Random Variables

1 Introduction

Recall that we introduced *random variables* as "functions defined on a sample space". In other words, we have a sample space Ω, and functions (often denoted by $X, Y, Z, \ldots$) which take specific values depending on the outcome of our observations. The point is that we will mostly deal with *events* defined by values taken by one or more random variables, as in

$$\{X = x\}, \{a \leq X \leq b\}, \{X = x, a \leq Y \leq b\}$$

etc. We'll sometimes write r.v. for "random variable", for brevity. In the following we start by considering variables that can only take a listed (finite or "countable", as in, as an example, all integers) set of values: $x_1, x_2, \ldots, x_{n,\ldots}$.

2 Describing Random Variables

The collection of probabilities

$$P[X = x] = p_X(x) \tag{1}$$

(for all possible values of x)is called the (*probability*) *distribution* of the random variable X. When we consider more than on random variable, we may speak of the collection

$$P[X = x, Y = y] = p_{X,Y}(x, y)$$

(for all possible values of x and y) as the *joint distribution* of X and Y.

Since a full description of the distribution is sometimes difficult to obtain, and at other times is not needed, we often use some *parameters* of the distribution. Specifically, we will often be interested in

- The *mean* (or *expected value*) of the r.v. X

$$EX = \sum_x p_X(x)\, x \tag{2}$$

where the sum is over all possible values of x.

- The mean (or expected value) of some interesting function of X: sometimes we consider a function of X,$f(X)$ (common examples are $X^2, X^3, \ldots$, e^{kX}, and so on). It is a random variable too, and with a little reflection it is not difficult to show that

$$Ef(X) = \sum_x p_X(x) f(x) \tag{3}$$

- As special cases of the previous definition, we have the *absolute moments* of the distribution:

$$EX^k = \sum_x p_X(x) x^k \tag{4}$$

for $k = 2, 3, \ldots$ and the *centered moments*

$$E(X - EX)^k = \sum_x p_X(x) (x - EX)^k = \sum_x p_X(x) \left(x - \sum_x p_X(x) x \right)^k \tag{5}$$

- An especially commonly used centered moment is the one with $k = 2$, called the *Variance* of the r.v.:

$$Var(X) = \sum_x p_X(x) (x - EX)^2 \tag{6}$$

Using the formulas above and a little algebra, it is not hard to show a few properties of the operation of "taking the expected value". For example:

- $E(aX + bY) = aEX + bEY$ (where a and b are two numbers)
- $E(aX + bY)^2 = a^2EX^2 + b^2EY^2 + 2abE(XY)$
- $Var(X + Y) = Var(X) + Var(Y) + 2cov(X, Y)$, where $cov(X, Y) = E[(X - EX)(Y - EY)]$ is usually called the *covariance* of X and Y

We'll see several cases where the whole point of our experiments will reduce to estimating the "true" value of the expected values and the variance of one or more random variables.

3 A Few Consequences

Concentrating a moment on the first few moments, in particular the mean and the variance, there a re a few consequences that we may want to draw.

3.1 Linear Transformations

Given a r.v. X we are sometimes interested in working with a new r.v. defined as $aX + b$, where a and b are two numbers. Note that

- $X + b$ is the same r.v., but with its values "shifted by b". Suppose, for example, that X represents the time until a certain event occurs (e.g., a bus arrives). To study it we need to decide a "starting time", when is it that $X = 0$. $X + b$ shifts the starting time to $-b$.
- aX is the same r.v., but using a different *scale*. Suppose we measure the time X mentioned above in hours. If we decide to change are units to minutes, all readings will be multiplied by 60, hence we will be considering the r.v. $60X$.

A little algebra, and the definitions show that

- $E(aX + b) = aEX + b$
- $Var(aX + b) = a^2 Var(X)$

In particular, a shift does not change the variance, since we are computing it with respect to EX which is shifted by the same amount.

3.2 Using Moments to Get Estimates

A simple (and very rough, because of its vast generality) estimate illustrates one use of moments. Consider

$$E|X|^k$$

for some k (we use absolute values, so as to deal only with non negative values, whether k is even or odd). Applying the formula we saw, we have

$$E|X|^k = \sum_x p_X(x)|x|^k \tag{7}$$

Now, suppose we are interested in the probability that $|X|$ exceed a certain value, $P[|X| > M]$. To get a rough handle on it, we can split the sum in (7) in two parts: for $|x| < M$, and $|x| \geq M$

$$\sum_x p_X(x)|x|^k = \sum_{x<M} p_X(x)|x|^k + \sum_{x\geq M} p_X(x)|x|^k$$

Now, if $|x| \geq M$, we lower the value of the sum if we write M in place of x:

$$\sum_{|x|\geq M} p_X(x)|x|^k \geq \sum_{|x|\geq M} p_X(x)M^k = M^k \sum_{|x|\geq M} p_X(x) = M^k P[|X| \geq M]$$

Also, no matter what, we'll have that $\sum_{|x|<M} p_X(x)|x|^k \geq 0$ (that's a quite rough estimate, but we are assuming almost nothing on p_X, so we can only apply very rough information). Combining the two,

$$E|X|^k \geq M^k P[|X| \geq M]$$

$$P[|X| \geq M] \leq \frac{E|X|^k}{M^k}$$

This is known as "Markov's Inequality". In particular, consider a r.v. Y and define $X = Y - EY$. Then $EX^2 = Var(Y)$, and Markov's inequality, for $k = 2$, becomes

$$P[|Y - EY| \geq M] \leq \frac{Var(Y)}{M^2}$$

This is known as "Chebyshev's Inequality". Hence, knowing the variance of a r.v. allows us a worst case estimate of the probability of ending up far from the mean.

3.3 What is "Expected" in the "Expected Value"?

Actually, nothing is expected. EX is not (necessarily) the most likely outcome, and, quite often, it is not even a value that X will ever take (think of X, equal to 0 or 1, each with probability $\frac{1}{2}$: the expected value is $0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$ which is not a value X can take).

The significance of EX, in practical terms, is given by the (mathematical) *Law of Large Numbers*, which we will discuss when turning to statistical applications. In a hand-waving way, it is a good approximation to the average of very many independent observations of X. For example, if you play a lottery, with probability of win p (say, $p = 10^{-5}$). If a win will gain you \$100, we can represent a win as a random variable X such that $P[X = 100] = 10^{-5}$, and $P[X = 0] = 1 - 10^{-5}$. Hence,

$$EX = 100 \cdot 10^{-5} = 10^{-3}$$

If n successive attempts at this lottery can be considered as independent, identically distributed copies of X, and n is sufficiently large, we may expect to end up with an *average win* (that is, total dollars won, divided by number of attempts) approximately equal to 10^{-3}.

This *does not mean that you can be pretty sure that after, say,* $10,000$ *attempts, you will end up with \$10!* To make this clear, consider a simpler calculation: suppose you play a "fair game" of chance, one in which the probability of winning is $\frac{1}{2}$, and you are looking at the number of wins in this game, over a large number of attempts. If the Law of Large Numbers applies, the percentage of wins will be close to 50%. This means that if you play N times, and you win n times

$$\frac{n}{N} \approx \frac{1}{2}$$

You would think that this implies that $n \approx N - n$, but that's not so! The statement above means, in precise language, that

$$\left| \frac{n}{N} - \frac{1}{2} \right| < \varepsilon \tag{8}$$

for any ε, provided N is sufficiently large. If, for example, we had $n = \frac{N}{2} + \sqrt{N}$, (8) would hold very well:

$$\left| \frac{n}{N} - \frac{1}{2} \right| = \left| \frac{\frac{N}{2}}{N} + \frac{\sqrt{N}}{N} - \frac{1}{2} \right| = \left| \frac{\sqrt{N}}{N} \right| = \frac{1}{\sqrt{N}}$$

and if $N > \frac{1}{\varepsilon^2}$, indeed

$$\left| \frac{n}{N} - \frac{1}{2} \right| < \varepsilon$$

Consequently, for, say $N = 10^6$ (you play a million games), you would be winning

$$\frac{10^6}{2} + 10^3$$

times, and so the difference between your wins and those of your opponent would be

$$\frac{10^6}{2} + 10^3 - \left(\frac{10^6}{2} - 10^3 \right) = 2 \cdot 10^3$$

a relatively small number compared to 10^6, but marking a significant difference in number of wins – and, under this assumption, things would get worse and worse as the game proceeded!

4 A Note on "Continuous" Random Variables

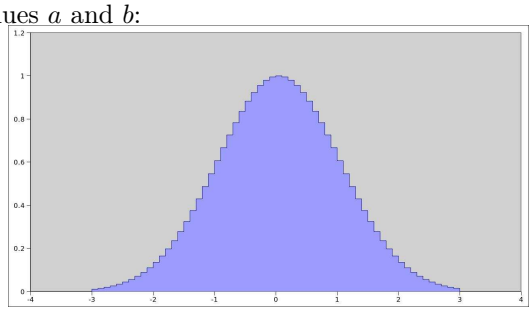
We may consider any r.v. as "discrete", i.e., taking a number of values (maybe, theoretically, infinite), that we can list. In some cases, this is obvious. For instance, the toss of two dice results in 11 possible values for $Z = X + Y$ where X and Y are the points shown by the first and second die. Also, if we play a game repeatedly, and for a very long time, the attempt number of our first win, let's call it N, can take value 1, 2, 3, ... and so on, potentially without bound, if we keep losing and are very persistent.

However, in other cases this idea of "listing all values" is a bit of a stretch. Suppose we are measuring the time needed for a piece of equipment to fail, T. In principle, T can take any non negative real number as a value. However, we may still treat it as a "discrete r.v.", if we take into account that our measurements will be inevitably limited in precision. Hence, if we can be accurate to the minute, and T is measured in hours, the only values we can observe will be $0, \frac{1}{60}, \frac{2}{60}, \cdots, 1, \frac{61}{60,}, \cdots$. Note that, in theory, this sequence can go on indefinitely.

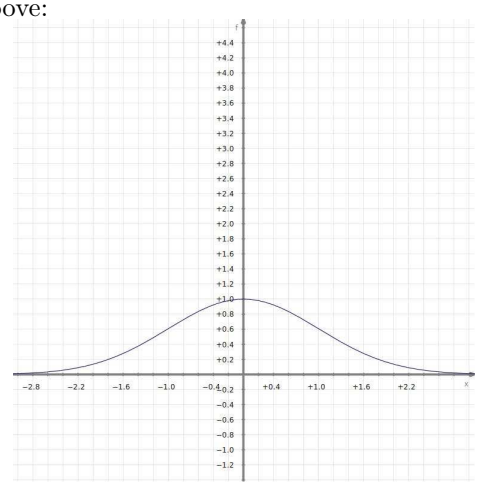
Hence, formulas like (1), (2), (3), (4), (5), (6), and (7) can always be thought as making sense. Of course, if we are dealing with a huge number of values, each

with a very small probability, calculating these formulas can become exceedingly difficult. For this reason, mathematicians have developed a tool to evaluate these sums to great accuracy, without having to actually add all these many small terms. If you will take a calculus class, you will learn how this problem of computing long sums of very small addends has been solved by the introduction of *integrals*. Since we will not really need to perform those sums (we'll rely on the work of others who bothered to do that), we won't go into this field. Of course, if you should pursue the study of statistics beyond introductory courses, you will definitely need to include calculus in your bag of tools!

The one thing that we need to remember is the following: suppose that we have a r.v. taking very many values, very close one to the next, each with very small probability. Draw a histogram of this distribution: the probability of the event $a \leq X \leq b$ is given by the area below the histogram between the two values a and b:



Now, if we think of each step as being very small, almost invisible (and note how this can be done by selecting a suitable *scale* for the units used to measure our variable!), we may substitute a smooth curve for the ragged line we have above:



And, just as with the ragged curve, the probability for X to be between two values will be the area under the curve, between these two values. More often than not this is not a simple calculation, but there are plenty of tables and computer programs that can do the work for us.

5 A Few Examples

The On Line Stat Book concentrates on two cases: the binomial distribution, as an example discrete distribution, and the so-called *normal distribution*, as the example continuous distribution. There are good reasons to concentrate on the latter (the main one being a deep theorem known as the *Central Limit Theorem*, which we will discuss when applying all this material to statistical problems), and less so for the former (that is, the binomial distribution is a useful model, but it is extremely far from "the" exclusive discrete model). Here are a few different examples, with a brief mention of where they may arise.

5.1 Discrete Distributions

5.1.1 The Geometric Distribution

Suppose an event may occur with probability p. If we repeat it over and over, in an independent way, the time of first occurrence will be given by

$$P[N = k] = p(1-p)^{k-1}$$

If you are curious, you may inquire for the first time the event will have occurred 2, 3, ... m times. You will find the result in any book on probability, as the *hypergeometric* distribution

5.1.2 The Multinomial Distribution

The binomial distribution is nice, but it considers two outcomes only (say, win or lose). What if we have several different possible outcomes (for instance we repeatedly toss a die, in an independent way – what is the probability of having a certain number of 1's, of 2's, ...? The formula is a slight complication of the binomial formula: if you have possible outcomes $a_1, a_2, \ldots, a_m$, each with probability $p_1, p_2, \ldots, p_m$and you repeat this experiment n times, you'll find that the probability of having n_1 times the outcome a_1, n_2 times the outcome a_2, and so on, will be given by the formula (not that, necessarily, $n_1 + n_2 + \cdots n_m = n$

$$\frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

For $m = 2$ this is just the binomial formula.

5.1.3 The Poisson Distribution

This is a distribution associated with "rare events". In fact, it can be deduced as the limiting case (in an appropriate sense) of the binomial distribution when the probability of one of the outcomes is very small, and the number of attempts is very large. The classical example is "the number of calls arriving at a switchboard over a fixed amount of time". This distribution is extremely useful in crucial applications since, for example, it is a simple but not unrealistic model

for the number of requests to use a resource in a network (a computer network, an electrical network, the number of customers arriving at a teller, ...). It turns out that the number N in question has distribution

$$P[N = k] = e^{-\lambda}\frac{\lambda^k}{k!}$$

where λ is a parameter that, in a sense, is connected the average time between requests (the higher the value of λ the more intense the flow of requests). This is a probability distribution because of the remarkable formula (which won't make sense until you learn some calculus)

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda}$$

Incidentally, it turns out that $EN = \lambda$, and $Var(N) = \lambda$ as well!

5.2 Continuous Distributions

Depending on the applications, there are myriads of distributions that have been considered. A few examples follow.

5.2.1 The Exponential Distribution

This is the continuous analog of the Geometric Distribution above (and it is connected to it in a very precise mathematical sense). Here you have a possible model of "first time to an event", like the breakdown of a piece of equipment. One way of giving a formula is through

$$P[T > t] = e^{-\lambda t}$$

from which you can easily deduce that

$$P[a < T < b] = e^{-\lambda a} - e^{-\lambda b}$$

With more calculations than we can perform here, you would find that

$$ET = \frac{1}{\lambda}, Var(T) = \frac{1}{\lambda^2}$$

If this reminds you of the Poisson distribution's formulas, you are right: there is a strong connection between the two distributions: if the number of arrivals has a Poisson distribution, the time between arrivals has an exponential distribution, and the "λ" is the same!

5.2.2 The Weibull Distribution

This a variation to the exponential distribution, used in survival analysis, when the exponential model is not appropriate;

$$P[T > t] = e^{-\lambda t^{\alpha}}$$

where $\alpha > 0$ is another parameter

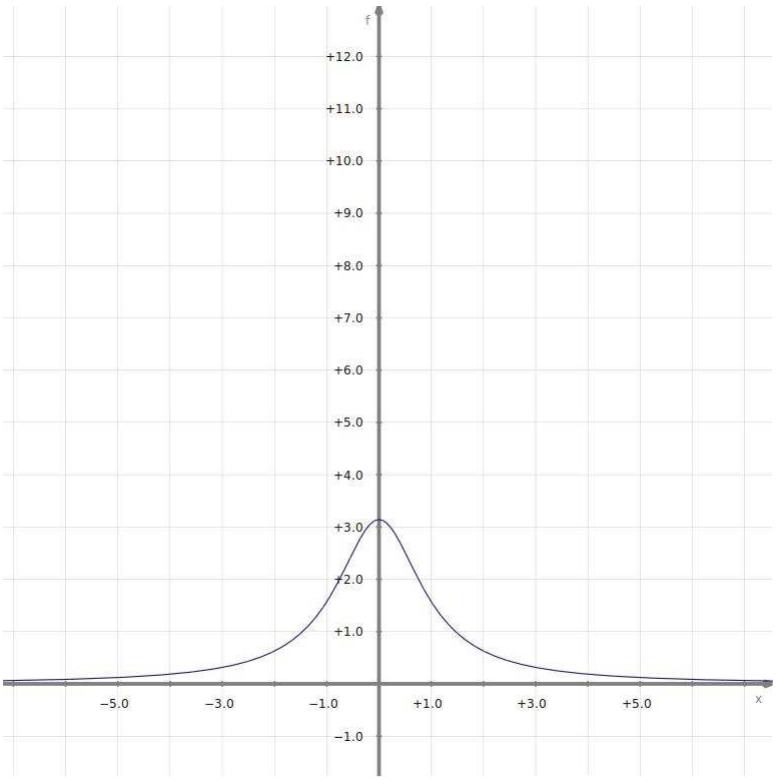
5.2.3 The Beta and the Gamma Distribution

Just as the exponential distribution concerns the first arrival time in a Poisson flow, and, in the discrete case, considering multiple arrivals led from the geometric to the hypergeometric distribution, in the "continuous" case the analog is the so-called Gamma distribution. Also, in a number of problems more or less connected to the same setup, another distribution comes up, called the Beta distribution. We won't be concerned with these more complex cases.

5.2.4 The Cauchy Distribution

Once you allow for "infinitely many" outcomes, you can't be really sure that the parameters we defined, EX, EX^2, and so on make sense. In fact, there are many examples where they don't. These examples have become of greater interest since the explosion of "Financial Mathematics", where the modeling of the distribution of stock prices has led to consider some of these "exotic" examples. The grand-daddy of these examples is the *Cauchy* distribution. This is a distribution where the probability of the random variable to fall between a and b is given by the area under the curve

$$\frac{1}{\pi} \cdot \frac{1}{1+x^2}$$



If you think this looks very much like the Normal Distribution, you are only seeing a superficial similarity. In fact, while for a normally distributed variable X, with parameters μ and σ,

$$EX = \mu, Var\left(X\right) = \sigma$$

attempts to compute expected value and variance for a Cauchy variable are fruitless: the numbers become infinitely large...