NRES 798 — Statistical Methods for Ecologists Chapter 2: Random Variables and Probability Distributions

Oscar García

February 2, 2013

Contents

1 Random variables

A *random variable* (RV) is a variable taking values determined by a random phenomenon. More precisely, as indicated before, it is a function assigning a number to each outcome (sample point). Or one might think of the outcomes as being numerical from the start. Anyway, from now on we deal only with RVs, without concerning ourselves with what might be behind those numbers (at least for mathematical manipulation purposes).

For RVs, the probability model (sample space and probability function) is commonly called a *probability distribution*. The sample space is taken as the whole real line (the set $\mathbf{R}$ of all the real numbers), even though many or most of the events there might have probability 0. Usually the RV (the function) is represented by capital letters near the end of the alphabet, like X and Y, while the corresponding lower-case letters are used for the RV values (numbers).

It should be clear from the definition that any function of RVs is also an RV. The original outcomes (e.g., 1 for observing a female moose and 0 for observing a male) are mapped into numerical outcomes in a new sample space (e.g., the number of females in a sequence of 4 observations). Actually, the new sample space looks the same as the old one(s), it is $\mathbf{R}$, so we do not need to worry about it (although the events with non-zero probability may be different). The new probability distribution can be derived from the original one(s). Much of Statistics deals with these RV transformations.

The analogy of a probability distribution to the distribution of mass over a line (string or wire) is almost perfect, and should help in understanding what follows. *Continuous* RV's can take any values over a (finite or infinite) interval, corresponding to a typical solid wire. If the density (or thickness) of the wire varies smoothly over its length, one could draw that density as a smooth curve, and the mass of any piece of wire is the area under the curve. Same with the probability density function (PDF), the only difference is that the total mass is 1.

A *discrete* RV has non-zero probability only on a finite or countably infinite set of points (usually integers). The mass on the line in concentrated into those points: the density is 0, except for a number of “spikes”. Of course, there is also the possibility of an RV being partly continuous and partly discrete, but we will only consider the two simple types[1].

One speaks also of discrete and continuous distributions, referring to the probability models based on the respective RV types.

[1] Almost everything will be also valid for the multivariate case of more than one RV, substituting vectors for the simple numbers (scalars). Think of the wire changing to a flat plate, or to a solid volume. But we will stick to univariate RVs here.

2 Specifying the probability

The probability function, defined over sets (events), is not very convenient to work with. However, set probabilities can be calculated using two simpler functions from numbers to numbers, the *probability density function* (PDF), or the *cumulative distribution function* (CDF). The domain of these functions, the set over which they are defined, is $\mathbf{R}$.

2.1 Discrete

For discrete RVs, it is easy to give simply the probabilities of the discrete outcomes. The probability for any other event can be obtained by adding these over the relevant set. The function that gives these probabilities for each point (the height of the spikes) is usually called the PDF, by analogy with the continuous case[2]. Typical notations are

$$f(x) \ , \quad f(x_i) \ , \quad p(x) \ , \quad p(x_i) \ , \text{ or } \quad p_i \ .$$

Here $\{x_i\}$ or $(x_1, x_2, \dots)$ are the possible outcomes, the function is 0 elsewhere.

The CDF $F(x)$ is the probability of $X \leq x$, for any real number x. In other words, the probability of the event consisting of all the sample points x_i that are smaller than or equal to the given number x. Therefore,

$$F(x) = \sum_{x_i \leq x} f(x_i) \ .$$

It is seen that this is a step function.

The CDF is more useful with continuous distributions. Even in the discrete case, however, it can simplify some calculations. Suppose, for instance, that the x_i are in increasing order, and that we want to obtain the probability of the values between x_5 and x_{12}, inclusive. With the PDF it would be $\sum_{i=5}^{12} f(x_i)$. With the CDF, the probability is $F(x_{12}) - F(x_4)$ (why?).

[2] Sometimes called the probability mass function, among other names. The textbook uses *distribution function*, which elsewhere normally refers to the CDF. I will stick to PDF and CDF to avoid confusion.

2.2 Continuous

For continuous RVs (continuous distributions), it may be easier to start with the CDF. It is the same as before, $F(x) = P(X \leq x)$. This is shorthand for $P(\{X \text{ such that } X \leq x\})$, or $P((-\infty, x])$, where $(-\infty, x]$ is the interval of the real line containing x and all the points to the left of it[3].

The CDF can be used to calculate probabilities for any set (event) obtainable through set operations on intervals, which should include all the events of practical interest. For instance,

$$P(X > x) = P((x, \infty)) = P(\overline{(-\infty, x]}) = 1 - F(x)$$
$$P(a < X \leq b) = P((-\infty, b] \setminus (-\infty, a]) = F(b) - F(a)$$

(here (a, b) denotes an open interval, not a pair!)

In particular, for an interval of length Δx starting at x, the average probability density, that is, the probability per unit length, is

$$\frac{P(x < X \leq x + \Delta x)}{\Delta x} = \frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{\Delta F}{\Delta x} \; .$$

This is the slope of the curve $F(x)$ over the interval, height difference divided by distance. As Δx becomes smaller, this becomes the *probability density function* (PDF)

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x} \; ,$$

the slope of $F(x)$ at x. With some care, it can be shown that this works also for discrete distributions and gives the same PDF that we had before, so we are OK using the same name.

The probability of an interval can be obtained as the area under the PDF curve, in addition to as the difference between the CDF values at the ends as shown before. This can be seen by dividing the interval into small segments; the probability of each of them equals roughly its density multiplied by its length (think of the piece of wire). So that the probability for the interval is a sum like

$$\sum_x f(x) \Delta x \; ,$$

[3] The textbook writes $F(x) = P(X < x)$, which is the same for continuous RVs because then the probability of x is 0. It may not be if the RV is not continuous; the definition above is the standard one.

which corresponds to the (approximate) area under the f curve. As Δx gets smaller the approximation improves, and the sum tends to an integral. Or one can use the fact that the derivative and the integral are inverses of each other. Either way,

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x \ .$$

We conclude that either the PDF or the CDF characterize a probability distribution model, and can be used to obtain the probability for any events of interest[4].

3 Expectation and variance

The *expected value* or *expectation* of a random variable is essentially a weighted average, defined for discrete RVs as

$$E[X] = \sum_i x_i f(x_i) \ ,$$

and similarly,

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

for continuous. The notation $E[X] = \mu$ (for *mean* or population mean) is sometimes used. It is a characteristic of the distribution that describes a central point around which the probability is distributed. In the mass distribution example, it is the centre of gravity.

From the linearity of sums and integrals, it follows that E is linear:

$$E[X+Y] = E[X] + E[Y] \ , \quad E[aX] = aE[X] \ ,$$
$$E[aX + bY + c] = aE[X] + bE[Y] + c \ ,$$

for any RVs X and Y, and any constants a, b, c.

The *variance* is a measure of spread or variability, defined as the expectation of squared deviations:

$$V[X] = \sigma^2(X) = E[(X-\mu)^2] \ .$$

[4] For a vector RV, $\boldsymbol{X} \leq \boldsymbol{x}$ is defined as $X_i \leq x_i$ for all the vector components. Draw the two-dimensional interval for a two-dimensional RV. Then $F(\boldsymbol{x})$ and $f(\boldsymbol{x})$ are surfaces, and similarly in higher dimensions.

Sometimes its square root, the *standard deviation* σ, is more useful, carrying the same measurement units as X and the mean.

From the definition, it is easy to see that

$$V[aX+b] = a^2 V[X] \; .$$

If X and Y are independent RVs, then[5]

$$V[X+Y] = V[X] + V[Y] \qquad (X, Y \text{independent}) \; .$$

4 Common distributions

4.1 Discrete

The *Bernouilli* applies to RVs that can only take the values 1 or 0. The PDF is $f(1) = p$, $f(0) = 1 - p$, and $f(x) = 0$ elsewhere; p is the distribution parameter.

The *binomial* distribution usually arises as the number X of ones in n independent Bernouilli trials. The PDF is

$$\begin{aligned} f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x \in \{0, 1, \ldots, n\} \; , \\ f(x) &= 0 \text{ otherwise.} \end{aligned}$$

The mean is $E[X] = np$.

The *Poisson* distribution approximates the binomial when n is large and p is small, with $np = \lambda$. It is also appropriate when things happen "randomly" at a constant rate λ over space or time. The PDF is $f(x) = \lambda^x \exp(-\lambda)/x!$ over the non-negative integers. It has one parameter, λ, which is also the mean.

4.2 Continuous

A *uniform* RV has a flat PDF over an interval $[a, b]$, 0 elsewhere. Since the area under f must be 1, $f(x) = 1/(b-a)$ over $[a, b]$. The CDF $F(x)$ must

[5] $V[X+Y] = E[((X+Y) - (\mu_x + \mu_y))^2] = E[((X - \mu_x) + (Y - \mu_y))^2] = E[(X - \mu_x)^2 + 2(X - \mu_x)(Y - \mu_y) + (Y - \mu_y)^2] = V(X) + 2\,\mathrm{Cov}(X, Y) + V[Y]$, where Cov is the covariance; $\mathrm{Cov}(X, Y) = 0$ if X and Y are independent.

be 0 for $x < a$, it has a constant slope $f(x) = 1/(b - c)$ over the interval, and then it must be 1 for $x > b$. That is,

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x > b \end{cases}$$

The mean is clearly the interval mid-point $E[X] = (b - a)/2$.

The *normal* or *Gaussian* is the most important distribution in Statistics. The PDF is positive over the whole real line, although fast decreasing in value away from the centre, in the form of a bell-shape curve. The *standard normal* PDF is

$$f(x) = \frac{\exp(-\frac{1}{2}x^2)}{\sqrt{2\pi}} .$$

$\sqrt{2\pi}$ is the integral of $\exp(-x^2/2)$ over $(-\infty, \infty)$, and needs to be included for the area under $f(x)$ to be 1. The mean is 0, and the variance is 1. This RV can be scaled and shifted:

$$\sigma X + \mu = Y ,$$

and Y has then the general normal distribution with parameters μ and σ. From the properties of expectation and variance, $E[Y] = \sigma E[X] + \mu = \mu$, and $V[Y] = \sigma^2 V[X] = \sigma^2$, justifying the choice of symbols. The notation $X \sim N(\mu, \sigma)$ may be used to mean "the RV X has a normal distribution with parameters μ and σ".

The sum of two normal RVs is also normal. More generally, any linear combination of normals is normal. The mean and variance of linear functions have already been given in Section 3.

The normal CDF $F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$ cannot be written in terms of elementary functions (logs, exponentials, square roots, and such)[6]. The integral has to be evaluated numerically for specific values of μ and σ, or statistical tables or software can be used (`pnorm` in *R*).

If Y is normal, then $X = \exp(Y)$ has the *log-normal* distribution. Or the other way around, an RV is log-normal if its logarithm is normal. Because exp is always positive, the PDF $f(x)$ and CDF $F(x)$ are non-zero only for positive x.

[6] It can be written in terms of *special functions*, such as the *complementary error function* erfc: $F(x) = \frac{1}{2}\,\mathrm{erfc}(\frac{\mu - x}{\sqrt{2}\sigma})$.

The *exponential* is another distribution with non-negative RVs (including 0 this time). The PDF is $f(x) = \beta \exp(-\beta x)$. Integrating (Maxima, Maple, Wolfram Alpha!), we obtain the CDF $F(x) = 1 - \exp(-\beta x)$. There is a relationship with Poisson processes, where things happen or appear randomly at a constant rate in space or time: the distance between successive occurrences in one dimension, or the distance between nearest neighbours in the plane, are exponentially distributed.

5 The Central Limit Theorem

Consider a sequence of independent an identically distributed RVs $X_1, X_2, \ldots, X_n$, each having mean μ and variance σ^2. For instance, a sample of n observations. The sum is $S_n = \sum_i X_i$, and the (sample) mean is $\overline{X}_n = S_n/n$.

From Section 3, we know that

$$\begin{array}{ll} E[S_n] = n\mu \ , & E[\overline{X}_n] = \mu \ , \\ V[S_n] = n\sigma^2 \ , & V[\overline{X}_n] = \sigma^2/n \ . \end{array}$$

Note that as n increases, the mean becomes more precise (its variance decreases), as one might expect.

From the previous section, we know that linear functions of normals are normal. Therefore, if $X_i \sim N(\mu, \sigma)$, then $S_n \sim N(n\mu, \sqrt{n}\sigma)$, and $\overline{X}_n \sim N(\mu, \sigma/\sqrt{n})$. The *Central Limit Theorem* says that this is also approximately true for non-normal RVs, when n is large enough.

More precisely, the Central Limit Theorem says that the distribution of the sum or average of n independent and identically distributed RVs tend to a normal distribution as n tends to infinity[7]. Depending on the exact wording and method of proof, we might want to avoid the mean or variance tending to infinity, or the variance tending to 0, by scaling the RVs. So, another version of the theorem says that the standardized sum $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ tends to a standard normal $N(0, 1)$ as $n \to \infty$.

Under some additional conditions, the theorem is also true for non-identically distributed RVs, and even under some forms of dependency. It justifies using the normal distribution as a model in situations where one might think of variability as the result of many unspecified causes. It justifies also taking

[7] The RVs must have finite variance, there are some unusual distributions that do not.

of a sample mean as approximately normal in “large samples”, regardless of the distribution of the underlying RV.