Some notes on different families of distributions, related to the Poisson.

Assistant Professor
Dept of Sociology
Stanford University
http://www.stanford.edu/~mrosenfe

Revision 11/26/2005

1) Poisson:

Formal Definition of Poisson(λ)

Mean=λ
Variance=λ
Std Deviation = $\sqrt{\lambda}$

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (1.1)$$

for k=0,1,2...

Intuitive explanation:

a) The Poisson distribution has one parameter, λ. This parameter describes the expected mean number of events. Let's say we have a hundred apple trees, all identical. During harvest season, we put a basket under each tree and we catch a few apples a day. Let's say the average is 2 apples per tree. The Poisson distribution tells us how many of the baskets we should expect to have exactly 2 apples each day, how many will have 0, 1, 2, 3, 4, and so on. As long as each tree is identical, and as long as each apple's fall into the basket is a separate and independent event (i.e. nobody shakes the branches to make a lot of apples fall at once), then the Poisson distribution will do a good job of describing the distribution of apples in the baskets. If we expected every tree to produce the same number of apples every day, our model would only need a single parameter. If, on the other hand, the trees in the last row get more sun and produce more apples, we should add a 'last row' parameter into the model, and test whether this additional term makes a significant improvement in the goodness of fit of the model.

A process that may more easily fit the model is the capture of neutrinos (a tiny subatomic particle) from the sun in a special detector. Every day the sun emits the same number of neutrinos, and each neutrino has a small but constant chance of hitting the detector on your desk. If the average count is 2, the Poisson distribution tells you how likely you are to catch 10 (or 2, or 1) on any given day. If a single day's neutrinos have a Poisson(2) distribution, it seems reasonable that if you checked the machine every other

day, you would have a Poisson(4) distribution. So, as long as the events are independent, Poisson(A)+ Poisson (B)=Poisson(A+B).

b) See my Poisson distribution examples.

c) Also notice, in the examples I demonstrate above, that Poisson(15) looks a lot like a Normal distribution, but Poisson(.1) and Poisson(2) do not. Why is this? For one thing, the Poisson distribution only allows nonnegative values (the smallest number of apples or neutrinos you can get is zero), whereas the Normal distribution does not discriminate against our friends the negatives. When the mean is close to zero, the Poisson distributions are more obviously skewed. When the mean is large, say 15, the Poisson(15) distribution looks fairly symmetrical because the mean is large and values as far away as zero are implausible anyway. Another reason why Poisson(15) looks like a Normal distribution with mean 15, is because of the Central Limit Theorem. One simple version of the Central Limit Theorem says:

If variable X has mean μ and variance σ^2, and $S_n = \sum_{i=1}^{n} X_i$

where the X_i are independent and identically distributed,

then $\lim_{n\to\infty}$ of S_n is $N(n\mu, n\sigma^2)$.

What this means is simply that if you take (almost) any old kind of variable X, and you take enough samples of it, and you add them up (or, more commonly, take their average), you end up with something that has a Normal distribution. The Central Limit Theorem in its various guises is of Central importance, hence the name. Think of Poisson(15) as Poisson(1)+Poisson(1), each independent,..15 times. The Central Limit Theorem suggests that Poisson(15) should start to look a lot like Normal (15,15). Of course, one could also note that Poisson(2), which looks very non-Normal is really just 15 separate and independent instances of Poisson(2/15) summed together. So why doesn't Poisson(2) look Normal? Well, the Central Limit Theorem says as n goes to infinity, S_n will be Normally distributed. It doesn't say how fast you'll get there. If you start with Poisson(2/15), you need to add together a lot more than 15 independent copies to get to something that looks like a Normal distribution.

d) Based on my description above, in part (a), you can see that the Poisson distribution is the summary results of many individual events at the apple or neutrino level. You can think of each apple having a fixed chance of falling into a basket, and each neutrino of having a fixed chance of falling into the detector. At the level of the apple or neutrino, we are dealing with a family of distributions called the binomials (binomial because each apple has 2 choices- in the basket or not, and each neutrino has 2 choices- hit the detector or not). To be very specific, the Poisson distribution is a special case of the Negative Binomial distribution, and Poisson regression is a special case of negative binomial regression. Whereas the Poisson distribution has only one parameter that fixes both the mean and variance (similar to the Chisquare but different from the Normal), the Negative Binomial distribution has more flexibility- the mean and variance

are determined by two separate parameters. Stata has a number of functions for negative binomial regression.

e) It is also worth noting, that if X is distributed as Poisson(λ), then the square root of X has an approximately Normal distribution, with constant variance. I'm not going to justify why this is, but I'll simply point out that when you want things to behave more Normally, you sometimes have to enforce a transformation. What a square root transformation does (see, again, my examples) is it reels in the high value outliers. When you have variables that take on only positive values, and have some high value outliers (income is one example, counts of events is another example), it is common to take the log or the square root of those variables in order to bring the high values down, and make the distribution less skewed.

2) Chisquare Distribution:

with n (integer) degrees of freedom,

$$f(x) = \frac{(1/2)^{n/2} x^{(n/2)-1} e^{-1/2x}}{\Gamma(n/2)} \quad \text{(2.1)}$$

for x≥0

Mean=n
Variance=2n
Standard Deviation $= \sqrt{2n}$

a) Unlike the Poisson distribution, which can take on only integer values or zero, the chisquare distribution has a range of all positive real numbers. The Greek letter Γ in the denominator is just Gamma, indicating the Gamma function. The Gamma function is nothing more than a fancy way of extending the factorial function to all the real numbers. Γ(x)= (x-1)!, when x is an integer. So if you think of Gamma as just an extended version of the factorial function, you'll see that the Chisquare distribution and the Poisson distribution have some similarities in the way their probability densities are defined. They also have similar shapes, see below.

b) $\chi^2(1)$, or the Chisquare distribution with one degree of freedom is defined as the square of a Standard Normal variable. In other words, if z has the familiar N(0,1) distribution whose cumulative distribution is the source of tables in the back of every statistics text book (i.e. Normal with mean of zero and variance of 1), and if $y=z^2$, then y has a $\chi^2(1)$ distribution. This also means that if you have a statistic expressed a value from a $\chi^2(1)$ distribution, you can take the square root and you will have the familiar z-

score. When switching back and forth from $\chi^2(1)$ and N(0,1) you do have to keep in mind that the Normal distribution has two tails (in the positive and negative directions), whereas as the Chisquare distribution only has the one tail.

c) Under independence, $\chi^2(a)+ \chi^2(b)= \chi^2(a+b)$. Another way to look at this is that $\chi^2(a)= \chi^2(1)+ \chi^2(1)+ \chi^2(1)+....$ a times (with each component being independent). Given what we know about the Central Limit Theorem, you would expect that $\chi^2(n)$ would look more and more like the Normal distribution, the larger n gets (since $\chi^2(n)$ is just n combinations of independent $\chi^2(1)$ variables). The examples of the chisquare distribution will verify that that $\chi^2(16)$ looks quite Normal (and in this case it approximates N(16,32)). Also note that, comparing the $\chi^2(n)$ and Poisson(n) distributions when n is the same or similar, shows that the distributions have some similarity in shape, which is not surprising since their probability density functions have some similar elements, and since the Central Limit Theorem forces both distributions to be more Normal when n grows large. We do know that $\chi^2(n)$ has a variance of 2n, and Poisson(n) has a variance of n, so clearly the Chisquare distribution has longer tails.

d) One property of the Chisquare distribution we've used throughout the class is that if Model 1 has goodness of fit chisquare $\chi^2(n)=V$, and Model 2 adds m additional terms and has a goodness of fit Chisquare of $\chi^2(n-m)=U$, then the comparison of Model 1 and Model 2 is $\chi^2(m)=V-U$. We have also said that this comparison only works if Model 1 is nested within Model 2 (that is, if Model 2 contains Model 1).

3) Loglinear Models, Interaction Terms, and the Odds Ratio

a) It's not immediately obvious why the interaction terms from a loglinear model are, in fact, log odds ratios. So here's a simple illustration. Let's start with a simple 2x2 table, the basic unit of all comparisons.

a	b
c	d

The odds ratio is defined as the cross product

$$OR = ad/bc \tag{3.1}$$

And the log odds ratio is defined as

(3.2) $$\ln(OR) = \ln(ad/bc) = \ln(a) + \ln(d) - \ln(b) - \ln(c)$$

Where *ln* is the natural (i.e. base *e*) logarithm. The standard error of the log odds ratio is

(3.3) $$\text{Std Error of ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Note that however one orders the rows and columns of this 2x2 table (and for nominal categories the order is arbitrary), the only two log odds ratios one can generate are either *ln*(ad/bc) or *ln*(bc/ad), and these are additive inverses. The standard error is the same in either case.

Let's take a simple saturated loglinear model for this 2x2 table.

(3.4) $$\ln(U) = C + \lambda_{Row} + \lambda_{Col} + \lambda_{Interaction}$$

Where *U* are the predicted values of the model, C is the constant term, and the lambdas can be defined in a variety of equivalent ways. Let's show that the $\lambda_{Interaction}$ term will actually be the log odds ratio. There are lots of equivalent ways to set this up, but let's assume (as STATA does) that the first category for every categorical variable is the excluded comparison category. Then:

Constant term applies to all cells

1	1
1	1

λ_{Row}

0	0
1	1

λ_{Column}

0	1
0	1

$\lambda_{Interaction}$

0	0
0	1

Since this model has 4 terms, and the dataset has only 4 data points, the model will fit the data exactly.

Predicted values will be as follows:

$$\ln(a) = C \tag{3.5}$$

$$\ln(b) = C + \lambda_{Column} \tag{3.6}$$

$$\ln(c) = \mathrm{C} + \lambda_{\mathrm{Row}} \tag{3.7}$$

$$\ln(d) = \mathrm{C} + \lambda_{\mathrm{Row}} + \lambda_{\mathrm{Column}} + \lambda_{\mathrm{Interaction}} \tag{3.8}$$

Now let's take the formulas for the predicted values and re-arrange them, keeping in mind equation (3.2)

$$\begin{aligned} \ln(OR) &= \ln(a) + \ln(d) - \ln(b) - \ln(c) \\ &= C + C + \lambda_{Row} + \lambda_{Column} + \lambda_{Interaction} - C - \lambda_{Column} - C - \lambda_{Row} \\ &= \lambda_{Interaction} \end{aligned} \tag{3.9}$$

That's what we wanted to show.

Interpreting the coefficients of loglinear models using relative risk or incidence rate ratio.

1) Starting point.

Let's say we have a simple model,

1a) $Log(U) = \text{Const} + B_1 X_1 + B_2 X_2 + ...$

Where the B's are model coefficients, and the X's are the variables (usually dummy variables) and the U are predicted counts.

When $X_1=0$, we have:

1b) $Log(U) = \text{Const} + 0 + B_2 X_2 + ...$

and when $X_1=1$

We have

1c) $Log(U) = \text{Const} + B_1 + B_2 X_2 + ...$

So we can always say, as a simple function, that the coefficient B_1 represents an increase in the log of predicted counts. If $B_1=2$, for instance, we could say that 'this model shows that factor X_1 increases the predicted log count by 2 (all other factors held constant)' because equation (1c)- equation (1b)= B_1. This is true but not the most illuminating thing to say.

Remembering that $e^0=1$, we can exponentiate equation (1b) to get

1d) $U = e^{Const} e^{B_2 X_2}$

and when $X_1=1$, we can exponentiate equation (1c) to get

1e) $U = e^{Const} e^{B_1} e^{B_2 X_2}$

If we take the ratio of (1e)/(1d), we get e^{B1}. If for the sake of discussion we give B_1 the arbitrary value of 2, $e^2=7.4$, we could say that 'variable X_1 increases the predicted counts by a factor of 7.4' (all other factors being held constant). The cells that have $X_1=1$ have predicted counts, or an incidence rate ratio (using Stata's terminology) or a relative risk (using Agresti's terminology) of 7.4 compared to the cells that have $X_1=0$. Think of equation (1d) as the predicted value of events or counts for one cell (where $X_1=0$) in a big table, and equation (1e) as the predicted valued of events or counts for another cell where dummy variable X_1 takes on a value of 1.

---

2) See also my "Notes on different families of distributions" document for a section on why loglinear model coefficients are also (in some cases) log odds ratios.

A brief look at Likelihood Maximization with the Poisson Distribution.

Formal Definition of Poisson(λ)

Mean=λ
Variance=λ
$\text{Std Deviation} = \sqrt{\lambda}$

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.1}$$

for k=0,1,2...

As you will recall, the Poisson distribution has one parameter, lambda. If we know lambda, the Poisson distribution tells us the probability that the number of events we observe will be equal to k, for any k among the counting numbers. In theory, we know lambda and calculate the probability for X=k. In practice, we have a series of observations, X_1, X_2, X_3,...X_n and we want to estimate lambda from the observations. If we imagine, for instance, that each observation X_i corresponds to one day's number of neutrinos detected, or one basket's worth of apples, and if we further assume that the X's are independent and identically distributed, then intuition tells us that the best estimate for lambda would be the average of the X's, or

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.2}$$

Now let's show that this is in fact the maximum likelihood estimate as well.

The definition of the likelihood is the product of the individual likelihood functions, because the X's are independent, their joint probability is simply the product of the individual probabilities,

$$\text{Likelihood}(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} \tag{4.3}$$

In calculus, the way we find the lambda which maximizes this function is by taking its derivative (with respect to lambda) and then finding the lambda which makes the derivative zero, since the derivates are zero at the maxima (and at the minima as well). Unfortunately, the product of n functions can be overly complex to differentiate.

So what to do? The answer is that we take the natural logarithm of the likelihood function, which transforms the product of probability functions into a sum, and try to maximize that instead. This step relies on a property logarithms, namely that log(abc...n)=log(a)+log(b)+log(c)+...+log(n). This step also relies on knowing that the lambda which maximizes the log likelihood must be the same lambda that maximizes the likelihood. Why must this be? Because the log is a monotonic function, it has no maxima of its own.

For each single observation, the log likelihood is

$$\text{LogLikelihood}(\lambda) = k_i \log(\lambda) - \lambda - \log(k_i!) \tag{4.4}$$

For the product of the probabilities of all the observations together, the log likelihood is simply the sum:

$$\text{LogLikelihood}(\lambda) = \sum_{i=1}^{n} [k_i \log(\lambda) - \lambda - \log(k_i!)] \tag{4.5}$$

For those of you who are worried about finding the derivative of the factorial function don't worry. We will trying to find the lambda which maximizes the function, so we will be differentiating with respect to lambda, and anything that isn't a function of lambda, such as log(k!) will disappear.

We can group the log likelihood function into three separate terms:

$$\text{LogLikelihood}(\lambda) = \log(\lambda) \sum_{i=1}^{n} (k_i) - n\lambda - \sum_{i=1}^{n} (\log(k_i!)) \tag{4.6}$$

Now we take the derivative with respect to lambda, and set it equal to zero:

$$\frac{d(\text{LogLikelihood}(\lambda))}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} (X_i) - n = 0 \tag{4.7}$$

Which means

$$\sum_{i=1}^{n} (X_i) = \lambda n \tag{4.8}$$

And therefore

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}(X_i) \tag{4.9}$$

Which is what we wanted to show. How do we know this lambda will maximize rather than minimize the log likelihood (and therefore also the likelihood) function? We could graph the log likelihood function, or we can look at the second derivative.

$$\frac{d^2(\text{LogLikelihood}(\lambda))}{d\lambda^2} = \frac{-1}{\lambda^2}\sum_{i=1}^{n}(X_i) \tag{4.10}$$

Since lambda squared is always positive, and since the sum of X's must be positive, the second derivative is always negative (because of the factor of -1), so there can only be one maximum and no minima, so our lambda is the maximum likelihood solution.