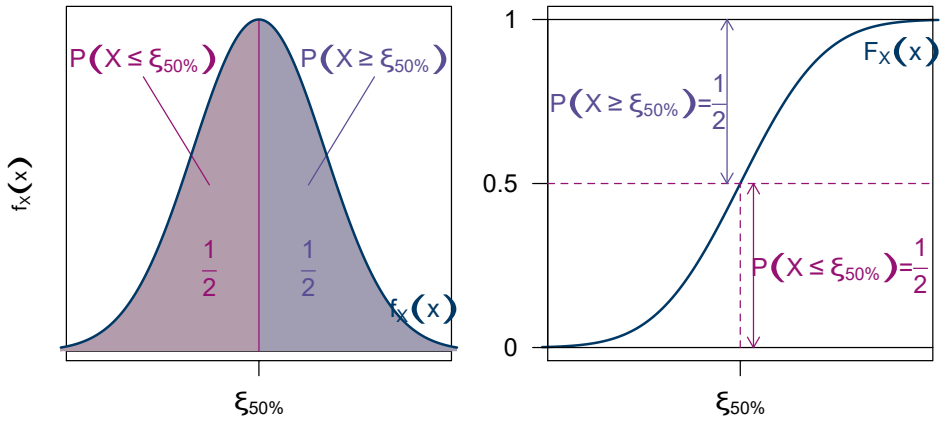


Figure 14

Note that the definition of the median says nothing other than

$$F_X(\xi_{50\%}) = \frac{1}{2}.$$

In other words, we can calculate the median using the inverse of the cumulative distribution function $F_X(X)$:

$$\xi_{50\%} = F_X^{-1}\left(\frac{1}{2}\right)$$

In contrast to the expected value, the median is guaranteed to exist.

It is more difficult to define the median of a discrete random variable, because the cumulative distribution function of a discrete random variable is a step function, rather than a non-decreasing continuous function.

Example 9

Let's return to the case from Example 2 where a man is waiting for a train on a very reliable train line on which trains run exactly every 20 mins between 6am and 6pm. A man 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 man expect to wait on his train?

We may assume

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

To find the median, we have to first find the cumulative distribution function (cdf). For $0 \leq x \leq 20$,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_(t)\, dt \\ &= \int_{0}^{x} \frac{1}{20}\, dt \\ &= \left[\frac{t}{20}\right]_{t=0}^{x} \\ &= \frac{x}{20} \end{aligned}$$

To find its inverse at $\frac{1}{2}$, we set the cumulative distribution function to $\frac{1}{2}$ and solve for x (which we will call $\xi_{50\%}$), i.e. we need to solve the equation

$$\frac{1}{2} = F_X(\xi_{50\%}) = \frac{\xi_{50\%}}{20},$$

which gives $\xi_{50\%} = 10$, i.e. the median waiting time is 10 minutes.

In this example the median is identical to the expected value. This is due to the distribution being symmetric.

Relationship to expected value and mode

In this section we look at how the median relates to other measures of the central location of a distribution. We have already looked at the mean. A third measure of the location of a distribution is the *mode*, which is the value of x for which the probability density function $f_X(x)$ is the largest, i.e. the most likely value of X.

If the distribution is symmetric around some centre of symmetry c, i.e. $f_X(c+\delta) = f_X(c-\delta)$ for all $\delta \in \mathbb{R}$, then the median $\xi_{50\%}$ and the mean μ are both c (provided the mean exists). If the probability density function is also unimodal (i.e. it only has one maximum), then the mode is also identical to the mean and median.

If the distribution is left-skewed (like the one shown on the left graph below), then the expected value μ is typically less than the median $\xi_{50\%}$, which in turn is typically less than the mode.

If the distribution is right-skewed (like the one shown on the right graph below), then the mode is typically less than the median $\xi_{50\%}$, which in turn is typically less than the expected value μ.

To better illustrate the difference between the median and the mean, imagine we have a random variable X measuring the salary of a randomly chosen employee in a company. The expected value of X would be the "average salary", whereas the median would be the salary of the "average employee", in the sense that half the employees would be paid less than the "average employee" and the other half would be paid more. The salary of the CEO will affect the expected value ("average salary"), but not the median ("salary of the average employee"). Given that the distribution of salaries is very likely to be right-skewed, the median (salary of the "average employee") would be less than the expected value ("average salary").

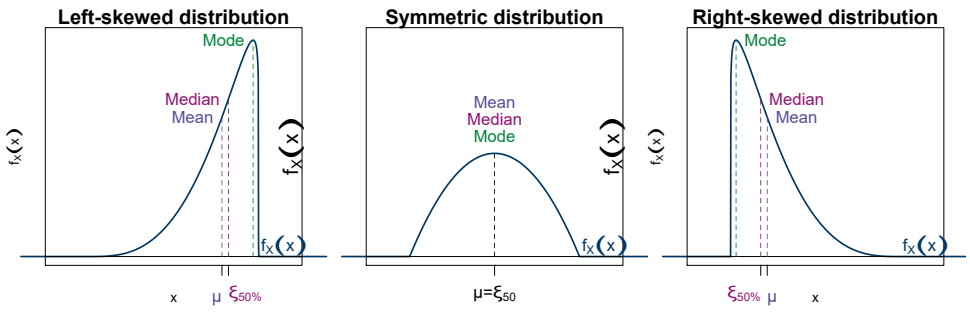


Figure 15

Quantiles and percentiles

Quantiles and percentiles are a generalisation of the median. Rather than being interested in the point in the middle of a distribution, we might be interested in more extreme outcomes. For example it might be of interest to know where the "most extreme" 5% lie. Quantiles and percentiles allow answering these questions.

Definition 7

Quantiles and percentiles

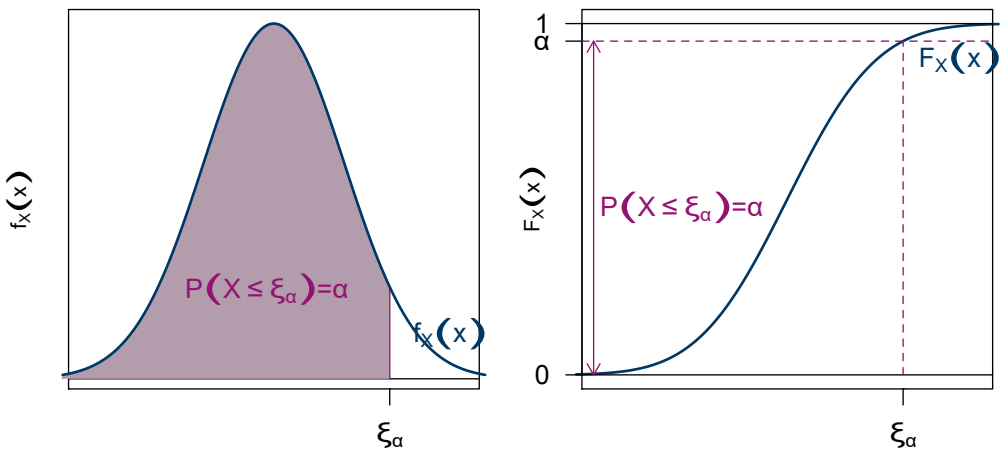
The α-quantile (percentile) of a continuous random variable X is the value ξ_α for which

$$P(X \leq \xi_\alpha) = \alpha.$$

Quantiles are typically expressed in terms of a probability, whereas percentiles are typically expressed in terms of a percentage.

The median is a special case of quantiles: it is the 0.5-quantile or, equivalently, the 50th-percentile.

The figure below illustrates this more general definition.



Like the median, we can calculate quantiles (and percentiles) from the inverse of the cumulative distribution function :inlineMath[230]: we require that :inlineMath[231], i.e.

$$\xi_\alpha = F_X^{-1}(\alpha).$$

Summary of results from week 5

Continuous Random Variable

A random variable with an uncountable range space is a continuous random variable.

Probability density function

For any $a \leq b$ the probability $P(a < X \leq b)$ is given by the area under the probability density function (pdf) $f_X(x)$ between a and b:

$$P(a < X \leq b) = \int_a^b f_X(x)\, dx$$

A valid probability density function must be

- non-negative, and
- integrate to 1, i.e. $\int_{x \in R_X} f(x)\, dx = 1$.

Relationship to cumulative distribution function

The cumulative distribution function (cdf) for a continuous random variable is

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} F_X(t)\, dt,$$

which implies

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Computing probabilities

Using the probability density function (pdf)	Using the cumulative distribution function (cdf)
$P(X \leq b) = \int_{-\infty}^{b} f_X(x)\, dx$	$P(X \leq b) = F_X(b)$
$P(X > a) = \int_{a}^{+\infty} f_X(x)\, dx$	$P(X > a) = 1 - F_X(a)$

Using the probability density function (pdf)	Using the cumulative distribution function (cdf)
$P(a < X \leq b) = \int_a^b f_X(x)\,dx$	$P(a < X \leq b) = F_X(b) - F_X(a)$

For a continuous random variable $P(X = x) = 0$.

Expectation of continuous random variable

Let X be a continuous random variable, then $E(X) = \int_{x \in R_X} x f_X(x) dx$.

Variance of continuous random variable

The variance of a random variable X is defined as $\mathrm{Var}(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$.

Answer 1

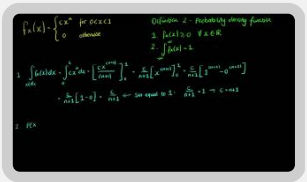
1. In order for this to be a valid pdf

$$\begin{aligned}\int_0^1 cx^n &= \frac{c}{n+1}[x^{n+1}]_{x=0}^1 \\ &= \frac{c}{n+1}[1^{n+1} - 0^{n+1}] \\ &= \frac{c}{n+1} \\ &= 1 \text{ when} \\ c &= n+1\end{aligned}$$

2.

$$\begin{aligned}P(X > x) &= \int_x^1 (n+1)x^n \\ &= [x^{n+1}]_x^1 \\ &= 1^{n+1} - x^{n+1} \\ &= 1 - x^{n+1}\end{aligned}$$

Video



Video model answer

Duration 4:57

Answer 2

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

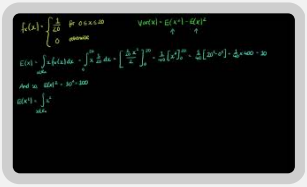
$$\begin{aligned} E(X^2) &= \int_{x \in R_X} x^2 f_X(x) dx \\ &= \int_0^{20} x^2 \frac{1}{20} dx \\ &= \frac{1}{20}\left[\frac{1}{3}x^3\right]_{x=0}^{20} \\ &= \frac{1}{20}\frac{1}{3}\left[(20)^3 - (0)^3\right] \\ &= \frac{1}{60}[8000] \\ &= \frac{400}{3} \end{aligned}$$

Therefore

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

Hence $\mathrm{sd}(X) = \sqrt{33.33} = 5.77$

Video



Video model answer

Duration 4:40