ST 516: Foundations of Data Analytics

Finding probabilities as area under curves

The advantage of thinking about probability distribution functions for random variables is that we can use them to easily find probabilities.

The probability of the outcome of our random variable being between two values a and b is described by the area under our probability distirbution function between those two values.

Conceptualy, it doesn't matter if we are talking about discrete or continuous random variable.

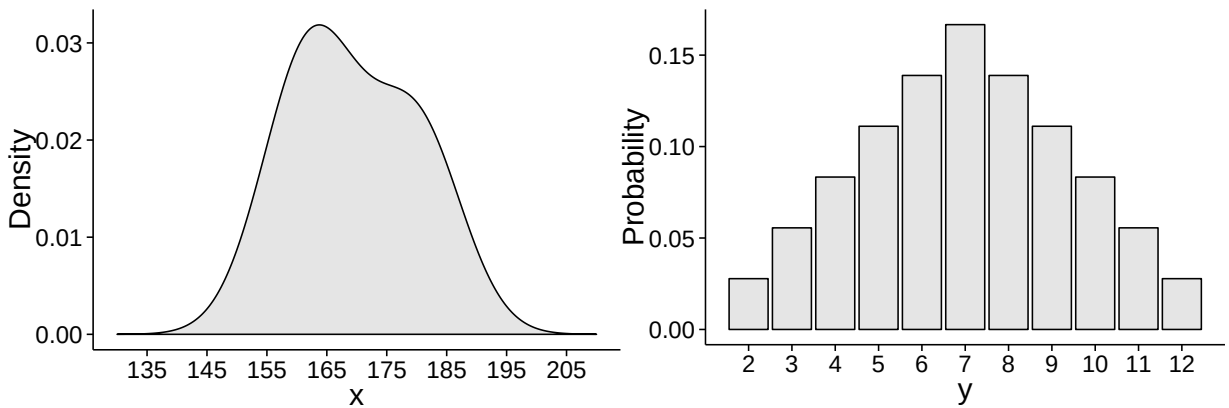
Examples

Let X be the continuous random variable,

$$X = \text{the height in cm of a randomly sampled US adult}$$

Let Y be the discrete random variable,

$$Y = \text{the sum of two dice}$$

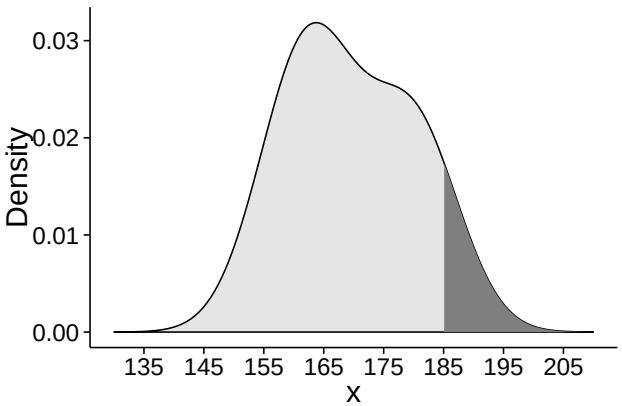


In each of the following examples, the probability of interest is the area shaded in dark grey.

One sided range

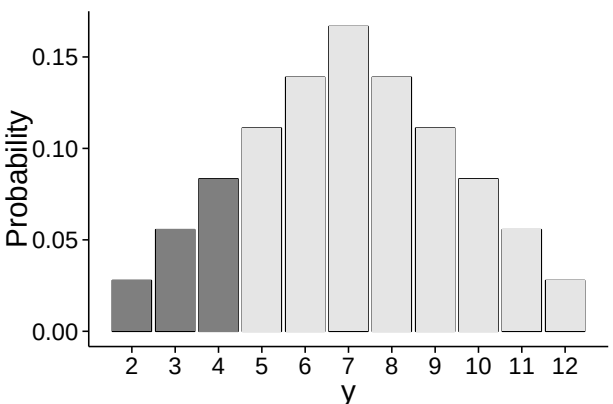
The probability that our randomly sampled US adult is taller than 185cm.

$P(X > 185)$



The probability that the sum of our two dice is less than 5.

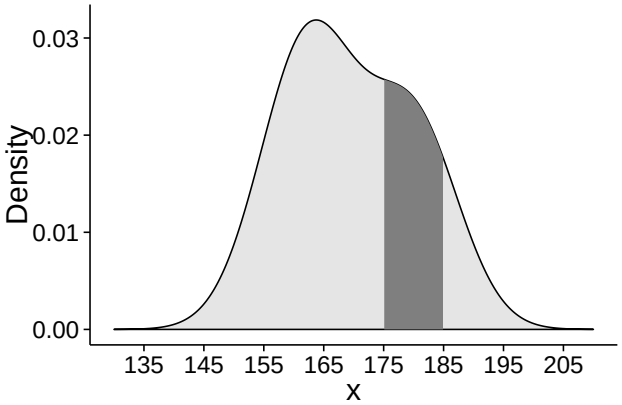
$P(Y < 5)$



Two sided range

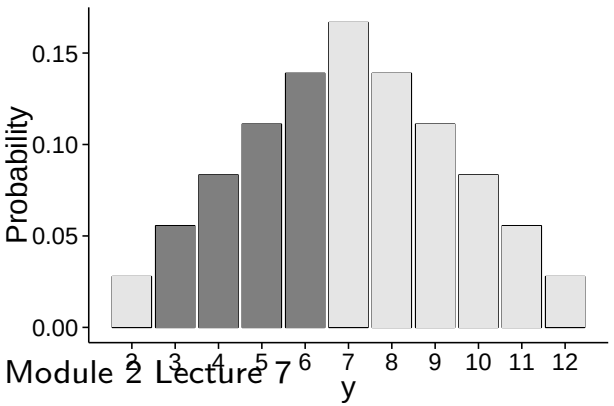
The probability that our randomly sampled US adult is between 175cm and 185cm tall.

$P(175 < X < 185)$



The probability that the sum of our two dice is more than 2 but less than 7

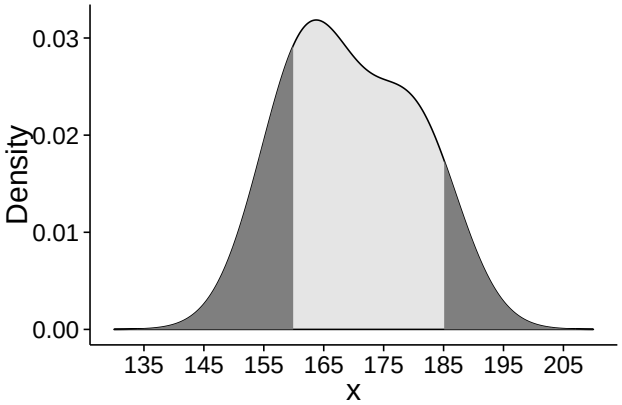
$P(2 < Y < 7)$



Disjoint ranges

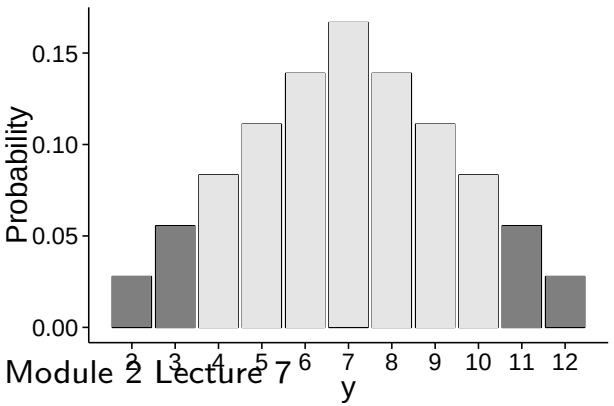
The probability that our randomly sampled US adult is shorter than 160cm or taller than 185cm.

$P(X < 160 \text{ or } X > 185)$



The probability that the sum of our two dice is less than 4 or greater than 10.

$P(Y < 4 \text{ or } Y > 10)$



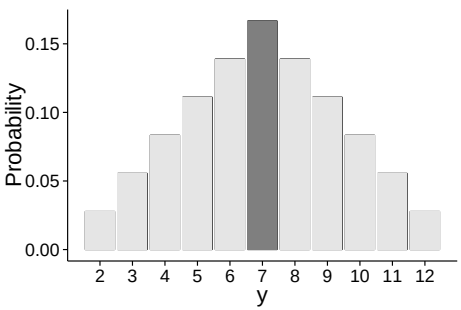
A tricky case

One case that is a little different between discrete and continuous is when we want the probability of an exact number.

For discrete variables, it's exactly what you might expect.

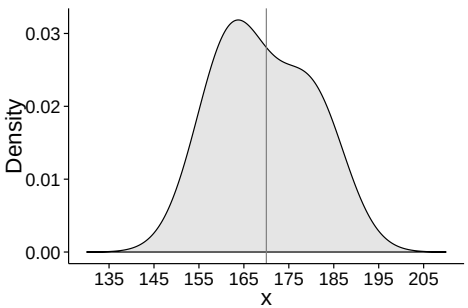
Probability of getting exactly a sum of 7 from my two dice.

$$P(Y = 7) = \frac{6}{36}$$



But, for continuous variables, the probability of an exact number is always zero, because the area we are trying to calculate has zero width.

The probability that our randomly sampled US adult is exactly 170cm tall. $P(X = 170) = 0$



In practice this means for continuous variables it doesn't matter if you use $\leq$ or $<$, but for discrete variables it does.

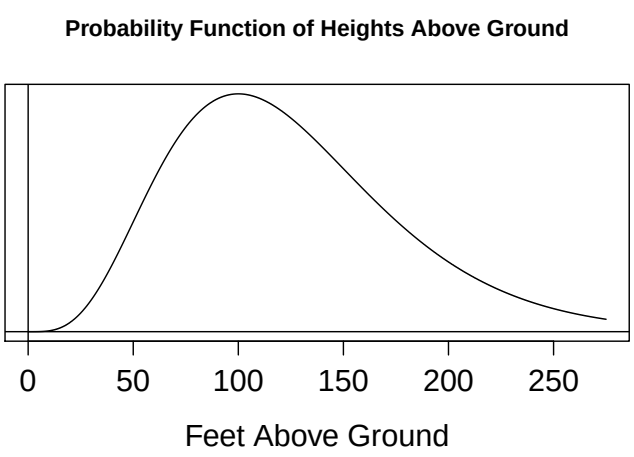
Combining with the rules of probability

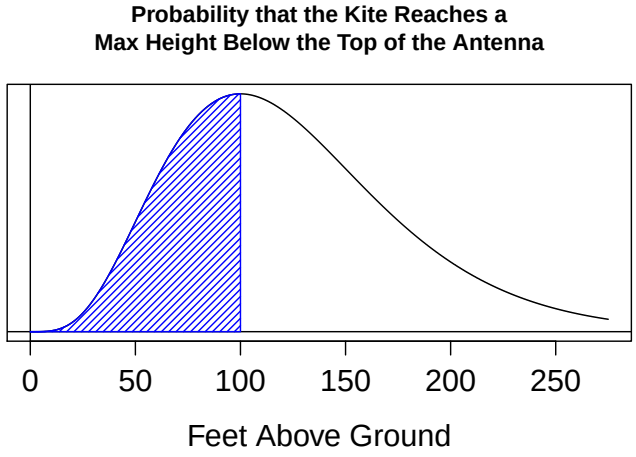
Sandra decides to fly a kite at the Pacific Ocean beach in Florence, Oregon on a windy day. There is a building nearby whose roof is 50 ft above the ground. On top of the roof there is an antenna that extends 50 ft above the building's roof.

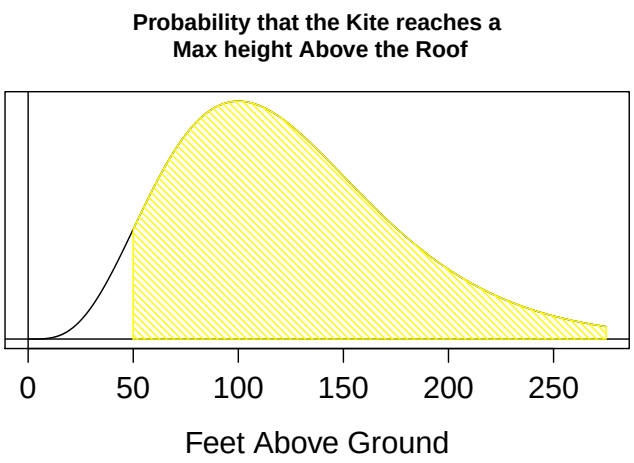
The kite reaches a maximum height above the building's roof with a probability of 0.8 and a maximum height below the top of the antenna with a probability of 0.4.

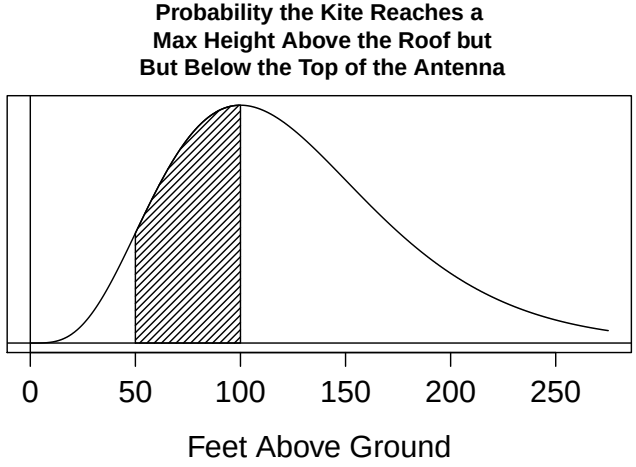
What is the probability that the kite reaches a height between the roof and the top of the antenna?

Combining with the rules of probability









Answer: $0.4 - (1 - 0.8) = 0.2$

Finding the probability

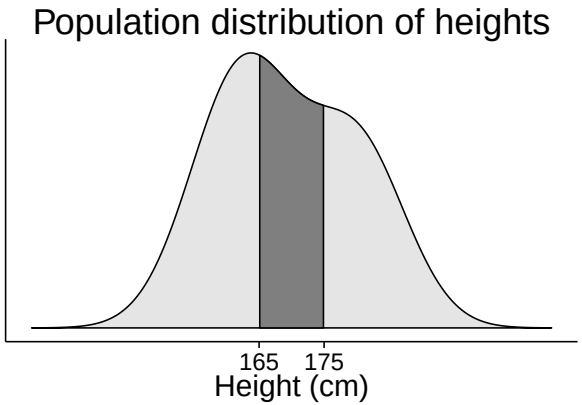
Finding an area under a curve involves integration (or summing for discrete variables). How can we estimate the area with simulation?

The general procedure is to:

- Simulate the random variable many times
- Find the proportion of simulated values that satisfy the conditions for our event. This is our simulation based estimate of the probability of the event.

Example

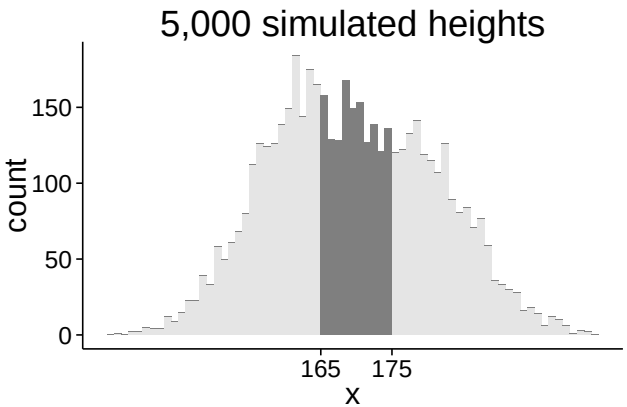
I want to find the probability a randomly sampled US adult is between 165cm and 175cm tall.



If we have a function `sim_height` that produces a single realization of the random variable, we can simulate it many times.

```
x <- replicate(5000, sim_height())
```

Our simulated values approximate the population density,



We can find the proportion of our simulated values that satisfy the criteria for our event of interest,

```
sum(x > 165 & x < 175)/5000
```

```
## [1] 0.2816
```

Our simulation estimate of the probability that a randomly sampled US adult is beteen 165 and 175 cm tall is 0.2816.

Simulation Error

If we simulated again, would we get the same number? No, because we would get a slightly different set of 5,000 heights, and consequently a slightly different proportion.

How close is our simulation estimate to the true probability? In this case, the true probability is 0.2805. So, with 5000 simulations we are pretty close. In general the more simulations we do the closer our estimate will be to the true value (our simulation results follow the Law of Large Numbers!).