**Probability Distributions**

Just going to throw some general stuff in here.

**Single variable probability distribution**

Generally just something like p(x). If we take repeated measurements of x, and get a list of data, say, xi. Then we can create a histogram with x on the ‘x’ axis, and probability density of that x, p(x), on the y-axis. The probability density would be p(x) = number(x, x+dx)/(N∙dx), i.e., the number of elements, xi, within the interval (x, x+dx), divided by the total number of elements, N, divided by the interval/bin-size dx.

**Changing variables**

Consider two coordinate systems ua, and uα, with their respective metrics, measures, probability distributions dP, probability distribution functions P, and probability densities p. We can say:



So the relationship between their probability distribution functions is:



Note |Xαa| is none other than the Jacobian, J.

**Probability distribution of a statistic**

Suppose we have a statistic s = s(u), and we want to know it’s probability distribution function. We can just form,



We can also do this through the moment generating function. Consider:



So the probability distribution of a statistic is just the inverse Laplace transform of its moment generating function. This way of getting P(s) is often much easier than directly evaluating it with delta functions. The same could be done with the characteristic function. Note we can also get moments of a statistic from the moment generating function per usual. And if we want to get the probability distribution of some other set of variables we can modify the method, by putting in two delta functions.

**Moment and Cumulant Generating Functions**

Let X be a random variable. The moment generating function is:



There is generally speaking a 1-1 correspondance between probability distributions and moment generating functions. Note that if we make the replacement *t* → -*t*, or *t* → i*t* then the moment generating function is simply the Laplace or Fourier transform of the probability distribution. And thus, having M(t) we can go back and get p(x) with the appropriate inverse transform. This is at least a plausible basis for the 1-1 property. Relatedly, the Characteristic function is defined as:



And the Taylor series expansion of M(t) gives us the moments sans n! stuff.



The cumulant generating function is defined as:



and its Taylor series expansion gives us the cumulants,



Or in other words:



Working out a few terms, we find:



Might note the following property,



Suppose we have an M(t) which agrees with another M’(t) for small t. What does that mean about the respective p.d.? Well, it at least implies that they’ll have approximately the same mean and variance. The further out in t that they agree, the greater number of moments will agree.

**Chebycheff’s Theorem**

Chebycheff’s theorem gives us lower bound for the probability a random variable will be k standard deviations from its mean. So this applies to any random variable. It says,



I wonder what is the distribution which would satisfy the equality?

**Multivariable Probability Distributions**

For a two-variable distribution, say, we’d have a list of *paired* data (xi, yi), i = 1,2,3,….,N. And we’d create a histogram, dividing up x-y space into bins (x, x+dx)(y, y+dy). We’d count up the number of pairs (xi, yi) within the interval, and then divide by the total number of pairs, and by the bin area dxdy. So in other words p(x,y) = number[(x, x+dx)∩(y, y+dy)]/[N∙dx∙dy]. The pairs matter. We cannot separately shuffle the x data or the y data without changing p(x,y). I’ve got a few other questions. Say have two independent sets of data. For instance, we measure the heights of wolves, x, and then the heights of giraffes, y. We want the joint probability distribution p(x,y). How would this be constructed? I think we would take our list of wolf heights, xi=1,2,…M, and our list of giraffe heights, yj=1,2…,N, and we’d form every pair (xi, yj) that we can – M×N pairs in total. This would be the so-called ‘*event space*’, i.e., the number of ways we can draw a pair of heights from wolf bag and giraffe bag. And then we construct a histogram of the pairs as described above. Since the events are independent, we should find p(x,y) = p(x)p(y). Say we then wanted to construct p(x-y). How would we do this? Well, then I think we would take every one of the M×N pairs that we had constructed (xi, yj), and we’d subtract them to get dij = (xi – yj), and then construct a histogram of all the M×N differences.

Suppose we have a probability distribution, p(x,y). This is defined to be the probability of the intersection of the events x and y. So,



So it’s the probability of x and y occuring together. The probability of x irrespective of y would be given by,



The marginal probability distribution: p(x|y) would be given by the relation,



Probability of x or y would be:



etc. Note,



**Changing variables**

Consider two coordinate systems ua, and uα, with their respective metrics, measures, probability distributions dP, probability distribution functions P, and probability densities p. We can say:



So the relationship between their probability distribution functions is:



Note det(Xαa) is none other than the Jacobian, J.

**Probability distribution of a statistic**

Suppose two events have a joint probability distribution p(x,y). Then if we wish to calculate the probability distribution of a certain statistic, s(x,y), we could proceed in the following manner. We can just insert a delta function. For instance, suppose x and y are Cartesian coordinates, and we wanted to get the probability distribution of the distance from the origin, R. Then we’d do:



But note we wouldn’t have gotten the right answer if we had formulated the δ function constraint thing differently.



Yeah nope. This isn’t normalized. And its missing stuff, like a factor of R. So how do you know how to formulate the δ function? We should use the one for which we recover identity:



So basically, it needs to be in the aforementioned form δ(s-s(x,y)). Let’s consider another one. Let’s go backwards and get P(x) from our P(r,θ).



So that checks out. So to generalize, if you have a probability distribution P(uα) and you want the probability distribution of s = s(uα), then I guess constraints need take the form δ(s - s(uα)), i.e.,



Note we could use the delta function method to go all the back to P(x,y) if we want. You just have to put 2 δ’s in there. So let’s take our P(r,θ) and get P(x,y). We already know we should get the product of two Gaussians.



So there we go! Could construct weird multi-variable distributions like P(x,ξ = x2), and would also put two δ’s in there: δ(x-rcosθ)δ(ξ – r2cos2θ). This should give you,



So generalizing a bit, say we had ua = fa(uα). Then,



which is well, true I guess. Could also do:



So it all works out the same. Here’s another unusual case, which did show up in physics when we were working out the probability distribution of the scattering angle. So say we had a particle constrained along a path (x(s), y(s)) and it had a probability distribution P(s) (wouldn’t interpret s as time but rather arc length or something). And we’d want the probability distribution of x, and y, rather. What would that be? Well, it’d be:



Now consider in general how the *moment generating function* method can be used to calculate the p.d. of a statistic. Consider a statistic s = s(uα) of a multi-dimensional probability distribution. Then,



So the probability distribution of a statistic is just the inverse Laplace transform of its moment generating function. This way of getting P(s) is often much easier than directly evaluating it with delta functions. The same could be done with the characteristic function. Note we can also get moments of a statistic from the moment generating function per usual. And if we want to get the probability distribution of some other set of variables we can modify the method, by putting in two delta functions.



So its just the multi-variate inverse Laplace transform of the characteristic function of the new variables.

**Moment and Cumulant Generating Functions**

We can generalize the concept of the moment generating function to multi-variable distributions.



Derivatives w/r to tj would give us the moment of Xj. Note that if the variables are independent, then we have,



So the multivariate moment generating functions separates into the product of each variables separate moment generating function. The cumulant generating function is defined as before,



For the sum of two independent variables, the moment and cumulant generating functions are:



This latter property is extremely useful, as it implies that the cumulants of N independent variables simply equals the sum of the N respective cumulants. Further, suppose we looked for a linear combination of two independent variables:



So for instance, for the first, second and third cumulants, we’d have:



**Some General Results for Multivariable Distributions**

Suppose that we have n independently distributed random variables, with means μi and std. devs. σi. Then



and,



and so,



It seems to be a general rule that for independent variables cumulants are additive – well we showed it was under the moment generating function section. Covariance of two random variables is defined as,



Can see that if the variables are independent, then Cov(X,Y) = 0. If two random variables have non-zero covariance, then their cumulants are not additive. For instance, revisiting our expression above,



So,



**Central Limit Theorem**

Suppose that we have n identically distributed random variables with mean μ and std. dev. σ, each following an otherwise undetermined distribution function. Then consider the average statistic:



From the discussion in the moment generating function/cumulant generating function section, we know its cumulants will be:



etc. One can say that the successive moments become irrelevant. If we keep just the first, then we get a δ function. If we keep the second too, then we get a Gaussian. A question is, suppose we just define X = ΣXi. Then can we conclude anything? Do we see a well defined limit? Judging by other examples I’ve worked out, like the stochastic product example Xm+1 = XmWm, (take the ln of both sides and we get a sum of independent variables thing) it seems that the sum of independent variables will converge to a Gaussian distribution, for a given n, but that the mean (of course), and std (of course) increase with n. So the limit isn’t n-independent per seʹ, so there isn’t really a limit per seʹ, and it kind of diverges, but the Gaussian approximation becomes better and better with larger n. I suppose this is why, to be mathematically precise, one formulates the CLT in terms of the statistic ΔX = (X1 + X2 + … + Xn - nXave)/√n.