Notes on Chapter 5

Continuous RVs as limits

The basic idea is that "every RV is discrete", and continuous RVs are limits. It was Kolmogorov's genius to recognize that abstract measure theory is tailor made for this purpose.

Specifically we have to assume that $X\colon \Omega \longrightarrow E$ maps a sample space into a measurable space, with some well defined $\sigma-$ algebra $\mathfrak{B}$. We need to make sure that, for any $B \in \mathfrak{B}$, $X^{-1}(B) \in \mathfrak{F}$, where $\mathfrak{F}$ is the $\sigma-$ algebra of events of Ω (we say that X is a *measurable* function, or, equivalently, a *random variable*). The idea is then to find a sub-collection of sets B in $\mathfrak{B}$, small enough that it is feasible to assign a (consistent) probability to all events of the form $\{X \in B\}$, and large enough so that it is possible to generate all other sets in $\mathfrak{B}$ through appropriate limits of such sets, allowing to extend by continuity a probability to all events of the form $\{X \in C\}$, where C is *any* set in $\mathfrak{B}$.

The logic of this approach is that *we assume we can reasonably assign probabilities to "simple" events*, where "simple" could mean, for example, to events defined by a convenient collection of subsets of E. Any other probability must be recoverable from this data. Hence, assuming we are good at taking limits, the necessity to restrict to collection of events that form a $\sigma-$algebra, and, more importantly, to probabilities that are $\sigma-$ additive (that is, such that for a sequence of disjoint events A_j, $P[\cup A_j] = \sum P[A_j]$). This last requirement allows us to define probabilities in a consistent way through limits. Dropping this requirement for probabilities leads to interesting theories (naturally, more constrained than "standard" probability theory), but is of little help in addressing actual applications.

The prototype of this construction is when $E = \mathbb{R}$, and we choose as $\mathfrak{B}$ the smallest $\sigma-$algebra to contain the intervals (it doesn't really matter if we start with open, closed, half-open, etc. intervals). This $\sigma-$ algebra is called the *Borel* $\sigma-$ algebra, and, if we consider the random variable X, uniform on $[0, 1]$ (i.e., such that $P[\{a < X < b\}] = b - a$, for any $0 \leqslant a \leqslant b \leqslant 1$) it turns out that what we end up with is that the probability of any event $\{X \in C\}$, for any *Borel set* C is the *Lebesgue measure of* C (since we require $\sigma-$ algebras, we end up with a pretty vast collection of acceptable sets, some of which are not very intuitive).

You may notice that the immediate reduction to *absolutely continuous* RVs, as presented in the book, has a logical flaw: we haven't any proof that the various distributions we see are in fact legitimate distributions! To be reassured, we should have shown that the probability assignments do indeed extend nicely to the smallest $\sigma-$ algebra containing the intervals, and that the resulting probability is $\sigma-$ additive (or, equivalently, continuous) over this $\sigma-$algebra. This is in fact the case, but we would need Lebesgue measure and integration theory to prove this.

Functions of Random Variables

There is another finesse that you may notice in the way the book handles "functions of a random variable" (page 242 and following, for the one-dimensional case, and 300 and following for the vector case). First of all, examples 7b and 7c do not satisfy the condition of theorem 7.1 that follows immediately, since the functions playing the role of g in the theorem are not invertible at all, and the second is not even differentiable. Second, theorem 7.1 (and the corresponding multi-dimensional version) seems a bit overly restrictive on its face: are we going to consider only invertible functions of random variables? Of course, not: we just looked at examples that are not.

The problem lies in being constrained into the very tight realm of Riemann integration theory to do anything. Examples 7b and 7c are easily decoded by noticing that the functions g are *piecewise* invertible (so theorem 7.1 has to be tediously applied to its pieces, which then have to be glued together), and that, even though the absolute value function is not differentiable, it is so only in one point, and isolated points "do not count" when dealing with continuous RVs. However, things would get much more intricate in the multidimensional case, so it might be worth checking what we can say in general.

Suppose we have a random variable $X\colon \Omega \longrightarrow E$, and a *measurable function* $g\colon E \longrightarrow F$ ("measurable" means, just as it did above, that the counter-image of any set in the appropriate $\sigma-$algebra of F, belongs to the appropriate $\sigma-$algebra of E). Then

$$P[g(X)\in K]=P\big[X\in g^{-1}(K)\big]$$

Note that the set $g^{-1}(K)$ is well-defined – the only problem is that it may be very complicated, if g is. Such a "change of variable" equation is, in general, far from trivial. In the very special case when g is invertible (except, possibly, at a "small" set of points – we won't go into the technical details), and, additionally, the setup is "easy", we can go and use "elementary" calculus to get us through. For example, if E and F are the real line, g is continuous, and K is an interval, then $g^{-1}(K)$ is also an interval, with endpoints mapped to endpoints. Still, we also need g^{-1} to have a continuous derivative to come up with an explicit formula (which, by the way, ensures that the monotone function g is strictly monotone).

The Exponential Distribution as the Limit of the Geometric Distribution

We can show that the same construction the leads from the binomial to the Poisson distribution leads form the geometric to the exponential distribution.

In fact, suppose we have a sequence of independent Bernoulli variables, X_1, X_2, ..., all with parameter p. Suppose now, we consider the case when $p=\frac{\lambda}{n}$ (λ is constant, as $n\longrightarrow\infty, p\longrightarrow 0$), and label each variable by $t=\frac{k}{n}$, instead of k, for a fixed n (but as n grows, we "move" k as well, so as to keep its relative position over the whole sequence of observations fixed). For example, we measure time in very short intervals, of length $\frac{1}{n}$, and look at the time of first success, which will be some $t=\frac{k}{n}$, a multiple of our basic unit of time. If T is the "time" at which we have the first "success",we have

$$P\left[T>t=\frac{k}{n}\right]=(1-p)^k=\left(1-\frac{\lambda}{n}\right)^{nt}$$

and the limit, as $n\longrightarrow\infty$, we end up with

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

which is the "survival function" for the exponential distribution. There is more to this observation. Let's discuss this in a slightly informal way. Consider a binomial experiment, in the Poisson limit. The time to the first "success" is the (random) time T, such that, for $t<T$, the Poisson variable is $N_t=0$, while $T_1=\inf\{t\colon N_t>0\}$. Now, if we look at what happens after time T, at least as long as we look at the binomial model, "things start afresh", since all trials are independent. So, the time to the *second* success T_2, will be some time after T_1, call it τ_1, that will be distributed just like T_1, since the circumstances are the same. We conclude that the time to the nth success will be distributed like $S_n=T_1+\tau_1+\tau_2+...+\tau_{n-1}$, where all variables have exponential distribution with parameter λ. This could be verified by working through the corresponding limit theorem for the *negative binomial* distribution (section 4.8.2 in the book). It also has interesting consequences for the distribution of this sum. In fact, you can see immediately that $\{S_n\geqslant t\}=\{N_t\leqslant n\}$. Since the probability of the right hand side is the cdf of the Poisson distribution, $\sum_{k=0}^{n}\frac{\lambda^k}{k!}e^{-\lambda}$, and, on the right hand side, there is no loss of generality in substituting $>$ for $\geqslant$ (S_n is obviously a continuous random variable), we have found a way to write the distribution of the sum of n i.i.d. exponential random variables.

Remark 1. You should really check out Example 5c, on page 232, as it is an extremely interesting example of application of the exponential distribution to a classic problem in *Queuing Theory*, another major application area of probability.

"At Random" Becomes Even More Vague

Example 3d, on page 217, illustrates two ways of solving "Bertrand's Paradox", whose basis is the undetermined meaning, in its context, of the phrase "at random". To make matters worse, there is even a third way (which is discussed, for example, in the old classic textbook on probability by B.V. Gnedenko):

> Another way to determine the position of the chord is to indicate the position of its midpoint. For the chord to satisfy the condition, the midpoint has to lie within a circle, concentric to the original one, whose radius is one half of the original one. Such a circle has an area equal to $\frac{1}{4}$ pf the big circle, so the "requested probability is equal to $\frac{1}{4}$".

The point of this "paradox" is that "at random" means very little, until you make precise what the sample space is, and how it is you specify the probabilities (in these old examples, the point is to specify which events are "equally likely").

Incidentally, this problem is one of the first introducing so-called "geometric probability" questions. A very classic example is Buffon's needle problem (example 6.2d, on page 270), which was historically introduced as a "Monte Carlo method" for the evaluation of π.