4.4: Continuous Probability Distributions

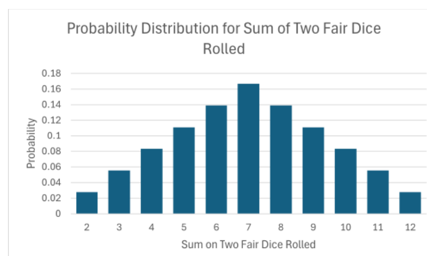
Learning Objectives

- Define the meaning of a continuous random variable probability distribution and its associated probability density function
- Use graphs to represent continuous random variables' probability distribution
- Connect area under the probability density function to probability measures for a continuous random variable
- Find area/probability measures for distributions with basic shapes

Review and Preview

We have introduced the concept of probability distributions for random variables: a distribution that represents all possible outcomes of a random variable and the associated probabilities for each. For example, we examined the discrete random variable of the sum of two rolled dice. The outcomes were sums of value 2 through 12, and the probability of each is given in the table below. A table is one way to represent the probability distribution; another is to produce a bar graph to have a pictorial representation of the distribution. We noted that the sum of the probabilities must total $1 = 100\%$ to have a complete probability distribution.

Table $4.4.1$: Probability distribution of the sum of two fair dice in graphical and tabular formats

X: Sum on Two Dice Rolled	**Probability $P(X = x_j)$**	**Graphic Representation**
2	$\frac{1}{36}$	
3	$\frac{1}{18}$	
4	$\frac{1}{12}$	
5	$\frac{1}{9}$	
6	$\frac{5}{36}$	
7	$\frac{1}{6}$	
8	$\frac{5}{36}$	
9	$\frac{4}{36}$	
10	$\frac{3}{36}$	
11	$\frac{2}{36}$	
12	$\frac{1}{36}$	
Total:	$\sum P(x_j) = \frac{36}{36} = 1.0000 = 100\%$	

Another critical concept in the above example was that the random variable was discrete. Each outcome could be listed, and the probability of each outcome was determined. Other examples include the random variable "number of days adults exercise per week" or the random variable "amount of change in teenagers' pockets."

Next, we discussed finding the mean ("expected value"), variance, and standard deviation measures from our discrete probability distribution tables. We saw how the computation concepts of grouped data (Sections 2.8 and 2.9) are used to find these measures in our probability distributions.

We also discussed in Section 4.3 a unique collection of discrete probability distributions called Binomial Distributions, distributions whose random variable is the number of successes in a given situation. If we have a well-defined success and failure in the situation, a fixed number of independent trials, and a fixed probability of success in the trials, then the probability distribution for the binomial situation is reasonably easy to construct.

Once we have the probability distribution table for a discrete random variable, we can use that information, along with our probability rules, to determine probability measures in relation to any outcomes of interest.

Now, we turn our focus to probability distributions of continuous random variables. Recall the example from Section 4.1 about the random variable of "the time (in seconds) it takes both dice in a two-dice roll to come to a complete stop after one die leaves our hand." We can no longer get accurate probability measures from a table listing outcomes and associated probabilities as in the discrete cases above. For example, there are always possible outcomes on a continuous variable between other values. Although we might build an estimated probability distribution table using intervals on the continuous random variable, doing so causes us to lose information about the distribution of the variable. We must use a different approach to maintain reasonable accuracy in dealing with continuous random variables.

Continuous Probability Distributions

Recall that a continuous variable can take on any numerical value in an interval of real numbers; in particular, another value exists between any two possible values. Examples of such variables included height, weight, ounces of water consumed, time elapsed, age, amount of electricity consumed, and many more. We must be aware that even though another height measure exists between any two heights, we measure using some chosen discrete scale, such as to the nearest inch. This rounded height measure does not make the variable discrete, the variable is still continuous. We use this information in our following theory on probability distributions on continuous variables.

A **continuous random variable probability distribution** assigns probability to an interval of values of the continuous random variable. For example, the probability distribution on the continuous variable height should give us the probability of randomly selecting a person whose height is between 5 feet and 6 feet; it should also assign a probability to any other interval of choice. This is where we move away from histograms and relative frequency tables that have specifically chosen intervals for the classes.

In Section 2.7, we demonstrated the use of a continuous mathematical function that matches the shape of a histogram graphic. Our example from that section is given below.

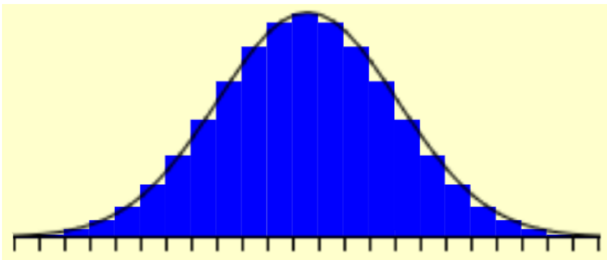


Figure $4.4.1$: Histogram with fitted curve

In modeling a continuous variable's distribution, we produce a curve that matches the behavior of the various classes in the histogram. If we move to more and more narrow class intervals, the variable will follow a function's curve. Another example is given below, in which we demonstrate the curve matching to a distribution that is positively skewed.

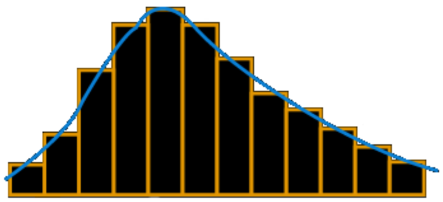


Figure $4.4.2$: A second histogram with fitted curve

To have a probability distribution, our variable's distribution and the curve fitting pf that distribution must be tied to probability (which is closely related to relative frequency). Relative frequency histograms tend to lose meaning as the width of their classes decreases, as pictured below on the left column in Figure $4.4.3$. For example, recalling that this data set represents the heights of people, notice that a little over 20% of people are 68 inches tall when their height is rounded to the nearest inch. If we instead measured height to the nearest tenth of an inch, we see approximately 2% are 68.1 inches tall, 2% are 68.2 inches tall, and so on. As we get more precise with our measurements, the proportion of people in any particular class gets smaller; hence, the relative frequencies go to 0. If the heights of the bars are the relative frequencies, then the picture degenerates. A way to overcome this issue is to represent the relative frequencies as areas instead of as heights. This is shown again below.

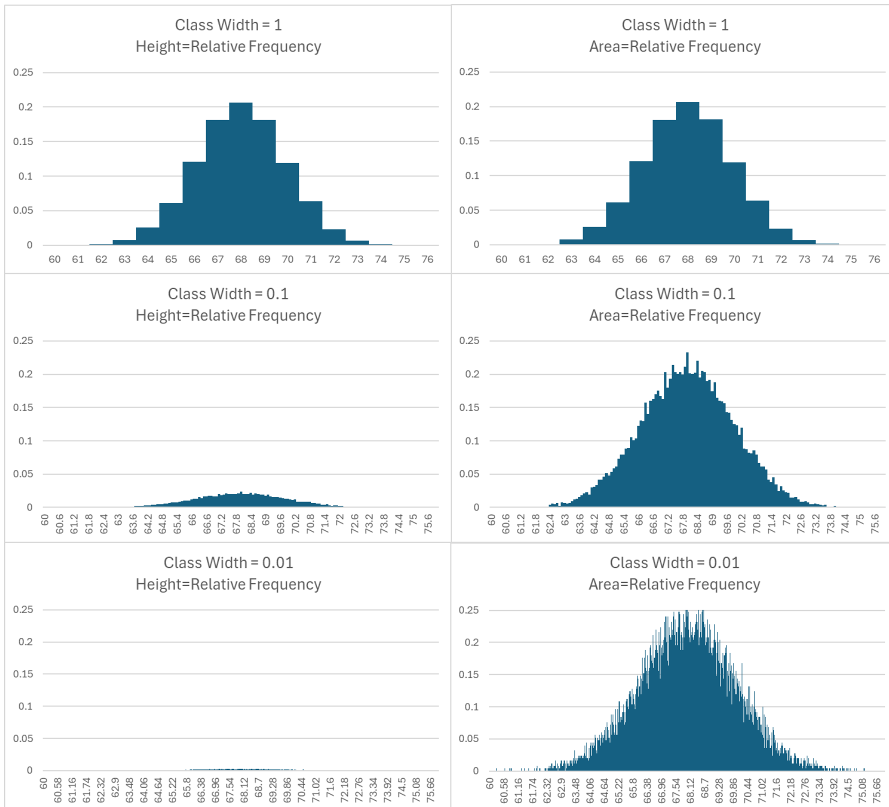


Figure $4.4.3$: Probabilities of classes as class width decreases (height of bar on left and area of bar on the right)

The curves that fit the area graphics are called **probability density functions** (PDFs for short.) The function $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ from the symmetric bell-shaped curve (commonly called the normal distribution) is the probability density function for the normal curve with a population mean of μ and standard deviation of σ. The curves are called density functions because the curve values are not directly probability measures but are measures of the denseness of probability. To find probability values, we measure the area under the density function values over an interval of values. We build regions under the probability density curve whose area measures equate to probability measures. This connection and its use will become more evident in the following sections.

All probability density functions for continuous random variables will always have three key features.

1. The domain of the curve (even if the continuous random variable has a smaller domain) can be all real numbers (in interval notation: $(-\infty,\infty)$).
2. The function values $f(x)$ for the density function will always be non-negative values; that is $f(x)\geq 0$ for all values of the continuous random variable x.
3. The total area under the curve is equal to $1=100\%$, and the area under the curve over an interval $a\leq x\leq b$ of the continuous variable will produce the probability measure $P(a\leq x\leq b)$.

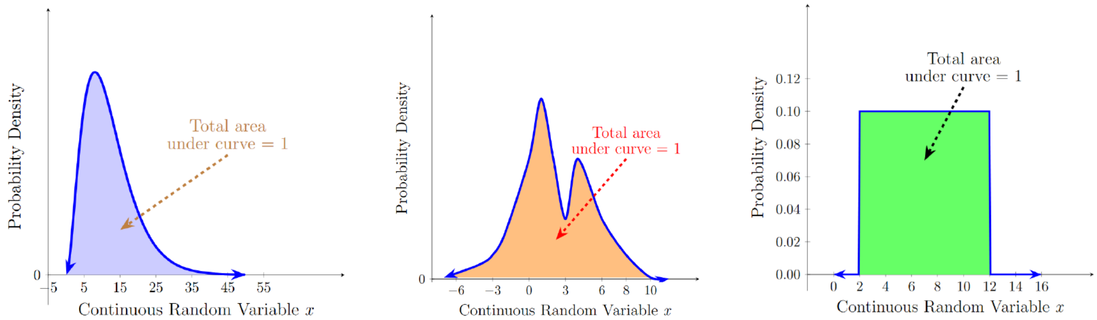


Figure 4.4.4: Examples of probability density functions

The horizontal scale will depend on the random variable being investigated. For example, the random variable represented by the left density function in Figure $4.4.4$ has its most commonly occurring outcomes between 0 and 50, the random variable of the center density function has most outcomes between -7 and 10, the random variable of the right density function has outcomes between 2 and 12. These horizontal scales are very important to the meaning of each random variable and that variable's distribution and should be included. We also note that, at times, vertical axis scaling will not be explicitly given when working with PDF graphs (compare the left two above with the right one); in general, this should not cause us concern provided we know the curve is a PDF satisfying our three requirements.

We also briefly note that these probability density functions approximate probability measures for discrete cases due to the many mathematical benefits of such curves. For example, if dealing with a binomial distribution situation in which the number of trials is large, say 500 trials, instead of building a binomial distribution table of variable values from 0 to 500--a huge table to work with--we can approximate that distribution with a single appropriate density function. This allows us to use functions instead of building a large table to examine the distribution.

Let's examine this connection between area and probability with continuous variable probability distributions.

Probability Measures from Continuous Probability Distributions

We first examine graphs of probability distributions and answer some questions concerning those distributions. For example, we might be given the graph below as a proposed probability distribution of a continuous random variable x.

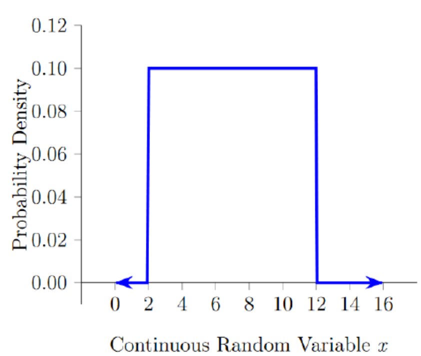


Figure 4.4.5: Example Probability Density Function

Notice for all values $2 \leq x \leq 12$; the graph shows a density function value of $PDF(x) = 0.10$, and for all other real number values of x, we have $PDF(x) = 0$. This graph implies that the continuous variable x only has possible outcomes between 2 and 12; all other real values are "impossible" outcomes since their probability density is 0. In such graphs, we focus on those intervals of variable values where $PDF(x) \neq 0$.

We should check that the three requirements of a $PDF(x)$ are met in this graphic. First, the curve's domain is all real numbers, as implied by the arrows at the end of the blue curve. Next, for all x, we see that $f(x) \geq 0$. Finally, we notice the rectangular region between the curve and the x-axis over the interval $2 \leq x \leq 12$. The width of this rectangle is $12 - 2 = 10$ units, and the height of this rectangle is a probability density measure of 0.10 units. The area calculation finds the enclosed area between the curve and x-

axis on the rectangle: $\text{Area} = \text{base} \cdot \text{height} = 10 \cdot \frac{1}{10} = 1 = 100\%$. We have a total probability measure of $1.00 = 100\%$ in this curve's area measure.

Even if we don't know the specific real-life context, this curve mathematically represents the probability distribution of some continuous random variable x. This graphic will allow us to find probability measures for different interval values; again, we focus only on intervals in which the PDF is non-zero to eliminate unnecessary work involving impossible outcomes for the variable.

For this variable x, with the given probability distribution shown above, we may wonder what the probability of randomly selecting outcomes over the interval $7 < x < 10$ would be; that is, we wish to determine $P(7 < x < 10)$. To illustrate, we can color the area within this distribution that coincides with the x values of the interval.

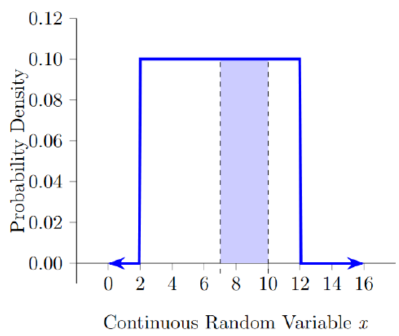


Figure 4.4.6: Finding $P(7 < x < 10)$

We notice that our shaded region is rectangular. The area of this shaded rectangle is the measure of the probability. The width of this rectangle is $10 - 7 = 3$ units in the continuous variable, and the height of this rectangle is a probability density measure of 0.10 units. The shaded area is again found by calculating the area of the rectangle:

$$\begin{aligned} \text{Area} &= \text{base} \cdot \text{height} \\ &= 3 \cdot \frac{1}{10} \\ &= 0.30 = 30\%. \end{aligned}$$

In this distribution, $P(7 < x < 10) = 0.30 = 30\%$. If we randomly select an outcome in this situation, then 30% of the time, we would expect to see an outcome between 7 and 10. Stated equivalently, 30% of outcomes in this $x-$variable's distribution are between 7 and 10.

? Text Exercise 4.4.1

Using our distribution of Figure $4.4.5$, find the following probability measures.

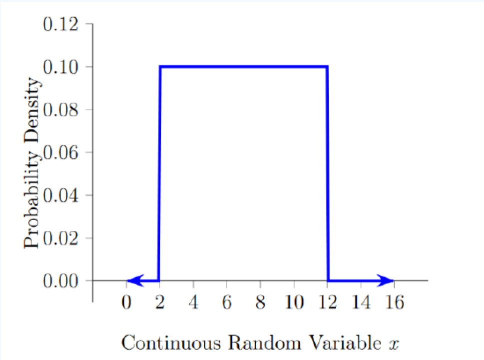


Figure 4.4.5: Replication of previous probability distribution

1. Determine $P(x \leq 8.5)$.

Answer

We shade the region under the density curve over the variable's interval $x < 8.5$.

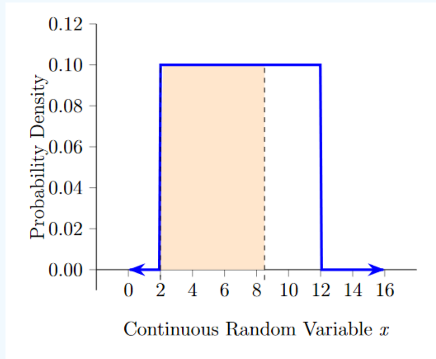


Figure 4.4.7: Finding $P(x < 8.5)$

Our shaded region is a rectangle with width $8.5 - 2 = 6.5$ and height of 0.10. So,

$$\begin{aligned} P(x < 8.5) &= \text{area of the region} \\ &= 6.5 \cdot 0.10 \\ &= 0.65 = 65\%. \end{aligned}$$

About 65% of this continuous variable's outcomes are less than 8.5 units.

2. Determine $P(x > 8.5)$.

Answer

We will take two approaches to make a critical point. Using the same approach, we shade under the density curve over the variable's interval $x > 8.5$.

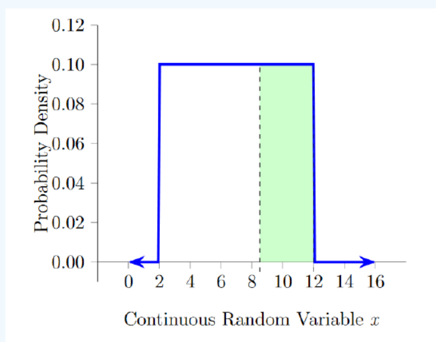


Figure 4.4.8: Finding $P(x > 8.5)$

Our region is a rectangle with width $12 - 8.5 = 3.5$ and height of 0.10. So,

$$\begin{aligned} P(x \geq 8.5) &= \text{area of the shaded region} \\ &= 3.5 \cdot 0.1 \\ &= 0.35 = 35\% \end{aligned}$$

We might show or not show solid or dashed vertical boundary lines on our regions; inclusion or exclusion will not make a measurement difference in the area.

3. Determine $P(2.75 < x < 5.5)$.

Answer

We shade under the density curve over the variable's interval $2.75 < x < 5.5$.

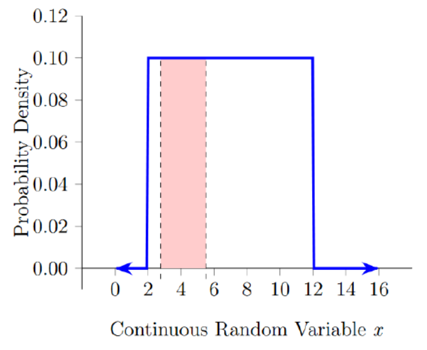


Figure 4.4.9: Finding $P(2.75 < x < 5.5)$

Our shaded region is a rectangle with width $5.5 - 2.75 = 2.75$ and height of 0.10.

$$\begin{aligned} P(2.75 < x \leq 5.5) &= \text{area of the region} \\ &= 2.75 \cdot 0.10 \\ &= 0.275 = 27.5\% \end{aligned}$$

About 27.5% of this continuous variable's outcomes are between 2.75 and 5.5 units.

The above examples and exercises were relatively straightforward since the regions of interest were always rectangles. Naturally, not all continuous random variables will have this same distribution shape.

Note: Strict and Non-Strict Inequalities

Here, we emphasize a crucial point. In our work, we make no distinction in area measures from regions formed on strict inequalities, such as $<$ or $>$, on a continuous random variable and other inequalities, such as $\leq$ or $\geq$. With continuous distributions, there is 0 area under the curve over a single value, that is, technically $P(x = a) = 0$ for any single outcome a. Therefore, the area measure of regions such as $P(x < a)$ is the same as for $P(x \leq a)$.

Due to this, when dealing with regions under continuous probability distribution functions, strict inequalities can be used interchangeably with non-strict inequalities. In our graphics of regions, we may or may not show dark or dashed vertical boundary lines on our regions; inclusion or exclusion will not make a measurement difference in the area.

We also remind ourselves that there is a difference, in general, between the use of strict and non-strict inequalities in discrete distribution probabilities discussed in earlier sections of this chapter. This demonstrates another reason why it is important to know if the random variable being analyzed is continuous or discrete.

Now, let us examine a different continuous probability distribution.

Note: Pertinent Common Area Formulas

Rectangle: $\text{base} \cdot \text{height}$

Triangle: $\frac{1}{2}\text{base} \cdot \text{height}$

Trapezoid: $\frac{\text{base}_1 + \text{base}_2}{2}\text{height}$

Text Exercise $4.4.2$

Suppose the following continuous variable distribution is given. Answer the following questions concerning this distribution.

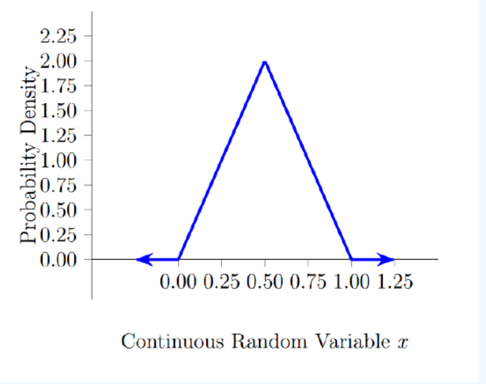


Figure $4.4.10$: Another continuous variable distribution

1. Determine if this graph can be a probability density function for a continuous variable.

Answer

We notice for all values $0 \leq x \leq 1$, the graph shows a changing density value, increasing from 0 to 2 and then decreasing back to 0. For all other real number values of x, we have $PDF(x) = 0$. This graph implies that the continuous variable x only has possible outcomes between 0 and 1; all other real values are "impossible" outcomes since their probability density is 0.

We also check that the three requirements of a $PDF(x)$ are truly met in this graphic. Notice that the domain of the curve is all real numbers, as implied by the arrows at the end of the blue curve. Next, for all x, we see that $f(x) \geq 0$. Finally, we notice a triangular region between the curve and the x-axis over the interval $0 \leq x \leq 1$. The base of this triangle is $1 - 0 = 1$ unit in the continuous variable, and the height of this rectangle is a probability density measure of 2.00units. So the enclosed area between the curve and x-axis is found by the area calculation on triangles:

$$\begin{aligned} \text{Area} &= \frac{\text{base} \cdot \text{height}}{2} \\ &= \frac{1 \cdot 2}{2} \\ &= 1 = 100\% \end{aligned}$$

Thus we have total probability measure of $1.00 = 100\%$in this curve's area. This analysis establishes that this curve does represent a probability density function for some continuous variable.

2. Determine $P(x \leq 0.50)$.

Answer

To find $P(x \leq 0.50)$, we shade under the density curve over the variable's interval $x \leq 0.50$.

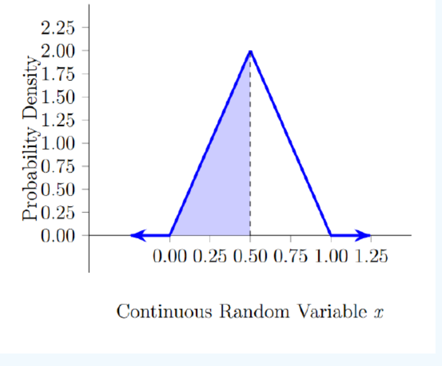


Figure $4.4.11$: Finding $P(x \leq 0.50)$

We might easily notice that our shaded region is half of the total region. This implies that $P(x \leq 0.50) = 0.50 = 50\%$.

To check this, we notice that our shaded region is a triangle with base $0.50-0.00=0.50$and height of 2.00.So,

$$\begin{aligned} P(x \leq 0.50) &= \text{area of the shaded region} \\ &= \frac{0.50 \cdot 2.00}{2} \\ &= 0.50 = 50\%. \end{aligned}$$

About 50% of this continuous variable's values are at most 0.50 units on the continuous scale. We note that many similar numbers are involved in this problem; often, we must focus more on the meaning of the values we use as we compute instead of the actual values themselves.

3. Determine $P(x \geq 0.75)$.

Answer

To find $P(x \geq 0.75)$, we again shade the related region under the density curve over the variable's interval $x \geq 0.75$.

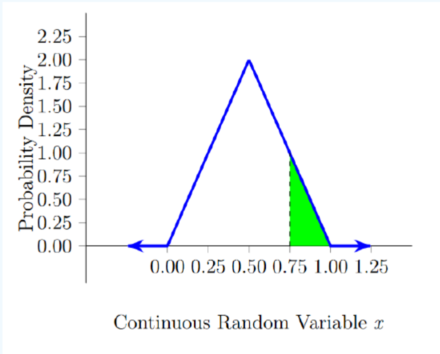


Figure $4.4.12$ Finding $P(x \geq 0.75)$

4. Determine $P(0.25 \leq x \leq 0.50)$.

Answer

To find $P(0.25 \leq x \leq 0.50)$, we shade under the density curve over the variable's interval $0.25 \leq x \leq 0.50$.

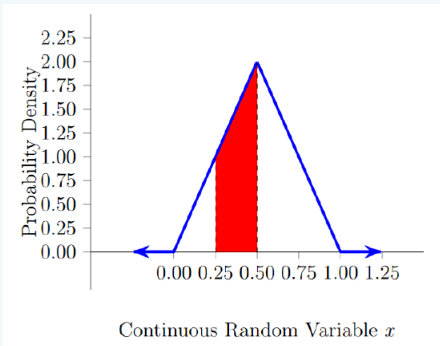


Figure $4.4.13$ Finding $P(0.25 \leq x \leq 0.50)$

We will find the area of this shaded region using two approaches. It is not about one way of thinking but finding a reasonable way to determine the shaded region's area.

For our first approach, we recall the total area under all the density functions is $1.00=100\%$. We also notice that the shaded region is surrounded by two triangles (white regions in the graphic). If we subtract the area of the two white triangles from the total area of 1.00, we will be left with the area of the red region. That is,

$$
\begin{aligned}
P(0.25 \leq x \leq 0.50) &= 1.00 - (\text{area of left white triangle} + \text{area of right white triangle}) \\
&= 1.00 - \left(\frac{0.25 \cdot 1.00}{2} + \frac{0.50 \cdot 2.00}{2}\right) \\
&= 1.00 - (0.125 + 0.50) \\
&= 1.00 - 0.625 = 0.375 = 37.5\%.
\end{aligned}
$$

So 37.5% of this continuous variable's values are between 0.25 and 0.50 units.

As a different approach, we might notice that the red-shaded region is a trapezoid, and the area of a trapezoid is found by the average of the parallel sides (commonly called the trapezoid bases) multiplied by the distance between the parallel sides (commonly called the height of the trapezoid). Using knowledge of trapezoids,

$$
\begin{aligned}
P(0.25 \leq x \leq 0.50) &= \text{area of red trapezoid} \\
&= \frac{1.00 + 2.00}{2} \cdot (0.50 - 0.25) \\
&= \frac{3}{2} \cdot 0.25 \\
&= 0.375 = 37.5\%.
\end{aligned}
$$

Although we found the area using a different approach, we see that 37.5% of this continuous variable's values are between 0.25 and 0.50 units.

As long as we have a probability distribution on a continuous variable with an appropriate probability density function, we can answer any probability question for that variable by finding the area of the related regions. Since we are naturally curious, our minds wonder what happens if our regions of interest are not always simple geometric figures. We examine such distributions at times in the following two text sections.

Summary

This section has connected probability distribution graphs on continuous random variables, probability density functions, and areas under probability density functions. Specifically, to find the probability of an interval of values for a continuous random variable, we must find the area under the related probability density function over the interval of interest. The following section will examine some of statistical analysis' most common continuous probability distributions.

4.4: Continuous Probability Distributions is shared under a Public Domain license and was authored, remixed, and/or curated by The Math Department at Fort Hays State University.

- **3.3: Measures of Central Tendency** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.