Data preprocessing and sampling

Part 3: Probability primer

By: Noureddin Sadawi, PhD

University of London

What is probability?

- Probability may be viewed as the numerical measure of likelihood or chance that an event will occur.
- We are interested in defining what is meant by the probability of an event *A*, that is, the likelihood of *A* occurring denoted by *P(A)*.
- A natural question to ask is: what values can probability take or how do we measure probability?

Probability scale

- Probability values are assigned on a scale from 0 to 1.
- A probability value near 0 indicates an event is unlikely to occur.
- A probability value of 0.5 indicates an event is as likely as it is unlikely to occur.
- A probability value close to 1 indicates an event is very likely to occur.
- A probability value of 0 indicates an **impossible event** while a probability value of 1 indicates a **certain event**.

Assigning probabilities

- We wish to look at how probabilities can be assigned to experimental outcomes.
- We will consider two approaches: the classical approach and the relative frequency approach.
- Regardless of the approach used, there are two basic requirements for assigning probabilities.

Assigning probabilities

- Let B_i denote the *i*th experimental outcome and $P(B_i)$ its probability.

$$0 \leq P(B_i) \leq 1, \qquad for\ all\ i. 0 \leq P(B_i) \leq 1$$

$$, \qquad for\ all\ i.$$

Assigning probabilities

- The sum of probabilities of all experimental outcomes is equal to 1. For *n* experimental outcomes, this can be written as

$$P(B_1) + P(B_2) + \cdots + P(B_n) = 1$$

Assigning probabilities

- The sum of probabilities of all experimental outcomes is equal to 1. For *n* experimental outcomes, this can be written as

$$P(B_1) + P(B_2) + \cdots + P(B_n) = 1$$

The classical approach

- This method is appropriate for assigning probabilities when there are a finite number of experimental outcomes, each of which are equally likely.
- If there are *n* possible experimental outcomes, then a probability of *1/n* is assigned to each experimental outcome.
- Note that the two basic requirements for assigning probabilities are automatically satisfied using this approach.

Example:

Consider the experiment of tossing a six-sided die. There are six possible outcomes each of which we can assume are equally likely provided that the die is a fair die. That is, there is no physical evidence to suggest it favours one side more than the others.

The sample space is

$$S = \{1,2,3,4,5,6\}$$

and

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.$$

The relative frequency approach

- This approach is appropriate when data are available to estimate the proportion of time the experimental outcome will occur if we repeat the experiment a large number of times.
- It is often used to determine probabilities for a particular population.

The relative frequency approach

- The relative frequency probability is the limit of the proportion of times event *A* occurs in a large number of *n* trials. Its given by:

$$P(A) = \frac{n_A}{n} \; P(A) = \frac{n_A}{n}$$

where n_A n_A denotes the number of outcomes for A and n denotes the total number of trials.

Example

- Consider a study of waiting times in the X-ray department for a local hospital.
- A clerk recorded the number of patients waiting for service at 10 p.m. on 20 consecutive days and obtained the following results.

Number waiting	Number of days outcome occurred
0	1
1	5
2	7
3	4
4	3
Total	20

Example

- From the table, we see that three patients were waiting on four days. So using the relative frequency approach, we assign a probability of:

$$\frac{4}{20} = 0.2$$

- To the experimental outcome of three patients waiting for service.