2 Random Variables and Probability

2.1 Note

At a number of points in this lecture and the following lecture, ideas will be expressed from what is known as a "frequentist" or "limiting relative frequency" viewpoint. There is another separate and arguably better system of first principals but, to avoid confusion, we will stick to that which is most commonly used at present.

2.2 Random variables

Suppose, for example, we are interested in individuals or items which belong to a certain group (e.g. the 25 ships listed in the example population). We call the group a *population*. Let the number of items in the population be N, where $N > 0$ and, for the moment, $N < \infty$. Suppose we will pick one item from the population. Let us choose our item in such a way that

1. We can not predict with certainty which item will be chosen.
2. Every item is "equally likely" to be chosen.

By 2 we mean the following. If we repeat the process m times and if the number of times we pick a particular item is f, then $f/m \to 1/N$ as $m \to \infty$. (Here f is the *frequency* of choosing the item and f/m is the *relative frequency*).

We call choosing in this way "choosing *at random*".

Suppose we will observe X, a particular characteristic of the chosen item, e.g. the tonnage of the ship. Choosing an item and observing X is an *experiment*. (Any activity which involves making an observation is an experiment). If the value of X depends on which item is chosen then X is a *random variable*.

2.3 Expectation

Suppose X is a quantitative variable. We define the *population mean* of X to be the arithmetic mean of the values of X taken by the items in the population. The *expected value* or *expectation*, $\mathrm{E}(X)$, of X is defined as follows. If we repeat m times the process of choosing at random from the whole population and observing X, then $\mathrm{E}(X)$ is the limit as $m \to \infty$ of the arithmetic mean of the observed values of X. Clearly this is the same as the population mean. If x_i is the value of X for item i in the population and that item is observed f_i times out of m then

$$\begin{aligned} \mathrm{E}(X) &= \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{N} x_i f_i \\ &= \sum_{i=1}^{N} \lim_{m\to\infty} x_i \frac{f_i}{m} \\ &= \frac{1}{N} \sum_{i=1}^{N} x_i \end{aligned}$$

2.4 Example

Consider the following experiment. "Pick a ship at random from the example population and determine the crew size, X, of the chosen ship." Here X is a discrete random variable. Let x_i be the crew size of ship i and f_j be the number of ships in the population with crew size j. Then

$$\mathrm{E}(X) = \frac{1}{25} \sum_{i=1}^{25} x_i = \sum_{j=-\infty}^{\infty} j \frac{f_j}{25} = \frac{320}{25} = 12.8.$$

Similarly, if $Y = X^2$, then

$$\mathrm{E}(Y) = \frac{1}{25}\sum_{i=1}^{25} x_i^2 = \sum_{j=-\infty}^{\infty} j^2 \frac{f_j}{25} = \frac{5932}{25} = 237.28.$$

2.5 Expectation of sums

Let X and Y be two quantitative random variables. Let a and b be two constants. Then

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y).$$

E.g. a tanker carries a cargo made up of oil of two grades. If X and Y are the volumes of the two grades loaded and a and b are the densities (assuming we know the densities) then $aX + bY$ is the total weight of oil.

2.6 Infinite population

For statistical purposes we would normally observe a *sample* of more than one item. If the population size N is finite then removing an item alters the conditions under which the next item is chosen. However, in practice, the population size is sometimes so large compared to the sample size that the effect is negligible. For example, if we measured the shoe sizes of 100 individual people from a population of ten million, the relative frequencies, f_i/n, in the population would be hardly affected by the removal of the sample of 100. In cases like this, we may, as an approximation, consider the population size to be infinite, in which case the behaviour is as if we replaced items after observing them.

In other cases, e.g. experimental error or other "random", i.e. unpredictable, behaviour, we use the notion of an "infinite population" when no physical population really exists. For example, consider the experiment, "Toss a coin five times and count the number of heads." We can repeat this as many times as we like without affecting the behaviour of the next experiment. We can consider there to be an "infinite population" of possible repetitions of this experiment from which we pick one out at random each time we do the experiment.

2.7 Events

An experiment may have several possible *outcomes*. E.g., if we select a ship from the example population and record the crew-size, we may regard observing a particular crew-size, e.g. 5, as an outcome (or alternatively we might regard selecting a particular ship, e.g. number 3, as an outcome if we record this). Outcomes are also called *elementary events*. The outcome contains all of the information which that experiment can give us. One and only one outcome can occur each time we do the experiment.

The set of all possible outcomes is called the *sample space*. For example, the set of all possible crew-sizes might be a sample space.

A collection of one or more outcomes, i.e. a subset of the sample space, is called an *event*. For example, choosing a ship with a crew-size of at least 5 is an event. This event does not tell us exactly what the crew-size is, whereas the outcome would.

2.8 Probability

Let A be an event. Define the variable I_A as follows.

$$\begin{array}{ll} I_A = 1 & \text{if } A \text{ occurs.} \\ I_A = 0 & \text{if } A \text{ does not occur.} \end{array}$$

We call I_A the *indicator* for A.
Then

$$\Pr(A) = E(I_A)$$

is called the *probability* of A.

For example, if C is the event "choose a ship with crew-size of at least 5", then $\Pr(C) = 19/25 = 0.76$, the proportion of the ships with crew-sizes of at least 5.

Clearly, for any event A,

$$0 \leq \Pr(A) \leq 1.$$

Impossible events will have probability 0 and events which are certain to occur will have probability 1. If S is the sample space then $\Pr(S) = 1$.

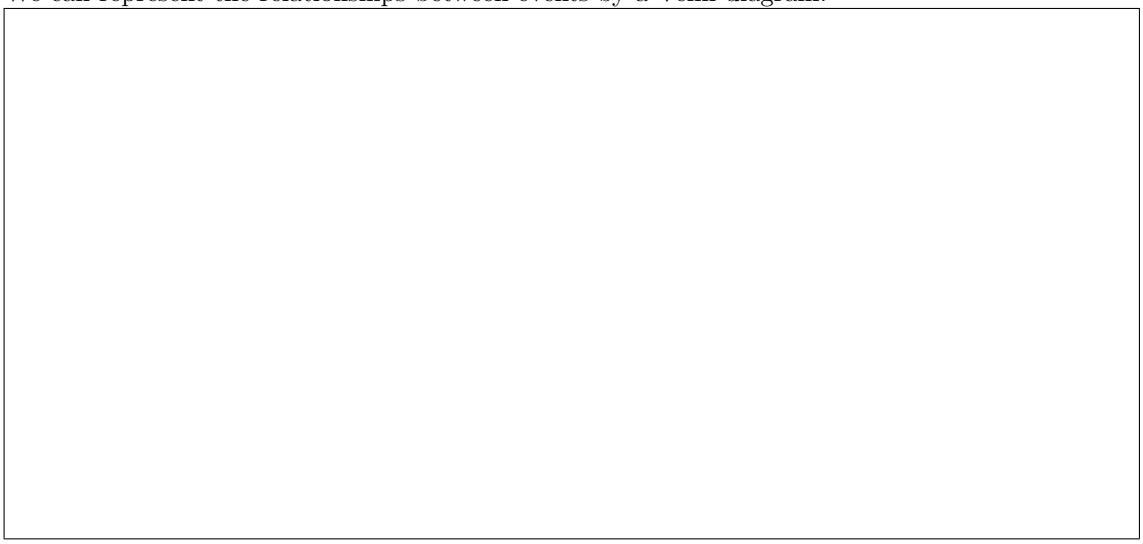
Let A be an event. Write $\bar{A}$ for the event "not A", i.e. "A does not happen". Then clearly $\Pr(\bar{A}) = 1 - \Pr(A)$.

2.9 Example

Suppose we pick a ship at random from our example population and observe the variables listed. Let A, B and C be events defined as follows.

A : "The ship is known to be steam-powered."
B : "The ship's tonnage is less than 100. "
C : "The ship's crew-size is greater than 4."

We can represent the relationships between events by a Venn diagram.



We can see that

$$\begin{aligned} \Pr(A) &= \frac{16}{25} = 0.64 \\ \Pr(B) &= \frac{5}{25} = 0.2 \\ \Pr(C) &= \frac{19}{25} = 0.76 \end{aligned}$$

2.10 Addition of probabilities

If U and V are two events then $U \wedge V$ is the event "U and V" and $\Pr(U \wedge V)$ is the probability that both U and V occur. In the example

$$\begin{aligned} \Pr(B \wedge C) &= 0 \\ \Pr(A \wedge B) &= 0 \\ \Pr(A \wedge C) &= \frac{16}{25} = 0.64. \end{aligned}$$

If two events U and V can not both occur, i.e. $\Pr(U \wedge V) = 0$, then U and V are said to be *mutually exclusive.* In the example A and B are mutually exclusive and B and C are mutually exclusive.

If U and V are two events then $U \vee V$ is the event "U or V", i.e. "at least one of U, V occurs", which has probability $\Pr(U \vee V)$. In the example

$$\Pr(A \vee B) = \frac{21}{25} = 0.84.$$

Clearly in general, bearing in mind the need to avoid counting outcomes twice,

$$\Pr(U \vee V) = \Pr(U) + \Pr(V) - \Pr(U \wedge V).$$

This is called the *generalised addition rule for probabilities.*
In the example

$$\Pr(A \vee C) = \frac{16}{25} + \frac{19}{25} - \frac{16}{25} = \frac{19}{25} = 0.76.$$

The identity

$$\Pr(U \vee V) = 1 - \Pr(\bar{U} \wedge \bar{V})$$

is often useful, where $\bar{U} \wedge \bar{V}$ is the event "neither U nor V". In the example

$$\Pr(B \vee C) = 1 - \frac{1}{25} = \frac{24}{25} = 0.96.$$

Clearly, if two events X, Y are mutually exclusive then $\Pr(X \vee Y) = \Pr(X) + \Pr(Y)$. In the example

$$\Pr(A \vee C) = \frac{16}{25} + \frac{5}{25} = \frac{21}{25} = 0.84.$$

2.11 Problems

1. Let X be the tonnage of a ship chosen at random from the example population. Find the expected value of X.

2. Let I_A, I_B, I_C be variables defined as follows.

   - If a ship is known to be sail powered, then $I_A = 1$, otherwise $I_A = 0$.
   - If a ship has a tonnage greater than 500, then $I_B = 1$, otherwise $I_B = 0$.
   - If a ship has a crew size greater than 5, then $I_C = 1$, otherwise $I_C = 0$.

   Find the expected values of I_A, I_B and I_C for a ship chosen at random from the example population.

3. Let I_A, I_B, I_C be as in Question 2. Find the expected values of

   (a) $I_A I_B$,
   (b) $I_A I_C$,
   (c) $I_B I_C$,
   (d) $I_A + I_B - I_A I_B$,

   for a ship chosen at random from the example population.

4. A fair coin is equally likely to show "heads" and "tails" when tossed. What would you say is the expected value of the total number of "heads" obtained in five tosses?

5. Define a suitable sample space for the variables in question 1 of lecture 1.

6. A ship is to be chosen at random from the example population. Find the probabilities for the following events.

   (a) The ship is known to be sail powered.

   (b) The ship has a tonnage greater than 500.

   (c) The ship has a crew size greater than 5.

7. A fair coin is equally likely to show "heads" or "tails" when tossed. What is the probability that "heads" is obtained

   (a) once only,

   (b) at least once,

   when a fair coin is tossed twice?

References

Floud, R., 1973. *An Introduction to Quantitative Methods for Historians*, 2nd edition. London: Methuen.