Common Probability Distributions for Simulation Modeling*

Jérémie Gallien
MIT Sloan School of Management
Cambridge, MA 02142

October 25, 2003

Which probability distribution should be used to most appropriately represent a given stochastic event? In this note I provide a short review of a few common distributions, mentioning for each some of the typical random phenomena it is often used to model.

In general, the process of identifying a distribution accurately representing a given data set is called *fitting*, and there are a number of statistical software tools performing this function. There is also a number of statistical tests designed to assess whether a given distribution faithfully describes a data set. However, in practice it may not always be possible to collect a large data set for each relevant source of variability because of time or financial constraints. In some cases, an extensive data collection may not even be useful, as the output of a given model may not be sensitive to the detailed probabilistic structure of the random input considered beyond basic mean and variance information. I thus believe that it is important to develop an intuitive feel for the likely stochastic structure of typical sources of variability in industry - hopefully this note will be helpful in that regard. For a (considerably) more exhaustive reference on theoretical properties of probability distributions and their uses in stochastic modeling, see the handbooks by Johnson, Kotz and Balakrishnan (ed. Wiley Interscience), and for a catalogue of distributions (and many other mathematical concepts/formulas) on the web, see http://mathworld.wolfram.com/. The book by Law and Kelton *Simulation Modeling and Analysis* also includes a description of probability distribution for simulation purposes.

1 Continuous Distributions

1.1 Uniform $U[a, b]$

1.1.1 Structure

Support (a, b), $F(x) = \frac{x-a}{b-a}$, $f(x) = \frac{1}{b-a}$, $E[X] = \frac{b-a}{2}$, $\sigma[X] = \frac{(b-a)^2}{12}$

Use This is the model of choice when the only information available about some random variable are the limits of its support. This distribution also has the property that all values within the support (a, b) are equally likely (hence its name). So this distribution is often used when very little information is available, but it is believed that the variability of the corresponding phenomenon is still an important feature to model. Finally, this distribution is in many settings is easy to deal with analytically, and it also plays an important theoretical role (generation of random numbers for other distributions in simulation software, distribution of arrival epochs of a Poisson process within a specified time interval).

1.2 Triangular $Tri[a, c, b]$

1.2.1 Structure

Support (a, b), $F(x) = \begin{cases} \frac{(x-a)^2}{(b-a)(c-a)} \text{ for } a \leq x \leq c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} \text{ for } c \leq x \leq b \end{cases}$, $f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} \text{ for } a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} \text{ for } c \leq x \leq b \end{cases}$,
$E[X] = \frac{1}{3}(a + b + c)$, $\sigma[X] = \frac{1}{3\sqrt{2}}\sqrt{a^2 + b^2 + c^2 - ab - ac - bc}$

1.2.2 Use

Like the uniform distribution, this distribution is used in cases when the limits of the support are known, but it allows in addition to model an asymetric probabilistic structure (values close to the mode parameter c are more likely). It is also useful when limited available information except a given finite support would suggest to use a uniform distribution, but the real phenomenon seems to exhibit less variability than is imposed by the uniform.

1.3 Exponential $Exp[\lambda]$

1.3.1 Structure

Support $(0, +\infty)$, $F(x) = 1 - e^{-\lambda x}$, $f(x) = \lambda e^{-\lambda x}$, $E[X] = \frac{1}{\lambda}$, $\sigma[X] = \frac{1}{\lambda}$

1.3.2 Use

This very important distribution can be used to model any phenomenon with a positive value and a known mean $\frac{1}{\lambda}$. In particular, it is often used to model time

to failure/breakdown for a component or a machine, time between two consecutive customer orders, spatial distance between two objects... Its fundamental property (and claim to fame) is the memoryless property, i.e. $P(X \geq x+t|X \geq t) = P(X \geq x)$. In words, how long you've waited already is completely irrelevant when trying to predict how much longer you will have to wait... The exponential distribution is the only continuous distribution with this property! A process of consecutive arrivals when the inter-arrival times are independent exponential distribution is also known as a Poisson process.

1.4 Normal $N(\mu, \sigma)$

1.4.1 Structure

Support $(-\infty, +\infty)$, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $E[X] = \mu$, $\sigma[X] = \sigma$

1.4.2 Use

This is the famous Gaussian or "Bell curve". It is used very widely to model any random variable with a symmetric structure and specified mean μ and standard deviation σ. A fundamental theoretical result associated with the normal distribution is the Central Limit Theorem (CLT), which is sometimes used abusively to justify the statement that "everything can be modeled as a normal", or to assume that any distribution is a priori normal. Because from the CLT the mean of a random sample quickly converges to a normal distribution, the normal distribution is also key to the construction of confidence intervals associated with simulation experiments and estimations. In part because of this, it is probably a good idea to keep in mind the fractile value $z'_{0.95} = 1.96$ for the standard normal distribution $Y \sim N(0, 1)$ defined by $P(|Y| \leq z'_{0.95}) = 0.95$ (in layman's terms: there is only a 5% chance that you will find yourself more than 1.96 standard deviations away from the mean).

2 Discrete Distributions

2.1 Fixed Value (Deterministic)

When you have good reasons to believe that the variability of a given phenomenon is very small compared to other factors and has only little impact on the output of your model, don't bother with a distribution!!! For example, in many manufacturing environment with a relatively high level of automation, the variability of individual machine processing times may be very small compared to demand or breakdown periods, and thus suitably represented by just a fixed number.

2.2 Bernouilli $Ber[p]$

2.2.1 Structure

Support $\{0,1\}$, $P(X=1)=p$, $E[X]=p$, $\sigma[X]=\sqrt{p(1-p)}$

2.2.2 Use

Whenever you are dealing with a binary event, this is the one! For instance, quality issues (piece is defective with a given probability), branching decisions (25% of parts go to machine 1, 75% to machine 2), representation of test power and accuracy through decision trees (with type I and type II errors), etc...

2.3 Geometric $Geom[p]$

2.3.1 Structure

Support $\{0,1,2,...,+\infty\}$, $P(X=k)=(1-p)^k p$, $E[X]=\frac{1-p}{p}$, $\sigma[X]=\frac{\sqrt{1-p}}{p^2}$

2.3.2 Use

Think of the russian roulette "game" when the gun barrel is spun again after every turn, or the number of successive bernoulli trials it takes to obtain a successful outcome. This is the discrete memoryless equivalent to the exponential distribution, in that $P(X=t+k|X\geq k)=P(X=t)$. Suppose that there is an independent probability of p that a system will crash at each period, then the number of periods before the crash follows $Geom[p]$. This can also represent for example the number of cycles that an item realizes in a process with a feedback loop, when there is an independent branching probability p that the item goes back through the loop again after each cycle. There is also a connection between this distribution and time discounting: If a given investment provides a constant return of S and the interest rate is r, a classical expression for the time-discounted stream of revenues is $\sum_{t=0}^{+\infty}\frac{S}{(1+r)^t}=S\frac{(1+r)}{r}$. This can also be interpeted as the expected total return when there is no interest rate, but at every period a probability $p=\frac{r}{1+r}$ that the stream of revenue will be terminated forever.

2.4 Binomial $Bin[N,p]$

2.4.1 Structure

Support $\{0,1,...,N\}$, $P(X=k)=\binom{N}{k}p^k(1-p)^{N-k}$, $E[X]=Np$, $\sigma[X]=\sqrt{Np(1-p)}$

2.4.2 Use

This is the distribution obtained when summing up N outcomes of Bernoulli variables with parameter p, and for this reason is very widely used. In the field

of quality for example, if the overall proportion of defective pieces is p, then the number of defective pieces in a sample of N will follow a distribution $Bin[N, p]$. If the independent probability that any plane will crash in a given year is p, then the number of crashes for a fleet of N aircrafts in that year will be $Bin[N, p]$... etc, etc.

2.5 Poisson $Poisson[\lambda]$

2.5.1 Structure

Support $\{0, 1, 2, ..., +\infty\}$, $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $E[X] = \lambda$, $\sigma[X] = \sqrt{\lambda}$

2.5.2 Use

The Poisson distribution is most frequently used to represent the occurence of a random number of events in a given time period. It is related to the Poisson process in the following way: Consider a stochastic arrival stream where the time between two consecutive arrivals follows an exponential distribution $Exp(\lambda)$ (this is the exact definition of a Poisson process), then the number of arrivals in any interval of length T follows a distribution $Poisson[\lambda T]$.