**Probability Distributions**

Now going to consider multivariable probability distributions.

**Multinomial distribution**

Suppose we conduct an experiment with 3 possible outcomes – a, b, c – with probabilities p1, p2, p3, which must add up to 1 of course, so p3 = 1-p1-p2 (hopefully, generalization to arbitrary number of outcomes will be straightforward). An example would be tossing a dice. There are six possible outcomes for each throw, each equiprobable. Then say we conduct n trials. Well, in any particular string of n experiments, the outcome will look like,



a total of n outcomes. The probability of this occurrence is equal to p1n1p2n2p3n3, where n1 is the number of p1’s, n2 the number of p2’s, and n3 = n-n1-n­2 the number of p3’s. Now the number of experiments that we can perform where we have this same number of p1’s, p2’s, and p3’s is equal to the number of distinct permutations of this series. That number is Cnn1,n2,n3. Therefore the probability of n1, n2, n3 outcomes of p1, p2 and p3 would be:



Let n1 = x1, n2 = x2, and n3 = x3 = n – n1 – n2. The moment generating function is:



So,



We can get various expectations from this,



and second power,



And variances,



And let’s do the covariance too,



So altogether, we have:



Why is the covariance negative? Basically because the larger one outcome is, the smaller, generally, the other must be, so as to keep the sum constraint satisfied.

***Gaussian Approximation***

Could attempt a Gaussian approximation. I’ll start from the moment generating function.



I’ll take a ln of both sides, and expand out to 2nd order in t1, t2,



Then exponentiating,



This looks exactly like our Gaussian moment generating function with the means and variances, covariances given by our multinomial distributions corresponding values (see below apropos two-variable Gaussian distribution). So we can say,



where,



Moving on,

**Multinomial Z2 distribution**

Suppose we have X = multinomial data, i.e., we have a category with k = 1, 2, 3, …, kmax outcomes. And a table listing the outcomes of all our measurements. Could be which side of dice came up, or what age bracket a person is in, or political affiliation, or letter grade, etc. Let nk be the number in each category, and nk(pred) the predicted number in each category, according to some hypothetical set of p’s. So nk(pred) would just be nk(pred) = npk. We might wish to determine whether our data supports the predictions. Accordingly, we can formulate a Z2 statistic, given below, which is kind of analogous to the Z-statistic of the Gaussian distribution (squared):



Turns out Z2 follows a χ2 distribution with ν d.o.f., where ν = # of independent p’s in the model (see previous file for demonstration of this fact for simpler case of k = 2). Letting x = Z2, we have:



For the least restrictive multinomial model, the number of independent p’s will be kmax – 1 (the last p is fixed because all p’s must add to 1). FWIW, it looks like Z2 will follow a χ2 distribution only *approximately*. As long as nk > 5 for every category, then we should be good. We can ascribe this to the fact that for a binomial distribution (or multinomial distribution) to follow the Gaussian approximation, we need the pi’s to be not too small, and the n to be fairly large.

**Sum of Poisson distributed variables**

Apparently the sum of 2 Poisson distributed variables, characterized by λ1 and λ2, is also a Poisson distribution, characterized by λ = λ1 + λ2.

**N dimensional Gaussian distribution**

Let’s consider the generalization of a 1D Gaussian distribution to arbitrary N dimensions. Actually we’ll drop down to 2 dimensions, but we’ll frame the discussion in such a way that generalization to N dimensions will be transparent. Something like this (most easily written in matrix form):



where we’ll observe the exponent looks like this:



and N (not the dimension N, a normalization N, rather) is the as-yet unknown normalization factor. The matrix A is the most general positive semidefinite matrix in 2 dimensions, as long as a > 0, c > 0, and ac – b2 > 0. We need the matrix to be positive sem-definite in order for P(x1,x2) to go to 0, not ∞, as x1 and x2 get large. What’s the normalization? We want to evaluate,



It helps to change variables to work out this integral. First, we can just shift variables by the means, without changing anything,



Then we’ll find eigenvalues/vectors, of A. Eigenvalues are:



and eigenvectors are:



Form the unitary matrix U, and eigenvalue matrix Λ,



Then we can write,



Now change variables to u = (u1 u2), given by:



and we can write,



Last, we need the Jacobian,



So our integral becomes,



Also note that from the property of determinants, we can say,



So we can say,



And therefore, the normalized probability distribution is:



The moment generating function is given by,



To work this out, we’ll shift variables again,



And we’ll change variables, like we did before,



recalling the Jacobian of the transformation was 1,



Then we use the identity,



to conclude,



Now we can simplify the exponent.



And come to:



Clearly the formula for P and M can be generalized to arbitrary dimensions, beyond just two. But explicitly, for our two dimensional case, we have:



where A-1 is given by:



So what are averages?



This is reassuring. What about the 2nd moment?



Therefore, the variances are given by:



(this would’ve been more easily obtained by taking derivatives of the cumulant generating function K = ln(M)) And last, let’s get the covariance,



So,



Summarizing our results,



Let’s introduce the covariance matrix:



We can see that when i = j, we just get the regular variance. So from our results above, we can that:



So



Before moving on, it might be nice to put A in terms of (σ2)-1.



So anyway, given this, we could rewrite our probability distribution more transparently,



where we use the property that det(A) = 1/det(A-1) for any matrix A, and that det(aA) = adim(A)A, where *a* is a constant and dim(A) is the dimension of A. So we can say,



where σ2 is the covariance matrix. And we could update our moment generating function to,



and our expectations to:



These expressions compare well with our 1D Gaussian distribution results. Basically then, to go from 1D to any D, we just have to update our x to a column vector x, and our average μ to a column vector μ, and our variance σ2 to a matrix of covariances σ2. By the way, I guess we could say the total variance is:



**Wick’s Theorem**

What is something like, <(x1-μ1)(x2-μ2)(x3-μ3)(x4-μ4)>? Borrowing from the path integral stuff, this should be:



Basically, the sum of the propagators. Let’s consider,



And look at <x1x2x3>. It’s clearly zero, but



because matrix is symmetric. Now next derivative:



And now last derivative,



Now setting t = 0,



So it looks like we can break down all higher order correlations into products of one and two coordinate correlations. Okay, now let’s do the fourth moment,



Now set t = 0,



which is again, the sum over all possible products of partitions into single and double point correlations.

**Example**

Let’s specialize to the bivariate distribution. Then,



Let’s write this out in terms of the σ1,2 and r = Cov(x1,x2)/σ1σ2. Well,



So,



And now introducing the correlation coefficient, r,



And now using more typical notation,



Can see that as r 🡪 1 (perfect correlation), the distribution gets narrower and narrower. I guess it would basically go to a diagonal delta function.

**Example**

Two sports teams’ scores have been tabulated for over 100 years, and a statistical analysis has been performed on them. Both teams’ scores are approximately normally distributed. Throughout the years, the Robbins have had a mean score of μ1 = 95, while the Blue Jays have had a mean score of μ2 = 92. The variances of their scores are σ21 = 100 and σ22 = 81. Their correlation coefficient is r = 0.87. What is the joint probability distribution of their scores? What is the probability that the Blue Jays will win any given game?



Want to calculate P(x2 > x1). This is:



We can only do this numerically, as far as I’m aware. Let’s simplify some more by scaling each coordinate by 10.



**Example**

Two ants roam along a sidewalk crack in search of food. Each’s position is described via:



What is their joint probability distribution function?



What is the probability density of their being an ant at position x?



where the last line presumes the probability distribution is symmetric.