A brief review of basic probability theory*

Winfried Just†

December 23, 2015

The spread of infectious diseases is inherently a stochastic process and the materials posted at this website[1] heavily rely on probability theory. Here we review some basic concepts of probability theory for easy reference. The material is restricted to notions that are used in teaching materials posted here or in our book chapters [1, 2]. While this posting cannot replace a regular textbook on probability, it can serve as a short refresher course.

1 Modeling stochastic systems: sample spaces and events

We are interested in studying the probability of random events that may occur in a natural system. In order to do this in a mathematically rigorous way, one first needs to define a *sample space* Ω that comprises all *elementary outcomes* that could possibly be observed. The choice of a suitable sample space depends both on the natural system that we want to study and on the particular questions we are interested in.

Let us consider an example from the study of the spread of infectious diseases. Consider a very small population of $N = 3$ *hosts* (humans, animals or plants) and a disease from which hosts never recover. Assume also that the disease can be both introduced from outside the population and transmitted from host to host within the population, and that currently all three hosts are *susceptible* to the disease, which means that they are neither immune nor already infected. Assume furthermore that the disease gets transmitted by direct contact between an *infectious* and a *susceptible* host during which a sufficient number of *pathogens* get transferred so that the formerly susceptible host becomes infectious.

If we are only interested in which of the hosts will be infectious exactly one year from now, then we could define

$$\Omega_1 = \{SSS, SSI, SIS, SII, ISS, ISI, IIS, III\}, \tag{1}$$

where a letter S in position i indicates that host i will still be susceptible one year from now and a letter I in position i indicates that host i will be infectious at that time.

†Department of Mathematics, Ohio University, Athens, OH 45701 E-mail: mathjust@gmail.com

[1] `http://www.ohio.edu/people/just/IONTW/`

The elements of a sample space are called *elementary outcomes* (of an "experiment" run by nature or perhaps a computer simulation). In Ω_1 they are represented by letter sequences. We want to conceptualize the elementary outcomes in such a way that they will give us all the information that we are interested in. The elementary outcomes in Ω_1 certainly tell us who will and who will not be infectious one year from now, but Ω_1 would be inadequate for modeling the time course of an outbreak in the population. For example, let T_2^I denote the time when host 2 becomes infectious and let time be measured in years. In our interpretation of the sample space Ω_1 above, the elementary outcome SII allows us to deduce that the inequality $T_2^I \leq 1$ holds, but it does not allow us to pinpoint T_2^I with any greater accuracy. For modeling the time course of an outbreak with arbitrary precision, a better choice for the sample space would be

$$\Omega_2 = [0, \infty)^3 = \{(x, y, z) \in \mathbf{R}^3 : \ 0 \leq x, y, z\}. \tag{2}$$

An elementary outcome $(x, y, z) \in \Omega_2$ indicates that the times of onset of infectiousness of the three hosts are $T_1^I = x, T_2^I = y$, and $T_3^I = z$.

Thus we have some flexibility in choosing the set of elementary outcomes, but there are some restrictions. We need to make sure that the sample space comprises *all possible* observations. This is not always obvious. For example, Ω_2 will be a *bona fide* sample space only if it is certain that every one of the three hosts *will* eventually become infectious. This is rather dubious in the real-world examples, but for our purpose of illustrating basic concepts we may pretend it were true.

We also need to make sure that no two different elementary outcomes could possibly occur simultaneously. This will be the case for Ω_1 and Ω_2 as defined above. But suppose we had tried instead to conceptualize our sample space as $\Omega^* = \{S_1, I_1, S_2, I_2, S_3, I_3\}$, where Si signifies that host i will still be susceptible one year from now and Ii signifies that host i will be infectious at that time. But, for example, S_1, I_2, S_3 could all occur simultaneously. Thus Ω^* would not be a valid option for a sample space.

However, I_1 as defined above is certainly a possible *event* that could occur during an outbreak, and the same is true for the other elements of Ω^*. In probability theory events are conceptualized as subsets of the sample space that comprise all outcomes that are *favorable* to the event. For example, in terms of Ω_1 the event I_1 would be treated as the subset $I_1 = \{ISS, ISI, IIS, III\} \subset \Omega_1$, in terms of Ω_2 we would have $I_1 = \{(x, y, z) \in \Omega_2 : \ x \leq 1\}$. Treating events as sets takes some getting used to. It is often more intuitive to think of events in terms of verbal descriptions and of the corresponding subsets of the sample space as a kind of elaborate mathematical symbols for them.

Each subset of a finite sample space like Ω_1 is considered an event and has a corresponding verbal description, albeit possibly a rather tortuous one (think how you would verbally describe the event $\{SIS, SII, ISI, SIS\} \subset \Omega_1$). This will no longer be true in infinite sample spaces like Ω_2 where some subsets simply defy any kind of description. For infinite sample spaces only so-called *measurable subsets* qualify as events, but not necessarily all subsets. We can think of measurable subsets as subsets that have a certain kind of

mathematical description.[2]

The treatment of events as subsets of the sample space allows us to translate some verbal descriptions into operations on sets. The *complement* $\bar{E}$ of an event $E \subseteq \Omega$ is the set $\Omega \backslash E$ of all elementary outcomes that are *not* favorable to E. It signifies that the event E did not occur. For example, the complement of I_1 in Ω_1 is the set

$$\bar{I}_1 = \Omega_1 \backslash \{ISS, ISI, IIS, III\} = \{SSS, SSI, SIS, SII\} = S_1.$$

Similarly, the *intersection* $E \cap F$ signifies that both events E and F occurred simultaneously, while the *union* $E \cup F$ signifies that at least one of the events E and F occurred. For example, in Ω_1 we get

$$\begin{aligned} I_1 \cap S_2 &= \{ISS, ISI, IIS, III\} \cap \{SSS, SSI, ISS, ISI\} = \{ISS, ISI\}; \\ I_1 \cap S_1 &= \{ISS, ISI, IIS, III\} \cap \{SSS, SSI, SSI, SII\} = \emptyset; \\ I_1 \cup S_1 &= \{ISS, ISI, IIS, III\} \cup \{SSS, SSI, SSI, SII\} = \Omega_1. \end{aligned} \tag{3}$$

The empty set $\emptyset$ is called *the impossible event.* The definite article is used since there is only one empty set, although the impossible event (like any other event, by the way) usually has many different verbal descriptions. The second line of (3) shows that the events I_1 and S_1 cannot occur simultaneously; they are *mutually exclusive.* This makes intuitive sense if we consider their verbal descriptions given above.

Note that the events I_1 and S_1 are complements of each other. It is true for any event E that $E \cap \bar{E} = \emptyset$ while $E \cup \bar{E}$ comprises the whole sample space.

2 Probability functions and probability measures

Assume we are given a sample space Ω and a family of subsets of Ω that we consider events. A *probability measure* is a function that assigns a *probability* $P(E)$ to each event E and has certain properties that we will discuss shortly.

The nature of the sample space imposes some restriction on how we can construct probability functions. A sample space Ω is *discrete* if all its elements can be arranged in a sequence $(e_1, \ldots, e_k, \ldots)$ of pairwise distinct entries that are indexed either by all positive integers k or by all integers $1 \leq k \leq K$, where K is the total number of elementary outcomes. The space Ω_1 as defined by (1) is an example of a discrete space, while Ω_2 as defined by (2) is not. In a discrete sample space, all subsets are considered events.

Now assume Ω is discrete and $(e_1, \ldots, e_k, \ldots)$ is an enumeration of its elementary outcomes. A *probability function* is any function f that assigns nonnegative reals to the elementary outcomes and satisfies

(f1) $f(e_k) \in [0, 1]$ for all $e_k \in \Omega$.

[2]This will be literally true for so-called *Borel measurable* subsets.

(f2) $\sum_k f(e_k) = 1$.

Condition (f1) simply asserts that the values of f are *bona fide* probabilities; the role of (f2) will become clear shortly. We want to point out that *any* function f that satisfies (f1) and (f2) qualifies as a probability function. Even if Ω is finite, the values $f(e_1)$ do not have to be all equal, and some of elementary outcomes may even be assigned probability 0.

Note that a *probability function* is defined on the set Ω of all elementary outcomes, while a *probability measure* is defined on the set of all *subsets* of Ω that are events. Given a probability function f, we can define a probability measure by assigning probabilities to events E as follows:

$$P(E) = \sum_{\{k:\ e_k \in E\}} f(e_k). \tag{4}$$

The resulting probability measure has the following properties:

(P1) If $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$.

(P2) $P(\Omega) = 1$.

Property (P2) follows immediately from Property (f2). It asserts that always *some* elementary outcome will be observed.

Property (P1) is called *additivity*. In particular, the complement $\bar{E}$ of an event E and the event E itself are always mutually exclusive, so that (P1) and (P2) together imply that $P(E \cup \bar{E}) = P(\Omega) = 1$, or, equivalently,

$$P(\bar{E}) = 1 - P(E). \tag{5}$$

The complement $\bar{\Omega}$ of the whole sample space is the empty set $\emptyset$. It follows from (5) that

$$P(\emptyset) = 1 - P(\Omega) = 1 - 1 = 0, \tag{6}$$

as befits the impossible event. Since we allowed 0 as the a legitimate value of the probability function f, it may not be the case though that $\emptyset$ is the *only* event that occurs with probability zero,

It is important to realize that $P(E \cup F) = P(E) + P(F)$ will hold *only* if $P(E \cap F) = 0$. By (6), this will be true if E and F are mutually exclusive as in $(P1)$, but in the general case we will need to use the formula

$$P(E \cup F) = P(E) + P(F) - P(E \cap F). \tag{7}$$

If the sample space Ω is not discrete (as in our example Ω_2 of a *continuous* sample space), then we may have a situation where *each* elementary outcome occurs with probability 0. In this case we can no longer define probability measures in terms of probability functions. The construction of probability measures for sample spaces like Ω_2 requires more advanced

mathematical tools than we need here. Let us only mention that the resulting probability measures are functions that satisfy a stronger version of Property (P1) called *countable additivity* in addition to Property (P2). In particular, Equations (5)–(7) remain valid for continuous sample spaces.

3 Conditional probabilities and independence

The previous section described mathematical constructions of probability measures and some of their properties, but did not say anything about the actual probabilities of events other than $\emptyset$ and Ω. In practice, it may be quite hard to determine the probabilities of events that may or may not occur in a natural system. In a sense, much of the work in the exercises on this web site and in the chapters [1, 2] can be understood as trying to estimate such probabilities for events in more elaborate versions of our toy sample spaces Ω_1, Ω_2 of Section 1 by simulating outbreaks.

For the purpose of estimating probabilities it is often useful to consider them as measures of the degree of certainty that a given event is going to occur. Let us return to our example of Section 1 and work with Ω_1. Assume that initially no host is infectious, but the infection may be introduced from outside the population and subsequently also be spread within it. Then $P(I_1)$ would represent our degree of certainty (or strength of belief) that host 1 will be infectious after one year in the absence of any additional information. Now suppose that we learn that host 2 is infectious after one year, in other words, that event I_2 has occurred. This would not tell us directly whether host 1 also became infectious. But the additional information would most likely *alter* our estimate of $P(I_1)$: It tells us that the infection must have been introduced into the population during the year and *could* have subsequently spread from I_2 to I_1 or the other way round (possibly through host 3 as an intermediary). None of these scenarios could have occurred if we knew that host 2 was still susceptible after one year. Thus the additional information would *increase* the strength of our belief in event I_1, albeit not to absolute certainty, which would translate into $P(I_1) = 1$. The revised probability estimate for the event $P(I_1)$ is called the *conditional probability* of I_1 *given* that event I_2 occurred and denoted by $P(I_1|I_2)$.

Statisticians are bitterly divided about whether conceptualizing probabilities entirely in terms of strength of beliefs is legitimate. One camp, the *Bayesians,* holds that it is, while the opposing camp of *frequentists* objects. To keep our neutrality, let us give the definition of conditional probability in terms of the framework that was developed in Section 1. Assume E, F are events such that $P(F) > 0$. Then the *conditional probability* of E *given* that event F occurred is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}. \tag{8}$$

If we multiply both sides of (8) by $P(F)$ we obtain the following multiplication law for conditional probabilities:

$$P(E \cap F) = P(E|F)P(F). \tag{9}$$

The definition given by (8) is less intuitive than the description of conditional probability in terms of changing beliefs, but it is easy to work with. Assume, for example, that introduction of the disease from outside the population is relatively rare, but once it gets introduced, it would quickly spread within our population. A probability function f that takes the following values might reasonably well fit this description:

$$\begin{aligned} f(SSS) = 0.3; \quad f(SSI) = f(SIS) = f(ISS) = 0.05; \\ f(SII) = f(IIS) = f(IIS) = 0.1; \quad f(III) = 0.25. \end{aligned} \tag{10}$$

Then from (4) and (8) we get:

$$P(I_1|I_2) = \frac{P(\{IIS, III\})}{P(\{SIS, SII, IIS, III\})} = \frac{0.1 + 0.25}{0.05 + 0.1 + 0.1 + 0.25} = \frac{0.35}{0.5} = 0.7, \tag{11}$$

while $P(I_1) = 0.5$. Thus given the information that I_2 has occurred we should revise our estimate of $P(I_1)$ up from 0.5 to 0.7.

Now suppose that we are modeling a visit by an infectious host from outside the population who interacts with each of hosts 1, 2, 3 and we are only interested which of these hosts gets infected by this visitor. We can study this question in terms of the same sample space Ω_1 as before, but the possible subsequent spread of the infection within the population becomes irrelevant. It may be reasonable to model the probability function f as if it resulted from tossing a biased coin three times, so that host i becomes infectious if, and only if, the coin comes up heads in toss number i. If the coin comes up heads with probability p on a single toss, then

$$\begin{aligned} f(SSS) = (1-p)^3; \quad f(SSI) = f(SIS) = f(ISS) = p(1-p)^2; \\ f(SII) = f(IIS) = f(IIS) = p^2(1-p); \quad f(III) = p^3. \end{aligned} \tag{12}$$

From (4) and (8) we get:

$$\begin{aligned} P(I_1) &= P(\{ISS, ISI, IIS, III\}) = p(1-p)^2 + 2p^2(1-p) + p^3 = p, \\ P(I_1|I_2) &= \frac{P(\{IIS, III\})}{P(\{SIS, SII, IIS, III\})} = \frac{p^2(1-p) + p^3}{p(1-p)^2 + 2p^2(1-p) + p^3} = \frac{p^2}{p} = p. \end{aligned} \tag{13}$$

In this case the given the information that I_2 has occurred would not alter our estimate of $P(I_1)$. The events I_1 and I_2 are *independent.*

The most natural formal definition would be to call two events E and F independent if, and only if, $P(E|F) = P(E)$. However, $P(E|F)$ is defined only if $P(F) > 0$. Here is a more elegant definition that always makes sense.

Definition 1 *Two events E and F are independent if, and only if,*

$$P(E \cap F) = P(E)P(F). \tag{14}$$

Equation (9) shows that as long as $P(F) > 0$, Definition 1 coincides with the requirement that $P(E|F) = P(E)$.

The notion of independence is quite subtle. For more than two events, say E_1, E_2, E_3, E_4, *independence* of all these events requires that

$$\begin{aligned}
&P(E_1 \cap E_2) = P(E_1)P(E_2), P(E_1 \cap E_3) = P(E_1)P(E_3), \ldots, P(E_3 \cap E_4) = P(E_3)P(E_4);\\
&P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3), \ldots, P(E_2 \cap E_3 \cap E_4) = P(E_2)P(E_3)P(E_4);\\
&P(E_1 \cap E_2 \cap E_3 \cap E_4) = P(E_1)P(E_2)P(E_3)P(E_4).
\end{aligned} \tag{15}$$

If only the first line of (15) is satisfied, then we say that the events E_1, E_2, E_3, E_4 are *pairwise independent.* This property does not automatically imply independence of the events; examples can be found in any introductory textbook on probability or statistics.

For our purposes it suffices to know that in experiments based on repeatedly tossing a coin or rolling a die we can always assume independence of events $E_1, \ldots, E_n$ *as long as there are no two events E_j, E_k with $j \neq k$ whose verbal descriptions involve the same toss or roll.* Recall the motivating story that we told in the paragraph that precedes Equation (12). In terms of this story, event I_1 is entirely determined by the first toss, while event I_2 is entirely determined by toss number two. Sure enough, our calculations showed that I_1 and I_2 are independent. By the same principle, events I_1, I_2 and S_3 will be independent. Trying to verify this by using the analogue of (15) and direct calculations would be quite tedious.

4 Random variables and probability distributions

When constructing sample spaces we need to choose our set of elementary outcomes so that it covers all possibilities. But in actual modeling we may not be interested in all aspects of the outcomes, at least not at the same time. For example, for our toy population of Section 1 we may just be interested in the total number ξ_I of hosts that will be infectious one year from now, or in the time T_1^I when host 1 will become infectious. Such numerical aspects of the outcomes can be expressed in terms of *random variables,* abbreviated *r.v.s.*

Technically, a r.v. is a function ξ that is defined on the sample space and takes values that are real numbers. For continuous sample spaces, there is an additional requirement that r.v.s must be *measurable* functions, but this does not concern us here as all functions that have a reasonably nice mathematical description are measurable.

We will need to make a distinction between *discrete* and *continuous* r.v.s. For example, the r.v. ξ_I that counts the total number of infectious hosts after one year in our example is discrete; it can take only values in the set $\{0, 1, 2, 3\}$. In contrast, the r.v. T_1^I is continuous; it can take any values in the interval $[0, \infty)$.

The *(probability) distribution* of a r.v. ξ tells us how likely it is that ξ takes certain values. The formal definition of this notion depends on whether the r.v. is discrete or continuous.[3]

Let us discuss the discrete case first. For our purposes we can restrict our attention to discrete r.v.s ξ that take only values in the set $\mathbb{N} = \{0, 1, 2, \dots\}$ of nonnegative integers. In this case the distribution of ξ simply specifies for each k the probability $p_k = P(\xi = k)$ that ξ will take the value k. Let us illustrate this construction with the distribution of the r.v. $\xi_I : \Omega_1 \to \mathbb{N}$ that counts the number of infectious hosts after one year. In this case we have

$$\begin{aligned} &\xi_I(SSS) = 0; \quad \xi_I(SSI) = \xi_I(SIS) = \xi_I(ISS) = 1; \\ &\xi_I(SII) = \xi_I(ISI) = \xi_I(IIS) = 2; \quad \xi_I(III) = 3. \end{aligned} \tag{16}$$

If we assume the probability measure that is generated by the probability function of (10), then

$$\begin{aligned} &p_0 = P(\xi_I = 0) = 0.3; \qquad p_1 = P(\xi_I = 1) = 0.15; \\ &p_2 = P(\xi_I = 2) = 0.3; \qquad p_3 = P(\xi_I = 3) = 0.25. \end{aligned} \tag{17}$$

For continuous r.v.s like T_1^I the definition of the distribution is more complicated, as for any given potential value x we will have $P(T_1^I) = 0$. But for any interval $[a, b]$ with $a < b$, the probability $P(a \le T_1^I \le b)$ that T_1^I falls into this interval will be positive. Thus the distribution of a continuous r.v. ξ should specify, for each interval $[a, b]$, the probability that ξ takes values in this interval.

This can be achieved as follows: Each continuous r.v. ξ has a so-called *probability density function* g. This function g is defined on the real line, takes nonnegative real values, and needs to be integrable. Typically g will be continuous except at a few points where it may have jump discontinuities. Moreover, g needs to satisfy

$$\int_{-\infty}^{\infty} g(x)\, dx = 1. \tag{18}$$

The relation between ξ and g is such that for each pair of real numbers $a < b$ we have:

$$P(a \le \xi \le b) = \int_a^b g(x)\, dx. \tag{19}$$

Thus g determines the distribution of ξ in the sense that we outlined above. Notice that (18) assures that with probability 1 the value of ξ is *some* real number. Since the probability that ξ takes one of the two endpoints of $[a, b]$ as its value is zero, we could equally well have written the left-hand side of (19) in the form $P(a < \xi < b)$.

[3]There are also *mixed* r.v.s that are neither discrete nor continuous, but we don't need them here.

Let us illustrate this construction with a simple example. Consider the function g_u that takes the value 1 for every x in the unit interval and the value 0 for all other x. This is a legitimate probability density function as $\int_{-\infty}^{\infty} g_u(x)\,dx = \int_0^1 1\,dx = 1$. If ξ_u is a r.v. with this probability density function, then for every $0 \le a < b \le 1$ we get

$$P(a \le \xi_u \le b) = \int_a^b 1\,dx = b - a. \tag{20}$$

Thus for every subinterval $[a, b]$ of $[0, 1]$ the probability that the value of ξ_u falls into $[a, b]$ is equal to the length of the interval. There is no particular area inside $[0, 1]$ where values of ξ_u tend to occur more frequently than elsewhere. We say that the r.v. ξ_u is *uniformly distributed over the interval* $[0, 1]$. The most basic random number generators that come with standard computer software generate values of r.v.s that can be, with very good approximation, assumed to be sampled from the uniform distribution over $[0, 1]$.

5 Mean and variance of a random variable

There are several notions of *averages* for a given r.v. ξ: the mean, the median, and the mode. Here we will discuss the mean. In the literature the mean is denoted by $\mu, E(\xi)$ or $\langle\xi\rangle$. We will adopt the notation $\langle\xi\rangle$.

If $\xi : \Omega \to \mathbb{N}$ is a discrete r.v. that takes values in the set $\mathbb{N}$ of nonnegative integers and if p_k denotes the probability that ξ takes the value k, then the mean $\langle\xi\rangle$ is given by the formula

$$\langle\xi\rangle = \sum_{k=1}^{\infty} k p_k. \tag{21}$$

Note that even if $p_0 > 0$, it does not matter whether the summation in (21) starts with 0 or 1 as $0p_0 = 0$. For example, the mean of the r.v. ξ_I whose distribution is given by (17) evaluates to

$$\langle\xi_I\rangle = 0(0.3) + 1(0.15) + 2(0.3) + 3(0.25) = 1(0.15) + 2(0.3) + 3(0.25) = 1.5. \tag{22}$$

If $\xi : \Omega \to \mathbf{R}$ is a continuous probability distribution with probability density function g, then the mean $\langle\xi\rangle$ is given by the formula

$$\langle\xi\rangle = \int_{-\infty}^{\infty} x g(x)\,dx. \tag{23}$$

For example, the mean of continuous r.v. ξ_u with uniform distribution on the interval $[0, 1]$ evaluates to

$$\langle\xi_u\rangle = \int_{-\infty}^{\infty} x g_u(x)\,dx = \int_0^1 x\,dx = \frac{x^2}{2}\Big]_0^1 = 0.5; \tag{24}$$

exactly as one would expect.

Functions such as ξ^2 or $\xi+\eta$ that are obtained by applying algebraic operations to one or more r.v.s are also r.v.s. The mean behaves very nicely with respect to *linear combinations.* If $\xi_1, \xi_2, \ldots, \xi_n$ are r.v.s that are defined on the same sample space Ω and $c_1, \ldots, c_n$ are constants, then

$$\langle c_1\xi_1 + c_2\xi_2 + \cdots + c_n\xi_n\rangle = c_1\langle\xi_1\rangle + c_2\langle\xi_2\rangle + \cdots + c_n\langle\xi_n\rangle. \tag{25}$$

For example, consider an experiment of tossing n coins and let $\xi_i = 1$ if coin number i comes up heads and $\xi_i = 0$ otherwise. Such r.v.s ξ_i are called *indicator r.v.s.* Then the total number ξ of heads that come up in the n tosses can be expressed as $\xi = \xi_1 + \xi_2 + \cdots + \xi_n$, and it follows from (25) that

$$\langle\xi\rangle = \langle\xi_1\rangle + \langle\xi_2\rangle + \cdots + \langle\xi_n\rangle. \tag{26}$$

The mean of a r.v. ξ is also often called the *first moment* of ξ. The *second moment* of ξ is defined as the mean $\langle x^2\rangle$, and the *variance* $Var(\xi)$ of ξ is defined as

$$Var(\xi) = \langle(\xi - \langle\xi\rangle)^2\rangle. \tag{27}$$

Note that the r.v. $(\xi - \langle\xi\rangle)^2$ takes only nonnegative values. Hence $Var(\xi)$ will always be nonnegative, and will be large if ξ takes on values that differ significantly from the mean with high probability. Thus the variance of ξ is a *measure of variability* of ξ. It will be zero only if $P(\xi = \langle\xi\rangle) = 1$, that is, if there is no variability whatsoever in the values of ξ. Another measure of variability of ξ is the *standard deviation* $\sigma = \sqrt{Var(\xi)}$. The notation σ^2 is also often used in the literature for the variance of a r.v.

The variance of any r.v. ξ is equal to the difference between the second moment of ξ and the square of its mean. Since the mean $\langle\xi\rangle$ can be treated as a constant, this result can be derived from (25) as follows:

$$Var(\xi) = \langle(\xi - \langle\xi\rangle)^2\rangle = \langle\xi^2 - 2\langle\xi\rangle\xi + \langle\xi\rangle^2\rangle = \langle\xi^2\rangle - 2\langle\xi\rangle\langle\xi\rangle + \langle\xi\rangle^2 = \langle\xi^2\rangle - \langle\xi\rangle^2. \tag{28}$$

The expression on the right of (28) often allows for easier calculation of the variance than (27). For a discrete r.v. ξ that takes values in $\mathbb{N}$, Equation (28) translates into

$$Var(\xi) = \langle\xi^2\rangle - \langle\xi\rangle^2 = \left(\sum_{k=1}^{\infty} k^2 p_k\right) - \left(\sum_{k=1}^{\infty} k p_k\right)^2. \tag{29}$$

For a continuous r.v. ξ with probability density function g, Equation (28) translates into

$$Var(\xi) = \langle\xi^2\rangle - \langle\xi\rangle^2 = \int_{-\infty}^{\infty} x^2 g(x)\,dx - \left(\int_{-\infty}^{\infty} x g(x)\,dx\right)^2. \tag{30}$$

6 Medians, quartiles, and percentiles

Let ξ be a random variable. In the previous section we defined the mean, variance, and standard deviation of ξ. The mean is a *measure of central tendency* while the variance and standard deviation of ξ are *measures of dispersion.* Another important measure of central tendency is the *median* Q_2 of ξ. It is defined in such a way that ξ will take values $\leq Q_2$ with probability at least 0.5 and values $\geq Q_2$ also with probability at least 0.5. A more general notion is the P-th percentile. It is a number Q such that that ξ will take values $\leq Q$ with probability at least $\frac{P}{100}$ and values $\geq Q$ also with probability at least $\frac{P}{100}$. The 25-th percentile is commonly denoted by Q_1, the 50-th percentile is the median Q_2, and the 75-th percentile is commonly denoted by Q_3. The numbers Q_1, Q_2, Q_3 are called *quartiles* as they typically divide a large data set of values of the r.v. ξ into four parts of roughly equal sizes. The difference $IQR = Q_3 - Q_1$ is called the *interquartile range* and is another measure of dispersion.

For r.v.s ξ with uniform distributions, the mean $\langle k \rangle$ is the same as the median Q_2. The same will be true if ξ has a binomial or a normal distribution that we will define in the next two sections. For r.v.s with other distributions the median can be larger or smaller than the mean. If ξ has an exponential distribution as defined in Subsection 8.1 below, then the mean will be larger than the median, as exponentially distributed r.v.s occasionally take unusually large values, while the minimum is bounded from below by 0. The mean may be strongly influenced by such outliers, while the median does not discriminate between "larger than average" and "much, much larger than average" values.

It is interesting to see what happens if we stretch the definition of r.v.s a bit so as to allow infinite values. Suppose ξ is such a generalized r.v. that takes no negative values and takes the value ∞ with positive probability. Then the mean of ξ must be infinite. However, if $P(\xi = \infty) < 0.5$, the median will still be finite. If $0.5 < P(\xi = \infty) < 1$, then the median will also be infinite, but there will be some positive number P such that the P-th percentiles is finite.

7 Selected discrete probability distributions

7.1 Bernoulli distributions

Think about tossing a biased coin that comes up heads with probability p and tails with probability $1 - p$. Let ξ be the r.v. that counts the number of "successes," that is, the number of times heads comes up in this one toss. This r.v. has a *Bernoulli distribution with parameter p.*

Note that $p_1 = p$ and $p_0 = 1 - p$, while $p_k = 0$ for $k > 1$. By substituting these probabilities into (21) and (29) we find that

$$\langle \xi \rangle = p; \qquad Var(\xi) = p - p^2 = p(1-p). \tag{31}$$

7.2 Binomial distributions

Think about tossing a biased coin n times. Assume that the coin comes up heads with probability p and tails with probability $1-p$. Let ξ be the r.v. that counts the number of "successes," that is, times heads comes up in these tosses. This r.v. has a *binomial distribution with parameters n and p*. We get:

$$\begin{aligned} p_k &= \binom{n}{k} p^k (1-p)^{n-k}; \\ \langle \xi \rangle &= np; \qquad Var(\xi) = np(1-p). \end{aligned} \tag{32}$$

If $0 < p < 1$, then $p_k > 0$ for all $k \in \{0, 1, \ldots, n\}$. For $k > n$ the binomial coefficient $\binom{n}{k}$ evaluates to 0, which makes sense, since the number of successes cannot exceed the number of coin tosses. The second line of (32) can be derived from the first. Alternatively, the formula $\langle \xi \rangle = np$ can be derived from (31) and (26).

7.3 Poisson distributions

For large n the probabilities p_k defined in the first line of (32) can be difficult to calculate directly. Often it is better to work with an approximation to the binomial distribution that gives similar probabilities. If n is large and np is of moderate size ($np < 10$ is often quoted as a rule of thumb), the *Poisson distribution* usually works very well. This has one parameter λ which should be set equal to np in approximations of the binomial distribution with parameters n and p. The next formula summarizes its properties.

$$\begin{aligned} p_k &= \frac{\lambda^k e^{-\lambda}}{k!} \\ \langle \xi \rangle &= Var(\xi) = \lambda. \end{aligned} \tag{33}$$

Notice that in a Poisson distribution we have $p_k > 0$ for all k. If we want to assume that a r.v. ξ has *exactly* a Poisson distribution, we need to allow that it can take on arbitrary nonnegative integers as its values. This of course does not, strictly speaking, make sense in the standard interpretation of binomial r.v.s. But for large k the probability p_k will be very close to 0 for moderate values of λ, which will give a rather good approximation to the binomial distribution where $p_k = 0$ for $k > n$.

8 Selected continuous probability distributions

8.1 Exponential distributions

Exponentially distributed r.v.s take only values that are positive reals. Their probability density functions have the form

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0, \end{cases} \tag{34}$$

where $\lambda > 0$ is the parameter of the distribution.

If ξ has an exponential distribution with parameter λ, then for all $T \geq 0$:

$$\begin{aligned} P(\xi \leq T) = 1 - e^{-\lambda T}, \quad \text{or, equivalently,} \quad P(\xi > T) = e^{-\lambda T}; \\ \langle \xi \rangle = \frac{1}{\lambda}; \qquad Var(\xi) = \frac{1}{\lambda^2}. \end{aligned} \tag{35}$$

Continuous r.v.s that represent waiting times often have exponential distributions. More precisely, such r.v.s are exponentially distributed if, and only if, the random variable is *memoryless,* that is, if the time one still needs to wait is independent of how long one has already been waiting. The first line of (35) shows that exponentially distributed r.v.s are indeed memoryless: The conditional probability that one will need to wait at least another t units of time *given* that one has already waited T time units can be calculated as

$$P(\xi \geq T + t | \xi > T) = \frac{e^{-\lambda(T+t)}}{e^{-\lambda T}} = e^{-\lambda t}. \tag{36}$$

The rightmost expression of (36) does not depend on T, which means that the future (how long one still needs to wait) is independent of the past (the fact that one has already been waiting T time units). This is exactly what the expression "ξ is memoryless" means.

The second line of (35) shows that larger values of the parameter λ in exponential distributions translate into shorter expected waiting times.

The first line gives us another interesting perspective on this: By L'Hospital's Rule, $\lim_{\Delta t \to 0^+} \frac{1 - e^{-\lambda \Delta t}}{\Delta t} = \lambda$. It follows that as long as Δt is sufficiently small and ξ is exponentially distributed with parameter λ, we have

$$P(t \leq \xi \leq t + \Delta t) = P(\xi \leq \Delta t) \approx \lambda \Delta t, \tag{37}$$

8.2 Normal distributions and the Central Limit Theorem

The probability density function of a r.v. ξ with *normal* or *Gaussian* distribution with parameters μ and σ is given by

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}. \tag{38}$$

The graph of g has a characteristic bell shape. The parameter μ represents the mean value and the parameter σ the standard deviation. In our notation this translates into $\langle \xi \rangle = \mu$ and $Var(\xi) = \sigma^2$.

Normal distributions often give good approximations to binomial distributions with parameters n and p in situations where np is too large for the Poisson approximation. Here

is how this works: Let ξ be a r.v. that has a binomial distribution with parameters n and p. Suppose we want to estimate the probability $P(k_1 \leq \xi < k_2)$. If n is sufficiently large, then we can use the fact that this probability will be very close to the probability that a Gaussian r.v. η with the same mean and standard variation as ξ takes values in the corresponding interval.

To perform these calculations, we first need to find the parameters μ and σ of η. By (32), we must set $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Next we must find the "corresponding" interval for η. The natural impulse would be to estimate $P(k_1 \leq \eta < k_2)$, but this may not work too well. While ξ can take only integer values, all real values are allowed for η. Thus $P(k_1 \leq \xi < k_2) - P(k_1 < \xi < k_2) = P(\xi = k_1) > 0$, while $P(k_1 \leq \eta < k_2) = P(k_1 < \eta < k_2)$. This undesirable effect can be largely avoided by making the so-called *continuity correction.* It is based on identifying the event that the discrete r.v. ξ takes the integer value k with the event that the continuous r.v. η takes a value in the interval $(k-0.5, k+0.5]$ that is centered around k and has length 1. We can then estimating $P(k_1 \leq \xi < k_2)$ by the probability $P(k_1-0.5 < \xi < k_2-0.5)$ that η takes values in the union of those unit-length intervals that are centered at $k_1, k_1+1, \ldots, k_2-1$.

Now we need to deal with a third problem: The integral of g that expresses $P(k_1-0.5 < \xi < k_2-0.5)$ cannot be evaluated in closed form. We need to convert the endpoints of the interval for η into so-called *z-scores* $z_1 = \frac{k_1-0.5-\mu}{\sigma}$ and $z_2 = \frac{k_2-0.5-\mu}{\sigma}$ so that $P(k_1 - 0.5 < \xi < k_2 - 0.5) = P(z_1 < \zeta < z_2)$ for a *standard normal r.v.* ζ, that is, for a random variable ζ that is normally distributed with parameters $\mu = 0$ and $\sigma = 1$. The probability $P(z_1 < \zeta < z_2)$ can then be looked up in a table or found with standard statistical software.

Normal distributions are ubiquitous in mathematical statistics. The main reason is that almost all r.v.s ξ that can be expressed in the form

$$\xi = \frac{\eta_1 + \eta_2 + \cdots + \eta_n}{n}, \tag{39}$$

where the η_m's are independent r.v.s with the same distribution (abbreviated *iid* for *independent and identically distributed*) have distributions that are very close to normal ones as long as n is large enough.

This is what the *Central Limit Theorem (CLT)* asserts. It almost doesn't matter what kind of distribution the r.v.s η_m themselves have, all we need is that it is the same for all m with some fixed and finite mean $\langle \eta_m \rangle = \mu$ and some fixed and finite variance $Var(\eta_m) = \sigma_\eta^2$. Then $\langle \xi \rangle = \mu$ and $Var(\xi) = \frac{\sigma_\eta^2}{n}$, so that ξ will have an approximately normal distribution with parameters μ and $\sigma = \frac{\sigma_\eta}{\sqrt{n}}$.

For example, the outcomes of repeated simulations of disease outbreaks can often be expressed in terms of numbers, such as the final size or total duration of the outbreak. These numbers will generally differ from simulation to simulation, so that the outcome of simulation number m becomes a r.v. η_m. We can assume that these r.v.s are independent, and as long as we keep the parameters of the simulation fixed, all of these r.v.s will have the same distribution. We don't know the actual distribution of the r.v.s η_m though; in fact, the purpose of running simulations is usually to find out something about this distribution.

In most cases, we want to estimate the mean $\langle \eta_m \rangle$. This should be the same number μ for all m.

The CLT tells us that if we run a *very large number* n of simulations, then the mean $\xi = \frac{\eta_1 + \ldots \eta_n}{n}$ of the observed outcomes will be close to the true mean μ. But how large does the number n of simulations need to be so that we can, say, be 95% sure that the observed mean ξ differs by less than ε from the true mean μ, where ε is our tolerance for error?

Here a nice property of normally distributed r.v.s ξ helps: With probability very close to 0.95, the value of ξ will differ from the mean by less than two standard deviations. In mathematical terms:

$$P(\mu - 2\sigma < \xi < \mu + 2\sigma) \approx 0.95. \tag{40}$$

It follows that all we need to do is choose n large enough so that for ξ defined as in (39) we get

$$\sigma \leq \frac{\varepsilon}{2}. \tag{41}$$

By the CLT, the equality

$$\sigma = \frac{\sigma_\eta}{\sqrt{n}}. \tag{42}$$

will hold with very good approximation. By substituting the right hand side of (42) into (41) and solving for n we conclude that we need to choose n large enough so that

$$n \geq \frac{4\sigma_\eta^2}{\varepsilon^2}. \tag{43}$$

Looks simple and straightforward, doesn't it? Unfortunately, there is a snag: We usually don't know the value of σ_η^2, so we cannot plug it into (43)!

The way out of this conundrum is to first run a few preliminary simulations, use the outputs to calculate a rough estimate of the standard deviation σ_η, plug this rough estimate into (43) and use it to derive an estimate of the minimum number n of simulations that we need to perform so as to be 95% confident that the observed mean will differ from the true mean of η by less than our error tolerance ε. This procedure usually works reasonably well.

References

[1] Winfried Just, Hannah Callender, and M Drew LaMar. Disease transmission dynamics on networks: Network structure *vs.* disease dynamics. In Raina Robeva, editor, *Algebraic and Discrete Mathematical Methods for Modern Biology*, pages 217–235. Academic Press, 2015.

[2] Winfried Just, Hannah Callender, M Drew LaMar, and Natalia Toporikova. Transmission of infectious diseases: Data, models, and simulations. In Raina Robeva, editor,

Algebraic and Discrete Mathematical Methods for Modern Biology, pages 193–215. Academic Press, 2015.