Lecture 1: A rapid overview of probability theory

Biology 429
Carl Bergstrom

January 2, 2008

Sources: These lecture notes draw upon material from Taylor and Karlin *An Introduction to Stochastic Modeling 3rd Edition* (1998), Parzen (1962) *Stochastic Processes*, Pitman (1993) *Probability*, Van Campen (1997) *Stochastic Processes in Physics and Chemistry*, and Dill and Bromberg (2003) *Molecular Driving Forces*. In places, definitions and statements of theorems may be taken directly from these sources.

Probability theory can be developed in a completely formal, rigorous fashion grounded in set theory and measure theory; this is known as *axiomatic probability theory.* Alternatively, one can be develop probability theory in a common-sense fashion, allowing our our basic intuitions about chance to replace much of the formalism. I will take the latter approach throughout.

1 Discrete probability

1.1 Random Variables

Definition 1 *A* random variable *is a variable that takes on its value probabilistically.*

A random variable X is defined by a set of possible values x, and a distribution of probabilities $p(x)$ over this set, such that

1. $p(x) \geq 0$ for all x, and
2. $\sum_x p(x) = 1$.

For example, if X is the random variable representing the outcome of rolling a fair die, the random variable X has a set of possible values $\{1, 2, 3, 4, 5, 6\}$ each with probability $p(x) = 1/6$. If X drawn is from distribution F, we write $x \sim F$. In this die example, X is a discrete random variable because it takes one of a set of discrete values. Random variables can also be continuous; they can take on any of a set of continuous values. Here we will treat discrete random variables first, and then move on to continuous random variables.

1.2 Events

Closely related to the notion of a random variable is the concept of an event.

Definition 2 *An* event *is the case that a random variable takes on a value within a described subset of possible values.*

For example, the event that I roll a die and get an odd number can be represented as the event $X \in \{1, 3, 5\}$. Events can include single values of random variables, e.g. the event that $X = 4$; they can include all possible values $X \in \{1, 2, \ldots, 6\}$, and they can include no possible values $X \in \{\}$.

If events $A_1, A_2, \ldots, A_n$ are mutually exclusive events with probabilities $P[A_i]$, the probability that any one of them occurs is

$$P[A_1 \text{ or } A_2 \ldots \text{ or } A_n] = \sum_i P[A_i].$$

1.3 Conditional probability

Conditional probability lets us talk about the chance that one event occurs given that another occurs.

Definition 3 *The* conditional probability *of A given B is*

$$P[A|B] = \frac{P[A \textit{ and } B]}{P[B]}$$

(We will write $P[AB]$ as shorthand for $P[A \text{ and } B]$). Note that we can also write

$$P[AB] = P[A|B]P[B]$$

We can extend this to write down a chain rule for probabilities. This rule tells us the probability that a series of events A_1, $A_2 \ldots$, A_n happen in succession.

$$P[A_1A_2A_3 \ldots A_n] = P[A_1]P[A_2|A_1]P[A_3|A_2A_1] \ldots P[A_n|A_{n-1} \ldots A_3A_2A_1]$$

Definition 4 *A* partition *is a collection of non-overlapping subsets of events* $\{B_1, B_2, B_3\}$ *that together cover the whole such that* $\sum_i B_i = 1$

Theorem 1 *The* Law of total probability *states that if* B *is a partition, the probability of an event* A *is the probability-weighted sum of the conditional probabilities of* A *given* B_i*:*

$$P[A] = \sum_i P[A|B_i]P[B_i]$$

1.4 Independence

We say that two events A and B are independent if the probability that A occurs does not depend on whether B has occured, and visa versa. That is, $P[A|B] = P[A| \text{ not } B] = P[A]$.

When two events are independent, the probability that both occur is simply the product of the probabilies that each occur.

$$P[AB] = P[A]P[B]$$

Similarly, we can extend this to more than 2 events. (Proof: define a new event as the event ”A and B”, then apply again.)

1.5 Bayes' Rule

Bayes' Rule allows us to work out the conditional probabilities of A given B from the conditional probability of B given A:

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

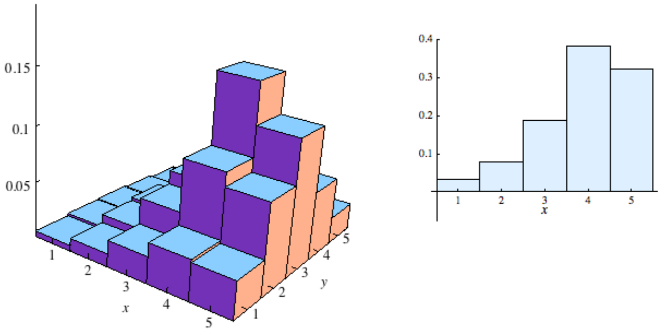
1.6 Joint Distributions and Marginal Distributions

The joint distribution F of two random variables X and Y can be written as

$$F[x, y] = P[X = x \text{ and } Y = y]$$

The marginal distribution $P_X[x] = \sum_i P[Y = y_i, X = x]$ gives us a single variable distribution on the values of X.

In the figure below, we see a the joint distribution (left panel) and marginal distribution function (with respect to X, right panel) for a simple probability distribution[1].



Notice that If X and Y are independent, $F[x, y] = P_X[x]P_Y[y]$.

1.7 Mean and variance

The expected value of a discrete random variable X is defined as:

$$E[X] = \sum_i x_i \, P[X = x]$$

Expectations are additive, even for random variables that are not independent. Thus

$$E[X + Y] = E[X] + E[Y]$$

[1] This image is a snapshot from the demonstration "Discrete Marginal Distributions". This is part of the The Wolfram Demonstrations Project, online at `http://demonstrations.wolfram.com/DiscreteMarginalDistributions/`

If two random variables are independent, the expectation of their product is equal to the product of their expectations:

$$E[XY] = E[X]E[Y] \text{ for } X, Y \text{ independent}$$

However, note that when X and Y are not independent, this is generally not true. Moreover the expectation of a function is not generally equal to the function of the expectation, i.e.,

$$E[f(x)] \neq f(E[x])$$

The variance of a random variable X is defined as

$$\mathrm{Var}[X] = E\left[(X - E[X])^2\right]$$

We can also write this as

$$\mathrm{Var}[X] = E[X^2] - E[X]^2$$

by expanding the definition of variance and recognizing that $E[XE[X]] = E[X]^2$.

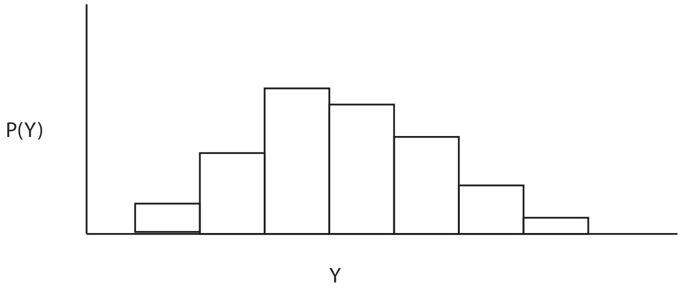
If two random variables are independent, the variance of their sum is equal to the sum of their variances.

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

2 Continuous probability

We can just about any continuous probability distribution as the limit of a sequence of discrete probability distributions. Suppose we have a random variable, such as the length of a peacock's tail, which takes on a continuous value rather than a discrete one. How can we describe and work with the probability distribution for this random variable?

One good way to start would be by "binning" the possible values of the random variable X. For example, we might bin the distribution into ten centimeter intervals; we can then construction a probability distribution for a binned random variable Y where $Y = 5$ means that the tail is than 0–10 cm long, $Y = 15$ means that the tail is 10–20 cm long, $Y = 25$ means that the tail is 20-30 cm long, etc. This procedure would give us a discrete distribution such as that shown below, and we know how to work with these.



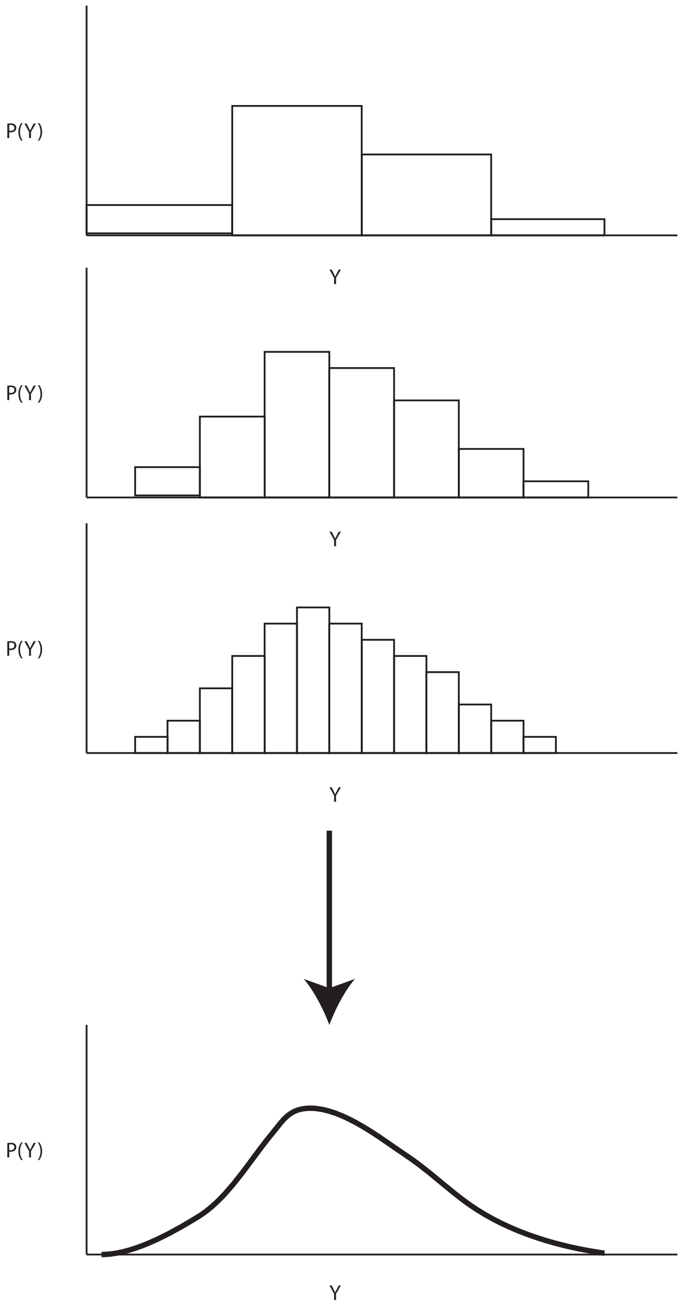
More formally, we break the distribution up into n bins of width Δ, and we consider the new discrete random variable Y to be the value of the bin into which the value of X falls:

$$Y = x_i, \text{ if } i\Delta \leq X < (i+1)\Delta$$

By the mean value theorem, we can thus select a set of values x_i such that $i\Delta < x_i < (i+1)\Delta$ and the probability distribution of Y is given by

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

Now we can formally see the connection between this definition and the idea of continuous distributions as the limit of n-division discrete spaces. We take the conventional Riemann integration view of the integral as the limit of a sequence of discrete sums over smaller and smaller subdivisions.



Our definitions of probability density, expectation, and variance for continuous distributions then following naturally.

Definition 5 *The probability density function or PDF $f(x)$ for a continuous random variable x is defined such that:*

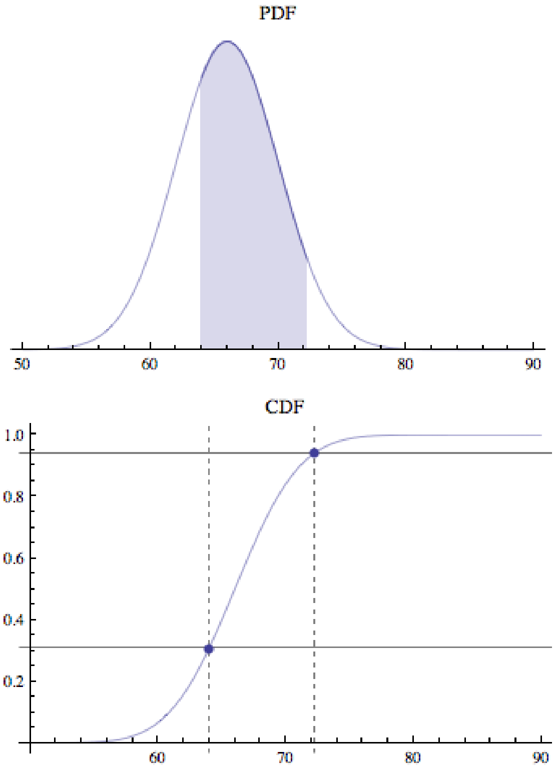
$$P[a \leq X \leq b] = \int_a^b f(x)dx$$

Here we must have $\int_{-\infty}^{\infty} f(x)dx = 1$ and $f(x) \geq 0$ for all x.

Definition 6 *The cumulative density function or CDF $F(x)$ for a continuous random variable x is defined such that*

$$F(x) = P[x \geq X] = \int_{-\infty}^{a} f(x)dx$$

Thus the CDF is the integral of the PDF, as illustrated below[2].



[2] This image is a snapshot from the demonstration "Connecting the CDF and the PDF". This is part of the The Wolfram Demonstrations Project, online at `http://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/`

Definition 7 *The expectation of a continuous random variable x with density function $f(x)$ is*

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

Definition 8 *The variance of a continuous random variable x with density function $f(x)$ is*

$$Var[X] = E[X^2] - E[X]^2$$

We can also develop an analogous set of definitions for joint probability distributions for continuous random variables.

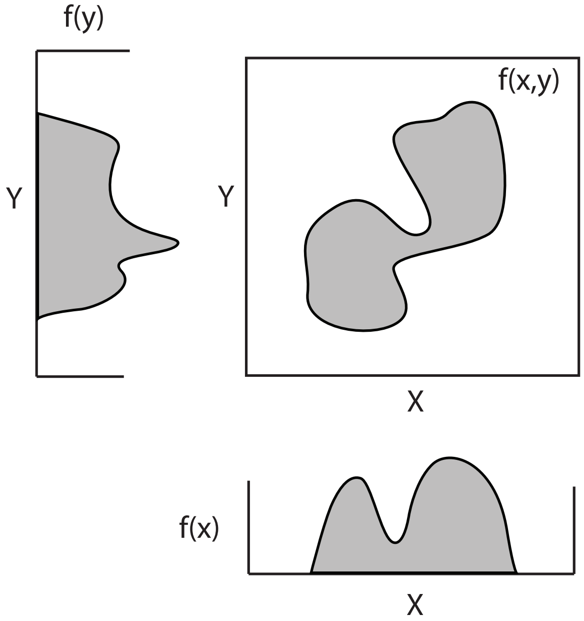
Definition 9 *The joint probability density function $f(x, y)$ for a continuous random variables x and y is defined such that:*

$$P[a \leq X \leq b, c \leq Y \leq d] = \int_{b}^{a} \int_{d}^{c} f(x, y)\, dy\, dx$$

Here we must have $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = 1$ and $f(x, y) \geq 0$ for all x and y.

If anything, marginal distributions are easier to understand for continuous probability distributions than for discrete ones.

Think of the marginal distributions of a jointly distributed random variable (X, Y) as the distribution of X (whatever Y is) and the distribution of Y (whatever X is).



That is,

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx
\end{aligned}
$$

Another way to envision this is to envision shaping the joint probability distribution out of play-doh. Taking the marginal distribution of x then corresponds to collapsing the y axis by squishing the playdoh along the y axis, e.g. between two books.

If X and Y are independent random variables, their joint distribution is equal to the product of the marginals:

$$f(x,y) = f_X(y)f_Y(y).$$

A Reference: Discrete distributions

A.1 Bernoulli Distribution

The Bernoulli distribution takes only two possible values, 0 or 1, with probabilities p and $1-p$ respectively. The mean of this distribution is $E[x] = p(1) + (1-p)0 = p$ and the variance is $E[X - E[X]]^2 = E[X-p]^2 = p(1-p)^2 + (1-p)(0-p)^2 = p(1-2p+p^2) + (1-p)p^2 = p - 2p^2 + p^3 + p^2 - p^3 = p - p^2 = p(1-p)$. We'll often call the event $X = 1$ a "success" and $X = 0$ a "failure."

A.2 Binomial Distribution

Suppose a lynx chases $n = 20$ hares in a day, and during each chase the probability of success (for the lynx, not the hare!) is $p = 0.1$. How many hares does the lynx catch on average? What is the probability that the lynx goes hungry on any given day? What is the variance in the number of catches?

The binomial distribution $B(n, p)$ represents the outcome of n trials each of which is an independent, identically distributed (iid) Bernoulli random variable with probability p of success (i.e., taking value 1 instead of value 0). The distribution function is for $X \sim B(n, p)$ is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

This is because the probability of getting any specific sequence of k success and $n-k$ failures is $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ different sequences with k success and $n-k$ failures.

The mean of this distribution is np and the variance is $np(1-p)$.

A.3 Geometric Distribution

Let us again consider our hunting lynx, who chases $n = 20$ hares a day and catches each with $p = 0.1$. Instead of asking about the distribution of catches in a given day, suppose we want to model the failed chases that the lynx attempts before the next catch.

Notice that thus far, we've looked at discrete random variables with a finite number of possible outcomes. Here, for the first time, we are looking at a discrete random variable with an infinite number of possible outcomes. The lynx could fail 1, 2, 3, . . . or any number of times before the next success.

Here we turn to the geometric distribution, which gives us the average number of failures before the first success. (According to Taylor and Karlin, so I'll use that formulation. More often, I see the geometric distribution defined as giving the average weighting time until the n-th success, which is simply the Taylor and Karlin value plus 1. But we'll use the Taylor and Karlin definition here.

The distribution function for $X\ G(p)$ is given by

$$p(k) = (1 - p)^k \, p$$

The first term is the probability of k consecutive failures; the second term is the probability of success on the $k + 1$-th trial. The expected value of X is $\frac{1-p}{p}$ and the variance is $\frac{1-p}{p^2}$. The tail probability is the probability that more than x consecutive trials fail, i.e., $P[X > x] = (1 - p)^{x+1}$.

So the average number of failures for the lynx is 9, the variance is 90. The probability that the lynx goes at least 20 trials without catching anything is $.9^{20}$.

A.4 Negative Binomial

This distribution tells us how many failures we see before the r-th success. The easiest way to think about the negative binomial is as the sum of r iid random variables each drawn from the geometric distribution.

The expectation and variance then follow immediately, given our previous rules about additivity of expectations and, for independent random variables, additivity of variances as well. The expectation is simply $r(1-p)/p$ and the variance is $r(1-p)/p^2$. The distribution function gives the probability of seeing exactly k failures before the $r-th$ success:

$$p(k) = p^r(1-p)^k \binom{k+r-1}{r-1}$$

Exercise: why?

A.5 Hypergeometric

Suppose that you have a population of n individuals, of whom k are of one type and $n-k$ are of a second type. If you sample j individuals from this population, how many of the first type will you get? This is closely related to the binomial distribution — but here we are sampling from a fixed population instead of a fixed frequency distribution. Another way to think about this is that in the hypergeometric distribution, we are sampling for a fixed population without replacement whereas in the binomial distribution we are sampling from a fixed population with replacement.

The distribution function for the hypergeometric gives us the probability of getting m samples of the first type:

$$P(m) = \frac{\binom{k}{m}\binom{n-k}{j-m}}{\binom{n}{j}}$$

Note that this is really just a formula in combinatorics. We write out the full set of equally likely members of the ensemble and look at how probable different combinations of these are; we do not even assign a probability parameter anywhere.

The expectation is $j(k/n)$ and the variance is $j(k/n)(1-k/n)\sqrt{\frac{n-j}{n-1}}$

A.6 Poisson

We can think of the Poisson distribution gives the distribution of successes in a large number of samples when successes are rare events.

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The expectation is λ and the variance is also λ.

We can view the probability distribution for a Poisson distribution with parameter λ as the limiting case of the binomial distribution with $np = \lambda$, as n gets very large and p gets very small.

A.7 Multinomial distribution

In the binomial distribution, you have two possible outcomes: success, or failure. Suppose that instead we are looking at a case where you have k different possible outcomes. For example, when I work on methylation patterns, a double-stranded CpG site can be methylated on both sides, unmethylated on both strands, or "hemimethylated", i.e. methylated on one side but not the other. If the probabilities of these three states are p_1, p_2, and p_3 respectively, what is the distribution function for the number sites of each type in a sample of n sites?

$$Pr(X = \{x_1, x_2, n - x_1 - x_2\}) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} p_3^{1-x_1-x_2}$$

The expectation is $\{np_1, np_2, np_3\}$ and the variance is $\{np_1(1-p_1), np_2(1-p_2), np_3(1-p_3)\}$. Of course this generalizes to more than three outcomes.

B Reference: Continuous Probability Distributions

Uniform Distribution

One of the simplest continuous distributions is the uniform, in which the random variable is equally likely to be anywhere within a fixed interval $[a, b]$. The density function is

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

The cumulative density function is

$$F(x) = \frac{x-a}{b-a}$$

The expectation is the midway point $(a + b)/2$.

Normal Distribution

The normal distribution is one of the most important distributions in statistics, because of the Central Limit Theorem.

Theorem 2 *Central Limit Theorem. Let Y be the sum of n independent identically distributed random variables X_i with expectation μ and variance σ^2. As n gets large, the distribution of Y approaches a normal distribution with mean $n\mu$ and variance $n\sigma^2$.*

Thus any random variable that we can view as the sum of a large number of iid elements should take on a normal distribution. Similarly, the log of any random variable that we can view as the product of a large number of iid elements should take on a normal distribution, because the log of the product is equal to the sum of the logs.

The density function for the normal $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

We cannot compute an explicit form for the cdf, because we cannot integrate $e^{-\frac{1}{2}x^2}$.

We can also view the normal distribution as a continuous analogue of the binomial distribution. For large n, the binomial with n trials each with probability p is approximated by the normal distribution with mean $n\,p$ and variance $n\,p\,(1-p)$.

Exponential distribution

The exponential distribution provides the distribution of waiting times until an event that occurs with a constant hazard rate λ; we can see it as the continuous time analogue of the geometric distribution. On its range $[0, \infty)$, the distribution function is

$$f(x) = \lambda e^{-\lambda t}$$

On the same range, the cumulative density function is

$$F(x) = 1 - e^{-\lambda t}$$

The expectation is $1/\lambda$ and the variance is $1/\lambda^2$.

Survival given a general expression for hazard

The exponential corresponds to a constant failure rate. If the failure rate varies over time as $r(t)$, we can still write down a density function as

$$f(x) = r(x)e^{-\int_0^x r(t)dt}$$

The corresponding cumulative distribution function is

$$F(x) = 1 - e^{-\int_0^x r(t)dt}.$$

Gamma distribution

The sum of k iid exponential random variables with rate parameter λ is a Gamma distribution; this can be seen as the waiting time until the k-th event under a constant hazard model, and thus analogous to the negative binomial distribution from discrete probability.

Its density function is

$$f(x) = \frac{\lambda}{(k-1)!}(\lambda x)^{k-1}e^{-\lambda x}$$

Notice that the gamma distribution $(1, \lambda)$ is simply the exponential with parameter λ.

The cdf is

$$F(x) = 1 - e^{-\lambda\, x}\sum_{k=0}^{r-1}\frac{(\lambda x)^k}{k!}$$

The above formulae hold only for integer k, though we can generalize this by replacing $(k-1)!$ which requires integer k with the gamma function $\Lambda(k)$ which does not. We cannot then compute the cdf explicitly, however.

As expected given that the gamma is the sum of iid exponentials, its mean and variance are k/λ and k/λ^2.

Convolutions

Often in dealing with random processes, we will be interested in the sum of independent random variables X and Y with probability densities of $f(x)$ and $g(y)$ respectively. If $Z = X + Y$, the density $h(z)$ of Z is given by the convolution integral

$$h(z) = \int_{-\infty}^{\infty} f(z-t)g(t)dt$$

The idea behind this integral is that it sums the various ways that Z could equal any particular value z: it could do so because $X = z - t$ and $Y = t$, for any value of t. So we integrate the product of the densities of X and Y with respect to these quantities, over all possible values of t. (We can take the product because we've assumed that X and Y are independent.)

We can also take a convolution of k random variables, by integrating over $k-1$ indices t_i:

$$h(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f_1(t_1)f_2(t_2)\dots f_k(z-t_1-t_2\dots-t_{k-1})(t)dt_1dt_2\dots dt_{k-1}$$

In this vein, the gamma distribution with shape parameter k is the k-fold convolution of the exponential. Also notice that the sum of two gamma distributions with shape parameters k_1 and k_2 and the same rate parameter λ is the gamma with shape parameter k_1+k_2 and rate parameter λ; thus the sum of two gammas is gamma so long as they have the same rate parameter, irrespective of their shape parameter.

Changes of variable

Suppose that we want to rescale the x-axis by some scaling function g, such that $Y = g(X)$. (We assume g is a proper rescaling, i.e., it is monotone increasing and differentiable.) Notice that — because the integral of the pdf must be 1 — the value of the pdf of a random variable changes when we change the scaling of the x axis. The density function for Y can be written as

$$f_Y(y) = \frac{f_X(x)}{g'(x)} = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}.$$

The value of the cdf does not to be corrected accordingly:

$$F_Y(y) = F_X(g^{-1}(y))$$

Recognizing this, we often do well to work with cdfs instead of pdfs when we are changing variables or rescaling the x-axis.