Review of Probability Without Measure Theory

This material is taken from a book I started writing a few years ago on measure theory.

Introduction

This chapter has a somewhat different purpose to its two predecessors. The aim is to review probability theory as it is generally taught in school and first year university courses. We emphasise the word "review". The author's assumption is that you have met this subject before, and that you would appreciate a refresher course. This is not the place to learn the subject from scratch. At this level, probability is not a rigorously grounded subject that can be compared in its logical cohesion to set theory and analysis (as discussed in the previous chapters), or to say, group theory and geometry. It is one of the tasks of this book to demonstrate that this can be done; but that will come later, after we've met some measure theory. As this chapter develops, we'll identify certain gaps in the presentation that need to be filled, and we will return to plug these gaps later in the book.

Probability theory is the mathematical theory of chance. It begins in the real world with the observation of phenomena that can be conceptualised within the language of mathematics. As shape and form give rise to geometry, change and motion inspire calculus and analysis, and symmetry motivates group theory; so chance underlies probability theory. It's worth emphasising, as this is often a source of confusion, that probability is a distinct subject to statistics. The latter is the science (or art?) of making inferences based on data. Of course, probability is an essential tool in statistics; but it also finds many other areas of application in a diverse range of topics, including biology, physics, economics, and even pure mathematics (particularly analysis and number theory)[1].

The Concepts of Probability

To make the study of chance phenomena into a mathematical theory, we begin with a set Ω of outcomes that chance may select from. We may imagine betting money on each of the individual outcomes, but we may also bet on events, which are combinations of outcomes, e.g. if $\Omega = \{1, 2, 3, 4, 5, 6\}$ represents the outcomes of throwing a fair die, then we may bet on the die showing an even number, and you would win if any outcome from the subset $A = \{2, 4, 6\}$ is thrown. So an event should be a subset of Ω. In general, the probability of an event should be a number lying between 0 (impossible), and 1 (certain).

[1] One can, and should, think about the probability that an arbitrarily selected natural number is prime.

The above suggests that we should consider probability as a set function, i.e. as a mapping $P : \mathcal{P}(\Omega) \to [0,1]$. A very important property that probability must have is suggested by experience: it should be additive on disjoint events, i.e. for $A, B \in \mathcal{P}(\Omega)$, with $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B). \tag{0.0.1}$$

Since the empty set is an event that can never happen, we must surely have

$$P(\varnothing) = 0. \tag{0.0.2}$$

If we treat (0.0.1) and (0.0.2) as defining axioms for probability, we can go quite a long way. In particular we can easily derive

$$P(A^c) = 1 - P(A), \text{ and } P(\Omega) = 1.$$

In principle, S can be any set; for example, we can take it to be the uncountable set $[0,1]$. Let us imagine that we choose a number at random from this set so that all numbers are equally likely. How can we formalise this? Is it reasonable to bet on our number coming from any subset of S, are are there subsets that are so strange that we could never select a number from them. For example, we spent some time in Chapter 1 studying the Cantor set. We proved that it is uncountable, contains no intervals, and yet it is perfect. Surely a subset that falls into a class that is labelled "perfect" should be able to occur in this random experiment? This leads to our first question.

Question 1 If Ω is uncountable, is $\mathcal{P}(\Omega)$ too large to support a theory of probability? If so, what can replace it, i.e. what structure (mathematically) best captures the notion of event?

We do not just want to assign probabilities to random events; but we also want to make quantitative measurements of variables that are of interest to us, and that are subject to the play of chance. For example, a fair coin is tossed M times in succession. Suppose that we want to bet on the number of heads that occur. This can be any number from 0 to M. We describe this succinctly by introducing the random variable X, which gives the number of heads in the M tosses. Then X is a mapping from Ω to $\mathbb{Z}_+$. To see why this is so, we first observe that we can take $\Omega = \{0,1\}^M$, where 0 and 1 represent "tail" and "head", respectively. Then a generic element of Ω is $\omega = (\omega_1, \omega_2, \ldots, \omega_M)$, where $\omega_i \in \{0,1\}$ $(i = 1, 2, \ldots, M)$. A precise description of X is

$$X(\omega) = \sum_{i=1}^{M} \omega_i.$$

If all outcomes are equally likely, we can calculate the probabilities:

$$P(X = r) = \binom{M}{r} \frac{1}{2^M},$$

for $0 \leq r \leq M$; indeed, this is a special case of the binomial distribution (see below). Writing $P(X = r)$ is a natural notation for the probability that x takes the value r. But since P can only be applied to sets, then we may enquire what this really means? We define $P(X = r)$ to mean $P(X^{-1}(\{r\})$, where $X^{-1}(\{r\})$ is the inverse image of $\{r\}$ under the mapping X, i.e.

$$X^{-1}(\{r\}) = \{\omega \in \Omega; X(\omega) = r\},$$

and you should convince yourself that this agrees with your intuition.

More generally, we may define a random variable to be a function from Ω to $\mathbb{R}$. For example, we may want to consider X to be the position of a randomly moving molecule that takes a continuous range of values on the whole real line. Then we may want to calculate the *law* or *distribution* of the random variable:

$$p_x(A) = P(X \in A) = P(X^{-1}(A)),$$

where $A \in \mathcal{P}(\mathbb{R})$ is a "suitable" set. This brings us back to the territory of question 1. Giving a precise meaning to the vague idea of "suitable set" is one of the key tasks of measure theory.

Note that the set of all random variables on Ω forms a (real) vector space in that if X and Y are two random variables, and $a, b \in \mathbb{R}$, then $aX + bY$ is another random variable, whose value at $\omega \in \Omega$ is $aX(\omega) + bY(\omega)$.

Question 2 Is the definition of random variable adequate? Is it enough to consider any function from Ω to $\mathbb{R}$, or should we restrict to functions that have some nice properties?

To proceed further, we should distinguish between discrete and continuous random variables. A random variable is said to be *discrete* if its range is at most countable. In this case, everything we need to know, in order to compute probabilities involving X, is contained within the *probability mass function*:

$$p_r = P(X = r) \text{ for } r \in R_X,$$

where R_X is the range of X, i.e.

$$R_X = \{r \in \mathbb{R}; r = X(\omega), \text{for some } \omega \in \Omega\}.$$

Note that for all $r \in R_X$,

$$0 \leq p_r \leq 1 \text{ and } \sum_{r \in R_X} p_r = 1.$$

We now present some of the best known examples of random variables. In fact each of these is a family of random variables, labeled by a parameter that typically has a probabilistic interpretation. In some cases, we will learn more about such parameters in the next section.

Examples of Discrete Random Variables

1. *Discrete Uniform.* Here the parameter is $N \in \mathbb{N} \setminus \{1\}$. The random variable X has range $\{1, 2, \ldots, N\}$ and $p_r = 1/N$ for all $r = 1, \ldots, N$. This describes selecting an item at random from N equally likely alternatives.

2. *Bernoulli.* Here the parameter is $0 < p < 1$. The random variable X has range $\{0, 1\}$ and $p_1 = p$, so that $p_0 = 1 - p$. This describes, for example, individual tosses of a biased coin. The special case of $p = 1/2$ is called a *symmetric Bernoulli random variable.*

3. *Binomial.* Here there are two parameters: $0 < p < 1$ and $n \in \mathbb{N}$. The random variable X has range $\{0, 1, \ldots, n\}$ and

$$p_r = \binom{n}{r} p^r (1-p)^{n-r}.$$

   This may describe, for example, the inspection of n items in succession, each of which is defective with probability p. Then X is the total number of defective items that are found.

4. *Poisson.* Here there is again one parameter $\lambda > 0$. The random variable X has range $\mathbb{Z}_+$ and

$$p_r = \frac{\lambda^r}{r!} e^{-\lambda}.$$

   This may describe a succession of random events where it is very unlikely that two or more events will occur close together, e.g. X could be the number of particles emitted by a radioactive source after a specified time.

A random variable is *continuous*[2] if its range is uncountable For example, consider two sticks of equal length which are nailed together at one end in

[2]The terminology is unfortunate, but it is so well established that we will have to live with it. Be aware that continuous random variables have nothing to do with continuous functions. We have two distinct uses of the word "continuous".

such a way that one of the sticks is fixed, and the other may rotate freely around it. If we apply a force to the stick that is free to rotate, when it finally comes to rest, the angle that it makes with the first stick is a random variable which takes values in the interval $[0, 2\pi)$.

In all such cases, the notion of probability mass function breaks down. We argue that it is impossible to locate an isolated point on the real number line with absolute precision, and so

$$P(X = x) = 0,$$

for all $x \in R_X$. In this case, a more useful tool is the *cumulative distribution function* (or c.d.f., for short), defined by

$$F_X(x) := P(X \leq x) = P(X^{-1}([x, \infty))),$$

for $x \in \mathbb{R}$. We can calculate all the probabilities that we need using F_X, e.g.

$$P(a \leq X \leq b) = F_X(b) - F_X(a).$$

Of course we could have also brought in the c.d.f. in the discrete case, e.g. if $R_X = \{a_n, n \in \mathbb{N}\}$ with $a_{n+1} \geq a_n$ for all $n \in \mathbb{N}$, then writing $p_i = p(a_i)$ for convenience, we have

$$F(a_n) = \sum_{i=1}^{n} p_i,$$

and for $i < j$, $P(a_i \leq X \leq a_j) = F(a_j) - F(a_{i-1})$.

Question 3. How fundamental is the division between discrete and continuous random variables? How far can we develop the general notion, without having to choose either of these alternatives?

Returning to the case of a general continuous random variable, in many examples of interest, the c.d.f. F_X is determined by a more fundamental object, called the *probability density function* (or p.d.f. for short), that is denoted by f_X. The relationship between these is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy,$$

and if f_X is continuous at x, then by Theorem 1.7.1 (1), F_X is differentiable there and

$$f_X(x) = F'(x).$$

If f_X is a p.d.f., we must have $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, and $\int_{-\infty}^{\infty} f_X(y)dy = 1$.

Question 4. How fundamental is the notion of p.d.f. Do all continuous random variables have a p.d.f., or do we have to restrict to a subclass. If such a subclass exists, how is it characterised?

Important examples are the *uniform distribution* U on $[a, b]$ where the p.d.f. is

$$f_U(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b]. \end{cases},$$

and the *normal distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, where for all $x \in \mathbb{R}$,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}. \tag{0.0.3}$$

If X has a normal distribution with parameters μ and σ as in (0.0.3), it is common to use the notation $X \sim N(\mu, \sigma^2)$. Of particular importance, is the *standard normal*, which is usually denoted Z. It is "standardised" to have $\mu = 0$ and $\sigma = 1$, i.e. $Z \sim N(0, 1)$ and for all $x \in \mathbb{R}$,

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

Note that if $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0, 1)$ where $Z = \frac{X-\mu}{\sigma}$. The importance and applicability of the normal distribution will be discussed in section 2.4.

In both the discrete and continuous cases, it is very useful to consider functions of random variables. Suppose that X is a random variable having range $S \subseteq \mathbb{R}$. We do not need to specify if it is discrete or continuous for this discussion. Let $g : S \to \mathbb{R}$ be a function. Then $g(X)$ is the random variable given by $g(X)(\omega) = g(X(\omega))$, for each $\omega \in \Omega$. As a simple example, let us return to the random angle θ obtained by rotations of a free rod about a fixed rod. Suppose that both rods have given length l. We may be interested in the area $A(\theta)$ of the isosceles triangle obtained after rotation through the angle θ, and this is given by a function of θ, i.e.

$$A(\theta) = \frac{1}{2} l^2 \sin(\theta).$$

Question 5. Is it legitimate to choose any function g to obtain a random variable $g(Y)$?

Expectation and Variance

This section seeks to investigate how we can attach numbers to random variables that summarise some essential features. For example, suppose that

we have a random variable taking a range of different values, with various probabilities. One intuitively appealing way of of obtaining some information that can be expressed by a single number, is to average out over all possible outcomes. To be precise, let X be a discrete random variable taking the values $a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, \ldots, p_n$. Then the *expectation* of X is defined to be

$$\mathbb{E}(X) = \sum_{i=1}^{n} a_i p_i. \tag{0.0.4}$$

If the range is infinite, then $\mathbb{E}(X) = \sum_{i=\infty}^{n} a_i p_i$, only makes sense if the infinite series converges. Returning to the finite case (0.0.4), if X has the discrete uniform distribution, then

$$\mathbb{E}(X) = \frac{1}{n}(a_1 + a_2 + \cdots + a_n),$$

which is the very simple form of average that we first encounter at school, say. More generally, (0.0.4) is a "weighted average", in that outcomes with high values and high probability will contribute more in the calculation than those with low values and low probabilities.

In the continuous case, writing F_X for the cdf of X , we obtain

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x dF_X(x), \tag{0.0.5}$$

which, since F_X is monotonic increasing, has meaning as a Stieltjes integral, provided that it exists. If X has a pdf f_X then we can rewrite (0.0.5) as a Riemann integral:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Note that if X and Y both have finite expectation, then linearity holds in that for all $a, b \in \mathbb{R}$,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

The number $\mathbb{E}(X)$ is often referred to as the *mean*[3] of the random variable X and denoted μ, or μ_X.

We may also average functions of random variables $g(X)$. As these are themselves random variables, the same formulae apply, i.e. (0.0.4) becomes

$$\mathbb{E}(g(X)) = \sum_{i=1}^{n} g(a_i) p_i,$$

[3]It should not be confused with the notion of *sample mean* in statistics, which is a different concept.

while (0.0.5) yields

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) dF_X(x).$$

An important case is obtained by taking $g(x) = x^2 - \mu$. This leads to the *variance* of X, which is denoted $\mathrm{Var}(X)$, and defined by

$$\mathrm{Var}(X) = \mathbb{E}((X - \mu)^2).$$

Note that $\mathrm{Var}(X) \in [0, \infty]$. Some straightforward algebra yields the useful formula:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2.$$

It is not hard to see that for all $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

The variance of X measures the average deviation of a random variable from its mean. If $\mathrm{Var}(X)$ is a small number, we expect that the values of X are clustered close to μ with high probability; if $\mathrm{Var}(X)$ is large, the values of X may be scattered far from μ with high probability. Just as we write μ, or μ_X for $\mathbb{E}(X)$, so we write σ^2 or σ_X^2 for $\mathrm{Var}(X)$. The *standard deviation* σ, or σ_X is precisely $\sqrt{\mathrm{Var}(X)}$ and has the advantage of being measured in the same units as the values of X. The table below lists the mean and variance of some standard random variables.

Random Variable	Mean	Variance
Bernoulli (p)	p	$p(1-p)$
Binomial (n, p)	np	$np(1-p)$
Poisson (λ)	λ	λ
Uniform on $[a, b]$	$(b+a)/2$	$(b-a)^2/12$
Normal (μ, σ^2)	μ	σ^2

There are other important quantities, apart from the mean and variance, that can be obtained by taking g to be a suitable function and calculating $\mathbb{E}(g(X))$. These include:

1. The *nth moment* of X, which is $\mathbb{E}(X^n)$, obtained by taking $g(x) = x^n$ for $n \in \mathbb{N}$,

2. The *moment generating function* of $X, M_X(t) := \mathbb{E}(e^{tX})$, obtained by taking $g(x) = e^{tx}$, for $t \in \mathbb{R}$,

3. The *characteristic function* of $X, \Phi_X(t) := \mathbb{E}(e^{itX})$, obtained by taking $g(x) = e^{itx}$, for $t \in \mathbb{R}$, where $i = \sqrt{-1}$.

The Cauchy random variable, which has range $\mathbb{R}$ and pdf. $f(x) = \frac{1}{\pi(1+x^2)}$, for $x \in \mathbb{R}$ is an example of a random variable that does not have a mean, i.e. the Riemann integral $\mathbb{E}(X)$ does not exist.

In physics, averaging over microstates yields emergent macroscopic behaviour, e.g. we model a gas as a collection of random motions of individual molecules; the concept of temperature has no meaning at the molecular level; we can only make sense of it as a bulk property of the gas as a whole, after we've averaged out all of the different molecular motions. We can similarly describe the behaviour of crowds of people, and flocks of birds by using averages that have no meaning at the individual level.

Independence, Sums of Random Variables, Limit Theorems

Consider two events, A and B having probabilities $P(A)$ and $P(B)$, respectively. Assume for now that $P(B) > 0$. To investigate the relationship between A and B at the level of probabilities, we might ask how the probability of A would change if we knew that B had occurred. To answer this question, we introduce the *conditional probability* of A given B, which is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \tag{0.0.6}$$

We might suspect that something special has happened if $P(A|B) = P(A)$, i.e. the occurrence of B has no effect on the probability of A. Simple algebra from (0.0.6) tells us that if both $P(A), P(B) > 0$ then $P(A|B) = P(A)$ if and only if $P(B|A) = P(B)$. We then say that the events A and B are *independent* if

$$P(A \cap B) = P(A)P(B). \tag{0.0.7}$$

Though less intuitive than using conditional probability, (0.0.7) has the advantage of treating A and B symmetrically. It also continues to be valid when either (or both) of $P(A)$ or $P(B)$ vanish. The essence of (0.0.7) is captured in the memorable phrase:

Independence means multiply.

We similarly say that two random variables X and Y are independent if for any suitable subsets of $\mathbb{R}$, A and B, the events $(X \in A)$ and $(Y \in B)$ are independent, i.e.

$$P((X \in A) \cap (Y \in B)) = P(X \in A)P(Y \in B).$$

The left hand side of the last display is a little clumsy, and we will generally replace it with the notationally simpler, $P(X \in A, Y \in B)$. If X and Y are independent, it can be shown that

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \text{ and } \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \tag{0.0.8}$$

We might also want to discuss independence for finite and countable sets of events and of random variables. We will deal with the countable case here, from which the restriction to the finite is obvious. We say that a family $\{A_n, n \in \mathbb{N}\}$ of events is *independent* if

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_N}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_N}),$$

for every finite subset $\{i_1, i_2, \ldots, i_N\}$ of $\mathbb{N}$. Similarly a sequence $(X_n, n \in \mathbb{N})$ of random variables is said to be independent if every family of events $\{(X_n \in A_n); n \in \mathbb{N}\}$ is independent, in the sense just give, for all possible choices of suitable families of sets $\{A_n, n \in \mathbb{N}\} \in \mathcal{P}(\mathbb{R})$.

Two or more random variables are said to be *identically distributed* if they have the same law, e.g. X and Y are identically distributed if

$$P(X \in A) = P(Y \in A)$$

for all $A \in \mathcal{P}(\mathbb{R})$. It is of great theoretical and practical interest to investigate sequences of random variables $(X_n, n \in \mathbb{N})$ that are both independent and identically distributed (usually written i.i.d. for short). This can for example describe idealised models of measurements from a sequence of experiments that are carried out under identical conditions, wherein the outcome of any particular experiment cannot influence our prediction of the outcome of any other. Note that if the mean and variance of X_n exist for some particular n, then they will exist and take the same values for all $n \in \mathbb{N}$.

Now consider the associated sequence of partial sums $(S_n, n \in \mathbb{N})$ defined for each $n \in \mathbb{N}$ by

$$S_n = X_1 + X_2 + \cdots + X_n.$$

If we think of X_n as being the random kinetic energy of a molecule of a gas, then S_n is the total kinetic energy of n molecules, and as n gets larger and larger, we should get more and more information about the macroscopic behaviour of the gas as a whole, such as its temperature.

If $\mathbb{E}(X_n) = \mu$ and $\mathrm{Var}(X_n) = \sigma^2$ for all $n \in \mathbb{N}$, we can deduce by linearity of expectation, and the second formula in (0.0.8) that

$$\mathbb{E}(S_n) = n\mu \text{ and } \mathrm{Var}(S_n) = n\sigma^2. \tag{0.0.9}$$

We may also consider the "empirical average", and define

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Note that $\overline{X}_n$ is itself a random variable, which in the physical scenario considered above, describes the kinetic energy of a "typical" or "average" particle within the ensemble of molecules.

By (0.0.9) we have

$$\mathbb{E}(\overline{X}_n) = \mu \text{ and } \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{0.0.10}$$

In particular the standard deviation of $\overline{X}_n$ is $\sigma/\sqrt{n}$, and this quantity is called the *standard error* by statisticians.

Limit Theorems

It is very important to be able to obtain information about S_n and $\overline{X}_n$ as n becomes very large. For example as n gets larger and larger, we begin to approximate the gas more accurately from its collection of constituent molecules. There are three basic *limit theorems* that give the asymptotic behaviour of these sequences some mathematical substance; these being the weak law of large numbers (WLLN), the strong law of large numbers (SLLN), and the central limit theorem (CLT). We will briefly discuss the SLLN and CLT here, and give a fuller, more rigorous treatment in Chapter ? Let us first state the SLLN. It says that

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

This is interesting for a number of reasons. Firstly, one might expect that, in general, the limit of a sequence of random variables would be a random variable, but here we get a number, and that number has a meaning. Indeed, in the physical example discussed above, it tells us that as we aggregate more and more molecules, the typical behaviour of a molecule can be found by just going back to any one of the constituent molecules, and finding its average behaviour. WLLN is a weaker statement than SLLN, but it is considerably easier to prove. We will not state it here, but it will be established, in all its glory, in Chapter?

The SLLN tells us nothing about the distribution of the fluctuations $\overline{X}_n - \mu$ as n becomes large. It is convenient to measure these random variables in units of the standard error $\sigma/\sqrt{n}$, and so we define a new sequence of random variables $(Y_n, n \in \mathbb{N})$ by

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

These are *standardised* in that $\mathbb{E}(Y_n) = 0$ and $\mathrm{Var}(Y_n) = 1$, for all $n \in \mathbb{N}$.

The CLT is deep and beautiful. It is one of the most profound and useful theorems in elementary mathematics. Recall that $Z \sim N(0,1)$ is the standard normal. Then the CLT tells us that the for all $-\infty \leq a < b \leq \infty$,

$$\lim_{n\to\infty} P(a \leq Y_n \leq b) = P(a \leq Z \leq b).$$

Equivalently,

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dx.$$

In words, as n gets larger and larger, then the distribution of Y_n gets closer and closer to that of the standard normal Z. A key observation here is that there was no assumption on normality on the "input random variables" $(X_n, n \in \mathbb{N})$ that we started with. They could have any distribution we like, discrete or continuous, binomial, Poisson, uniform, or any other distribution that we might consider. The only constraint is that they must have a finite mean and variance (so the Cauchy distribution is ruled out). Thus the normal distribution is revealed to have a universal character, as the attractor for the distribution of normalised sums of i.i.d. random variables with finite mean and variance. This is a good explanation as to why it is so important, in both theory and applications.

Application to Statistical Inference – Confidence Intervals

In statistics we seek to learn about a *population* which is some quantitative attribute of a large number of people, fish, machines or even cans of beans. Let N be the size of the population. We treat the population as unknowable, as it is either too large, inaccessible or expensive for us to individually measure each element of it. From the population, we extract a *sample* $\{x_1, x_2, \ldots, x_n\}$, where n is much smaller than N. We introduce a random variable X whose range is precisely the values of the unknown population. The distribution p_X of X should also capture the frequency with which various values occur in the population. Of course we do not know p_X either. Finally we can, at least in principle, define the *population mean* $\mu = \mathbb{E}(X)$, and the *population variance* $\sigma^2 = \mathrm{Var}(X)$. In this section, we will for simplicity, make the assumption (which may be quite unrealistic in general) that σ^2 is known. It is not a difficult task to drop that assumption, but it would make this section longer and more complicated to do so. Our task is to gain as much information as we can about the unknown μ. The *sample mean*

$$\overline{x}_n := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

may be a crude and unreliable estimate of μ. In statistics, it is called a *point estimate*. We will seek to use our knowledge of probability theory to sharpen this to an *interval estimate*, i.e. an interval in which μ lies with high probability.

Let us consider the process of choosing the sample from the population. In principle we might choose any member of the population to be x_1. But the distribution p_X will determine which values are more likely than others to be chosen. To quantify this, we introduce a new random variable X_1, which is simply a copy of X, and take X_1 to be the random variable that describes the act of choosing x_1. Similarly X_2 will be the random variable that chooses x_2, and thus we generate a set $\{X_1, X_2, \ldots, X_n\}$ of random variables. Although the population decreases by one every time we select a sample point, if N is sufficiently large, it is reasonable in a simple idealised model to choose $X_1, X_2, \ldots, X_n$ to be identically distributed. We will also regard them as independent. In that case the random variable $\overline{X}_n$ has an interesting interpretation: its range consists of all possible values of the sample mean $\overline{x}_n$ which are obtained by considering all possible samples of size n. By the central limit theorem, we know that if n is sufficiently large then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \text{ is approximately } N(0,1),$$

and in practice, statisticians are relaxed about taking $n = 30$. To construct an interval in which μ lies with high probability, we begin by choosing that probability (or at least, one minus it). Let α be a small number (typically $\alpha = 0.01, 0.05$ or 0.1). Using normal distribution tables, or a software package such as R, we can find the *critical value* $z_{\alpha/2} > 0$ so that

$$P(Z > z_{\alpha/2}) = \alpha/2,$$

for example, if $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$. By symmetry of the standard normal, $P(Z < z_{\alpha/2}) = \alpha/2$, and so

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

But from the central limit theorem, we can write (being a little cavalier about approximation

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha,$$

and some straightforward algebra then yields

$$P\left(\overline{X}_n - \frac{\sigma z_{\alpha/2}}{\sqrt{n}} < \mu < \overline{X}_n + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha. \tag{0.0.11}$$

The inequalities (0.0.11) give a *probability interval* for (0.0.11). Notice how our ignorance of the distribution p_X has not been an obstacle to obtaining (0.0.11), due to the universal character of the normal distribution. However (0.0.11) cannot be used in practice, as it involves the random variable $\overline{X}_n$. But statisticians are happy to replace this with a number that they can measure precisely - the sample mean $\overline{x}_n$. We then say that a $100(1-\alpha)\%$ *confidence interval* for μ is

$$(\overline{x}_n - \sigma z_{\alpha/2}/\sqrt{n}, \overline{x}_n + \sigma z_{\alpha/2}/\sqrt{n}).$$

We will not say more here about the justification or further theory of confidence intervals. The main purpose of this section has been to illustrate how important the central limit theorem is in enabling us to build sensible schemes for estimating unknown quantities. So it is highly desirable to find a mathematically rigorous proof of the CLT, and this will be a major focus of Chapter ?.