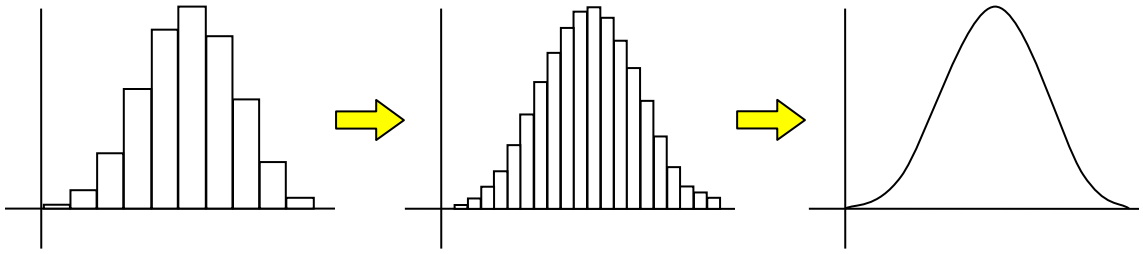
Probability Density Curves

Recall that a continuous statistical variable is a variable that can take on any value within a certain range of numbers to however many decimal places you can measure. To make a histogram graph out of the data for a continuous variable, we must divide the data into classes.

The histogram can give us an idea of how likely it is that a randomly chosen response from a survey will fall into one of the classes: the higher the bar for a particular class, the more likely that result is. We can also use a histogram to tell us how likely we are to choose a response that falls within a given range of classes. We do this by considering the heights of all the bars within that range — the sum of these bar heights tells you the chances of a random response falling within the range of interest. More properly, the *area* taken up by the bars gives us the probability, since the sum of the areas of all bars represents 100% of the data.

The more data you have, the more detailed you can make your histogram. You can make the classes narrower and the bars that result will still have meaningful heights. What if you had an infinite amount of data, and the data came from a regular, consistent source? What we mean by regular and consistent is that all the data has a similar centre (mean, μ) and a similar spread (standard deviation, σ). We could make narrower and narrower classes until they had no width at all, and the jagged corners at the top of the histogram would become smooth curves:



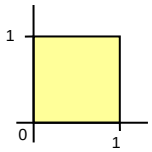
A curve like this is called a **density curve**. It is an idealized mathematical model of the behaviour of data. **The total area under any density curve is 1** because it represents all the possible data values that the variable can take (the sum of all possible outcomes of an event must equal 1). Also, a density curve will always be above the horizontal axis since having any part of it be below the axis would imply a negative probability of a value occurring, which is meaningless.

We can use a density curve to find the probability that a randomly chosen representative of a data set has a value higher or lower than a specified number, or between specified numbers. **To find a probability using a density curve, we make a vertical slice through the curve** at any value at the edge of an interval we're interested in, **and we calculate the area under the curve** within the interval. This is an extremely important concept to master, because the entire rest of the stats course will

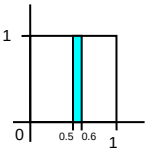
deal with calculating probabilities with density curves! This is a tool in statistics for taking information about a real-world sample and extrapolating information about a population.

Example 1: A random number generator on a calculator displays numbers from 0 to 1 with a uniform distribution, i.e. a flat distribution. Find the probability that the calculator will display a number between 0.5 and 0.6 when prompted.

Solution: First, we need to draw the density curve for the situation. The range of the generator is from 0 to 1, so the curve exists over the x-axis from 0 to 1 only. We are also told that the distribution is flat, so our "curve" is really a rectangle. (Don't worry about the use of the word "curve" — to a mathematician curves don't have to bend, so straight lines are all "curves". You'll also find that when we talk about density curves we may say "curve" to mean the area or shape under the curve.) The area within the rectangle has to be equal to 1. We calculate area for a rectangle using the familiar formula: length times height. The length of the rectangle is 1 ($x_{final} - x_{initial} = 1 - 0 = 1$), so the height must also be 1. The density curve looks like the one at the right:

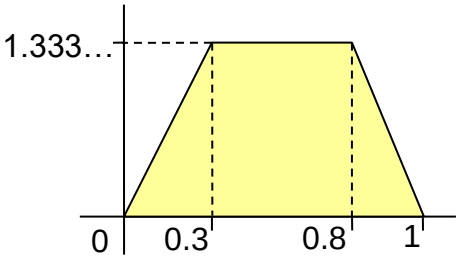


We're interested in values between 0.5 and 0.6. We make vertical slices through the curve at these locations on the x-axis. Think of it like cutting a cake. You can make two slices and pull out a piece of the cake, or you can make one slice and take one side of what's left of the cake. In our case, we are taking a small slice of cake out of the middle. The area of the rectangle between the slices is the probability of getting an answer between those values. The length of the rectangle is 0.1 ($x_{final} - x_{initial} = 0.6 - 0.5 = 0.1$) and the height hasn't changed, so the area is $0.1 \times 1 = 0.1$. This means that there is a probability of 0.1 of getting a value between 0.5 and 0.6.



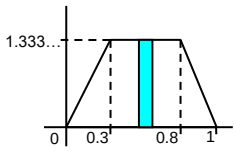
One issue that was never addressed in this example: is the range inclusive or exclusive? Do 0.5 and 0.6 count in that probability? It doesn't matter whether they do or not; the answer will be the same. The probability of the calculator displaying 0.50000000… is so negligible that it doesn't affect our answer.

Example 2: If you ask a person to pick random numbers between 0 and 1, the results will not in fact be random. Humans tend to pick numbers between 3 and 7 inclusive for the first decimal place, and then pick different digits after that. The probability density curve of the resulting "random" numbers will look like the graph at the right. Find the probability that a bingo caller from Abbotsford, when asked to write down a random number, will write a number between (a) 0.5 and 0.6. (b) 0 and 0.3. (c) 0.7 and 1.2. (d) 0.3 and 1.

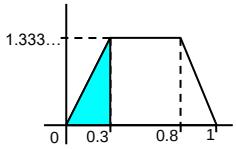


Solution: Our method of solution will be the same: take vertical slices through the density curve, and find the area between the slices. It will be useful for this question to remember the formula for the area of a triangle: $A = \frac{1}{2}bh$ (base × height).

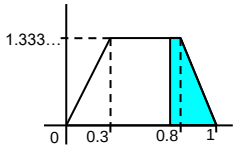
(a) We're interested in values between 0.5 and 0.6. We make vertical slices through the curve at these locations on the x-axis. The area of the rectangle between the slices is the probability of getting an answer between those values. The length of the rectangle is 0.1 ($x_{final} - x_{initial} = 0.6 - 0.5 = 0.1$) and the height is 1.333, so the area is $0.1 \times 1.333 = 0.1333$. This means that there is a probability of 0.1333 (or 13.33%) of getting a value between 0.5 and 0.6.



(b) We're interested in values between 0 and 0.3. This section of the density curve is a triangle. Its width is 0.3 ($x_{final} - x_{initial} = 0.3 - 0 = 0.3$) and the height is 1.333. The area is $\frac{1}{2} \cdot 0.3 \cdot 1.333 = 0.2$, so the probability of getting a value between 0 and 0.3 is 0.2 (20%).



(c) We're interested in values between 0.7 and 1.2, but since there are no possible values from 1 to 1.2, the area there is 0 and we can ignore it. We make a slice at 0.7 and consider the shape to the right of this slice. We can cut the shape into two pieces whose areas are easier to calculate: a rectangle to the left of the dotted line and a triangle to the right. Using the same techniques as above, the rectangle from 0.7 to 0.8 has an area of $0.1 \times 1.333 = 0.1333$, and the triangle from 0.8 to 1 has an area of $\frac{1}{2} \times 0.2 \times 1.333 = 0.1333$, so the total area is 0.2666, which is the probability of getting a value from 0.7 to 1.2.



(d) We're interested in values between 0.3 and 1, which is the opposite of part (b). We're making the slice in the same place, but this time we want the other side of it. Since the area under the curve is equal to 1 (otherwise it's not a real density curve), we can take a short cut to the answer by subtracting the area from part (b) from 1. The probability of getting a value from 0.3 to 1 is $1 - 0.2 = 0.8$. This method of finding the value on the other side of a line by subtracting from 1 is an important one. We could easily calculate this area by cutting it into pieces as in part (c), but in future applications this won't be so easy.

NOTES

There are two numbers involved in each of these problems: the value we can get from a real-world response to a survey and the probability of getting a result within a particular interval. The first manifests itself as a mark on the horizontal axis below a graph, and the other as an area under a curve. A common mistake students make is to get these numbers confused in a problem.

All of the examples below have density "curves" made entirely out of straight lines. In truth, the density curves of real events tend to be actual curves. It requires calculus to take slices of curves in the manner we've been describing here, but you won't be responsible for that in this course. Instead, you'll be looking these numbers up in tables.

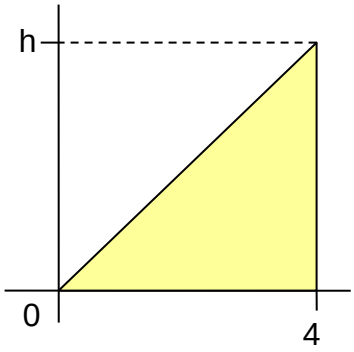
EXERCISES

A. A random-number generator is established that will give decimal numbers between 1 and 3. The distribution of the numbers it generates is uniform (flat).

1) What is the height of the density curve that represents this situation?

2) Estimate the probability that this generator will produce a result between 2 and 3. Determine the probability of this event. Did the calculation match your prediction?

B. Another random number generator is set up to produce numbers between 0 and 4, but with a distribution that favours higher numbers. The density curve comes out to be a right triangle, as shown in the diagram.



1) Determine the value of h, the height of the density curve at its highest point.

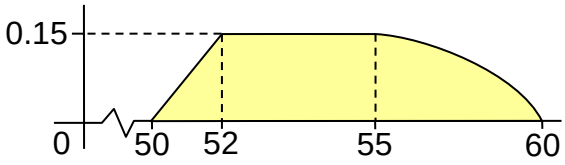
2) Determine the probability of the generator producing a number between 0 and 2. [*Hint: since the curve is a triangle, the top line of the shape has a consistent slope. This means that whatever fraction of the way along the horizontal axis your mark is, the height at that point will be that fraction of the overall height.*]

3) Determine the probability of the generator producing a number between 2 and 4.

C. Verify the answer to part (d) of Example 2 by calculating the total area under the curve from 0.3 to 1.

D. A poll asks people to select a number from 1 to 10. Suggest why these results could not be modelled by a density curve.

E. A statistical variable has values that range from 50 to 60. Its probability density curve, shown to the right, is sloped from the values 50 to 52, flat from 52 from 55 and curved from 55 to 60. Find the probability of a result greater than 54 for this variable. [*Hint: You will need to calculate this answer indirectly.*]



SOLUTIONS

A. (1) Since the length of the rectangle is 2 (3 − 1) and the area is 1, the height is ½.
(2) You might predict that the probability of getting a number from 2 to 3 (half the range of values) is ½. This is supported by the area calculation: 1 × ½ = ½.

B. (1) The base of the triangle is 4 and the area is 1, so the height is ½.
(2) The height of the slice at 2 is ¼. The area is ½ × 2 × ¼ = ¼. (3) 1 − ¼ = ¾

C. The area from 0.3 to 0.8 is 0.5 × 1.333 = 0.6667 to four decimal places.
0.6667 + 0.1333 = 0.8.

D. This poll would likely result in discrete data. Most people would say a whole number, not continuous decimal numbers. A density curve can only describe continuous data.

E. The area under the curve for values below less than 54 is ½ × 2 × 0.15 + 2 × 0.15 = 0.45. Therefore the area above 54 is 1 − 0.45 = 0.55.