Probability: Random Variables

Introduction to Probability: Random Variables

Often when we make careful measurements of the same thing on more than one occasion, it is very rare that we get the same values. It could be because our measurements are made on samples, and different samples give different values, or because our measurements only give us an approximation to what we are trying to capture. Statistical methodology was developed to allow us to learn from noisy measurements. In order to do this, we need mathematical models for the randomness that results in the noise. Therefore to understand statistical methods, we need to understand some probability.

Probability theory is what makes statistical methods work. Probability models give us mathematical ways to describe how this randomness occurs and allow us to describe the variability that occurs in our data.
Random results from data can occur because:

- Samples differ from the populations from which they were randomly drawn, and from one random sample to another drawn from the same population.
- The error in some measurements can be modeled as random noise.
- In experiments, subjects have been randomly assigned to treatments.

We use probability to quantify the uncertainty of a random circumstance; that is, a process with random outcome.

The probability models for some random experiments are well-known.
Suppose we flip a typical coin 10 times. The coin is Heads on one side, and Tails on the other and probability of each is 0.5.

Probability model for 1 flip: H with probability 0.5*, T with probability* 0.5

But that does not mean we will get exactly 5 heads every time we flip the coin 10 times. We can get 6 heads in our 10 flips. If we did this again, we might get 4 or 5 or 7 and even 0 or 10 heads (although those extreme values are less likely).

Sometimes we perform random experiments without having background knowledge of the outcome we expect. Instead of flipping a coin, we can flip a beer cap that is red on one side and silver on the other. A beer cap has an irregular shape, so we don't know which side is more likely to come up, and by how much.

Probability model for 1 flip: Red with probability ?? , Silver with probability ??

We flip the beer cap 10 times, and we get 4 reds and 6 silvers. We could use this to estimate

a probability model for the probability of getting red.

Estimate of probability of Red is 0.4

But are we convinced that the true probability is 0.4? These are the types of statistical questions that we need probability to answer.

Some other examples of measurements with random outcome:

- The number of reds we get when we flip a beer cap 10 times
- The birth weight of the next baby born in a hospital
- The number of babies born at the hospital today that are boys
- How long a rider needs to wait at a bus stop until the next bus arrives

These can all be described by a numerical random quantity, that is called a **Random variable**. A random variable takes an outcome from a random experiment and gives it a numerical value. Typically random variables are denoted by upper-case letters, later in the alphabet. So we can define:

- V = The number of reds we get when we flip a beer cap 10 times
- W = The birth weight of the next baby born in a hospital
- X = The number of babies born at the hospital today that are boys
- Y = How long a rider needs to wait at a bus stop until the next bus arrives

Suppose now we are interested in the probability that we need to wait more than 5 minutes for the next bus.

$$P(Y > 5) = ?$$

To be a valid probability, $P(Y > 5)$ must be a number between 0 and 1.
If it has probability 1, it is sure to occur .
If it has probability 0, it is sure not to occur.
Also note, the probability of waiting 5 minutes or less is

$$P(Y \leq 5) = 1 - P(Y > 5)$$

In general, the probability that something does not happen, is 1 minus the probability that it does happen.

Random variables are **independent** if knowing the value of one does not affect the probability of the other.
Example of independent random variables: The outcome on the first and the outcome on the second roll of a die.
Example of random variables that are not independent: The birth weight of a baby and the

baby's sex (0 for boys, 1 for girls) (since if we know the birth weight of a baby, that may give us some information about whether that baby is a boy or a girl, as, on average, baby boys tend to be a little larger than newborn girls).

Random variables can be **discrete** or **continuous**. A discrete random variable can take one of a countable list of distinct values.
Examples:

- The number of times the red side comes up when we flip a beer cap 10 times
- The number we get when we roll a die
- The number of people in line when you arrive at the coffee shop.

A continuous random variable can take any value in an interval (or collection of intervals).
Examples:

- Birth weights of babies
- The wait time for the next bus.

We need to be able to distinguish between situations where our measurements can be thought of as discrete or continuous, so that we can choose an appropriate probability model.