NRES 798 — Statistical Methods for Ecologists
Chapter 1: Introduction to Probability

Oscar García

February 9, 2013

Contents

1 What

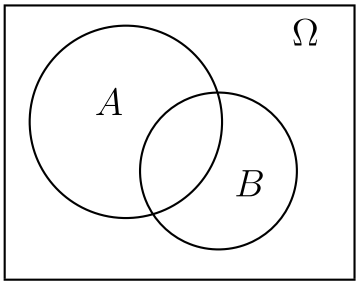
A real-world system may be modelled as a (random) *experiment* or *trial*. The set of possible *outcomes* or *sample points* is the *sample space*. The sample space may be *discrete*, if the number of sample points is countable (finite or countably infinite), or *continuous* if not[1]. A set of sample points is called an *event*. A probability is a function that maps events into real numbers, and that satisfies three axioms. If Ω is the sample space, P is the probability, and A and B are events, the axioms are

Axiom 0: $P(A) \geq 0$.

[1] Hybrids of the two are possible, but not all that common.

Axiom 1: $P(\Omega) = 1$ [2].

Axiom 2: $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint (i.e., if $A \cap B = \emptyset$, where $\emptyset$ is the empty set).



Reminder (visualize in the Venn diagram): $\cup$ is **u**nion, the points that belong to any of the two sets, $\cap$ is intersection, the points that belong to both sets. Another operation is the set difference $A \setminus B$, the points of A that are not in B. The complement $\bar{A} = \Omega \setminus A$ denotes the points outside A; other notations are A^C or A'. The symbols $\subset$ or $\subseteq$ mean "is a subset of", and $\in$ means "is an element of".

Other properties can be derived from the axioms:

$$P(A) \leq 1\,, \quad P(\bar{A}) = 1 - P(A)\,, \quad P(\emptyset) = 0\,, \quad \text{etc.}$$

You should be able to figure out all this from looking at Venn diagrams, no need to memorize. Sometimes a different set of axioms is chosen, and some of those above are taken as derived properties[3].

[2] The wording in the textbook, "The sum of all the probabilities of outcomes within a single sample space = 1.0", is true for discrete sample spaces. In a continuous sample space those probabilities are 0, and the sum of infinite zeroes is undefined. There is a similar problem with the formulation of Axiom 2.

[3] More advanced treatments talk of a *probability space* $(\Omega, \mathcal{F}, \mathcal{P})$ consisting of three things (a triple): Ω, the set of events $\mathcal{F}$ (set of subsets of Ω), and P. The probability is a function taking events $A \in \mathcal{F}$ into real numbers $p \in \boldsymbol{R}$, that satisfies certain axioms. In the usual notation,

$$\begin{aligned} P : \mathcal{F} &\to \boldsymbol{R} \\ A &\mapsto p \end{aligned}$$

P is a special case of a *measure*. Other examples of measures are areas, volumes, and weights, which do not necessarily satisfy Axiom 1.

For continuous Ω there are some technicalities about the sets allowed in $\mathcal{F}$, because for some weird sets, like fractals or Cantor sets, it is not possible to define a probability (they are not *mesurable*). Anyway, $\mathcal{F}$ must contain $\emptyset$ and Ω, and must include any sets that can

In applications, probability gives weights to events according to frequency, belief, or any other interpretation. A good mental model for it is the weight of portions of an object where density varies smoothly throughout. The unit of weight is scaled so that the total weight is 1. More specifically, this would correspond to a 3-dimensional continuous sample space. In 2-D one might think of a sheet made of that material, or in 1-D, a wire. In a discrete sample space, the “stuff” is clumped into discrete chunks.

Terminology: As pointed out in the textbook, in some circumstances a random experiment or trial may be called a *replicate*. It also uses the word “event” for a trial a couple of times, perhaps by mistake. For the elements of Ω, *sample point* is perhaps the less ambiguous term, although not the shortest. And in general it is not the same as a *sample*, see later. *Outcome* is commonly used, although some authors use outcome for event. The textbook uses *event*, or later *simple event*, for a sample point, and *complex event* for what is normally known as *event* (the term used here)[4]; be careful! Some refer to sample points as *atomic events*. Of course, a sample point is also a special case of event (in the usual sense), as are also $\emptyset$ and Ω.

2 Examples

1. Pitcher plants. We model the visit of an insect to a pitcher plant as a random experiment or trial with two possible outcomes: $\Omega = \{$capture, escape$\}$. There are 4 possible events: $\emptyset$, capture, escape, and Ω (sets of size 0, 1, 1, and 2, respectively; one could also have written $\{$capture$\}$ and $\{$escape$\}$). The 4th event is interpreted as “either capture or escape”.

   Assume that $P(\{\text{capture}\})$, or simply $P(\text{capture})$, is some unknown

be obtained from others through unions, intersections and complements, what is known as a *σ-algebra*. In the special case of continuous random variables, where Ω is a set of real numbers or numeric vectors, the allowable events are intervals (possibly multi-dimensional intervals), and any other sets obtainable from intervals through unions, intersections, and complements. These are called *Borel sets*, and should include everything likely to be useful in practice. In the main text above, we are just trying to be reasonably precise without being overly pedantic.

[4] Is this wrong? Not really. Does it matter? Not if this is the only thing you will ever read on the topic. In math, we are free to define anything in any way we want, provided we define it clearly. I could say “Let $\pi = 42$”, and proceed through three pages of derivations representing 42 by π. Of course, likely that would confuse the hell out of everybody. But strictly speaking, it would not be wrong, just bad manners.

number θ. Then $P(\text{escape}) = 1 - \theta$, and obviously $P(\emptyset) = 0$ and $P(\Omega) = 1$.

2. Two pitcher plant visits. Using pairs to show the outcomes of the first and second visits, the sample space for this composite experiment has 4 outcomes:

$$\Omega = \left\{ \begin{array}{cc} (\text{capture}, \text{capture}) & (\text{capture}, \text{escape}) \\ (\text{escape}, \text{capture}) & (\text{escape}, \text{escape}) \end{array} \right\}.$$

   There are 16 possible events (list them!). For instance, the event "one capture" = {(capture, escape), (escape, capture)}.

   If the outcome from the second visit is not affected by what happens in the first one, the probability of a composite outcome equals the product of the probabilities for the individual visits (*independence*, more on this later). The probabilities for the outcomes are then

$$\begin{array}{cc} \theta^2 & \theta(1-\theta) \\ (1-\theta)\theta & (1-\theta)^2 \end{array}$$

   The probabilities for other events can be worked out from Axiom 2.

3. Milkweeds and caterpillars. In a certain location we can find a population of milkweed that is resistant to caterpillars (R), one that is not resistant ($\sim R$), or there may be no milkweed (N). In the same location one can find caterpillars (C), or no caterpillars ($\sim C$). For the composite of combined occurrences we have the sample space

$$\Omega = \left\{ \begin{array}{cc} (R, C) & (R, \sim C) \\ (\sim R, C) & (\sim R, \sim C) \\ (N, C) & (N, \sim C) \end{array} \right\}.$$

   An example of an event other than these "simple events" is the presence of resistant milkweed, $\{(R, C), (R, \sim C)\}$.

   Assume that the probabilities of milkweed occurrence are $P(R) = 0.2$, $P(\sim R) = 0.8$, $P(N) = 0$. For caterpillars, $P(C) = 0.7$, $P(\sim C) = 0.3$. Initially, the milkweed and caterpillars disperse independently, so that the probability of a combination is the product of the probabilities of the components, giving the composite outcome probabilities

$$\begin{array}{cc} 0.14 & 0.06 \\ 0.56 & 0.24 \\ 0 & 0 \end{array}$$

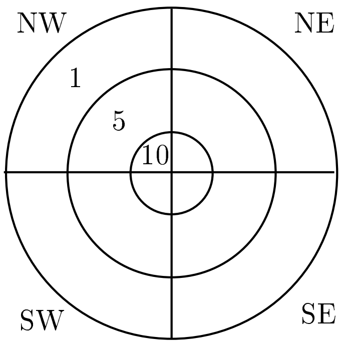
The probability of occurrence of resistant milkweed is found to be $0.14 + 0.06 = 0.20$, agreeing with the original $P(R)$.

Suppose that after that, the caterpillars eat all the non-resistant milkweed. The probabilities change to

0.14	0.06
0	0.24
0.56	0

The probability of finding any milkweed is now $0.14+0.06+0+0.24 = 1 - (0.56 + 0) = 0.44$.

4. Throwing a dart. We throw a dart at this dartboard:



Only throws where the dart sticks to the dartboard are valid[5]. Assuming that any point can be hit, the sample space Ω (in the model!) is now continuous, consisting of all the points within the outer circle[6]. Example events are hitting inside the central circle (10), or hitting the 5-ring within the 2nd quadrant, which might be coded as (5, NW). Note that this last event can be seen as the intersection of the event "5-ring" and the event "NW quadrant".

On discrete sample spaces, it is true that events with probability 0 will never happen, and events with probability 1 are sure to happen. This is not necessarily true if the sample space is continuous. In the darts example, any individual point has probability zero. But when the dart is thrown, it hits some point, which had probability zero. Similarly,

[5] If you must have biological examples, translate to a seed falling on a field.

[6] The whole wall would also be an acceptable sample space, with the area outside the dartboard having probability 0 (hitting there does not count).

the rest of the board excluding that point had probability 1, but it was not hit. Probabilists are careful in speaking of *sets of probability zero*, which are not necessarily impossible, and of things happening *almost surely*, meaning with probability 1 [7].

3 Why

The way probability is used in Statistics goes more or less like this: A (probabilistic, aka *stochastic*) model of the real system is made up, like in the examples above. If not already numerical, the outcomes are coded as numbers called random variables (RVs). Said slightly differently, an RV is a function mapping outcomes into numbers (or into numeric vectors in multivariate statistics). For instance, dart throws could be described by their Cartesian (x, y) or polar (r, α) coordinates. Note that functions of RVs are also RVs. The word *distribution* is normally used for the probability model of an RV. Then, some kind of *sample* is considered, where hypothetical observations of the model RVs are made, usually from a sequence of repetitions of the experiment or trial. The sample is a set (or vector, or table) of RVs; each of these may correspond to the RV in the original model, or may be another RV derived from it. Probability theory is used to derive the distribution of the sample (a probability model), starting from the original system model. The next step is to choose a *statistic*, that is, some function of the sample, which is therefore also an RV. Probability theory is then used again to obtain the distribution of this statistic, called its *sampling distribution.* This sampling distribution is then used for making inferences about the unknowns in the model.

We demonstrate with the pitcher plants example. Define an RV that is 1 for *capture*, and 0 for *escape.* Assume that in one day we make $m = 1000$ visits, and we observe the outcomes. A result would look like this: $(0, 0, 0, 1, 0, 0, 1, 0, \ldots, 0)$, a vector of length m. It is assumed that the visits are independent, the result from a visit is not affected by the results in other visits.

[7] Weird things can happen when dealing with infinities and infinitesimals. Such models can be convenient (remember, they are only models), but one has to tread carefully. Mathematicians of the *Constructivism School* insist that everything should be done in finite terms, possibly passing to a limit at the end when all else has been done (http://en.wikipedia.org/wiki/Finitism). That may be too much work, however, and professional mathematicians enjoy showing off their skills in avoiding potential pitfalls.

Instead of recording all this, we decide to record only the number of captures X, that is, the number of ones, or the sum of the vector elements[8]. We need the distribution of X, that is, its probability model consisting of a sample space and a probability function. In m visits we can observe anywhere between 0 and m captures, so that the sample space of X is $\Omega = \{0, 1, 2, \ldots, m\}$. It is shown in the next chapter that in this instance the probability of an outcome x is given by a *binomial* probability density

$$P(X = x) = \binom{m}{x} \theta^x (1 - \theta)^{1-x} \ .$$

This is commonly written $X \sim \text{Binomial}(m, \theta)$, or something like that, read as "X is distributed as a binomial with parameters m and θ".

An observation is made every week over one year, so that one collects 52 values "from" that binomial distribution. The mean $\bar{X}$ of these $n = 52$ numbers[9] (a statistic) is used as an estimate of θ. Note that $n\bar{X}$ is simply the number of captures in $mn = 52000$ visits, so that $nX \sim \text{Binomial}(mn, \theta)$ (sampling distribution). This can be used to justify the use of the sample mean to estimate θ, and to compute standard errors, confidence limits, or to test hypotheses.

4 More properties

4.1 Unions

Axiom 2 says that if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$. What if the intersection is not empty? Look at the Venn diagram above (you may think of that as hitting parts of the dartboard in Example 4). It is seen that on adding the points in A and the points in B, the points in the intersection are counted twice. Therefore, correcting for the double counting,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \ , \tag{1}$$

generalizing Axiom 2.

[8] It can be shown that the order of the 0's and 1's does not provide any information on the unknown θ, the count is a *sufficient statistic*.

[9] The bar notation is commonly used for sample means, do not confuse with set complement.

4.2 Conditional probability, independence

$P(A|B)$ denotes the probability of an outcome being in A, knowing that it is in B. Think of this in the Venn diagram and/or dartboard: The sample space is first restricted to B, which contains all the relevant outcomes. P must be divided by $P(B)$, so that the new probability adds to 1 over the new sample space. The relevant part of A is $A \cap B$. Therefore, the *conditional probability* is defined as

$$P(A|B) = \frac{A \cap B}{P(B)} \ . \tag{2}$$

Solving for $A \cap B$, one can write the *joint probability* as

$$P(A \cap B) = P(A|B)P(B) \ . \tag{3}$$

An event A is said to be *independent* of B if $P(A|B) = P(A)$. That is, knowledge of B does not change our knowledge of A. If A is independent of B, then equation (3) becomes

$$P(A \cap B) = P(A)P(B) \ ,$$

which is often taken as an alternative definition of independence. From this, it is clear that if A is independent of B, then B is independent af A, so that we can simply say that A and B are independent[10].

Independence is most common in models of composite experiments or of sampling. In example 3, the events "resistant milkweed" $= \{(R, C), (R, \sim C)\}$ and "no caterpillars" $= \{(R, \sim C), (\sim R, \sim C), (N, \sim C)\}$ are independent. In the previous Section, the numbers of captures observed in different days are (assumed to be) independent. The results of throwing two darts (or the same dart twice) might be assumed to be independent. Occasionally, independence may be built into a non-composite model; for example, in the darts model we might assume that variability along the x-axis is unrelated to variability along the y-axis.

4.3 Bayes Theorem

From (3),

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \ .$$

[10] This is not exactly the same as causal, physical, or other kinds of independence. What happens today may depend on what happened yesterday, but usually not the other way around. Statistical independence is about joint (although not necessarily simultaneous) occurrence $P(A \cap B)$.

Solving in the last two terms for $P(A|B)$, we obtain (one form of) *Bayes Theorem*:

$$P(A|B) = \frac{P(B|A)}{P(B)} \; . \tag{4}$$

The *marginal probability* $P(B)$ is sometimes included in the theorem in other forms. For instance, in the textbook

$$P(B) = P(B|A) + P(B|\bar{A})$$

(check on the Venn diagram).

The theorem works for any definition of probability, but it is most often used with subjective probabilities in Bayesian inference (hence the term *Bayesian*). There, A corresponds to parameters, and B to data. A probability for parameters makes no sense in the frequentist interpretation.