Introduction to Statistical Ideas and Methods

Probability: Random Variables

Probabilities for Continuous Random Variables

For some random variables, it is impossible to list all of the possible values they can take. For example, if we were measuring birth weights of babies or life expectancies of countries, and we wanted to measure to any number of decimal places, we could not write down all of the possible values of birthweight or life expectancy. Therefore if we wanted to assign probabilities to the possible values of the random variable, we need to use a mathematical model.

A random variable takes each outcome from a random experiment (such as flipping a coin, spinning a roulette wheel, or measuring the birth weight of a randomly chosen baby), and gives it a numerical value. We say a random variable is **continuous** if the numbers it can be are anything in an interval. This interval can extend as low as negative infinity and up to positive infinity.

Example

Suppose buses arrive at a bus stop every 10 minutes. We want a probability model for how long a rider has to wait for a bus. Suppose we do not know when the last bus arrived. Then a good model for this situation would be a model that for a bus it is equally likely to arrive at any time between 0 and 10 minutes. We can illustrate the probability of its arrival with this line over the interval from 0 to 10 minutes:

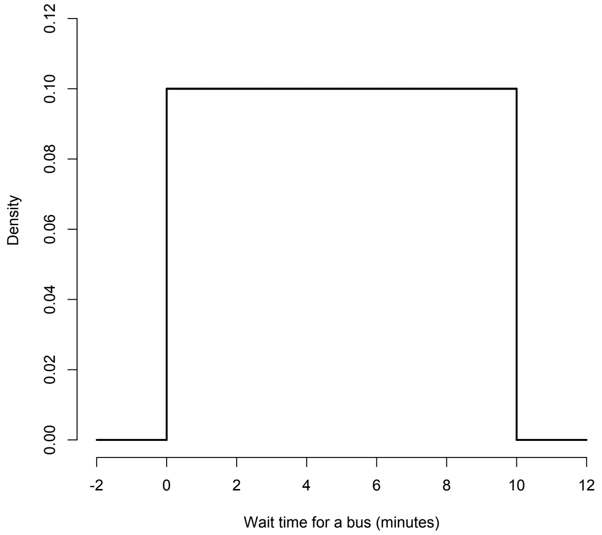

Figure 1: Theoretical model for 'Bus arrival' example

The line is horizontal, because the bus is equally likely to arrive at any time in the interval. Since the interval is 10 minutes long, the probability the bus arrives in the first minute, should be $\frac{1}{10} = 0.10$. We will represent that probability as the area under our horizontal line, from 0 to 1 minutes. For the area of this rectangle to be 0.10, the height of this line must also be 0.10 (since *Area of Rectangle*= *height* × *base*).

This is an example of a **Uniform** distribution for a continuous random variable. Let's call this random variable X. So that:

$X =$ wait time until the next bus

Hence X has a uniform distribution on the interval 0 to 10. Mathematically we write this as:

$$X \sim Uniform[0, 10]$$

Here $\sim$ means X has distribution. Note that X is a continuous random variable because it can be any number in the 0 to 10 interval.

The horizontal line from 0 to 10 at height 0.10 is an example of a **probability density function** (PDF). Probability density functions are mathematical functions that describe the probability distribution of continuous random variables.

A probability density function is a curve that must satisfy two conditions:

- It must always be greater than or equal to 0.
- The total area under the curve must be equal to 1.

See that the PDF above satisfies these conditions. Note that when we sketch a PDF, the values of the random variables are on the horizontal axis, and values of the density function on the vertical axis. The actual values of the density function are not typically interesting. Instead, the interesting aspect is areas underneath the function. Probabilities for continuous random variable are found by finding areas under its probability density function. The probability that a random variable falls into any particular interval is equal to the area under the density curve above this interval.

Examples

Suppose we want the probability our rider waits less than 2 minutes for the bus ($P(X < 2)$). We need to find the area under the density curve from 0 to 2, which is the area of the rectangle with width 2 and height $\frac{1}{10}$. Therefore:

$$P(X < 2) = 2 \times \frac{1}{10} = 0.20$$

The probability of waiting between 2 and 6 minutes ($P(2 < X < 6)$), is the area of the rectangle with width 4, and height $\frac{1}{10}$ so:

$$P(2 < X < 6) = 4 \times \frac{1}{10} = \frac{4}{10} = 0.40$$

Note: the probability that our rider must wait exactly 2 minutes for the bus is 0 ($P(X = 2) = 0$). Since if we tried to find the area under our density curve for the wait times at 2 minutes, it would be 0 since we have 0 width. Actually for any continuous random variable, the probability that it is equal to any one particular value is 0. Thus we have next results, for any continuous random variable X and any numeric value a,

$$P(X = a) = 0$$

and

$$P(X < a) = P(X \leq a)$$

Here is another way to think about the concept of a probability density function.

Suppose our bus rider recorded how long she waited for a bus for her last 200 rides. We could picture these data in a histogram, showing the frequencies, which are the counts of how many times she waited between 0 and 1 minute, between 1 and 2 minutes, and so on.

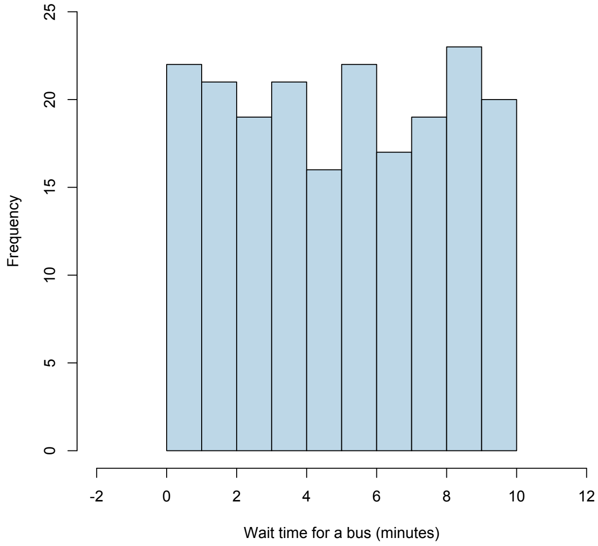

Figure 2: Histogram of wait times in ‘Bus arrival’ example (counts)

If we change the scale on the vertical axis so that the total area of all of the rectangles in the histogram is 1, we get a density histogram.

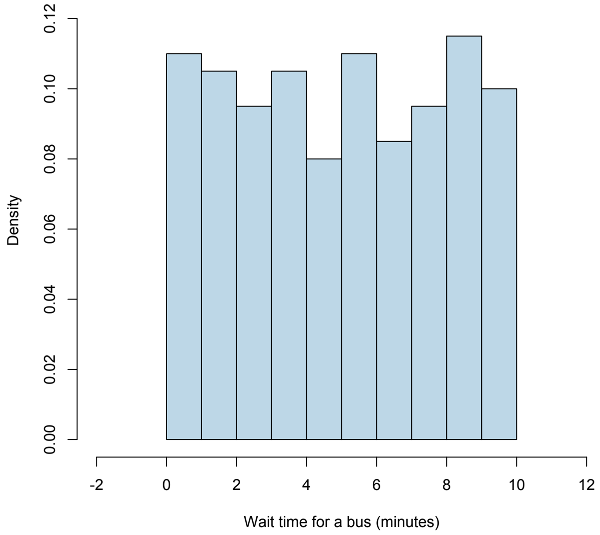

Figure 3: Histogram of wait times in 'Bus arrival' example (density)

Areas under this scaled histogram give the proportion of wait times that are in an interval. For example, the area between 8 and 10 minutes is a total of about 0.110 plus 0.105 and equals to 0.215.

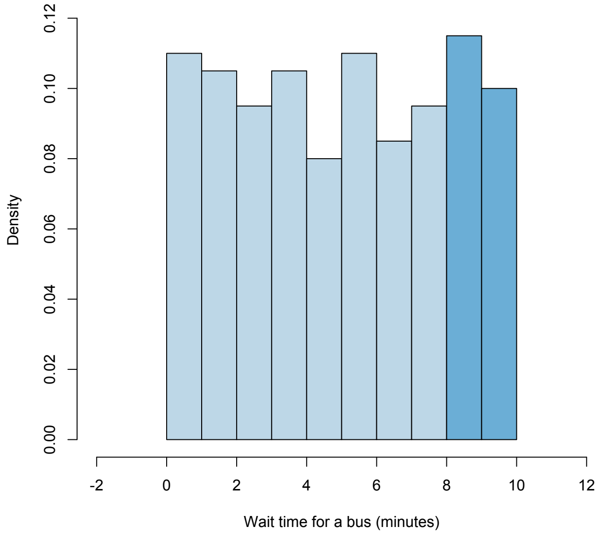

Figure 4: Estimated probability of waiting more than 8 minutes

So we estimate that the probability that our rider waited more than 8 minutes is 0.215. We can approximate this density histogram by a smooth curve to capture the overall shape of the distribution of wait times.

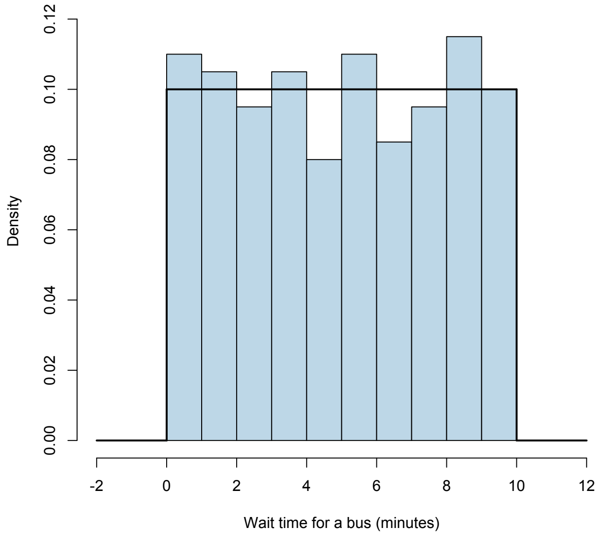

Figure 5: Density histogram with uniform probability density function

Smoothing this density histogram gives us a mathematical model for the probability density function for our bus rider's wait times. That smoothed curve would be the $Uniform[0, 10]$ probability density function, which we discussed above. We can then use this model to estimate the probabilities that the wait times are in any interval. In this way, we are thinking of our data as actual observations of a continuous random variable and the probability distribution of this continuous random variable is described by this density function. So the density function is a mathematical model, telling us how individual observed values of the data behave, in terms of the probability that an individual value is in a certain range.

Example

The birthweight of a baby is another continuous random variable but it does not have a uniform distribution. Here is a histogram of the birth weights of 200 male, full term babies at a hospital.

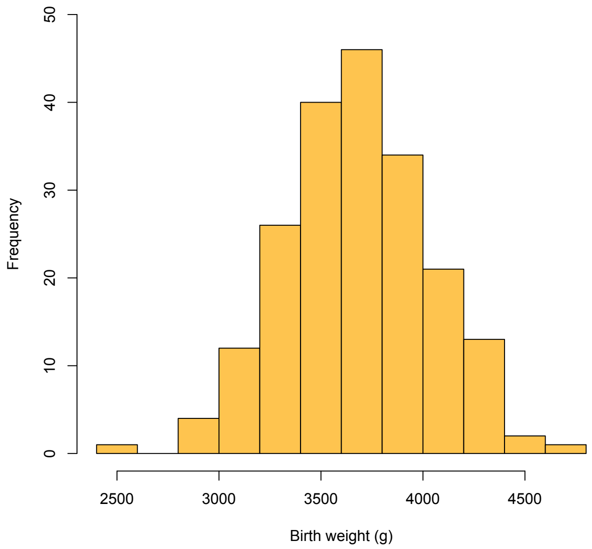

Figure 6: Histogram of the birth weights (counts)

We will change the scale of the histogram so that the total area is 1.

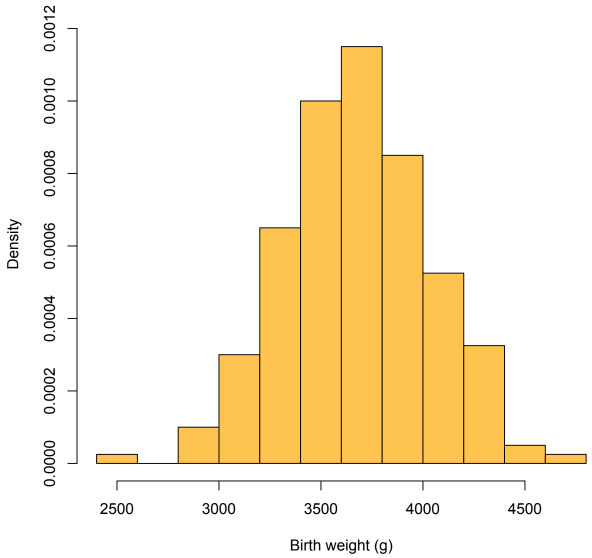

Figure 7: Density Histogram of the birth weights

This distribution has a familiar shape: symmetric, unimodal and bell-shaped. Approximating this with a smooth curve gives a mathematical model for the probability density function of birth weights. In this case it would described well by a bell curve.

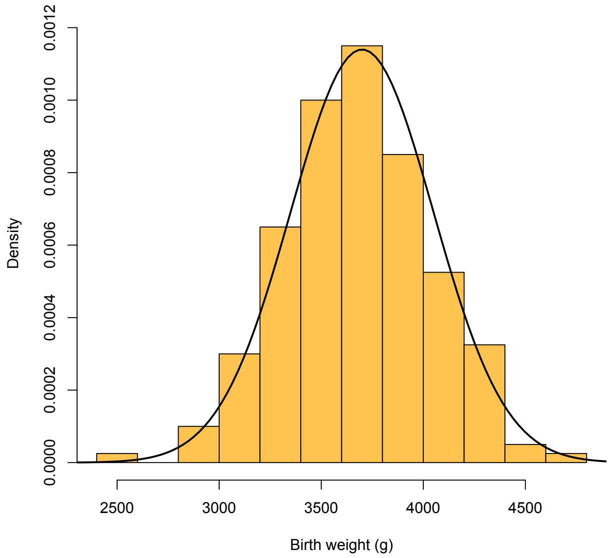

Figure 8: Density Histogram with a smooth probability density function

This smoothed curve is our density function and a probability density function with this shape is called the **Normal** density. Mathematically, this probability densify function is more complex than a uniform density function, but probabilities are calculated the same way, as areas under the density curve. So if we wanted the probability that a male full-term baby born at this hospital had a birth weight between 4000 and 4500 grams, we would find the area of this shaded region.

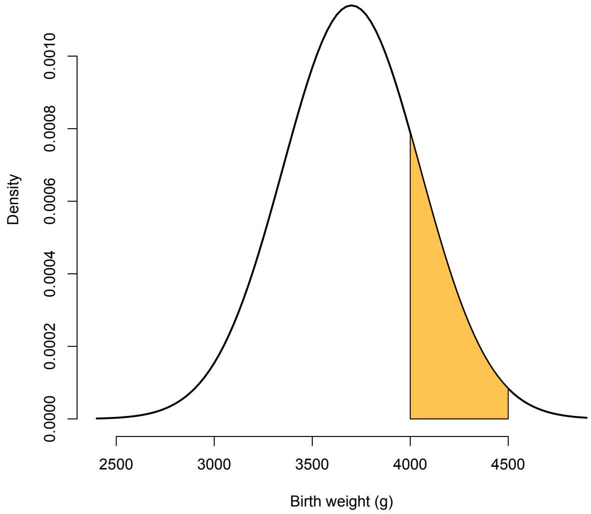

Figure 9: Probability of birth weight between 4000 and 4500 grams